MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-298

Inter-Annotator Agreement and the Upper Limit on Machine Performance: Evidence from Biomedical Natural Language Processing

Mayla Boguslav and Kevin Bretonnel Cohen

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

Abstract

Human-annotated data is a fundamental part of natural language processing system development and evaluation. The quality of that data is typically assessed by calculating the agreement between the annotators. It is widely assumed that this agreement between annotators is the upper limit on system performance in natural language processing: if humans can't agree with each other about the classification more than some percentage of the time, we don't expect a computer to do any better. We trace the logical positivist roots of the motivation for measuring inter-annotator agreement, demonstrate the prevalence of the widely-held assumption about the relationship between inter-annotator agreement and system performance, and present data that suggest that inter-annotator agreement is not, in fact, an upper bound on language processing system performance.

Keywords:

Natural Language Processing; Supervised Machine Learning; Evaluation Studies

Introduction

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]. In natural language processing, human-annotated data is the gold standard for most evaluation studies [2], and therefore it is crucial for understanding the limits of our work. It is standard practice to measure the quality of that data by assessing the extent to which humans agree with each other in the task of producing it [3]. This is called inter-annotator agreement (IAA) [4]. A standard assumption in the field is that the inter-annotator agreement establishes an upper bound on system performance [5-10]. In fact, 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 searching for the basis of that assumption; demonstrating that it is, in fact, a widely held assumption; and then collecting published findings in which system performance has exceeded the inter-annotator agreement and building simple statistical models of their relationships. This is important because if the assumption turns out not to be supported, 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 potentially be.

Background

The calculation of inter-annotator agreement (often known outside of corpus linguistics as *inter-rater agreement*) is motivated by the need to deal with the problem of subjectivity in judgments about things that are not observable with the senses, a classic case of this being semantics. Lenaars [11] traces its roots back to logical positivism, and Krippendorff brought it to linguistic data in particular in the context of content analysis [12]. Exactly how inter-annotator agreement should be calculated remains an open topic of discussion.

Cohen [13], focusing on research in psychology, proposed quantifying the reproducibility and reliability of categorization by calculating the agreement between two annotators and correcting it for the probability of agreement by chance. This measure is known as Cohen's Kappa:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

…where *Pr(a)* is the 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 than chance [13]. Typically, language processing researchers compare the IAA score to the F_1 measure obtained by the system, including all of the papers discussed here except for [14], which uses precision (positive predictive value).

The F_1 is the harmonic mean of precision (P) and recall (R) (sensitivity). It is calculated on the basis of the numbers of true positives, false positives, and false negatives in a system's output:

$$F_1 = \frac{2PR}{P+R}, F_\alpha = \frac{(\alpha+1)\mathrm{PR}}{\alpha\mathrm{P}+\mathrm{R}}$$

In the case of annotating linguistic data, it is often the case that the expected chance agreement (*Pr(e)* in the formula for kappa) is effectively zero, since there is no clear definition of what would count as a false positive, e.g. in the case of any task that requires the labelling of boundaries, such as in named entity recognition or any task involving scope (e.g. syntactic analysis). When this is the case, kappa is equivalent to F-measure, and this observation is the justification for their comparison here [15].

Methods

This paper approaches the assumption of inter-annotator agreement as the upper limit on system performance in three steps. First, we seek to answer the question of whether it is, indeed, a widely held assumption in the natural language processing community. Then, we try to find the source for this

assumption—the definitive citation. Finally, we look for counter-examples to the assumption; having found some, we do a statistical analysis of the papers that report results that contradict the assumption, on the rationale that if those results are just "noise," it should show up in the descriptive statistics.

Since we are trying to characterize the community's shared assumptions and to find the source of those assumptions, we took the literature as a proxy for those assumptions. We did a search for papers that explicitly assert that inter-annotator agreement is an upper bound on machine performance, and we looked for the sources that are cited in support of that claim. We carried out two separate searches. One was done by a natural language processing researcher. The other was done by a literature search service. We had worked with them in the past, and knew them to be quite competent in researching questions related to natural language processing. The full instructions given to the literature search service are available on this project's GitHub site [16]. Briefly, they were as follows:

1. Find papers that assert explicitly that inter-annotator agreement is an upper bound on system performance.
2. Identify the source citation for that assertion.

We then asked the literature search service to find examples of papers that reported inter-annotator agreement *and* results from a natural language processing system, such that the system performed higher than the inter-annotator agreement.

