Improving precision in concept normalization

Mayla Boguslav†, K. Bretonnel Cohen, William A. Baumgartner Jr., and Lawrence E. Hunter

Computational Bioscience Program, University of Colorado School of Medicine,
Aurora, CO 80045, USA
†E-mail: Mayla.Boguslav@ucdenver.edu
compbio.ucdenver.edu

Most natural language processing applications exhibit a trade-off between precision and recall. In some use cases for natural language processing, there are reasons to prefer to tilt that trade-off toward high precision. Relying on the Zipfian distribution of false positive results, we describe a strategy for increasing precision, using a variety of both pre-processing and post-processing methods. They draw on both knowledge-based and frequentist approaches to modeling language. Based on an existing high-performance biomedical concept recognition pipeline and a previously published manually annotated corpus, we apply this hybrid rationalist/empiricist strategy to concept normalization for eight different ontologies. Which approaches did and did not improve precision varied widely between the ontologies.

Keywords: ontologies; precision medicine; natural language processing; text mining; concept normalization

1. Introduction

Ambiguity is a fundamental feature of all known human languages.[1] That ambiguity is one of the fundamental challenges for natural language processing systems.[2] One form of ambiguity relevant to text mining for precision medicine is *polysemy*—the phenomenon of words having more than one meaning. For example, Johnson et al.[3] point out that the word *cone*, concept BTO:0000280 in the BRENDA Tissue Ontology, refers to an ovule- or pollen-bearing structure that participates in reproductive functions of pine trees. The word *cone* also appears in terms from the Gene Ontology, where it refers to a kind of cell found in the retina. Since it has more than one potential meaning, it is ambiguous—specifically, it is polysemous.

Concept normalization is the language processing task of mapping mentions of concepts in text to an independently defined terminology or ontology.[4,5] Polysemy-based ambiguity of terms in the terminology can cause errors in that mapping. These include both false negatives—a failure to recognize a mention of a concept—and false positives—incorrectly identifying a concept as being mentioned, when in fact it has not. In some applications in precision medicine, false positives are more harmful than false negatives.[6–11] For example, false positives in tumor profiling in cancer precision medicine can lead to somatic and germline mutation confusion.[12] Despite this general challenge, dictionary-based methods are effective for mapping text to ontologies. For example, in normalizing genes, chemicals, cell types, and tissue types.[13] This motivates the error-analysis-based strategy for reducing the false positive rate for concept normalization systems.

Most systems[14] (and most evaluations of them[4]) attempt to optimize a measure called F_1, which weights false positives (i.e. precision) and false negatives (i.e. recall) equally. However, there are applications of concept normalization in which equal weighting is not ideal. For example, emphasizing precision is useful in protein protein interaction tasks involving large corpora because false positives will degrade the accuracy and reliability of the knowledge inferred.[15] High precision is also important for user acceptance, as the presence of obvious errors in system output reduces user confidence, for example, as shown in a study of information retrieval from medical textbooks.[16] Further, large corpora are often redundant, the same information is present multiple times.[17] If the goal is to identify *types* of concepts in large corpora, then high precision, at the cost of some loss of recall, improves overall performance. There are situations in which recall can be more important.[18,19] For example, recall is important "for practical applications like semantic search, since such applications need to recognise as many events as possible [in biomedical event extraction]."[18] If the results of text mining will be further used in computation, it is advantageous to have high recall for further training steps,[18] and it may be possible to filter out false positives in later processing, making recall more important.[19]

Here though, we focus on precision because of the application: precision medicine. Concept recognition has found many applications in precision medicine, for example information extraction from cancer clinical trials[6,7] or cancer pathology reports,[10] prediction of cancer[8] or cancer stage[9] in electronic medical records (EMRs), and in patient-centered decision support.[11] In each of these applications, improved precision (even at the cost of some recall) has significant advantages. One concern with many false positives is user acceptance: a system with many errors is less likely to be implemented. Thus, reducing false positives increases the likelihood that these systems will be used for precision medicine.

As with many other phenomena in NLP, the distribution of false positives is often Zipfian: a relatively small number of errors occur frequently, while most occur rarely. Johnson, et al.[3] observed that precision can often be improved significantly by addressing the common errors: "The Zipf-like distribution of error counts across terms suggest that filtering a small number of terms would have a beneficial effect on the error rates due to... ambiguity-related errors."[3] Thus, fixing the top errors can have a beneficial effect on overall performance. This is the motivation here.

