computational bioscience program
school of medicine at the university of colorado denver

December 9, 2016

Inter-Annotator Agreement and the Upper Bound on System Performance in Biomedical and General Domain Natural Language Processing

Authors: Mayla Boguslav* and Kevin Cohen*

Background and the Problem

Is categorization reproducible and reliable?

Jacob Cohen (clinical-social-personality areas of psychology) quantified reproducibility and reliability as having ≥2 independent judges categorize a sample and determine degree, significance, and sampling stability of their agreement.

Cohen's Kappa[1]: Pr(a) - probability that the 2 annotators agree, Pr(e) - probability that the annotators agree by chance.

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

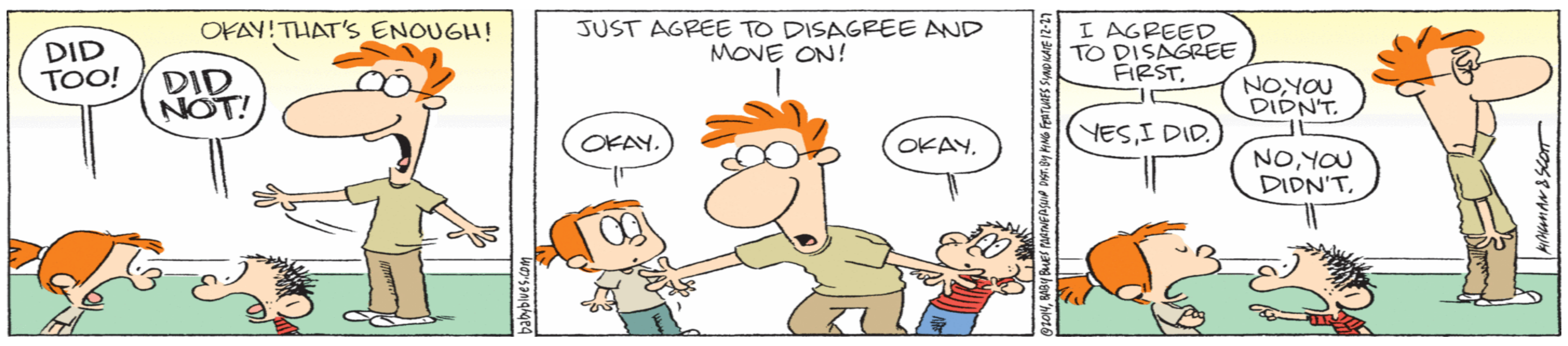
Natural Language Processing (NLP), specifically Machine Learning[2]:

We use data labeled by humans (annotators) with correct answers to train and validate categorization algorithms. We then compare the computer performance against the **Inter-Annotator Agreement (IAA)** score: how often do 2 annotators agree about the classification of the same text? Cohen's Kappa is one quantification of the IAA for 2 annotators (Fleiss's Kappa for more than 2). To compare to computer performance, use precision, recall, or F1-measure.

The Problem:[3-8]

It's often thought that the IAA (agreement between annotators) is the upper limit on how well a system can perform: if humans can't agree with each other about the classification more than some percentage of the time, then it's not reasonable to expect a computer to do any better. In fact though, the IAA is not an upper bound on system performance in NLP in both the biomedical and general domains.

Agreeing to Disagree[9]


DID TOO!
DID NOT!
OKAY! THAT'S ENOUGH!
babyblues.com
JUST AGREE TO DISAGREE AND MOVE ON!
OKAY.
OKAY.
©2014, BABY BLUES PARTNERSHIP DIST. BY KING FEATURES SYNDICATE 12-27
I AGREED TO DISAGREE FIRST.
YES, I DID.
NO, YOU DIDN'T.
NO, YOU DIDN'T.
KIRKMAN & SCOTT

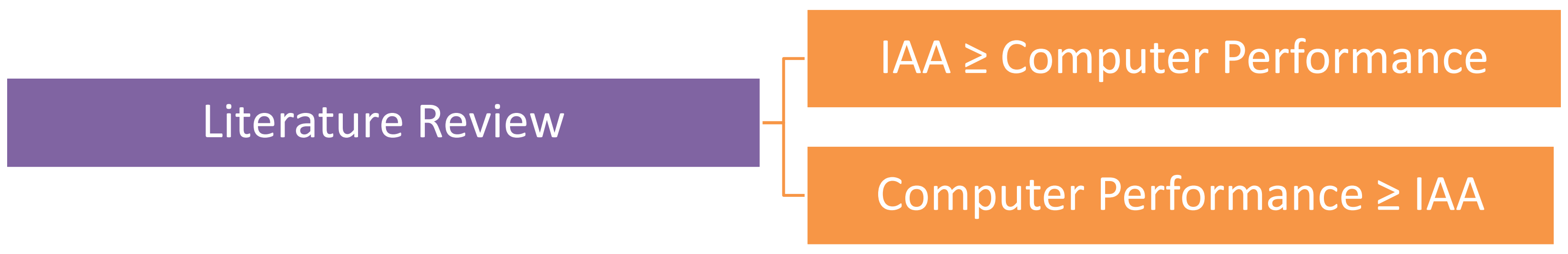
Methods and Materials

Methods:

If the assumption is true that IAA is the upper bound on computers, then there should not be findings of system performance higher than the IAA in the literature. We know about system performance to the extent that it gets published, so we did a literature review.

