Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing

Mayla Boguslava, Kevin Bretonnel Cohena

**a** Computational Bioscience Program, University Colorado School of Medicine, Aurora, CO, USA

Abstract

To classify texts in natural language processing, we compute the agreement between annotators – do annotators classify texts the same? It’s often thought that the agreement between annotators is the upper limit on system performance: if humans can’t agree with each other about the classification more than some percentage of the time, then we don’t expect a computer to do any better. We trace the logical positivist roots of the motivation for measuring inter-annotator agreement, trace the origins of the widely-held belief about the relationship between inter-annotator agreement and system performance, and present data on 6 articles that suggest that inter-annotator agreement is not in fact an upper bound, with evidence from the biomedical and general domains. Further, we found a significantly positive correlation between inter-annotator agreement and system performance.

Keywords: Natural language processing L01.224.050.375.580; Supervised machine learning G17.035.250.500.500; Evaluation Studies N05.715.360.335

Introduction

Machine learning has become one of the dominant paradigms in natural language processing both in the biomedical and general domain [9]. Within machine learning, the dominant paradigm is supervised machine learning: the use of labeled data to train a classifier, which is then tested on more labeled data. The data is a fundamental ingredient in the system, and it has become standard practice to measure its quality by assessing the extent to which humans agreed with each other in the task of producing it [17]. This agreement is called inter-annotator agreement (IAA) [2], and it is measurable by a variety of metrics, including F1 measure, precision, and recall {Anderson, 1992 #23}[10]. A standard assumption in the field is that the inter-annotator agreement (IAA) establishes an upper bound on system performance [7,11,13-15,19]. In fact, although we know a fair amount about IAA in theory, the assumption that it is an upper bound on system performance turns out to be just that—a heretofore-untested assumption. The goal of the work reported here is to test that assumption. We do so by collecting data on IAA and machine performance of the same task, then building simple statistical models of their relationships. This is important to do because if the assumption turns out not to be supported, then as a community, we may be mis-estimating the actual performance of our systems. In particular, we may be over-estimating the quality of their performance by under-estimating how good it could be.

Background

Jacob Cohen, focusing on clinical-social-personality areas of psychology, quantified the reproducibility and reliability of categorization as having two or more judges independently categorize a sample and then determine the degree, significance, and sampling stability of the agreement. This measure is known as Cohen's Kappa and has been extended to annotation in general (inter-annotator agreement or IAA):

With *Pr(a)* as the relative observed agreement between two annotators and *Pr(e)* as the expected agreement between the annotators if each annotator randomly picked a category for each annotation. Thus Cohen’s Kappa adjusts for chance to determine how much better the annotators did from chance [4]. Indeed, a higher κ means the annotation task is well-defined and reproducible by other annotators.

At the same time, a high IAA score simply means the annotators interpreted the instructions consistently in the same way, not that the annotations are necessarily correct. "Annotators are probably the most variable aspect of an annotation task" [17]. Typically we compare the IAA score to the precision, recall, or F1 measure of the computer's performance.

Precision and recall gained favor in the 1960s when researchers began to compare multiple systems [9]. Precision (P) is the proportion of actual correct answers to computer output correct answers (true positives). Recall (R) is the proportion of actual correct answers to computer output correct answers (true positives) plus answers that the computer deemed incorrect but are actually correct.

There is a trade-off between precision and recall and ideally we want to use both, thus there are various composite measures. The most common is the F1 measure, which is the harmonic mean of precision and recall. We can also weight precision or recall more by adding 𝛼.

Lastly, F1 is easy to calculate and interpret [9]. Most papers discussed here utilized F1 for comparison to the IAA [3,16,20,22,23], with one using precision [8].

Method

We are questioning a belief in the literature, so it would be good to verify what is asserted in the literature through a literature review and analysis.