To assess the accuracy of a concept normalization system, a manually annotated gold standard corpus is generally required. Here, we use the Colorado Richly Annotated Full Text Corpus (CRAFT) of full text biomedical journal articles, annotated with concepts from eight different ontologies.[20] As a baseline concept normalization system, we used the best performing systems from Funk, et al.,[21] with the precision maximizing parameters for each ontology. For each ontology, we tested for a Zipfian distribution, identified the most common concept errors in PubMed Central Open Access, and tested a set of five different potential pre- and post-processing steps that could improve precision.

2. Materials and Methods

To assess whether biomedical concepts have a Zipf-like distribution, and to identify the most common false positives, we use the PubMed Central Open Access (PMCOA) corpus. PMCOA is under the Creative Commons or similar license that generally allows more liberal redistribution and reuse than a traditional copyrighted work.[22] As of this writing there are 1,494,227 articles in this set.

To calculate precision and recall, we used The Colorado Richly Annotated Full Text Corpus (CRAFT): 67 public full-text biomedical articles from PMCOA that are manually annotated, with approximately 100,000 concept annotations to eight different biomedical ontologies, including Chemical Entities of Biological Interest (ChEBI), Cell Ontology (CL), Gene Ontology (GO) which includes biological processes (GO-BP), cellular components (GO-CC), and molecular function (GO-MF), NCBI Taxonomy (NCBI Taxon), Protein Ontology (PRO), and Sequence Ontology (SO).[20] Our concept normalization system is ConceptMapper, a high-performance customizable dictionary look-up tool implemented as a UIMA component.[23] Funk et al. determined that ConceptMapper is the best performing (highest F_1 measure) concept recognition software as compared to others.[21] They also determined the best parameter settings for it to obtain the highest F_1 measure, precision, or recall.[21] We use the high-precision parameter settings as baseline concept normalization. Also, the dictionaries for ConceptMapper are taken from ontology concept names and synonyms. See `https://github.com/UCDenver-ccp/ccp-nlp-pipelines` for the ConceptMapper pipeline.

To test the distributional characteristics of the PMCOA corpus, we ran the ConceptMapper system over it and calculated concept frequencies for each ontology. To identify the most frequent false positives, we manually reviewed the 20 most frequently occurring concepts for each of the eight ontologies. Based on this manual review, we applied five different pre- and post-processing strategies (and their combinations) in no particular order, in order to improve precision for each ontology. Each strategy's effectiveness was evaluated using CRAFT. The strategies included:

- Pre-processing ConceptMapper dictionaries to remove problematic concepts/synonyms

  Some false positives are due to errors in ConceptMapper dictionary entries, so delete concepts or synonyms if they grossly misidentify the definition of the concept. For example, CHEBI:90880, the tipiracil cation, had as a synonym "(1+)," which matched all instances of the numeral "1" in PMCOA. Although 91% of the papers in PMCOA had at least one match, none of those were references to the ChEBI term.
- Pre-processing to remove single word concepts that are in the general English dictionary

  One type of ambiguity stems from the fact that some concepts have acronyms that are general English words (single words), even though they are scientifically specialized words. For example, the PRO concept, PR:00003444, "carbonic anhydrase 2," with acronym "CAN," triggers annotation of PR:00003444 to all instances of the general English word "can". This repair filters out concepts and synonyms that are in a general English dictionary in hopes to only capture the scientifically specialized words in the ontologies.
- Post-processing to check for the presence of acronym apositives

  To ensure that acronyms are found correctly, keep concepts using capitalization: all upper case or first letter capitalized. For example, the NCBI Taxon concept NCBITaxon:3702, "Arabidopsis thaliana," with acronym "AT," not the general English word "at," is commonly recognized as a false positive. Further, some acronym apositives have the first letter uppercase such as the GO-MF concept GO:0043336, "site-specific telomere resolvase activity," with acronym "ResT." This filter keeps concepts with either full capitalization or the first letter capitalized.
- Post-processing to keep only the canonical form of the concept

Keep concepts that are in the canonical form in the ConceptMapper dictionaries. For example, the ChEBI concept CHEBI:27889, “lead(0),” in its canonical form, is kept, but its synonym “lead” is not.

