From Inter-Annotation to Intra-Publication Inconsistency

Alaa Abi Haidar, Mihnea Tufis, Jean-Gabriel Ganascia

▶ **To cite this version:**

Alaa Abi Haidar, Mihnea Tufis, Jean-Gabriel Ganascia. From Inter-Annotation to Intra-Publication Inconsistency. Inconsistency Robustness, 52, College Publications, pp.614, 2015, Studies in Logic, 978-1-84890-159-9. hal-01245137

HAL Id: hal-01245137

https://hal.sorbonne-universite.fr/hal-01245137v1

Submitted on 16 Mar 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

From Inter-Annotation to Intra-Publication Inconsistency

Alaa Abi Haidar (Université Pierre et Marie Curie Paris 6)
Mihnea Tufiș (Université Pierre et Marie Curie Paris 6)

What are effective ways to help people with chronic illness, e.g. diabetes and heart disease? Computational linguistics relies on human-annotated data to train machine learners. Inconsistency among the human annotators must be carefully managed (otherwise, the annotations are useless in computation). How can this annotation process be made scalable?

ABSTRACT

Curing chronic illnesses and diseases requires the huge effort of collecting all available information on this matter and piecing it together with the aids of mathematical and computer modeling. Both phases of information collection and piecing together are prone to error. Errors may result from human annotation inconsistency, machine learning and parameterization when using supervised learning. On a different scale, published results that need to be collected may suffer from another kind of disagreement either due to varying experimental methodologies or assumptions. Here, we discuss these inconsistencies and disagreements in scientific literature and we investigate those of the inter-annotation of named entities in bioliterature from empirical perspectives.

INTRODUCTION

Mathematical modeling and simulation help us understand the underlying mechanisms behind complex and barely understood systems, such as immune systems in order to advance biomedical and drug studies and cure diseases [10]. However, mathematical modeling requires huge amounts of parametric data, usually published in experimental and theoretical manuscripts and dispersed in the scientific literature. Pubmed comprises more than 19 million scientific articles [9] and this amount is growing at astounding rates. The manual extraction of valuable information and their classification into predefined labels, such as parametric

values, units and species names, is very costly and inefficient. Hence, we use text mining, and more specifically, named entity recognition [8], in order to automatically and accurately extract and classify numerical and textual entities, that can later be plugged into mathematical models and simulated.

Nevertheless, a significant number of parametric values describing experimental results in the bio-literature are inconsistent due to variations in experimental approaches or imperfection in the experiments [19]. For example, the amount of T cells that mature from the thymus through the process of negative selection[1] remains a huge debate [1-7] and may vary between "less than 5%" to "10%". Our study focuses on analysing all reported rates for such processes in biological and other complex systems in order to study their statistical variations while identifying average values and outliers.

BACKGROUND

Most techniques for named entity recognition rely on supervised machine learning that require human-annotated training data. Studies have shown that there is at least 25% human inter-annotation disagreement (inconsistency) when annotating biological named entities [24]. According to the G-theory [20], sources of inconsistency might be of external influences like alterations in the tools used for annotations, increasing time pressure, removal or adding of rewards, or changes in the annotation scheme. A study reports 55% and 82% F1 scores for exact and relaxed[2] inter-annotation agreement respectively when the task was to extract interactions between enzymes and marine drugs in over 230 full-text articles [11]. In another study, the relaxed inter-annotator agreement showed that 94% of the time curators were precisely extracting GO[3] annotation from the literature and 72% of the time curators recalled all possible valid GO terms from the text [14]. Yet another study reports an inter-annotator agreement rate of over 60% for triggers and of over 80% for arguments using an exact match constraint [16]. More recently, a study reported inter-annotator agreement (IAA) F-measures for medication names and medication types, 94.2% and 88.2% respectively [13]. However, the rates of inter-annotation agreement vary not only from study to study but also from domain to domain. For instance, Wiebe et al. report 82.0 F1

[1] T cells that recognize self antigens are eliminated in the thymus through a process known as negative selection so that they do not bind to self and cause auto-immune diseases

[2] Unlike exact matches, relaxed matches may span over less or more words to describe the same concept.

