Part-of-speech tagging for clinical text: wall or bridge between institutions?

Jung-wei Fan, PhD[1], Rashmi Prasad, PhD[2], Rommel M. Yabut, MD[1], Richard M. Loomis, MD[1], Daniel S. Zisook, MD[1], John E. Mattison, MD[1], Yang Huang, PhD[1]
[1]Kaiser Permanente Southern California, Pasadena, CA
[2]University of Pennsylvania, Philadelphia, PA

Abstract

Part-of-speech (POS) tagging is a fundamental step required by various NLP systems. The training of a POS tagger relies on sufficient quality annotations. However, the annotation process is both knowledge-intensive and time-consuming in the clinical domain. A promising solution appears to be for institutions to share their annotation efforts, and yet there is little research on associated issues. We performed experiments to understand how POS tagging performance would be affected by using a pre-trained tagger versus raw training data across different institutions. We manually annotated a set of clinical notes at Kaiser Permanente Southern California (KPSC) and a set from the University of Pittsburg Medical Center (UPMC), and trained/tested POS taggers with intra- and inter-institution settings. The cTAKES POS tagger was also included in the comparison to represent a tagger partially trained from the notes of a third institution, Mayo Clinic at Rochester.

Intra-institution 5-fold cross-validation estimated an accuracy of 0.953 and 0.945 on the KPSC and UPMC notes respectively. Trained purely on KPSC notes, the accuracy was 0.897 when tested on UPMC notes. Trained purely on UPMC notes, the accuracy was 0.904 when tested on KPSC notes. Applying the cTAKES tagger pre-trained with Mayo Clinic's notes, the accuracy was 0.881 on KPSC notes and 0.883 on UPMC notes. After adding UPMC annotations to KPSC training data, the average accuracy on tested KPSC notes increased to 0.965. After adding KPSC annotations to UPMC training data, the average accuracy on tested UPMC notes increased to 0.953. The results indicated: first, the performance of pre-trained POS taggers dropped about 5% when applied directly across the institutions; second, mixing annotations from another institution following the same guideline increased tagging accuracy for about 1%. Our findings suggest that institutions can benefit more from sharing raw annotations but less from sharing pre-trained models for the POS tagging task. We believe the study could also provide general insights on cross-institution data sharing for other types of NLP tasks.

Introduction

With the adoption of Electronic Health Record (EHR) systems and advances in healthcare information technologies, narrative clinical documents have become valuable resources for decision support,[1] quality-of-care measurements[2] and various other research and clinical applications.[3,4,5] Natural Language Processing (NLP) is considered one of the key technologies to unlock the valuable information in biomedical narratives ("biomedical" collectively means clinical, medical, and biological in this paper), by analyzing the syntactical structure and semantics of text, and transforming ambiguous text into computable discrete data.

A typical NLP system consists of pipelined components handling multiple tasks such as tokenization, sentence delimitation, part-of-speech (POS) tagging, phrase chunking, parsing, and concept mapping. As one of the initial steps, POS tagging determines the part of speech (or word class, such as noun, verb, adjective, preposition, etc.) for each token in a sentence. These word classes serve as a fundamental type of features to be consumed by downstream NLP components for critical tasks including phrase chunking. Kazama et al.[6] reported positive effect on the performance in recognizing biomedical named entities by adding POS features. Since POS tags are essential input for most syntactic parsers, accuracy of POS tagging transitively affects applications that rely on accurate parsing (e.g., relation extraction from biomedical text).[7] In evaluating a parser for clinical notes, Fan and Friedman[8] reported that 20% of the known parsing errors resulted from incorrect POS tags. For example, in

"history of significant left lower quadrant pain" mis-tagging the "left" as a noun and "lower" as a verb was found to disrupt the entire phrase structure.