Materials:

Conference papers, Journal articles, natural language processing books


Literature Review

IAA ≥ Computer Performance

Computer Performance ≥ IAA

IAA Scores

Interpretation of IAA[2]:

A measure for how hard a problem is. High IAA score indicates the task is well-defined and other annotators will be able to continue the work (reproducible).

Note: Having a high IAA only means the annotators interpreted the instructions consistently in the same way, it does not mean that annotations are correct. In general, annotators are probably the most variable part of an annotation.

K	Agreement Level
<0	Poor
0.01-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Perfect

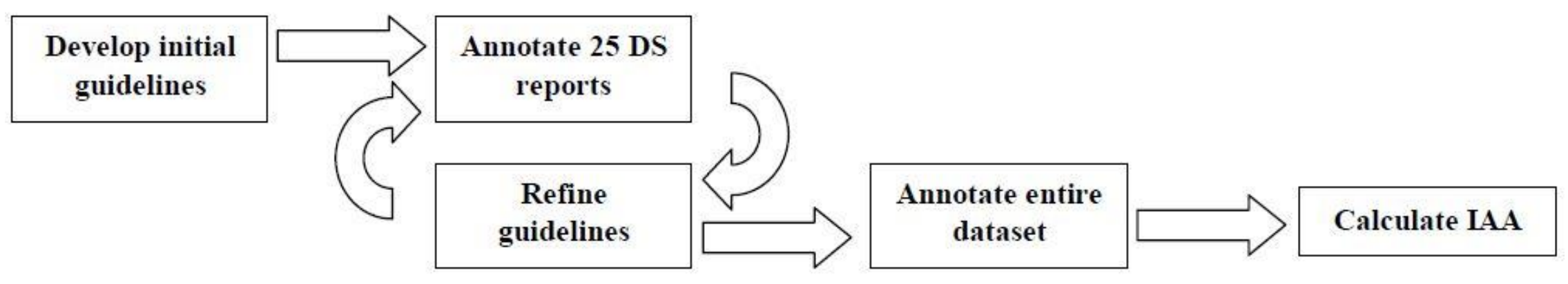
Examples of Computers Outperforming Human Annotators

Biomedical Examples:

- *Combining Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation*[10]
  a) Machine learning to recognize specific entities within clinical notes
  b) F1 measure vs. IAA: classifying intervention had lowest IAA and F1 ≥ IAA with multiple methods
- *Disambiguation of Occurrences of Reformulation Markers*[11]
  a) Reformulation vs. non-reformulation in French with specific markers
  b) 2 annotators for ESLO1/2 (spoken scenarios) and hypertension illness/life forum (hybrid of spoken and written language)

Corpus	Agreement (Cohen's Kappa)	Interpretation	Precision
ESLO1	0.617	Substantial	0.630
ESLO2	0.526	Moderate	0.664
Forum	0.784	Substantial	0.752

- *SemEval-2015 Task 6: Clinical TempEval*[12]
  a) Systems compete to identify critical timeline components of clinical notes and pathology reports from the Mayo Clinic
  b) **IAA1** between 2 independent annotators (Kappa statistic)
  c) **IAA2** between adjudicator and 2 annotators
  d) Many systems F1 ≥ IAA1 and a few better than IAA2 (stronger)
- *Automatically Detecting Acute Myocardial Infarction (AMI) Events from EHR Text: a Preliminary Study*[13]
  a) Automate the annotation of Worcester Heart Attack Study for AMI
  b) Annotation process: iterative and complex