- Post-processing based on the frequency of occurrence of the concept in the document
  Calculate how frequent a concept appears in each document and test whether keeping an annotation if the concept was more or less frequent than a threshold improved precision. For example, if found that deleting concept annotations that appeared under 40 times in a text improved precision, and the PRO concept PR:000009431, “kinetochore-associated protein 1,” with acronym “ROD,” appeared 28 times in a document, then the PRO concept annotation would be removed. To determine the threshold for each ontology, optimize precision over a range of thresholds from 0 to 50 on CRAFT. 50 was the upper threshold because manually, the max threshold found was 40. Note that each ontology may have its own threshold.

3. Results

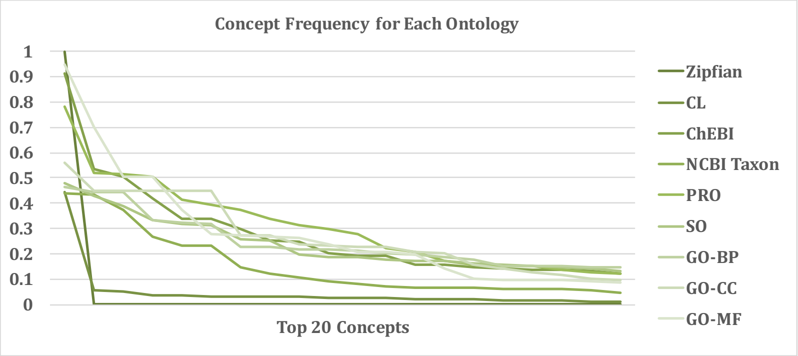


Fig. 1: The concept frequencies for all 8 ontologies for the top 20 concepts. The rates decline rapidly in a Zipfian distribution using the Kolmogorov-Smirnov test[24] based on no significant difference between each ontology distribution and the Zipf distribution (p-value $>$ 0.2).

The distribution of the 20 most frequent concepts in the PMCOA literature obeys a Zipfian distribution (see figure 1) using the Kolmogorov-Smirnov test[24] (igraph package in R[25]) based on seeing no significant difference between each ontology distribution and a Zipfian distribution (see `github.com/mboguslav/Concept-Normalization`). Thus, fixing the most-frequent errors could be expected to increase precision. The 20 most frequently recognized concepts from each ontology in the PMCOA corpus were evaluated manually. Many of the most frequently occurring concepts were errors. The top two terms for each ontology are listed below with false positives in red and the ontology concept ID and frequency in parentheses. The top 20 are available at `https://github.com/mboguslav/Concept-Normalization`. This manual evaluation was used to determine the pre- and post-processing strategies that aim to get rid of false positives and improve precision.

- CL: cell (CL_0000000, 44%); T cell (CL_0000084, 5.6%)
- ChEBI: (1+) (CHEBI_90880, 91%); group (CHEBI_24433, 54%)

- NCBI Taxon: Homo Sapiens (NCBITaxon_9606, 44%); order (NCBITaxon_order, 43%)
- PRO: carbonic anhydrase 2 (PR_000034449, 78%); Golgi-associated PDZ and coiled-coil motif-containing protein (PR_000008147, 52%)
- SO: single (SO_0000984, 48%); region (SO_0000001, 43%)
- GO-BP: developmental process (GO_0032502, 47%); cell aging (GO_0007569, 44%)
- GO-CC: cell and encapsulating structures (GO_0005623, 56%); cell part (GO_0044464, 45%)
- GO-MF: 3-methyl-2-oxobutanoate dehydrogenase (ferredoxin) activity (GO_0043807, 95%); enoyl-CoA hydratase activity (GO_0004300, 70%)

The manually determined problematic concepts in the top two are in red above, such as the “(1+)” issue already mentioned. With this list of problematic terms, an error analysis (looking to the original text) was performed to determine the cause of the false positives. This demonstrated that many of the false positives were also terms in general English (e.g. can, lead, fig). We then assessed all the terms in each ontology that were in a general English dictionary using the Enchant spell-checker library[26] (see figure 3). With this information, we tried to post-process the general English words (as mentioned in methods), but the precision did not increase and this was better captured by the other post-processing methods.