Similar to many data-driven NLP programs, an automated POS tagger needs sufficient quality training data that are manually tagged by knowledgeable experts. For general English, POS taggers trained on the Penn Treebank corpus[9] achieved accuracies of higher than 0.97, [10] benefiting tremendously from the abundant high-quality POS annotations of the corpus. However, biomedical language is quite different from general English, and it has been reported that the high tagging accuracy did not carry over when Treebank-trained taggers were applied to biomedical text directly (accuracy dropped ~4% on clinical notes[11] and dropped ~12% on Medline abstracts).[12] Encouragingly, Coden et al.[11] showed tagging accuracy on clinical notes increased ~2% when the Treebank corpus was augmented with a set of Mayo Clinic's own notes. Based on their study, Mayo Clinic subsequently released a POS tagger trained on a mixed corpus comprising Treebank, Medline abstracts, and local clinical notes.[13] Based upon the above studies, it is widely expected that the community will substantially benefit from sharing such annotated data and/or pre-trained programs.

In this study, we performed experiments to determine the potential benefit of collaboration to ease the resource-intensive annotation process. Specifically, we attempted to answer the following two questions:

1. Can we benefit from directly using POS taggers trained on other institutions' clinical notes?
2. Can we benefit from using other institutions' notes to train our own tagger, if there are more POS-annotated clinical notes available to the community?

In order to answer these questions, we investigated the effect of training/testing POS taggers with different institution's notes as well as combining training data from different institutions. In summary, the results indicated that tagging accuracy dropped markedly if a tagger was trained purely on one institution's notes and tested on another's. Interestingly, when another institution's notes were added to the local training data (about equal sizes in our experiments), the tagger gained accuracy consistently. These findings suggest that institutions may benefit more from sharing raw annotations rather than canned models. Our study also touched upon other practical issues such as variance of guidelines and clinical genres in sharing annotations.

Background

To support biomedical NLP, many POS taggers have been developed and made available to public. Most of the taggers target biomedical literature and involve annotation of a training corpus. MedPost[14] was trained with 5,700 manually tagged Medline sentences (available for research) and achieved an accuracy of 0.969. Trained also with the MedPost corpus, dTagger[15] achieved an accuracy of 0.951 and was featured with leveraging the comprehensive UMLS Specialist Lexicon[16] for reducing unknown tokens and handling phrases. The GENIA tagger[12] was experimented with training data from different combinations of Wall Street Journal (WSJ), PennBioIE corpora, the GENIA corpus,[17] and was reported to achieve accuracies higher than 0.97. Both the PennBioIE and GENIA POS-tagged corpora are available to the research community. The former contains 2,257 Medline abstracts about CYP450 or oncology; the latter has 2,000 Medline abstracts about human transcription factors.

Less work has been reported on clinical POS tagging, likely due to strict regulations (e.g., HIPAA) on the access and use of clinical texts. Pestian et al.[18] reported POS annotation for 390,000 words of pediatric text at the Cincinnati Children's Hospital Medical Center. A tagger was trained, achieved an accuracy of 0.915, and (similar to dTagger) was reported to benefit from incorporating the Specialist Lexicon into the tagger dictionary. However, neither the corpus nor the tagger was noted to be available. Liu et al.[19] developed heuristic sampling methods to reduce the size of required clinical text annotation when co-training a POS tagger with the WSJ corpus. Evaluated in tagging surgical pathology reports, one of the sampling methods was reported to reduce 84% of the training data and achieved an accuracy of 0.927 (compared to 0.939 achieved by the full set). Similarly, the annotated corpus and the trained tagger were not noted to be available to the community. The MED corpus[20] developed by the Mayo Clinic at Rochester, Minnesota, contains 273 clinical notes with 100,650 POS-tagged tokens. The

annotations were pooled with POS-tagged WSJ and GENIA corpora in training a tagger, which was reported to achieve an accuracy of 0.936 on the Clinic's notes. Although the corpus is not available, the tagger is distributed as a pre-trained reusable model in a full-fledged biomedical NLP package, cTAKES,[13] contributed by Mayo Clinic.