To search the full text of publications, the service used Google Scholar. Phrasal search for *inter-annotator agreement* and *F-measure* and proximity operators to find cases where they occur near each other were used to retrieve an initial set of around 100 papers. Those papers were then examined manually, and any papers in which the inter-annotator agreement was higher than system performance *or* there was no explicit discussion of the relationship between them were excluded. This resulted in a set of 6 papers that included data on 20 systems that outperformed the inter-annotator agreement.

We next extracted the IAA and system performance measure for all 20 systems described within those articles. To evaluate the possibility that these values were noise, rather than an actual finding, we used simple statistical models to test for structure in the relation between IAA and system performance in three data sets: the systems that outperformed the IAA, other systems that did not, and both combined. The reasoning here is that if the findings are noise, that should be reflected as random variation in the F-measure, the IAA, or both; on the other hand, if it is not just noise, that would be reflected by structured relation.

For all three datasets, we used the Shapiro-Wilk test [17] to determine if they were normally distributed. We calculated the correlation between IAA and F-measure, reasoning that if the papers that report outperforming IAA are just observing noise, there should be no relationship between them. Because most of the distributions were not normal, Spearman's correlation, a non-parametric test [18], was used to calculation the correlations. The details are available on the GitHub site [16].

Results

There is no citation that establishes inter-annotator agreement as an upper bound on system performance. Neither we nor a professional literature search service found an authoritative citation for the idea that inter-annotator agreement is the upper bound on language processing system performance. It is often asserted, but we have not found a cited source that establishes it to be the case. None of the papers that explicitly asserted the assumption cited a source for the assertion.

Nonetheless, explicit statements of the assumption in multiple papers demonstrate that this assumption is widespread. We give six explicit statements of the assumption, including in papers by some of the most prominent researchers in the field—see the quotes in Table 1.

Table 1 – Explicit statements of the asssumption of IAA as an upper bound in the natural language processing literature

Paper	Quote
Resnik and Lin [8]	"It is generally agreed that human inter-annotator agreement defines the upper limit on our ability to measure automated performance"
Gale, Church, and Yarowsky [10]	"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 [7]	"…the average inter-annotator agreement…should be seen as upper bounds on the accuracy of any classifier"
Navigli [6]	"[Inter-annotator agreement] numbers lead us to believe that a credible upper bound for unrestricted find-grained word sense disambiguation is…"
Meyer and Gurevych [5]	"Besides the inter-annotator agreement A–B, which serves as an upper bound…"
Padó and Lapata [9]	"..the upper bound given by the inter-annotator agreement on the calibration data set"

Six papers (four from the biomedical domain and two from the general domain) reported at least one system that outperformed the IAA (see Table 2), for a total of 20 systems. Note that generally system performance was measured using F_1 measure, so we will use those terms interchangeably. The small number of papers reflects the fact that this is not a commonly reported phenomenon. However, neither is it unattested—this was not just a single counter-example, and those six papers reported on 20 systems that outperformed the inter-annotator agreement.

*Note that all articles use F_1 for system performance except for [14], which uses precision.

Table 2 – Systems that outperform the IAA

Paper	Systems that outperformed the IAA
Combining Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation [19]	• Machine learning to recognize specific entities within clinical notes • Classifying intervention had lowest IAA and $F_1 \geq$ IAA with multiple methods
Disambiguation of Occurrences of Reformulation Markers [14]	• Reformulation vs. non-reformulation in French with specific markers • ESLO1/2 (spoken scenarios): Precision* > IAA
SemEval-2015 Task 6: Clinical TempEval [20]	• 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 $F_1 \geq$ IAA and a few better than Adj-Ann (stronger)

Paper	Systems that outperformed the IAA
Automatically Detecting Acute Myocardial Infarction (AMI) Events from EHR Text: a Preliminary Study [21]	• Automate the annotation of Worcester Heart Attack Study for AMI • F_1 of system for ICD Diagnosis outperformed the IAA
Deception Detection using Real-Life Trial Data [22]	• Deception detection • System performance using decision trees significantly higher than annotator agreement and kappa statistic (0.01-0.20) • Humans detect deception only slightly above chance
Automatic Classification of Lexical Stress Errors for German CAPT (Computer-Assisted Pronunciation Training) [23]	• Classify non-native German lexical stress errors from manually annotated corpus of German word utterances by native French speakers • IAA only fair

Figure 1 shows the F1 and inter-annotator agreement for the 20 systems. The Shapiro-Wilk normality test [17] showed that only the system performance measure is normally distributed. IAA is not, and skewed left. Therefore, we calculated the Spearman correlation, which is non-parametric. This showed that IAA and F_1 measure are significantly positively correlated (rho = 0.807, p-value = 8.56 X 10^{-6}) (see Figure 2).