We did an initial literature search to find the source of the belief that inter-annotator agreement is the upper bound on system performance. Then we sought to determine whether or not this really is a widely held belief in the community. Next we looked to see if we could find examples of when this is not the case. To perform a more thorough search, we hired an independent third-party research firm, asking them to research two questions. The full text of the questions and analysis structure can be found in Kevin Bretonnel Cohen’s GitHub [5]. Briefly, they were as follows:

1. We asked them to find citations for the claim that inter-annotator agreement is an upper bound on system performance in language processing.
2. We then asked them to find papers reporting results that exceeded inter-annotator agreement in both the biomedical and general domains.

The results of the literature search found a total of twelve articles: six articles claiming that human performance is the upper bound of computer performance, and six articles in the biomedical or general domains where at least one system outperformed the annotators (IAA) to the extent that we understand the agreement and performance metrics. There were twenty systems tested against the IAA within the six papers where the systems outperformed the IAA. We next collected the IAA and system performance measure for all systems within the twelve articles, if they calculated both. With this data, we wanted to understand whether there were any correlations, not simply noise, between IAA and system performance in three data sets: systems that outperformed the IAA, systems that did not, and both combined.

For all three datasets, we first analyzed the distributions of these data using the Shapiro-Wilk normality test [21] to determine if they were normally distributed. Next, since most were not normally distributed, we calculated Spearman’s correlation [12] between the IAA and system performance (generally F1 measure) in R with the cor.test package [18].

Results

Neither we, nor the literature search service found an authoritative citation of the assumption that the inter-annotator agreement (IAA) is the upper bound on computer annotation performance. However we found six papers that state the belief that the IAA is an upper bound (see table 1).

Table 1 – Assumption of IAA as upper bound in the literature

|  |  |
| --- | --- |
| **Paper** | **Quote** |
| Resnik and Lin [19] | “It is generally agreed that human inter-annotator agreement defines the upper limit on our ability to measure automated performance” |
| Gale, Church, and Yarowsky [7] | “An estimate of the upper bound is obtained by assuming that our ability to measure performance is largly limited by our ability to obtain reliable judgements from human informants” |
| Ormandjieva, Hussain, and Kosseim [14] | “A high level of precision should not be expected for this task [because] the average inter-annotator agreement [0.66 and 0.64] for conceptual understanding should be seen as upper bounds on the accuracy of any classifier” |
| Navigli [13] | “[Previous calculated inter-annotator agreement] numbers lead us to believe that a credible upper bound for unrestricted find-grained word sense disambiguation is around 70%, a figure that state-of-the-art automatic systems find it difficult to ouperform” |
| Meyer and Gurevych [11] | "We rather measured the agreement between our algorithm M and both the human annotators A and B. Besides the inter-annotator agreement A–B, which serves as an upper bound, we tried the naive baseline approach 0 that always chooses the first target word sense." |
| Padó and Lapata [15] | “We also compared our results to the upper bound given by the inter-annotator agreement on the calibration data set” |

Next we found six papers (four from the biomedical domain and two from the general domain) where at least one system outperformed the IAA (see table 2). Note that generally system performance was measured using F1 measure, so we will use those terms interchangeably from now on. Further some articles evaluated multiple systems.

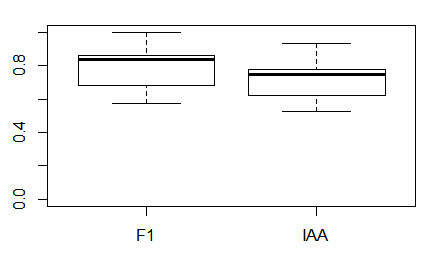
\*Note that all articles use F1 measure for system performance except for *Disambiguation of Occurrences of Reformulation Markers*, in which they used precision and explained how their results are comparable to previous ones [8].