At Kaiser Permanente (KP), we have a constant need and interest in leveraging NLP to process unstructured data in our large-scale electronic medical record system. For a fundamental task like POS tagging, we are interested in investigating how an individual institution could benefit from existing experience and resources shared by the community. In the clinical domain, there are few publically available POS taggers and POS-annotated training data. The cTAKES tagger appears to be the only POS tagger available to the community that is adapted closely for clinical text. In terms of annotated corpora, it appears currently no institution has released POS-tagged clinical notes for others to use. Possibly due to a dearth of such resources, several institutions have taken a makeshift approach that pools non-clinical training annotations and reported slightly positive to slightly negative effect on the accuracy in tagging clinical notes.[11,13] On the other hand, we are more interested in knowing the effect of pooling annotations from different institutions, with the anticipation that cross-institutional collaboration will be able to annotate sufficient clinical notes in the future. A promising resource is the i2b2/VA NLP Challenge shared corpora, of which the 2010 corpus contains 826 de-identified clinical notes contributed by Partners HealthCare, Beth Israel Deaconess Medical Center, and University of Pittsburgh Medical Center. Although the released annotations are on concept, assertion, and relation extraction, the community is encouraged to add other levels of annotations (including POS tags) and feedback to the repository.

Considering the available resources for studying cross-institution POS tagging, we decided to perform POS annotation for some of our own KP notes and some of the i2b2/VA Challenge shared notes, and investigate the performance changes as a POS tagger is trained/tested on notes of different institutions. The pre-trained cTAKES tagger was also included to (roughly) represent Mayo Clinic's annotations in the comparative study. Details of our experiment design are described in Methods.

Methods

I. Gold standard annotation

In this study we utilized progress notes written in general medicine. Twenty five Kaiser Permanente Southern California (KPSC) progress notes were randomly selected and manually de-identified. Sentence boundaries were manually annotated for the KPSC notes.

An equivalent number of progress notes from University of Pittsburg Medical Center (UPMC) were manually selected from the 2010 i2b2/VA NLP Challenge shared corpus. To enhance diversity, the manual selection involved avoiding notes with apparent copied/pasted contents. Sentence boundaries in the UPMC notes were already provided with the shared corpus. We also manually fixed fragmented tokens in the shared corpus, e.g., "non - alcoholic" and "**AGE[in 30s]" were grouped into single tokens for appropriate POS tagging.

For each set, five random notes were pre-tagged by a model trained with the MedPost corpus and manually corrected by a computational linguist (2nd author, RP). The ten annotated notes were then reviewed and discussed by the authors to develop an annotation guideline for clinical POS tagging. Our guideline conforms to the Penn Treebank POS tagging guideline[21] in most cases, but with adjustments for clinical writing style (i.e., abundance of abbreviations, shorthand notations, and misspellings). For example:

- "Plan for abd ultrasound" → "abd" was tagged as JJ (adjective) for "abdominal"
- "Patient c/o sore throat" → "c/o" was tagged as VBZ (verb, 3rd person singular present) for "complains of"
- "HR is belw 60" → "belw" was tagged as IN (preposition) for misspelled "below"

It was also observed that some of the abbreviations were used productively and required context-dependent interpretation for precise tagging. For example:

- "Will check thyroid US to f/u status of goiter" → "f/u" was tagged as VB (verb, base form) for "follow up"
- "Pt will need outpatient f/u for DM" → "f/u" was tagged as NN (noun, singular) for the nominal sense of the "follow-up" action

Following the guideline, the linguist annotated the remaining forty notes, which along with the ten group-reviewed notes formed a gold standard of fifty notes. To measure three-way inter-annotator agreement, the first and last authors also independently annotated another ten random notes (5 KPSC, 5 UPMC) that did not overlap with the group-reviewed.