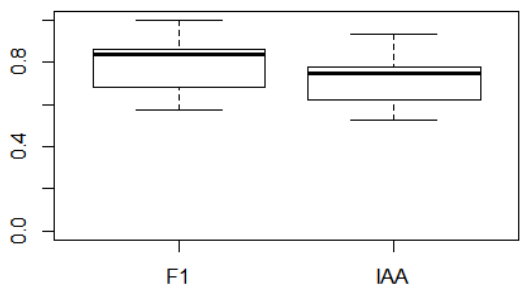

Figure 1 – Boxplot of F_1 measure (system performance) and IAA (annotator agreement) for systems that outperform the IAA

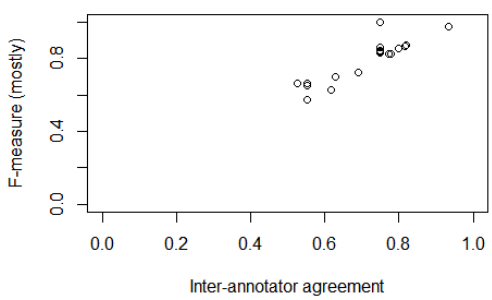

Figure 2 – Positive correlation between system performance and inter-annotator agreement for systems that outperform the IAA.

We then did the same analysis both for systems that did not beat the IAA, and for all systems together. In systems that did not outperform the IAA, neither IAA nor F-measure were normally distributed. There was a signficant positive correlation between IAA and system performance (rho = 0.653, p-value = 1.449 x 10^{-11}) (see Figure 3).

For the combined data combined for systems that did and did not outperform the IAA, the IAA and system performance were significantly postively correlated, but less so compared to only the systems that outperformed the IAA (rho = 0.513, p-value = 1.81 x 10^{-8}) (see Figure 4).

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).

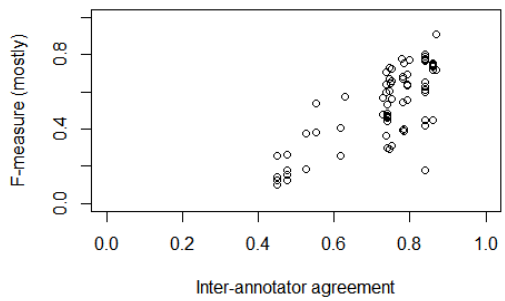

Figure 3 – Positive correlation between F_1 measure (system performance) and IAA (annotator agreement) for systems that do not outperform the IAA.

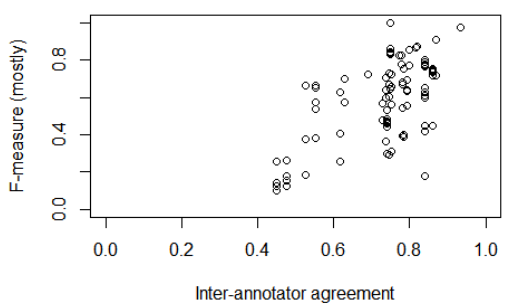

Figure 4 – Positive correlation between the IAA and F_1 measure for all data combined (both systems that do and do not outperform IAA).

Comparing the relationships between inter-annotator agreement and F-measure in the three sets of systems—ones which did outperform the IAA, ones which did not outperform the IAA, and the combination of those two, the relationships were the same—significantly positively correlated. This similarity across the three groups suggests that the cases of outperforming the inter-annotator agreement are not just noise.

Table 3 – Medians of IAA, system performance, and their difference across all data sets

Data	Median IAA	Median performance	Median difference
System > IAA	0.75	0.836	0.0785
System < IAA	0.7647	0.5865	-0.1655
All systems	0.7504	0.6380	-0.1383

Discussion

Providing accurate information to the public about technology and research results—as well as making funding decisions—requires the ability to accurately interpret measures of performance. The data presented here shows that a common standard for assessing natural language processing tools may overestimate their performance: contrary to a widely-shared and hitherto unexamined assumption in the field, the inter-annotator agreement is not necessarily the upper bound on performance in natural language processing. A lack of awareness of this can lead to the belief that systems are performing as well as they can, when in fact they are not.

