# Learning Representation for Information Access

Benjamin Piwowarski

# Habilitation à Diriger des Recherches

spécialité **Informatique**

# Learning Representation for Information Access

*Defense date: October 23, 2020*

**Benjamin Piwowarski**
**CNRS, UMR 7606 (LIP6), Sorbonne Université**

**Reviewers**

Prof. Éric Gaussier (Université Grenoble-Alpes, France)

Prof. Jian-Yun Nie (University of Montreal, Canada)
Prof. Fabrizio Sebastiani (Consiglio Nazionale delle Ricerche, Italy)

**Examiners**

Prof. Eneko Agirre (University of the Basque Country, Spain)
Prof. Matthieu Cord (Sorbonne Université, France)
Prof. Mounia Lalmas (Spotify Research, UK)



l

# Abstract

The field of information access is of vital importance in our modern societies, since most of the information is now accessible in a digital form and is increasing in volume at a fast pace. Techniques from this domain allow to query this information and to access it in an appropriate form (e.g. summary, list of documents, etc).

Data representation for many different entities, such as a query from a user, the text of a document, an image, is key to the success of information access models based on machine learning techniques. Throughout the years, there has been a shift from hand-crafted models of data to automated methods for learning such an appropriate representation of data. The latter, i.e. the problem of how to represent *raw* data, has undergone a revolution in the last ten years, driven by deep learning. Such works have developed a series of models and techniques to represent complex data as vectors in a vector space, empowering the notion of distance/angle in such spaces to represent semantic relationships between the entities.

The work I present in this manuscript focuses on the problem of data representation in the context of information access. In particular, I present works dealing with (1) probabilistic representations of textual and graph data; and (2) the problem of grounding textual representation in the "real" world.

Le domaine de l'accès à l'information est d'une importance vitale dans nos sociétés modernes où la majeure partie de l'information est accessible sous forme digitale. La représentation de données (question d'un utilisateur, texte d'un document) est une clef du succès de l'ensemble des modèles basés sur des techniques d'apprentissage automatique. Le problème de la représentation de données brutes a subi une révolution ces dix dernières années, sous l'impulsion de l'apprentissage profond, en développant une série de modèles et techniques permettant de représenter des données complexes sous la forme d'éléments d'un espace vectoriel, se reposant sur le lien entre distances/angles dans l'espace vectoriel et les relations sémantiques qu'entretiennent les entités représentées. Ce manuscrit présente mes travaux dans le cadre de la représentation de données. En particulier,

1. La représentation de données, en utilisant des formalismes tels que les probabilités quantiques ou les distributions gaussiennes ;
2. L'ancrage du texte dans la réalité (ou tout du moins dans une réalité moins biaisée).

# Résumé

# Contents

# Chapter 1

# Introduction and organization

Given the ever increasing amount of stored digital information, techniques and models allowing users to *access* and/or *query* this information are of tremendous importance. The *information Access* research field precisely deal with models and techniques that can help a user to access/query digital information.

There exist many Information Access tasks. Among those, we can cite the following representative ones (those tackled in this manuscript are written in bold):

- **Information Retrieval** – retrieve information by finding the subset of documents relevant to a user information need (this includes interactive/discussion-based retrieval systems).
- Question Answering – Answer to specific questions, such as "What is the height of the Eiffel tower?", by retrieving the document containing the answer (optional), and then locating the answer within the document.
- **Recommendation** – Recommend items (e.g. movies) to users given their implicit or explicit past preferences.
- Document filtering – Filter a stream of documents that match a pre-specified criterion, to tame the incoming information flow (e.g. Twitter and Facebook are offering a personal view over the flow of produced information).
- Document Clustering – Cluster documents to explore the information space. While this task is not a main research topic nowadays, since the systems so far have not been adopted by end users, it might regain attention if the quality of the results and presentation improves, especially to support complex searches.
- Document Classification – Classify documents into a pre-defined topics, e.g. predict which inbox a mail should go in or which node in a classification ontology a web page should be placed in.
- **User Modeling** – User models allowing to predict their actions. This has many applications both for prognostic and diagnostic. For instance, with a prognostic view, web search users can be modeled so as to predict their behavior (and estimate their satisfaction); with a diagnostic view, the same model can be used to estimate the relevance of the document to a user information need given its behavior (i.e. clicks or non clicks).
- **Summarization** – Summarize one or more documents. The first practical applications of summarization were web search result snippets (so the user can estimate the relevance of the proposed documents), but in the future this might lead to more interesting applications such as summarizing a set of results.

- Translation – Translate text to allow users accessing information in foreign languages. This area has been transformed recently with deep learning, and has the potential of allowing people to search information in languages they do not know, or to present more diversified opinions around the globe.

All the above end user tasks have been studied, or are still studied nowadays, by the Information Access field, and the impact on society has been tremendous, especially with the rise of search engines in the 1990s and recommendation systems in the 2000s. This field has an even longer history that traces back to the first computers and the idea of digitizing information.

The way those tasks have been handled has evolved with time, from model-based models to representation-based learning ones. We can distinguish three phases in the evolution of models for Information Access:

1. Probabilistic and heuristic models (roughly 1960-2000);
2. Machine learning models (roughly 1990-2010);
3. Representation Learning (roughly 2010 to Today).

Below, we discuss briefly and illustrate each phase with examples from the Information Retrieval (IR) research field.

**Towards model-based** Spanning 1960–2000, heuristic or probabilistic models were developed with a task at hand. All these models rely on specific data representations, which are considered as mathematical objects from which the models can be expressed. Models are then expressed based on heuristics (e.g. the more a term occurs, the more it is representative of the topic of the document) or probabilistic modeling (e.g. the distribution of the number of occurrences of a term follows a Poisson law). In all these models, the underlying representation can be said to be "hidden away" within the theoretical or heuristic parts since the actual implemented model might use it or not.

***Illustration*** In information retrieval, Cleverdon experiments in the 60s have first shown that by using a simple bag-of-words representation, where each term corresponds to a dimension in a large vector space, automated systems were able to retrieve documents automatically given a user query (Cleverdon 1967). Salton, Wong, and C. S. Yang (1975) proposed heuristics to associate an importance with each word of a document. Finally, probabilistic models were developed to estimate the relevance of a document (within a dataset) for a given query – one of the most successful model being BM25 (Robertson and Walker 1994).

**Machine Learning** Spanning from 1990 to 2010, with the development of robust machine learning models like Support Vector Machines (SVMs), Conditional Random Fields or Gradient Boosted Trees (Murphy 2012), there has been a shift in the way of representing information: instead of pursuing ever-more sophisticated models for representing data, the goal is to produce a set of (hopefully) independent features *correlated* with the task. Good data representations should emphasize explanatory factors – at least for the task at hand. These features can then be used by a classifier that *learns* to combine the different features so as to take a decision.

***Illustration*** In IR, learning to rank models were initiated by (Fuhr and Buckley 1991), who suggested the use of a logistic classifier based on various features, among which classic relevance models such as BM25. This work has been followed by many others in IR – proposing pairwise and list-wise losses (T.-Y. Liu 2011). Most major Web search engines are based on these ideas nowadays.

**Representation Learning** Even though the question of representation has been outshone by learning to rank approaches, the question of how to represent data remains central since the model depends on these representations. The process of designing new features is inherently slow and costly. More importantly, it might hinder models from capturing important relationships between data and the task at hand: Going beyond handcrafted features is a natural evolution of learning to rank techniques. This evolution has been fostered those last ten years with the "rebirth" of neural networks, i.e. with *deep learning*.

Deep learning, and more generally representation learning models (Bengio, Courville, and Vincent 2014), aim at associating with any object described by raw features (e.g. pixel RGB values or simply one-hot encoding like for text) a latent representation in a (vector) space such as $\mathbb{R}^d$. Geometric relationships between entities encode existing relations or similarities of raw data.

Continuous representations of information are in the spotlight of many data-related communities (Goodfellow, Bengio, and Courville 2016) – image, text, and audio signals are processed and transformed automatically into high level continuous representations. A specific conference, ICLR (*International Conference on Learning Representations*) exists since 2013, and many workshops and papers in the field of machine learning follow this direction.

In a schematic way, representation learning techniques aim at obtaining a simple continuous representation of complex objects (element of a sequence, node of a graph, image, text, etc.) – simple in the sense that it is possible to use classifiers such as neural networks or support vector machines. Typically, this representation corresponds to a vector in $\mathbb{R}^n$ and is called *distributed representation*, or *continuous representation.* This representation is ideally of sufficiently high level that each component of the vector represents a latent factor of variance of the space of observations (Bengio, Courville, and Vincent 2013). A very illustrative example given by LeCun is that of the representation of the face of a person. With an image, the raw representation amounts to millions of pixels. However, a face is controlled by 50 muscles, and its position in space can be represented with 6 measures (position and orientation). A perfect representation of the factor of variation would be to represent a face with only 56 values (plus some others representing the environment, i.e. the light).

An additional assumption, which is true in practice in many areas, is that the representation space has a (local) geometry where the objects are grouped in areas of space that have common properties – such as belonging to the same class (e.g. the class of agricultural landscapes for images), being relevant to the same user queries, etc.

**Illustration** In IR, models based on representation learning have been developed into two orthogonal directions. The first represents separately the question and the document in a common vector space, before matching them e.g. by computing their cosine (P. Huang et al. 2013). The second instead uses the interactions between pre-computed representations of query and document words (i.e. inner product between keyword representations), which can be seen as a representation of the degree of match between each word from the query and each word from the document. This match can then be processed using neural networks (Guo et al. 2016). In both cases, the document or word representations can be learned from data (e.g. query-document-relevance triples).

## 1.1 Summary of Contributions

In the context of representation learning, the work reported in this manuscript is centered around the question of how to *represent information with continuous representations* in the application context of *information access,* with an emphasis on the representation of text – whose related works are discussed in Chapter 2.

The manuscript presents my research along to orthogonal directions, i.e. working on the properties of a chosen representation scheme (i.e. quantum and probabilistic embeddings) and improving the representation of textual data (grounding).

**Quantum and Gaussian Representations (Part I)** Most works in representation learning assume that the representation space is Euclidean equipped with the usual inner product. However, such spaces do not have an obvious definition of how probability distributions should be defined, and have no way to express how certain is a representation. The two first chapters present works conducted around the development of alternative representation spaces, namely through the use of a quantum probabilistic representation space (chapter 3) and the exploitation of Gaussian distributions in $\mathbb{R}^n$ (chapter 4).

**Grounding (Part II)** Representing textual information mostly relies on word co-occurrence information. However, since it is known that human relies on representations which are *grounded* in reality Barsalou (2008), it has been hypothesized that any worthy representation should also be *grounded* – in particular to capture *common sense knowledge.* Chapter 5 discusses the problem of using multimodal information to improve representations – i.e. can we use images to integrate common sense, such as "the sky is blue", into word and sentence embeddings?

**Other contributions (Part III)** Finally, my other contributions are described succinctly in Chapter 6, and deal with user modeling, evaluation, information extraction and document summarization.

Finally, Chapter 7 concludes the manuscript, and presents my current research direction.

## 1.2 Notations

The following notations will be used throughout the manuscript (in most parts):

- We use boldface to denote vectors $\boldsymbol{x}$ and matrices $\boldsymbol{A}$
- $\underline{x} \in \mathbb{K}^d$ is the representation of $x$ in a vector space of dimension $d$ of field $\mathbb{K}$ ($\mathbb{R}$ or $\mathbb{C}$ in this manuscript).
- When there are various possible representation for $x$, we note $\underline{R}_x$ the $R$ representation of $x$. For instance, we denote $\underline{\text{cnn}}_x$ the CNN representation of an image $x$. We also use the notation $\underline{\text{cnn}}_\theta(x)$ or $\underline{\text{cnn}}(x;\theta)$ when parameters need to be specified.
- $\text{card}(X)$ is the number of elements of the set $X$
- $\text{softmax}_k \mathbf{y}$ is the $k$th component of the softmax operator of a vector $\mathbf{y}$

$$\text{softmax}_k \mathbf{y} = \frac{\exp(\mathbf{y}_k)}{\sum_{i=1}^{n} \exp(\mathbf{y}_i)}$$

  we also denote softmax $\mathbf{y}$ the vector of the probability simplex $\Delta^n$ (i.e. the probability distribution over all the outputs).

# Chapter 2

# Background: The evolution of textual representations

In this chapter, we describe how the representation of textual data has evolved in the last fifty years – starting from bag-of-word representations in the 1960s to the rise of deep learning in the 2000s.

## 2.1 Feature-based Representations

The first attempts at representing a text were based on statistics about occurring words in a document and within a collection. The most famous example is that of Salton (Salton and Lesk 1965) who developed an automated term extraction system for IR. The interest and potential of such a representation was latter confirmed to be competitive with many sophisticated indexing schemes in the Cleverdon experiments (Cleverdon 1967). Many attempts have been conducted to improve this representation, but successful attempts have mainly focused on improving *how to compute the importance of a term* (term weights) rather than changing altogether the fact that terms were the basic representation units (i.e. bag of word representation).

With the rise of machine learning in Natural Language Processing (NLP) and Information Access fields, there has been a focus on the development of meaningful features, carefully chosen for the task at hand. To cite a few,

- In Web Information Retrieval, learning to rank approaches rely on features such as relevance scores from model-based IR such as BM25 (Robertson and Zaragoza 2009), the PageRank (Page et al. 1999) of the given page, the number of incoming/outgoing links, the length of the page, and so on (T.-Y. Liu 2011).
- In emotion recognition, the average valence of words (dictionary based) is used in many works (Polanyi and Zaenen 2006), but more advanced features such as those based on neural networks (A. Y. Ng et al. 2011) can be used.
- In information extraction, or more specifically, in relation classification where the task is to predict which relation (between two entities) occurs in a sentence, a representation of the path between the two entities at hand (e.g. part-of-speech tags of words between the two entities) can be used as features (Yao, Haghighi, et al. 2011).

In all these cases, designing and evaluating the different proposed features is time consuming – both for humans and machines. The trend towards continuous representations, that we describe in the next sections, is tackling this main issue. The recent success of neural architectures show that it is possible to learn automatically how to represent texts, relying on the expressive power of such architectures, as well as on a wealth of knowledge about what architecture can be used on which task, and about how to train their parameters successfully.

## 2.2 Towards continuous representations

Text representation has since then evolved from being non continuous (e.g. a set of words) to being continuous (aka *distributed*). This evolution is centered around two different phases.

**Latent Topic Spaces** The first phase is in the 1990s, Deerwester et al. (1990) have shown that it is possible to use word co-occurrence within documents to represent both documents and words in a *latent topical space.* Their method, named Latent Semantic Indexing (LSI), is based on the singular value decomposition (SVD) of a term by document matrix[1], allowing to uncover latent factors that represent a term or a document.

Probabilistic approaches have also been proposed, such as Probabilistic Latent Semantic Allocation (Hofmann 2001) or Latent Dirichlet Allocation (Blei, A. Ng, and M. Jordan 2003), which are generative models based on the idea that the (latent) topic explain the word distribution over some regions of a document. There has been a good number of works extending LDA, and an overview of this type of models is presented in (Blei, Carin, and Dunson 2010).

Latent Topic Spaces have had two main applications: document clustering and term analysis (see e.g. Blei, A. Ng, and M. Jordan 2003). In the former, the latent topics are good indicator of the theme the document is discussing, and for the latter, it is possible to represent the term specific topics – sometimes even looking at their evolution through time (C. Wang, Blei, and Heckerman 2012).

Apart from these specific tasks, this type of representation has never been successful in tackling other ones. A reason of this failure is that latent spaces are mostly topical spaces, and do not allow to capture important information such as which specific named entity is used, which is essential for many information access tasks. The consequence is that they have been used has a part of other models, e.g. to smooth word distributions in language model based approaches as in (Deveaud, SanJuan, and Bellot 2013).

Another reason is that powerful supervised models exploiting such a representation were not developed at this time: most approaches relied on using a classifier on top of a continuous representation, the latter being computed with no supervision (language modeling task). The representation is thus not task-specific enough for being successfully used in the different tasks.

**Neural Networks** The second phase in the continuous word representation debuts with the resurgence of neural networks for language modeling, and more precisely the work of Bengio, Ducharme, et al. (2003). In this generative model, each word is associated with a continuous representation. A recurrent neural network (RNN) is used to compute a probability distribution over words $w$, conditioned over the previous words $w_1$ to $w_{t-1}$,

$$p\left(w|w_{1\ldots t-1}\right) = p\left(w|s_t = f_\theta\left(s_{t-1}, \underline{w}_{t-1}\right)\right)$$

The RNN is used to summarize the information of the previous words into a fixed size *state* vector $s_t$ in a latent continuous space through the inductive formula $s_t = f_\theta\left(s_{t-1}, \underline{w}_{t-1}\right)$. The RNN can thus be thought (in theory at least) as a Markov chain with infinite order: The state vectors $s_t$ contain syntactic and semantic information necessary for the generative task. This model both learns the representation of each word (word embeddings) as well as the parameters of the RNN.

Following Bengio, Ducharme, et al. (2003), many works have tried to adapt (recurrent) neural networks for specific tasks. Once this model has been trained, it is possible to use the RNN state $s_t$ as a representation of a text up to the processed word $w_t$. The last state of a text $s_T$ corresponds to the representation $\underline{rnn}_d$ of the text $d$. Since states are of fixed dimensions,

[1] each term corresponds to a row $i$, each column to a document $j$, and a value in the matrix corresponds to the importance of the term $i$ for the document $j$

they can be used for classification (e.g. topics) and/or regression tasks (e.g. sentiment). Such RNNs were used directly, e.g. for speech recognition (Schwenk 2007) and morpho-syntactical labeling (Collobert et al. 2011).

There are many variations around recurrent neural networks:

- Increasing the representational power by processing the document in the opposite time direction (Schuster and Paliwal 1997). With Bi-directional RNNs, the representation of a document $d$ is the concatenation of the last states of the RNNs in both directions $\underline{rnn}_d^{\rightarrow} \oplus \underline{rnn}_d^{\leftarrow}$.
- Allowing them to capture more distant relationships in the input. It has been shown early in the history of RNNs that they were prone to the vanishing gradient problem, i.e. that the magnitude of the gradient decrease exponentially when going back in time with the back-propagation mechanism. Solutions such as Long Short Term Memories (LSTM) were proposed by Hochreiter and Schmidhuber (1997). These have only been picked in NLP and information access much latter, giving rise to variations such as Gated Recurrent Units (GRU, defined in Cho et al. 2014).

**Convolutional Neural Networks** Coming from the vision research field, Convolutional Neural Networks (CNNs) have been proposed to leverage the (two-dimensional) equivariance of data (LeCun et al. 1989), i.e. the idea that the processing of an image region should not depend on its location. This makes sense since an object identity will not be changed by moving it in the picture. The second idea in convolutional networks is that the analysis should be compositional – i.e. go from regions, composed of basic shapes, to larger ones, composed of real world objects.

This idea can be transposed for texts, the only difference being that the translation equivariance is unidimensional and not bi-dimensional. The basic 1D convolution can be expressed as:

$$\boldsymbol{y}_i = \sum_{j=1}^{w} \boldsymbol{K}_j \boldsymbol{x}_{i-j+1} + \boldsymbol{b} \in \mathbb{R}^{d_{\text{output}}}$$

where $\boldsymbol{x}_{i-j+1} \in \mathbb{R}^{d_{\text{input}}}$ is 0 when $i-j$ is out of the sequence bounds (i.e. less than 0 or greater than the the sequence length $l$), $\boldsymbol{K}_j \in \mathbb{R}^{d_{\text{output}} \times d_{\text{input}}}$ is the $j^{\text{th}}$component of the kernel of width $w$, and $\boldsymbol{b}$ the bias.

In the first layers of a CNN, word embeddings are used, i.e. $\boldsymbol{x}_i$ is equal to the embedding $\underline{w_i}$ of the word $w_i$. As for RNNs, this embedding is learned when training the model, eventually starting with pre-trained word embeddings (see next section).

Because of its robustness and its ability to relate distant words (compared to RNNs), CNNs have been used for many (classification) tasks such as:

- Information Extraction, as for example the P-CNN (Y. Y. Huang and W. Y. Wang 2017; Sahu et al. 2016; D. Zeng et al. 2015), where the text before, between and after the entities is processed with three different CNNs, before classifying the relationship into one of the predefined categories.
- Sentiment Analysis where CNN are quite popular since they can easily detect patterns such as "I like", "This ... was great" or "This ... was particularly awful" (Kalchbrenner, Grefenstette, and Blunsom 2014).
- Entity recognition (Adel, B. Roth, and Schütze 2016), where each token of the sentence is classified as being either an entity or not.
- Information Retrieval, to classify whether a document is relevant for a query given its interaction map, i.e. the inner product value between each term of the query and each term of the document (Yin et al. 2016).

As shown above, recurrent/Convolution Neural Networks have been studied and experimented with intensively these last fifteen years for a variety of tasks. We now focus on models and techniques that try to represent words or sentences from raw data, i.e. from text alone.

## 2.3 Word Embeddings

Most – if not all – neural network-based models nowadays rely on pre-trained embeddings[2], which form an important part of the model parameters. Examples are sentiment analysis (A. Y. Ng et al. 2011), information extraction (Y. Y. Huang and W. Y. Wang 2017; C. N. d. Santos, B. Xiang, and B. Zhou 2015; Y. Xu et al. 2015; Z. Zhang 2004), response to questions (Iyyer, Boyd-Graber, et al. 2014; A. Kumar et al. 2015), machine translation (Bradbury and Socher 2017; Sennrich, Haddow, and Birch 2016) and information retrieval (Mitra and Craswell 2017).

Apart from specializing neural architectures on specific tasks, and given the importance of the quality of word/sentence embeddings, as well as the difficulty to train models such as (Bengio, Ducharme, et al. 2003) on large quantities of text, methods have been developed to learn word or sentence embeddings with easier to train models. The question is now of being able to produce *high* quality embeddings, e.g. useful as a starting point to train a supervised model on a given task.

Bengio, Ducharme, et al. (2003) embeddings are slow to compute, and hence big datasets cannot be used to learn the word embeddings. The seminal work of Mikolov et al. (2013) proposed an unsupervised model to learn word representations in a vector space, with the goal of using them in natural language processing tasks, by grouping semantically similar words in the same region of the vector space. This has been for many years the basic representation for many neural network models manipulating text.

The goal of the Skip-Gram model (the simplest and most used model from Mikolov et al. 2013) is to learn word representations from which the context can be inferred, i.e. that can predict surrounding words. Thus words that appear in similar contexts have similar representations. More formally, the training objective of the model consists in maximizing the log-probability:

$$\sum_{(t,c)\in\mathcal{D}} \log p(t \text{ appears in the context of c}) \stackrel{\text{def}}{=} \sum_{(t,c)\in\mathcal{D}} \log p\left(t|c\right) \tag{2.1}$$

where $(t, c) \in \mathcal{D}$ is the set of terms $t$ associated with the context $c$ in which they occur. For example, Mikolov et al. (2013), for the Skip-Gram model, take a window of a fixed size centered on $t$, and consider that any word but $t$ in this window are part of the context. Again, various conditional probability functions may be used for $p(t|c)$. The basic formulation (Mikolov et al. 2013) consists in using a softmax function and learning two representations, one $\underline{c}$ for the context $c$ and one $\underline{t}$ for the target word $t$:

$$p(t|c) = \text{softmax}_t\ \left(\underline{t'} \cdot \underline{c}\right)_{t'} = \frac{\exp(\underline{t} \cdot \underline{c})}{\sum_{t'\in\mathcal{V}} \exp(\underline{t'} \cdot \underline{c})} \tag{2.2}$$

where $\mathcal{V}$ is the set of all words. In practice, the sum involves too many terms, and a classification loss can be used instead:

$$p(\text{co-occur}|t, c) = \sigma(\underline{t} \cdot \underline{c})$$

This necessitates in turn requires finding "negative" samples, i.e. words that should not appear in this context. In Mikolov et al. (2013), random words are sampled as negative contexts since the probability of picking a *wrong* context word at random is very low. Formally, in

[2] At least, up to the recent works on Self-Attention Networks, aka Transformers (Vaswani et al. 2017)

Skip-Gram, the optimized loss is

$$- \mathbb{E}_{(t,c)} \left[ \sigma(\underline{t} \cdot \underline{c}) + \sum_{\substack{k=1 \\ c^- \sim \mathcal{U}(\mathcal{V})}}^{N} \left(1 - \sigma(\underline{t} \cdot \underline{c^-})\right) \right] \tag{2.3}$$

where $c^-$ is a negative context, i.e. a context where the term $t$ should *not appear*. In practice, in Mikolov et al. (2013) and all its derived works (including ours), the negative context is simply sampled uniformly from the set of terms – which is a true negative context often enough.

An interesting property of the learned representations, for this model as well as other related ones, is that some basic algebraic operations on word representation were found to be meaningful and understandable from a natural language point of view (Mikolov et al. 2013). For example, a simple operation such as France + capital is close to Paris. Even complex operations can be done, for example, the closest word representation Lisbon-Portugal+France is Paris. This shows that the learned representations not only bear semantics, but that the geometry of the space is meaningful.

The models proposed in (Mikolov et al. 2013), namely Skip-Gram and C-BOW, can be set within the wider framework of *exponential family embeddings* (M. R. Rudolph et al. 2016). Such a family defines a context function, a conditional probability function (generally modeled from the exponential family) and an embedding structure to form the objective function.

The most well-known alternative models to Word2Vec include Glove (Pennington, Socher, and C. Manning 2014) that approximates the log-probability of co-occurrence as the inner product between the word representations (up to a constant), and FastText (Bojanowski et al. 2016) that solves the out-of-vocabulary problem by including subword information to represent a word similarly to (Schütze 1993), that is, as a sum of n-gram representations.

Studies trying to compare the different word embedding models have not found high discrepancies between those, in terms of performance on the task at hand (Baroni, Dinu, and Kruszewski 2014) – which is sensible since theoretical works have shown how close these ideas are. Levy and Goldberg (2014) have shown that Word2Vec can be interpreted as the matrix factorization of the point-wise mutual information (PMI) matrix, which is close to both Glove and previous works on latent topic models. Arora et al. (2015) gives another theoretical view of these different models, whereby the text is obtained by a generative process driven by a context vector – this allows to justify why the linear structure of word embeddings (e.g. France+capital) is meaningful.

## 2.4 Sentence Embeddings

In the previous section, we described techniques for learning to represent words from raw text. At the sentence level, apart from topics models such as Latent Semantic Analysis (LSA, Deerwester et al. 1990) or Latent Dirichlet Allocation (LDA, Blei, A. Ng, and M. Jordan 2003), most sentence embeddings are an indirect outcome of a task-specific training. This includes supervised and task-specific techniques with recursive networks (Socher et al. 2013), convolution networks (Kalchbrenner, Grefenstette, and Blunsom 2014), averaging followed by several non-linear transformation in Deep Averaging Networks (Iyyer, Manjunatha, et al. 2015) but also unsupervised methods producing universal representations given large text corpora. We focus on the latter in this section, since they are more connected to the works we conducted.

Direct approaches such as averaging the embeddings of words present in the sentence are limited by the fact that they (**L1**) do not include syntactic information; (**L2**) do not take into account the order of words in the sentence; and (**L3**) do not take into account the interaction of words in the sentence.

Trying to tackle those limitations, sentence-specific approaches were thus developed along two axes:

| Paper | Tackled | | | Encoding | Loss |
|---|---|---|---|---|---|
| | L1 | L2 | L3 | | |
| Sent2Vec (Le and Mikolov 2014) | X | | | WE (average) | **classification**: word in the sentence (AE) |
| Skip Thought (Kiros et al. 2015) | X | X | X | RNN | **generation**: next/previous sentence |
| FastSent (Hill, Cho, and Korhonen 2016) | | | X | WE (sum) | **classification**: word in previous/next/same sentence (AE) |
| Kenter, Borisov, and Rijke (2016) | X | | | WE (average) | **classification**: next sentence |
| Quick Thoughts(Logeswaran and H. Lee 2018) | X | X | X | RNN | **classification**: next/previous sentence |
| Universal Sentence (Cer et al. 2018) | X | X | X | WE (average) + MLP / Contextual WE (average) | Multi-Task (**generation**, **classification**) |
| Arroyo-Fernández et al. 2019 | X | ? | X | WE (convex) | **No learning** (mutual information) |
| Deep Averaging Network (Iyyer, Manjunatha, et al. 2015) | | | X | WE (sum, followed by an MLP) | **No learning** (finetuning on task) |

Table 2.1 – Different approaches for unsupervised sentence representation (WE stands for Word Embeddings, AE for auto-encoder)

1. Modifying the way word representations are aggregated: with RNNs (tackles **L1, L2 and L3**), as a simple combination of word embeddings (sum/average, tackles **L1**) or convex combination of words (tackles *somehow* **L1, L2 and L3**).

2. Modifying the optimization task(s) which are used to learn the sentence representations – generative (e.g. generate the next/previous sentence) vs classification. The latter can be useful to learn semantic representations through classifications tasks such as "can this sentence follow this one?".

The table 2.1 summarizes the different approaches, and shows that most combinations have been tried so far, with experimental results showing that despite of their simplicity, approaches based on simple encodings (sum/average) perform quite well because they are much easier to train. A special mention should be paid to (Cer et al. 2018) since they use contextual word embeddings (see next section), and perform much better than previous approaches.

## 2.5 Contextual Word Embeddings

Performances of unsupervised sentence encoders for end-tasks were not better than using word embeddings and fine-tuning a model. At the same time, word embeddings are not fully satisfying since ideal word embeddings should model both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). However, these characteristics are highly context dependent, and cannot be captured with the models presented above.

Recently, *contextual* word embeddings models have been proposed and used with success on end tasks – yielding significant improvements over past works – with the seminal works of M. E.

Peters et al. (2018) with ELMO, and, more significantly, of Transformers architectures such as BERT (Bidirectional Transformers for Language proposed in Devlin et al. 2018).

ELMO is a language model based on stacked RNNs trained on the standard auto-regressive language modeling task – the deepest layers are supposed to contain more contextual information about the word at hand, and can be fine-tuned on specific task. However, this work does not depart significantly from previous works, and the limitations of recurrent neural networks are still present (e.g. limited dependency range).

Departing from RNNs, and relying rather on the attention mechanism developed by memory-based neural networks (Graves, Wayne, and Danihelka 2014; Weston, Chopra, and Bordes 2015), BERT (Devlin et al. 2018) is based on the transformer architecture proposed by Vaswani et al. (2017). A transformer is composed of parametric functions that successively refine the representation of a sequence of embeddings. More precisely, each layer $\ell$ transforms the sequence $x$ composed of $n$ vectors $\underline{x}_1^{(\ell-1)}, \ldots, \underline{x}_n^{(\ell-1)}$ into a sequence $\underline{y}_1^{(\ell)}, \ldots, \underline{y}_n^{(\ell)}$ of the same length, through an attention over the context sequence $c$ composed of $m$ vectors $c_1, \ldots, c_m$ with

$$\underline{x}_i^{(\ell)} = \underline{x}_i^{(\ell-1)} + \sum_j p(j|i) v_\ell(c_j^{(\ell)})$$

with $\log p(j|i) \stackrel{+}{=} \left\langle q_\ell\left(\underline{x}_i^{(\ell-1)}\right), k_\ell\left(c_j\right)\right\rangle$. Three functions are defined, $q_\ell$ (the query), $k_\ell$ (the key) and $v_\ell$ (the value). The first two are used to select which part of the context should be used to update the contextual representation, while the third ($v_\ell$) is the residual value to add to the contextual representation.

BERT uses a self-attention mechanism, i.e. $c_j^{(\ell)} = \underline{x}_j^{(\ell-1)}$ and various other mechanism to allow efficient learning, and is trained on two unsupervised tasks:

1. Fill-in-the-blank task, where one or more tokens are removed from the input – the goal being to recover them;

2. Predict if a sentence could come next (similar to unsupervised sentence embeddings losses, see previous section).

GPT models (Radford, Narasimhan, et al. 2018; Radford, Wu, et al. 2018) were proposed around the same time as (Devlin et al. 2018), but use a language model objective rather than a fill-in-the-blanks task – showing similar performance on an array of tasks. This shows that the pre-training task(s), while important, are not primordial for achieving a good performance when fine-tuning – rather, the self-attention mechanism and the amount of data on which these models are trained are the key to the improved performance.

Since BERT and GPT, many research directions have been followed around Self-Attention Networks (SANs) – see (Rogers, Kovaleva, and Rumshisky 2020) for a recent overview of self-attention models:

1. The first following works have proposed different optimization objectives to enhance further the quality of the contextual representations, as measured on end tasks such as translation, question-answering or summarization:

   (a) RoBERTa (Y. Liu et al. 2019) removes the sentence classification task and uses a bigger dataset – leading to a better pre-trained BERT version;

   (b) HUBERT (Moradshahi et al. 2019) proposes to bind symbols and roles (in transformers and LSTMs) by adding an extra layer to BERT;

   (c) BART (Lewis et al. 2019) generalizes BERT by using more perturbations (beyond token masking) such as sentence permutation, token deletion, etc.

   (d) ALBERT (Lan et al. 2020) uses a loss that asks whether two sentences are in the right order

(e) T5 (Raffel et al. 2019) uses several high level tasks (translation, question answering, ...) to train the model.

(f) Electra (Clark et al. 2019) uses an adversarial approach by using a discriminator (token level)

2. Some models have been trained on specific datasets (language or domain), e.g.

   (a) SciBert (Beltagy, Lo, and Cohan 2019) for scientific texts;

   (b) Camembert (Martin et al. 2020) for French language

3. As BERT has a complexity which is quadratic with respect to the sequence length (both in memory and time), another line of research seeks to reduce the computational requirements:

   (a) by lowering the number of parameters: ALBERT (Lan et al. 2020) by sharing parameters (across layers) and reducing the rank of the embedding matrix and DistilBERT (Sanh et al. 2019) that approximates a BERT model with less parameters,

   (b) by trying to approximate the key/value matching: Reformer (Kitaev, Kaiser, and Levskaya 2020) leverages Locally Sensitive Hashing (LSH) and reversible layers (Gomez et al. 2017) to allow SANs to process large texts; more recently, (S. Wang et al. 2020) proposed LinFormer where the matching process is occurring in a space of fixed dimension (independent of the sequence length), and (Katharopoulos et al. 2020) uses polynomial kernels.

   (c) by using a sparse attention: Transformer-XL uses a sliding window for self-attention (Dai et al. 2019) and (Yang et al. 2019) is based on Transformer-XL

   (d) by compressing the context, as in Compressive Transformers (Rae et al. 2019)

The Self-Attention Networks models are now prevalent in NLP and in IR (e.g. W. Yang, H. Zhang, and J. Lin 2019). Most of my conducted works presented in the manuscript do not use such approaches – but the main principles of the approaches could be to a large extent adapted to this formalism; however, the study of the properties of transformers is an active area I intend to tackle so as to gain insights on what are the desirable properties that should be taken back to "simpler" models.

# Part I

# Probabilistic Representation Spaces

The objective of most representation learning approaches is to map input instances (such as images, relations, words or nodes in a graph) to vectorial representations in a low-dimensional space. The goal is that the geometry of this low-dimensional latent space be smooth with respect to some measure of similarity in the target domain. That is, objects with similar properties (e.g. class) should be mapped to nearby points in the embedded space. While this approach is highly successful, representing instances as vectors in the latent space carries some important limitations:

- Vector representations do not naturally express uncertainty about the learned representations;
- We cannot properly model inclusion or entailment by comparing vector representations (usually done by inner products or Euclidean distance which are symmetric).

In this part of the manuscript, we report our works on probabilistic representation spaces, i.e. representations that are associated with a probability distribution in the representation space. More precisely, we present our works on representations leveraging the quantum probability formalism (chapter 3 on the following page) and multivariate Gaussian distributions (chapter 4 on page 36).

# Chapter 3

# Quantum Information Access Framework

## 3.1 Introduction

The representation of documents and questions in Information Retrieval (IR) has remained predominantly uni-dimensional (i.e. a document or a query is a vector). This representation has limitations. For example, it is not easy to represent an ambiguous question or a document that deals with several topics. These issues are important for developing interactive IR systems or seeking to diversify results – which are two sides of the same coin, since they both need to represent faithfully the different topics of a document and/or a user information need.

Latent topic models (Blei, A. Ng, and M. Jordan 2003; Deerwester et al. 1990; Dumais 2004) address this multi-topicality issue, but at the cost of a drop in the precision of the represented information. Using such models (alone) for IR have already been shown to decrease the performance of the systems (Hoenkamp 2011).

Another more successful approach is to diversify the search results. Current models for diversification have not evolved much in the last years, and are either based on diversifying the query (J. He, Hollink, and Vries 2012; R. Santos et al. 2010), or on using a scoring function that accounts for the novelty (Carbonell and Goldstein 1998; Xia et al. 2017) of each document relative to the previously retrieved ones. In both cases, the representation of the document itself remains unchanged, and diversification is achieved by leveraging the relationship between the query and the document. The representation of a document lacks precision, since only a fixed set of topics can be used to describe any document. In both cases, this might be a problem for documents that are not mono-topical.

The approach based on "quantum probabilities" — the mathematical formalism of quantum physics — provides formal bases for a multi-dimensional representation of documents (or more generally, information objects) that exceeds the above mentioned limits.

The quantum probabilistic framework generalizes the theoretical framework of classical probabilities by generalizing distributions on a *set of elements* to distributions over *vector subspaces* of a Hilbert space. This generalization is interesting because it combines geometry and probabilities, two components present in information retrieval models.

To leverage such formalism, we proposed the Quantum Information Access framework (QIA), that posits that *there exists an information space where texts can be represented.* Taking inspiration from K. v. Rijsbergen (2004), who proposed to use the quantum formalism for IR, QIA provides both theoretical and experimental insights on the relationships between quantum physics and information access.

More precisely, and relying on precise definitions that we will detail latter, the QIA framework relies on a multidimensional representation of text fragments, both to represent the (quantum) probabilistic distribution of *information units* present in a text fragment and to represent the

topics covered by this fragment (using quantum probabilistic events). Using a multidimensional representation of documents has been shown to be important in IR, e.g. to deal with multi-topical documents (Zuccon, L. Azzopardi, and K. Rijsbergen 2009), to build up semantic spaces or to cope with contextual IR (Melucci 2008).

In the following, we discuss related works, and then define the different notions used to describe the QIA framework, before proceeding to the results we obtained in two information access tasks, namely document retrieval and extractive summarization.

## 3.2 Related works

The QIA framework relies on a multi-dimensional representation of texts (quantum densities and subspaces). Multi-dimensional representations have been implicitly used in IR to handle negative feedback. In interactive Information Retrieval, Dunlop (1997) showed that positive feedback could be used easily, but this was not the case for negative one. Based on these results, X. Wang, Fang, and Zhai (2008) found that negative feedback could be handled by describing the information need as a set of vectors. More in details, when estimating the relevance of a document $d$, they (1) first cluster the negative documents in the representation space; and (2) compute the maximum similarity between the document $d$ and the different cluster centers. The QIA framework encompasses this approach in the sense that a query is represented as a weighted set of vectors, too.

A more explicit use of multi-dimensional representations is the work from L. Chen, J. Zeng, and Tokuda (2006) who proposed to randomly split a document into two parts, and to use a two-dimensional representation of documents to obtain a "stereoscopic" view of a document. Our framework can be thought of as a principled extension of this work, where we do not limit ourselves to two dimensions and, in addition, rely on the probabilistic framework of quantum physics to compute the relevance of a document to a query.

Explicit multidimensional representations of texts have also been explored. Zuccon, L. A. Azzopardi, and C. v. Rijsbergen (2009) showed that the cluster hypothesis still holds when representing documents as subspaces. Their methodology to build subspaces is close to QIA, since to represent documents they compute the subspace spanned by a set of vectors (albeit implicitly). In our work, we provide an explicit methodology to construct the subspaces.

Using a quantum formalism, Melucci (2008) also uses a subspace for representing a user's information need (the subspace where relevant document vectors should lie), and a vector representation for documents. The probability that a document is relevant to a user's information need is determined by the length of the projection of its vector representation onto the corresponding (information need) subspace. Following quantum physics, we interchanged the role of document and user's information need. This is motivated by the fact that the user's information need should be represented as a dynamic component, as advocated in e.g. Ingwersen and Järvelin (2005).

Our work also bears some similarity with Latent Semantic Indexing (LSI, Dumais 2004) since we use spectral analysis to extract document and query representations. However, we do not represent objects in a low-dimensional space as in LSI, but use a spectral analysis to obtain a compact representation of our document subspaces and query densities. Hyperspace Analogous to Language (HAL) spaces Burgess, Livesay, and Lund (1998) are also closely related to our work. In this work, each word $w$ is represented by a term vector where non null components correspond to words co-occurring within a small window centered around word w. Our representation of a term is inspired by this approach, but, for each word w, we use spectral analysis to summarize the information brought by the set of vectors associated with it.

Outside IR, in the face detection domain, subspaces are commonly used to solve recognition problems. In Belhumeur, Hespanha, and Kriegman (1997), a face is represented by a subspace (generated from different picture vectors of the same face) and recognition involves computing

the distance between a vector (representing the face to be recognized) and the subspace. Different from (Belhumeur, Hespanha, and Kriegman 1997), we also leverage mixture of multivariate densities.

Quantum IR is a field that has been initiated by van Rijsbergen's seminal work (K. v. Rijsbergen 2004). Besides works presented above, various research directions have been followed (the list is not exhaustive):

- Zuccon, L. Azzopardi, and K. Rijsbergen (2009) experimented with a quantum inspired principle for ranking documents (interferences). Our work proposes another approach to the problem of diversity, whereby the representation itself gives a principled way to rank documents;
- These are represented by a subspace generated from different aspects of the document, to represent documents in a space different from the standard term space Huertas-Rosero, L. Azzopardi, and C. J. v. Rijsbergen (2008);
- Sordoni, Bengio, et al. (2013) has proposed to take into account term dependencies using quantum densities;
- Sordoni, J. He, and Nie (2013) proposed a quantum-based latent topic model.