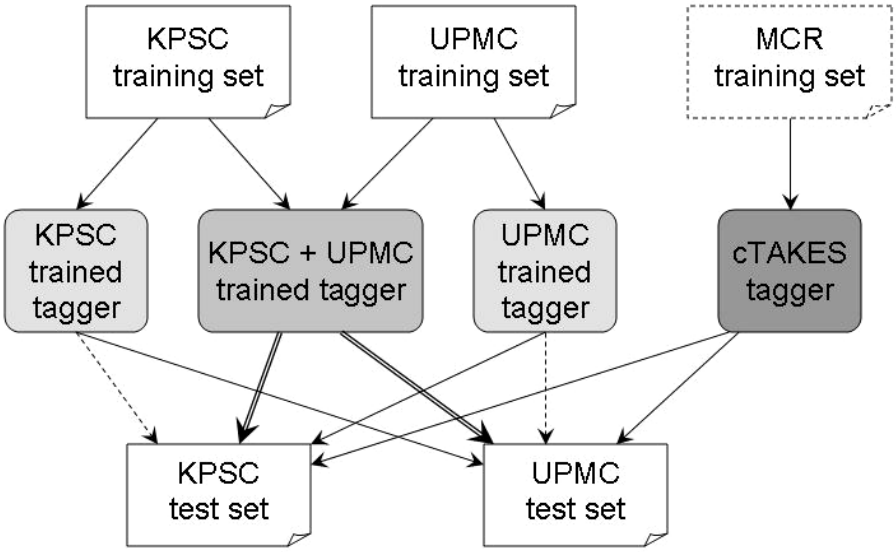

Figure 1. Different training/testing scenarios for intra-, inter-, and mix-institution evaluations. MCR stands for Mayo Clinic at Rochester, and the dashed box means the original training set was not available. Three types of arrows pointing to the test sets: a) the dashed means intra-institution cross-validation, b) the double-lined means mix-institution-trained model evaluated on mono-institution, c) the solid means inter-institution evaluation.

II. Evaluation with different training/testing combinations

The POS tagging program we used in the experiments was a machine learning model implemented in the OpenNLP package (version 1.4.3), with event cutoff value of 4 and 50 iterations in training. The base token-tag dictionary used by the model was prepared from the UMLS Specialist Lexicon,[16] coupled with a variable dictionary generated from the training set that differed depending on experiment setting. Three types of training/testing combinations were experimented (refer to Figure 1 above):

1. Intra-institution cross-validation

   To estimate the tagger's performance on a single institution's notes, 5-fold cross-validation was performed within the twenty five KPSC and twenty five UPMC notes respectively.

2. Inter-institution training/testing

   To estimate the tagger's performance when trained on notes from one institution and tested on notes from another, we trained the tagger with twenty five KPSC notes and tested it on the twenty five

UPMC notes, and vice versa. The two experiments meant to inspect the effect of notes from different institutions, with the single annotator following a single guideline.

As the cTAKES POS model/dictionary was available to the community, we used it conveniently as a tagger representing a tagger trained on Mayo Clinic's notes and tested on the KPSC and UPMC sets respectively. Note that due to the mixture of non-clinical texts in Mayo Clinic's training corpus and their different annotation guideline, the experiments were considered as loosely controlled and should be interpreted with caveats. Additionally, we computed an extra set of accuracies to accommodate known guideline differences. For example, in the loosened evaluation we allowed cTAKES-tagged FW (foreign word) to match any tag in our gold standard because we always annotated foreign words with the closest non-FW tag (e.g., "b.i.d." was tagged as RB - adverb).

3. Mix-institution training and mono-institution testing

To understand the effect of incorporating training data from another institution, we merged the entire annotated set from institution X with institution Y's training partition of each fold in the 5-fold intra-institutional cross-validation experiments. For example, in each fold of testing on five KPSC notes, the training data consisted of the remaining twenty KPSC notes plus the twenty five UPMC notes.

Results

I. Annotated gold standard

The twenty five KPSC notes contain 2,576 sentences with 15,556 tokens, of which 3,306 are unique. The twenty five UPMC notes contain 2,707 sentences with 15,844 tokens, of which 3,298 are unique. There is about one third (1,040) overlapping unique tokens between these two document sets. We summarize more comparison between the two annotated corpora in Table 1. Side-by-side tag distributions of the two corpora are displayed in Figure 2. The inter-annotator agreement rates are shown in Table 2, which indicates high agreement among the annotators following our POS tagging guideline. We plan to contribute our POS tagging for the 25 UPMC notes back to the i2b2 repository.