Table 2 – Systems that outperform the IAA

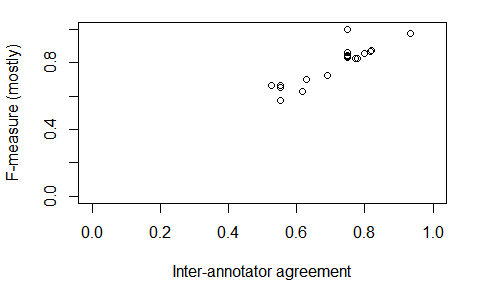
|  |  |
| --- | --- |
| **Paper** | **Systems that outperformed the IAA** |
| *Combining Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation [20]* | * Machine learning to recognize specific entities within clinical notes * Classifying intervention had lowest IAA and F1 ≥ IAA with multiple methods |
| *Disambiguation of Occurrences of Reformulation Markers [8]* | * Reformulation vs. non-reformulation in French with specific markers * ESLO1/2 (spoken scenarios): Precision\* > IAA |
| *SemEval-2015 Task 6: Clinical TempEval [3]* | * Multiple Systems compete to identify critical timeline components of clinical notes and pathology reports from the Mayo Clinic * **Adj-Ann**: IAA between adjudicator (final judge of the data to generally be used to train the system) and 2 annotators * Many systems F1 ≥ IAA and a few better than Adj-Ann (stronger) |
| *Automatically Detecting Acute Myocardial Infarction (AMI) Events from EHR Text: a Preliminary Study [23]* | * Automate the annotation of Worcester Heart Attack Study for AMI * F1 of system for ICD Diagnosis outperformed the IAA |
| *Deception Detection using Real-Life Trial Data [16]* | * Multimodal system detects deception using text/gesture modalities * System performance using decision trees significantly higher than annotator agreement and kappa statistic (0.01-0.20) * Humans detect deception slightly above chance |
| *Automatic Classification of Lexical Stress Errors for German CAPT (Computer-Assisted Pronunciation Training) [22]* | * Classify non-native German lexical stress errors from manually annotated corpus of German word utterances by native French speakers * IAA = relatively low with a fair agreement overall * System outperformed IAA: low agreement may be a signal that some annotators are not very reliable or task is difficult for humans |

With the articles mentioned above using Shapiro-Wilk normality test [21], we found that only the system performance measure is normally distributed, whereas IAA is not and instead skewed right. We then looked at the distribution of the data (see figure 1). We expect to see the F1 measure higher than IAA since we chose those articles only. Thus to look for correlation, we used a Spearman correlation since our data is not normally distributed.

We calculated Spearman’s correlation for both IAA versus F1 measure (rho = 0.807, p-value = 8.56 X 10-6). We found that IAA and F1 measure are significantly positively correlated (p-value < 0.05) (see figure 2).



*Figure 1 – Boxplot of F1 measure (system performance) and IAA (annotator agreement) for systems that outperform the IAA*

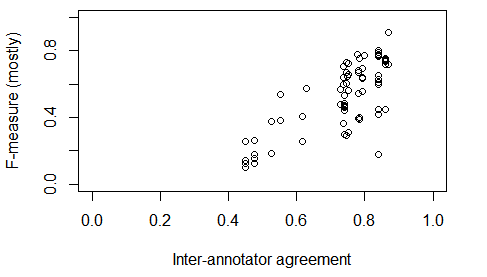


*Figure 2 – Positive correlation between F1 measure (system performance) and IAA (annotator agreement) for systems that outperform the IAA.*

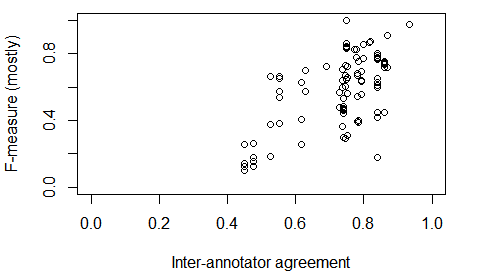
We then did the same analysis both for systems that do not beat the IAA and all systems together. We found that in systems that do not outperform the IAA that both IAA and F-measure are not normally distributed, but skew right (boxplot figure not shown). Furthermore, there was a signficant positive correlation (p-value < 0.05) between IAA and system performance (rho = 0.653, p-value = 1.449 x 10-11) (see figure 3).