[3] The Gene Ontology, or GO, is a bioinformatics attempt to unify the representation of genes and gene product attributes across various species

score of human annotation agreement for opinion expression [12].

Consequently, the human inter-annotation inconsistency creates a gray area of uncertainty that the machine learner depends on to create fine-tuned rules and exceptions. Furthermore, human annotation is often used as the gold standard for evaluating machine learning methods [15] and therefore it is very important to have as few disagreements as possible. Nevertheless, the human inter-annotation disagreement can be reduced by using strictly agreed-upon annotations, the reasonings of a single annotator, majority rules or by identifying mislabeled annotations [17, 18].

RESULTS AND DISCUSSION

Here, we attempt to quantify human annotation inconsistency based on a biological article annotated by several annotators. More specifically, we study the annotation inconsistency of a biomedical article [24] that is annotated by three experts for 9 categories of named entities. The following table lists the number of annotated entities for each of the 9 categories by each of the three annotators:

almeida_annotation_Al.tag	almeida_annotation_Veronique.tag	almeida_annotation_Florence.tag
17 UNIT	17 UNIT	28 LOCATION
22 LOCATION	20 LOCATION	30 UNIT
34 NUM	65 NUM	38 INDIVIDUAL
63 INDIVIDUAL	74 INDIVIDUAL	58 NUM
227 METHOD	226 METHOD	161 METHOD
255 COFACTOR	231 COFACTOR	356 COFACTOR
344 PROCESS	360 PROCESS	454 PROCESS
950 POPULATION	907 POPULATION	904 POPULATION
8140 O	8152 O	8014 O

Table 1. The number of annotated entities in a biomedical article [24] for each of the 9 categories by each of the three annotators

The categories describe biological concepts. In the following example “The T cell proliferation is 0.4 cells/hr” can be annotated as follows: “T cell proliferation” describes a PROCESS, “0.4” a NUM (numerical value) and “cells/hr” a UNIT. More complex entities are harder to classify into concepts which may create annotation inconsistencies between several annotators.

For the example at hand, we identify inconsistencies varying from 1.5% to 8.3% out of a total of 10052 terms as shown in table 2. Our average values are below those reported by [24]. However, that might be due to the fact of working on a different dataset and with different entities.

	Al	Ver	Flo
Al	0%	1.5%	7.8%
Ver		0%	8.3%
Flo			0%

Table 2. Inter-annotator inconsistencies varying from 1.5% to 8.3% out of a total of 10052 terms

Next, we study to what extent inter-annotation inconsistency in training data can influence the robustness of machine learning and the predicted results. We hypothesize that factors such as the size of annotated training data, the number of annotators, the number of class labels, and the over-fitness of supervised machine learners play major roles in the robustness of the learning and the classification results. We attempt to answer these questions using empirical approaches inspired by [21] a study of robustness when classifying noisy land rover data from satellite images.

EVALUATING INTER ANNOTATOR AGREEMENT

Determining inter annotator agreement (IAA) can be a daunting task for reasons such as proper choice of agreement coefficients and lack of consensus on the interpretation of such coefficients [26].

Taking into account the recommendations of Artstein and Poesio [26 - 590] considering the better quality measure given by chance corrected coefficients as compared to simple percentage agreement, we have performed the reliability testing over the annotations performed over our corpus of text. Given that we are in a scenario of multi-annotators (namely, three) using a nominal variable, we will discuss the results we've obtained for the computation of the adapted versions of Cohen's κ and Fleiss' multi-π as well as for Krippendorff's α.

As a reminder, the first 2 coefficients above are based on the basic coefficients used in 2-annotators scenarios: Scott's π (1955), Cohen's κ (1960).

$$\pi,\ k = \frac{A_o - A_e}{1 - A_e},$$

where A_o is the observed agreement and A_e is the expected agreement. [26 - 559]

As explained in [26 - 560], the difference between π and κ lies in the assumptions made to compute the probability for a coder (annotator) to categorize an utterance in a certain category.