Our work is different from the above in the sense that it proposes a general methodology to represent document and queries, and more generally to represent textual information.

Finally, there are links with probabilistic representations, e.g. Gaussian-based representations (see section 4). In both the quantum and the Gaussian cases, a representation is linked with a density in a probabilistic space. The main differences stems from the fact that the space is not the same, namely a set for standard probabilities, Hilbert spaces for quantum probabilities (or rather, a *-algebra). In practice, the advantages of the quantum formalism are in how easy it is to deal with the manipulation of multidimensional spaces:

- Conditional distributions are projections in quantum probabilities – they can thus be computed easily (when the dimension is small enough);
- Representing events in quantum probabilities is somehow easy, since it amounts at defining a Hilbert subspace;
- Mixture of quantum probabilities are multimodal – this is not the case for continuous distributions for which the mixture is easy to compute (e.g. a mixture of Gaussians is also a Gaussian). Multimodal distributions are particularly interesting when it comes to modeling diversity (of the user information need), and we observed positive results in extractive summarization where (one of the goal) is to diversify the selected sentences (see section 3.4.2 below).

The price to pay is that there is no control over the variance of random variable for quantum representations – it is determined by the quantum formalism, and that there is less expressive power than for multivariate distributions.

We now turn to the QIA framework and how it is used in two applications: ad-hoc retrieval and extractive summarization. Before doing so, we introduce quickly the probabilistic framework of quantum physics.

## 3.3 Quantum Probabilities

The quantum probability formalism is a generalization of standard probability theory, which makes use of Hilbert spaces, unit vectors and subspaces. We present the components of the formalism used in our work.

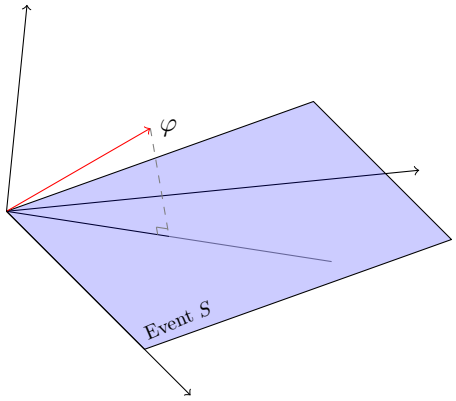



Figure 3.1 – Quantum probabilities - The projection of a simple probability distribution defined by a unit vector $\varphi$ on an Hilbert subspace (event $S$).

### 3.3.1 Systems and States

Quantum theory describes the behavior of matter at atomic and subatomic scales, by representing the state of a physical system (e.g. the position of an electron) as a state in a probability distribution space. Formally, this space is a Hilbert space $\mathcal{H}$, i.e. a vector space defined on the complex field $\mathbb{C}$ and equipped with its natural inner product. The state of the system is described by a unit vector in the state space, called (improperly) the state vector[1]. States determine statistically the measures obtained on the system, for instance the position of a particle. In this case, the state vector determines the *probability* that the particle is at a given position (but not its exact position).

A system state may be fully known, in which case the system is described by exactly one state vector, but the formalism also allows to be uncertain about its state, in which case the system state can be seen as a distribution over the possible state vectors. This distribution is called a density operator (K. v. Rijsbergen 2004). Depending on the information access task, we may regard single-part systems (a quantum probability distribution space), but very often we need to discuss multi-part systems. We therefore introduce single-part systems, including the cases where states are known or uncertain, and then proceed to multi-part ones (a product of quantum probability distribution spaces).

**States and Probabilities** Given a Hilbert space $\mathcal{H}$ and a state vector $\varphi$, a probabilistic event is represented as a subspace $S$ of $\mathcal{H}$. A state vector $\varphi$ induces a probability distribution on events (i.e., the subspaces). The probability of an event is given by the square of the length of the projection of $\varphi$ onto the corresponding event subspace $S$, that is by computing the value $\|\hat{S}\varphi\|^2$ where $\hat{S}$ is the orthogonal projector onto the subspace $S$, i.e. it is defined as

$$\hat{S}\varphi = \underset{\varphi' \in S}{\operatorname{argmin}} \left\| \varphi' - \hat{S}\varphi \right\|$$

In case of a finite Hilbert space of dimension $d$ (which will be our case), the projector $\hat{S}$ is a linear transformation and can thus be identified with a matrix $\mathbb{C}^d$. This value is the probability of the event $S$ with respect to the probability distribution defined by $\varphi$ (see figure 3.1):

$$\mathrm{q}\left(S|\varphi\right) = \left\| \hat{S}\varphi \right\|^2 \tag{3.1}$$

which is by definition between 0 and 1, since $\|\varphi\| = 1$ and $\|\hat{S}\| = 1$

[1] since multiplying with any complex number $z$ of norm 1 does not change the underlying state, the state $\varphi$ and $z\varphi$ correspond to the same quantum state.

**Uncertain States** There is often some uncertainty on the preparation process of the system, which in turn induces some uncertainty about the state the system is in at the beginning of the experiment. This corresponds (see K. v. Rijsbergen 2004, p. 83) to a distribution over the states $p(\varphi)$. Given this distribution, we can then define the quantum probability of an event $S$ as:

$$\mathrm{q}(S|V) = \int_{\varphi} p(\varphi)\, \varphi^{\dagger} \hat{S} \varphi$$

where tr is the trace operator, $\varphi^{\dagger}$ the transpose of $\varphi$ and $\hat{S}$ the orthogonal projector on the subspace $S$. Using the property of the trace (linearity and cycle permutation), it can be shown that:

$$\mathrm{q}(S|V) = \int_{\varphi} p(\varphi)\, \varphi^{\dagger} \hat{S} \varphi = \mathrm{tr}\left(\left(\underbrace{\int_{\varphi} p(\varphi)\, \varphi\varphi^{\dagger}}_{\rho_V}\right) \hat{S}\right) = \mathrm{tr}\left(\rho_V \hat{S}\right) \tag{3.2}$$

We can observe that we can isolate the event $S$ and a linear operator $\rho_V$ that represents the quantum probability distribution. Note that equation (3.2) reduces to equation (3.1) when the state is certain, i.e. when the distribution over the states $\varphi$ is a Dirac.

**Conditioning** To compute the conditional distribution of $\rho_V$ knowing $S$, in the quantum formalism, we simply project the matrix $\rho_V$ onto the subspace $S$, before normalizing. Formally, the conditional probability distribution obtained after observing $S$ with an initial quantum distribution $\rho$ is given by:

$$\rho_{V/S} = \frac{\hat{S}\rho_V}{\mathrm{tr}\left(\rho_V \hat{S}\right)}$$

This shows that is very easy to condition in the quantum probability framework (compared for example as a conditionalization in the case of continuous distributions). Moreover, it also shows that observing an event corresponds at simply restricting the quantum density to the observed subspace.

**Multi-part systems** Like in standard probabilities, it is possible to combine measurable spaces into a product of measurable spaces. Without entering into details, we apply standard probability rules when the event and density can be decomposed into independent factors:

$$q\left(\bigotimes_i S_i | \bigotimes_i \rho_i\right) = \prod_i q(S_i|\rho_i) \tag{3.3}$$

where $\bigotimes_i S_i$ is a product of events and $\bigotimes_i \rho_i$ is the product of densities. In this case, both behave like a simple a Cartesian product of standard events and probability distribution. Densities and subspaces can be combined, but we won't enter here into details which deal with problem of entanglement, a very powerful quantum concept, and refer the interested reader to (Gudder 1988).

## 3.4 The Quantum Information Access framework

The Quantum Information Access framework (QIA) aims at defining a general methodology to represent any text, both as an event and a density. This process was first introduced in Piwowarski, Frommholz, Moshfeghi, et al. (2010) in a document filtering scenario. In Piwowarski, Frommholz, Lalmas, et al. (2010), we further showed how to construct a document subspace representation by experimenting with a number of strategies and associated parameters. The

process was abstracted in Piwowarski, M.-R. Amini, and Lalmas (2012) and in this section, we follow this latter presentation.

The text representation is based on the assumption that a typical text corresponds to a set of "*information units*". Conceptually, information units can be understood as the set of assertions made by the document. This can be related to the notion of "nugget" (Clarke et al. 2008), used in summarization and question-answering to assess the amount of relevant information a summary or an answer contains, or in summarization to the notion of *factoids* (Teufel and Halteren 2004).

QIA also assumes that each text can be split into possibly overlapping and non-contiguous semantic fragments, where each fragment addresses a specific information unit. We believe that if information units exist, this hypothesis is quite natural. Given that it is true, there are still two main problems that need to be solved:

1. How to extract these fragments from text? In QIA, we relied in all our experimental works on a very crude assumptions, namely that semantic fragments correspond to sliding windows over the text.

2. How to represent those fragments? In QIA, we chose a simple bag-of-word representations as a starting point, even though we experimented with other possibilities (see Section 3.4.3).

Based on those two assumptions, we can now model a text $\mathbf{t}$ as a probability distribution $\mathrm{p}(\mathbf{n}|\mathbf{t})$ over *information units or nuggets* $\mathbf{n} \in \mathfrak{N}$. This space is infinite since it corresponds to all the unit vectors in the Hilbert space. We suppose that each information unit $\mathbf{n}$ corresponds to a representation/state $\underline{\varphi}_{\mathbf{n}}$ in the topical space.

**Two views of a text** We first describe how a text can be defined either as a quantum probability distribution (density) or as an event. According to equation (3.2), we can represent the text as a density $\underline{\rho}_{\mathbf{t}}$ defined as:

$$\underline{\rho}_{\mathbf{t}} = \int_{\mathbf{n}\in\mathfrak{N}} p(\mathbf{n}|\mathbf{t})\, \underline{\varphi}_{\mathbf{n}} \underline{\varphi}_{\mathbf{n}}^{\dagger} \tag{3.4}$$

where we suppose that the set of information units in a text is finite, i.e. formally that the $p(\mathbf{n}|\mathbf{t})$ is either 0 or a Dirac such that

$$\int_{\mathbf{n}\in\mathfrak{N}} p(\mathbf{n}|\mathbf{t}) = 1$$

The first view over a text is thus a density over information units. As the information unit space can be large, we rely on an eigenvalue decomposition of the density $\rho_t$, i.e.

$$\underline{\rho}_{\mathbf{t}} = \sum_i \lambda_i b_i b_i^{\top} \tag{3.5}$$

where the $\{b_i\}$ form a basis of a subspace of $\mathcal{H}$. This eigenvalue decomposition can be used to filter out noise by discarding eigenvectors whose eigenvalue $\lambda_i$ is below a given threshold.

Given an event $S$ (a subspace), the probability is given by a "weighted" length of the projection of the density on the subspace $S$, i.e.

$$\mathrm{q}\left(S|\underline{\rho}_{\mathbf{t}}\right) = \mathrm{tr}\left(\underline{\rho}_{\mathbf{t}}\hat{S}\right) = \sum_i \lambda_i \left\|\hat{S}b_i\right\|^2$$

where $\hat{S}$ is the projector on $S$.

Another view of a text in the QIA framework is as an event, where a text event is the subspace spanned by the information unit representations that the text contain

$$\underline{S_t} = \text{span}\left\{\underline{\varphi}_{\mathbf{n}} | p\left(\mathbf{n}|\mathbf{t}\right) > 0\right\}$$

In practice, we don't compute the real span of the information unit vectors; we rather rely on the above eigenvalue decomposition, and use the span of the eigenvectors whose eigenvalue is above a given threshold. This allows to filter out noise, i.e. directions for which $p\left(b_i|\mathbf{t}\right)$ is close to 0, i.e. using the eigendecomposition of $\underline{\rho}_{\mathbf{t}}$ – see eq. (3.5), we define the subspace representation of a text as

$$\underline{S}_{\mathbf{t}} = \text{span}\left\{b_i | \lambda_i > \text{threshold}\right\}$$

**Fundamental QIA hypothesis** The event representation of a text $\underline{S}_{\mathbf{t}}$ has an important implication that we wish to discuss now. More precisely, we implicitly suppose that *any linear combination of information units presents in a text* $\mathbf{t}$ *is also a topic of the document.*

For example, supposing information units are represented as bag of words vectors (term frequency representation), we consider a text made of two information units (pizza 1, cambridge 1, uk 1) and (cheese 1, cambridge 1, uk 1). Given the QIA hypothesis, this text is supposed to contain the information unit (pizza 1, cheese -1), which means in practice that it will be either discussing pizza or cheese, but not both at the same time.

The QIA formalism depends on this strong assumption. It is obviously hard to disprove formally such an assertion, and only experimental data can draw a clear line on whether this holds or not.

**Information Units in practice** Having defined the density and event views of information units, we now turn to a more practical issue: How can we extract from text information units, and how to turn them into unit vectors?

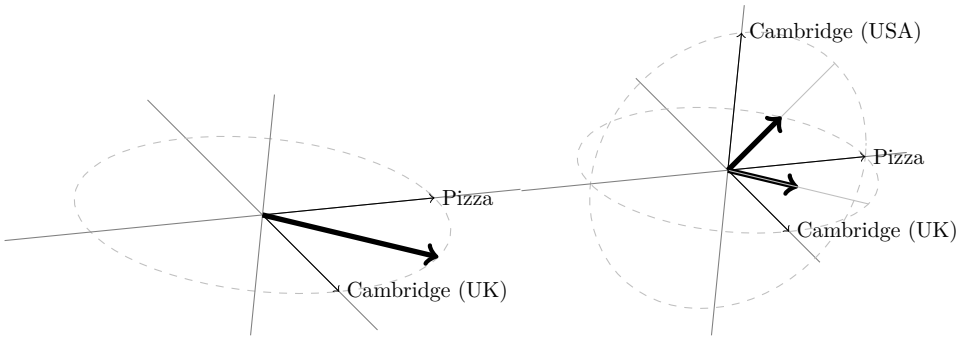
There exists many different ways to extract information units from a document. We can choose to use sentences, paragraphs or any unit that we suppose to contain a single information unit. This is consistent with other approaches for building word representations, such as the seminal work of Mikolov et al. (2013). In all the conducted experimental works but summarization (where we used sentences), we used a simple approach based on sliding windows, and used a uniform prior over windows, i.e.

$$p\left(\varphi\right) = \frac{1}{\text{card}\left(windows(d)\right)} \sum_{\mathbf{n} \in windows(d)} \underline{\varphi}_{\mathbf{n}} \underline{\varphi}_{\mathbf{n}}$$

where $\underline{\varphi}_{\mathbf{n}}$ is the representation of a sequence of terms of length $\ell$ extracted from the document $d$. This corresponds to a uniform prior over the extracted units.

**Superposition vs mixture** Finally, to illustrate the interest of the quantum formalism, we discuss two ways to combine information units representation that have no direct analogue in the standard probability formalism, namely the *superposition*. In the following, for illustration purposes, we use an information retrieval setting where the user has expressed his/her information need with a sequence of keywords. We further suppose that each keyword is associated with a density over the information unit space (that corresponds to the potential information units that should be found in relevant documents).

Superposition amounts at computing a linear combination of two or more unit vectors in order to obtain a new one. Superposition is commonly used in vectorial IR, since to represent a query or a document, one combines linearly the representation of each keyword. In the simplest case, this corresponds to a vector which is naught everywhere but on the component that corresponds to the word.




(a) A superposition of two information needs (b) A mixture of two information needs

Figure 3.2 – Combining information needs

An example shall illustrate this. As depicted in Figure 3.2, let $\underline{\varphi}_p = (1,0,0)^\top$, $\underline{\varphi}_{c/uk} = (0,0,1)^\top$ and $\underline{\varphi}_{c/usa} = (0,1,0)^\top$ be three vectors in a three dimensional space that respectively represent the information needs "I want a pizza", "I want to be delivered in Cambridge (UK)" and "I want to be delivered in Cambridge (USA)". The information need "I want a pizza to be delivered in Cambridge (UK)" would be represented by the vector $\frac{1}{\sqrt{2}}(1,0,1)^\top$ which is the (normalized) linear combination of $\underline{\varphi}_p$ and $\underline{\varphi}_{c/uk}$. In order to represent the information need of a user typing "pizza delivery in Cambridge" (where we don't know whether Cambridge is in USA or in UK), we would use a mixture of the two possible information needs (assuming there is no other source of ambiguity). Similarly, a document answering the information need "pizza delivery in Cambridge (UK)" would be represented as the subspace defined by the vector $(1,0,1)^\top$ while for Cambridge (USA) it would be by the vector $(1,1,0)$. Finally, a document answering both information needs would be defined as a plane in this vector space.

This shows how superposition can be used to express specific information units (e.g. pizza delivery in Cambridge (UK)) whereas mixture is used to represented the uncertainty about the user information need.

Now that we have defined precisely the QIA framework, we show how it has been applied to two Information Access tasks, namely information retrieval and summarization.

### 3.4.1 Ad-hoc Information Retrieval

To adapt the QIA framework to ad-hoc IR, we already have a mean to compute the representation of documents as explained above. The main remaining question is how to represent the information need – i.e., the query $q$.

**Single term queries** As a query in its simplest form consists of a set of terms, we are first interested in building the query representation for a query composed of a single term, $t$. This representation is later needed for constructing the representation of multi-term queries. We assume that a query term $t$ can be represented as a distribution over fragments that would be possible answers to a query containing term $t$. In the QIA framework, each fragment can be associated with an atomic information need.

We suppose that the set of fragments for a query term $t$ corresponds to the set of text excerpts centered on a term $t$. That is, we suppose that *a fragment containing term $t$ is an information unit that answers a query containing $t$.* We then use the immediate surroundings of the term $t$ occurrences in documents to build that term representation. This methodology is similar to pseudo-relevance feedback using passages from retrieved documents containing the query terms (Allan 1995). The difference is that we use all the passages to build the query

representation as we want to consider all possible information units associated with the query term (word) $w$. Once this is done, we can use equation (3.4) to represent the term as density using the distribution over information units defined as:

$$p(\mathbf{n}|t) = \frac{\mathbb{1}\left[w \in \mathbf{n}\right]}{\text{card}\left(\{\mathbf{n}' \in \mathfrak{N} | w \in \mathbf{n}'\}\right)}$$

where $\mathfrak{N}$ is the set of information units.

**Multi-term queries** We used three different ways to represent a multi-term query – either by using a mixture over fragments defined by a probability distribution over query terms, as a superposition and as a tensor product (the equivalent of a probability space product for quantum probabilities).

The *mixture* is simply a mixture of the term specific probability distributions,

$$p(\mathbf{n}|q) = \sum_{w \in q} p(\mathbf{n}|w)p(w|q) \tag{3.6}$$

where we suppose that the information need is expressed by one of the terms with an importance defined by $p(w|q)$. The *mixture* corresponds to queries where each query term is as important as the other, and where the presence of either concept associated with a query term is enough for a document to be relevant (e.g. an hypothetical query "food eat").

The *mixture of superposition* makes a different hypothesis, by supposing that a possible information unit answering the question is a superposition of the information units coming from the different words $w_i$ of the query $q$. This corresponds to the case where query terms should be understood as a concept, e.g. "pizza in Cambridge (UK)", but where a document discussing Cambridge or pizza is still OK. Formally, this corresponds to the following distribution over information units:

$$p(\mathbf{n}|q) = \sum_{\underline{\phi}_{\oplus_i \mathbf{n}_i} = \underline{\varphi}_{\mathbf{n}}} \prod_i p\left(n_i|w_i\right) \text{ with } \underline{\phi}_{\oplus_i \mathbf{n}_i} \propto \sum p(w|q)\underline{\varphi}_{n_i} \tag{3.7}$$

In practice, the above distribution probability cannot be computed sufficiently fast, and we rely on an approximation of it, detailed in (Piwowarski, Frommholz, Lalmas, et al. 2010).

For both mixtures (simple or of superpositions, equations 3.6 and 3.7), the above distribution probabilities over information units define a density, following equation (3.4).

The difference between the mixture approaches and the tensor product one, that we describe below, lies in their ability to handle queries containing different aspects. The above representations provide no explicit means to distinguish between aspects; they operate in one aspect space and treat each aspect associated with a term equally, even if the terms describe different aspects. To support aspects explicitly, we present in (Piwowarski, Frommholz, Lalmas, et al. 2010) a quantum analogue of the "weighted and" (#wand) operator proposed in (Metzler and Croft 2004).

However, user's INs (*Information Needs*) often consist of several "aspects" that relevant documents should address. For example, in the TREC-8 topic 408, "What tropical storms (hurricanes and typhoons) have caused significant property damage and loss of life?", we can identify two (topical) IN aspects, namely tropical storms and significant damage/loss of life. Each IN aspect can be defined within an IN aspect space, where the state vectors are now called information units. Examples of information units are the vectors representing "hurricane" and "typhoon" for the first IN aspect (tropical storms). We use the terminology information unit, since one information unit addresses one aspect of the IN (tropical storms) in the same way that a information unit addresses an IN.

We now suppose that, to be relevant to a query that comprises several aspects of an IN, a document should satisfy ideally all of its aspects. In the quantum formalism, we suppose that the

information need space is represented by a product of probability spaces, i.e. each information need expressed by a query term should be answered independently for the document to be relevant.

The *tensor product* representation of a query is thus defined as (see Eq. 3.3 for its definition).

$$\underline{\rho}_q = \bigotimes_i \underline{\rho}_{t_i}$$

From a practical point of view, in (Piwowarski, Frommholz, Lalmas, et al. 2010), we extend this representation by allowing users to give different importance to certain aspects, by introducing a weighting scheme for aspects, relying on the introduction of an additional dimension of the representation space for which every document is relevant.

Finally, as discussed above, ideally queries (or query parts) should be handled by one of three representations depending on their nature. For instance, the query "food italian french Paris France" should be interpreted as having three aspects (food $\otimes$ Italian **or** French $\otimes$ Paris France), where the second aspect is a disjunction (either can match) which should be handled by a mixture and the third a mixture of superposition (the concept "Paris France").

To investigate this, in a set of preliminary experiments, we manually "translated" TREC keyword queries into structured queries. Experimental results where disappointing though, the best approach being consider each keyword as a different aspect (as shown experimentally below). As with any negative experimental result, it is difficult to know whether this is due to a deficient representation space or information unit extraction methodology, or if this is a more fundamental issue.

**Experimental results** We used the TREC 1 to 8 collections (with the exception of TREC-4 since it did not contain the "title" field), and the TREC ROBUST 2004 collection. We compare the performance of BM25 (with standard parameters, see Robertson and Zaragoza 2009), TF-IDF (without document normalization) and, for the QIA framework, those instantiations corresponding to each query construction process, i.e. mixture, mixture of superpositions, and tensor product (T1 and T2). We used a window span of 5.

Results are given in table 3.1. Overall the results were consistent across all collections. The MAP values are below that of BM25 for mixture and mixture of superpositions, and comparable for both tensor approaches. Given the novelty of our framework, and its still unexplored parameters and their effect, we are satisfied with its performance.

The performance of the QIA framework is well above that of a simple TF-IDF model (which is at its basis). This shows that the QIA framework includes some document length normalization, most probably because information units are meaningful units of sense.

The tensor-based approach T performed better than mixture M and mixture of superpositions MS. The fact that MS works worse than M can be due to the fact that in general terms denote different components of the information needs – it was observed that using mixture of superpositions was better suited for phrase queries.

Finally, we were interested by the influence of the query length. We plotted the difference in average precision between BM25 and the performance of the QIA framework, depending on the query length and on the query representation. We kept constant the span of the window (5). We observed that in all cases, the performance degrades with longer queries. This shows that the representation of multi-term queries is not well handled in the QIA framework. Understanding why this is the case would be important for developing further such type of approaches, but we believe this might be due to the fact that as all the approaches based on semantic representations (e.g. LSA), the representation of a term is not specific enough. For queries composed of a few words, this is not important since most senses should be expanded – however, this is not the case anymore for longer queries. A simple solution would be to use a linear combination of

| | TREC-1 | TREC-2 | TREC-3 | TREC-5 | TREC-6 | TREC-7 | TREC-8 | RB-2004 |
|---|---|---|---|---|---|---|---|---|
| BM25 | 0.230 | 0.209 | 0.282 | 0.148 | 0.224 | 0.182 | 0.236 | 0.242 |
| TF-IDF | 0.084† | 0.041† | 0.056† | 0.035† | 0.088† | 0.056† | 0.082† | 0.074† |
| M | 0.205† | 0.184† | 0.226† | 0.115† | 0.173† | 0.142† | 0.165† | 0.180† |
| MS | 0.209† | 0.167† | 0.206† | 0.112∗ | 0.157† | 0.117† | 0.159† | 0.165† |
| T | 0.232 | 0.195† | 0.281 | 0.148 | 0.214 | 0.182 | 0.234 | 0.240 |

Table 3.1 – QIA - information retrieval task: this table reports mean average precision (MAP). The first line shows the test collection. The second and third lines show the MAP value for BM25 and TF-IDF, respectively. For the query construction, M stands for mixture, MS for mixture of superpositions, T for tensor product. For completeness, significance of the difference with BM25 is shown for the 0.05 level (∗) and the 0.01 level (†).

term-centered models such as BM25 with the QIA-based value, similarly to other words based on latent representations (see e.g. Deveaud 2013).

### 3.4.2 Extractive Summarization

Extractive summarization aims at selecting sentences from one or more documents to summarize (hence the *extractive*). Summarization techniques can be broadly categorized into three groups, feature-based (M. R. Amini and Usunier 2011; Harabagiu and Lacatusu 2005; Radev et al. 2004), graph-based (Erkan and Radev 2004; Mihalcea 2005; D. Wang, T. Li, et al. 2008) and lexical chain based (Barzilay and Elhadad 1997; Y. Chen, X. Wang, and B. Liu 2005; J. Li and Sun 2008) methods. The former first identifies themes and then assigns scores to sentences in each of these themes based on sentence-level and inter-sentence features, e.g. sentence similarity, position, cluster centroids, etc. Graph-based techniques begin by characterizing a set of documents as a weighted text graph and then recursively compute sentence significance, globally, from the entire text graph rather than using single sentences as in feature-based methods. The underlying hypothesis of both methods is that summary sentences are those belonging to an identified theme or to a sentence cluster found in the graph. Therefore, sentences relevant to more than one theme or those midway between two clusters in the graph are never extracted and hence are never part of the summary. Finally, lexical chain approaches first construct different sequences of semantically related words, chains relevant to the topic at hand are identified and eventually sentences matching these identified chains are extracted from the collection of documents.

Our proposed QIA-based approach to summarization belongs to the first group (feature-based) and bears similarity with `LSA`-based approaches, a group of successful approaches first proposed for single document summarization. They aim at extracting salient sentences of a given document within a reduced term space[2] and are based on the singular value decomposition (`SVD`) of a term-sentence matrix.

There are two groups of LSA-based approaches. The first (Gong and X. Lin 2001; Murray, Renals, and Carletta 2005) assumes that each topic found by SVD should be present in the final summary and select sentences having the highest entry along each of the extracted topics. Steinberger and Ježek (2004) found that sentences belonging to several "latent" topics may be good candidates for extraction but are never selected by `LSA`-based approaches to form the summary. To overcome this, they compute a score for each sentence that depends on the most salient extracted latent topics. Our approach re-interprets `LSA`-based methods under the `QIA` framework, and naturally paves the way for selecting those sentences falling into one theme or more.

For extractive summarization, we leveraged the fact that the quantum probability handles

[2]Sentences are represented in a term space, and singular value decomposition (SVD) is used to find the main latent topics, i.e. the "cluster" representatives, in the original term space.

naturally distribution over topics and diversity – be it for events or distributions. We defined the task of extractive summarization as finding the set of sentences $S^* = \{s_1^*, \ldots, s_n^*\}$ that covers the most the topic of the document(s) $\mathcal{D}$ to summarize, i.e. the set $S^*$ such that:

$$S^* = \text{argmax}_{s_1,\ldots,s_n} \text{q}\left(\underline{S}_{s_1,\ldots,s_n} | \underline{\rho}_{\mathcal{D}}\right) \tag{3.8}$$

where $\underline{S}_{s_1,\ldots,s_n}$ is the subspace spanned by the information unit vectors associated with sentences $s_1, \ldots, s_n$, and where the density $\underline{\rho}_{\mathcal{D}}$ is defined by equation (3.4), i.e.

$$\underline{\rho}_{\mathcal{D}} = \int_{\mathbf{n} \in \mathfrak{N}} p\left(\mathbf{n}|\mathcal{D}\right) \underline{\varphi}_{\mathbf{n}} \underline{\varphi}_{\mathbf{n}}^{\dagger}$$

where we experimented with various definitions of $p\left(\mathbf{n}|\mathcal{D}\right)$ – taking into account (or not) the length of the sentences or the query that was used to select the documents to summarize. Details are reported in (Piwowarski, M.-R. Amini, and Lalmas 2012).

**Links with LSA-based approaches** It is illuminating to look at the link with the two main LSA-based approaches with our QIA-based model.

To form a summary, Gong and Lin (Gong and X. Lin 2001) use the $k$ information units associated with the $k$ highest singular values[3], i.e. with $\sigma_1, \ldots, \sigma_k$. The $j^{\text{th}}$ information unit is represented in the sentence space by the $j^{\text{th}}$ column of the matrix $V$ (Equation 3.5). The $i^{\text{th}}$ entry $V_{ij}$ of this vector corresponds to the importance of the $i^{\text{th}}$ sentence for the $j^{\text{th}}$ information unit. Formally, for the $j^{\text{th}}$ information unit, Gong and Lin (Gong and X. Lin 2001) select the $i_*^{\text{th}}$ sentence such that:

$$i_* = \underset{i}{\text{argmax}}\, V_{ij}^2$$

Using the fact that $V = A^\top U \Sigma^{-1}$, we can rewrite this selection criterion as:

$$\begin{aligned}
\underset{i}{\text{argmax}}\, V_{ij}^2 &= \underset{i}{\text{argmax}} \left(A^\top U \Sigma^{-1}\right)_{ij}^2 = \underset{i}{\text{argmax}} \left(A^\top U\right)_{ij}^2 \Sigma_{jj}^{-1} \\
&= \underset{i}{\text{argmax}} \left(s_i^\top U_{\bullet j}\right)^2 = \underset{i}{\text{argmax}}\, \text{tr}(U_{\bullet j} U_{\bullet j}^\top s_i s_i^\top) \\
&= \underset{i}{\text{argmax}}\, \text{q}\left(\underline{S}_{\mathcal{D}}^{(j)} | \rho = s_i s_i^\top\right)
\end{aligned} \tag{3.9}$$

where $\mathcal{S}_{\mathcal{D}}^{(j)}$ is the one-dimensional subspace associated with the $j^{\text{th}}$ column of $U$, i.e. to the $j^{\text{th}}$ latent information unit. Hence, the selection process corresponds to maximizing the probability associated with the $j^{\text{th}}$ dimension of the subspace $S_{\mathcal{D}}$ that represents the salient topics of the documents to summarize. This means that a sentence that is a combination of two information units ($j_1$) and ($j_2$) might not be selected because it lies half way between the subspaces $\mathcal{S}_{\mathcal{D}}^{(j_1)}$ and $\mathcal{S}_{\mathcal{D}}^{(j_2)}$. However, this topic, according to the hypotheses of the `QIA` framework, is fully contained with the topics of the documents, and would constitute a good candidate for the summary.

This is an illustration of the problem of the hard clustering existing in Gong and Lin selection method. This problem is further exacerbated when singular values are close to each other. In the extreme case where they are equal, i.e. $\sigma_{j_1}$ and $\sigma_{j_2}$, the `SVD` problem is degenerate, i.e. the two vectors can be any two that define the same two-dimensional subspace, making the criterion arbitrary and sensitive to numerical approximations.

Steinberger and Ježek (2004) also noticed this problem. Although they did not give a principled explanation of the underlying reason, they noted that a sentence can be highly ranked for many information units but never sufficiently to be selected. The approach they proposed

[3] If there are less that $k$ non-null singular values, the method cycles through the singular values, beginning with the highest ones.

is to first select an appropriate rank $k$ for approximation of the matrix $A$. Then, they proposed to select the $i^{\text{th}}$ sentence that maximizes the following criterion:

$$g_i = \sum_{j=1}^{k} V_{ij}^2 \sigma_j^2 = \text{tr}\left(V_{i\bullet}\Sigma^2 V_{i\bullet}^\top\right)$$

Since $V_{i\bullet}$ equals $s_i^\top U\Sigma^{-1}$, we have

$$g_i = \text{tr}\left(s_i^\top UU^\top s_i\right) = \text{q}\left(\underline{S_\mathcal{D}}|\underline{\rho}_{s_i}\right) \tag{3.10}$$

where $s_i$ is a pure information unit state, i.e. we know that the information unit is $s_i$. Hence, this criterion selects sentences maximizing the probability of being present in the most important (i.e. $k$) document topics.

This method has two shortcomings. First, it assumes that the dimension of $\mathcal{S}_\mathcal{D}$ is correctly chosen, because if the rank is maximal the probability defined by Equation 3.10 is always equal to 1 since $S_\mathcal{D}$ is a subspace that contains all the atomic vectors present in the documents of $\mathcal{D}$. Second, different to (Gong and X. Lin 2001), sentences close to only one SVD information unit can be selected repeatedly. While for important information units, i.e. those with high singular values, this can be a good property, it may lead to too much homogeneity in the summary. In the worst case, a sentence that occurs more than one time in the documents to summarize can be chosen repeatedly.

The QIA-based approach, defined in Equation (3.8), caters for information units that (1) are combination of the SVD information units, hence overcoming (Gong and X. Lin 2001) problems, and that (2) extract sentences from different topics, hence overcoming the limitations of (Steinberger and Ježek 2004).

**Experimental results** We conducted our experiments on the DUC 2005 to 2007 datasets[4]. Documents consist of news articles collected from TREC for DUC 2005 and the AQUAINT corpus for DUC 2006 and 2007. We were interested in the main task[5] of DUC 2007, i.e. providing a summary of no more than 250 words for each topic to answer the associated question. For a given question, a summary is to be formed on the basis of a subset of documents to its corresponding topic.

Table 3.2 reports the results on the models where hyperparameters were optimized on held-out training sets. Thus, the results reflect the ones we would have obtained on one DUC collection, when the summaries of the two others are available for parameter tuning.

We also compare the results with two graph-based models (symmetric non-negative matrix factorization (SNMF) and Lexrank); two baseline systems, namely lead and random; and the best competing summarization system in DUC 2005, DUC 2006 and DUC 2007, denoted by Best@DUC.

The lead baseline returns all the first sentences (up to 250 words) in the most recent document for each topic and the random baseline selects sentences randomly.

**SNMF** conducts symmetric non-negative matrix factorization on a sentence-sentence similarity matrix (D. Wang, T. Li, et al. 2008), the hyper-parameter $\lambda$ for computing sentence scores was fixed to 0.7 which gave best results on all three DUC collections.

**Lexrank** defines a random walk model on top of a graph where sentences to be summarized define its nodes and the edges represent the similarity measures between the nodes of the graph. Sentences are then scored by the expected probability of a random walker visiting each sentence Erkan and Radev 2004. Here, the cosine threshold $t$ was fixed to 0.1 leading to best results with this approach.

[4] http://www-nlpir.nist.gov/projects/duc/data.html

[5] We ignored the short summary task (less than 100 words), which was abandoned in 2008 because of its difficulty for extractive summarization methods.

| Metric | DUC 2005 | | DUC 2006 | | DUC 2007 | |
|---|---|---|---|---|---|---|
| **Model / DUC** | ROUGE-2 | ROUGE-SU4 | ROUGE-2 | ROUGE-SU4 | ROUGE-2 | ROUGE-SU4 |
| Best@DUC | 0.072 | 0.133 | 0.095 | 0.155 | 0.123 | 0.175 |
| Average@DUC | 0.060 | 0.115 | 0.075 | 0.132 | 0.096 | 0.150 |
| Lead | 0.043 | 0.093 | 0.053 | 0.104 | 0.065 | 0.113 |
| Random | 0.041 | 0.091 | 0.049 | 0.101 | 0.060 | 0.110 |
| LexRank | 0.076 | 0.136 | 0.093 | 0.150 | 0.120 | 0.172 |
| **SNMF** | 0.060 | 0.121 | 0.085 | 0.140 | 0.110 | 0.158 |
| Gong | 0.072 | 0.133 | 0.087 | 0.148 | 0.118 | 0.180 |
| Murray | 0.073 | 0.135 | 0.086 | 0.147 | 0.120 | 0.181 |
| Ozsoy | 0.071 | 0.133 | 0.085 | 0.145 | 0.111 | 0.173 |
| Steinberger | 0.071 | 0.133 | 0.081 | 0.144 | 0.111 | 0.169 |
| **QIA** | **0.077** | **0.135** | **0.091** | **0.151** | **0.127** | **0.185** |

Table 3.2 – Final evaluation on the held-out corpus. The first three rows give respectively the performance of the best system in DUC, the random and lead strategies.

**QIA** We used a greedy approach, i.e. at each step we select the sentence $s_n^*$ that maximizes the criterion given by Equation 3.8 if added to an already constructed set of sentences $s_1^*, \ldots, s_{n-1}^*$. That is, $s_n^*$ is given by

$$s_n^* = \operatorname*{argmax}_s \mathrm{q}\left(\mathcal{S}_{s_1^*,\ldots,s_{n-1}^*,s}|\mathcal{D}\right)$$

We can see that our results match the main conclusion drawn in the previous sections, although parameters vary slightly depending on the specific corpora on which they were optimized. More precisely, for LSA-based approaches, TF and unigram/strict bi-grams with POS filtering perform the best, and including the different priors was important. The parameters are quite different for POS-based models, where a TF-IDF weighting scheme on unigrams, with uniform prior over sentences, perform the best in general.

From a performance point of view, we improved substantially all the LSA-based models by selecting appropriate indexing units (in particular, using part-of-speech tagging, as suggested in Ozsoy, Cicekli, and Alpaslan 2010) and using priors on sentences in the document be summarized, as suggested by the QIA approach. Those priors are biased towards the topic and the length of the sentences.

In all cases, we can observe that the QIA-based model performs the best for both metrics. The performance of both QIA-based models are over those of the best systems in DUC for the corresponding years (not significant except for ROUGE-SU4 in DUC 2005 and 2007); in particular, this means that in 2007 QIA-based models would have been ranked first since the data from 2005 and 2006 was available.

Finally, QIA-based models are in most of the cases performing better (significantly in 2007 for Lexrank and 2005-07 for SNMF) than two state of the art extractive summarization methods, namely SNMF and Lexrank, thus showing that the QIAframework is a very promising approach for extractive summarization.

In summary, our experimental results show that when summarization is performed on a set of relevant documents to a given topic (topic-oriented documents), as it is the case with the DUC collections, QIA-based models are able to implicitly capture the topics covered by the set of documents and are less sensitive to varying documents or sentence lengths. This is an important result as it means that the similarity estimations between sentences and the topic, performed by most systems in these competitions, is not required by the QIA-based models. Indeed, the latter uncover automatically, without relying explicitly on the DUC-provided topic at hand, the important information units covered by a set of topic-oriented documents

More precisely, we showed that even though LSA and QIA-based techniques are based on spectral decomposition, these models differ in the choice of their optimal parameters. LSA-based approaches benefit from the various pre-processing steps (part-of-speech, bi-grams, topic and length bias, rank selection) whereas QIA-based approaches rely on the standard IR TF-IDF scheme and a few (typically one) information units that represent the important topics of the documents to summarize. This difference is due to the criteria used to select sentences. LSA-based models do not consider the "topical" space covered by a set of extracted sentences, whereas the ones from QIA-based models do.

This leads to an important conclusion. The topical space, in the case of summarization, resembles more a TF-IDF term space than a TF term space, which can be linked to the QIA hypothesis on the linear combination of information units. Such a linear combination makes more sense when less important terms (i.e. low IDF) do not influence much the result of the linear combination.

### 3.4.3 Kernel approach

The results in information retrieval being somewhat unsatisfactory (Piwowarski, Frommholz, Lalmas, et al. 2010), it was necessary to try to exploit spaces of different natures to represent the documents. Kernels (Lanckriet et al. 2004) allow to work in Hilbert spaces of potentially infinite dimension. I created tools to compute and manipulate « quantum probability distributions » (Piwowarski 2012), which I expanded to allow learning the parameters of the linear combination of kernels. This work was based on:

1. Techniques for updating quantum densities based on singular value decomposition updates (Grant et al. 2018) ;

2. Techniques for reducing the number of pre-images used by the kernel-based learning methods (Weston, Schölkopf, and Bakir 2004) .

These tools were presented during two tutorials in conferences in Information Retrieval (ECIR 2012 and ICTIR 2013), and aim to facilitate experiments with this type of models.

These tools also allowed me to conduct new experiences in information retrieval. Rather than assuming that the space representing documents is reduced to the space of terms (or to a reduced space), the purpose of this work was to explore potentially infinite spaces for representing documents. Based on annotated databases (documents, questions and the associated assessed documents), I tried to use different types of criteria for learning: scheduling (Burges et al. 2005), direct optimization of measures evaluation in search of information using black box optimization techniques (Le Digabel 2011), but these experiments were unsuccessful. There are several reasons for this:

1. The numerical complexity of the algorithms does not make it possible to use sufficiently fine approximations ;

2. The initial representation is not suitable, and the use of kernels does not correct this problem

## 3.5 Discussion and perspective

In this chapter, I presented the work conducted on leveraging the quantum probability formalism, which crystallized around the Quantum Information Access framework – in which information "units" are equated to physical states of quantum physics. In addition to the presented results, the QIA framework has been applied to document filtering (Piwowarski, Frommholz, Moshfeghi, et al. 2010) and to support poly-representation (Frommholz, Larsen, et al. 2010), i.e. information

objects that are heterogeneous (e.g. a page describing a product, with its associated reviews and technical characteristics).