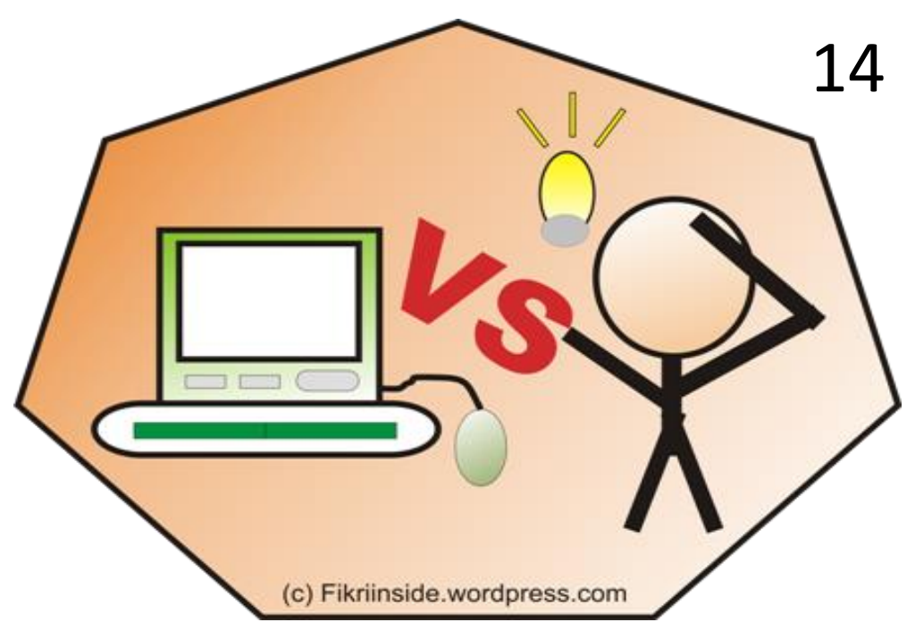


  c) F1 of system for ICD Diagnosis outperformed the IAA

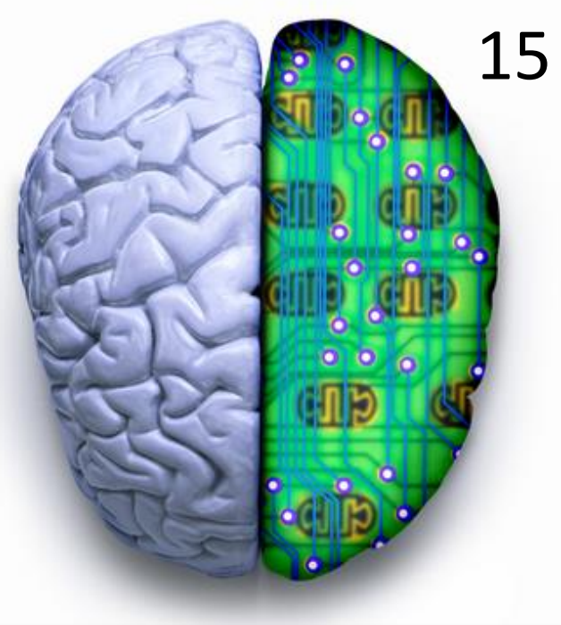
Conclusions

1. IAA is not an upper bound of system performance
2. Low IAA leads investigators to not trust annotators or provides evidence that the task is difficult for humans
3. IAA2 (adjudicator-annotators) may be a better upper bound since this includes the data analyzed

Computer ≥ IAA!



14



15

References

1. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20(1):37–46.
2. Pustejovsky J, Stubbs A. Natural language annotation for machine learning. Sebastopol, CA: O'Reilly Media; 2013.
3. Resnik, Philip, and Jimmy Lin. "Evaluation of NLP Systems." The handbook of computational linguistics and natural language processing 57 (2010): 271.
4. Gale, William, Kenneth Ward Church, and David Yarowsky. "Estimating upper and lower bounds on the performance of word-sense disambiguation programs." Proceedings of the 30th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1992.
5. Ormandjieva, Olga, Ishrar Hussain, and Leila Kosseim. "Toward a text classification system for the quality assessment of software requirements written in natural language." Fourth international workshop on Software quality assurance: in conjunction with the 6th ESEC/FSE joint meeting. ACM, 2007.
6. Navigli, Roberto. "Meaningful clustering of senses helps boost word sense disambiguation performance." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006.
7. Gurevych, Iryna, and Hendrik Niederlich. "Computing semantic relatedness in German with revised information content metrics." Proceedings of" OntoLex. 2005.
8. Meyer, Christian M, and Iryna Gurevych. "Worth its weight in gold or yet another resource—A comparative study of Wiktionary, OpenThesaurus and GermaNet." Computational linguistics and intelligent text processing. Springer Berlin Heidelberg, 2010. 38-49.
9. Kirkman R, Scott J. agree to disagree, arguing, bickering, Dad, Darryl, Hammie, siblings, Zoe [Internet]. Baby Blues. 2014 [cited 2016Nov30]. Available from: http://babyblues.com/comics/december-27-2014/
10. Roberts A, Gaizauskas RJ, Hepple M, Guo Y. 2008. Combining Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation. LREC: European Language Resources Association.
11. Grabar N, Eshkol-Taravela, I. 2016. Disambiguation of occurrences of reformulation markers c'est-à-dire, disons, ça veut dire. JADT 2016.7-10 June 2016, Nice, France.
12. Bethard S. et al. SemEval-2015 Task 6: Clinical TempEval. 2015. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado, June 4-5, 2015. pp. 806–814.
13. Zheng J. et al. Automatically detecting acute myocardial infarction events from EHR text: a preliminary study. 2014. In AMIA Annu Symp Proc. pp. 1286-93.
14. Fikriinside. Human vs Computer (which one has a better knowledge??) [Internet]. fikriinside – Let share our knowledge. Wordpress.com; 2012 [cited 2016Nov30]. Available from: https://fikriinside.wordpress.com/2012/04/26/human-vs-computer-which-one-has-a-better-knowledge/
15. Artificial Intelligence [Internet]. Oscar Education. 2012 [cited 2016Nov30]. Available from: http://oscareducation.blogspot.in/2013/01/artificial-intelligence.html

*University of Colorado School of Medicine