The highest precision criteria were chosen, and statistical significance of the difference was calculated according to Yeh.[27] This criteria was implemented on PMCOA to get the new top 20 most frequent concepts, in hope that the new concept frequencies make more sense then the previous (see `https://github.com/mboguslav/Concept-Normalization` for the new top 20 concepts).

In general, only two ontologies needed pre-processing due to problematic concepts or synonyms: ChEBI and PRO. This increased precision significantly for ChEBI and had little impact on PRO (see section 3.1 with figure 2).

Within post-processing, focusing on precision only, the best criteria to improve precision was checking for the presence of appositives: keeping concept annotations if the annotation was all upper case (an acronym) or the first letter of the annotation was uppercase (acronym or proper noun), and discarding them otherwise (see section 3.3 with figure 4a). With these criteria, precision improved significantly for 6 out of the 8 ontologies. However, recall plummets at the same time, except for PRO: the change in recall is minimal. Thus different criteria is needed that improves precision and maintains recall (remains the same or slightly decreases).

Combining the post-processing that keeps only the canonical forms and the post-processing based on the frequency of occurrence, improves precision and only slightly lowers recall. Thus keep concept annotations if they are in the canonical form of the ontology concept or the number of times the annotation concept appears per document is over a specific threshold as explained above. With the optimal thresholds for the highest precision for each ontology, ConceptMapper ran over CRAFT (see thresholds in table 1). This significantly improved precision for 5 of the 8 ontologies and did nothing for the others (see section 3.4). Overall precision improves, while recall decreases slightly leading to a decrease in F_1 measure.

3.1. *Results from deleting concepts and synonyms*

Since deleting concepts or synonyms is likely to cause an increase in false negatives, this approach is likely to lead to performance improvements only in the case of obvious ontological errors, such as the case of "(1+)" as a synonym for the tipiracil cation. These errors were found in ChEBI and PRO. Changes led to a modest improvement for ChEBI, but not for PRO. This is likely because the erroneous terms do not appear in CRAFT (although they did in PMCOA). For example, the PRO concept PR:000015574, "small proline-rich protein 2A," has the synonym "2-1," which was deleted, but the synonym is not in CRAFT. In no case did this reduce performance (see figure 2).

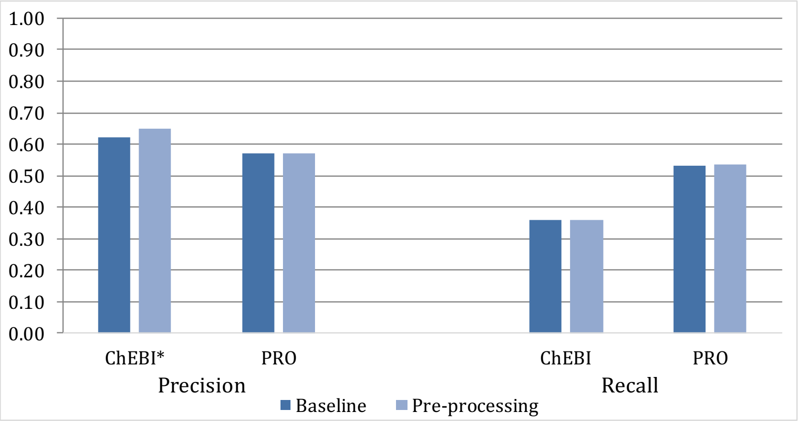


Fig. 2: Precision and recall after the pre-processing step as compared to the baseline. *Statistically significant difference in precision between pre-processed and baseline using χ^2 test.[27]

3.2. *Results from general English dictionary*

Some of the ambiguity issues involve concept acronyms that are words in the general English dictionary. Thus, we quantified the number of single word concepts that are in the general English dictionary for each ontology. As figure 3 shows, GO-BP has the most single word concepts that are general English words, followed by GO-CC. The caveat to this is that some concepts are general English words and correctly identify the ontology concept, such as sulfation in GO-BP. Due to this caveat, more concept annotations were deleted that were identifying the correct concept, rather than identifying the wrong concept, leading to a decrease in precision. Further, the issue with single word concepts as general English words is somewhat solved in the other post-processing criteria.