In spite of positive results in extractive summarization (Piwowarski 2012), the experiments in information retrieval did not allow to improve the performances nor to find new directions which would make it possible to solve the problems met, and this, despite the development of kernel-based approaches which could bring more flexibility into the quantum formalism.

At that time, the development of neural-based approaches to language processing, which share many concepts with QIA (superposition, geometric relationships within vector spaces linked to a probability distributions) together with a much simpler and direct optimization approach, is much more promising.

There are however many ideas that could be useful and applied to current neural network models. For instance, there is a link with recent works in continuous language models (S. Kumar and Tsvetkov 2018). At a more fundamental level, there is a link with unitary transformations (Arjovsky, Shah, and Bengio 2015) which preserve the norm of the input vectors (and could act on densities, i.e. by allowing them to evolve).

**Outcomes**

- B. Piwowarski and M. Lalmas (2009a). "A Quantum-based Model for Interactive Information Retrieval." In: *Proceeedings of the 2nd International Conference on the Theory of Information Retrieval*. Ed. by L. Azzopardi et al. Vol. 5766. Lecture Notes in Computer Science. Cambridge, United Kingdom: Springer
- B. Piwowarski and M. Lalmas (2009b). "Structured Information Retrieval and Quantum Theory." In: *Proceedings of the 3rd QI Symposium*. Ed. by P. Bruza et al. Vol. 5494. Springer
- I. Frommholz, B. Larsen, et al. (Aug. 2010). "Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework." In: *Proceedings of the third symposium on Information interaction in context*. New Brunswick, NJ, USA. DOI: 10.1145/1840784.1840802
- B. Piwowarski, I. Frommholz, M. Lalmas, et al. (2010). "What can Quantum Theory bring to IR?." In: *Proceedings of the nineteenth ACM conference on Conference on information and knowledge management*. Ed. by J. Huang et al. ACM. DOI: 10.1145/1871437.1871450
- B. Piwowarski, I. Frommholz, Y. Moshfeghi, et al. (2010). "Filtering documents with subspaces." In: *ECIR*. ed. by C. Gurrin et al. Vol. 5993. Advances in Information Retrieval. Springer
- A. Caputo, B. Piwowarski, and M. Lalmas (2011). "A Query Algebra for Quantum Information Retrieval." In: *Proceedings of the 2nd Italian Information Retrieval Workshop*
- I. Frommholz, B. Piwowarski, et al. (Apr. 2011). "Processing Queries in Session in a Quantum-inspired IR Framework." In: *Proceedings of the 33rd European conference on Advances in information retrieval*
- G. Zuccon, B. Piwowarski, and L. Azzopardi (2011). "On the use of Complex Numbers in Quantum Models for Information Retrieval." In: pp. 346–350. DOI: 10.1007/978-3-642-23318-0_36
- B. Piwowarski (2012). *The Kernel Quantum Probabilities (KQP) Library*. Tech. rep. 1203.6005v2

- B. Piwowarski, M.-R. Amini, and M. Lalmas (2012). "On Using a Quantum Physics Formalism for Multidocument Summarization." In: *Journal of the American Society for Information Science and Technology* 63, pp. 865–888. DOI: 10.1002/asi.21713
- Tutorials: ECIR 2012 and ICTIR 2013

# Chapter 4

# Graphs and Gaussian Representations

Gaussian representations, or more generally probabilistic representations, are interesting because they allow to represent the uncertainty of information – this was e.g. not possible with the quantum probability framework presented in the previous chapter.

This information about the uncertainty of representations is particularly suited for representing each node of a graph – since for some nodes, a lot of information is known (they have many neighbors and/or have many meta-information associated with them). For some others, almost nothing is known. In the former case, nodes can be represented with a distribution with a low variance, and in the latter, with a distribution with high variance.

In this chapter, after presenting related works, we focus on the use of Gaussian Representations for two information access tasks:

- Node Classification (section 4.2): In the context of social networks, such models can be used to classify items or users into sets of predefined categories. An example of application is to automatically label musicians in a musical social graph (e.g. LastFM), allowing users to browse categories.
- Recommendation (section 4.3), or more precisely collaborative filtering, where the goal is to recommend items to users based on the interaction history (e.g. given ratings) of all users.

These two tasks were chosen because they are representative of the case where uncertain representations are useful since in this type of social networks, many entities (e.g. users, items) are not associated with much information (e.g. a user has only rated two movies).

## 4.1 Capturing uncertainties in representations

Before presenting our own models, we first describe different approaches and models able to modeling representation uncertainty (besides the quantum formalism described in the previous chapter, which lacks flexibility since the variance is "built in" the formalism) through the learning of (latent) densities. There is a long line of previous work in mapping entities to probability distributions, but few represent entities directly as density distributions in a latent function space. Density representations allow to map entities not only to vectors but to regions in space, modeling uncertainty, inclusion, and entailment, as well as providing a rich geometry of the latent space.

In the following, we give a state of the art of models using densities to represent entities. We firstly give a short view of Bayesian approaches, whose posteriors can be interpreted as density representations of nodes. We then focus on Gaussian embeddings, in which both means and variances are directly learned from specific loss function over the densities.

### 4.1.1 Bayesian Approaches

Leveraging the Bayesian framework, learning latent distributions is equaled to finding the posterior marginal distributions over all individual latent variables. We suppose that each entity $e$ is represented by a random variable $\underline{Z}_e$ in $\mathbb{R}^d$. The goal is to infer the posterior $P(\{\underline{Z}_e\}_{e\in\mathcal{E}}|\mathcal{D})$, where $\mathcal{D}$ is the observed data – which in the case of graphs can encompass a variety of information such as links between entities or meta-information associated with nodes.

In the context of relational data, this type of approach has been used in task such as classification (Airoldi et al. 2005; R. Xiang and Neville 2013), data collaborative completion (Kim and Choi 2014) or collaborative filtering (Stern, Herbrich, and Graepel 2009). Airoldi et al. (2005) and R. Xiang and Neville (2013) address collective classification by inferring the posterior distribution of the class degrees of membership for each node. Kim and Choi (2014) propose a graphical model for collaborative completion that handle non-random missing data. For collaborative filtering, Stern, Herbrich, and Graepel (2009) proposed an inference model and approximate the posterior distribution by a product of Gaussian distributions. Given a set of rating tuples (user, item, rating), they learn the approximate posterior distributions. The optimization is accomplished using variational message passing (Bishop 2006). Thus, they achieve to learn posteriors mean and covariance of each latent representations, which they use to approximate the distribution as a normal. They then use these learned representations to predict the rating of an item for a user.

For most of the interesting models, algorithms performing exact inference are computationally intractable for all but the simplest models. There exists two ways to approximate probability distribution, either by sampling (Monte Carlo) or by introducing a distribution approximating the original one, but simpler to compute (variational approaches).

On the one hand, it is possible to use Monte Carlo Markov Chain (MCMC) sampling from the posterior distribution and then train a model on those samples. However, MCMC methods are often slow to converge and in many cases variational methods are preferred since they are more sample efficient, and are usually faster.

Variational inference aims at approximating the (intractable) posterior distribution, denoted $P$, by a (simpler and tractable, e.g. making independence assumptions between variables) variational one, denoted $Q$, i.e. $Q(Z|X) \approx P(Z|X)$, where $Z$ is the set of latent variables and $X$ the observations. Different variational families have been studied to complete this optimization, but they all require thoughtful decisions on how to introduce independence assumptions into the model.

### 4.1.2 Probabilistic Embeddings

Recently, instead of using Bayesian approaches, several works have proposed to learn these densities "directly". The main idea is to derive a cost function which is directly based on the densities, and to train discriminatively instead of estimating posteriors. This allows to use standard optimization toolchains, and, more importantly, allows for much more freedom in the expressive power of the optimized cost. For example, they allow to compare two distributions, which is much hard to express in a Bayesian setting.

As far as I know, all the works using probabilistic embeddings use Gaussian embeddings, since they are simple to estimate and to optimize for a variety of distances between distributions, such as KL-divergence or the inner product distance. The KL-divergence is particularly interesting naturally asymmetric, and has a simple geometric interpretation as an inclusion between families of ellipses.

We begin by describing this new framework, used in our contributions, namely Gaussian representation learning. Introduced in (Vilnis and McCallum 2014), it consists in learning representations in the space of Gaussian distributions. We denote $Z \sim \mathcal{N}(\mu, \Sigma)$ a Gaussian representation characterized by its mean $\mu$ and its covariance matrix $\Sigma$.

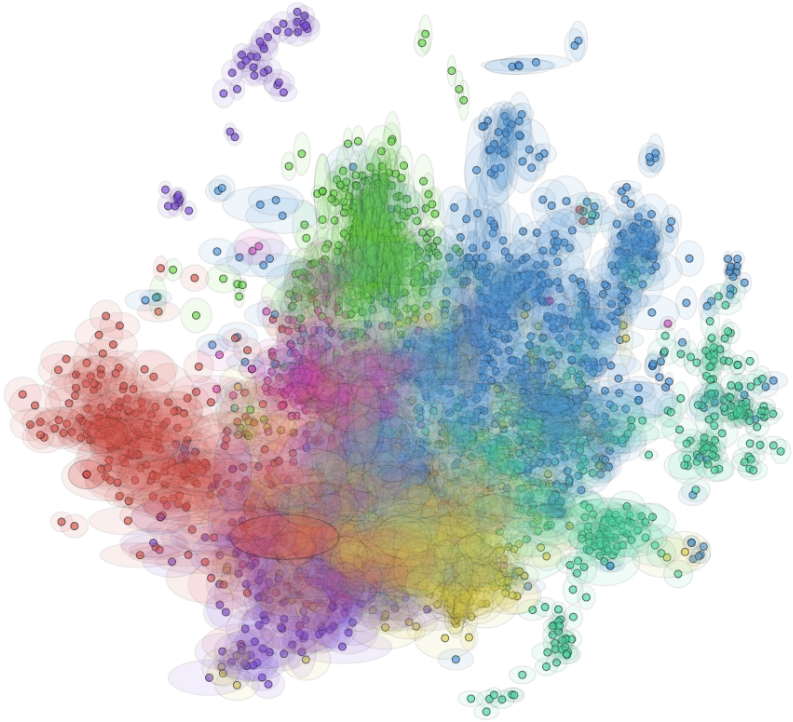


Figure 4.1 – Learned two dimensional Gaussian representations of nodes on the Cora dataset from (Bojchevski and Günnemann 2017a). Color indicates the class label, and ellipses around points indicate one standard deviation.

Figure 4.1 shows a visualization of Gaussian representations learned by the unsupervised model presented in (Bojchevski and Günnemann 2017a) for the Cora dataset[1]. Colors indicate the class label (not used during training).

The work presented in (Vilnis and McCallum 2014) has the same goal as (Mikolov et al. 2013), i.e. learning unsupervised representations of words using their context. To learn Gaussian representations, Vilnis and McCallum (2014) proposed a ranking-based loss, following (Bordes, Usunier, et al. 2013). They chose a max-margin ranking objective, which pushes the energy of positive pairs below negative ones by a margin:

$$L(i, p, n) = \max(0, 1 - E(\underline{Z}_i, \underline{Z}_p) + E(\underline{Z}_i, \underline{Z}_n))$$

where $E(\underline{Z}_i, \underline{Z}_{p/n})$ is the energy associated to the pair of words $i$ and $p/n$ given their Gaussian representations. For a given word $i$, $\underline{Z}_p$ is a Gaussian representation of a positive context for word $p$ (i.e. that occurs within a given window around term $i$) and $\underline{Z}_n$ a Gaussian representation of a negative one (randomly sampled, as in Mikolov et al. 2013).

They proposed two functional forms for the energy function $E$, one symmetric and one asymmetric. The symmetric form did not perform as well, and is less interesting since it does not express the fact that the context of a word should *contain* the word representation. To circumvent this limitation, they proposed to use the Kullback-Leibler divergence, denoted KL-divergence or $D_{\mathrm{KL}}$:

[1] The Cora dataset consists of scientific publications classified into one of seven classes together with a citation network.

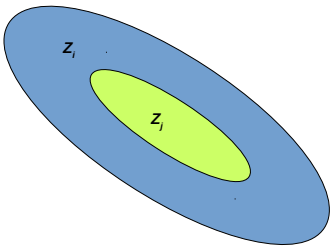



Figure 4.2 – Example of a low $D_{\mathrm{KL}}(\underline{Z}_j||\underline{Z}_i)$.

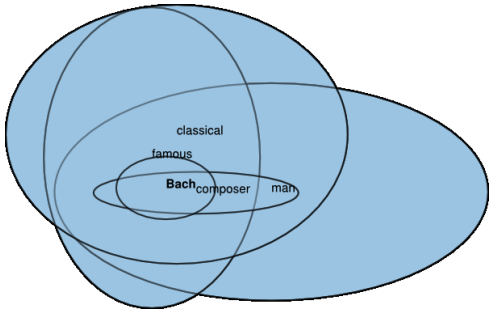



Figure 4.3 – Learned Gaussian representations with diagonal variances, from (Vilnis and McCallum 2014). The first letter of each word indicating the position of its mean. We can see for instance that "Bach" is "included" in "composer", which in turn is "included" in *man*.

$$\begin{aligned} -E(\underline{Z}_i, \underline{Z}_j) \overset{def}{=} D_{\mathrm{KL}}(\underline{Z}_j||\underline{Z}_i) &= \int_{z\in\mathbb{R}^d} \mathcal{N}(z;\underline{\mu}_j,\underline{\Sigma}_j) \log \frac{\mathcal{N}(z;\underline{\mu}_j,\underline{\Sigma}_j)}{\mathcal{N}(z;\underline{\mu}_i,\underline{\Sigma}_i)} dz \\ &= \frac{1}{2}\left(\mathrm{tr}(\underline{\Sigma}_i^{-1}\underline{\Sigma}_j) + (\underline{\mu}_i - \underline{\mu}_j)^T \underline{\Sigma}_i^{-1}(\underline{\mu}_i - \underline{\mu}_j) - d - \log \frac{\det(\underline{\Sigma}_j)}{\det(\underline{\Sigma}_i)}\right) \end{aligned} \tag{4.1}$$

KL-divergence is a natural energy function for modeling entailment between concepts. A low KL-divergence $D_{\mathrm{KL}}(\underline{Z}_j||\underline{Z}_i)$ indicates that we can encode $\underline{Z}_j$ easily as $\underline{Z}_i$, implying that $\underline{Z}_j$ entails $\underline{Z}_i$. This can be more intuitively visualized and interpreted as a soft form of inclusion between the Gaussian distributions of the two Gaussian representations – if there is a low KL-divergence, then most of the mass of $\underline{Z}_j$ lies inside $\underline{Z}_i$ as shown in Figure 4.2. Figure 4.3 shows what kind of Gaussian representations are learned by the model proposed by (Vilnis and McCallum 2014).

This allows (Vilnis and McCallum 2014) to learn the Gaussian representations by gradient descent. (Vilnis and McCallum 2014) has inspired several recent work using the same framework on different tasks and datasets such as (S. He et al. 2015) for knowledge graphs datasets, (Mukherjee and Hospedales 2016) for image classification, (Bojchevski and Günnemann 2017a) for attributed graphs with an inductive unsupervised framework and (Purpura et al. 2019) for information retrieval.

This work has also inspired three of our contributions on graph node classification (section 4.2.2), forecasting relational time series (not presented here) and collaborative filtering (section 4.3). We focus next on node classification and collaborative filtering, since they are more related to the information access field.

## 4.2 Node classification

The need for graph node classification stems from several application domains, like web data mining or biology. For example, web page classification may be formulated as a homogeneous

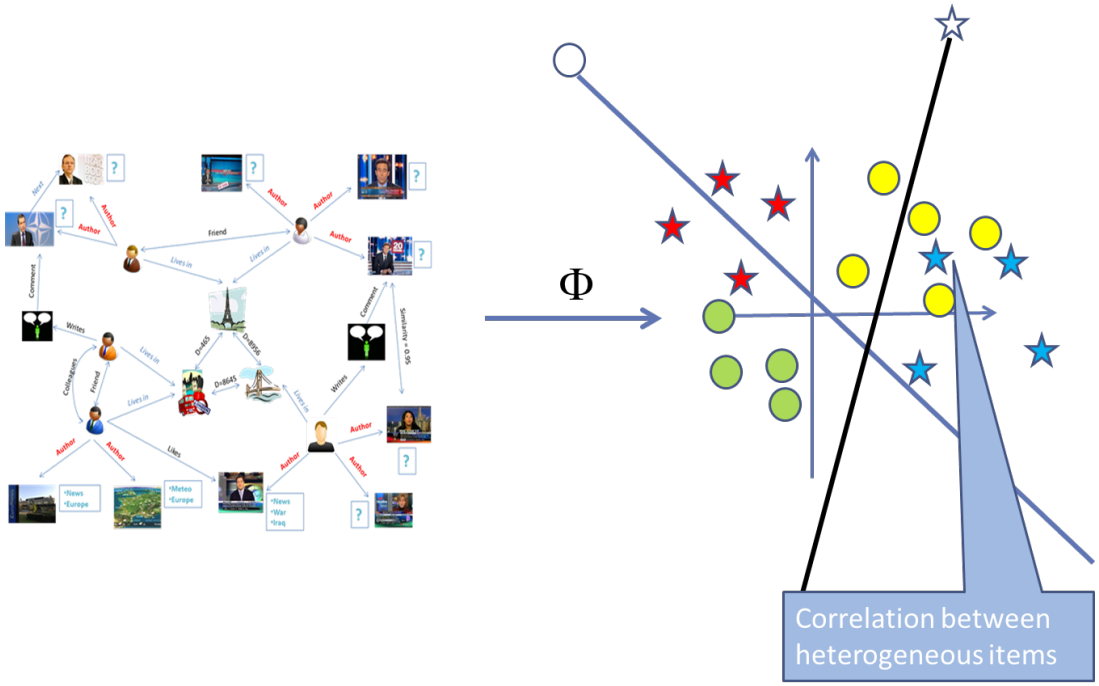



Figure 4.4 – Drawing of representation learning for relational data where nodes are projected from the initial graph (left) to the common latent space (right) where classifiers are learned.

graph node classification task where nodes are web pages, edges between web pages are hyperlinks and node labels are the web page topic.

Work in this domain firstly focused on homogeneous graphs, i.e. with only one type of node and one type of relation. The analysis of more complex data, e.g. coming from social sources or knowledge bases, where nodes may be of different types and share different and sometimes multiple relations, requires new models and techniques. Heterogeneous graph nodes classification is a recent trend and an open problem. We review below the main directions for graph nodes classification.

An illustration of representation learning for relational graphs is given in Figure 4.4. Each of the nodes of the graph is projected onto $\mathbb{R}^d$, and is thus associated with a learned vector representation. In this example, we consider two types of nodes and distinguish them in the figure representing the latent space: the first type is represented as circles and the second as stars. This can be used in classification, as shown in the figure where two classifiers are learned, one for the stars and the other one for the circles. In this example, circles and stars correspond to two different types of nodes. Each type has to be classified among two classes, green/yellow for circles and red/blue for stars, the two types of nodes are linked so that their representations influence each other.

For graphs, supervised learning models learn representations and classifiers at the same time. The classification task has some specificities. For example, even if labels can be seen as a new type of node (with each node linked to its corresponding labels), the node-label relation is very particular and handling them separately in the model performs better. That is why the node-label relation needs to be treated differently in the loss function.

### 4.2.1 Related works

Most of the models share the same basic intuition (Getoor 2007): neighboring nodes have similar properties, e.g. they tend to be classified similarly. In the following, we distinguish models that are:

**Propagation-based** Associate with each node a classification score for each possible label, and propagate this information to neighboring nodes.

**Representation-based** Associate a vector in $\mathbb{R}^d$ with each node of the graph, and propagate this representation to neighboring nodes.

**Propagation-based method** *Iterative Classification Algorithms* (ICA) (Sen et al. 2008) are extensions of classical inductive classification schemes to relational data. They consist in iteratively building a local classifier at each node, using as inputs both node characteristics and statistics on the node neighbors current labels. Standard models, like Bayesian classifiers (Neville and Jensen 2000) or logistic regression (Q. Lu and Getoor 2003), are used to perform the classification task. As ICA methods can be heavily influenced by the absence of links, Gallagher et al. (2008) proposed a way to predict new links by connecting unlabeled nodes to labeled ones to circumvents the potential graph sparsity. ICA methods have been extended to multi-relationship graphs, by estimating how labels propagate through each type of relationship or type of nodes (S. Peters, Denoyer, and Gallinari 2010; X. Wang and Sukthankar 2013), but dealing with multiple node and relation types is still an issue for this family of methods.

*Random walks* have been used for graph nodes classification. Labels are propagated from labeled to unlabeled nodes using the graph structure: each label has a given probability to be propagated to the neighboring nodes, and the stationary distribution corresponds to the scores given to the different labels. The most typical work, for homogeneous graphs, is that of X. Zhu and Ghahramani (2002) who proposed *Homogeneous Label Propagation (HLP)*, the stationary distribution of a random walk is used for classification. This branch of research has motivated a large number of extensions, as for example taking into consideration more specific information, such as graph communities (Devooght et al. 2014) or label distribution (Nandanwar and Murty 2016). For heterogeneous graphs, there are two approaches that amount at modifying the random walk by taking into account different types of nodes and relationships. One defines specific random walks (Y. Zhou and L. Liu 2014) based on input features and graph statistics. The specific definition of the random walk has to be done manually, which is a clear limit of this type of approach. General models, like *Graffiti* (Angelova, Kasneci, and Weikum 2012a), are based on simple extensions of the random walk process: *Graffiti* is based on two intertwined random walks, the first one operating between nodes of the same type which are directly connected in the graph while the other one is defined between nodes of the same type which are connected through another node type. While simple, *Graffiti* is very competitive: according to our experiments with different models, it represents the state of the art in the domain and is thus one of our baselines.

*Regularized models*, while related to random walks, are different since the objective is formulated as loss optimization, when random walks do not make use of an explicit loss. A diffusion equation, appearing as a regularization term in this loss propagates through connected nodes. For instance, (D. Zhou, J. Huang, and Schölkopf 2005) use a regularization inspired by random walks, and (Pimplikar et al. 2014; J. Wang, Jebara, and Chang 2008) minimize the graph-based distance between the nodes using graph Laplacian matrices. Recently, (Ye and Akoglu 2015) tackled the homogeneous multi-relational classification case, by assuming that relations have varying levels of informativeness, and, associated with them, weights that are learned. Models for heterogeneous graphs have been proposed in the case where the label set is the same for all node types (Hwang and Kuang 2010; Ji et al. 2010b). In that case, simply ignoring the node type allows one to use frameworks developed for the homogeneous case (Hwang and Kuang 2010). Ji et al. (2010b) have nevertheless made a step towards heterogeneous networks by introducing weights on relations between node types and thus differentiating in some way the different types of nodes.

**Representation learning and Deep Learning** In the context of graphs, representation learning amounts at learning a representation in $\mathbb{R}^d$ of each node of the graph. This representation can then be used as an input for any appropriate classifier. Inspired from deep learning techniques, (Fan and B. Huang 2017; Moore and Neville 2017) tackle the homogeneous graph node classification using Recurrent Neural Networks (RNN). They transform each set of neighbors for a given node into a sequence of their attributes and use a RNN to predict the label of the last (test) node of the sequence. (T. Pham et al. 2016) use a neural network where the hidden representations are computed based on the input graph. Contrary to our model, these methods require node characteristics (inductive learning), whereas we only depend on the structure (though content could be integrated within the proposed models quite easily).

Several works focused on learning a node representation *without* supervision. Inspired by work on word representation learning of (Mikolov et al. 2013), (Perozzi, Al-Rfou, and Skiena 2014; Tang et al. 2015) learn distributed representations of graph nodes using a random walk. In *LINE*, (Tang et al. 2015), the joint probability of neighbor nodes is a function of the inner product of their representations. In DeepWalk (Perozzi, Al-Rfou, and Skiena 2014), the authors propose to sample *paths* within the graph, and then use them as sentences fed to Skip-Gram (Mikolov et al. 2013) to learn the representations of the nodes. Based on this work, (Cao, W. Lu, and Q. Xu 2015) learns node representations that incorporate global and local structural information. These representations can then be used as inputs to a classifier. Similarly, Grover and Leskovec 2016 proposed an unsupervised learning model that maximizes the likelihood of preserving a network neighborhood of nodes. In practice, this neighborhood is defined through a biased random walk. Deep learning architectures have been used in recent works with convolution neural networks Niepert, Ahmed, and Kutzkov 2016 or auto-encoders (D. Wang, Cui, and W. Zhu 2016).

Finally, our model belongs to the class of *semi-supervised transductive representation-based models*. Closely related to our model are the works of (Tu et al. 2016; Z. Yang, Cohen, and Salakhutdinov 2016), both based on DeepWalk (Perozzi, Al-Rfou, and Skiena 2014). Z. Yang, Cohen, and Salakhutdinov (2016) have recently proposed a model that learns two different representations for a node, one from features associated with the node (e.g. the text of a web page), and another one from the graph alone. The main focus of this model was to take into account external information, while in our work the focus is on handling various types of relationships between nodes without considering node characteristics. Tu et al. (Tu et al. 2016) proposed a Maximum-Margin Deep Walk (MMDW) procedure that couples DeepWalk with an SVM. There are several differences between our model and MMDW: their regularization term is based on an inner product and not on a distance like ours, our model uses a simpler optimization procedure and learns to weight the relationship. We compared with this model in our experiments, and have shown that our choices lead to improved performance in classification.

Finally, in the field of knowledge-based representation, a lot of work has been conducted (Bordes, Collobert, et al. 2011; Bordes, Usunier, et al. 2013; B. Yang et al. 2014), where the goal is to project each node of a graph into a vector space by exploiting the data related to the graph. Although they deal with heterogeneous graphs, these methods are specialized for the processing of specific graphs derived from the semantic Web, and are not very adapted to the processing of social graphs. Moreover, the criterion they optimize is not specifically related to the classification task that is the focus of our work in many cases, and the links between nodes of the graph all have the same importance in the optimization process. It is not the case in any graphs.

### 4.2.2 Formalization

A heterogeneous network is modeled as a directed graph $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ where $\mathcal{U}$ is the set of nodes and $\mathcal{E}$ the set of edges. Each node $u \in \mathcal{U}$ of the graph has a type $t_u \in \mathcal{T}$ where $\mathcal{T}$ is the set of types. Similarly, each edge in $\mathcal{E}$ is defined as a triplet $(u, v, r) \in \mathcal{U} \times \mathcal{U} \times \mathcal{R}$ where $\mathcal{R}$ is the set

of of relationships that can hold between two nodes. We denote $\mathcal{U}_u \subseteq \mathcal{U} \times \mathcal{R}$ the neighbors of $u$.

Regarding the classification task, let $\mathcal{Y}^t$ denote the set of categories associated with nodes of type $t$. For a node $u$, the label $y$ indicates whether $u$ belongs to category $c$ (if $y_c = 1$) or not (if $y_c = -1$). A supervised dataset is noted $\mathcal{D} = \{(u, y)\}$ where $u \in \mathcal{U}$ and $y \in \mathcal{Y}$.

We learn the representations of nodes and classifiers parameters by minimizing an objective loss function. It takes the general form of transductive regularized loss (Ji et al. 2010a; D. Zhou, Bousquet, et al. 2003):

$$L(\underline{Z}, \theta; \mathcal{D}, \mathcal{G}) = L_C\left(\underline{Z}, \theta; \mathcal{D}\right) + \lambda L_G\left(\underline{Z}; \mathcal{G}\right) \tag{4.2}$$

where $L_C$ is a classification loss and $L_G$ is a graph regularization loss with $\lambda \in \mathbb{R}$ is a regularization weight.

The idea of using a projection of nodes in a continuous space was presented in (Jacob, Denoyer, and Gallinari 2014). We extended this work in (L. D. Santos et al. 2018), by learning continuous hyperparameters and providing extensive experiments (section 4.2.3), and finally in (Dos Santos, Piwowarski, and Gallinari 2017), using probabilistic representations – this latter work is the focus of this chapter.

As for classical transductive graph losses, the minimization in Equation (4.2) aims at finding a trade-off between the difference between observed and predicted labels $y$ in $\mathcal{D} = \{(x, y)\}$, and the amount of information shared between two connected nodes. There are however major differences, since here $\underline{Z}$ is not a label as in classical formulations (e.g. ICA), but (a distribution over a) node embedding. Finally, the function $f_\theta(.)$ is a parametric classifier for a node of type $t$ – there is one such classifier for each node type. Since we are using Gaussian embeddings, the $\underline{Z}$s are random variables and the regularization term is a dissimilarity measure between distributions.

#### 4.2.2.1 Classifier Loss

The mapping onto the latent space is learned so that the labels of each type of node can be predicted from their embedding. We set the problem as a multi-label classification problem, i.e. each node can belong to zero or more categories $c$ in those allowed by the node type $t_u$.

For that, we use a parametric classification function $f_\theta$ whose parameters depend on the label $c$. In this probabilistic setting, this multivariate function takes as input a node representation and outputs a vector of score distributions for each label corresponding to the node type. The parameters $\theta$ of the classifier are learned by minimizing the following loss on labeled data:

$$L_C\left(\underline{Z}, \theta; \mathcal{D}\right) = \sum_{(x,y)\in\mathcal{D}} \sum_{c\in\mathcal{Y}^{t_u}} \Delta_C(f(\underline{Z}_x; \theta_c), y_c) \tag{4.3}$$

where $\mathcal{Y}^{t_u}$ are the potential classes of a node of type $t_u$ and $\Delta_C(f(\underline{Z}_u; \theta_c), y)$ is the loss associated with predicting labels $f(\underline{Z}_u; \theta_c)$ given the observed label $y_c$.

For each category $c$, we use a linear classifier function $f^c_\theta$. Even though more complex functions could be used, we make the hypothesis that the representation space is sufficiently large, and the geometry sufficiently euclidean, to cater for the different degrees of variation of data. This allows to obtain closed form formulas for the different derivatives used for the optimization of the global cost.

Given a representation $\underline{Z}_u$, the classifier score is thus defined then as

$$f\left(\underline{Z}_u; \theta_c\right) = \theta_c \cdot \underline{Z}_u$$

which is a random variable defined over $\mathbb{R}$, with positive values associated with the positive class, i.e. the node $u$ belongs to category $c$ if $f\left(\underline{Z}_u; \theta_c\right) > 0$.

In our experiments, we used different losses for $\Delta_C$ – one directly inspired by the deterministic classification loss (Jacob, Denoyer, and Gallinari 2014), and the other exploiting the variance information.

We first considered the case where a class decision is simply the expectation of the classifier score together with a hinge loss, adapting the loss proposed in (Jacob, Denoyer, and Gallinari 2014). For a given node $u$ of type $t_u$ with an embedding $\underline{Z}_u$, the *Expected Value* classification loss (C-EV) is defined as:

$$\Delta_{C-EV}\left[f(\underline{Z}_u, \theta_c), y_c\right] \stackrel{\text{def}}{=} \max\left(0; 1 - y_c \mathbb{E}_{\underline{Z}_u}\left[f\left(\underline{Z}_u; \theta_c\right)\right]\right) \tag{4.4}$$

where $y_c$ is 1 if $x$ belongs to category $c$ and $-1$ otherwise, and $f\left(\underline{Z}_u; \theta_c\right)$ is a random variable which is greater than 0 if $u$ belongs to class $c$. The main issue with this formulation is that the information encoded in the variance is lost since

$$\mathbb{E}_{\underline{Z}_u}\left[f\left(\underline{Z}_u; \theta_c\right)\right] = \theta_c \cdot \underline{\mu}_u$$

Alternatively, the probabilistic formulation allows us to leverage the density-based representation through a probabilistic criterion, even in the case of linear classifiers. This "probabilistic" classification loss (C-PR) used the log-probability that $y_c f\left(\underline{Z}_u; \theta_c\right)$ take a positive value:

$$\Delta_{C-PR}\left[f\left(\underline{Z}_u; \theta_c\right), y_c\right] \stackrel{\text{def}}{=} -\sum_c \log p\left(y_c f\left(\underline{Z}_u; \theta_c\right) > 0\right) \tag{4.5}$$

In this case, the variance will be influenced by the two loss terms (see next section for the precise definition of the graph loss). If the two terms act in opposite directions, i.e. if there is a conflict between the graph regularization loss and the classification loss C-PR, one possible solution is to increase the variance of $\underline{Z}_u$. This is confirmed by our experimental results.

#### 4.2.2.2 Graph Embedding Loss

We make the hypothesis that two nodes connected in the graph should have similar representations, whatever their type is. Intuitively, this will force nodes of the same type which are close in the graph to be close in the representation space. The strength of this attraction between nodes of the same class is proportional to their closeness in the graph and to the weight of the path(s) linking them. We use the asymmetric loss proposed in (Vilnis and McCallum 2014):

$$L_G\left(\underline{Z}; \mathcal{G}\right) = \sum_{(u,v,r)\in\mathcal{E}} \boldsymbol{w}_r D_{KL}(\underline{Z}_u || \underline{Z}_v) \tag{4.6}$$

where $\boldsymbol{w}_r$ is a weight specific to the relationship $r$, and $D_{KL}(\underline{Z}_j || \underline{Z}_i)$ is the Kullback-Leibler divergence between the distributions of $\underline{Z}_u$ from $\underline{Z}_v$. Other similarity measures between distributions could be used as well (e.g. the inner product distance), the Kullback-Leibler divergence having the advantage of being asymmetric, which fits well the social network datasets used in the experiments since relations are not necessarily symmetric.

### 4.2.3 Prior Parameters and Learned Relation Specific Parameters

The graph regularization coefficients $\boldsymbol{w}_r$ are specific to a relationship $r$. They reflect the importance of relation $r$ for the inference task. For example, if inference consists in classifying an author research domain, then the authorship relation between authors and their published papers is probably more important than their affiliation relationship. For the model, this means that authors' representations should be close to their papers' representations, while the reverse is not true.

The $w_r$ are hyperparameters that could be learned by grid search and cross-validation. Since there might be several relation types for a heterogeneous graph, and hence a high number of

potential values to experiment with, this is not a relevant option here. We used instead the framework of continuous optimization of hyperparameters (Bengio 2000; Luketina et al. 2016).

This framework has been developed for learning regularization hyperparameters. Given a regularized loss such as (4.2), hyperparameters are learned along with the model parameters, by optimizing the unregularized loss on a distinct training set $\mathcal{D}_{\boldsymbol{w}}$.

Contrarily to grid search, which also selects regularization hyperparameters using an unregularized loss on a validation set, by testing different preset hyperparameters values, here parameters and hyperparameters are learned simultaneously and their dynamics are intertwined. There is no formal proof that such procedures converge to an optimal choice of the hyperparameters, but they offer an approximate solution which performs well in many cases (see Luketina et al. 2016 for a discussion).

There have been several instances of this general framework, and we derive below our own version for the specific problem handled here. Our inference problem is classification, and hence, following (Bengio 2000; Luketina et al. 2016), our loss function for hyperparameter training is the classification objective denoted $L_W$. This loss $L_W$ is defined on $\mathcal{D}_w$, a set of labeled nodes distinct from the labeled set $\mathcal{L}_c$ used in equation (4.2), i.e. we ensure that $\mathcal{D}_w \cap \mathcal{D}_c = \emptyset$.

The loss we optimize with respect to the weights $\boldsymbol{w} = \{w_r\}$ is similar to $L_C$ but defined on a different set of labeled nodes:

$$L_W(\boldsymbol{w}; \mathcal{D}_w) = \sum_{(x,y)\in\mathcal{D}} \sum_{c\in\mathcal{Y}^{t_u}} \Delta_C(f(\underline{Z}_x(\boldsymbol{w}); \theta_c(\boldsymbol{w})), y_c) \tag{4.7}$$

where the most important difference with $L_c$ is that this loss depends on the weights $\boldsymbol{w}$ and not on the representations $\underline{Z}$ or the classifier parameters $\theta$, and, more importantly, where $\underline{Z}$ and $\theta$ are both a function of $\boldsymbol{w}$, i.e. there are one of the solutions that minimize (4.2).

The proposed learning scheme uses an alternating optimization algorithm:

1. learning the parameters $\underline{Z}$ and $\theta$ by optimizing the loss $L_C + \lambda L_G$ of equation (4.2), with $\boldsymbol{w}$ fixed;

2. learning the hyperparameters $\boldsymbol{w}$ by optimizing the loss of equation (4.7).

At each iteration of this process, the value of the parameters are dependent on the current $\boldsymbol{w}$: more precisely, $\theta(\boldsymbol{w})$ and $\underline{Z}(\boldsymbol{w})$ are the parameters that minimize the loss of equation (4.2). This dependency is emphasized in equation (4.7) by the notations $z_i(\boldsymbol{w})$ and $\theta(\boldsymbol{w})$.

Minimizing (4.7) is performed by gradient descent. In (L. D. Santos et al. 2018), we derive a closed form for $\frac{\partial L_W}{\partial \boldsymbol{w}}$, that relies on a series of assumptions about how the classifier and embeddings change with respect to a change of the hyperparameter $\boldsymbol{w}$. More precisely, we suppose that the parameters of the classifiers $\theta$ remain mostly unchanged (for a small change in $\boldsymbol{w}$) compared to the representations of the nodes $\underline{Z}$ and we further suppose that each node position can be approximated as a linear combination of its neighbors previous representations. Note that the latter is a reasonable assumptions, since for a node in $\mathcal{D}_{\boldsymbol{w}}$, the only expressed loss in equation (4.2) is the graph regularization.

This allows us to find closed formulas for the representations of nodes. For instance, when using the $\Delta_{EV}$ variant, we need to express $\mu_u$ as an explicit function of $w$ by zeroing $\nabla_z L$ on $\mathcal{D}_w$. We obtain closed-form equations for the parameters.

### 4.2.4 Results

To investigate whether using probabilistic embeddings for graph node classification is beneficial, we performed extensive experiments on three datasets respectively extracted from DBLP (author/paper network), FlickR (image, tags and authors), and LastFM (artists, users, albums and songs). For all but the first dataset (DBLP), each node can have multiple labels. Results

| Train size | Model | DBLP | | FlickR | | LastFM | |
|---|---|---|---|---|---|---|---|
| | | **Micro** | Macro | **Micro** | Macro | **Micro** | Macro |
| 10% | LINE | 19.5 | 23.0 | 20.7 | 23.2 | 20.4 | 15.9 |
| | HLP | 24.1 | 27.2 | 26.3 | 27.8 | 38.4 | 30.0 |
| | Graffiti | 30.9 | 38.1 | 24.5 | 27.0 | 40.0 | 31.4 |
| | LaHNet | **32.1** | **40.0** | 29.3 | 29.1 | 36.3 | 27.2 |
| | HCGE($\Delta_{EV}$,S) | 30.9 | 38.5 | **32.7** | **32.6** | **44.0** | **34.1** |
| | HCGE($\Delta_{EV}$,D) | 30.4 | 37.4 | 32.6 | **32.6** | 43.6 | 34.0 |
| | HCGE($\Delta_{Pr}$,S) | 27.9 | 34.3 | 29.7 | 29.2 | 27.8 | 20.7 |
| | HCGE($\Delta_{Pr}$,D) | 28.3 | 34.3 | 31.9 | 32.2 | 29.4 | 21.9 |
| 30% | LINE | 21.9 | 24.8 | 21.5 | 24.2 | 20.5 | 17.0 |
| | HLP | 36.0 | 41.9 | 47.7 | 43.7 | 49.7 | 40.0 |
| | Graffiti | 38.5 | 46.6 | 47.0 | 43.7 | 50.3 | 40.4 |
| | LaHNet | 41.2 | 52.9 | 48.4 | 43.6 | 53.3 | 40.3 |
| | HCGE($\Delta_{EV}$,S) | **42.3** | 52.6 | 50.0 | 45.6 | **57.3** | 45.0 |
| | HCGE($\Delta_{EV}$,D) | 41.2 | 50.8 | **50.1** | 45.7 | 57.0 | **45.3** |
| | HCGE($\Delta_{Pr}$,S) | 41.3 | 52.1 | 49.0 | 44.4 | 50.4 | 37.7 |
| | HCGE($\Delta_{Pr}$,D) | **42.3** | **54.1** | 50.0 | **45.8** | 50.8 | 38.5 |
| 50% | LINE | 22.3 | 25.0 | 21.8 | 24.6 | 20.5 | 17.0 |
| | HLP | 39.4 | 46.5 | 54.1 | 48.6 | 52.1 | 42.3 |
| | Graffiti | 41.2 | 49.4 | 54.0 | 48.8 | 53.5 | 43.2 |
| | LaHNet | 44.4 | 56.8 | 54.0 | 47.9 | 56.7 | 43.2 |
| | HCGE($\Delta_{EV}$,S) | 44.6 | 55.2 | 55.8 | 50.0 | **60.4** | **48.7** |
| | HCGE($\Delta_{EV}$,D) | 43.9 | 53.7 | 55.8 | 50.0 | 60.3 | 48.6 |
| | HCGE($\Delta_{Pr}$,S) | 45.5 | 57.1 | 54.8 | 49.0 | 58.5 | 45.0 |
| | HCGE($\Delta_{Pr}$,D) | **45.7** | **57.7** | **55.9** | **50.3** | 58.9 | 47.2 |

Table 4.1 – P@1 results of the model HCGE and baselines on DBLP, FlickR and LastFM

are reported in Table 4.1 – details about the datasets and the full results can be found in L. D. Santos et al. 2018.

We compared the results of four variants of our Gaussian embedding model (HCGE), and compared to various state-of-the-art baselines:

- LINE (Tang et al. 2015), which is representative of unsupervised learning of graph embeddings suitable for various tasks such as classification. We performed a logistic regression with the learned representations as inputs.
- Graffiti (Angelova, Kasneci, and Weikum 2012b) which is representative of transductive graph algorithms developed for semi-supervised learning.
- HLP (X. Zhu and Ghahramani 2002) which is representative of transductive graph algorithms developed for semi-supervised learning. As HLP is designed for homogeneous graphs, we perform as many random walks as the number of node types, considering each time that all the nodes are of a same given type.
- LaHNet (L. D. Santos et al. 2018) which corresponds to the classification $\Delta_{C-EV}$ with Dirac distributions and where the graph regularization is based on the Euclidean distance rather than the Kullback-Leibler divergence.