The indices we used are the generalizations of Scott's π and Cohen's κ made by Fleiss (1971) and Davies and Fleiss (1982) respectively.
To generalize, Fleiss' multi-π uses a different interpretation of the observed annotation A_o, namely the pairwise agreement, which is the number of pairs agreeing on an utterance out of the total number of pairs of coders.
Equally, multi-κ involves the computation of the expected agreement A_e based on individual coder marginals.

Finally, Krippendorff's α is a versatile coefficient which addresses the limitations of (multi-)κ and (multi-)π regarding the equal treatment of all disagreements [26 - 564]

$$\alpha = 1 - \frac{D_o}{D_e}$$

, where D_o is the observed disagreement and D_e is the expected disagreement.
[26 - 565, 566]

We computed the coefficients in the following 2 situations [27, 28]:

1. Full annotations, taking into account the large number of utterances classified as *O(ther)* by all annotators (Coders = 3, Utterances = 10043).

Average Pairwise Agreement [%]	Pairwise (1-2) Agreement	Pairwise (1-3) Agreement	Pairwise (2-3) Agreement
94.162%	98.457%	92.283%	91.745%

Table 3a. Average Pairwise Agreement [%]

Avg. Pairwise CK	Pairwise (1-2) CK	Pairwise (1-3) CK	Pairwise (2-3) CK
0.828	0.953	0.774	0.758

Table 3b. Average Pairwise Multi-κ (based on Cohen's κ)

multi-π	Ao	Ae
0.827	0.942	0.662

Table 3c. multi-π (based on Scott's π)

Krippendorff's α	No. of decisions
0.827	30129

Table 3d. Krippendorff's α

Looking at the "classical" agreement reporting, the average pairwise percentage agreement is at 94.2%, with the best agreement rate being between coder 1 and coder 2 at 98.5%.

As expected, the values for multi-π and average pairwise multi-κ are approximately equal and since all disagreements are treated equally, Krippendorff's α is also nearly equal to the two before. The approximate value of 0.82 for multi-κ, classifies the annotation process as "perfect" according to the strength scale given by Landis and Koch (1977) [26 - 576].

2. Altered annotations, discarding the utterances classified as *O(ther)* by all annotators (Coders = 3, Utterances = 2349).

Average Pairwise Agreement [%]	Pairwise (1-2) Agreement	Pairwise (1-3) Agreement	Pairwise (2-3) Agreement
75.039%	93.401%	67.007%	64.708%

Table 4a. Average Pairwise Agreement [%]

Avg. Pairwise CK	Pairwise (1-2) CK	Pairwise (1-3) CK	Pairwise (2-3) CK
0.675	0.914	0.569	0.543

Table 4b. Average Pairwise Multi-κ (based on Cohen's κ)

multi-π	Ao	Ae
0.674	0.75	0.233

Table 4c. multi-π (based on Scott's π)

Krippendorff's α	No. of decisions
0.674	7047

Table 4d. Krippendorff's α

In this case, almost 8000 utterances which have been all annotated as *Other* by all three coders were completely discarded from the reliability study.

The classic pairwise percentage agreement thus records a drop of the average value (75%); however, it is interesting to notice that the agreement between coders 1 and 2 stays as high as in the original situation (93.4%) which indicates a very high agreement for their annotations on the "main" classes (excepting the all-*O(ther)*) as well.

As before, the values of the three coefficients are (not surprisingly) almost equal, but are dropping this time, somewhere around 0.67, which on the strength scale of Landis and Koch (1977) [26 - 576] indicates only a "substantial" agreement in the annotation process, making the annotation suitable for “tentative conclusions”. Interesting enough, this is exactly the value which was set as the original threshold by Krippendorff (although he referred to it as "highly tentative and cautious") before he reviewed it and later set it at 0.8.

CONCLUSION

In this manuscript, we identify and discuss two forms of inconsistencies, one in published parametric results that we are studying in collaboration with immunologists[4], and another in manually human annotated data for machine learning. Both forms of inconsistencies influence the accuracy of biomedical research to a significant extent that we are interested in quantifying in our research. We expect our study to shed a light on both forms of inconsistencies, ones resulting from human inter-annotation and those published in biological literature.