Table 1. Content comparison of the two annotated corpora (SB stands for sentence begin)

	25 KPSC notes	25 UPMC notes
Number of unique tokens	3,306	3,298
Unique tokens unknown to the base tagger dictionary	1,841	1,467
Average number of tags per unique token	1.03	1.04
Top 10 tag transitions	SB → NN JJ → NN NN → : NN → NN SB → JJ NN → CD NN → , IN → NN NN → IN NN → .	SB → NN JJ → NN NN → NN CD → NNS NN → . NN → , NN → : NN → CD NN → IN IN → NN

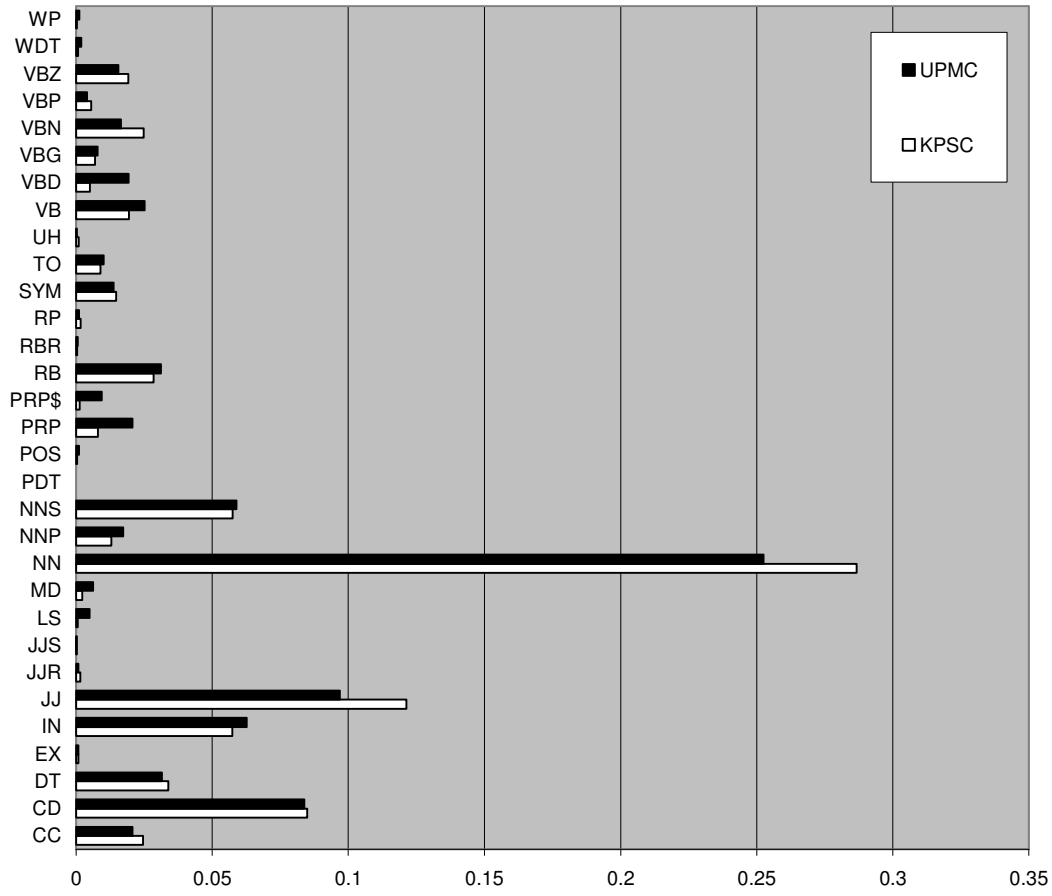

Figure 2. Tag distributions of the two annotated corpora (punctuation tags are omitted)

Table 2. Inter-annotator agreement rates in POS-tagging ten clinical notes

	5 KPSC notes	5 UPMC notes
Raw agreement ratio	0.966	0.959
Kappa statistics	0.960	0.953

II. Intra-institution cross-validation

The 5-fold cross-validation within the twenty five KPSC notes estimated an accuracy of 0.953, and it was 0.945 within the twenty five UPMC notes (see the first row of Table 5). The results indicate that a machine-learning POS tagger can yield promising performance given training data from the same institution following the same annotation guideline.