Also, when we look at all the data combined for systems that did and did not outperform the IAA (see figure 4), IAA and system performance (rho = 0.513, p-value = 1.81 x 10-8) are significantly postively correlated (p-value < 0.05) but are less positively correlated compared to just systems that outperform the IAA.

We can also see how the medians are affected depending on which systems are included: those that outperform the IAA, those that do not, or both (see table 3).



*Figure 3 – Positive correlation between F1 measure (system performance) and IAA (annotator agreement) for systems that do not outperform the IAA.*

**

*Figure 4 – Positive correlation between the IAA and F1 measure for all data combined (both systems that do and do not outperform IAA).*

Table 3 – Medians of IAA, System performance, and the difference across all data analyses

|  |  |  |  |
| --- | --- | --- | --- |
| **Data** | **IAA Median** | **System performance Median** | **Difference Median** |
| System performance > IAA | 0.75 | 0.836 | 0.0785 |
| System performance < IAA | 0.7647 | 0.5865 | -0.1655 |
| All Data | 0.7504 | 0.6380 | -0.1383 |

Discussion

We must have tools to evaluate our research in order to make funding decisions and provide accurate information to the public about technology, specifically in the realm of annotation tasks in natural language processing. Here we discuss the limitations of the current way, using inter-annotator agreement, and suggest a stronger method, using adjudicator-annotators agreement.

All supervised classification tasks in natural language processing are predicated on a gold standard (supervised learning) which rely heavily on annotators and the scientists that created the rules for classification. Thus a commonly held belief in the literature is that a system can only classify as well as its annotators or scientists that determined the rules. This claims that humans are better than computers at natural language. In many annotation tasks, scientists enlist experts in linguistics and or the specific field of interest to ensure the texts are annotated and or the rules are defined carefully and correctly. This means that experts should be better than the computer system.

As far as we can tell though, based on two literature searches and the fact that no one ever cites one, **there is no authoritative citation** for the idea that IAA is the upper bound on system performance – in fact, it is just **an untested assumption**. It is a **widely-held assumption,** as shown by the fact that we did not have a problem finding multiple explicit statements of it.Further **it is not true** since we found multiple papers that contradicted the assumption. Thus the computer system can outperform the annotators. **We cannot assume the IAA is an upper bound on performance of annotation systems because we showed here it was false.**

In some of the cases where the system outperformed the annotators, the scientists questioned their annotators and the task itself [3,7,20,22,23]. We currently use the IAA to determine how well-defined a problem is (higher IAA is better defined), how difficult the problem is (lower IAA means more difficult), and as a measure of how well a computer system can perform in comparison to human annotators with the same rules. The scientists in these cases explain that they hope to have better annotators in the future, better define the problem, or the problem itself is difficult (low IAA). These are all interpretations of the IAA though. The higher the IAA only tells you that the annotators interpreted the rules consistenly in the same way, not that the annotators are correct or the rules are correct [17]. Thus the notion that the IAA is an upper bound of computer performance falls short because a computer system makes objective and calculated decisions whereas annotators may not always follow the rules given. Further, not all annotators had rules to follow and were not experts, simply everyday people [16].

We also wanted to understand if there is something unique about the systems that outperform the IAAs as compared to when they do not and both situations aggregated together. First, there were more situations where the IAA was higher than the system which seems to be an indication as to why scientists believe that the IAA was an upper bound. Second, in situations where the system outperformed the IAA, the signficiant positive correlation between IAA and system performance was higher than both the systems that did lower than the IAA and all sytems together. This seems to suggest that the higher the IAA, the better the system performs. The higher the IAA, the more likely the system will perform better because the rules are defined well and the task is not as difficult. At the same time, both the systems that did lower than the IAA and all sytems together also had significant positive correlations between IAA and system performance. Overall, it seems that the better the IAA, the better the system will perform.