ACKNOWLEDGEMENT

"This work has been done within the LABEX OBVIL project, and received financial state aid managed by the Agence Nationale de la Recherche, as part of the programme

[4] The team of Integrative Immunology (I2) is based at the Pitié Salpêtrière. 83, Bld de l'hôpital. 75013 Paris, France

"Investissements d'avenir" under the reference ANR-11-IDEX-0004-02". We thank the expertise and effort of Veronique THOMAS-VASLIN and Florence GIESEN for the help with the annotation and valuable discussions.

BIBLIOGRAPHY

1. Faro, J. & Velasco, S. González-Fernández, Á., & Bandeira, A. (2004). *The Journal of Immunology*, *172*(4), 2247-2255.
2. Starr, T. K., Jameson, S. C., & Hogquist, K. A. (2003). *Annual review of immunology*, *21*(1), 139-176.
3. Egerton, M., Scollay, R. & Shortman, K. (1990) Proc. ntnl. Acad. Sci., U.S.A. 87, 2579–2582.
4. Huesmann, M., Scott, B., Kisielow, P. & von Boehmer, H. (1991) Cell 66, 533–540.
5. Scollay, R. & Godfrey, D.I. (1995) Immun. Today 16, 268–273.
6. Surh, C.D. & Sprent, J. (1994) Nature 372, 100–103.
7. Bevan, M.J. (1977) Nature 269, 417–418.
8. Nadeau, D., & Sekine, S. (2007). *Lingvisticae Investigationes*, *30*(1), 3-26.
9. Eliot, T. S. (2011).. *Evidence-Based Public Health*, 158.
10. Kitano, H. (2002). *Science*, *295*(5560), 1662-1664.
11. Rafal Rak, Andrew Rowley, William Black, Sophia Ananiadou (2012). Database (Oxford) . doi: 10.1093/database/bas010
12. J. Wiebe and T. Wilson and C. Cardie (2005).. In Language Resources and Evaluation, volume 39, issue 2-3.
13. Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., & Solti, I. (2013). *Journal of medical Internet research*, *15*(4).
14. Camon, E. B., Barrell, D. G., Dimmer, E. C., Lee, V., Magrane, M., Maslen, J., ... & Apweiler, R. (2005). *BMC bioinformatics*, *6*(Suppl 1), S17.
15. Gaudan S, Jimeno Yepes A, Lee V, Rebholz-Schuhmann D (2008). EURASIP journal on bioinformatics & systems biology.
16. Mihaila, C., Ohta, T., Pyysalo, S., & Ananiadou, S. (2013). *BMC bioinformatics*, *14*(1), 2.
17. Brodley, C. E., & Friedl, M. A. (2011). *arXiv preprint arXiv:1106.0219*.
18. Guan, D., Yuan, W., Lee, Y. K., & Lee, S. (2011).. *Applied Intelligence*, *35*(3), 345-358.
19. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, et al. (2000). *Bioinformatics* 16: 906–914
20. Bayerl, P. S., & Paul, K. I. (2007). *Computational Linguistics*, *33*(1), 3-8.
21. DeFries, R. S., & Chan, J. C. W. (2000). *Remote Sensing of Environment*, *74*(3), 503-515.
22. Valiant, L. G. (1984).. *Communications of the ACM*, *27*(11), 1134-1142.
23. Kearns, M. J., & Vazirani, U. V. (1994). The MIT Press.
24. Van Mulligen, Erik M., et al. *Journal of biomedical informatics* 45.5 (2012): 879-884.
25. Almeida, Afonso RM, et al. *Frontiers in immunology* 3 (2012).
26. Artstein, R., & Poesio, M.(2008). *Computational Linguistics*, 34(4), 555-596.
27. Freelon, D. *ReCal3: Reliability for 3+ Coders*. [Online] Available: http://dfreelon.org/utils/recalfront/recal3/ [2013, March 6]
28. Freelon, D. (2010) *International Journal of Internet Science*, 5(1), 20-33