III. Inter-institution training/testing

To visualize performance changes along with the inter-institution results, we also included the intra-institution 5-fold cross-validation accuracies in Table 3. Though annotated by following the same

guideline, it was found that the tagging accuracies did not carry over different institutions (see first two rows of Table 3). Trained purely on KPSC notes, the accuracy dropped from 0.953 to 0.897 when tested on the UPMC notes. Trained purely on UPMC notes, the accuracy dropped from 0.945 to 0.904 when tested on the KPSC notes. The first two columns of Table 4 display frequent types of mis-tagging by training/testing across the KPSC and UPMC corpora.

The cTAKES POS tagging model was trained with a set of clinical notes from Mayo Clinic as well as non-clinical texts (biology literature and newspaper). Directly evaluating the cTAKES tagger resulted in an accuracy of 0.855 and 0.852 on the KPSC and UPMC notes respectively (third row of Table 3). As described in Methods, the other evaluation was computed to reduce the penalty from known difference in annotation criteria (e.g., on foreign words). The loosened evaluation resulted in an accuracy of 0.881 on the KPSC notes and 0.883 on the UPMC notes (fourth row of Table 3). Both still indicate a more than 5% drop of accuracy from the 0.936 measured on Mayo Clinic's own notes. The decreased accuracies appear to be consistent with the findings from the KPSC and UPMC experiments, suggesting that pre-trained POS model would not hold performance across institution boundaries. The last two columns of Table 4 display frequent types of mis-tagging by cTAKES (evaluated with loosened criteria) when tested on the KPSC and UPMC sets.

Table 3. Tagging accuracy with training/testing on different institutions

	Tested on KPSC notes	Tested on UPMC notes
Trained on KPSC notes	0.953 (5-fold CV)	0.897
Trained on UPMC notes	0.904	0.945 (5-fold CV)
cTAKES tagger	0.855	0.852
cTAKES tagger (looser evaluation)	0.881	0.883

Table 4. Frequent mis-tagging across institutions

Gold tag ~> Mis-tag	KPSC train / UPMC test	UPMC train / KPSC test	cTAKES / KPSC test	cTAKES / UPMC test
NNP ~> JJ	21	11	9	18
NNP ~> NN	1	2	1	3
JJ ~> NN	1	1	2	2
NN ~> VB	0	3	2	0
NN ~> VBP	0	0	1	1

IV. Mix-institution training and mono-institution testing

Despite the observed decrease of accuracies in applying one institution's tagger directly to another, a question remained to be answered was how an institution could benefit from another institution's annotations. By mixing another institution's POS annotations in training of the cross-validation experiments, interestingly we observed consistent improvement in tagging accuracies (see the second row compared to first row of Table 5). Inspection into the accuracy of each note revealed that the mix-institution trained models increased the number of correct tags for most of the notes. For example, in one of the UPMC notes the accuracy gain reached as high as 3% (32 more correct tags). The phenomenon indicated that pooling training data from another institution with the same guideline benefited POS tagger, and a certain amount of base training data from the local institution appeared essential for good performance.

Table 5. Five-fold cross-validation accuracy (with 95% confidence interval) based on mono-institution and mix-institution training data

	Tested on KPSC notes	Tested on UPMC notes
Mono-institution training	0.953 (0.949, 0.956)	0.945 (0.942, 0.949)
Mix-institution training	0.965 (0.962, 0.968)	0.953 (0.950, 0.957)

Discussion

POS tagging is a fundamental step for various NLP systems, rule-based or statistical. The training of POS taggers requires sufficient high-quality annotations, which are scarce especially in the clinical domain because of limited resources and stringent regulations. As the community has been seeking collaborative efforts to address this constraint, we wanted to discover (non-policy) issues in sharing POS annotations and gain insights on how we may benefit from such collaborations. Specifically, we looked into how POS tagging accuracy could be affected by different training/testing combinations involving data from different institutions.