The results presented here might give us some insight into the origins of the idea that IAA is an upper limit on system performance. As we see here, the higher the IAA the better the system performs, leading to the belief that the IAA determines how well-defined and difficult a task is. It is believed that the higher the IAA the more well-defined the task is because if there are clear defined rules, then we can simply input those into a computer and the computer should perform quite well. Also, the IAA defines how difficult the task is because we believe that the more the annotators agree on the classification, the easier the problem must be in that the rules capture most of the situations. This then leads us to believe that human annotation is the upper bound on computer annotation performance because we use the human annotated gold-standard texts to train the computer. The computer cannot be “smarter” than the human annotators that trained it. However this is a false belief to the extent that we understand the agreement metrics and the performance metrics, systems presented here outperformed the IAA.

This summons the question: how we measure system performance if IAA is no longer an upper bound? *The SemEval-2015 Task 6: Clinical TempEval* paper possibly provides an answer by suggesting a different type of annotator agreement, **Adj-Ann** from the results. The Adj-Ann calculated the IAA between the adjudicator and annotators. The adjudicator decides which answers are correct in order to either give the system sample correct answers to learn from or to assess how well the system did. The paper points out that usually the system is trained on adjudicated data, not annotated data. Thus a better upper bound on the system performance may be a combination of the adjudicated annotations and the annotators as suggested in this paper [3].

These results are relevant to the small but growing body of work on the ethics of NLP [6]. The Code of Ethics and Professional Conduct of the Association for Computing Machinery includes the imperative to share knowledge of the limitations of computer systems (ACM Code of Ethics and Professional Conduct 2.7) [1]. It is often suggested in the literature [7,11,13-15,19] that IAA is the upper bound on system performance, yet these results call that assumption into question. It is incumbent upon NLP researchers to be as well informed as possible on the way we discuss the relationship between IAA and achievable NLP performance.

The limitations of this work include the literature search and not finding all situations where the system outperforms the IAA and vice versa. More data could lead to other correlation results, but these are statistically significant. This work also did not differentiate between different measures of performance (precision, recall, F1 measure, and others). The majority of systems here used F1 measure, but not all. Future work could include gathering more systems that both outperform and do not outperform their IAA to have a greater depth of data, and differentiating between systems performance measures.

Conclusion

Overall, we determined that the inter-annotator agreement is not an upper bound of system performance showing examples from both the biomedical and general domain. We found significant positive correlations between the IAA and the system performance amongst all types of system performance (greater or lower than IAA). We further suggest a stronger measurement from one of the articles [3] that uses supervised learning that takes into account the initial annotated data (gold-standard data) the system learns on and outside annotators combined. This is not an upper bound however, as seen in the article [3]. Hopefully, this will better allow scientists to evaluate their research, make funding decisions, and provide accurate information to the public about technology.

Acknowledgements

Thank you to the Dean’s Fund at University of Colorado Anschutz Medical Campus for first year funding and to the Computational Bioscience Program there. Also thank you to the grants from Kevin Bretonnel Cohen and Larry Hunter: NIH LM008111, LM009254 and NSF IIS-1207592.

References

Uncategorized References

[1] R.E. Anderson, G. Engel, D. Gotterbarn, G.C. Hertlein, A. Hoffman, B. Jawer, D.G. Johnson, D.K. Lidtke, J.C. Little, D. Martin, D.B. Parker, J.A. Perrolle, and R.S. Rosenberg, ACM Code of Ethics and Professional Conduct, in, ACM, 1992.

[2] R. Artstein and M. Poesio, Inter-Coder Agreement for Computational Linguistics, *Association for Computational Linguistics* **34** (2008), 555-596.

[3] S. Bethard, L. Derczynski, G. Savova, J. Pustejovsky, and M. Verhagen, SemEval-2015 Task 6: Clinical TempEval, in: *SemEval 2015*, Denver, Colorado, 2015, pp. 806-814.