3.3. *Results from Case post-processing*

Requiring annotations to be acronyms based on case, improved precision significantly for six ontologies, but hurt two (NCBI Taxon and GO-CC). The first letter capitalized also finds proper nouns, which are also more likely to be concepts of interest. For all but NCBI Taxon, precision was above 0.74 with most between 0.8 and 0.98 (see figure 4a). Looking at F_1 measure, overall it drastically declined, except for PRO which increased from 0.55 to 0.58, suggesting that this may be a good fix for PRO at least (see figure 5).

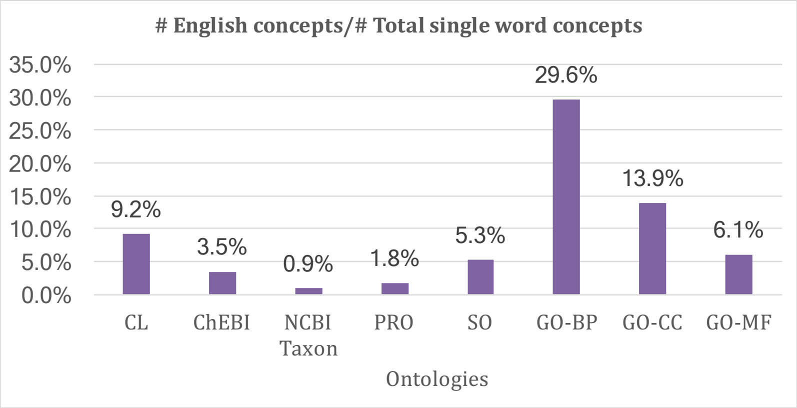
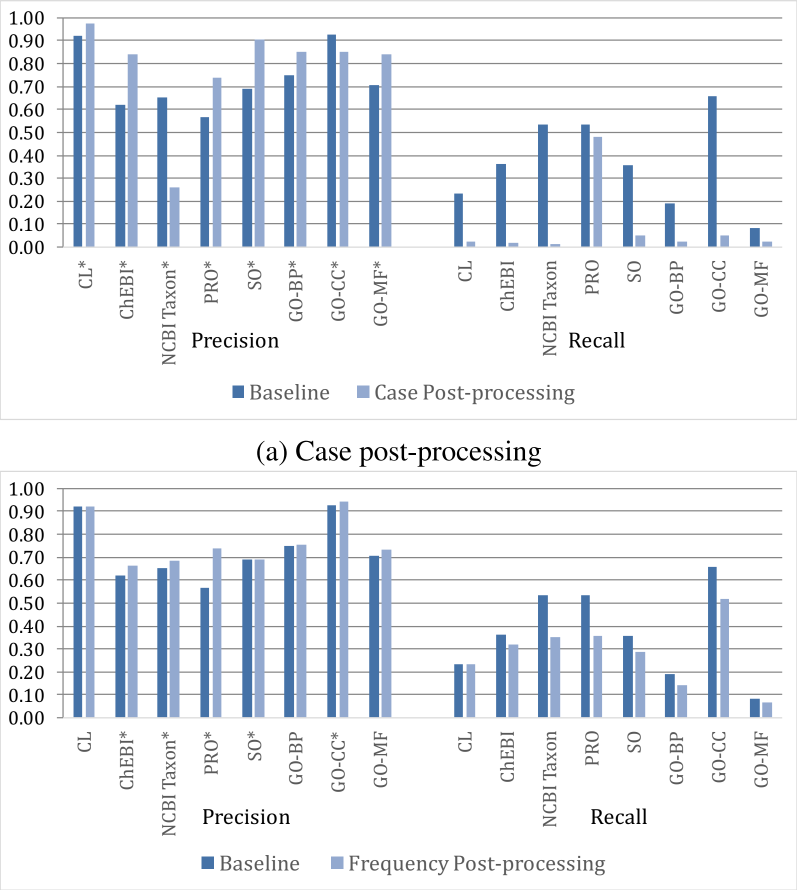


Fig. 3: The percentage of single word concepts for each ontology in the general English dictionary.