The main conclusions that we can draw from the various experiments are that:

- Supervised models (HLP, Graffiti, LaHNet and HCGE) using the class information outperform unsupervised representation learning, which matches the results reported in (Jacob,

Denoyer, and Gallinari 2014; L. D. Santos et al. 2018);

- Modeling the heterogeneity of the graph brings noteworthy improvements: On all datasets, the performances of HLP are below the performances of Graffiti, LaHNet and HCGE;
- Learning node representations improve the results: comparing the heterogeneous models, both LaHNet and HCGE outperform Graffiti on all datasets. We can note that the more complex the dataset, the higher the gap compared to the baselines;
- Finally, introducing uncertainty in representations clearly improves results, as acknowledge by the comparison with LaHNet. Let us also point out that, according to our initial intuition, the effect of using uncertainty has more impact when the amount of training data is lower: the difference between LaHNet and HCGE decreases in general when more training data is available.

Among the different variants of our model HCGE (classification loss and form of the covariance matrix), there were no clear global trend.

We noticed that globally, $\Delta_{Pr}$ seems to be disadvantaged by a low number of training examples, when $\Delta_{EV}$ seems to be more stable in comparison to other baselines. However, the more training data, the closer the $\Delta_{Pr}$ variant is to $\Delta_{EV}$. We believe that this is due to the fact that the covariance matrix is only optimized in the graph regularization term in the case of $\Delta_{EV}$.

With respect to the use of a spherical and a diagonal covariance matrix. For the $\Delta_{EV}$ variant, it looks like moving from a spherical covariance matrix to a diagonal one brings no improvement. It even decreases the performance on DBLP. Concerning the $\Delta_{Pr}$ variant, for which the covariance matrix plays a role in the classification cost, conclusions are reversed and using diagonal covariance matrices improves the results.

Overall, probabilistic embeddings did improve the results over all the baselines we compared it to. It seems that there is no real difference between the different variants – or that they are dependent on the dataset or/and on the amount of training data. Further work would be needed in order to clear this out.

## 4.3 Recommendation

We now turn to the use of probabilistic embeddings for recommender systems. These systems help users find items they might like in large repositories, thus alleviating the information overload problem. Methods for recommending text items can be broadly classified into collaborative filtering (CF), content-based, and hybrid methods (Ricci et al. 2011). Among recommender systems, collaborative filtering systems are the most successful and widely used because they leverage past ratings from users and items, while content-based approaches typically build users (or items) profiles from items (or users) predefined representations. Among collaborative recommender systems we are interested in those that represent users and items as vectors in a common latent recommendation space. Matrix factorization derivatives are a typical example of such models. The main assumption is that the inner product between a user and an item representation is correlated with the rating the user would give to the item. The main interest of these models is their potential for integrating different sources of information (F. Zhang et al. 2016) as well as their good performance, both in efficiency and effectiveness (Ricci et al. 2011). The most successful approaches are based on learning-to-rank approaches such as (Rendle et al. 2009; Weimer et al. 2007).

Despite their successes, one limitation of those models is their lack of handling of uncertainty. Even when they are based on a probabilistic model, e.g. (Rendle et al. 2009; Salakhutdinov and Mnih 2007), they only suppose a Gaussian prior over the user and item embeddings, but still

learn a deterministic representation. The prior is thus only used as a regularization term on user and item representations.

Uncertain representations can have various causes, either related to the lack of information (new users or items, or users that did not rate any movie of a given genre), or to contradictions between user or item ratings. To illustrate the latter, let us take a user that likes only a part of Kung Fu movies, but with no clear pattern, i.e. no sub-genre. With usual approaches, such a user will have a component set to zero for the direction of the space corresponding to Kung Fu. With our proposed approach, we would still have a zero mean, but with a high variance. Our hypothesis is that, because of these two factors, namely cold start and conflicting information, training will result in learned representations with different confidences, and that this uncertainty is important for recommending.

Moreover, using a density rather than a fixed point has an important potential for developing new approaches to recommendation. First, instead of computing a score for each item, the model can compute a probability distribution, as well as the covariance between two different item scores. This can be interesting when trying to diversify result lists, since this information can be leveraged e.g. by Portfolio approaches (J. Wang and J. Zhu 2009): The model could propose diversified lists with different degrees of risk. Secondly, such models are interesting because they can serve as a basis for integrating different sources of external information, and thus serve as a better bridge between content-based approaches and collaborative filtering ones. Thirdly, time could be modeled by supposing that the variance of the representation increases with time if no new information (i.e. user ratings) are provided.

In the following, we first describe our Gaussian Embeddings Ranking Model, and then show experimentally that it performs very well compared to state-of-the-art approaches.

### 4.3.1 The Gaussian Embeddings Ranking Model

Our model is learned through a pairwise learning-to-rank criteria that was first proposed by (Rendle et al. 2009) for collaborative filtering (Bayesian Personalized Ranking, BPR). Formally, they optimize the *maximum a posteriori* of the training dataset $\mathcal{D}$, i.e. $p(\Theta|\mathcal{D}) \propto p(\mathcal{D}|\Theta)\, p(\Theta)$ where $\mathcal{D}$ represents the set of all ordered triples $(u, i, j)$ bearing the semantic that the user $u \in \mathcal{U}$ prefers item $i \in \mathcal{I}$ to item $j \in \mathcal{I}$, and $\Theta$, the model parameters. The factor $p(\Theta)$ corresponds to the prior on the item and user representations (a Gaussian prior in Rendle et al. 2009). Then, using the standard hypothesis of the independence of samples given the model parameters, we have

$$p(\mathcal{D}|\Theta) = \prod_{(u,i,j)\in\mathcal{D}} p(i >_u j|\Theta) \tag{4.8}$$

In Rendle et al. 2009, the probability that user $u$ prefers item $i$ to item $j$, noted by $i >_u j$, is given by the sigmoid function of the difference of the inner products, that is:

$$p(i >_u j|\Theta) = \sigma\left(\underline{x}_i \cdot \underline{x}_u + \underline{b}_i - \underline{x}_j \cdot \underline{x}_u - \underline{b}_j\right) \tag{4.9}$$

where $\underline{b}_i$ (resp. $\underline{b}_j$) is the bias for item $i$ (resp. $j$), i.e. it can be interpreted as the average user rating (or popularity) for this specific item.

In this work, while we start with Eq. (4.8), the different variables $x_\bullet$ ($\bullet$ is either a user $u$ or an item $j$) are random variables, denoted $\underline{\mathbf{X}}_\bullet$, and not elements of a vector space. We can hence estimate directly probability $p(i >_u j|\Theta)$ of the inner product $\underline{\mathbf{X}}_u \cdot \underline{\mathbf{X}}_i$ being higher than $\underline{\mathbf{X}}_u \cdot \underline{\mathbf{X}}_j$.

We start by supposing that user and item representations follow a normal distribution. That is, for any user $u$ and item $i$,

$$\underline{\mathbf{X}}_i \sim \mathcal{N}\left(\underline{\mu}_i, \underline{\Sigma}_i\right) \text{ and } \underline{\mathbf{X}}_u \sim \mathcal{N}\left(\underline{\mu}_u, \underline{\Sigma}_u\right)$$

We suppose that $\underline{\Sigma}_\bullet = \text{diag}\left(\underline{\sigma}_{\bullet 1}, \ldots, \underline{\sigma}_{\bullet N}\right)$ is a diagonal covariance matrix to limit the complexity of the model ($N$ is the dimension of the latent space): in practice, we have $2N$ parameters for each user, and $2N+1$ for each item ($+1$ because of the bias). The variance is associated with each element of the canonical basis of the vector space (diagonal covariance matrix). In practice, we hypothesize that this helps the model to learn meaningful dimensions, since it forces the use of the canonical latent space basis.

Since $\underline{\mathbf{X}}_i$ and $\underline{\mathbf{X}}_u$ are random variables, their inner product is also a random variable, and we do not need to use the sigmoid function of Eq. (4.9) as for BPR to compute the probability that user $u$ prefers $i$ to $j$. We can instead directly use the random variables defined above, which leads to:

$$\begin{aligned} p\left(i >_u j | \Theta\right) &= p\left(\underline{\mathbf{X}}_i \cdot \underline{\mathbf{X}}_u + \underline{b}_i > \underline{\mathbf{X}}_j \cdot \underline{\mathbf{X}}_u + \underline{b}_j | \Theta\right) \\ &= p\Big(\underbrace{\underline{\mathbf{X}}_u \cdot \left(\underline{\mathbf{X}}_j - \underline{\mathbf{X}}_i\right)}_{\mathbf{Z}_{uij}} < \underline{b}_i - \underline{b}_j \Big| \Theta\Big) \end{aligned}$$

where $\underline{b}_i$ and $\underline{b}_j$ are the item biases[2]. To compute the above equation, we use the following two simplifications:

1. Item representations are independent given the model parameters. This implies that the difference $\underline{\mathbf{X}}_j - \underline{\mathbf{X}}_i$ follows a normal distribution $\mathcal{N}\left(\underline{\mu}_j - \underline{\mu}_i, \underline{\Sigma}_i + \underline{\Sigma}_j\right)$
2. We approximate the inner product distribution by a Gaussian using moment matching (mean and variance), using the results from Brown and Rutemiller (1977) who defined the moments of the inner product of two Gaussian random variables; with our notations, the two first moments of $\mathbf{Z}_{uij}$ are defined as:

$$\mathbb{E}\left[\mathbf{Z}_{uij}\right] = \underline{\mu}_u^\top (\underline{\mu}_j - \underline{\mu}_i) \tag{4.10}$$

$$\begin{aligned} \text{Var}\left[\mathbf{Z}_{uij}\right] \underset{=}{} & 2\underline{\mu}_u^\top \left(\underline{\Sigma}_i + \underline{\Sigma}_j\right) \underline{\mu}_u + \left(\underline{\mu}_j - \underline{\mu}_i\right)^\top \Sigma_u (\underline{\mu}_j - \underline{\mu}_i) \\ & + \text{tr}\left(\underline{\Sigma}_u \left(\underline{\Sigma}_i + \underline{\Sigma}_j\right)\right) \end{aligned} \tag{4.11}$$

The above assumptions allow us to write:

$$p\left(i >_u j | \Theta\right) \approx \int_{-\infty}^{b_i - b_j} \mathcal{N}\left(x; \mathbb{E}\left[\mathbf{Z}_{uij}\right], \text{Var}\left[\mathbf{Z}_{uij}\right]\right) \, \mathrm{d}x \tag{4.12}$$

which is the Normal cumulative distribution function. This function can be easily differentiated with respect to the $\underline{\mu}_\bullet$ and $\underline{\Sigma}_\bullet$ ($\bullet$ is a user or an item).

From Equations (4.12) and (4.11) and ignoring the biases, we can see that the difference between the inner products $\underline{\mu}_u \cdot \underline{\mu}_i$ and $\underline{\mu}_u \cdot \underline{\mu}_j$ is controlled by the variance of the user and the item (especially for the components of the means that have a high magnitude) – the larger, the closer to 0.5 the probability. This is the main difference with other matrix factorization-based approaches.

Finally, we have to define the prior distribution over the parameters $\Theta$, i.e. the means and variances. As we consider only a diagonal covariance matrix, we can consider an independent prior for each mean and variance, i.e. for any $\mu_{\bullet k}$ and $\Sigma_{\bullet kk}$. In this case, using a normal-gamma prior is a natural choice since it is the conjugate distribution of a normal. More specifically, we suppose that

$$\left(\underline{\mu}_{\bullet k}, \underline{\Sigma}_{\bullet kk}^{-1}\right) \sim \text{NormalGamma}\left(\nu, \lambda, \alpha, \beta\right) \tag{4.13}$$

[2]In this work, we did not consider them to be random variables, but they could be

with $\nu = 0$, $\lambda = 1$, $\alpha = 2$ and $\beta = 2$ – these parameters do not change much the solution, and were chosen so that user and item representations are not too much constrained. In the absence of data, this would set, for each component, the mean to 0 and the variance to 1, which corresponds to the mode of the normal-gamma distribution.

Finally, using Eq. (4.13) and (4.8) to express the MAP criterion gives:

$$\begin{aligned}\mathcal{L} &= \log p\left(\Theta\right) + \sum_{(u,i,j)\in D_S} \log p\left(i >_u j|\Theta\right) \\ &\stackrel{+}{=} \sum_{\bullet,i} \left[\left(\frac{1}{2} - \alpha\right) \log\left(\underline{\Sigma}_{\bullet ii}\right) - \frac{2\beta + \lambda \underline{\mu}_{\bullet i}}{2\underline{\Sigma}_{\bullet i}}\right] - c\sum_i b_i^2 \\ &\quad + \sum_{(u,i,j)\in\mathcal{D}} \log p\left(i >_u j|\Theta\right)\end{aligned}$$

where $\bullet$ is a user or an item, $\stackrel{+}{=}$ means equal up to a constant, and $p\left(i >_u j|\Theta\right)$ is given by equations (4.12), (4.10) and (4.11), and where we suppose a normal prior for the biases (to avoid overfitting). The parameters $\Theta = \left\{\underline{\Sigma}_\bullet, \underline{\mu}_\bullet\right\}_{\bullet\in\mathcal{U}\cup\mathcal{I}}$ are optimized with stochastic gradient update, picking a random triple at each step, and updating the correspond means and variances (for user $u$, and items $i$ and $j$).

### 4.3.2 Ordering items

We could have used pairwise comparison from our model since the relation $i >_u j \iff p\left(i >_u j|\Theta\right) > 0.5$ defines a total order over items, but it does not make any difference in the ranking if $p\left(i >_u j|\Theta\right)$ equals 0.51 or 0.99, and this difference could be due to the sole difference between variances. To avoid this problem, we order items by their probability of having a positive score, i.e. $s_{ui} = p(\underline{\mathbf{X}}_u \cdot \underline{\mathbf{X}}_j + b_i > 0)$[3]. This ordering preserves the variance information in contrast to the pairwise comparison – a score with a high variance will be associated with a score $s_{ui}$ close to 0.5. Item scheduling that depends on the covariance between items, using « portfolio techniques » where the variance and covariance associated with earnings (here the relevance of a recommendation) are taken into account (J. Wang and J. Zhu 2009) .

### 4.3.3 Results

Results are reported in Table 4.2. There are roughly three groups of models: (1) Soft Margin (SM) that performed the worst; (2) BPRMF and most popular (MP); (3) GER and CofiRank (CR). Most popular (MP) performs well because of the chosen experimental evaluation where only items *seen* by the user are evaluated and ranked.

Our model (GER) outperforms the others on the MovieLens and Yelp dataset. It is usually above CofiRank (nDCG difference between 0.01 and 0.08), with the exception of the Yahoo dataset for train sizes 20 and 50 (nDCG difference between 0.001 and 0.02). Overall, GER is performing very well on three datasets with different characteristics in terms of rating behavior and number of ratings.

We now turn to the analysis of results, based on the study of the learned representations for the Yahoo dataset. We used the hyper-parameters that were selected on the validation set, and looked at the set of mean-variance couples, i.e. the $(\mu_{\bullet k}, \Sigma_{\bullet k})$ for $k \in \{1, \ldots, N\}$ and $\bullet \in \mathcal{U} \cup \mathcal{I}$.

In Figure 4.5, we plotted the different couples $(\mu_{\bullet k}, \Sigma_{\bullet k})$, as well as the histogram of mean and variance. We can see that the model exploits the different ranges of mean (-0.3 to 0.3) and, more importantly, variance (0.4 to 1.2) around the priors. This was satisfying since it was not obvious that the variance component would be used by the model, i.e. that it would deviate from its prior.

[3] We use the same moment matching approximation to compute the inner product distribution.

| Train Size → | 10 | | | 20 | | | 50 | | |
|---|---|---|---|---|---|---|---|---|---|
| Model ↓ | N@1 | N@5 | N@10 | N@1 | N@5 | N@10 | N@1 | N@5 | N@10 |
| **Yahoo!** | | | | | | | | | |
| MP | 53.0 | 59.1 | 67.3 | 52.5 | 58.3 | 66.4 | 53.6 | 57.8 | 64.0 |
| BPRMF | 52.8 | 59.0 | 67.2 | 52.2 | 58.3 | 66.4 | 52.2 | 57.7 | 63.5 |
| SM | 50.9 | 56.7 | 65.4 | 49.7 | 55.6 | 64.2 | 49.9 | 54.1 | 60.3 |
| CR | **53.5** | **60.3** | 68.2 | **57.8** | **61.7** | **68.9** | **56.0** | **60.0** | **65.6** |
| GER | **53.5** | **60.3** | **68.3** | 53.8 | 60.7 | 68.2 | 54.3 | 59.6 | 65.3 |
| **MovieLens** | | | | | | | | | |
| MP | 66.0 | 64.7 | 65.8 | 68.4 | 65.3 | 66.3 | 69.1 | 67.4 | 67.5 |
| BPRMF | 66.1 | 64.6 | 65.7 | 66.3 | 64.3 | 65.8 | 66.9 | 65.0 | 66.2 |
| SM | 55.9 | 57.5 | 60.3 | 58.3 | 59.6 | 61.6 | 58.6 | 60.4 | 62.5 |
| CR | 69.0 | 67.3 | 68.6 | 69.7 | 68.5 | 69.5 | 71.4 | 69.4 | 70.6 |
| GER | **70.3** | **67.7** | **70.0** | **72.0** | **69.5** | **71.1** | **72.5** | **71.3** | **71.5** |
| **Yelp** | | | | | | | | | |
| MP | 40.7 | 41.5 | 46.9 | 39.5 | 39.9 | 44.7 | 37.4 | 37.6 | 41.4 |
| BPRMF | 40.8 | 41.3 | 46.8 | 39.6 | 39.8 | 44.6 | 37.3 | 37.2 | 40.9 |
| SM | 37.3 | 38.3 | 44.4 | 35.8 | 36.9 | 41.9 | 33.4 | 34.1 | 38.0 |
| CR | 47.1 | 46.9 | 51.1 | 46.5 | 46.6 | 50.4 | 46.2 | 45.8 | 48.6 |
| GER | **55.2** | **52.2** | **56.2** | **57.4** | **53.5** | **56.4** | **58.2** | **53.7** | **55.3** |

Table 4.2 – Collaborative Ranking results. nDCG values (×100) at different truncation levels are shown within the main columns, which are split based on the amount of training ratings.

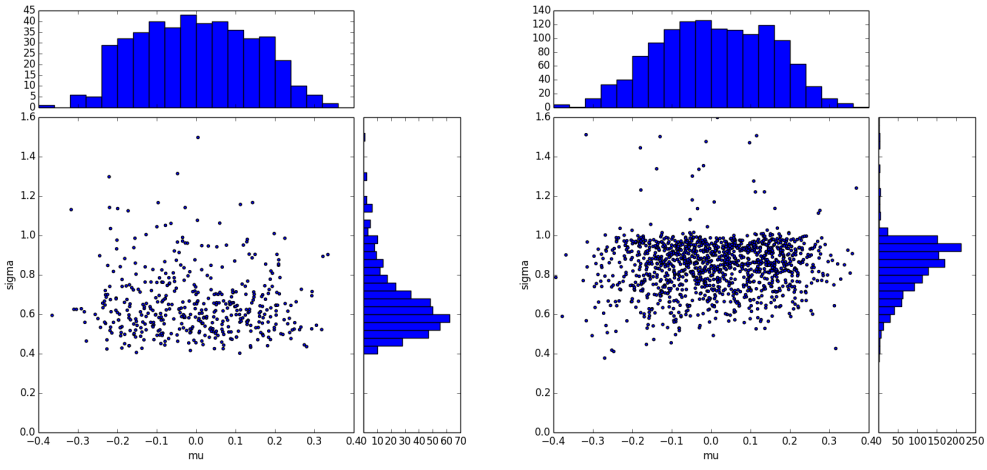



Figure 4.5 – User (left) and item (right) learned first component representation plots for T50 and a representation dimension of 50.

#### 4.3.3.1 Conclusion

In this work, we have shown that using probabilistic embeddings again allowed to outperform previous approaches.

This work can be extended in various directions:

- By using other learning criterion that make use of the variance information
- By incorporating meta-information about users and items, leveraging the work conducted in graph node representations

With respect to the former, a possibility that we explored is to use a learning criterion, SoftRank (Taylor et al. 2008), which takes into account the covariance between the predicted scores, as well as the set of scores distributions of a set of items (rather than two to two). This was published in (Titeux, Piwowarski, and Gallinari 2018a), but the preliminary results did not show a strong improvement with respect to the MAP cost we used in this work. The most likely reason is that in the context of recommendation, taking into account popularity is already a strong baseline; the extra performance brought by list-wise learning-to-rank models does not outweigh the extra computational cost and instability arises in optimizing such criteria. However, more experiments are needed to check whether this holds true in different experimental settings.

With respect to the latter, this corresponds to a problem that still does not have a satisfactory solution is the integration of meta-information on items or users, in order to solve cold start problems when no judgment is known for a new item or a new user. Content-based methods rely on the definition of distance based on meta-information, which is neither efficient nor easily transferable to new data sources. Works that combine representational learning and taking into account meta-information started to appear (Vasile, Smirnova, and Conneau 2016), but do not take into account the problem of the uncertainty related to the information provided.

Using the formalism of graph representation exposed in section 4.2, we can express meta-information as triplets $(s, r, t)$ where s is the source entity, $r$ the relation, and $t$ the target entity; for example, a film that is « comedy » will be expressed in the form of a triplet (film, « genre » , « comedy » ). In the case of a categorical type entity (eg. « genre of the film » ), it is possible to use also a representation in the form of a density, and, following the works (Dos Santos, Piwowarski, and Gallinari 2016; S. He et al. 2015) , to express the relation between $s$ and $t$ as a «distance» such as Kullback-Leibler:

$$\Delta(s, r, t) = KL(\mathbf{X}_t || \mathbf{X}_s)$$

where $\mathbf{X}_s$ and $\mathbf{X}_t$ correspond to probability distributions. To repeat the previous example, when $t$ is a genus, $\mathbf{X}_t$ represents in some way the places in the space where a movie can be found whose genus is $t$ . For vague genera, the variance of $\mathbf{X}_t$ will be large, and vice versa for a specific genre.

More complex relationships, for example linking a movie to a text describing it, can also be considered by learning to project a text in the distribution space. One possibility we will explore is to learn the parameters of a Gaussian, i.e. for a text $t$, representing it as

$$\underline{\mathbf{X}}_t \sim \mathcal{N}(\mu_\theta(\underline{t}), \Sigma_\theta(\underline{t}))$$

where $\theta$ are the parameters of the projection models: the goal of learning is to predict the center and covariance matrix of each text $t$ . This model has the possibility to learn if a text makes it possible to specify or not (via the variance) the locus of an entity which is described by this text.

## 4.4 Gaussian Embeddings: Conclusion

In this chapter, we presented two of our works leveraging Gaussian embeddings. In both cases, we obtained better results than state-of-the art works, with the same number of parameters (even when taking into account the cost of adding a variance information). A third work was conducted on time-series models (Ziat et al. 2016), by trying to model the evolution of a distribution in a state space, but the quantitative results were not improving much compared to deterministic approaches.

Even with the success of recent neural approaches, this problematic has nevertheless continued to be explored in various works (Bojchevski and Günnemann 2017b; Y. Chen, Pu, et al. 2019; Pei et al. 2020), and is still especially useful when modeling nodes in a graph. I believe that given the simplicity of the optimization scheme (as opposed to a Bayesian modeling) and its flexibility (e.g. using the KL distance), this type of probabilistic embeddings can be leveraged in many cases successfully.

**Outcomes**

- I co-supervised (with P. Gallinari) the thesis of Ludovic Dos Santos (defended in December, 2017).
- Graph node classification
  - L. Dos Santos, B. Piwowarski, and P. Gallinari (2016). "Multilabel Classification on Heterogeneous Graphs with Gaussian Embeddings." English. In: *ECML*. Springer International Publishing, pp. 606–622. DOI: 10.1007/978-3-319-46227-1_38
  - L. D. Santos et al. (July 2018). "Representation Learning for Classification in Heterogeneous Graphs with Application to Social Networks." In: *ACM Transactions on Knowledge Discovery from Data* 12.5, 62:1–62:33. DOI: 10/gdvmmq
- Gaussian Embeddings
  - A. Ziat et al. (2016). "Modeling Relational Time Series Using Gaussian Embeddings." In: *NIPS Time Series Workshop*
  - L. Dos Santos, B. Piwowarski, and P. Gallinari (2017). "Gaussian Embeddings for Collaborative Filtering." In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '17. New York, NY, USA: ACM, pp. 1065–1068. DOI: 10.1145/3077136.3080722
  - H. Titeux, B. Piwowarski, and P. Gallinari (Mar. 2018b). "Représentations Gaussiennes pour le Filtrage Collaboratif." In:

# Part II

# Language Grounding

# Chapter 5

# Language Grounding

## 5.1 Introduction

To a large extent, Natural Language Processing and Information Access tasks are mostly concerned with textual information, and do not consider other sources of information. Said otherwise, extracting information from a document only relies on textual clues and statistical patterns. However, a wealth of information is available in images, videos or audio signals, and especially a type of information that is seldom made explicit in texts – namely, common sense knowledge.

A few years ago, in NLP, some works have underlined the limitations of using purely textual data. Bruni et al. (2012b) showed for instance that purely textual models learn representations that do not contain the typical color of common concrete objects. Analyzing a purely-textual distributed semantic models (DSMs), they found for example that the closest color (in terms of cosine similarity) in the semantic space of the word *pig* is *brown*, and that DSMs are confused by non-literal uses of color words (e.g. *white wine*). Following this work, Gordon and Durme (2013) noted discrepancies between real word frequencies and concept frequencies in texts.

These empirical observations are supported by works in grounded cognition that reject the traditional view of cognition as computation on amodal symbols, independent of the brain's modal systems for perception, action, and introspection. This traditional view is linked to works viewing language as a purely symbolic system based on logical grammar (Burgess and Lund 1997; Chomsky 1980). A body of knowledge (for references, see Barsalou 2008) has shown that this view is far from reality, and have advocated for *grounded cognition* theories, i.e. theories that supposes that modal (motor or perceptual) simulations, bodily states and situated actions must be an integral part of any theory of cognition. In particular, Barsalou (2008) discusses a number of experiments that have shown that comprehension is linked to actual modal simulations, as for instance:

- A spatial representation of a scene is constructed when reading it (interfering with this construction slows down comprehension);
- Abstract concepts are grounded metaphorically in embodied and situated knowledge: specifically, these researchers argued that people possess extensive knowledge about their bodies (e.g., eating) and situations (e.g., verticality), and that abstract concepts draw on this knowledge metaphorically.

Whether this embodiment is necessary for natural language processing is an open question, but it is clear from these works, and from our daily life experience, that text and image (among other modalities) are complementary. Some works have already been trying to squeeze out of images some real world knowledge (Bagherinezhad et al. 2016; Vedantam et al. 2015; Yatskar, Ordonez, and Farhadi 2016).

An example is the work of Bagherinezhad et al. (2016), who seek to find what is the typical size of an animal or object. Their model is based on the assumption that each object size follows

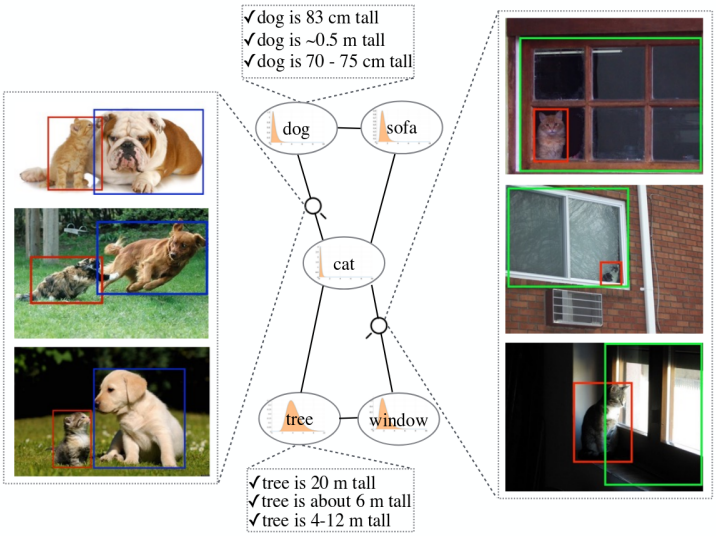



Figure 5.1 – **Exploiting textual and visual resources to reason about sizes of common objects**. In this example, language gives direct information about sizes of objects, while images gives information about the relative sizes of objects. Combining both direct and relative size knowledge allows to infer size of new objects. Illustration taken from Bagherinezhad et al. (2016)

a log-normal distribution, and that (1) language gives access to *absolute size of some objects*, while (2) images provide *relative sizes of objects* when two objects are contained within the same image, by comparing the depth-adjusted size of the two bounding boxed estimated with webly-supervised detectors (Divvala, Farhadi, and Guestrin 2014). In their work, they show how one can infer the size of each object using a probabilistic modeling of reality (see Figure 5.1).

Finally, at a more application level, there are several (multi-modal) tasks related to textual information (Beinborn, Botschen, and Gurevych 2018) where information from other modalities can be useful (or is mandatory):

**Cross-modal transfer** by learning how to transform a representation from one modality to another (e.g. cross-modal retrieval, searching an image from a textual query). Sometimes, the goal is to obtain a compressed and structured intermediate representation of the input to generate a useful interpretation in the target modality (e.g. image captioning).

**Joint multimodal processing** are tasks which explicitly require the combination of knowledge from different modalities. Examples of such tasks emotion recognition from video and sound, disambiguation with image and text, etc.

**General tasks** Finally, and more importantly, general natural language processing related tasks rely on word or sentence representations. These representations, when *grounded*, can potentially bring common-sense information into the natural language processing models.

In this chapter, we focus mainly on grounding. We investigate whether visual information can be integrated into text representation, and present two works aiming at grounding *words* in Section 5.3.1 and *sentences* in Section 5.3.2. We also look at the question of grounding with temporal information in Section 5.3.3.

## 5.2 Visual and textual representation spaces

Before presenting our works in grounding, we first study the difference between the visual modality, which is the main source of information we consider in this chapter, and the textual repre-

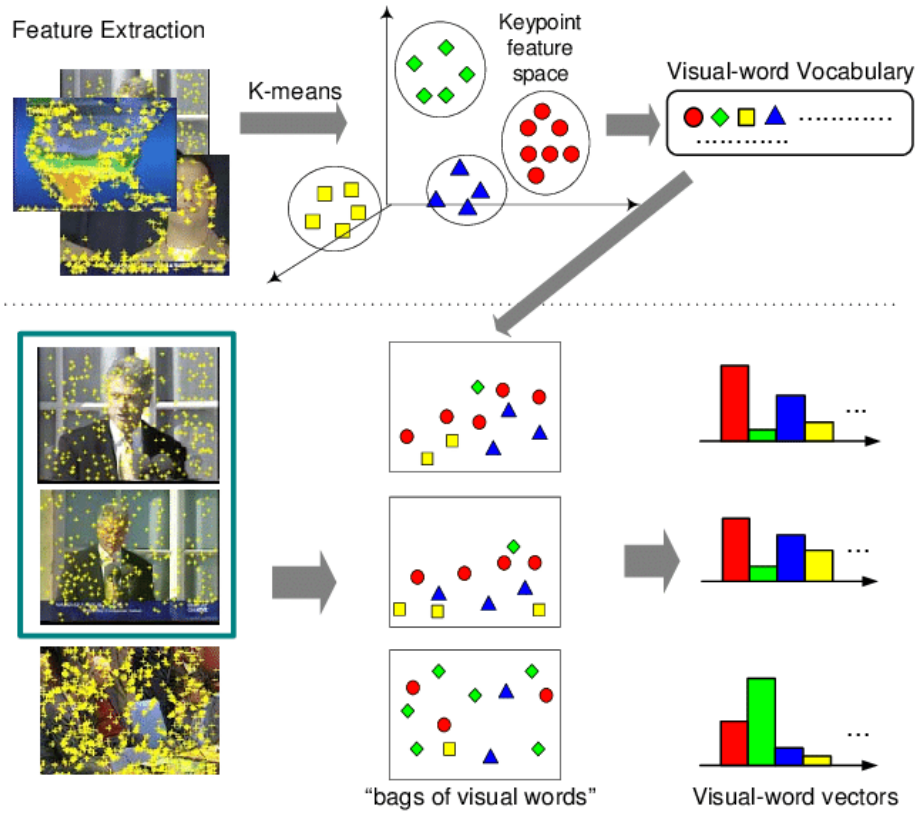



Figure 5.2 – **Illustration of the BoVW approach.** Image taken from (Jiang et al. 2010)

sentation space.

### 5.2.1 Representing images

In this section, we study the evolution of the representation of images. We first present the pre-neural canonical representation of images, before focusing on convolution neural networks.

#### 5.2.1.1 Bag of Visual Features

In traditional computer vision pipelines, no learning is involved to compute image representations. The main idea is to compute descriptors for interesting parts of the images before aggregating them. For example, these local descriptors could be obtained by convolving kernels on images, with kernels coming from filter banks. Based on convolving Gaussian kernels on images at different scales, the SIFT (Lowe 1999) descriptors provide features invariant to rotations and image scaling. The obtained features have been shown to be robust across shifts in intensity and illumination, and, to some extent, to changes in 3D viewpoint and distortions.

To obtain a global representation, local image descriptors can be aggregated as *bags of visual words* (G. Qiu 2002), through a four-step process illustrated in Figure 5.2. First, the feature detection algorithm is used to detect key points, i.e. the important regions in the image. Second, the feature description part associates with an image region a local representation, usually with SIFT (Lowe 1999), HOG (Freeman and M. Roth 1995) or SURF (Bay et al. 2008). Third, the codebook generation clusters local feature descriptors, giving rise to a codebook that associates with each feature a cluster. Finally, in the fourth step, the image is represented as a bag of words (or more precisely, of codewords discovered during the previous step).

Once an image is represented, any discriminative model such as a naive Bayes model or a SVM can be used to to tackle the task of interest. This traditional pipeline has however several limitations. As for the BoW model that represents documents using one hot word representations

ignoring the sequential information, the BoVW model ignores the (spatial) structure of image patches. Second, and this is more image specific, designing good feature descriptors is task-dependent, and, generally expert knowledge is incorporated in the used filter banks (Y. M. Lu and Do 2007). This creates models that poorly generalize to new domains and that are costly to design as intensive human intervention is needed. Finally, like any human-crafted representation, the expressive power might be limited and might not capture the important variations in the input source, here the images. These limitations were solved to a large extent with the advent of deep neural networks that we present next.

#### 5.2.1.2 Representation Learning: The ConvNet era

As mentioned in the introduction, the deep learning revolution comes from the ability of neural networks to learn rich and expressive data representation. This emerging paradigm has shown to be much more efficient and effective than using handcrafted features, and this has begun with image classification tasks.

Inspired by biological processes, and how certain neurons in the visual cortex were found to fire when certain images are shown, Fukushima and Miyake (1982) proposed the seminal idea of convolutional neural networks. The two key points are the 2D equivariance (the processing does not depend on the location of the region of interest in the image) and the hierarchical processing (basic characteristic are first detected, and their composition is used to detect higher order characteristics, etc.).

The 2D equivariance in CNN is based on convolution layers, i.e. layers that convolve an image with a kernel. Formally, we suppose that an image is a tensor $x \in \mathbb{R}^{W \times H \times C}$ where $W$ is the width, $H$ the height and $C$ the number of channels (3 for RGB images). The output is a tensor $y \in \mathbb{R}^{W' \times H' \times C'}$ where $C'$ is the number of kernels or filters (an hyperparameter) and the width/height $W'$ and $H'$ depends on the original width and height, and also on the characteristics of the kernel $K$. Each filter allows to detect one pattern in the image region (e.g. a rounded corner in the lower layers or a face in the highest). The output of one element of this tensor is formally given by a linear combination of the input region centered on the point $(i, j)$:

$$y_{ijc'} = \sum_{l,m,c \in \mathbb{N}^3} K^{(c)}_{lmc} x_{i-l,j-m,c} + b_{c'} \tag{5.1}$$

where $K$ is a 2D convolution kernel (with a finite support) and $x_{i-l,j-m,c}$ is the pixel value of the image $c^{\text{th}}$ channel at coordinates $i-l, j-m$ (or 0 if the point does not belong to the image). Using 2D kernels allows to capture the equivariance by translation that should be verified by image processing techniques.

The hierarchical processing is obtained through the composition of several layers (convolution, followed by some other normalization or non linear layers) that extract hierarchical deep representations. For example, given the image of a person's face, the first layers can detect edges and corners, the next layers can detect larger patterns such as an eye or a nose, and the final layers can recognize face's shape Zeiler and Fergus (2014). The Figure 5.3 illustrates the different layers of a CNN, or more precisely shows an image region that activates the most the feature map of each layer.

In practice, the last layer (here, layer 5) contains larger image regions, and bears more semantics that the first layers. This layer is aggregated into a fixed-sized representation (e.g. by taking the maximum value of each channel) and then processed further with a fully connected (multi-layer perceptron) architecture.

This obtained image representation can then be used to recognize the object(s) contained in the image, and, more generally, to represent the image. In the remaining of the chapter, for an image $v$, its CNN representation $\underline{CNN}_v \in \mathbb{R}^{c_L}$ (where $c_L$ is the number of channels of the last CNN filter) corresponds to the last layer before the classification part of the network. As the classification part of the network is usually a simple linear layer (followed by a softmax to obtain


Figure 5.3 – Visualization of ConvNet learned features (from Zeiler and Fergus 2014)

a distribution over classes), $\underline{CNN}_v$ can be seen as a linear combination of vectors representing the different classes on which the CNN has been trained. This is for example exploited by (Norouzi et al. 2013) for Zero-Shot Learning in image recognition.

Despite being introduced in the 1980s, CNNs have only become widespread only from the 2010s. On a theoretical point of view, the rise can be explained by regularization techniques such as *dropout* (N. Srivastava, I. Hinton, et al. 2014) and the use of suitable activation functions such as ReLU (Glorot, Bordes, and Bengio 2011). In addition, practical reasons also explain the rise of CNN-based methods, such as the use of large datasets, e.g. MS COCO (T.-Y. Lin et al. 2014), *ImageNet* (Deng et al. 2009), and *Visual Genome* (Krishna et al. 2016), as well as fast implementations of convolutions and linear algebra computations on graphical processing units (GPUs).

In 2012, the *ImageNet* challenge has been won by a CNN-based method which outperformed previous approaches by a large margin (Krizhevsky, Sutskever, and G. E. Hinton 2012). Since then, CNN-based architectures have won all of the subsequent *ImageNet* competitions. The convolutional architecture have been improved, with deeper and larger layers, such as with the VGG network Simonyan and Zisserman 2014, the inception network Szegedy et al. 2015 and resnet network K. He et al. 2015.

Not only does the use of CNN has spread to a wide variety of tasks in computer vision — object recognition, detection and segmentation, image generation with GAN, and video analysis —, but CNNs have also been widely adopted for text (see Section 2) and other fields such as board games (e.g. the game of *go* with AlphaGo).

### 5.2.2 The structure of images and text

Having presented the evolution of image representation, we now discuss the differences between text and image.

| *Word* | *Teraword* | *Knext* | *Word* | *Teraword* | *Knext* |
|---|---|---|---|---|---|
| Spoke | 11,577,917 | 372,042 | Hugged | 610,040 | 11,453 |
| Laughed | 3,905,519 | 179,395 | Blinked | 390,692 | 21,973 |
| Murdered | 2,843,529 | 16,890 | Was late | 368,922 | 31,168 |
| Inhaled | 984,613 | 5,617 | Exhaled | 168,985 | 4,052 |
| Breathed | 725,034 | 41,215 | Was on time | 23,997 | 14 |

Table 5.1 – **Illustration of the Human Reporting Bias**. Count of the number of times that *A person may ...*, in Teraword and Knext textual corpora. Reproduced from (Gordon and Durme 2013).

**Differences between images and text** First, textual and visual modalities are trivially different by the way information is represented: language is composed of sequences of tokens (words) arranged in sentences, paragraphs, document, etc., while images are composed of channels (e.g. RGB), with pixel values spatially arranged in two dimensions. Beyond these obvious differences, these modalities do not bear the same semantics – even though they are both a biased view of reality:

- Visual data are direct depictions of the reality and are less subject to interpretation: views of objects and spatial organization of scenes in images are unequivocal.
- Conversely, language refers to high-level concepts, is highly ambiguous, relies on context and background knowledge (e.g. common-sense). For example, in the sentence "Hector went to the bank and swam", the reader needs to understand that the bank is a river bank, and to infer that the river was large enough to swim in it, etc.

This discrepancy between the textual and visual modalities can be explained by the fact that natural language is a communication mean for humans, and has been "optimized" through a long evolutionary process. For example, when people talk or write, they make the underlying assumptions that the people to which the language is addressed know about the world: many implicit facts are not mentioned because they are taken for granted by both parties (Grice 1975).

They are many illustration of this discrepancy. A first one is the discrepancies between real-world and textual statistics (Gordon and Durme 2013) reported in table 5.1. For example, based on $n$-gram occurrences count, they observe that a person is three times more likely to get murdered that to inhale. This discrepancy is called the "reporting bias", and stands for the fact that the more expected a situation, the less likely people are to report it.

**Bridging image and text** Several works have shown that even though the information contained in images and texts is of a different nature, they could be bridged. More pragmatically, it is possible to project an image onto the textual space, a text to the visual space, or both text and image to a common shared space. Such projections are primarily used for retrieval, i.e. finding an image where an object appears, or finding which objects appear in an image.