[4] J. Cohen, A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement* **20** (1960), 37-46.

[5] K.B. Cohen and M. Boguslav, KevinBretonnelCohen/InterAnnotatorAgreement in, GitHub Inc., GitHub, 2016.

[6] K. Fort, G. Adda, and K.B.C. (accepted), Éthique et traitement automatique des langues et de la parole: truismes et tabous, *Traitement Automatique des Langues* (2016).

[7] W. Gale, K.W. Church, and D. Yarowsky, Estimating upper and lower bounds on the performance of word-sense disambiguation programs, in: *annual meeting on Association for Computational Linguistics*, 1992.

[8] N. Grabar and I. Eshkol-Taravela, Disambiguation of occurrences of reformulation markers c’est-à-dire, disons, ça veut dire, *JADT 2016* **7** (2016).

[9] P. Jackson and I. Moulinier, *Natural language processing for online applications: text retrieval, extraction, and categorization*, John Benjamins Publishing Company, Amsterdam, 2007.

[10] A.A. Leenaars, *Sucide Notes: Predictive Clues and Patterns*, Human Sciences Press, Inc., New York, 1988.

[11] C.M. Meyer and I. Gurevych, Worth its weight in gold or yet another resource—A comparative study of Wiktionary, OpenThesaurus and GermaNet, in: *Computational linguistics and intelligent text processing*, Springer Berlin Heidelberg, 2010, pp. 38-49.

[12] M. Mukaka, A guide to appropriate use of Correlation co-efficient in medical research, *Malawi Medical Journal: The Journal of Medical Association of Malawi* **24** (2012), 69-71.

[13] R. Navigli, Meaningful clustering of senses helps boost word sense disambiguation performance, in: *International Conference on Computational Linguistics and the annual meeting of the Association for Computational Linguistics*, 2006.

[14] O. Ormandjieva, I. Hussain, and L. Kosseim, Toward a text classification system for the quality as-sessment of software requirements written in natural language, in: *Fourth international workshop on Software quality assurance: in conjunction with the 6th ESEC/FSE joint meeting*, ACM, 2007.

[15] S. Padó and M. Lapata, Cross-linguistic projection of role-semantic information, in: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, 2005, pp. 859-866.

[16] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, Deception Detection using Real-life Trial Data, in: *ICMI 2015*, New York, NY, USA,, 2015, pp. 59-66.

[17] J. Pustejovsky and A. Stubbs, *Natural language annotation for machine learning*, O'Reilly Media ;, Sebastopol, CA, 2013.

[18] R-core, cor.test, in, RDocumentation.

[19] P. Resnik and J. Lin, Evaluation of NLP Sys-tems, *The handbook of computational linguistics and natural language processing* **57** (2010).

[20] A. Roberts, R. Gaizauskas, M. Hepple, and Y. Guo, Combining Terminology Resources and Statistical Methods for Entity Recognition: An Evaluation. , *LREC: European Language Resources Association.* (2008), 2974-2980.

[21] S.S. Shapiro and M.B. Wilk, An Analysis of Variance Test for Normality (complete Samples), *Biometrika* **52** (1965), 591-611.

[22] A.S. Vakil, J. Trouvain, and L. Automatic classification of lexical stress errors for German CAPT. In Proceeding of the Workshop on Speech and Language Technology for Education (SLaTE), pp. 47-52., Automatic classification of lexical stress errors for German CAPT

in: *SLaTE*, Leipzig, 2015, pp. 47-52.

[23] J. Zheng, J. Yarzebski, B.P. Ramesh, R.J. Goldberg, and H. Yu, Automatically Detecting Acute Myocardial Infarction Events from EHR Text: A Preliminary Study, *AMIA Annu Symp Proc* **2014** (2014), 1286-1293.

Address for correspondence

Kevin.Cohen@gmail.com