(a) Case post-processing

(b) Frequency post-processing

Fig. 4: Precision and recall for both case post-processing and frequency post-processing as compared to the baseline. *Statistically significant difference in precision between post-processing and baseline using χ^2 test.[27]

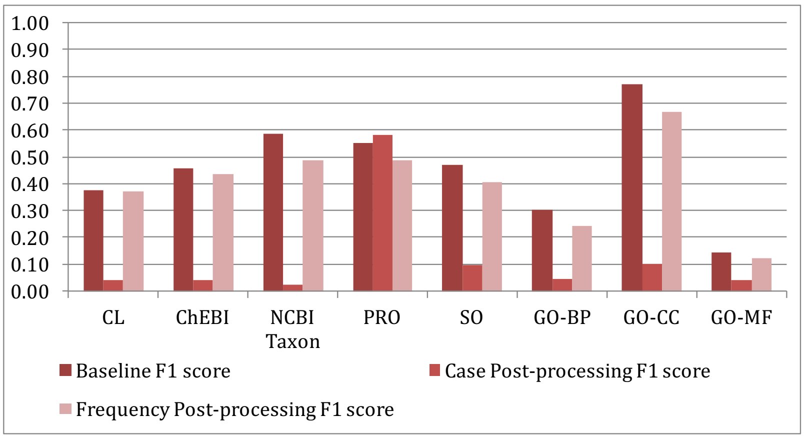


Fig. 5: F_1 scores for both post-processing steps as compared to baseline.

3.4. *Results from canonical form and frequency-based post-processing*

Requiring the canonical form of the concept is helpful: using the canonical forms improved precision because of its specificity to the concept. For example, in the “lead(0)” example above, the synonym “lead” causes false positive issues, finding annotations of the verb (“lead to”) instead of the noun. Thus, if we only keep the annotations that have the ”lead(0)” form, then we are more likely to annotate the concept correctly. At the same time, keeping only the canonical form drastically decreased recall because concept synonyms are widely used, including “lead”. Therefore, the canonical form post-processing must be used in combination with another post-processing technique: frequency-based post-processing. Using the concept annotation frequency per document (see table 1 for thresholds for each ontology) and the specificity of the canonical forms, enabled us to improve precision significantly for five ontologies (ChEBI, NCBI Taxon, PRO, SO, and GO-CC), not hurt the other three, and while only slightly lowering recall (see figure 4b). Precision also improved for the two ontologies, NCBI Taxon and GO-CC, where precision decreased using case post-processing (see figure 4a). Overall, F_1 measure remained the same or decreased slightly for each ontology (see figure 5), an improvement on the case post-processing.

Table 1: The concept frequency threshold for each ontology based on maximizing precision with thresholds between 0 and 50.

Ontology	**Frequency threshold**	**Ontology**	**Frequency threshold**
CL	2	SO	42
ChEBI	22	GO-BP	43
NCBI Taxon	40	GO-CC	22
PRO	43	GO-MF	3

4. Repeatability and reproducibility

The ConceptMapper pipeline is available at: `github.com/UCDenver-ccp/ccp-nlp-pipelines`. The intermediate data files for the analysis are available at `github.com/mboguslav/Concept-Normalization`.

Since ontologies are often updated, the results here are unlikely to be reproduced exactly the same in the future. In a more general sense, however, the effects of applying the pre- and post-processing rules to the concept annotations can change any time the contents of the ontologies change. For example, we reported the tipiracil cation error to the ChEBI maintainers. They have since replaced *(1+)* with *tipiracil(1+),* obviating the need for our handling of the false positives that resulted from the *(1+)* synonym. Although this could potentially change the measured performance of the current implementation, on balance it is difficult to see this state of affairs as anything other than good.

These are weaknesses of the current version of the implementation (and probably any future ones). Nonetheless, the elaboration of the hybrid rationalist/empiricist methodology of using distributional aspects of the errors to prioritize which ones to address, with knowledge-based approaches being applied to recover from them, remains a contribution. There is an oft-heard trope that purely statistical approaches to language processing "work" better than knowledge-driven ones. That claim is rarely, if ever, substantiated. The work presented here is consistent with what one actually observes in practice: hybrid systems are a reasonable approach to the challenges of the ambiguity of natural language.

5. Discussion and conclusions

Table 2: Summary of findings for each pre- or post-processing step. Y means the method increased precision, No means the method did not change precision, and Decrease means that the method decreased precision.