The fundamental hypothesis is that there exit a *smooth* mapping between the visual and the textual space, or more precisely a mapping that preserves the semantic of geodesics in both spaces. As illustrated in Figure 5.4, where we imagine that the space is regular enough so that even if a concept has not been learned (e.g. "tiger"), there is a high enough correlation between visual and textual features so that an approximate mapping can be learned. Formalizing this idea, (Costa Pereira et al. 2014) hypothesized that both semantic abstraction and cross-modal correlation are important, i.e. both image and text representations should bear high level semantics and be correlated in the common space.

Another view of the underlying hypothesis is that the identification of entities of the target domain (here, the visual domain) is made possible thanks to the implicit *principle of compositionality* (a.k.a. Frege's principle) — an object is formed by the composition of its attributes and

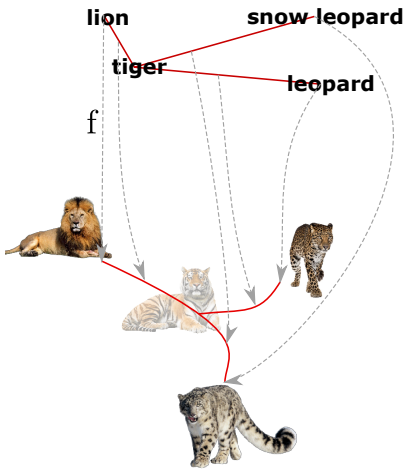



Figure 5.4 – Correlation between the image and textual spaces

characteristics — and the fact that other entities of the source domain share the same attributes. For example, if textual resources state that an apple is round and that it can be red or green, this knowledge can be used to identify apples in images because these characteristics (‘round‘, ‘red‘, ‘green‘) could be shared by classes of the source domain (e.g, ‘round‘ like a tennis ball, ‘red‘ like a strawberry...).

With neural-based representations (i.e. CNNs and word embeddings), the first hypothesis is somehow addressed, leaving the mapping between both spaces to be tackled. This does seem achievable since Collell and M.-F. Moens (2016) have shown that it was possible to predict feature norms from word embeddings – feature norms (McRae et al. 2005) are an attribute taxonomy where attributes are grouped by type (e.g., tactile or taxonomic). For example, `has_legs` is a form and surface attribute, while `is_a_bird` or `is_a_fruit` are instances of taxonomic attributes.

One of the first neural models exploiting this idea of a mapping is Devise (Deep Visual-Semantic Embedding Model) proposed by (Frome et al. 2013). This model learns to project images in the textual (word) representation space. Predicting the object in an image or the images containing an object amounts at finding the nearest neighbor in this space. This mapping between both spaces has been since then explored in various works (Akata et al. 2016; Bucher, Herbin, and Jurie 2017), and extended to captions in the image captioning models (Fang et al. 2014; Karpathy, Joulin, and F. F. F. Li 2014; Y. Li, Song, et al. 2016; Shekhar et al. 2017).

Predicting the visual appearance of a word (typical color or shape) is not the only information that can be squeezed out of word embeddings. For instance, Collell, Van Gool, and M.-F. Moens (2018) have shown that the relative position of two objects linked by a relation, e.g. “(boy, eating cheese)”, could be predicted from embeddings. In (E. Zablocki et al. 2019a), we show that textual embeddings can predict (to some extent) the visual frequency (i.e. how frequently an object appears in an image) and the contextual frequency (i.e. how frequently an object appears in an image provided another object is present).

However, due to their very different characteristics, the two spaces are still quite different. Collell and M.-F. Moens (2018a) have shown that, surprisingly, the neighborhood structure of the predicted vectors consistently resembles more that of the input vectors than that of the target vectors (i.e. words projected in the visual space), in the sense that for an entity $e$, the nearest neighbors of the projection of its textual representation in the visual space $f(\underline{t}_e)$ intersect more with the nearest neighbors of $\underline{t}_e$ than of $\underline{\text{cnn}}_e$. This shows that either the mapping between modalities is far from perfect with current models, or, more probably, that both spaces are too different to be mapped directly (as discussed in our work on grounded sentence embeddings presented in section 2.4).

## 5.3 Grounding Textual Embeddings

In this section, we now focus on (unsupervised) models that aim at grounding text representation. The general problematic is to take into account data coming from different sources and modalities (text, image, video) to learn to represent words, sentences or more generally texts, so that representation models for NLP can leverage common sense.

In section 5.3.1, we show that taking into account the visual context allows for a better word representation – in the sense that it increases the performances on tasks depending on this representation (synonyms, etc.), but this supposes a fine grained annotation (i.e. to know that an object appears in a picture at least). To really exploit the potentially available data, raw data must be used in each of the modalities and sources of information.

In section 5.3.2, we focus on sentence grounding, and study how to align images and sentences without requiring a common space, even with the hypothesis that the geometry of the visual world is quite different from the one of the textual one.

Finally, in section 5.3.3, we describe a model that tries to leverage temporal information (i.e. at what date was the text written), based on the idea that word embeddings describe trajectories in the representation space.

### 5.3.1 Grounding Words using Visual Context

In this section, we present works that seek to *ground word* representation using visual information. We can distinguish various models depending on the way to use the visual information, that we can classify into:

- *indirect approaches* that produce representations by learning to solve a particular task (e.g. representations as a byproduct of a captioning task);
- *fusion* approaches that (learn to) combine pre-learned visual and textual *representations*;
- *alignment* approaches that learn a multimodal representation from textual and visual *data*

**Indirect approaches** Many multi-modal tasks imply learning a multimodal or grounded representation of word or sentences, as, for instance, in captioning (Mansimov et al. 2015; Vinyals et al. 2014; K. Xu et al. 2015) or in Visual Question Answering (Goyal et al. 2016; Hu et al. 2017; Y. Zhu, Groth, et al. 2015; Y. Zhu, C. Zhang, et al. 2015). (Rohrbach et al. 2016) tackle the visual grounding task with the *GroundeR* model, which computes compatibility of an image region with a phrase with multimodal fusion. While these works could be useful in the future, we do not focus on those since these models are hard to train and not generic enough to be used as a valid pre-training procedure.

**Fusion approaches** The simplest approach for grounding words is to combine the representations from various modalities, which have been learned on mono-modal data, into a multimodal embedding.

In the simplest case, we suppose that for each entity $e$ there exists a visual representation $\underline{v}_e$ and a textual representation $\underline{t}_e$, and we wish to compute a grounded representation $\underline{g}_e$. The simplest fusion functions are concatenation: $f_\theta(e) = \underline{t}_e \oplus \underline{v}_e$, where $\oplus$ designates the concatenation operator (Kiela and Bottou 2014). To reduce the space and capture linear correlations between both modalities, a Singular Value Decomposition (SVD) can be conducted on the concatenation (Bruni et al. 2012a).

An alternative is to compute the projection into a common multimodal space. Silberer and Lapata (2012) use a Canonical Correlation Analysis (CCA) to compute a projection in a multimodal space that maximizes the correlation between projected representations. Extensions

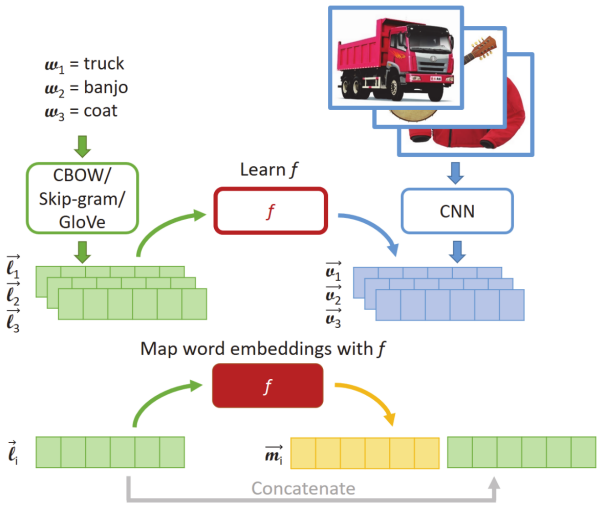



Figure 5.5 – **Example of fusion.** Image taken from (Collell, T. Zhang, and M.-F. Moens 2017).

of CCA include the use of kernelized version of CCA (Bach and M. I. Jordan 2002; Hill, Reichart, and Korhonen 2014), and the use of non-linear projections in the deep CCA (Andrew et al. 2013; Klein et al. 2015). As an alternative to projection, Hill, Reichart, and Korhonen (2014) use a Weighted Gram Matrix Combination where visual and textual similarity matrices are combined to produce a multimodal representation – thereby combining visual and textual kernels.

All the methods above are symmetric, in the sense that they give the same importance to textual and visual information. When learning grounded representations, it might be useful to focus on the textual information, and integrate into it some visual information. The approach taken by Collell, T. Zhang, and M.-F. Moens (2017) follows this path, combining models that bridge modalities (see section 5.2) and fusion approaches. In their work, illustrated in Figure 5.5, they first learn a non linear projection $f_{\mathcal{T}\to\mathcal{V}}$ function from the textual representation to the visual one (bridging), before conducting an SVD on the concatenation, for each entity $e$, of the textual representation $\underline{t}_e$ and the so-called imagined one $f_{\mathcal{T}\to\mathcal{V}}(\underline{t}_e)$. With this model, even abstract words, which do not have associated visual features, can benefit from grounding since $f_{\mathcal{T}\to\mathcal{V}}$ can be computed for any entity.

**Alignment models: the Strong Grounding Hypothesis** Multimodal fusion models prevent potentially beneficial interactions during training between the different modalities. This relies on the hypothesis that the visual and textual representation of concepts should *interact* to produce meaningful results. For instance, knowing that an apple is round can modify the visual representation space by biasing it towards representing "explicitly" the concept of being round. This in turn can modify the textual representation of all words referring to "round" objects. This follows the way humans are supposed to learn grounded meaning in semantics (Barsalou 2008; Glenberg and Kaschak 2002).

This hypothesis, that we name the "*Strong Grounding Hypothesis*", has important implications in terms of efficiency since this implies that optimal representations *must be computed by letting modal representations interact* – or said otherwise, visual and textual representations cannot be merged through a simple process.

An instance of models capturing this interaction are alignment models that directly learn a joint representation from textual and visual inputs. Some joint models require aligned texts and images. For example (Roller and Schulte im Walde 2013) use a Bayesian modeling approach based on the assumption that text and associated images are generated using a shared set of underlying latent topics and (Kottur et al. 2016) ground word representations into vision by

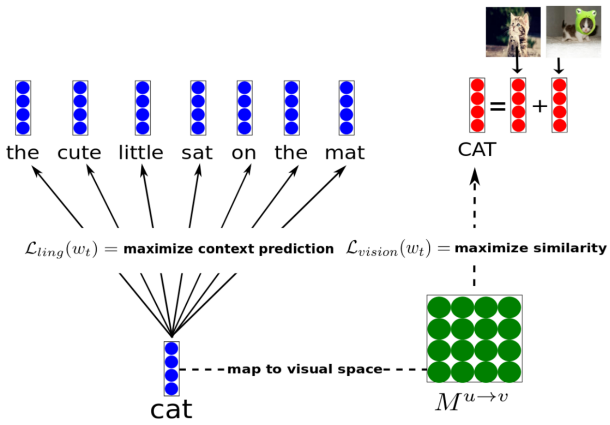



Figure 5.6 – **Example of alignment**. Image taken from (Lazaridou, N. T. Pham, and Baroni 2015).

trying to predict the abstract scene associated to a given sentence.

Extensions of the Skip-Gram algorithm have also been proposed, and they have the advantage of not relying on aligned text and images. For example, (Hill and Korhonen 2014b) base their model on the assumption that the frequency of appearance of concrete concepts correlates with the likelihood of "experiencing" it in the world. Perceptual information for concrete concepts is then introduced to the model whenever that concept is encountered in the textual modality. Representations of concrete words are trained to predict surrounding words (as in the classical Skip-Gram model) and the perceptual features, which are feature-norms (McRae et al. 2005) that describe objects as a set of features (typical color, usage, etc.).

The work by (Hill and Korhonen 2014b) was later followed by (Lazaridou, N. T. Pham, and Baroni 2015), whose method is uses natural images instead of the handcrafted feature-norms. They force the representation of words for which they have images to be close to their visual (pre-trained) representation — their approach is illustrated in Figure 5.6. More precisely, for any visual entity $e$, they build visual a vector $\underline{v_e}$ from the average of activations obtained with a pre-trained ResNet applied on 100 images (taken from ImageNet) of the entity $e$. During training, along with a purely textual Skip-Gram loss, the similarity between the embedding $\underline{t_e}$ of the entity $e$ and its visual appearance $\underline{v_e}$ is maximized in a max-margin framework:

$$\mathcal{L} = \sum_{e \in \mathcal{D}} \sum_{v^-} \max(0, \gamma - \cos(\underline{t_e}, \underline{v_e}) + \cos(\underline{t_e}, \underline{v^-})) \tag{5.2}$$

where $\gamma$ is the margin and $v^-$ is the visual appearance of a "negative" object (randomly sampled over all objects, with uniform distribution). The visual representation $\underline{v_e}$ of each entity $e$ is kept fixed throughout the optimization of the loss function.

#### 5.3.1.1 Model

The multimodal grounding models presented in the previous section mostly ignore the *visual context* of objects. This has to two drawbacks. First, there is an asymmetry in the consideration of the modalities: text defines a semantic context for each word — i.e. its surrounding words — while images are used to gather visual information about the object. Second, it ignores that context in which objects appear is informative and complementary to textual inputs to improve word representation.

Indeed, this fact is suggested by several works, such as (Bruni et al. 2012a) who propose a fusion approach where a visual embedding is built by factorizing the matrix counting visual words in images. This is the first attempt to apply the distributional hypothesis to images: *Semantically similar objects will tend to occur in similar environments in images.* Through

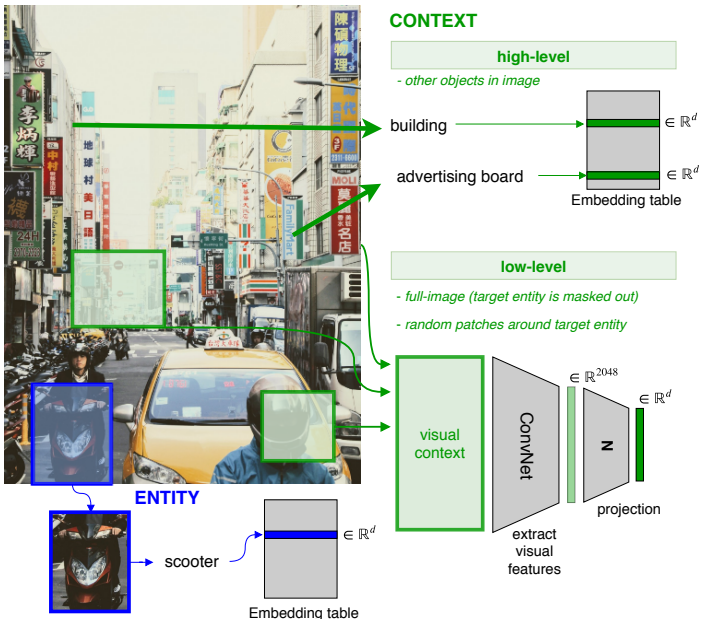



Figure 5.7 – **Overview of the model.** An object is chosen as the target entity and is represented with an embedding table. Two options are considered to define its context. (1) *high-level*: the list of the other objects in the image is known and the context is represented using an embedding table. (2) *low-level*: pixel values of surrounding image patches (or the full-image without the zone containing the entity) are processed by a ConvNet to compute a context vector.

their experiments, they come to the conclusion that the appearance of the surrounding objects is more informative for semantics than the appearance of the object itself. In comparison to the model we present in this chapter, their work does not propose to jointly learn embeddings from both visual and textual context.

We posit that contexts in different modalities should be a key component of grounding model, and propose a model that captures it. The different components of the model are illustrated in Figure 5.7.

Based on the original Skip-Gram algorithm that considers entities $e$ (words) and their contexts $\mathcal{C}_e = \{c_1, ..., c_n\}$ ($n$ surrounding words within a window centered on the entity), we translate in what follows the distributional hypothesis for images.

In our case, the contexts $\mathcal{C}_e$ are visual contexts. The choice for visual context elements $c \in \mathcal{C}_e$ does not need to correspond to a list of semantic entities, as shown by (M. R. Rudolph et al. 2016) who propose a generalization of the Skip-Gram algorithm with arbitrary contexts. For instance, visual context elements can be the surrounding objects, low-level features such as the visual appearance, or also the localization of the surrounding objects with respect to the considered entity.

With this in mind, we define a function $f_\theta$, parameterized by $\theta$ (learned), such that for any entity $e$ and visual context element $c \in \mathcal{C}_e$, $f_\theta(c)$ is a vector of $\mathbb{R}^d$ which corresponds to the representation of the context in the *textual* semantic space. These representations are then used in the negative-sampling loss:

$$\mathcal{L}_{\text{i}} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[ \log \sigma(f_\theta(\underline{c})^\top \underline{t}_e) + \sum_{c^-} \log \sigma(-f_\theta(\underline{c^-})^\top \underline{t}_e) \right] \tag{5.3}$$

where $\mathcal{D}$ is the set of entities, $\underline{t}_e$ is the (learned) embedding associated with the entity/word $e$, $c^-$ is a negative context (as in many other works, it is sampled uniformly from the set of possible contexts), and $\sigma$ is the sigmoid function. This loss formulation is very close to the

original Skip-Gram loss but integrates the learning of $f_\theta$, a function that projects the visual those tackled representation into the textual space.

Given an entity $e$, we explored different ways of modeling an instance of visual context elements $c \in \mathcal{C}_e$. First, we used a high-level context where we suppose that we know precisely which entities are present in the context of the object – in that case it is possible to use the textual embedding the context $c$, i.e. $\underline{c} = \underline{t_c}$. Second, we used a more realistic representation where we only assume that the bounding box of the different entities is known. In this case, the representation of a context element $c$ is the CNN representation $\underline{\text{cnn}_c}$. Third, in the most unfavorable case, only the bounding box of the entity $e$ is known. In this case, the context is the full image where the entity is wiped out using a black mask. These three models correspond respectively to $\mathbf{V}_O$, $\mathbf{V}_B$ and $\mathbf{V}_M$ in the experiments.

We now present our multimodal representation learning model that integrates the previously presented visual module with the textual Skip-Gram. The main idea is that while word embeddings should be shared across modalities, context is media-specific. Indeed, imposing shared context representations would be over-constraining, because of the bias that affects contexts distribution across modalities, The contribution of each modality is controlled by a linear combination (hyper-parameter $\alpha$, determined by cross-validation) of modality-specific costs, which gives the following global loss function:

$$\mathcal{L}(\underline{T}, \underline{U}, \theta) = \mathbf{T}(\underline{T}, \underline{U}) + \alpha \mathbf{V}(\underline{T}, \theta) \tag{5.4}$$

where $\underline{T}$ (resp. $\underline{U}$) denotes the textual entity (resp. context) embeddings, $\mathbf{T}(\underline{T}, \underline{U})$ the Word2Vec loss function and $\mathbf{V}(\underline{T}, \theta)$ the visual Skip-Gram. The main outcome of the model are the word embeddings $\underline{T}$.

A crucial point is that this model does not require aligned texts and images to train the model, or extra pre-trained representations on external datasets – we only require that entities identified in images be associated with a unique word of the vocabulary. Besides, we justify the use of a joint model as we think it is important that representations are learned both for entities and for contexts. Indeed, as the entities embeddings are affected by both modalities, the context representations should change and be updated by transitivity between modalities through the shared embeddings.

#### 5.3.1.2 Experiments

In this section, we evaluate the learned word embeddings on different tasks. In particular, we measure the performance of word embeddings built from visual data and multimodal data. We evaluate the following models:

1. The text-only Skip-Gram model (noted $\mathbf{T}$).

2. Our Visual Skip-Gram model, denoted $\mathbf{V}_O$, $\mathbf{V}_B$ and $\mathbf{V}_M$ for the three different context (known entities, bounding boxes, and masked) – we use $\mathbf{V}$ when denoting our model in general.

3. Our visually contextualized model denoted $\mathbf{V}_O{+}\mathbf{T}$, $\mathbf{V}_B{+}\mathbf{T}$ and $\mathbf{V}_M{+}\mathbf{T}$ where embeddings are learned by minimizing the full multimodal loss of Eq. (5.4).

4. A sequential model, noted $\mathbf{V} \oplus \mathbf{T}$, where embeddings of model $\mathbf{T}$ are concatenated with embeddings obtained from $\mathbf{V}$ and then projected in a lower-dimensional space with PCA. This serves as a comparison point between our joint approach and a sequential one.

5. A baseline inspired by the state-of-the-art model of (Lazaridou, N. T. Pham, and Baroni 2015). During training, along with the purely-textual Skip-Gram loss, they maximize a max-margin loss $\mathbf{A}$ (for visual **A**ppearance) the similarity between the embedding $t_e$ of the

entity $e$ and its visual appearance $v_e$ . We denote this model $\mathbf{L} + \mathbf{A}$ where $\mathbf{A}$ corresponds to the visual loss and $\mathbf{T}$ the text-only Skip-Gram loss.

6. Our full model denoted $\mathbf{A} + \mathbf{V}_O + \mathbf{T}$, that uses textual context and both visual appearance and context.

Similarly to previous work (Collell, T. Zhang, and M.-F. Moens 2017; Lazaridou, N. T. Pham, and Baroni 2015), we evaluate our model on three different semantic tasks, namely word similarity and relatedness, feature norm prediction (predicting object attributes, such as "can fly"? "can be eaten"? etc.), and abstractness/concreteness prediction (i.e. predict that apple is concrete while freedom is abstract). Note that the latter, concreteness, can only be evaluated when using the textual loss. Each task serves as a weak indicator of the quality of the embeddings.

| | | | Encyclopedic | Taste | Sound | Taxonomic | Function | Tactile | Color | Shape | Motion | Conc. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Baselines** | Text | $\mathbf{T}$ | *58* | *52* | *44* | *79* | *62* | *11* | *32* | *54* | *60* | *42.1* |
| | Sequential | $\mathbf{V}_O \oplus \mathbf{T}$ | 5 | 3 | -4 | -7 | -3 | 1 | 3 | 0 | -2 | 2 |
| | Appearance | $\mathbf{A}$ | -2 | -3 | -8 | -3 | -6 | 6 | 9 | 6 | -2 | *n/a* |
| | Joint | $\mathbf{A} + \mathbf{T}$ | 3 | 3 | -2 | 1 | -3 | 0 | -1 | 0 | 2 | 1 |
| **Our models** | Visual | $\mathbf{V}_B$ | -28 | -1 | -21 | -31 | -25 | -7 | -8 | -16 | -30 | *n/a* |
| | | $\mathbf{V}_M$ | -28 | -4 | -14 | -33 | -27 | -5 | -9 | -19 | -33 | *n/a* |
| | Objects | $\mathbf{V}_O$ | -10 | -6 | -9 | -17 | -14 | -8 | -11 | -11 | -24 | *n/a* |
| | Combinations | $\mathbf{A} + \mathbf{V}_O$ | 0 | 0 | -2 | -5 | -6 | -9 | -5 | -1 | -7 | *n/a* |
| | | $\mathbf{V}_B + \mathbf{T}$ | 2 | 4 | 5 | 3 | -2 | 1 | 0 | 1 | 1 | 1 |
| | | $\mathbf{V}_M + \mathbf{T}$ | 2 | 3 | 0 | 3 | 1 | 3 | 0 | 1 | -1 | 1 |
| | | $\mathbf{V}_O + \mathbf{T}$ | 4 | 3 | 2 | 3 | -1 | 2 | 1 | 1 | 1 | 1 |
| | | $\mathbf{A} + \mathbf{V}_O + \mathbf{T}$ | 5 | 3 | 6 | 3 | -2 | -1 | 1 | 1 | -1 | 2 |

Table 5.2 – **Grounded words: feature norm prediction and concreteness** (results are given for $\mathbf{T}$, and then relative to the performance of $\mathbf{T}$). Feature norm results are detailed by category. Scores for the feature-norm prediction task are f1-scores (multiplied by 100). Concreteness measures (conc.) are coefficients of determination ($R^2$) given in percentage.

We report results in table 5.2 (feature norm and concreteness) and table 5.3 (for word similarity). We report the main observations below.

**Overall** Results highlight that all of the trained multimodal models outperform the text-only baseline on all evaluation tasks (when used with the textual loss $\mathbf{T}$). For instance, $\mathbf{V}_O + \mathbf{T}$ shows an average improvement of 9% over $\mathbf{T}$. This is in-line with the conclusions of related works, and shows that contextual grounding has a positive impact on basic properties of word embeddings. Whether this translates into a better performance for extrinsic tasks such as Question-Answering was however not investigated in this work (we report positive results on sentences in the next section).

**Contextual information** Using low-level visual features is a challenging problem. However, they are promising since they are cheap to collect, do not require context annotations, and contain rich information if handled correctly. The difficulty lies in the natural noise in the surroundings of objects and the need for visual modules that automatically extract high-level information from raw pixel values.

In our experiments, we observe that the higher level the representation of the context is ($\mathbf{V}_O$, followed by $\mathbf{V}_M$ and $\mathbf{V}_B$), the better the performance – but this difference fades away when

| | | | VisSim | SemSim | Simlex | MEN | WordSim |
|---|---|---|---|---|---|---|---|
| **Baselines** | Text | $\mathbf{T}$ | *48* | *60* | *33* | *69* | *63* |
| | Sequential | $\mathbf{V}_O \oplus \mathbf{T}$ | 1 | 2 | 0 | 2 | 1 |
| | Appearance | $\mathbf{A}$ | 5 | -15 | -17 | -47 | -46 |
| | Joint | $\mathbf{A} + \mathbf{T}$ | 4 | 5 | 1 | 2 | 2 |
| **Our models** | Visual | $\mathbf{V}_B$ | -20 | -25 | -16 | -34 | -41 |
| | | $\mathbf{V}_M$ | -13 | -18 | -14 | -26 | -35 |
| | Objects | $\mathbf{V}_O$ | -5 | -6 | -2 | -5 | -36 |
| | Combinations | $\mathbf{A} + \mathbf{V}_O$ | -3 | -3 | 0 | -3 | -29 |
| | | $\mathbf{V}_B + \mathbf{T}$ | 5 | 5 | 2 | 3 | 4 |
| | | $\mathbf{V}_M + \mathbf{T}$ | 5 | 5 | 1 | 4 | 2 |
| | | $\mathbf{V}_O + \mathbf{T}$ | 5 | 6 | 2 | 6 | 4 |
| | | $\mathbf{A} + \mathbf{V}_O + \mathbf{T}$ | 6 | 6 | 2 | 6 | 2 |

Table 5.3 – **Word similarity evaluation** (results are given for $\mathbf{T}$, and then relative to the performance of $\mathbf{T}$). Scores are Spearman correlations (multiplied by 100) on the word similarity benchmarks, and relative to the Skip-Gram baseline $\mathbf{T}$

adding the textual loss $\mathbf{T}$, showing that even using a very crude representation of the context (an image where the target entity is blacked out), this information can still be leveraged.

**Contextual and visual appearance** The appearance and the context are complementary sources of information: this can be seen by observing that $\mathbf{A} + \mathbf{T}$ and $\mathbf{V}_O + \mathbf{T}$ perform both worse than $\mathbf{L} + \mathbf{V}_O + \mathbf{T}$. More in detail, we can see that the effect of each loss has a different impact depending on the tested properties:

- Visual appearance improves the feature norm prediction that describe visually the objects (e.g. *is_red* in "Color" category or *is_round* in the "Shape" category) but not for the other non visual categories such as "Encyclopedic", "Taste" and "Sound".
- Context improves the performance in word similarity benchmarks: for instance, results of $\mathbf{V}_M$ is on average 29% higher than those of the appearance baseline $\mathbf{A}$ (table 5.3).

**Interaction of the visual and textual representations** In the experiments conducted here, we observe that the *strong grounding hypothesis* (section 5.3.1 on page 63) holds in this case: using a joint optimization scheme, where textual and visual representations interact, is better since the performance of $\mathbf{V}_O + \mathbf{T}$ is better than that of $\mathbf{V}_O \oplus \mathbf{T}$.

**Concrete and abstract words** Finally, to get a deeper insight into learned embeddings, we aim at explaining the impact of the visual modality on the multimodal word representation. To do so, with the model $\mathbf{O} + \mathbf{T}$, we estimate the correlation between the shift measured on the embedding (the norm of the difference of the initial textual embedding and the final multimodal embedding), and the concreteness degree of a word. We measured a correlation $\rho_{\text{Spearman}} = 0.33$, showing that visual and concrete words see their embeddings being more changed than other non visual and abstract words. This was to be expected because the visual part only adds information to visual entities.

#### 5.3.1.3 Conclusion

In this work, we have shown that it was possible to ground words using two complementary sources of information, namely the visual appearance and the visual context in which an entity

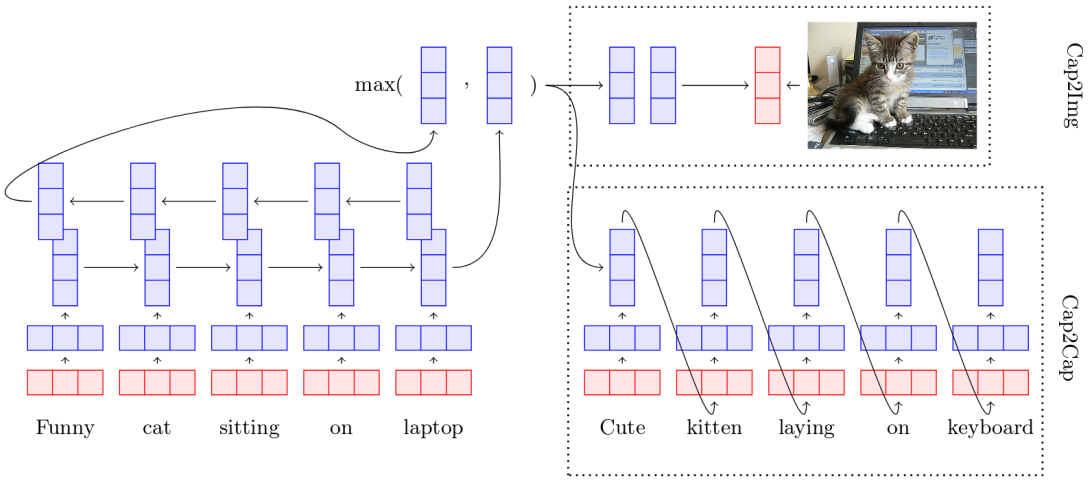



Figure 5.8 – **Learning visually grounded sentence representations**. Illustration taken from (Kiela, Conneau, et al. 2018)

appears. Through extensive experiments, and in line with related work, we observed the complementarity of visual and textual data to learn word representations. This work shows that visual information, in the form of visual contexts, can be integrated in a semantic space along with textual data.

In the paper that describes in details this work (É. Zablocki et al. 2018), we show that it is possible to use spatial relationship (e.g. measuring whether an object is on top of another one) to further improve the quality of embeddings (but with less impact that when using context). Extensions of this work could look at how to leverage other sources of information (e.g. knowledge bases), and looking at more sophisticated methods to compare the learned representations.

### 5.3.2 Grounding Sentences

While the literature is abundant about learning grounded *word* representations, there exist far less methods learning grounded *sentence* representations. Note that we exclude from the discussion *indirect* methods that learn representations through a task (e.g. captioning), since their representation is too biased towards the visual modality and perform worse (Chrupala, Kádár, and Alishahi 2015).

We are aware of two works which learn sentence representations, and both of them leverage images aligned with sentences in captioning datasets, which are based on works learning sentence representation (see section 2.4), i.e. works that try to leverage raw text by using losses inspired by word2vec but at the sentence level. Differently from captioning models, they also include a textual loss that allows the model not to be too biased by the visual modality. We describe the few existing works below.

(Chrupala, Kádár, and Alishahi 2015) propose Imaginet, a model where two recurrent neural network models, coupled through common word embeddings, are trained on a textual task (language model) and on a visual task (predicting the image, or more precisely its CNN representation).

Inspired by Imaginet, the model of (Kiela, Conneau, et al. 2018) additionally exploits the fact that the same image is associated with several images (in the dataset they use) by making a sentence predict both the image (Cap2Img, a grounding objective) and another caption of the same image (Cap2Cap, inspired by SkipThought by Kiros et al. 2015). An illustration of the model is given in 5.8. The final sentence representations are obtained by concatenating (1) purely-textual SkipThought representations, and (2) grounded sentence vectors obtained with the Cap2Cap or Cap2Img (or both).

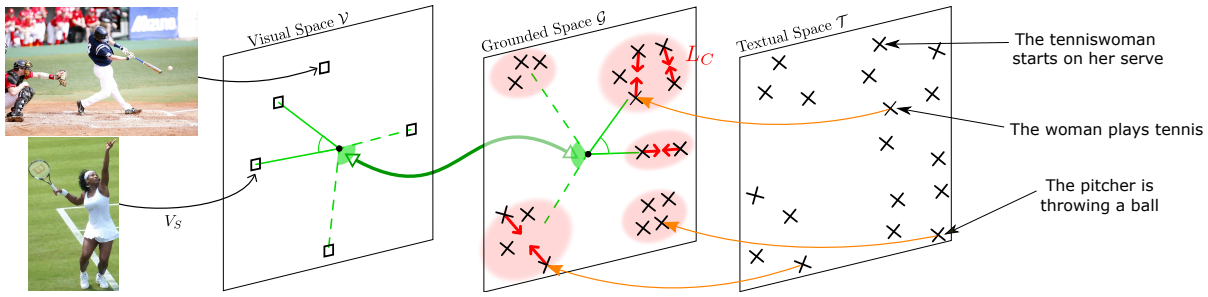



Figure 5.9 – Model overview. Red circles indicate visual clusters. Red arrows represent the gradient of the cluster loss, which gathers visually equivalent sentences — the contrastive term in loss $\mathcal{L}_C$ is not represented. The green arrow and angles illustrate the perceptual loss, ensuring that cosine similarities correlate across modalities. The origin is at the center of each space.

#### 5.3.2.1 Incorporating visual semantics within an intermediate grounded space

The previous models (Chrupala, Kádár, and Alishahi 2015; Kiela, Conneau, et al. 2018) rely on paired textual and visual data. Moreover, the hypothesis of a one-to-one correspondence between modalities is implicitly assumed. An image of an object unequivocally represents a sentence. However, there is no obvious reason implying that the structure of the two spaces should match. Indeed, (Collell and M.-F. Moens 2018b) empirically show that cross-modal projection of a source modality does not resemble the target modality in terms of its neighborhood structure. This is especially the case for sentences, where many different sentences can describe a similar image and where many images can be described by the same sentence. Therefore, we argue that learning grounded representations with projections to a visual space is particularly inadequate in the case of sentences.

To overcome this issue, we proposed an alternative approach where the structure of the visual space is *partially transferred* to the textual space. The first contribution is to preserve textual semantics and to avoid an over-constrained textual space by incorporating the visual information to textual representations using an intermediate representation space that we call *grounded space*. The second contribution is to distinguish two types of complementary information sources. First, the *cluster information*: the implicit knowledge that sentences associated with the same image refer to the same underlying reality. Second, the *perceptual information*, which is contained within high-level representations of images. These two sources of information aim at transferring the structure of the visual space to the textual space, through the grounding space.

**Model overview** We aim at learning sentence grounded representations by jointly leveraging the textual and visual contexts of a sentence. We note $s$ a sentence and $\underline{s} = f_{\mathcal{T}}(s; \theta_t)$ its representation computed with a sentence encoder $f_{\mathcal{T}}$ parameterized by $\theta_t$. We follow the classical approach developed in the language grounding literature at the word level (Mikolov et al. 2013; É. Zablocki et al. 2018), which balances a textual objective $\mathcal{L}_{\mathcal{T}}$ with an additional grounding objective $\mathcal{L}_{\mathcal{G}}$:

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathcal{T}}(\theta_t) + \mathcal{L}_{\mathcal{G}}(\theta_t, \theta_i) \tag{5.5}$$

The parameters $\theta_t$ of the sentence encoder $f_{\mathcal{T}}$ are shared in $\mathcal{L}_{\mathcal{T}}$ and $\mathcal{L}_{\mathcal{G}}$, and therefore benefit from both textual and grounding objectives. $\theta_v$ denotes extra grounding parameters, including the weights of the image encoder $f_{\mathcal{V}}$.

Note that any textual objective $\mathcal{L}_{\mathcal{T}}$ and sentence encoder $f_{\mathcal{T}}$ can be used. In our experiments, we choose the well-known SkipThought model (Kiros et al. 2015), trained on a corpus of ordered sentences.

In what follows, we therefore focus on the modeling of the grounding objective $\mathcal{L}_\mathcal{G}$, learned on a captioning corpus, where each image is associated with several captions. Grounding approaches generally leverage visual information by embedding textual and visual elements within the same multimodal space (Kiela, Conneau, et al. 2018; Silberer and Lapata 2014). However, it is not satisfying since texts and images are forced to be in one-to-one correspondence. Moreover, a caption can (1) have a wide variety of paraphrases and related sentences describing the same scene (e.g., *the kitten is devouring a mouse* versus *a cat eating a mouse*), (2) be visually ambiguous (e.g., *a cat is eating* can be associated with many different images, depending on the visual scene/context), or (3) carry non-visual information (e.g., *cats often think about their meals*). Usual grounding objectives, that embed sentences in the visual space, can discard non-visual information (3) through the projection function. They can handle (1) by projecting related sentences to the same location in the visual space. However, they are over-sensitive to visual ambiguity (2), because ambiguous sentences should be projected to different locations of the visual space, which is not possible with current grounding models.

To overcome this lack of flexibility, we propose the following approach, illustrated in Figure 5.9. To cope with (1), sentences associated with the same image should be close — we call this *cluster information*. To cope with (2), we want to avoid projecting sentences to a particular point of the visual space: instead, we require that the similarity between two images in the visual space (which is linked to the "context discrepancy") should be close to the similarity between their associated sentences in the textual space. We call this *perceptual information*. Finally, as we want to preserve non-visual information in sentence representations (3), we make use of an intermediate space, called *grounded space*, that allows textual representations to benefit from visual properties without degrading the semantics brought by the textual objective $\mathcal{L}_\mathcal{T}$.

**Grounding space and objectives** We now introduce more formally the grounded space and the different information (cluster and perceptual) captured in the grounding loss $\mathcal{L}_\mathcal{G}$. We suppose that we dispose of a set $\mathcal{D} = \{(s, v)\}$ of matching pairs of captions $s$ and images $v$.

**Grounded space** The grounded space relaxes the assumption that textual and visual representations should be in a one-to-one correspondence. It rather assumes that the structure of the textual space might be partially modeled on the structure of the visual space. The underlying hypothesis is that the representation of a sentence contain information which has nothing to do with the visual information, and that moreover the structure of the textual and visual representation spaces is very different.

In practice, instead of directly applying the grounding objectives on a sentence embedding $\underline{t}_s$, we propose to train the grounding objective $\mathcal{L}_\mathcal{G}$ on an intermediate space we name the "*grounded space*". Practically, we use a projection $\underline{g}_s = \underline{g}\,(s; \theta_g)$ of a sentence s from the textual space to the grounded space, where $g$ is a multi-layer perceptron applied to the textual representation $\underline{t}_s$ of the sentence s. The representation $\underline{t}_s$ can be obtained by any sentence embedding model. In our case, we use a RNN to obtain the representation $\underline{rnn}_s$ as the final state of the recurrent architecture.

**Visual equivalence: the cluster loss $\mathcal{L}_C$** Without considering any visual information, it is already possible to exploit the fact that two sentences describe, or not, the same underlying reality. For convenience, two sentences are said to be *visually equivalent* (resp. *visually different*) if they are associated with the same image (resp. different images), i.e. if they describe the same (resp. different) underlying reality – allowing to exploit dataset where several captions are associated with an image as in (Kiela, Conneau, et al. 2018). We call *cluster* a set of visually equivalent sentences. For instance, in Figure 5.9, the sentences *"The tennis-woman starts on her serve"* and *"The woman plays tennis"* are visually equivalent and belong to the same cluster.

Our hypothesis is that *the similarity between visually equivalent sentences* $(s, s^+)$ *should be higher than visually different sentences* $(s, s^-)$. We translate this hypothesis into the constraint in the grounded space. Following (Carvalho et al. 2018; Karpathy and F.-F. Li 2015), we use a max-margin ranking loss to ensure the gap between both terms is higher than a fixed margin $\gamma$ (cf. red elements in Figure 5.9) resulting in the *cluster loss* $\mathcal{L}_\mathcal{C}$:

$$\mathcal{L}_C = \sum_{(s,s^+,s^-)} \lfloor \gamma - \cos(\underline{g}_s, \underline{g}_{s^+}) + \cos(\underline{g}_s, \underline{g}_{s^-}) \rfloor_+ \tag{5.6}$$

where $s^+$ (resp. $s^-$) is a randomly sampled visually equivalent (resp. different) sentence to $s$. This loss function is also used in the cross-modal retrieval literature to enforce structure-preserving constraints between sentences describing a same image (L. Wang, Y. Li, and Lazebnik 2016).

**Leveraging the visual information: the perceptual loss $\mathcal{L}_\mathcal{P}$** The cluster hypothesis alone ignores the structure of the visual space and only uses the visual modality as a proxy to assess if two sentences are visually equivalent or different. Moreover, the ranking loss $\mathcal{L}_\mathcal{C}$ simply drives apart visually different sentences in the representation space, which can be a problem when two images have a closely related content. For instance, the baseball and tennis images in Figure 5.9 may be different, but they are both sports images, and thus their corresponding sentences should be somehow close in the grounded space. Finally, it supposes that we have a dataset of images associated with several captions.