Based on two literature searches and on 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 has been only an untested assumption. If there is some authoritative source that establishes this, not only have we not been able to find it, but also apparently no one else has, either, since no one cites one.

Despite being an untested assumption, it is nonetheless a widely-held assumption, as shown by the fact that we did not have any problem finding multiple explicit statements of it, some of them by the top people in our field.

Finally, the studies whose results are analyzed in this paper demonstrate that this widely-held assumption is not true. The distributional characteristics of the results—that is, their correlated, rather than unstructured nature—suggest that this is a real phenomenon, and not just noise. We cannot assume that the inter-annotator agreement is an upper bound on system performance, and in doing so, we may be over-stating how good natural language processing systems are.

In some of the papers that reported performance better than the inter-annotator agreement, the authors pointed out explicitly the unusualness of that finding. Some saw this as needing explanation, and they suggested explanations that were consistent with typical assumptions about inter-annotator agreement, such as that low inter-annotator agreement reflects a poor problem definition, an inherently difficult problem, poor guidelines, or—commonly—poor annotators [10; 20; 23-25].

If inter-annotator agreement does not establish the upper bound for system performance, what *should* we use in estimating the upper bound on system performance? Although a definitive answer to that question is outside of the scope of this paper, we discuss three possible solutions. They are based on changing what metric we use to quantify agreement; on changing who we define as the raters between whom the agreement is being calculated; and on replacing the agreement altogether with probabilistic estimates of a label quality.

One possibility is that we can safely use an inter-annotator agreement if we calculate it as something other than kappa. Although kappa is the most commonly reported measure of inter-annotator agreement, it has a number of problems. Some of these are essentially cultural—although there are a number of ways to calculate the expected chance agreement that is at the core of its claimed advantages, authors rarely report how they calculated the expected chance agreement. Consequently, it is often unclear what the kappa number actually reflects. When combined with the fact that the sensitivity of kappa to the probability of an estimated chance agreement is unstable—above an estimated chance agreement of about 0.5, kappa is extremely sensitive to small changes in the probability of chance agreement, while being relatively insensitive to small changes in the probability of a chance agreement below that value—it is clear that there are many reasons to be suspicious of reliance on this number.

While in the previous paragraph we have discussed calculating something other than kappa to characterize the inter-annotator agreement, Bethard et al. [20] suggest changing the definition of the raters, such that rather than calculating agreement between two annotators, we calculate agreement between an annotator and an adjudicator. This may provide an agreement value that is more reflective of the data on which the system will be trained and evaluated, since if adjudicated data is available, that is typically what is used for training and testing. However, a number of conditions must be met for this to be possible—at minimum, there has to be an adjudication step, which is not always the case. Furthermore, changing the definition of the raters between whom agreement is calculated does not answer the question of how to calculate the agreement between them.

Finally, Passoneau and Carpenter suggest abandoning an agreement entirely and building a probabilistic annotation model of label quality [26].

In the larger context of responsible conduct of science, the findings reported here are relevant to the small but growing body of work on the ethics of NLP [27-29]. As noted above, the ethical standards of the Association for Computing Machinery include the responsibility to communicate the *limitations* of computer systems [1]. In reporting performance, there is a common assumption that metrics that approach an inter-annotator agreement reflect *high* performance [5-10]. The data reported here suggest that such performance may not be as high as we think it is, relative to the best possible performance, suggesting that this assumption can lead—certainly inadvertently—to conduct that does not meet the Association for Computing Machinery standards.

Conclusion

This paper examines a common assumption in natural language processing. It is shown that the assumption is, indeed, widespread; that there is no established justification for that assumption; and that the assumption is not true. This last point is demonstrated both by multiple counterexamples, and by descriptive statistics that suggest that the counter examples are not random noise in the larger population of published papers on language processing, but rather reflect a real phenomenon. Responsible conduct of science will be enhanced by being aware of this.