Ontology	**Pre-processing**	**Case post-processing**	**Frequency post-processing**
CL		**Y**	No
ChEBI	**Y**	**Y**	**Y**
NCBI Taxon		**Decrease**	**Y**
PRO	No	**Y**	**Y**
SO		**Y**	No
GO-BP		**Y**	No
GO-CC		**Decrease**	**Y**
GO-MF		**Y**	No

At least one method improved precision for each ontology tested (see table 2), since there is a "Y" for at least one method for each ontology. Further, the ontology with the worst baseline precision was PRO and precision significantly improved using both post-processing methods, and F_1 measure even improved using the case method (see figure 4). The ontology with the second lowest precision was ChEBI, and precision improved significantly for all methods (see "Y" for all methods for ChEBI in table 2). Note that there is no single method that only increases or only decreases precision for all ontologies. Each method presented here though, helps at least one ontology, specifically pre-processing, the case post-processing, and frequency post-processing methods. So, we suggest trying these methods to improve precision either on a manual annotated corpus like CRAFT, or an unknown set of documents that one can review manually after, before singling out each ontology and finding its specific errors and fixes. Another option is to analyze the most frequent concepts as we did here to see if they are false positives and follow a Zipfian distribution.

Although these results are from the biomedical literature, there are reasons to believe this will apply to concept normalization in electronic patient records as well. For example, a general text mining NLP tool and ontologies have been used to extract information from electronic medical records.[28]

5.1. *Limitations*

Here we only evaluated our metrics (precision, recall, and F_1 measure) using CRAFT and do not know how a different corpus may affect these metrics.[29] Further, the limited size (and domain) of the CRAFT corpus means that the estimates of precision and recall will not perfectly reflect the changes in performance over the entire PMCOA corpus.[30] However, we believe that some of the changes that showed little effect in CRAFT are actually useful over the entire PMCOA corpus, suggested by the fact that the new top 20 concepts for each ontology have more realistic concept frequencies. Further, there are other possible post-processing rules to evaluate that were not included in this study. For example, for the PRO concept PR:000008147, "Golgi-associated PDZ and coiled-coil motif-containing protein," with acronym "FIG," which currently recognizes the same shorthand for figures in text, context could inform post-processing. For figures in the text, a number usually appears after it, whereas it would probably not for the protein. Future directions include trying other pre- and post-processing techniques that have the potential to improve precision.

5.2. *Conclusion*

The work here shows that it is possible to improve precision without losing much recall for existing systems, ConceptMapper, by exploring both pre- and post-processing methods of concepts and concept annotations, respectively. Further, this work provides some evidence that there is no single fix that will improve precision for concept recognition for all ontologies, based on exploring five different methods to do so. At the same time, these methods improve precision for at least one ontology, suggesting that trying these methods before examining each ontology closely could be worthwhile. For example, the case post-processing method improves both precision and F_1 measure for the protein ontology suggesting that this is a necessary step in concept normalization for it. Thus trying these different combinations of pre- and post-processing steps on other ontologies may prove fruitful for concept recognition as a whole.

6. Acknowledgements

Boguslav was supported by the Deans Fund at University of Colorado Anschutz Medical and now is supported by the training grant T15 LM009451. 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 dAlembert fellowship. Hunter is supported by the grants R01 LM008111 and R01 LM009254, which also supports Baumgartner. The work was aided by discussions with Michael Bada, Negacy Hailu, and Tiffany Callahan; all remaining faults are the authors.

7. Author contributions

MB: ran all experiments, analyzed the data, and wrote the final version of the paper. KBC: analyzed the data and wrote the first draft of the paper. WAB: performed precursor work and participated in running experiments. LEH: conceived of and directed the project. All authors participated in analyzing the data and approved the final version of the paper.