To cope with these limitations, we consider the structure of the visual space and use the content of images. The intuition is that the structure of the textual space should be modeled on the structure of the visual one to extract visual semantics. We choose to preserve *similarities* between related elements across spaces (cf. green elements in Figure 5.9). We thus assume that *the similarity between two sentences in the grounded space should be correlated with the similarity between their corresponding images in the visual space.* We translate this hypothesis into the perceptual loss $\mathcal{L}_\mathcal{P}$:

$$\mathcal{L}_\mathcal{P} = -\rho(\left\{ \left( \cos\left(\underline{g}_{v_1}, \underline{g}_{v_2}\right), \cos\left(\underline{\text{cnn}}_{v_1}, \underline{\text{cnn}}_{v_2}\right) \right) \mid (v_1, s_1), (v_2, s_2) \in \mathcal{D} \right\}) \tag{5.7}$$

where $\rho$ is the Pearson correlation computed for series of cosine between captions. Note that the choice of cosine is arbitrary, and it might interesting to look at different ways to match texts or images in their respective representation spaces.

**Grounded loss** Taking altogether, the grounded space and cluster/perceptual information leads to the grounding objective $\mathcal{L}_\mathcal{G}$ as a linear combination of the aforementioned objectives:

$$\mathcal{L}_\mathcal{G} = \alpha_C \mathcal{L}_C + \alpha_P \mathcal{L}_P \tag{5.8}$$

where $\alpha_C$ and $\alpha_P$ are hyper-parameters weighting contributions of $\mathcal{L}_C$ and $\mathcal{L}_P$.

#### 5.3.2.2 Evaluation protocol

For the textual corpus, following (Hill, Cho, and Korhonen 2016; Kiros et al. 2015), we use the Toronto BookCorpus dataset. This corpus consists of 11K books, and 74M ordered sentences, with an average of 13 words per sentence. For the visual corpus, we use the MS COCO (T.-Y. Lin et al. 2014). This image captioning dataset consists of 118K/5K/41K (train/val/test) images, each with five English descriptions.

In the experiments, we focus on one of the most established sentence models at the time of writing of the papers, namely SkipThought, noted **T** as the **T**extual baseline. The parameters of the sentence embedding model are obtained by minimizing $\mathcal{L}_\mathcal{T}$. All the grounding models (baselines and our own models) are then based on **T**.

| Model | Structural measures | | | | Semantic relatedness | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | MNNO | $\rho_{vis}$ | $C_{inter}$ | $C_{intra}$ | STS | | | | SICK |
| | | | | | All | Cap | News | Forum | |
| T | 10.0 | 4.1 | 54.2 | 70.1 | 30 | 41 | 36 | 21 | 51 |
| CM (text) | 24.2 | 12.8 | 41.7 | 74.8 | 52 | 76 | 42 | **37** | 55 |
| $\mathtt{P}_{id}$ | 21.1 | **37.9** | 42.2 | 69.3 | 45 | 66 | 41 | 34 | 54 |
| $\mathtt{C}_{id}$ | 27.5 | 10.5 | **2.9** | **84.7** | 60 | 83 | 45 | 20 | 55 |
| $\mathtt{C}_{id} + \mathtt{P}_{id}$ | **27.9** | 25.8 | 6.7 | 82.6 | **61** | **84** | **46** | 28 | **57** |
| CM (vis.) | 27.1 | 19.2 | 1.5 | 85.8 | 56 | 78 | 40 | 34 | 55 |
| $\mathtt{P}_g$ | 21.3 | **32.4** | 43.9 | 73.3 | 45 | 66 | 41 | **37** | 53 |
| $\mathtt{C}_g$ | 28.6 | 9.4 | **1.1** | **88.5** | 62 | 83 | 46 | 29 | 59 |
| $\mathtt{C}_g + \mathtt{P}_g$ | **28.9** | 29.1 | 4.7 | 87.5 | **63** | **84** | **48** | 33 | **60** |

Table 5.4 – Intrinsic evaluations carried out on the grounded space for models with $g$ = MLP; the textual space for **T**, **CM** (text) and models with $g = id$; and the visual space for **CM** (vis).

We adapt two classical multimodal word embedding models for sentences (two first items) and use the grounding model from (Kiela, Conneau, et al. 2018):

**Cross-modal Projection (CM)** Inspired by (Lazaridou, N. T. Pham, and Baroni 2015), this baseline learns to project sentences in the visual space using a max-margin loss.

**Sequential (SEQ)** Inspired by (Collell, T. Zhang, and M.-f. Moens 2017), we learn a linear regression model $(W, b)$ to predict the visual representation of an image, from the representation of a matching caption. The grounded word embedding is the concatenation of the original SkipThought vector $\underline{T}_s$ and its predicted ("*imagined*") representation $\underline{I}_s = W\underline{T}_s + b$, which is then projected through a PCA into dimension $d_t$. In both cases, the parameters to be learned, in addition to the sentence encoder, are the cross-modal projections – and the sentence representation is obtained by averaging word vectors.

**GroundSent Model** We re-implement the GroundSent models of (Kiela, Conneau, et al. 2018), obtaining comparable results. The authors propose two objectives to learn a grounded vector: (a) Cap2Img: the cross-modal projections of sentences are pushed towards their respective images via a max-margin ranking loss, and (b) Cap2Cap: a visually equivalent sentence is predicted via a LSTM sentence decoder. The Cap2Both objective is a combination of these two objectives. Once the grounded vectors are learned, they are concatenated with a textual vector (learned via a SkipThought objective) to form the GS-Img, GS-Cap and GS-Both vectors.

We tested variants of our grounding model, all based on **T**: $\mathbf{T} + \mathbf{C}_g$, $\mathbf{T} + \mathbf{P}_g$, $\mathbf{T} + \mathbf{C}_g + \mathbf{P}_g$, where $\mathbf{C}_g$ (resp. $\mathbf{P}_g$) represents the cluster loss $\mathcal{L}_C$ (resp. the perceptual loss $\mathcal{L}_P$). We also consider scenarios where $g$ equals the identity function (i.e., no grounding space), which we note $\mathbf{C}_{id}$, $\mathbf{P}_{id}$, $\mathbf{C}_{id} + \mathbf{P}_{id}$, etc. Finally, we also performed preliminary analysis learning only from the visual modality, i.e. the previous models without adding the textual loss (**T**), that we present next.

#### 5.3.2.3 Preliminary analysis: Study of the grounded space

To probe the learned grounded space, we define structural measures, and report their values on the validation set of MS COCO (5K images, 25K captions):

1. To study perceptual information, we define $\rho_{vis}$, the Pearson correlation

$$\rho(\cos(\underline{g}_s, \underline{g}_{s'}), \cos(\underline{v}_s, \underline{v}_{s'}))$$

between image representation $v$ and their corresponding sentences' similarities. For cluster information, we introduce $C_{intra} = \mathbb{E}_{v_s=v_{s'}}[cos(s, s')]$ where $v_s$ is the image associated with caption $s$, which measures the homogeneity of each cluster, and $C_{inter} = \mathbb{E}_{v_s \neq v_{s'}}[cos(s, s')]$, which measures how well clusters are separated from each other.

2. To study the neighboring structure, we rely on the *mean Nearest Neighbor Overlap* (mNNO) metric, as defined in (Collell and M.-F. Moens 2018b), that indicates the proportion of shared nearest neighbors between image representations and their corresponding captions in their respective spaces – this metric shows how much the local structure is common between text, visual and multimodal representations.

We now validate our hypotheses on the grounded space, using the Cross-Modal Projection baseline (**CM**) and our model scenarios. For fair comparison, metrics for the baseline **CM** are estimated either on the visual or the textual space depending on whether our models rely on the grounded space ($g$) or not ($id$). These results correspond to the rows **CM** (text) and **CM** (vis.).

The results highlight that:

1. Using a grounded space is beneficial; indeed, semantic relatedness and mNNO scores are higher in the lower half of Table 5.4, e.g., $\mathbf{C}_g > \mathbf{C}_{id}$, $\mathbf{P}_g > \mathbf{P}_{id}$ and $\mathbf{C}_g + \mathbf{P}_g > \mathbf{C}_{id} + \mathbf{P}_{id}$;

2. Solely using cluster information leads to the highest $C_{intra}$ and lowest $C_{inter}$, which suggests that $\mathbf{C}_\bullet$ is the most efficient model at separating visually different sentences;

3. Using only perceptual information in $\mathbf{P}_\bullet$ logically leads to highly correlated textual and visual spaces (highest $\rho_{vis}$), but the local neighborhood structure is not well preserved (lowest $C_{intra}$);

4. Our model $\mathbf{C}_\bullet + \mathbf{P}_\bullet$ is better than **CM** at capturing cluster information (higher $C_{intra}$, lower $C_{inter}$) and perceptual information (higher $\rho_{vis}$). This also translates in a higher mNNO measure for $\mathbf{C}_\bullet + \mathbf{P}_\bullet$, leading us to think that the conjunction of both perceptual and cluster information leads to high correlation of modalities, in terms of neighborhood structure. Moreover, this high mNNO score results in better performances for our model $\mathbf{C}_\bullet + \mathbf{P}_\bullet$ in terms of semantic relatedness.

***Influence of concreteness*** To understand in which cases grounding is useful, we compute the average visual concreteness of the STS benchmark, which is divided in three categories (*Captions*, *News*, *Forum*). This is done by using a concreteness dataset built by (Brysbaert, Warriner, and Kuperman 2014) consisting of human ratings of concreteness (between 0 and 5) for 40,000 English words; for a given benchmark, we compute the sum of these scores and average over all words that are in the concreteness dataset. The performance gain $\Delta$ between $\mathbf{C}_g + \mathbf{P}_g$ and $\mathbf{T}$ are observed when the visual concreteness is high: for *Captions* (average concreteness 3.10), the improvement is substantial: ($\Delta = +43$); for benchmarks with a lower concreteness (*News* with 2.61 and *Forum* with 2.39), the improvement is smaller ($\Delta = +12$). Thus, grounding brings useful complementary information, especially for concrete sentences.

***t-SNE visualization*** This finding is also supported by a qualitative experiment showing that grounding groups together similar visual situations. Using sentences from CMPlaces (Castrejon et al. 2016), which describe visual scenes (e.g., *coast*, *shoe-shop*, *plaza*, etc.) and are classified in 205 scene categories, we randomly sample 5 visual scenes and plot in Figure 5.10 the corresponding sentences using t-SNE (Maaten and G. Hinton 2008). We notice that our grounded model is better able to cluster sentences that have a close visual meaning than the text-only model. This is reinforced by the structural measures computed on the five clusters of Figure 5.10: $C_{inter} = 19, C_{intra} = 22$ for $\mathbf{T}$, $C_{inter} = 11, C_{intra} = 27$ for $\mathbf{C}_g + \mathbf{P}_g$. Indeed, $C_{inter}$

(resp. $C_{intra}$), is lower (resp. higher) for the grounded model $\mathbf{C}_g + \mathbf{P}_g$ compared to $\mathbf{T}$, which shows that clusters corresponding to different scenes are more clearly separated (resp. sentences corresponding to a given scene are more packed).

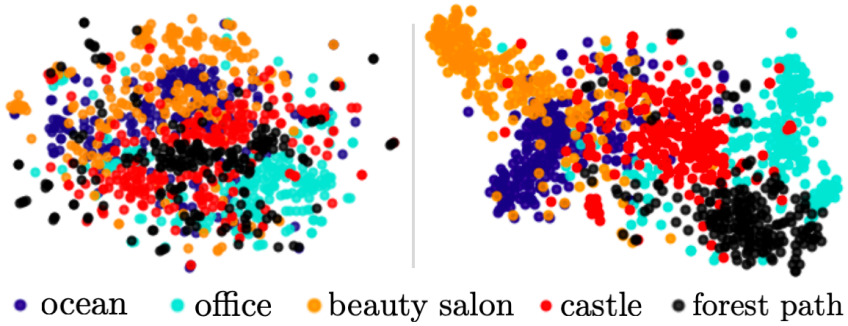



Figure 5.10 – t-SNE visualization on CMPlaces sentences for a set of randomly sampled visual scenes. Left: textual model $\mathbf{T}$. Right: grounded model $\mathbf{C}_g + \mathbf{P}_g$.

***Nearest neighbors search*** Furthermore, we show that concrete knowledge acquired via our grounded model can also be transferred to abstract sentences. To do so, we manually build sentences using words with low concreteness (between 2.5 and 3.5) from the USF dataset (Nelson, McEvoy, and Schreiber 2004). Then, nearest neighbors are retrieved from the set of sentences of Flickr30K Plummer et al. 2015. In this sample, we see that our grounded model is more accurate than the purely textual model to capture visual meaning. The observation that visual information propagates from concrete sentences to abstract ones is analogous to findings made in previous research on word embeddings (Hill and Korhonen 2014a).

***Neighboring structure*** To illustrate the discrepancy on the mNNO metric observed between $\mathbf{C}_g + \mathbf{P}_g$ and $\mathbf{T}$, we select a query image $Q$ in the validation set of MS COCO, along with its corresponding caption $S$; we display, in Figure 5.11, the nearest neighbor of $Q$ in the visual space, noted $N$, and the nearest neighbors of $S$ in the grounded space. With our grounded model, the neighborhood $S$ is mostly made of sentences corresponding to $Q$ or $N$.

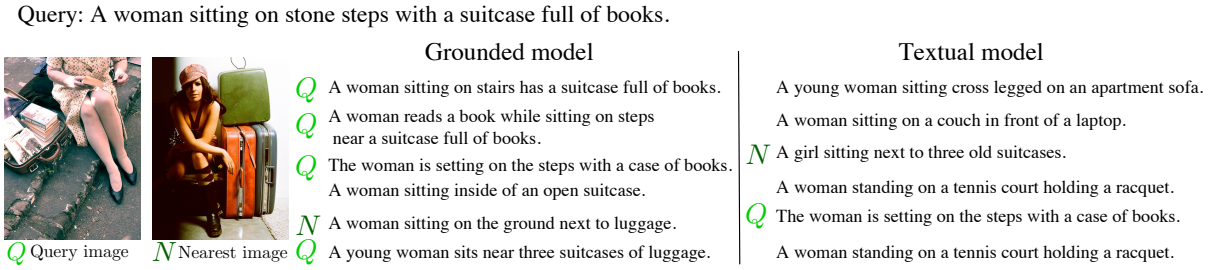



Figure 5.11 – Nearest neighbors of a selected sentence in the validation set of MS COCO, for both grounded and purely textual models. $Q$ is the query image, $N$ is the nearest neighbor of $Q$ in the visual space. Sentences that are caption of $Q$ or $N$ are prefixed with $Q$ or $N$.

#### 5.3.2.4 Experiments

We now focus on extrinsic evaluation of the embeddings. In line with previous works (Hill, Cho, and Korhonen 2016; Kiros et al. 2015), we consider several benchmarks to evaluate the quality of our grounded embeddings using SentEval (Conneau and Kiela 2018).

Table 5.5 reports evaluations of our baselines and scenarios on SentEval (Conneau and Kiela 2018), a classical benchmark used for evaluating sentence embeddings. Before further analysis,

| | Sentiment | | | Objectivity | Similarity | Paraphrase | Entailment | | |
|---|---|---|---|---|---|---|---|---|---|
| | MR | CR | MPQA | SUBJ | SST | MRPC | SNLI | SICK | **AVG** |
| GS-Cap | 72.0* | 76.8* | 85.5* | 90.7* | 76.7* | 72.9/80.6 | 73.7 | **82.9** | 78.4 |
| GS-Img | 74.5* | 79.3* | 87.8* | 90.8* | 80.0* | 73.0/80.3 | 72.2* | 80.9* | 79.8 |
| GS-Both | 72.5* | 75.7* | 85.4* | 90.7* | 76.7* | 72.9/81.3 | 72.2* | 81.4* | 78.4 |
| **T** | 75.9* | 79.2* | 86.7* | 92.0 | 81.8* | 72.2/80.2 | 72.0* | 81.1* | 80.1 |
| **T + CM** | 77.6 | 81.4 | 88.3 | 92.6 | 82.0* | 73.5/81.1 | 73.0 | 81.4* | 81.1 |
| **SEQ** | 76.1* | 79.8* | 86.7* | 92.5 | 81.7* | 70.0*/79.5* | 67.3* | 76.7* | 78.9 |
| $\mathbf{T} + \mathbf{P}_{id}$ | 77.5 | 81.5 | 88.4 | 92.7 | 82.4 | **73.7**/81.3 | 72.4 | 81.1 | 81.2 |
| $\mathbf{T} + \mathbf{P}_{g}$ | **77.8** | **81.8** | 88.1 | **93.0** | **83.5** | 73.3/**81.6** | 72.8 | 82.2 | **81.6** |
| $\mathbf{T} + \mathbf{C}_{id}$ | 77.5 | 81.6 | 88.3 | 92.8 | 82.2 | 72.9/80.5 | 73.1 | 82.3 | 81.3 |
| $\mathbf{T} + \mathbf{C}_{g}$ | 77.3 | 81.5 | **88.6** | 92.8 | 82.6 | 73.6/81.1 | **74.1** | 82.6 | **81.6** |
| $\mathbf{T} + \mathbf{C}_{id} + \mathbf{P}_{id}$ | 77.3 | 81.2 | 88.4 | **93.0** | 82.5 | 73.0/80.6 | 73.5 | 82.1 | 81.4 |
| $\mathbf{T} + \mathbf{C}_{g} + \mathbf{P}_{g}$ | 77.4 | 81.5 | 88.1 | **93.0** | 82.7 | 73.2/80.9 | 73.9 | **82.9** | **81.6** |

Table 5.5 – Extrinsic evaluations with SentEval (Conneau and Kiela 2018). All models represent sentences in a space of dimension $d_t = 2048$ (except for **T**, where the dimension is 1024). 'AVG' stands for the average accuracies reported in the other columns. '†': the model has been re-implemented (we obtained higher scores than the one given in the original papers). '‡': the baseline is an adaptation of the model to the case of sentences. '$*$': significantly differs from the best scenario among our models.

we find that our grounded models systematically outperform the textual baseline **T**, on all benchmarks, which shows the first substantial improvement brought by grounding and visual information in a sentence representation model. Indeed, models GS-Cap, GS-Img and GS-Both from (Kiela, Conneau, et al. 2018), despite improving over $\mathbf{T}_{1024}$, perform worse than the textual model of the same dimension **T** — this is consistent with what they report in their paper.

Our interpretation of the results is the following:

1. our joint approach shows superior performances over the sequential one S, confirming results reported at the word level (section 5.3.1) and more generally of the *strong grounding hypothesis*. Indeed, both sequential models, GS models (Kiela, Conneau, et al. 2018) and **SEQ** (inspired from (Collell, T. Zhang, and M.-f. Moens 2017)) are systematically worse than our grounded models for all benchmarks.

2. Preserving *the structure* of the visual space is more effective than learning cross-modal projections; indeed, all our models outperform $\mathbf{T} + \mathbf{CM}$ on average (column *AVG*).

3. Making use of a grounded space yields slightly improved sentence representations. Indeed, our models that use the grounded space ($g$ = MLP) can take advantage of more expression power provided by the trainable $g$ than models which integrate grounded information directly in the textual space ($g = id$). More particularly, we observe that these models improve the scores for semantic tasks (entailment and paraphrase).

4. Among our model scenarios, $\mathbf{T} + \mathbf{P}_g$ has maximal scores on the most tasks; however, it has lower scores on SNLI and SICK, which are entailment tasks. Models using cluster information $\mathbf{C}_g$ are naturally more suited for these tasks and hence obtain higher results. Finally, the combined model $\mathbf{T} + \mathbf{C}_g + \mathbf{P}_g$ shows a good balance between classification and entailment tasks.

Finally, from an experimental point of view, the importance of the hyper-parameter controlling the importance of the textual loss versus the grounding losses was crucial in obtaining a

good equilibrium of performances for high level tasks like entailment and paraphrase detection and relatively low level like sentiment, objectivity and similarity.

#### 5.3.2.5 Conclusion

In this section, we have exposed a methodology to learn sentence embeddings aiming at preserving the structure of visual and textual spaces to learn grounded sentence representations. The contributions of this model include (1) leveraging both perceptual and cluster information and (2) using an intermediate grounded space enabling to relax the constraints on the textual space. Our approach was the first to report consistent positive results against purely textual baselines on a variety of natural language tasks.

Note that the above described model could be improved by studying more in depth *how* the structure of the visual and textual space should be related. In this work, we used the correlation between intramodal similarities, but more principled and/or empirically validated approaches could be interesting to investigate.

As with grounded word embeddings, possible extensions of this work should include an analysis of what this type of model brings for end-user tasks like summarization, question answering, and more visual tasks, such as visual dialog or question answering.

### 5.3.3 Grounding words in Time

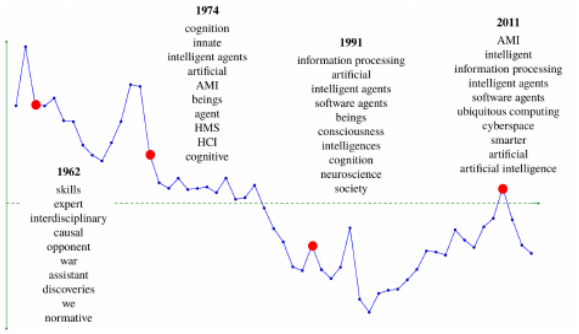
In this section, departing from the previously presented works, we discuss *temporal* grounding of words. More specifically, we try to capture the evolution of language over time, using representation-based models – this was a preliminary work detailed in (Kraljevic et al. 2016). This can be interesting to model the evolution of a word sense, as illustrated in Figure 5.12. A second potential use is for transfer learning: If we want to process texts which belong to a different time period, then using such models can be interesting. This also brings interesting questions in how to design models able to deal with evolving embeddings.

Time is an important factor that must be taken into account in representation models because the objects represented are not static and their distributed representation must adapt to this evolution. Most dynamic representation models consider sequences where the representation of an object evolves in a discrete way, in the form of a state in a vector space. For example, the evolution of the state in a phrase (Collobert et al. 2011) or in a video (N. Srivastava, Mansimov, and Salakhutdinov 2015). These models are not satisfactory because we want to be able to represent any object/entity at an arbitrary time $t$.

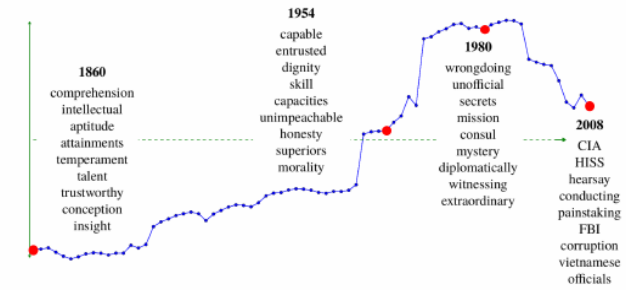
In the case of the text, representation models using continuous time have been proposed to deal with the problems of changing themes. Two types of approaches can be distinguished: the most common is to use a discrete time (Saha and Sindhwani 2012) and to learn the representations for each time interval under constraint of a regularity between the representation of a word between two intervals of time.

The biggest problem with this type of approach is that the choice of temporal granularity is fixed and strongly influences the results. In addition, in the context of learning representations, it would be necessary to store a vector per object and time interval, which is prohibitive. Another approach is that of X. Wang and McCallum (2006) who propose a probabilistic model where latent factors are associated with a theme and a distribution over time. It is this last type of approach that inspired us.

What is specific to our approach is the usage of a representation space for words. The evolution of the sense of words is translated by the word representations *moving* in the semantic space. As an example, we can imagine that the term *deep* and *neural* would begin to get closer, in the representation space, for a period starting around the beginning of the 21st century.




(a) INTELLIGENCE in ACM abstracts (1951–2014)




(b) INTELLIGENCE in U.S. Senate speeches (1858–2009)

Figure 5.12 – Example of the evolution of the term "Intelligence" in two datasets (figure taken from M. Rudolph and Blei 2017): words with similar embeddings are listed for at four points in time (y-axis is a one dimensional projection of the embedding vectors onto "meaning").

#### 5.3.3.1 Time as embedding transformations

In our work, we explored two simple dynamic models; both rely on the fact that the representation of a word $a$ evolves around a central position $\underline{x}_a^{(0)}$, and this evolution is driven by (1) the degree with which a word belongs to a group; and (2) the importance of the group at a given time $t$. More intuitively, we suppose that when a word belongs to a cluster, the cluster modifies the position of the word in the representation space *during a given time period*.

More formally, we use the following equation:

$$\underline{x}_a^{(t)} = \underline{x}_a^{(0)} + \sum_{c \in \mathcal{C}} \rho_{ac} g_c(f_c(t), \underline{x}_a^{(0)}) \tag{5.9}$$

where $\underline{x}_a^{(0)} \in \mathbb{R}^N$ is the central position of the word in the representation space, $\rho_{ac}$ is the degree with which the word $a$ belongs to the group $c$, $f_c$ corresponds to the probability that the group $c$ is active at time $t$, and $g_c$ corresponds to the action of the group $c$ on the representation $\underline{x}_a^{(0)}$ given the influence $f_c(t)$ at the time $t$ of the group $c$ on the representation of the objects belonging to it. We present below two simple instances of this model.

In the first model developed (*the attraction model*), words are attracted towards a cluster center $\underline{x}_c$, and the time influence is modeled with a cluster-specific Gaussian kernel centered on time $t_c$. Formally, this gives

$$\begin{aligned} f_c(t) &= \exp\left(-\tau_c \left\| t - t_c \right\|\right) \\ g_c(u, \underline{x}_a^{(0)}) &= u \times \left(\underline{x}_c - \underline{x}_a^{(0)}\right) \end{aligned}$$

where the dilation parameter $\tau_c$, the central time $t_c$ and the attractor $\underline{x}_c$ are the learned cluster parameters, and the central representation $\underline{x}_a^{(0)}$ is learned of each word $a$ of the vocabulary.

In the case of a word $a$ belonging fully to only one cluster $c$ (i.e. $\rho_{ac} = 1$), we have

$$\underline{x}_a^{(t)} = \underline{x}_a^{(0)} + \exp\left(-\tau_c \left|t - t_c\right|\right)\left(\underline{x}_c - \underline{x}_a^{(0)}\right)$$

and hence the representation of the word is $\underline{x}_c$ when $t = t_c$ and goes back to $\underline{x}_a^{(0)}$ when $t$ is sufficiently different from $t_c$.

Our second model (*the translation model*) assumes that words belonging to a group $c$ are translated by a cluster-specific vector $u_c$ in the semantic space – this vector is learned. Formally, the transformation on the central word embedding is simply defined by its translation vector:

$$g_c(x_a) = u_c \tag{5.10}$$

If we integrate (5.10) into (5.9), that gives us

$$\underline{x}_a^{(t)} = \underline{x}_a^{(0)} + \sum_{c \in \mathcal{C}} \rho_{ac} \exp\left(-\tau_c \left|t - t_c\right|\right) u_c$$

This model is less adapted to the trajectory of words because it only indirectly increases the similarity (in the sense of the scalar product) between two different words, contrary to the attraction model. However, it is a simpler model that can be learned more easily.

We then extend Word2Vec (Mikolov et al. 2013) which proposed a simple criterion to learn a distributed representation of words: the dot product between two representations of words must be even stronger when the two words appear together. Formally, the optimized criterion is the likelihood of the observed data. Using the parametric form of the equation (5.9) instead of a fixed representation, it is possible to use the same criterion as that proposed in (Mikolov et al. 2013) and to learn the representation of words as well as groups, i.e. by optimizing

$$\mathbb{E}_{(w,c^+,t)}\left[\sigma(\underline{x}_w^{(t)} \cdot \underline{y}_{c^+}^{(t)}) + \sum_{\substack{k=1 \\ c_k^- \sim \mathcal{U}}}^{N} \sigma(-\underline{x}_w^{(t)} \cdot \underline{y}_{c^-}^{(t)})\right]$$

Comparing with equation (2.3), we can see that the only change is that the representation of a word is varying with the time $t$, using the formulas given above.

Such models are capable of expressing complex evolutions in a representation space by using a small number of parameters with respect to a model where at each time step corresponds a new series of representations.

#### 5.3.3.2 Experiments

To validate the approach, we choose to evaluate the approach as an unsupervised time-dependent clustering. We used a collection of 12 million micro-blogs (J. Yang and Leskovec 2011), a subset of the June 2009 Twitter micro-blogs, which we downloaded using the twitter API. These micro-blogs can be written in any language, and deal with any topic. We first cleaned up the dataset as follows. We removed micro-blogs with less than 8 words, which reduced the set to 10 million documents. For the remaining ones, we removed numbers, punctuation, hyperlinks, user links (*@user*) and tags (`#tag`). We kept the association between micro-blogs and tags in order to evaluate the groups automatically found by the different algorithms, assuming that a tag corresponds to a given group.

The micro-blog groups were computed with k-means (baseline), and using our model to derive the estimation of a micro-blog belonging to one of the groups. Using the naive Bayes hypothesis, i.e. in assuming that the occurrence of a word and the time are independent if the group is known, and assuming that each group has the same probability, we can write:

$$p(c|d) = \frac{p(c)p(t_d|c)\prod_{a \in d} p(c|a)p(a)/p(c)}{p(d)} \propto f_c(t_d)\prod_{a \in d}\rho_{ac}$$

where $d$ is a document written at a time $t_d$, and where the product is defined on the set of words that appear in the document $d$. We then assign the document $d$ to the group $c_d$ which has the highest probability $p(c|d)$.

Preliminary experiments with the model (Y. Wang, Agichtein, and Benzi 2012) showed that it was difficult to obtain satisfactory results with it – given its numerical complexity. The corpus on which it was evaluated is much smaller; our first attempts to adapt the algorithm to a large collection of tweets were unsuccessful. For the k-means, we used a representation of each micro-blog in the thematic representation space given by Word2Vec, by summing the vectors representing each word contained in the micro-blog.

The reference groups were formed under the assumption that a group is defined by the micro-blogs containing a given label. Given the large number of tags in this corpus, we selected small subsets (about 100 tags), with the following methodology. First of all, the labels present in less than 500 micro-blogs were removed. Then, we generated the following sets:

**Random** A randomly chosen set (uniform probability) of labels;

**Top** The most frequent labels;

**H.C.** A set of the most frequent hand-picked labels – trying to favor those corresponding to an event covering a short period of time (a few days);

**G.I. 0.80, 0.85 and 0.90** Labels for which the Gini index is above a certain threshold. The Gini index was calculated based on the frequency of occurrence of a label for each day. The more uniformly a label appears over the days, the lower its Gini index, and the more skewed its distribution, the higher its Gini index. We have empirically chosen thresholds of 0.8, 0.85 and 0.90 which correspond to very "uneven" distributions (a few days).

The first two sets (Random and Top) are more likely to favor the K-means algorithm, because they correspond to themes that have been treated throughout the period corresponding to the month of June, contrary to the last ones (HC, GI 0.80, 0.85 and 0.90) which correspond more to themes extending over a few days, and which *a priori* should be better detected by our model.

To measure the quality of the groups found, we used the V-measure (Rosenberg and Hirschberg 2007), a measure that compares the found and reference groups by calculating the harmonic average of two measurements with values in $[0, 1]$ :

1. Homogeneity, which is related to the conditional entropy $H(C|K)$ of a $C$ group given the $K$ classes, is all the greater the more the group found is "pure", i.e. it contains only documents with the same labels;

2. Completeness, which is related to the conditional entropy $H(K|C)$, is all the greater as the documents associated with the same label are in the same found group.

Like accuracy and recall, increasing homogeneity tends to decrease completeness, and vice versa. The V-measure is then defined as the harmonic mean of homogeneity and completeness.

In table 5.6, we give the V-measure for the different selected label sets and models. As expected, our models behave better when the selected labels correspond (explicitly or implicitly) to events that cover few days. The difference with the k-means algorithm is very important for these four sets of labels, which means that the groups found, even if they are only a by-product of the word trajectory model, are able to satisfactorily capture the events as well as the words that describe them.

At the level of our models, the attraction model obtains results that are better (0.03 to 0.05 points of difference), except in the case of labels chosen randomly or according to their frequency. This corresponds to our hypothesis – the similarity of words in the thematic space is better taken into account when the words all converge towards the center of the group; what was less obvious was that this would influence the quality of the groups found. Note that this difference is reversed when the sets of labels are no longer those favoring temporal patterns.

| Labels \ Model | Attraction | Translation | K-mean |
|---|---|---|---|
| G.I. > 0.80 | 0.48 | 0.45 | 0.43 |
| G.I. > 0.85 | 0.52 | 0.48 | 0.44 |
| G.I. > 0.90 | 0.56 | 0.51 | 0.47 |
| H.C. | 0.47 | 0.44 | 0.43 |
| Top 100 | 0.16 | 0.18 | 0.30 |
| Random 100 | 0.15 | 0.18 | 0.29 |

Table 5.6 – V-measurement for different models and label sets - the best result for each label set is shown with a green background

#### 5.3.3.3 Conclusion

The results obtained on micro-blogs (Twitter) show that this approach makes it possible to automatically find temporal cluster of tweets. This type of model can be useful for modeling the evolution of various objects in a representation space, and therefore has great potential for all models of representations where objects can change their representation as a function of time.

More generally, we were interested by the problem of defining parametric trajectories in the semantic space. I believe that there is some unexploited potential in defining potential equations for the movement of words – and more generally of documents – in semantic spaces, maybe inspired by works in modeling physical phenomena such as which mix neural networks and differential equations (Bezenac, Pajot, and Gallinari 2018).

## 5.4 Conclusion

This chapter presented works dealing with *grounding* textual representations with modal information – mostly visual, but also temporal as in the last section. In the conclusion, we focus on the latter since it was the main part of our work.

We have first shown that even for representations inferred from purely textual information, there was still a link between this representation and the visual or common sense properties of the objects, even if the structure of the space is still very different. This implied that some more information could be brought to the textual representations by grounding them.

We then discussed three works. The two first respectively dealt with grounding words and sentences with visual information. The main conclusion that we draw from this line of research is that

1. The geometries of the visual and textual spaces are very different, and thus aligning them directly might not be the best way to do so. This motivated the use of a grounding space to align both spaces without requiring them to have the same structure. Even though this is only a step in this direction, we believe that it is important to go further in terms of *how* to measure the alignment beyond using the Pearson correlation between the similarities in both textual and visual spaces.

2. The visual context is useful in representing text since it allows to build textual representations that capture the visual context in which an entity appears. When grounding words, we did not exploit any grounded space, which might explain why we could not fully exploit this information.

3. The *strong grounding hypothesis* did hold for word and sentence embeddings; it is difficult still to judge whether this is due to the quite naive ways to fuse textual and visual modalities (in the literature in general), i.e. by concatenating and projecting, or if this hypothesis really holds.

In both cases, improvements were obtained in terms of performances on probing tasks such as word similarity or affordance. However, when applied to higher level tasks (e.g. retrieval, summarization, etc.), results tend to be not so different from using models and representations trained on text only – especially with new models such as the Transformer-based ones.

There are two potential causes for this problem. The first is that common sense knowledge, as captured by grounding models, is not really useful in most cases – because datasets are biased towards tasks that do not require grounding but rather direct natural language inference. The second is that current grounding models are too naive, and do not really integrate useful common sense information besides basic properties such as color, shape or size – which, again, are not so useful for high level tasks.

Addressing the former issue is quite straightforward, by designing new datasets and tasks (such as GuessWhat). The latter issue is more serious, and might imply more complex and/or principled models – and maybe working more on *grounded* image representation – more precisely, this would imply biasing the CNN processing the image to be grounded on text, allowing it to recognize new objects.

Even today, it is difficult to judge whether this is because no useful information can be gathered from modalities for those tasks, but this information could be useful in tasks relying on more common sense knowledge. More work is needed in investigating the real difference between grounded and not-grounded representations if one wants to draw any definitive conclusion.

**Publications and supervision**

- The work carried out is a collaboration with Laure Soulier (MLIA team, LIP6), as part of the European project MUSTER, through the supervision of two doctoral students E. Zablocki (October 2016 – October 2019) and P. Bordes (started in 2017 – defense expected in September, 2020).
- I have supervised a master student (Z. Kraljevic) on the temporal grounding work
- Z. Kraljevic et al. (Mar. 2016). "Représentation Temporelle Des Mots : Application Au Clustering de Micro-Blogs." In: *Conférence En Recherche d'Infomations et Applications*, pp. 531–544
- É. Zablocki et al. (Feb. 2018). "Learning Multi-Modal Word Representation Grounded in Visual Context." en. In: *Proceedings of the Association for the Advancement of Artificial Intelligence*
- E. Zablocki et al. (May 2019a). "Context-Aware Zero-Shot Learning for Object Recognition." en. In: *International Conference on Machine Learning*, pp. 7292–7303
- E. Zablocki et al. (2019b). "Incorporating Visual Semantics into Sentence Representations within a Grounded Space." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

# Part III

# Other contributions and conclusion

# Chapter 6

# Other contributions

In this chapter, I list the other contributions in terms of developed software, and around other Information Access themes I worked on after my PhD thesis, but which did not fit in the works presented in the previous chapters.

## 6.1 Software

The main software contributions are listed here

- Experimaestro
  https://experimaestro.github.io/experimaestro-python

  Experimaestro is a computer science experiment manager whose goals are:

  - To decompose experiments into a set of parameterizable tasks
  - Schedule tasks and handle dependencies between tasks
  - Avoids to re-run the same task two times by computing unique task IDs depending on the parameters
  - Handle experimental parameters through tags

- Datamaestro
  https://experimaestro.github.io/datamaestro

  This projects aims at grouping utilities to deal with the numerous and heterogeneous datasets present on the Web. It aims at being

  1. A reference for available resources, listing datasets
  2. A tool to automatically download (when freely available) and process resources
  3. Integration with the Experimaestro experiment manager.

  Each dataset is uniquely identified by a qualified name such as `com.lecun.mnist`, which is usually the inverted path to the domain name of the website associated with the dataset. The main repository only deals with very generic processing (downloading, basic pre-processing and data types). Plugins can then be registered that provide access to domain specific datasets (I have developed three so far, for text, image and generic machine learning datasets).

- Kernel Quantum Probabilities (KQP)
  https://github.com/bpiwowar/kqp

  The Kernel Quantum Probability library (KQP) aims to provide tools to effectively compute "quantum probabilities", that is to compute a representation of densities, events and to update the densities when events are observed (conditioning). It also provides access to generalization of standard probabilistic measure like entropy and divergence.

**Publications**

- Experimaestro and datamaestro have been publicized through the following publication: B. Piwowarski (July 2020). "Experimaestro and Datamaestro:Experiment and Dataset Managers (for IR)." in: *ACM SIGIR 2020*. Xian, China. DOI: 10.1145/3397271.3401410
- KQP has been described on arXiv, and presented during two ECIR tutorials B. Piwowarski (2012). *The Kernel Quantum Probabilities (KQP) Library*. Tech. rep. 1203.6005v2

## 6.2 Past works

In this section, I give a high level overviews of other research themes I have worked on since after my PhD.

### 6.2.1 Information Retrieval Metrics

The structured information retrieval (i.e. where IR models can retrieve subparts of documents, such as chapters, sections or paragraphs) paradigm implies to change the way systems are evaluated since they can return elements within a document instead of a full document. In INEX (Kazai 2009), a new assessment scale has been proposed along with new precision/recall metrics. We proposed several metrics, the latest being the most expressive (generalization of precision-recall) and simple to compute.

**Outcomes**

- B. Piwowarski, P. Gallinari, and G. Dupret (2007). "An Extension of Precision-Recall with User Modelling (PRUM): Application to XML Retrieval." In: *ACM Transactions On Information Systems* 25.1. DOI: 10.1145/1198296.1198297
- B. Piwowarski and G. Dupret (2006). "Evaluation in (XML) Information Retrieval: Expected Precision-Recall with User Modelling (EPRUM)." in: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. Ed. by E. N. Efthimiadis et al. Seattle, Washington, USA: ACM, pp. 260–267. DOI: 10.1145/1148170.1148218

### 6.2.2 User modeling

We worked on a model that is able to distinguish the attractiveness (how likely a user is going to click on it) of paper and the position effect (how likely is it that the user will even consider this document, i.e. look at the snippet). This work is described in a SIGIR paper (Dupret and Piwowarski 2008).

We developed a predictive model for navigational queries (Piwowarski and Zaragoza 2007), where the aim is to predict the document the user will click on along with the confidence we have for this prediction; the confidence is important if one wants to automatically redirect (for example) the user to the page he wanted to browse to when typing their query.

My last project aimed at modeling a whole search session (from the first query string to the last click, including reformulations) using layered Bayesian networks (Piwowarski, Dupret, and Jones 2009). The main potential applications of this work are the estimation of user satisfaction and of the relevance of documents (based on a single interaction between the user and the search engine).

**Outcomes**

- G. Dupret and B. Piwowarski (2008). "A user browsing model to predict search engine click data from past observations." In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. Ed. by S.-H. Myaeng et al. Singapore: ACM
- B. Piwowarski, G. Dupret, and R. Jones (Feb. 2009). "Mining User Web Search Activity with Layered Bayesian Networks or How to Capture a Click in its Context." In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. Ed. by R. A. Baeza-Yates et al. Barcelona, Spain: ACM, pp. 162–171. DOI: 10.1145/1498759.1498823 – lead to a patent[1]
- B. Piwowarski and H. Zaragoza (2007). "Predictive User Click Models Based on Click-through History." In: *Proceedings of the 16th ACM International Conference on Information and Knowledge Management*. Lisbon, Portugal: ACM, pp. 175–182 – lead to a patent[2]

### 6.2.3 Recommendation

I supervised the second part of the PhD thesis of Y. Moshfeghi which was about emotions and information retrieval. More specifically, we designed models based on Latent Dirichlet Allocation – LDA (Blei, A. Ng, and M. Jordan 2003) – to include various emotion-related signals (e.g. emotion expressed in the plot summary for a movie) to improve the quality of recommendation.