The results between KPSC and UPMC notes showed that, even based on a single guideline, a tagger trained on annotations from one institution suffered about 5% loss in accuracy when applied directly to tag the notes of another institution. The loss was likely due to unknown tokens and rare patterns in one institution versus the other. It is hypothesized that certain types of mis-tagging might actually be harmless to downstream processing tasks (e.g., parsing or entity recognition). Investigating this hypothesis exceeds the scope of this study and we leave it to future research.

When a tagger was trained not only with local notes but also another institution's notes annotated with same guideline, the accuracy could gain about 1% increase as shown in Table 5. It is inferred that the local training data served as a smoothing core for the model to capture frequent institution-specific patterns, and the other institution's notes augmented more training instances for the infrequent patterns. This suggests that when sharing POS annotations among institutions, each institution would benefit more from receiving raw annotations than receiving a pre-trained tagger with the original corpus already crunched into model statistics.

Although there were factors we could not control in experimenting with the pre-canned cTAKES tagger, the relatively lower performance gave a hint of the possible effect from those factors – specifically we identified three of them for discussion:

1. There were likely other unknown differences between Mayo Clinic's guideline and ours that affected the accuracies, even when we loosened the criteria in the evaluation. In this study we did not have first-hand results from purely controlling training data based on different guidelines, but we postulate that sticking to a single guideline will be critical in sharing annotations among institutions. Otherwise, automated convergence of annotations from different guidelines would be feasible only if the guidelines are sufficiently detailed so as to allow computing/altering all the differences. We suggest that further experiments should be performed to inspect whether a POS tagger could still benefit from training with annotations based on moderately different guidelines.
2. The cTAKES tagger was trained with Mayo Clinic's notes and non-clinical texts including the GENIA and Treebank corpora. It was reported that mixing non-clinical texts in training led to slightly lower accuracy than pure clinical training data (0.936 compared to 0.940).[13] Besides different linguistic properties, the non-clinical corpora were annotated by following different guidelines. For example, it is known that in the GENIA corpus abbreviated units such as "mg" are annotated simply as NN, while in Mayo Clinic's guideline the tagging depends on the

quantity preceding the unit. This suggests that mixing annotations from different guidelines can introduce inconsistent noise into the training. Therefore, we believe clinical POS taggers will benefit from replacing the non-clinical training data with more coherently annotated clinical training data in the future.

3. The Mayo Clinic's notes used in training cTAKES tagger appeared to have wider coverage of genres than our focus on progress notes. Possible effects of mixing diverse clinical genres in training might include introducing more ambiguous tokens and incoherent patterns, making the model unstable especially when the number of training notes was small. This raises an interesting question in sharing annotations, that is, should we also specify the shared genre (e.g., progress note or discharge summary) to optimize the training? We believe further experiments are needed to add to our knowledge on the effects of training/testing POS taggers across different clinical genres.

In addition to some of the limitations mentioned above, the current study is limited with respect to:

1. The scale of annotation (~31K tokens) was relatively small (e.g., Mayo Clinic had more than 100K tokens) and could limit the statistical power in interpreting the results. On the other hand, our corpus was better controlled with focus on general medical progress notes than others without controlling the confounding factor of document type. Therefore, 31K tokens could have qualified as a good sample size for a single document type. We believe it still requires future investigation to figure out the sufficiency issue.
2. Partially limited by the scale of our annotations, we did not explore various ratios for combining the local training data and that from another institution. It will be an interesting study to investigate if there is an optimal ratio for mixing cross-institution training sets.
3. Although we observed high inter-annotator agreement in a random subset, the experiments involving KPSC and UPMC notes were based on the annotations from a single annotator. It is inferred that the variance of annotation would be higher if multiple annotators were used. For future studies, we suggest a cross-institution, multi-annotator, and single-guideline setting can be experimented to simulate real-world collaboration.
4. This study did not include controlled experiments to directly compare the effect of adding cross-institution training data versus non-clinical (cross-domain) training data. The comparison was not the focus of the current study, so we decided to leave it to future work.
5. In terms of model optimization, we did not compare different tagging algorithms and did not explore ensemble learning approaches such as voting among differently trained taggers. We leave the exploration to future work.