Acknowledgements

Boguslav is supported by the Dean's Fund at University of Colorado Anschutz Medical. Cohen is supported by NIH grants LM008111, LM009254, and NSF IIS-1207592 to Lawrence E. Hunter, and by generous funding from Labex DigiCosme (project ANR11LABEX0045 DIGICOSME) operated by ANR as part of the program «Investissement d'Avenir» Idex ParisSaclay (ANR11 IDEX000302), as well as by a Jean d'Alembert fellowship. The work was aided by discussions with Patrick Paroubek, Bob Carpenter, and Tiffany Callahan; all remaining faults are the authors'.

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

[3] J. Pustejovsky and A. Stubbs, *Natural language annotation for machine learning*, O'Reilly Media ;, Sebastopol, CA, 2013.
[4] R. Artstein and M. Poesio, Inter-Coder Agreement for Computational Linguistics, *Association for Computational Linguistics* 34 (2008), 555-596.
[5] C.M. Meyer and I. Gurevych, Worth its weight in gold or yet another resource—A comparative study of Wiktionary, OpenThesaurus and GermaNet, in: *Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing*, Springer, Berlin, Heidelberg, 2010.
[6] 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*, Sydney, Australia, 2006.
[7] O. Ormandjieva, I. Hussain, and L. Kosseim, Toward a text classification system for the quality assessment of software requirements written in natural language, in: *Fourth International Workshop on Software Quality Assurance, SOQUA 2007, in conjunction with the 6th ESEC/FSE joint meeting*, Dubrovnik, Croatia, 2007.
[8] P. Resnik and J. Lin, Evaluation of NLP Systems, in: *The handbook of computational linguistics and natural language processing,* C.F. Alexander Clark, Shalom Lappin, ed., Wiley-Blackwell, 2010, pp. 271-295.
[9] S. Padó and M. Lapata, Cross-linguistic projection of role-semantic information, in: *HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Vancouver, British Columbia, Canada, 2005, pp. 859-866.
[10] W. Gale, K.W. Church, and D. Yarowsky, Estimating upper and lower bounds on the performance of word-sense disambiguation programs, in: *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 1992, pp. 249-256.
[11] A.A. Leenaars, *Sucide Notes: Predictive Clues and Patterns*, Human Sciences Press, Inc., New York, 1988.
[12] K. Krippendorff, *Content analysis : an introduction to its methodology*, Sage Publications, Beverly Hills, 1980.
[13] J. Cohen, A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement* 20 (1960), 37-46.
[14] N. Grabar and I. Eshkol-Taravela, Disambiguation of occurrences of reformulation markers c'est-à-dire, disons, ça veut dire, *JADT 2016* 7 (2016).
[15] G. Hripcsak and A.S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, *J Am Med Inform Assoc* 12 (2005), 296-298.
[16] K.B. Cohen and M. Boguslav, KevinBretonnelCohen/InterAnnotatorAgreement, in, GitHub, GitHub Inc., 2016.
[17] S.S. Shapiro and M.B. Wilk, An Analysis of Variance Test for Normality (complete Samples), *Biometrika* 52 (1965), 591-611.
[18] 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.
[19] 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.
[20] S. Bethard, L. Derczynski, G. Savova, J. Pustejovsky, and M. Verhagen, SemEval-2015 Task 6: Clinical TempEval, in: *SemEval 2015*, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 806-814.
[21] 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.
[22] 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.
[23] J.T. Anjana Sofia Vakil, Automatic classification of lexical stress errors for German CAPT, in: *Workshop on Speech and Language Technology in Education (SLaTE 2015)*, Leipzig, 2015, pp. 47-52.
[24] R.G. A. Roberts, M. Hepple, Y. Guo, Combining terminology resources and statistical methods for entity recognition: an evaluation, *LREC: European Language Resources Association* (2008), 2974-2980.
[25] 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, in: *AMIA Annual Symposium*, AMIA, 2014, pp. 1286-1293.
[26] R.J. Passonneau and B. Carpenter, The benefits of a model of annotation, *Transactions of the Association for Computational Linguistics 2* (2014), 311-326.
[27] K. Fort, G. Adda, and K.B. Cohen, Éthique et traitement automatique des langues et de la parole: truismes et tabous, *Traitement Automatique des Langues* 57 (2016), 7-19.
[28] D. Hovy, S. Spruit, M. Mitchell, E.M. Bender, M. Strube, and H. Wallach, Ethics in natural language processing, *Association for Computational Linguistics* (2017).
[29] D. Hovy and S.L. Spruit, The social impact of natural language processing, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

Address for correspondence

kevin.cohen@gmail.com