**Outcomes**

- Y. Moshfeghi, D. Agarwal, et al. (2009). "Movie Recommender: Semantically Enriched Unified Relevance Model for Rating Prediction in Collaborative Filtering Lecture Notes in Computer Science." In: *Proceedings of the 31th European Conference on Information Retrieval Conference*. Ed. by M. Boughanem et al. Toulouse, France: Springer, pp. 54–65. DOI: 10.1007/978-3-642-00958-7_8
- Y. Moshfeghi, B. Piwowarski, and J. Jose (2011). "Handling Data Sparsity in Collaborative Filtering using Emotion and Semantic Based Features." In: *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM SIGIR Conference on Research and Development in Information Retrieval. ZSCC: 0000084. ACM

## 6.3 Ongoing research – representation learning

Finally, I describe shortly the works I am actually conducting – and which are more linked with the works presented at length in this manuscript.

### 6.3.1 Weakly Supervised Information Extraction

A first theme is the development of information extraction models with (almost) no supervision. The most successful models have been using generative ones – where the generated text is conditioned by a (latent) relationship (Yao, Haghighi, et al. 2011; Yao, Riedel, and McCallum

[1] B. Piwowarski and G. Dupret (2009). "System and Method for Deducing User Interaction Patterns Based on Limited Activities."

[2] B. Piwowarski and H. Zaragoza (2009). "System and Method for Creating and Applying Predictive User Click Models to Predict a Target Page Associated with a Search Query."

2012). However, these models are limited in their expressiveness. More importantly, depending on the set of features, they might focus on features not related to the relation extraction task.

In the supervised setting, neural network models have demonstrated substantial improvement over approaches using hand-crafted features. In particular, piecewise convolutional neural networks (PCNN, D. Zeng et al. 2015) are now widely used as a basis for other improvements, such as the instance-level selective attention mechanism of (Y. Lin et al. 2016) which follows the multi-instance multi-label framework (Hoffmann et al. 2011; Surdeanu et al. 2012) . The recent NN approaches however need large amount of data to achieve good performances.

We posit that discriminative approaches can help in going further in expressiveness, especially considering recent results with neural network models. To the best of our knowledge, the only discriminative approach to unsupervised relation extraction is the variational auto-encoder approach (VAE) proposed by Marcheggiani and Titov (2016): the encoder extracts the semantic relation from hand-crafted features of the sentence (related to those of Rel-LDA), while the decoder tries to predict one of the two entities given the relation and the other entity, using a general triplet scoring function.

The model of Marcheggiani and Titov (2016) is however very unstable, and our first work has been to design unsupervised losses that make possible the learning of powerful discriminative models. In particular, we proposed two losses (named RelDist) to effectively train expressive relation extraction models by enforcing the distribution over relations to be uniform. We were able to successfully train a deep neural network classifier that performed well in a supervised setting. We demonstrated the effectiveness of our RelDist losses on three datasets and showcased its effect on cluster purity.

Future work will investigate more complex and recent neural network models such as Self-Attention Networks (Devlin et al. 2018), and work on new losses able to leverage signals even weaker than those that we were relying on so far.

**Outcomes** This work is conducted through the supervision of the PhD thesis of E. Simon (with P. Gallinari and V. Guigue) that started in October 2017 (expected defense in mid 2021):

- E. Simon, V. Guigue, and B. Piwowarski (July 2019). "Unsupervised Information Extraction: Regularizing Discriminative Approaches with Relation Distribution Losses." In: *Proceedings of ACL 2019*. Firenze, Italia

### 6.3.2 Summarization

Automatic summaries can be abstractive or extractive (Moreno et al. 2017). The extractive approach works by returning the most important sentences of the original text. Most of the methods are unsupervised and based on a comparison between the extracted sentences and all the sentences of the document(s) to summarize. However, they are a lot of limits, as the consideration of co-references for example (Genest 2013; Verma and D. Lee 2017).

On the author hand, the abstractive approaches generate a summary of the document(s) at hand. This area of research has until recently been focused on a post-processing step of extractive approaches – where coreference resolution is handled, and where sentences are simplified to remove non-essential content (Woodsend and Lapata 2012).

Since then, the generative capabilities of neural networks have been quickly adopted as a reference method for abstractive summarization (Q. Zhou et al. 2017). The models used are *Seq2Seq* models (input sequence, output sequence) based on recurrent neural networks (RNN). While these models excel at generalizing abstract concepts, it is more difficult for them to deal with facts. Faced with this difficulty, memory neural networks (Weston, Chopra, and Bordes 2015) can be used. These algorithms are today the state of art for the automatic generation of abstracts (See, P. J. Liu, and C. D. Manning 2017; Vaswani et al. 2017).

The question of automatic summary remains open despite the definite progress that has been made in this area. No model has yet achieved the quality of a human summary, i.e. able to transcribe the main information of the original text (*semantic* dimension) expressed with a correct syntax (*syntactic* dimension).

Finally, to train models, abstract approaches require the development of very important datasets (see New York Times for example) in the form of a set of documents and their associated summaries. In addition to being laborious to create and therefore not very adaptable to multilingualism, the latter are too constrained because they present only one abstract for a text while many other summaries are given while a large number of other summaries are acceptable. In addition, the evaluation of summaries is very complex, relying on golden summaries and far from perfect metrics.

As those metrics are also the target of recently-proposed models based on reinforcement learning techniques, it becomes necessary to develop approaches that do not rely on those imperfect metrics, and ideally, be able to train with no supervision – in particular for multilingualism or personalized summarization. Based on the idea of adversarial approaches, the overall goal of our models is to optimize two contradictory objectives:

1. Maximize the syntactic quality of the summary
2. Maximize the semantic content of the summary with respect to the document to summarize

The work conducted so far has explored the problem of generating a meaningful reward since current metrics such as ROUGE are only weakly correlated with human metrics and necessitate human annotated data. In (Scialom, Lamprier, et al. 2019), we proposed to predict the true performance of a model (as evaluated by a human) given indicators that necessitate or not human annotated data. The former is shown to improve the quality of the question generation (we choose this generative task since it is simpler than the summarization one). The latter case is particularly interesting since it opens the door for developing models that can learn from unlabelled data (or at least, pre-learned models than can be fine-tuned).

In our second contribution, we started to explore whether Generative Adversarial Networks (GANs) could be used for summarization. As a first step, we trained discriminators that can predict whether a generated summary is human-written or not. Instead of using them as a training signal (as this would be done with GANs), we used it to bias the generative process: at each step of the beam decoding process, instead of using the joint probability, we used the score provided by the discriminator, showing that it improved substantially the quality of the generated summaries.

Following this work, we are currently investigating how GANs can be trained based on these discriminators – many attempts have not produced the expected outcomes in adapting GANs from image to text (Rekabdar, Mousas, and Gupta 2019; L. Yu et al. 2016), mostly because of the very sparse rewards available in the textual setting (because of the discrete nature of text).

**Outcomes** This work is conducted through the supervision of the PhD thesis (CIFRE, with Recital) of Thomas Scialom (with J. Staiano from Recital, and S. Lamprier and P. Gallinari from LIP6), and has led to the following publications:

- T. Scialom, S. Lamprier, et al. (2019). "Answers Unite! Unsupervised Metrics for Reinforced Summarization Models." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. DOI: 10/ggchbh
- (accepted in ICML 2020) T. Scialom, P.-A. Dray, et al. (Feb. 2020). "Discriminative Adversarial Search for Abstractive Summarization." In: *arXiv:2002.10375 [cs]*. ZSCC: NoCitationData[s0] arXiv: 2002.10375

# Chapter 7

# Conclusion

## 7.1 Contributions

In this manuscript, I have presented works on the problem of learning semantic continuous representations of entities – focusing mostly on the problem of representing text. I summarize below the main contributions and outline the potential research directions related to the presented works.

### 7.1.1 Probabilistic Representation spaces

We first tackled the problem of choosing a representation space that allows working with specific properties such as uncertainty through the use of probabilistic embeddings, and more precisely using either quantum or Gaussian embeddings.

Quantum embeddings allowed us to define two views of the same information object, and were particularly well adapted for tasks in which one needs to compute the extent with which a text covers different topics, e.g. in summarization. This is because it is easy to compute the representation of a complex mixture of distributions, contrarily to most continuous distributions. The downside of such an approach its it lack of flexibility, and its numerical complexity.

This motivated the use of Gaussian embeddings for further works. They were particularly well adapted to graphs where many nodes should have an uncertain representation, because they either are at the frontier of two node groups (e.g. neighbors have different properties) or when not enough information is known on them. This allowed us to tackle tasks that have those properties, namely graph node classification, recommendation and time series prediction (with inter-related time series).

**Potential Research Directions** The question of how to represent uncertainty is a key to improve information access techniques, e.g. by modeling a potentially evolving user information need or to represent how much information is known about a specific item in a recommender system. The positive results obtained in specific cases show that this information can benefit from being directly modeled – as opposed of being implicitly captured, for instance by using some specific regions or directions of the representation space to represent uncertainty. A potential direction would be to investigate more probability distribution (families) and study their properties.

However, representing this uncertainty comes at a cost, and more precisely requires an explicit modeling from data science experts. Even though there is a renewal of probabilistic approaches in deep neural networks (Schulman et al. 2015), pushed by the integration of reinforcement learning and stochastic optimization techniques into more and more models, as for instance in summarization (Y.-C. Chen and Bansal 2018), the problem of the definition of how the uncertainty should be represented cannot be solved automatically. In many applications, this

is satisfying enough, but a potential research direction would be to investigate whether, through specific losses or interactions between representations, one can integrate some uncertainty in probability-agnostic models.

### 7.1.2 Grounding textual embeddings

The second problem that was exposed is the question of integrating information into textual representation *through grounding.* We started by analyzing the visual information that can be extracted from purely textual embeddings through a zero-short recognition task, showing that only a part of the information could be transferred, but that the structure of both space is too different to allow recognizing objects on which the image CNN was not trained.

This motivated further the need of grounding textual embeddings. We have shown through a word embeddings task that using the visual context of a word/entity allowed to improve the quality of the embeddings on various intrinsic tasks. We have further shown that to align the visual and textual representation spaces, using a grounding space (i.e. the "visual" component of a textual embedding) and a specific alignment loss (requiring that similarities in the grounded space be correlated with the similarities in the visual space) were key to obtaining good quality grounded embeddings, that preserved or improved the performance of extrinsic tasks where visual information might not be so important, and improved for those where it is more.

The main conclusion is that grounding can improve the quality of word or sentence embeddings, by integrating information from visual (and potentially other) modalities; the *strong grounding hypothesis,* whereby representations from various modalities should interact to get a better *grounded* representation, seems to be confirmed in our experiments.

**Potential Research Directions** Since these works were published, a huge gap in performance was observed with Self-Attention Architectures (section 2.5). It would be necessary to adapt our grounding methodology to those contextualized representation models. Note that some works mixing the visual and textual modalities with SAN architectures have already been conducted, such as VL-BERT (Su et al. 2019) using a pre-training task where masked words can be predicted either from their textual context or from visual cues.

Another potential research direction would be to leverage the probabilistic embeddings proposed in chapter 4, since this would allow to cater for the fact that a sentence can be visually ambiguous by using the variance information around the projected point in the visual space.

More fundamentally, the study of the discrepancy between the visual and textual modalities has only been very imperfectly tackled. The use of a grounding space is probably a key to a correct matching of both spaces (since each modality should keep its specific properties), if ones succeeds in designing the right alignment criterion. In the sentence embeddings work, the correlation between similarities might be too simplistic since it does not possess some necessary conditions a matching operator should, such as the fact that two images can be described by two different sentences.

Finally, the pre-training proposed in our work, and more recently in (Su et al. 2019) relying on SAN architecture, is probably one step towards the goal of leveraging the grounding process to build more efficient information access tools. The effect of grounding for such high-level tasks is understudied, mostly due to the fact that, similarly to many techniques designed to enhance the representation of documents in IR, the impact on performance is contrasted depending on the precise information access need: the benefits or drawbacks of grounding are not yet mastered.

## 7.2 Research Perspectives

In this section, after a quick restatement of the evolution of text representations, I argue for more structured representations, which I deem necessary for tackling complex information access

tasks such as discussion based retrieval (i.e. retrieving documents through an interaction with the user).

### 7.2.1 The evolution of representation

In the introduction, I have exposed how text representations evolved from model-based and/or hand-crafted features. For a long time, many models have attempted to leverage more sophisticated textual representations. For example taking into account word order is difficult, and models based on term proximity behave better (Zhao and Yun 2009). Another example is the use of NLP, such as part-of-speech and disambiguation tools that have never been an important feature for IR systems – mainly because the added complexity and the errors of the model do not help the retrieval task.

This began to change with the fast adoption of machine learning techniques and of large datasets (T.-Y. Liu 2011), that allowed the development of many features loosely correlated with relevance – or only correlated in some specific cases that can be determined by the machine learning algorithm.

This search for better representations – which in turn can produce more measures correlated useful for the task to solve – evolved into learning representations from (un)supervised data. Within this recent evolution, we have seen techniques to compute meaningful word embeddings, coupled with neural network architectures like convolution or recurrent neural networks.

In the last few years, neural models have been applied to a multitude of task for access to information: question-answering (W. Wang et al. 2017), information extraction (Y. Y. Huang and W. Y. Wang 2017; C. N. d. Santos, B. Xiang, and B. Zhou 2015; Z. Zhang 2004), automatic summarization (See, P. J. Liu, and C. D. Manning 2017), automated translation (Bradbury and Socher 2017; Britz et al. 2017; Conneau, Lample, et al. 2017; Delbrouck, Dupont, and Seddati 2017), document retrieval (Mitra and Craswell 2017; Mitra, Diaz, and Craswell 2017; Nalisnick et al. 2016) or question answering (X. Qiu and X. Huang 2015) .

Lastly, self-attention-based architectures (SANs, aka Transformers, Vaswani et al. 2017) have swept away almost all other neural-based techniques on any task related to language. These successes were drawn by the capacity of transformers to handle long-term relationship, and made possible by the huge progress in computing power. They also brought two major innovations:

1. They do not try to summarize information, but rather to contextualize it – and this is a key to their success in e.g. Information Retrieval where other neural architectures fell short of working.

2. They rely on a residual approach, whereby a term representation is modified by its context – this was already leveraged in residual CNNs for image recognition (R. Srivastava, Greff, and Schmidhuber 2015) but only partially in NLP and information access with LSTMs units (Hochreiter and Schmidhuber 1997).

### 7.2.2 Limits of current approaches

While the success of representation learning, and of SANs more particularly, is undeniable, these models have some drawbacks.

First, these models are computationally intensive, and only large organizations have the computational resources necessary to train them from scratch. Even if this is less true with recent works (see section 2.5), this limits the processing power to small sized documents.

Second, these models have probably more parameters than needed. Indeed, Kovaleva et al. (2019) show that these model are over-parameterized. More disturbingly, they show self-attention is in many cases not needed, opening the door for models that might rely on very sparse long-range attention.

Third, a particularly strong argument against SANs – and it is true to some extent for other (neural) models – is that their success is partly due to the fact that they exploit better artifacts of test collections (McCoy, Pavlick, and Linzen 2019; Niven and Kao 2019). These limits are underlined in a recent paper from Bender and Koller (2020) where the authors argue that *a "*system trained *only on form* has a priori no way to learn meaning". Said otherwise, it prevents NLP-based models to reason over text, past the reproduction of training patterns data. The authors take the example of manipulating a new concept, e.g. the "coconut catapult", that would be hard or impossible if the neural network has not seen such patterns in the training data.

There is a need to go from systems that are able to react appropriately to a given input to systems able to reason with text, and this, for a variety of reasons. First, this allows to go beyond the current state-of-the-art for many complex tasks, such as e.g. summarization, translation and interactive retrieval. Indeed, to faithfully capture and translate the meaning of a text, being able to infer knowledge from text and its context (the context including the world knowledge) is important so that systems are able to adapt to newly expressed concepts and/or complex semantics. Second, this is a step towards more intelligent systems – and models able to reason over language would probably be also systems that can react more appropriately in unforeseen conditions.

### 7.2.3 Towards structured representations

To tackle such a challenge, one possible direction is to develop models able to work with structured representations where the different content units of a document are explicitly represented and linked together. We give here a very quick overview of two related but different existing approaches to bring structure into language representation, which we name the *linguistic* and the *psycholinguistic* views.

#### 7.2.3.1 The linguistic view – structured *logical* representations

The works of Chomsky in the 1960s (summarized in Chomsky 1980) have stemmed an interest for building formal grammars, whereby a sentence can be fully analyzed by applying rules such as "S $\rightarrow$ NP VP", i.e. a sentence is composed of a noun and verbal phrase. These rules can in turn be interpreted as a parse tree which bears some semantics by relating sub-sequences together. Such approaches led to more powerful theories of meaning such as the Discourse Representation Theory (DRT, see Kamp and Reyle 1993), summarized in (Bos et al. 2017), where the analysis can be conducted over a full text (and not a single sentence).

A lot of work has been invested in trying to design models able to construct structured representations from text, and more specifically logical representations of text. In particular, the work on combinatorial categorial grammars is a recent example of the development of such approaches (Stanojević and Steedman 2020; Steedman and Baldridge 2005). Logical approaches to language representations have however problems dealing with the variability and ambiguity of language; even approaches based Markov logic still rely on a symbolic representation of knowledge (Kok and Domingos 2008), which limits their robustness and expressiveness.

#### 7.2.3.2 The psycholinguistic view – structured *distributed* representations

An alternative view is found in psycholinguistics. The most well known models (Broek et al. 1999; Kintsch 1998) propose to model the reading as a construction of a valued graph where "pseudo-predicates" are linked. The graph obtained at the end of the reading process corresponds to a faithful representation of the analyzed text. These models were only applied to toy data, and even though they seemed to reproduce laboratory observations, they were not readily applicable to real world data.

These models are however very close to the efforts conducted in neural approaches since in both cases there is no formal logic associated with the representation, and in both case the representation is distributed. It has been argued recently by Besold et al. (2017) that distributed representations are the appropriate representational language, since they provide the necessary robustness and effectiveness for robust intelligent systems.

In machine learning, this corresponds to works dealing with graphs of distributed representations. Works on graph-based neural approaches were proposed in (Gori et al. 2009), and then followed up by worked based on memory-based neural networks (Graves, Wayne, and Danihelka 2014; Weston, Chopra, and Bordes 2015). Note that Self-Attention Networks are in direct line with these works – here the memory is the text itself, i.e. each text unit bears a part of the semantics of the text. However, *no structure* is present explicitly, even if it can be recovered (to some extent) for syntax trees (Hewitt and C. D. Manning 2019).

The initial works (Graves, Wayne, and Danihelka 2014; Weston, Chopra, and Bordes 2015) were quickly exploited for complex tasks such as answers to questions (A. Kumar et al. 2015; Sukhbaatar et al. 2015), then expanded with structures such as piles (Joulin and Mikolov 2015), hierarchical memories (W. Zhang, Y. Yu, and B. Zhou 2015) and finally graph memories (Y. Li, Tarlow, et al. 2015; D. D. Johnson 2017). This latest work has shown that it is possible to learn models capable of making decisions such as adding a node, or a link between nodes.

Although they have a strong expressive power, structured representation-based models are hard to train, which explains partly the lack of works pursuing this direction: learning to construct graphs of individual representations is hard, and needs explicit supervision (Y. Li, Tarlow, et al. 2015; D. D. Johnson 2017), which is very costly.

### 7.2.4 Structured Representations

We have seen in the previous section that there are two competing approaches to structured representations of language.

On the one hand, structured *logical* representations are potentially very powerful, but are limited by the fact that texts are noisy and hard to process, and also because a logical representation is not adapted to all the type of texts (especially for ambiguous or unclear ones). On the other hand, structured *distributed* representations are more flexible, allowing to represent a variety of types of texts, but are hard to train since it is impossible to clearly label the sentences.

Because of their flexibility and their link to psycholinguistics who describe how a human processes a text, I strongly believe that this second type of approach are the most promising. This motivates the design of unsupervised means to learn structured *distributed* representations, motivated by the development of machine learning models based on invariants and hypotheses about the data.

This line of thought corresponds to the shift brought by representation-based machine learning. While machine learning has focused on designing by hand models closer to reality as a way to leverage human knowledge, neural approaches have focused on capturing the invariants of the problems (e.g. the translation equivariance in images or texts) and on designing strong hypothesis on the nature of the task. For instance, in information extraction, supposing that if two specific entities are present in a sentence, they always share the same relationship (Simon, Guigue, and Piwowarski 2019) can be a good starting point to learn an information extraction system. Most of the works (but quantum-based representations) presented in this manuscript are also based on this idea that better representations can be captured through the definition of appropriate learning schemes.

A possible direction would thus be to constrain models (for instance, self-attention models) to have structured attention mechanisms, whereby accessing some information would imply *using* this structure. Some works have already took a step in this direction, such as the LongFormer (Beltagy, M. E. Peters, and Cohan 2020) where the long-range attention is constrained to be on specific tokens.

A complementary approach would be in leveraging datasets and tasks where this structured representation could be necessary, such as document summarization or question answering, or better, designing unsupervised tasks equivalent to token masking used to train SANs, but in which structured representations would bring important information.

I also believe that another important aspect of building structured representations systems is to leverage some knowledge of the world. It is not surprising that some tasks benefit from the use of resources like WordNet (network of words and concepts) or Wikipedia. In tasks such as question answering, textual entailment, and information retrieval, they help to reduce the mismatch between vocabularies produced by different sources (for example, the question by the user and the response by the user). These resources can be used as a source of information (Müller and Gurevych 2009), where Wikipedia is used to identify concepts (defined as Wikipedia entries) that can be used to index documents. Neural models for question-answering (Y. Zhang et al. 2016) have also proposed to exploit the knowledge bases, but more from a research perspective than on the real modeling of an interdependence of a text and knowledge. There is also a link with the work conducted on *grounding* language, since there is always an interplay between trying to integrate world knowledge into the embeddings (as done in our works on grounding sentences and words), and using external databases that can be accessed by models.

### 7.2.5 Application in Information Access

Finally, developing more structured distributed representations of information is key to handling more complex tasks. In the following, I give hints on how two complex tasks such as interactive information retrieval and summarization could benefit from structured representations.

Interactive information retrieval systems – whose aim is to interact with a user who is searching for information – would benefit greatly with such memories since they might be key to allow counterfactual reasoning ("what if..."?) more appropriately than other kind of representations, as argued in (Bottou et al. 2013), for instance by manipulating a part of the structure of the representation and seeing how this changes the answer of the model. Another use of structure would be to represent the current knowledge state of the user, which is important to support a user searching for new knowledge (Belkin, Oddy, and Brooks 1982).

In summarization, observing that neural approaches fail to produce salient and logically entailed (by the document) summaries, there is trend towards getting back to the idea of extracting (salient sentences) and the summarizing. Some works are trying to use sentence entailment to obtain a better structure in the output (Pasunuru and Bansal 2018). Having access to a structured representation of the text would allow such a research direction.

# Bibliography


Adel, H., B. Roth, and H. Schütze (2016). "Comparing Convolutional Neural Networks to Traditional Models for Slot Filling." In:

Airoldi, E., D. Blei, E. Xing, and S. Fienberg (2005). "A Latent Mixed Membership Model for Relational Data." In: *Proceedings of the 3rd International Workshop on Link Discovery*. LinkKDD '05. New York, NY, USA: ACM, pp. 82–89. DOI: 10.1145/1134271.1134283.

Akata, Z., F. Perronnin, Z. Harchaoui, and C. Schmid (2016). "Label-embedding for image classification." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 38.7. ZSCC: NoCitationData[s0] tex.bibsource: dblp computer science bibliography, https://dblp.org tex.biburl: https://dblp.org/rec/bib/jou tex.timestamp: Thu, 08 Jun 2017 09:06:11 +0200, pp. 1425–1438. DOI: 10/gfkxc5.

Allan, J. (1995). "Relevance Feedback with Too Much Data." In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '95. New York, NY, USA: ACM, pp. 337–343. DOI: 10/bzhd4v.

Amini, M. R. and N. Usunier (2011). "Transductive Learning over Automatically Detected Themes for Multi-Document Summarization." In: *Proceedings of the* $34^{th}$ *annual international ACM SIGIR conference*.

Andrew, G., R. Arora, J. A. Bilmes, and K. Livescu (2013). "Deep Canonical Correlation Analysis." In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 1247–1255.

Angelova, R., G. Kasneci, and G. Weikum (2012a). "Graffiti: graph-based classification in heterogeneous networks." In: *World Wide Web* 15.2, pp. 139–170.

— (2012b). "Graffiti: graph-based classification in heterogeneous networks." English. In: 15.2, pp. 139–170. DOI: 10.1007/s11280-011-0126-4.

Arjovsky, M., A. Shah, and Y. Bengio (2015). "Unitary Evolution Recurrent Neural Networks." In:

Arora, S., Y. Li, Y. Liang, T. Ma, and A. Risteski (Feb. 11, 2015). "RAND-WALK: A Latent Variable Model Approach to Word Embeddings." In: *Transactions of the Association for Computational Linguistics*. arXiv: 1502.03520.

Arroyo-Fernández, I., C.-F. Méndez-Cruz, G. Sierra, J.-M. Torres-Moreno, and G. Sidorov (July 1, 2019). "Unsupervised Sentence Representations as Word Information Series: Revisiting TF–IDF." In: *Computer Speech & Language* 56, pp. 107–129. DOI: 10/ggcx54.

Bach, F. R. and M. I. Jordan (2002). "Kernel Independent Component Analysis." In: *Journal of Machine Learning Research* 3, pp. 1–48.

Bagherinezhad, H., H. Hajishirzi, Y. Choi, and A. Farhadi (2016). "Are Elephants Bigger than Butterflies? Reasoning about Sizes of Objects." In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. Pp. 3449–3456.

Baroni, M., G. Dinu, and G. Kruszewski (2014). "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors." In:

Barsalou, L. W. (2008). "Grounded Cognition." In: *Annual Review of Psychology*.

Barzilay, R. and M. Elhadad (1997). "Using Lexical Chains for Text Summarization." In: *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10–17.

Bay, H., A. Ess, T. Tuytelaars, and L. Van Gool (2008). "Speeded-up Robust Features (SURF)." In: *Computer vision and image understanding* 110.3, pp. 346–359. DOI: 10/ffsc9r.

Beinborn, L., T. Botschen, and I. Gurevych (June 17, 2018). "Multimodal Grounding for Language Processing." In: *arXiv:1806.06371 [cs]*. arXiv: 1806.06371.

Belhumeur, P., J. Hespanha, and D. Kriegman (July 1997). "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.7, pp. 711–720. DOI: 10.1109/34.598228.

Belkin, N., R. Oddy, and H. Brooks (Jan. 1982). "Ask for Information Retrieval: Part I. Background and Theory." In: *Journal of Documentation* 38.2, pp. 61–71. DOI: 10/fcc4qn.

Beltagy, I., K. Lo, and A. Cohan (Nov. 2019). "SciBERT: A Pretrained Language Model for Scientific Text." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3606–3611. DOI: 10/ggcgtm.

Beltagy, I., M. E. Peters, and A. Cohan (Apr. 10, 2020). "Longformer: The Long-Document Transformer." In: *arXiv:2004.05150 [cs]*. ZSCC: 0000004. arXiv: 2004.05150.

Bender, E. M. and A. Koller (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In: *ACL*, p. 13.

Bengio, Y. (2000). "Continuous optimization of hyper-parameters." In: *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*. Vol. 1. IEEE, pp. 305–310.

Bengio, Y., A. Courville, and P. Vincent (2014). "Representation Learning: A Review and New Perspectives." In:

Bengio, Y., A. Courville, and P. Vincent (2013). "Representation learning: A review and new perspectives." In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.8, pp. 1798–1828.

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin (2003). "A Neural Probabilistic Language Model." In: *The Journal of Machine Learning Research*, pp. 1137–1155.

Besold, T. R., A. d. Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kuehnberger, L. C. Lamb, D. Lowd, P. M. V. Lima, L. de Penning, G. Pinkas, H. Poon, and G. Zaverucha (Nov. 10, 2017). "Neural-Symbolic Learning and Reasoning: A Survey and Interpretation." In: *arXiv:1711.03902 [cs]*. arXiv: 1711.03902.

Bezenac, E. d., A. Pajot, and P. Gallinari (Feb. 15, 2018). "Deep Learning for Physical Processes: Incorporating Prior Scientific Knowledge." In: International Conference on Learning Representations.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*.

Blei, D., L. Carin, and D. Dunson (2010). "Probabilistic Topic Models." In:

Blei, D., A. Ng, and M. Jordan (2003). "Latent Dirichlet Allocation." In: *Journal of Machine Learning Research*.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (July 15, 2016). "Enriching Word Vectors with Subword Information." In: *TACL*. arXiv: 1607.04606.

Bojchevski, A. and S. Günnemann (July 12, 2017a). "Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking." In: arXiv: 1707.03815 [cs, stat].

— (July 12, 2017b). "Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking." In: *arXiv:1707.03815 [cs, stat]*. arXiv: 1707.03815.

Bordes, A., R. Collobert, J. Weston, and Y. Bengio (2011). "Learning Structured Embeddings of Knowledge Bases." In: *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Bordes, A., N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko (June 2013). "Translating Embeddings for Modeling Multi-relational Data." In: pp. 2787–2795.

Bos, J., V. Basile, K. Evang, N. J. Venhuizen, and J. Bjerva (2017). "The Groningen Meaning Bank." In: *Handbook of Linguistic Annotation*. Ed. by N. Ide and J. Pustejovsky. Dordrecht: Springer Netherlands, pp. 463–496. DOI: 10.1007/978-94-024-0881-2_18.

Bottou, L., J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson (July 27, 2013). "Counterfactual Reasoning and Learning Systems." In: *arXiv:1209.2355 [cs, math, stat]*. arXiv: 1209.2355.

Bradbury, J. and R. Socher (Sept. 2017). "Towards Neural Machine Translation with Latent Tree Attention." In: *arXiv:1709.01915 [cs]*.

Britz, D., A. Goldie, M.-T. Luong, and Q. Le (Mar. 2017). "Massive Exploration of Neural Machine Translation Architectures." In: *arXiv:1703.03906 [cs]*.

Broek, P., M. Young, Y. Tzeng, T. Linderholm, and H. v. Oostendorp (1999). "The landscape model of reading." In:

Brown, G. G. and H. C. Rutemiller (1977). "Means and Variances of Stochastic Vector Products with Applications to Random Linear Models." In: *Management Science* 24.2.

Bruni, E., G. Boleda, M. Baroni, and N.-K. Tran (2012a). "Distributional semantics in technicolor." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* 1.July, pp. 136–145.

— (2012b). "Distributional Semantics in Technicolor." In:

Brysbaert, M., A. B. Warriner, and V. Kuperman (Sept. 2014). "Concreteness ratings for 40 thousand generally known English word lemmas." In: *Behavior Research Methods* 46.3, pp. 904–911. DOI: 10.3758/s13428-013-0403-5.

Bucher, M., S. Herbin, and F. Jurie (Aug. 23, 2017). "Generating Visual Representations for Zero-Shot Classification." In: *arXiv:1708.06975 [cs]*. arXiv: 1708.06975.

Burges, C., T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender (2005). "Learning to rank using gradient descent." In: pp. 89–96.

Burgess, C., K. Livesay, and K. Lund (Jan. 1, 1998). "Explorations in Context Space: Words, Sentences, Discourse." In: *Discourse Processes* 25.2-3, pp. 211–257. DOI: 10/dvgjmn.

Burgess, C. and K. Lund (Mar. 1997). "Modelling parsing constraints with high-dimensional context space." In: 12.

Cao, S., W. Lu, and Q. Xu (2015). "GraRep: Learning Graph Representations with Global Structural Information." In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, pp. 891–900.

Caputo, A., B. Piwowarski, and M. Lalmas (2011). "A Query Algebra for Quantum Information Retrieval." In: *Proceedings of the 2nd Italian Information Retrieval Workshop*.

Carbonell, J. and J. Goldstein (1998). "The use of MMR, diversity-based reranking for reordering documents and producing summaries." In:

Carvalho, M., R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord (2018). "Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings." In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pp. 35–44. DOI: 10.1145/3209978.3210036.

Castrejon, L., Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba (2016). "Learning Aligned Cross-Modal Representations from Weakly Aligned Data." In:

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil (Apr. 12, 2018). "Universal Sentence Encoder." In: arXiv: 1803.11175 [cs].

Chen, L., J. Zeng, and N. Tokuda (2006). "A "Stereo" Document Representation for Textual Information Retrieval." In:

Chen, Y., X. Wang, and B. Liu (2005). "Multi-document summarization based on Lexical chains." In: *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, pp. 1937–1942.

Chen, Y.-C. and M. Bansal (July 2018). "Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2018. Melbourne, Australia: Association for Computational Linguistics, pp. 675–686. DOI: 10/ggm84n.

Chen, Y., J. Pu, X. Liu, and X. Zhang (2019). "Gaussian Mixture Embedding of Multiple Node Roles in Networks." In: p. 29.

Cho, K., B. Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." In:

Chomsky, N. (1980). "Rules and representations." In: *Behavioral and brain sciences* 3.1. ZSCC: 0006786 tex.publisher: Cambridge University Press, pp. 1–15. DOI: 10/djk9fk.

Chrupala, G., Á. Kádár, and A. Alishahi (2015). "Learning language through pictures." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pp. 112–118.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning (Sept. 25, 2019). "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." In: International Conference on Learning Representations.

Clarke, C. L., M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon (2008). "Novelty and Diversity in Information Retrieval Evaluation." In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. New York, NY, USA: ACM, pp. 659–666. DOI: 10.1145/1390334.1390446.

Cleverdon, C. (1967). "The Cranfield tests on index language devices." In: *Aslib proceedings*.

Collell, G. and M.-F. Moens (2016). "Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences between Visual and Linguistic Representations." en. In: *The 26th International Conference on Computational Linguistics*, p. 11.

— (2018a). "Do Neural Network Cross-Modal Mappings Really Bridge Modalities?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, p. 7.

— (2018b). "Do Neural Network Cross-Modal Mappings Really Bridge Modalities?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 462–468.

Collell, G., L. Van Gool, and M.-F. Moens (2018). "Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Collell, G., T. Zhang, and M.-F. Moens (2017). "Imagined Visual Representations as Multimodal Embeddings." In: *AAAI*, pp. 4378–4384.

Collell, G., T. Zhang, and M.-f. Moens (2017). "Imagined Visual Representations as Multimodal Embeddings." In:

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). "Natural language processing (almost) from scratch." In: 12, pp. 2493–2537.

Conneau, A. and D. Kiela (Mar. 2018). "SentEval: An Evaluation Toolkit for Universal Sentence Representations." In: *arXiv:1803.05449 [cs]*.

Conneau, A., G. Lample, M. Ranzato, L. Denoyer, and H. Jégou (Oct. 2017). "Word Translation Without Parallel Data." In: *arXiv:1710.04087 [cs]*.

Costa Pereira, J., E. Coviello, G. Doyle, N. Rasiwasia, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos (Mar. 2014). "On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.3, pp. 521–535. DOI: 10.1109/TPAMI.2013.142.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman (1990). "Indexing by latent semantic analysis." In:

Delbrouck, J.-B., S. Dupont, and O. Seddati (July 2017). "Visually Grounded Word Embeddings and Richer Visual Features for Improving Multimodal Neural Machine Translation." In: *arXiv:1707.01009 [cs]*. arXiv: 1707.01009 [cs].

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li (2009). "ImageNet: A Large-Scale Hierarchical Image Database." In:

Deveaud, R. (Nov. 29, 2013). "Vers une représentation du contexte thématique en Recherche d'Information." PhD thesis. Université d'Avignon.

Deveaud, R., E. SanJuan, and P. Bellot (Aug. 2013). "Are Semantically Coherent Topic Models Useful for Ad Hoc Information Retrieval?" In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL 2013. Sofia, Bulgaria: Association for Computational Linguistics, pp. 148–152.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (Oct. 10, 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *arXiv:1810.04805 [cs]*. arXiv: 1810.04805.

Devooght, R., A. Mantrach, I. Kivimäki, H. Bersini, A. Jaimes, and M. Saerens (2014). "Random walks based modularity: application to semi-supervised learning." In: *Proceedings of the 23rd international conference on World wide web*. ACM, pp. 213–224.

Divvala, S. K., A. Farhadi, and C. Guestrin (2014). "Learning Everything about Anything: Webly-Supervised Visual Concept Learning." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3270–3277. DOI: 10.1109/CVPR.2014.412.

Dos Santos, L., B. Piwowarski, and P. Gallinari (2016). "Multilabel Classification on Heterogeneous Graphs with Gaussian Embeddings." English. In: *ECML*. Springer International Publishing, pp. 606–622. DOI: 10.1007/978-3-319-46227-1_38.

— (2017). "Gaussian Embeddings for Collaborative Filtering." In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '17. New York, NY, USA: ACM, pp. 1065–1068. DOI: 10.1145/3077136.3080722.

Dumais, S. (2004). "Latent semantic analysis." In:

Dunlop, M. D. (Apr. 1, 1997). "The Effect of Accessing Nonmatching Documents on Relevance Feedback." In: *ACM Transactions on Information Systems* 15.2, pp. 137–153. DOI: 10/c2963b.

Dupret, G. and B. Piwowarski (2008). "A user browsing model to predict search engine click data from past observations." In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. Ed. by S.-H. Myaeng, F. Sebastiani, T.-S. Chua, and M.-K. Leong. Singapore: ACM.

Erkan, G. and D. R. Radev (2004). "LexRank: Graph-based Centrality as Salience in Text Summarization." In: *Journal of Artificial Intelligence Research* 22 (1), pp. 457–479.

Fan, S. and B. Huang (2017). "Recurrent Collective Classification." In: *arXiv preprint arXiv:1703.06514*.

Fang, H., S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig (Nov. 18, 2014). "From Captions to Visual Concepts and Back." In: *arXiv:1411.4952 [cs]*. arXiv: 1411.4952.

Freeman, W. T. and M. Roth (1995). "Orientation Histograms for Hand Gesture Recognition." In: *International Workshop on Automatic Face and Gesture Recognition*. Vol. 12, pp. 296–301.

Frome, A., G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov (2013). "DeViSE: A Deep Visual-Semantic Embedding Model." In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Pp. 2121–2129.

Frommholz, I., B. Larsen, B. Piwowarski, M. Lalmas, P. Ingwersen, and K. Rijsbergen (Aug. 2010). "Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework." In: *Proceedings of the third symposium on Information interaction in context*. New Brunswick, NJ, USA. DOI: 10.1145/1840784.1840802.

Frommholz, I., B. Piwowarski, M. Lalmas, and K. Rijsbergen (Apr. 2011). "Processing Queries in Session in a Quantum-inspired IR Framework." In: *Proceedings of the 33rd European conference on Advances in information retrieval*.

Fuhr, N. and C. Buckley (July 1991). "A Probabilistic Learning Approach for Document Indexing." In: *ACM Trans. Inf. Syst.* 9.3, pp. 223–248. DOI: 10/b73xx7.

Fukushima, K. and S. Miyake (1982). "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position." In: *Pattern Recognition* 15.6, pp. 455–469. DOI: 10/cpsr8j.

Gallagher, B., H. Tong, T. Eliassi-Rad, and C. Faloutsos (2008). "Using ghost edges for classification in sparsely labeled networks." In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 256–264.

Genest, P.-E. (2013). "Génération de résumés par abstraction." PhD thesis. Université de Montréal.

Getoor, L. (2007). *Introduction to statistical relational learning*. MIT press.

Glenberg, A. M. and M. P. Kaschak (2002). "Grounding language in action." In: *Psychonomic bulletin & review*.

Glorot, X., A. Bordes, and Y. Bengio (2011). "Deep Sparse Rectifier Neural Networks." In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pp. 315–323.

Gomez, A. N., M. Ren, R. Urtasun, and R. B. Grosse (2017). "The Reversible Residual Network: Backpropagation Without Storing Activations." In: *NIPS*. Long Beach, USA, p. 11.

Gong, Y. and X. Lin (2001). "Generic text summarization using relevance measure and Latent Semantic Analysis." In: *Proceedings of the* $24^{th}$ *annual international ACM SIGIR conference*, pp. 19–25.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*.

Gordon, J. and B. Durme (2013). "Reporting bias and knowledge acquisition." In:

Gori, M., F. Scarselli, A. Tsoi, M. Hagenbuchner, and G. Monfardini (2009). "The graph neural network model." In:

Goyal, Y., T. Khot, D. Summers-Stay, D. Batra, and D. Parikh (Dec. 2, 2016). "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering." In: *arXiv:1612.00837 [cs]*. arXiv: 1612.00837.

Grant, E., C. Finn, S. Levine, T. Darrell, and T. Griffiths (Jan. 2018). "Recasting Gradient-Based Meta-Learning as Hierarchical Bayes." In: *arXiv:1801.08930 [cs]*.

Graves, A., G. Wayne, and I. Danihelka (2014). "Neural Turing Machines." In:

Grice, H. P. (1975). "Logic and conversation." In: *1975*, pp. 41–58.

Grover, A. and J. Leskovec (2016). "Node2Vec: Scalable Feature Learning for Networks." In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864.

Gudder, S. P. (1988). *Quantum Probability*. Elsevier. DOI: 10.1016/C2009-0-22184-2.

Guo, J., Y. Fan, Q. Ai, and W. Croft (2016). "A Deep Relevance Matching Model for Ad-Hoc Retrieval." In:

Harabagiu, S. and F. Lacatusu (2005). "Topic themes for multi-document summarization." In: *Proceedings of the* $28^{th}$ *annual international ACM SIGIR conference*, pp. 202–209.

He, J., V. Hollink, and A. de Vries (2012). "Combining implicit and explicit topic representations for result diversification." In: *SIGIR*.