References

1. D. of Linguistics, *Language Files: Materials for an Introduction to Language and Linguistics.*, 12 edn. (The Ohio State University Press, 2016).
2. D. Jurafsky and J. H. Martin, *Speech and language processing* (Pearson London, 2014).
3. H. L. Johnson, K. B. Cohen and L. Hunter, A fault model for ontology mapping, alignment, and linking systems, in *Pacific symposium on biocomputing.*, 2007.
4. F. Rinaldi, T. R. Ellendorff, S. Madan, S. Clematide, A. Van der Lek, T. Mevissen and J. Fluck, *Database* **2016** (2016).
5. K. B. Cohen, G. K. Acquaah-Mensah, A. E. Dolbey and L. Hunter, Contrast and variability in gene names, in *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, 2002.
6. J. Zeng, Y. Wu, A. Bailey, A. Johnson, V. Holla, E. V. Bernstam, H. Xu and F. Meric-Bernstam, *AMIA Summits on Translational Science Proceedings* **2014**, p. 126 (2014).
7. J. Xu, H.-J. Lee, J. Zeng, Y. Wu, Y. Zhang, L.-C. Huang, A. Johnson, V. Holla, A. M. Bailey, T. Cohen *et al.*, *Journal of the American Medical Informatics Association* **23**, 750 (2016).
8. M. Hoogendoorn, P. Szolovits, L. M. Moons and M. E. Numans, *Artificial intelligence in medicine* **69**, 53 (2016).
9. J. L. Warner, M. A. Levy, M. N. Neuss, J. L. Warner, M. A. Levy and M. N. Neuss, *Journal of oncology practice* **12**, 157 (2015).
10. A. E. Wieneke, E. J. Bowles, D. Cronkite, K. J. Wernli, H. Gao, D. Carrell and D. S. Buist, *Journal of pathology informatics* **6** (2015).
11. M. Becker and B. Böckmann, *Studies in health technology and informatics* **235**, p. 271 (2017).
12. A. Garofalo, L. Sholl, B. Reardon, A. Taylor-Weiner, A. Amin-Mansour, D. Miao, D. Liu, N. Oliver, L. MacConaill, M. Ducar *et al.*, *Genome medicine* **8**, p. 79 (2016).
13. H. L. Johnson, K. B. Cohen, W. A. Baumgartner Jr, Z. Lu, M. Bada, T. Kester, H. Kim and L. Hunter, Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies, in *Pacific Symposium on Biocomputing*, 2006.
14. K. B. Cohen and D. Demner-Fushman, *Biomedical natural language processing* (John Benjamins Publishing Company, 2014).
15. J. Lee, S. Kim, S. Lee, K. Lee and J. Kang, *BMC medical informatics and decision making* **13**, p. S7 (2013).
16. D. C. Berrios, R. J. Cucina and L. M. Fagan, *Journal of the American Medical Informatics Association* **9**, 637 (2002).

17. K. B. Cohen, K. Verspoor, H. L. Johnson, C. Roeder, P. V. Ogren, W. A. Baumgartner Jr, E. White, H. Tipney and L. Hunter, *Computational intelligence* **27**, 681 (2011).
18. M. Miwa and S. Ananiadou, *BMC bioinformatics* **16**, p. S7 (2015).
19. W. A. Baumgartner, K. B. Cohen and L. Hunter, *Journal of biomedical discovery and collaboration* **3**, p. 1 (2008).
20. K. B. Cohen, K. Verspoor, K. Fort, C. Funk, M. Bada, M. Palmer and L. E. Hunter, *The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain*, in *Handbook of Linguistic Annotation*, (Springer, 2017), pp. 1379–1394.
21. C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter and K. Verspoor, *BMC bioinformatics* **15**, p. 59 (2014).
22. J. A. Fain, Nih public access policy how does information get uploaded to pubmed central and by whom? (2015).
23. M. A. Tanenblatt, A. Coden and I. L. Sominsky, The conceptmapper approach to named entity recognition, in *LREC*, 2010.
24. R. Wilcox, *Encyclopedia of biostatistics* (2005).
25. T. Nepusz and G. Csardi, R igraph manual pages (2003), `http://igraph.org/r/doc/fit_power_law.html`.
26. D. Lachowicz, Enchant (2008), `https://abiword.github.io/enchant/`.
27. A. Yeh, More accurate tests for the statistical significance of result differences, in *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 2000.
28. R. Batool, A. M. Khattak, T. S. Kim and S. Lee, Automatic extraction and mapping of discharge summarys concepts into SNOMED CT, in *Conf Proc IEEE Eng Med Biol Soc*, 2013.
29. S. Mehrabi, A. Krishnan, A. M. Roch, H. Schmidt, D. Li, J. Kesterson, C. Beesley, P. Dexter, M. Schmidt, M. Palakal *et al.*, *Studies in health technology and informatics* **216**, 604 (2015).
30. J. G. Caporaso, N. Deshpande, J. L. Fink, P. E. Bourne, K. B. Cohen and L. Hunter, Intrinsic evaluation of text mining tools may not predict performance on realistic tasks, in *Pacific symposium on biocomputing.*, 2008.