Conclusion

We performed experiments to investigate POS tagging accuracies achieved from sharing pre-trained taggers and from sharing raw training data. Following a single annotation guideline, a tagger trained purely on KPSC notes resulted in about 5% decrease in accuracy when tested on UPMC notes, and vice versa. Following a moderately different guideline and trained with Mayo Clinic's notes plus non-clinical texts, the cTAKES tagger also resulted in about a 5% decrease in accuracy when tested on KPSC and UPMC notes. For the KPSC and UPMC annotations following a single guideline, when the tagger was trained with a mixture of local training notes and those from the other institution, the accuracy gained about 1% increase on tested local notes. Our findings suggest that cross-institution collaboration will form a bridge through sharing raw annotations instead of a wall by sharing pre-trained models.

Acknowledgement

De-identified clinical records in i2b2/VA NLP Challenge were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY.

References

[1] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009 Oct;42(5):760-72.

[2] Voorham J, Denig P. Computerized extraction of information on the quality of diabetes care from free text in electronic patient records of general practitioners. J Am Med Inform Assoc. 2007 May-Jun;14(3):349-54.

[3] Tu SW, Peleg M, Carini S, Bobak M, Ross J, Rubin D, Sim I. A practical method for transforming free-text eligibility criteria into computable criteria. J Biomed Inform. 2011 Apr;44(2):239-50.

[4] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):392-402.

[5] Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2006 Jul 26;6:30.

[6] Kazama J, Makino T, Ohta Y, Tsujii J. Tuning support vector machines for biomedical named entity recognition. Workshop on Natural Language Processing in the Biomedical Domain. Philadelphia, USA; Association for Computational Linguistics; 2002. p.1-8.

[7] Fundel K, Küffner R, Zimmer R. RelEx – relation extraction using dependency parse trees. Bioinformatics. 2007;23(3):365-71.

[8] Fan J, Friedman C. Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. J Biomed Inform. 2011. doi:10.1016/j.jbi.2011.04.006

[9] Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: the Penn Treebank. Comput Linguist. 1993;19(2):313-30.

[10] Brants T. TnT: a statistical part-of-speech tagger. Sixth Conference on Applied Natural Language Processing. Seattle, USA; Association for Computational Linguistics; 2000. p.224-31.

[11] Coden AR, Pakhomov SV, Ando RK, Duffy PH, Chute CG. Domain-specific language models and lexicons for tagging. J Biomed Inform. 2005 Dec;38(6):422-30.

[12] Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, Tsujii J. Developing a robust part-of-speech tagger for biomedical text. 10th Panhellenic Conference on Informatics. Volos, Greece; Springer; 2005. p.382-92.

[13] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13.

[14] Smith L, Rindflesch T, Wilbur WJ. MedPost: a part-of-speech tagger for biomedical text. Bioinformatics. 2004;20(14):2320-1.

[15] Divita G, Browne AC, Loane R. dTagger: a POS tagger. AMIA Annu Symp Proc. Washington DC, USA; American Medical Informatics Association; 2006. p.200-3.

[16] McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994. p.235-9.

[17] Ohta T, Tateisi Y, Kim JD, Tsujii J. Genia corpus: an annotated research abstract corpus in molecular biology domain. Proc Hum Lang Technol Conf. 2002. p.73-7.

[18] Pestian JP, Itert L, Duch W. Development of a pediatric text-corpus for part-of-speech tagging. Proc Int Intell Inf Sys. Zakopane, Poland; Springer; 2004. p.219-26.

[19] Liu K, Chapman W, Hwa R, Crowley RS. Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. J Am Med Inform Assoc. 2007 Sep-Oct;14(5):641-50.

[20] Pakhomov SV, Coden A, Chute CG. Developing a corpus of clinical notes manually annotated for part-of-speech. Int J Med Inform. 2006 Jun;75(6):418-29.

[21] Santorini B. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical Report, Department of Computer & Information Science, University of Pennsylvania, 3rd revision, 2nd printing edition, 1990.