He, K., X. Zhang, S. Ren, and J. Sun (2015). "Deep Residual Learning for Image Recognition." In:

He, S., K. Liu, G. Ji, and J. Zhao (2015). "Learning to Represent Knowledge Graphs with Gaussian Embedding." In:

Hewitt, J. and C. D. Manning (June 2019). "A Structural Probe for Finding Syntax in Word Representations." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. NAACL-HLT 2019. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4129–4138. DOI: 10/ggxrg9.

Hill, F., K. Cho, and A. Korhonen (2016). "Learning Distributed Representations of Sentences from Unlabelled Data." In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 1367–1377.

Hill, F. and A. Korhonen (2014a). "Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 255–265.

— (2014b). "Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean." In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 255–265.

Hill, F., R. Reichart, and A. Korhonen (2014). "Multi-Modal Models for Concrete and Abstract Concept Meaning." In: *Transactions of the Association for Computational Linguistics* 2, pp. 285–296.

Hochreiter, S. and J. Schmidhuber (Nov. 1, 1997). "Long Short-Term Memory." In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Hoenkamp, E. (2011). "Trading Spaces: On the Lore and Limitations of Latent Semantic Analysis." In:

Hoffmann, R., C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld (2011). "Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 541–550.

Hofmann, T. (2001). "Unsupervised Learning by Probabilistic Latent Semantic Analysis." In:

Hu, R., J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko (Oct. 2017). "Learning to Reason: End-to-End Module Networks for Visual Question Answering." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, pp. 804–813. DOI: 10.1109/ICCV.2017.93.

Huang, P., X. He, J. Gao, L. Deng, A. Acero, and L. Heck (2013). "Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data." In:

Huang, Y. Y. and W. Y. Wang (2017). "Deep Residual Learning for Weakly-Supervised Relation Extraction." In: *EMNLP*.

Huertas-Rosero, Á. F., L. Azzopardi, and C. J. van Rijsbergen (Feb. 18, 2008). "Characterising through Erasing: A Theoretical Framework for Representing Documents Inspired by Quantum Theory." In: arXiv: 0802.1738 [quant-ph].

Hwang, T. and R. Kuang (2010). "A heterogeneous label propagation algorithm for disease gene discovery." In: *SDM*, p. 12.

Ingwersen, P. and K. Järvelin (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. The Information Retrieval Series. Springer Netherlands. DOI: 10.1007/1-4020-3851-8.

Iyyer, M., J. L. Boyd-Graber, L. M. B. Claudino, R. Socher, and H. Daumé III (2014). "A Neural Network for Factoid Question Answering over Paragraphs." In: *EMNLP*, pp. 633–644.

Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daumé III (July 2015). "Deep Unordered Composition Rivals Syntactic Methods for Text Classification." In: *Proceedings of the 53rd*

*Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL-IJCNLP 2015. Beijing, China: Association for Computational Linguistics, pp. 1681–1691. DOI: 10/gf332j.

Jacob, Y., L. Denoyer, and P. Gallinari (Jan. 2014). "Learning Latent Representations of Nodes for Classifying in Heterogeneous Social Networks." In: *WSDM '14: Proceedings of the 7th ACM international conference on Web search and data mining*. ACM Request Permissions, pp. 1–10. DOI: 10.1145/2556195.2556225.

Ji, M., Y. Sun, M. Danilevsky, J. Han, and J. Gao (2010a). "Graph regularized transductive classification on heterogeneous information networks." In: *European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 570–586.

— (2010b). "Graph regularized transductive classification on heterogeneous information networks." In: *ECML PKDD*. Vol. 0053. Springer, pp. 570–586.

Jiang, Y.-G., J. Yang, C.-W. Ngo, and A. G. Hauptmann (2010). "Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study." In: *IEEE Trans. Multimedia* 12.1, pp. 42–53. DOI: 10.1109/TMM.2009.2036235.

Johnson, D. D. (2017). "Learning Graphical State Transitions." In: *ICTIR*. Toulon, France.

Joulin, A. and T. Mikolov (2015). "Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets." In:

Kalchbrenner, N., E. Grefenstette, and P. Blunsom (2014). "A Convolutional Neural Network for Modelling Sentences." In:

Kamp, H. and U. Reyle (July 28, 1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. 1993 ed. Dordrecht: Springer. 717 pp.

Karpathy, A., A. Joulin, and F. F. F. Li (2014). "Deep fragment embeddings for bidirectional image sentence mapping." In: *Advances in neural information processing systems*, pp. 1889–1897.

Karpathy, A. and F.-F. Li (2015). "Deep visual-semantic alignments for generating image descriptions." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3128–3137. DOI: 10.1109/CVPR.2015.7298932.

Katharopoulos, A., A. Vyas, N. Pappas, and F. Fleuret (June 30, 2020). "Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention." In: *arXiv:2006.16236 [cs, stat]*. arXiv: 2006.16236.

Kazai, G. (2009). "INitiative for the Evaluation of XML Retrieval." In: *Encyclopedia of Database Systems*. Ed. by L. LIU and M. T. ÖZSU. Boston, MA: Springer US, pp. 1531–1537. DOI: 10.1007/978-0-387-39940-9_151.

Kenter, T., A. Borisov, and M. de Rijke (2016). "Siamese CBOW: Optimizing Word Embeddings for Sentence Representations." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL. Berlin, Germany.

Kiela, D. and L. Bottou (2014). "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics." In: *EMNLP 2014*.

Kiela, D., A. Conneau, A. Jabri, and M. Nickel (June 2018). "Learning Visually Grounded Sentence Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. NAACL-HLT 2018. ZSCC: NoCitationData[s0]. New Orleans, Louisiana: Association for Computational Linguistics, pp. 408–418. DOI: 10/ggjq5w.

Kim, Y.-D. and S. Choi (2014). "Bayesian binomial mixture model for collaborative prediction with non-random missing data." en. In: ACM Press, pp. 201–208. DOI: 10.1145/2645710.2645754.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*.

Kiros, R., Y. Zhu, R. Salakhutdinov, R. S. Zemel, R. Urtasun, A. Torralba, and S. Fidler (2015). "Skip-Thought Vectors." In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3294–3302.

Kitaev, N., Ł. Kaiser, and A. Levskaya (2020). "Reformer: The Efficient Transformer." In: *ICLR*. arXiv: 2001.04451.

Klein, B., G. Lev, G. Sadeh, and L. Wolf (2015). "Associating neural word embeddings with deep image representations using Fisher Vectors." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4437–4446. DOI: 10.1109/CVPR.2015.7299073.

Kok, S. and P. Domingos (Sept. 15, 2008). "Extracting Semantic Networks from Text Via Relational Clustering." In: *Machine Learning and Knowledge Discovery in Databases*. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, pp. 624–639. DOI: 10.1007/978-3-540-87479-9_59.

Kottur, S., R. Vedantam, J. M. F. Moura, and D. Parikh (2016). "VisualWord2Vec (Vis-W2V): Learning Visually Grounded Word Embeddings Using Abstract Scenes." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4985–4994. DOI: 10.1109/CVPR.2016.539.

Kovaleva, O., A. Romanov, A. Rogers, and A. Rumshisky (2019). "Revealing the Dark Secrets of BERT." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, pp. 4364–4373. DOI: 10.18653/v1/D19-1445.

Kraljevic, Z., N. Baskiotis, B. Piwowarski, and P. Gallinari (Mar. 2016). "Représentation Temporelle Des Mots : Application Au Clustering de Micro-Blogs." In: *Conférence En Recherche d'Infomations et Applications*, pp. 531–544.

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. (2016). "Visual genome: Connecting language and vision using crowdsourced dense image annotations." In: *arXiv preprint arXiv:1602.07332*.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks." In: *NIPS*, pp. 1106–1114.

Kumar, A., O. İrsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher (June 2015). "Ask Me Anything: Dynamic Memory Networks for Natural Language Processing." In: *arXiv.org*. arXiv: 1506.07285v1 [cs.CL].

Kumar, S. and Y. Tsvetkov (Sept. 2018). "Von Mises-Fisher Loss for Training Sequence to Sequence Models with Continuous Outputs." In: *International Conference on Learning Representations*.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut (2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." In: International Conference on Learning Representations.

Lanckriet, G. R. G., N. Cristianini, P. L. Bartlett, L. El Ghaoui, and M. I. Jordan (2004). "Learning the Kernel Matrix with Semidefinite Programming." In: *The Journal of Machine Learning Research*, pp. 27–72.

Lazaridou, A., N. T. Pham, and M. Baroni (2015). "Combining Language and Vision with a Multimodal Skip-gram Model." In: *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pp. 153–163.

Le, Q. V. and T. Mikolov (May 2014). "Distributed Representations of Sentences and Documents." In: *Proceedings of the 31st International Conference on Machine Learning*.

Le Digabel, S. (Feb. 2011). "Algorithm 909: NOMAD: Nonlinear Optimization with the MADS Algorithm." English. In: *ACM Transactions on Mathematical Software (TOMS)* 37.4, pp. 44–15. DOI: 10.1145/1916461.1916468.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (Dec. 1, 1989). "Backpropagation Applied to Handwritten Zip Code Recognition." In: *Neural Computation* 1.4, pp. 541–551. DOI: 10.1162/neco.1989.1.4.541.

Levy, O. and Y. Goldberg (2014). "Neural Word Embedding as Implicit Matrix Factorization." In:

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer (Oct. 2019). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." en. In: *arXiv:1910.13461 [cs, stat]*.

Li, J. and L. Sun (2008). "A Lexical chain approach for update-style query-focused multi-document summarization." In: *Proceedings of the 4$^{th}$ Asia Information Retrieval Conference on Information Retrieval Technology*, pp. 310–320.

Li, Y., D. Tarlow, M. Brockschmidt, and R. Zemel (Nov. 2015). "Gated Graph Sequence Neural Networks." In: *arXiv:1511.05493 [cs, stat]*.

Li, Y., Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo (2016). *TGIF: A New Dataset and Benchmark on Animated GIF Description*.

Lin, T.-Y., M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). "Microsoft COCO: Common Objects in Context." In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48.

Lin, Y., S. Shen, Z. Liu, H. Luan, and M. Sun (Aug. 2016). "Neural Relation Extraction with Selective Attention over Instances." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 2124–2133.

Liu, T.-Y. (2011). *Learning to Rank for Information Retrieval*. Springer.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (July 2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." en. In: *arXiv:1907.11692 [cs]*.

Logeswaran, L. and H. Lee (Mar. 2018). "An efficient framework for learning sentence representations." In: *arXiv:1803.02893 [cs]*.

Lowe, D. G. (1999). "Object Recognition from Local Scale-Invariant Features." In: *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference On*. Vol. 2. Ieee, pp. 1150–1157. DOI: 10/dqfs59.

Lu, Q. and L. Getoor (2003). "Link-based classification." In: *ICML*. Vol. 3, pp. 496–503.

Lu, Y. M. and M. N. Do (2007). "Multidimensional Directional Filter Banks and Surfacelets." In: *IEEE Trans. Image Processing* 16.4, pp. 918–931. DOI: 10/fgg3vz.

Luketina, J., T. Raiko, M. Berglund, and K. Greff (2016). "Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters." In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 2952–2960.

Maaten, L. v. d. and G. Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.Nov, pp. 2579–2605.

Mansimov, E., E. Parisotto, J. Ba, and R. Salakhutdinov (2015). "Generating Images from Captions with Attention." In:

Marcheggiani, D. and I. Titov (June 8, 2016). "Discrete-State Variational Autoencoders for Joint Discovery and Factorization of Relations." In: *Transactions of the Association for Computational Linguistics* 4.0, pp. 231–244.

Martin, L., B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot (May 1, 2020). "CamemBERT: a Tasty French Language Model." In: *arXiv:1911.03894 [cs]*. arXiv: 1911.03894.

McCoy, T., E. Pavlick, and T. Linzen (July 2019). "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 3428–3448. DOI: 10.18653/v1/P19-1334.

McRae, K., G. S. Cree, M. S. Seidenberg, and C. Mcnorgan (Nov. 2005). "Semantic feature production norms for a large set of living and nonliving things." en. In: *Behavior Research Methods* 37.4, pp. 547–559. DOI: 10.3758/BF03192726.

Melucci, M. (2008). "A basis for information retrieval in context." In:

Metzler, D. and W. B. Croft (Sept. 2004). "Combining the language model and inference network approaches to retrieval." In: *Information Processing & Management*. Bayesian Networks and Information Retrieval 40.5, pp. 735–750. DOI: 10/c8vdcz.

Mihalcea, R. (2005). "Language Independent Extractive Summarization." In: *Proceedings of the ACL 2005 conference*, pp. 49–52.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality." In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.* Pp. 3111–3119.

Mitra, B. and N. Craswell (2017). *An Introduction to Neural Information Retrieval*.

Mitra, B., F. Diaz, and N. Craswell (2017). "Learning to Match Using Local and Distributed Representations of Text for Web Search." In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, pp. 1291–1299. DOI: 10.1145/3038912.3052579.

Moore, J. and J. Neville (2017). "Deep collective inference." In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

Moradshahi, M., H. Palangi, M. S. Lam, P. Smolensky, and J. Gao (Oct. 2019). "HUBERT Untangles BERT to Improve Transfer across NLP Tasks." en. In: *arXiv:1910.12647 [cs, stat]*.

Moreno, L., G. Bavota, M. Di Penta, R. Oliveto, A. Marcus, and G. Canfora (2017). "ARENA: An approach for the automated generation of release notes." In: *Proc. of IEEE Trans. on Software Engineering* 43.2, pp. 106–127.

Moshfeghi, Y., D. Agarwal, B. Piwowarski, and J. Jose (2009). "Movie Recommender: Semantically Enriched Unified Relevance Model for Rating Prediction in Collaborative Filtering Lecture Notes in Computer Science." In: *Proceedings of the 31th European Conference on Information Retrieval Conference*. Ed. by M. Boughanem, C. Berrut, J. Mothe, and C. Soulé-Dupuy. Toulouse, France: Springer, pp. 54–65. DOI: 10.1007/978-3-642-00958-7_8.

Moshfeghi, Y., B. Piwowarski, and J. Jose (2011). "Handling Data Sparsity in Collaborative Filtering using Emotion and Semantic Based Features." In: *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM SIGIR Conference on Research and Development in Information Retrieval. ZSCC: 0000084. ACM.

Mukherjee, T. and T. Hospedales (2016). "Gaussian Visual-Linguistic Embedding for Zero-Shot Recognition." In:

Müller, C. and I. Gurevych (2009). "Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval." In: *Evaluating Systems for Multilingual and Multimodal Information Access*.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*.

Murray, G., S. Renals, and J. Carletta (2005). "Extractive Summarization of Meeting Recordings." In: *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 593–596.

Nalisnick, E., B. Mitra, N. Craswell, and R. Caruana (2016). "Improving Document Ranking with Dual Word Embeddings." In: *Proceedings of the 25th International Conference Companion on World Wide Web*. WWW '16 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, pp. 83–84. DOI: 10.1145/2872518.2889361.

Nandanwar, S. and M. N. Murty (2016). "Structural Neighborhood Based Classification of Nodes in a Network." In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, pp. 1085–1094. DOI: 10.1145/2939672.2939782.

Nelson, D. L., C. L. McEvoy, and T. A. Schreiber (2004). "The University of South Florida free association, rhyme, and word fragment norms." In: *Behavior Research Methods, Instruments, & Computers* 36.3, pp. 402–407.

Neville, J. and D. Jensen (2000). "Iterative classification in relational data." In: *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pp. 13–20.

Ng, A. Y., R. Socher, C. D. Manning, J. Pennington, and E. H. Huang (July 2011). "Semi-supervised recursive autoencoders for predicting sentiment distributions." In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Niepert, M., M. Ahmed, and K. Kutzkov (2016). "Learning Convolutional Neural Networks for Graphs." In: *ICML 2016*.

Niven, T. and H.-Y. Kao (July 2019). "Probing Neural Network Comprehension of Natural Language Arguments." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 4658–4664. DOI: 10.18653/v1/P19-1459.

Norouzi, M., T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean (Dec. 19, 2013). "Zero-Shot Learning by Convex Combination of Semantic Embeddings." In: *arXiv:1312.5650 [cs]*. arXiv: 1312.5650.

Ozsoy, M. G., I. Cicekli, and F. N. Alpaslan (2010). "Text summarization of Turkish texts using latent semantic analysis." In: *Proceedings of the* $23^{rd}$ *International Conference on Computational Linguistics*, pp. 869–876.

Page, L., S. Brin, R. Motwani, and T. Winograd (1999). "The PageRank citation ranking: bringing order to the web." In:

Pasunuru, R. and M. Bansal (Apr. 2018). "Multi-Reward Reinforced Summarization with Saliency and Entailment." In: *arXiv:1804.06451 [cs]*. DOI: 10/gfrh7x. arXiv: 1804.06451 [cs].

Pei, Y., X. Du, J. Zhang, G. Fletcher, and M. Pechenizkiy (May 12, 2020). "struc2gauss: Structural role preserving network embedding via Gaussian embedding." In: *Data Mining and Knowledge Discovery*. DOI: 10.1007/s10618-020-00684-x.

Pennington, J., R. Socher, and C. Manning (2014). "Glove: Global Vectors for Word Representation." In:

Perozzi, B., R. Al-Rfou, and S. Skiena (2014). "Deepwalk: Online learning of social representations." In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 701–710.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (Feb. 2018). "Deep contextualized word representations." In: *NAACL*.

Peters, S., L. Denoyer, and P. Gallinari (2010). "Iterative annotation of multi-relational social networks." In: *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*. IEEE, pp. 96–103.

Pham, T., T. Tran, D. Phung, and S. Venkatesh (2016). "Column Networks for Collective Classification." In: *arXiv preprint arXiv:1609.04508*.

Pimplikar, R., D. Garg, D. Bharani, and G. Parija (2014). "Learning to Propagate Rare Labels." In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, pp. 201–210.

Piwowarski, B. (2012). *The Kernel Quantum Probabilities (KQP) Library*. Tech. rep. 1203.6005v2.

— (July 2020). "Experimaestro and Datamaestro:Experiment and Dataset Managers (for IR)." In: *ACM SIGIR 2020*. Xian, China. DOI: 10.1145/3397271.3401410.

Piwowarski, B., M.-R. Amini, and M. Lalmas (2012). "On Using a Quantum Physics Formalism for Multidocument Summarization." In: *Journal of the American Society for Information Science and Technology* 63, pp. 865–888. DOI: 10.1002/asi.21713.

Piwowarski, B. and G. Dupret (2006). "Evaluation in (XML) Information Retrieval: Expected Precision-Recall with User Modelling (EPRUM)." In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. Ed. by E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Järvelin. Seattle, Washington, USA: ACM, pp. 260–267. DOI: 10.1145/1148170.1148218.

— (2009). "System and Method for Deducing User Interaction Patterns Based on Limited Activities."

Piwowarski, B., G. Dupret, and R. Jones (Feb. 2009). "Mining User Web Search Activity with Layered Bayesian Networks or How to Capture a Click in its Context." In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. Ed. by R. A. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, and B. B. Cambazoglu. Barcelona, Spain: ACM, pp. 162–171. DOI: 10.1145/1498759.1498823.

Piwowarski, B., I. Frommholz, M. Lalmas, and K. Rijsbergen (2010). "What can Quantum Theory bring to IR?" In: *Proceedings of the nineteenth ACM conference on Conference on information and knowledge management*. Ed. by J. Huang, N. Koudas, G. Jones, X. Wu, K. Collins-Thompson, and A. An. ACM. DOI: 10.1145/1871437.1871450.

Piwowarski, B., I. Frommholz, Y. Moshfeghi, M. Lalmas, and K. Rijsbergen (2010). "Filtering documents with subspaces." In: *ECIR*. Ed. by C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and K. van Rijsbergen. Vol. 5993. Advances in Information Retrieval. Springer.

Piwowarski, B., P. Gallinari, and G. Dupret (2007). "An Extension of Precision-Recall with User Modelling (PRUM): Application to XML Retrieval." In: *ACM Transactions On Information Systems* 25.1. DOI: 10.1145/1198296.1198297.

Piwowarski, B. and M. Lalmas (2009a). "A Quantum-based Model for Interactive Information Retrieval." In: *Proceeedings of the 2nd International Conference on the Theory of Information Retrieval*. Ed. by L. Azzopardi, G. Kazai, S. E. Robertson, S. M. Rüger, M. Shokouhi, and D. Song. Vol. 5766. Lecture Notes in Computer Science. Cambridge, United Kingdom: Springer.

— (2009b). "Structured Information Retrieval and Quantum Theory." In: *Proceedings of the 3rd QI Symposium*. Ed. by P. Bruza, D. Sofge, W. Lawless, K. van Rijsbergen, and M. Klusch. Vol. 5494. Springer.

Piwowarski, B. and H. Zaragoza (2007). "Predictive User Click Models Based on Click-through History." In: *Proceedings of the 16th ACM International Conference on Information and Knowledge Management*. Lisbon, Portugal: ACM, pp. 175–182.

— (2009). "System and Method for Creating and Applying Predictive User Click Models to Predict a Target Page Associated with a Search Query."

Plummer, B. A., L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik (2015). "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models." In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2641–2649. DOI: 10.1109/ICCV.2015.303.

Polanyi, L. and A. Zaenen (2006). "Contextual Valence Shifters." In: *Computing Attitude and Affect in Text: Theory and Applications*. Ed. by J. G. Shanahan, Y. Qu, and J. Wiebe. Vol. 20. Berlin/Heidelberg: Springer-Verlag, pp. 1–10. DOI: 10.1007/1-4020-4102-0_1.

Purpura, A., M. Maggipinto, G. Silvello, and G. A. Susto (2019). "Probabilistic Word Embeddings in Neural IR: A Promising Model That Does Not Work As Expected (For Now)." In: *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR '19. Santa Clara, CA, USA: ACM, pp. 3–10. DOI: 10/ggfqpb.

Qiu, G. (2002). "Indexing Chromatic and Achromatic Patterns for Content-Based Colour Image Retrieval." In: *Pattern Recognition* 35.8, pp. 1675–1686. DOI: 10/drzd79.

Qiu, X. and X. Huang (2015). "Convolutional Neural Tensor Network Architecture for Community-Based Question Answering." In: *IJCAI*, pp. 1305–1311.

Radev, D. R., H. Jing, M. Styś, and D. Tam (2004). "Centroid-based summarization of multiple documents." In: *Information Processing Management* 40 (6), pp. 919–938.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever (2018). *Improving Language Understanding by Generative Pre-Training*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2018). "Language Models are Unsupervised Multitask Learners." In: p. 24.

Rae, J. W., A. Potapenko, S. M. Jayakumar, C. Hillier, and T. P. Lillicrap (Sept. 25, 2019). "Compressive Transformers for Long-Range Sequence Modelling." In: International Conference on Learning Representations.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (Oct. 24, 2019). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." In: *arXiv:1910.10683 [cs, stat]*. ZSCC: 0000012 version: 2. arXiv: 1910.10683.

Rekabdar, B., C. Mousas, and B. Gupta (Jan. 2019). "Generative Adversarial Network with Policy Gradient for Text Summarization." In: *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. 2019 IEEE 13th International Conference on Semantic Computing (ICSC). Newport Beach, CA, USA: IEEE, pp. 204–207. DOI: 10/gfw8jd.

Rendle, S., C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme (June 2009). "BPR: Bayesian personalized ranking from implicit feedback." In: *Uncertainty in Artificial Intelligence*. AUAI Press, pp. 452–461.

Ricci, F., L. Rokach, B. Shapira, and K. Paul B. (2011). *Recommender Systems Handbook*. Vol. 532. DOI: 10.1007/978-0-387-85820-3.

Rijsbergen, K. van (2004). *The Geometry of Information Retrieval*.

Robertson, S. E. and S. Walker (1994). "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval." In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland). SIGIR '94. New York, NY, USA: Springer-Verlag New York, Inc., pp. 232–241.

Robertson, S. E. and H. Zaragoza (2009). *The Probabilistic Relevance Framework: BM25 and Beyond Foundations and Trends in Information Retrieval*.

Rogers, A., O. Kovaleva, and A. Rumshisky (Feb. 2020). "A Primer in BERTology: What we know about how BERT works." In: *arXiv:2002.12327 [cs]*. ZSCC: 0000000 arXiv: 2002.12327.

Rohrbach, A., M. Rohrbach, R. Hu, T. Darrell, and B. Schiele (2016). "Grounding of Textual Phrases in Images by Reconstruction." In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pp. 817–834. DOI: 10.1007/978-3-319-46448-0_49.

Roller, S. and S. Schulte im Walde (2013). "A Multimodal LDA Model Integrating Textual, Cognitive and Visual Modalities." In: *Emnlp* October, pp. 1146–1157.

Rosenberg, A. and J. Hirschberg (2007). "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure." In:

Rudolph, M. and D. Blei (Mar. 23, 2017). "Dynamic Bernoulli Embeddings for Language Evolution." In: arXiv: 1703.08052 [cs, stat].

Rudolph, M. R., F. J. R. Ruiz, S. Mandt, and D. M. Blei (Aug. 2016). "Exponential Family Embeddings." In: *arXiv:1608.00778 [cs, stat]*.

Saha, A. and V. Sindhwani (Feb. 2012). "Learning evolving and emerging topics in social media." In: *the fifth ACM international conference*. New York, New York, USA: ACM, pp. 693–702. DOI: 10.1145/2124295.2124376.

Sahu, S. K., A. Anand, K. Oruganty, and M. Gattu (June 2016). "Relation Extraction from Clinical Texts Using Domain Invariant Convolutional Neural Network." In:

Salakhutdinov, R. and A. Mnih (2007). "Probabilistic Matrix Factorization." In: *Proceedings of Advances in Neural Information Processing Systems*. Vol. 20, pp. 1257–1264.

Salton, G. and M. E. Lesk (June 1, 1965). "The SMART Automatic Document Retrieval Systems—an Illustration." In: *Communications of the ACM* 8.6, pp. 391–398. DOI: 10/c64rfr.

Salton, G., A. Wong, and C. S. Yang (Nov. 1975). "A Vector Space Model for Automatic Indexing." In: *Commun. ACM* 18.11, pp. 613–620. DOI: 10/fw8vv8.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf (Oct. 16, 2019). "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." In: arXiv: 1910.01108 [cs].

Santos, C. N. d., B. Xiang, and B. Zhou (2015). "Classifying relations by ranking with convolutional neural networks." In: *arXiv preprint arXiv:1504.06580*.

Santos, L. D., B. Piwowarski, L. Denoyer, and P. Gallinari (July 2018). "Representation Learning for Classification in Heterogeneous Graphs with Application to Social Networks." In: *ACM Transactions on Knowledge Discovery from Data* 12.5, 62:1–62:33. DOI: 10/gdvmmq.

Santos, R., J. Peng, C. Macdonald, and I. Ounis (2010). "Explicit Search Result Diversification through Sub-queries." In: *ECIR*.

Schulman, J., N. Heess, T. Weber, and P. Abbeel (June 17, 2015). "Gradient Estimation Using Stochastic Computation Graphs." In:

Schuster, M. and K. Paliwal (Dec. 1, 1997). "Bidirectional recurrent neural networks." In: *Signal Processing, IEEE Transactions on* 45, pp. 2673–2681. DOI: 10.1109/78.650093.

Schütze, H. (1993). "Word Space." In: *Advances in Neural Information Processing Systems 5*. Ed. by S. J. Hanson, J. D. Cowan, and C. L. Giles. Morgan-Kaufmann, pp. 895–902.

Schwenk, H. (July 1, 2007). "Continuous Space Language Models." In: *Computer Speech & Language* 21.3, pp. 492–518. DOI: 10/dzgfmr.

Scialom, T., P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano (Feb. 2020). "Discriminative Adversarial Search for Abstractive Summarization." In: *arXiv:2002.10375 [cs]*. ZSCC: NoCitationData[s0] arXiv: 2002.10375.

Scialom, T., S. Lamprier, B. Piwowarski, and J. Staiano (2019). "Answers Unite! Unsupervised Metrics for Reinforced Summarization Models." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. DOI: 10/ggchbh.

See, A., P. J. Liu, and C. D. Manning (Apr. 14, 2017). "Get To The Point: Summarization with Pointer-Generator Networks." In: *arXiv:1704.04368 [cs]*. arXiv: 1704.04368.

Sen, P., G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad (2008). "Collective classification in network data." In: *AI magazine* 29.3, p. 93.

Sennrich, R., B. Haddow, and A. Birch (2016). "Neural Machine Translation of Rare Words with Subword Units." In:

Shekhar, R., S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, and R. Bernardi (2017). "FOIL it! Find One mismatch between Image and Language caption." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1, pp. 255–265. DOI: 10.18653/v1/P17-1024.

Silberer, C. and M. Lapata (2012). "Grounded Models of Semantic Representation." In: *EMNLP-CoNLL*.

— (2014). "Learning Grounded Meaning Representations with Autoencoders." In:

Simon, E., V. Guigue, and B. Piwowarski (July 2019). "Unsupervised Information Extraction: Regularizing Discriminative Approaches with Relation Distribution Losses." In: *Proceedings of ACL 2019*. Firenze, Italia.

Simonyan, K. and A. Zisserman (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition." In:

Socher, R., A. Perelygin, J. Wu, and J. Chuang (2013). "Recursive deep models for semantic compositionality over a sentiment treebank." In:

Sordoni, A., Y. Bengio, J.-Y. Nie, and Y. Bengio (2013). "Modeling Term Dependencies with Quantum Language Models for IR." In:

Sordoni, A., J. He, and J.-Y. Nie (2013). "Modeling latent topic interactions using quantum interference for information retrieval." In: *CIKM*.

Srivastava, N., I. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." In:

Srivastava, N., E. Mansimov, and R. Salakhutdinov (Feb. 2015). "Unsupervised Learning of Video Representations using LSTMs." In: arXiv: 1502.04681.

Srivastava, R., K. Greff, and J. Schmidhuber (2015). "Highway Networks." In:

Stanojević, M. and M. Steedman (July 2020). "Max-Margin Incremental CCG Parsing." en-us. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4111–4122.

Steedman, M. and J. Baldridge (2005). "Combinatory Categorial Grammar." In:

Steinberger, J. and K. Ježek (2004). "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation." In: *Proceedings of ISIM*.

Stern, D., R. Herbrich, and T. Graepel (2009). "Matchbox: Large Scale Bayesian Recommendations." In:

Su, W., X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai (Sept. 25, 2019). "VL-BERT: Pre-Training of Generic Visual-Linguistic Representations." In: International Conference on Learning Representations.

Sukhbaatar, S., A. Szlam, J. Weston, and R. Fergus (2015). "End-To-End Memory Networks." In:

Surdeanu, M., J. Tibshirani, R. Nallapati, and C. D. Manning (2012). "Multi-instance multi-label learning for relation extraction." In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, pp. 455–465.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (Dec. 1, 2015). "Rethinking the Inception Architecture for Computer Vision." In: arXiv: 1512.00567 [cs].

Tang, J., M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei (2015). "LINE: Large-scale Information Network Embedding." In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1067–1077.

Taylor, M., J. Guiver, S. Robertson, and T. Minka (2008). "SoftRank: optimizing non-smooth rank metrics." In:

Teufel, S. and H. van Halteren (2004). "Evaluating information content by factoid analysis: human annotation and stability." en. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, p. 8.

Titeux, H., B. Piwowarski, and P. Gallinari (Mar. 2018a). "Représentations Gaussiennes pour le Filtrage Collaboratif." In: *Conférence en Recherche d'Information et Applications*.

— (Mar. 2018b). "Représentations Gaussiennes pour le Filtrage Collaboratif." In:

Tu, C., W. Zhang, Z. Liu, and M. Sun (2016). "Max-Margin DeepWalk: Discriminative Learning of Network Representation." In: *IJCAI*, pp. 3889–3895.

Vasile, F., E. Smirnova, and A. Conneau (2016). "Meta-Prod2Vec - Product Embeddings Using Side-Information for Recommendation." In: *arXiv:1607.07326 [cs]*, pp. 225–232. DOI: 10/gcvkdt.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (June 12, 2017). "Attention Is All You Need." In: *arXiv:1706.03762 [cs]*. arXiv: 1706.03762.

Vedantam, R., X. Lin, T. Batra, C. L. Zitnick, and D. Parikh (2015). "Learning Common Sense Through Visual Abstraction." In:

Verma, R. and D. Lee (2017). "Extractive Summarization: Limits, Compression, Generalized Model and Heuristics." In: *arXiv:1704.05550*.

Vilnis, L. and A. McCallum (2014). "Word Representations via Gaussian Embedding." In: *arXiv.org*.

Vinyals, O., A. Toshev, S. Bengio, and D. Erhan (2014). "Show and Tell: A Neural Image Caption Generator." In:

Wang, C., D. Blei, and D. Heckerman (2012). "Continuous Time Dynamic Topic Models." In:

Wang, D., T. Li, S. Zhu, and C. Ding (2008). "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization." In: *Proceedings of the* $31^{th}$ *annual international ACM SIGIR*, pp. 307–314.

Wang, D., P. Cui, and W. Zhu (2016). "Structural Deep Network Embedding." In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16, pp. 1225–1234.

Wang, J., T. Jebara, and S.-F. Chang (2008). "Graph Transduction via Alternating Minimization." In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. New York, NY, USA: ACM, pp. 1144–1151. DOI: 10.1145/1390156.1390300.

Wang, J. and J. Zhu (2009). "Portfolio theory of information retrieval." In: *SIGIR*. ACM Press. DOI: 10.1145/1571941.1571963.

Wang, L., Y. Li, and S. Lazebnik (2016). "Learning Deep Structure-Preserving Image-Text Embeddings." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 5005–5013. DOI: 10.1109/CVPR.2016.541.

Wang, S., B. Z. Li, M. Khabsa, H. Fang, and H. Ma (June 8, 2020). "Linformer: Self-Attention with Linear Complexity." In: *arXiv:2006.04768 [cs, stat]*. arXiv: 2006.04768.

Wang, W., N. Yang, F. Wei, B. Chang, and M. Zhou (2017). "Gated Self-Matching Networks for Reading Comprehension and Question Answering." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1, pp. 189–198. DOI: 10.18653/v1/P17-1018.

Wang, X. and G. Sukthankar (2013). "Multi-label relational neighbor classification using social context features." In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 464–472.

Wang, X., H. Fang, and C. Zhai (2008). "A Study of Methods for Negative Relevance Feedback." In:

Wang, X. and A. McCallum (Aug. 2006). "Topics over time: a non-Markov continuous-time model of topical trends." In: *the 12th ACM SIGKDD international conference*. New York, New York, USA: ACM, pp. 424–433. DOI: 10.1145/1150402.1150450.

Wang, Y., E. Agichtein, and M. Benzi (Aug. 2012). "TM-LDA: efficient online modeling of latent topic transitions in social media." In: *the 18th ACM SIGKDD international conference*, pp. 123–131. DOI: 10.1145/2339530.2339552.

Weimer, M., A. Karatzoglou, Q. V. Le, and A. Smola (2007). "CofiRank Maximum Margin Matrix Factorization for Collaborative Ranking." In: *Advances in Neural Information Processing Systems*, pp. 1–3.

Weston, J., S. Chopra, and A. Bordes (2015). "Memory Networks." In: *Proceedings of the International Conference on Learning Representations*.

Weston, J., B. Schölkopf, and G. H. Bakir (2004). "Learning to Find Pre-Images." In: *Advances in Neural Information Processing Systems*. Ed. by S. Thrun, L. K. Saul, and B. Schölkopf. Vol. 16. MIT Press, pp. 449–456.

Woodsend, K. and M. Lapata (2012). "Multiple Aspect Summarization Using Integer Linear Programming." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational NaturalLanguage Learning*, p. 11.

Xia, L., J. Xu, Y. Lan, J. Guo, W. Zeng, and X. Cheng (2017). "Adapting Markov Decision Process for Search Result Diversification." en. In: ACM Press, pp. 535–544. DOI: 10.1145/3077136.3080775.

Xiang, R. and J. Neville (2013). "Collective inference for network data with copula latent markov networks." In: *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, pp. 647–656.

Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio (2015). "Show, Attend and Tell - Neural Image Caption Generation with Visual Attention." In:

Xu, Y., L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin (2015). "Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths." In: *EMNLP*, pp. 1785–1794.

Yang, B., W.-t. Yih, X. He, J. Gao, and L. Deng (2014). "Learning Multi-Relational Semantics Using Neural-Embedding Models." In: *CoRR abs/1411.4072*.

Yang, J. and J. Leskovec (2011). "Patterns of temporal variation in online media." In: *WSDM*. ACM.

Yang, W., H. Zhang, and J. Lin (Mar. 26, 2019). "Simple Applications of BERT for Ad Hoc Document Retrieval." In: arXiv: 1903.10972 [cs].

Yang, Z., W. Cohen, and R. Salakhutdinov (2016). "Revisiting Semi-Supervised Learning with Graph Embeddings." In: *ICML 2016*.

Yao, L., A. Haghighi, S. Riedel, and A. McCallum (2011). "Structured Relation Discovery using Generative Models." In:

Yao, L., S. Riedel, and A. McCallum (July 2012). "Unsupervised Relation Discovery with Sense Disambiguation." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2012. Jeju Island, Korea: Association for Computational Linguistics, pp. 712–720.

Yatskar, M., V. Ordonez, and A. Farhadi (2016). "Stating the Obvious - Extracting Visual Common Sense Knowledge." In:

Ye, J. and L. Akoglu (2015). "Robust Semi-Supervised Classification for Multi-Relational Graphs." In: *arXiv preprint arXiv:1510.06024*.

Yin, W., H. Schütze, B. Xiang, and B. Zhou (Dec. 2016). "ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs." In: *Transactions of the Association for Computational Linguistics* 4, pp. 259–272. DOI: 10.1162/tacl_a_00097.

Yu, L., W. Zhang, J. Wang, and Y. Yu (Sept. 18, 2016). "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient." In: *arXiv:1609.05473 [cs]*. arXiv: 1609.05473.

Zablocki, E., P. Bordes, L. Soulier, B. Piwowarski, and P. Gallinari (May 2019a). "Context-Aware Zero-Shot Learning for Object Recognition." en. In: *International Conference on Machine Learning*, pp. 7292–7303.

— (2019b). "Incorporating Visual Semantics into Sentence Representations within a Grounded Space." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Zablocki, É., B. Piwowarski, L. Soulier, and P. Gallinari (Feb. 2018). "Learning Multi-Modal Word Representation Grounded in Visual Context." en. In: *Proceedings of the Association for the Advancement of Artificial Intelligence*.

Zeiler, M. D. and R. Fergus (2014). "Visualizing and Understanding Convolutional Networks." In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pp. 818–833. DOI: 10.1007/978-3-319-10590-1_53.

Zeng, D., K. Liu, Y. Chen, and J. Zhao (2015). "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks." In:

Zhang, F., N. Yuan, D. Lian, X. Xie, and W.-Y. Ma (2016). "Collaborative Knowledge Base Embedding for Recommender Systems." In: *KDD*. DOI: 10.1145/2939672.2939673.

Zhang, W., Y. Yu, and B. Zhou (2015). "Structured Memory for Neural Turing Machines." In:

Zhang, Y., K. Liu, S. He, G. Ji, Z. Liu, H. Wu, and J. Zhao (June 2016). "Question Answering over Knowledge Base with Neural Attention Combining Global Knowledge Information." In: *arXiv:1606.00979 [cs]*.

Zhang, Z. (2004). "Weakly-supervised Relation Classification for Information Extraction." In: *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. CIKM '04. New York, NY, USA: ACM, pp. 581–588. DOI: 10/fpsvqr.

Zhao, J. and Y. Yun (2009). "A proximity language model for information retrieval." In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*. the 32nd international ACM SIGIR conference. Boston, MA, USA: ACM Press, p. 291. DOI: 10.1145/1571941.1571993.

Zhou, D., O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf (2003). "Learning with Local and Global Consistency." In: *Proceedings of the 16th International Conference on Neural Information Processing Systems*. NIPS'03. Cambridge, MA, USA: MIT Press, pp. 321–328.

Zhou, D., J. Huang, and B. Schölkopf (2005). "Learning from labeled and unlabeled data on a directed graph." In: *Proc. of the 22nd intern. conf. on Machine learning*. ICML '05. Bonn, Germany: ACM, pp. 1036–1043. DOI: http://doi.acm.org/10.1145/1102351.1102482.

Zhou, Q., N. Yang, F. Wei, and M. Zhou (2017). "Selective Encoding for Abstractive Sentence Summarization." In: *arXiv:1704.07073*.

Zhou, Y. and L. Liu (2014). "Activity-edge centric multi-label classification for mining heterogeneous information networks." In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1276–1285.

Zhu, X. and Z. Ghahramani (2002). *Learning from labeled and unlabeled data with label propagation*. Tech. rep. Citeseer.

Zhu, Y., O. Groth, M. Bernstein, and L. Fei-Fei (2015). "Visual7W: Grounded Question Answering in Images." In:

Zhu, Y., C. Zhang, C. Re, and L. Fei-Fei (2015). "Building a Large-scale Multimodal Knowledge Base for Visual Question Answering." In:

Ziat, A., L. Denoyer, B. Piwowarski, P. Gallinari, and A. Ziat (2016). "Modeling Relational Time Series Using Gaussian Embeddings." In: *NIPS Time Series Workshop*.

Zuccon, G., L. Azzopardi, and K. Rijsbergen (2009). "The Quantum Probability Ranking Principle for Information Retrieval." In: *ECIR*.

Zuccon, G., L. A. Azzopardi, and C. van Rijsbergen (2009). "Semantic Spaces: Measuring the Distance between Different Subspaces." In: *Quantum Interaction*. Ed. by P. Bruza, D. Sofge, W. Lawless, K. van Rijsbergen, and M. Klusch. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 225–236. DOI: 10/dgw69v.

Zuccon, G., B. Piwowarski, and L. Azzopardi (2011). "On the use of Complex Numbers in Quantum Models for Information Retrieval." In: pp. 346–350. DOI: 10.1007/978-3-642-23318-0_36.