Annotating a Large Representative Corpus of Clinical Notes for Parts of Speech

Narayan Choudhary, Parth Pathak, Pinal Patel, Vishal Panchal
ezDI, LLC.
{narayan.c, parth.p, pinal.p, vishal.p}@ezdi.us

Abstract

[We report of the procedures of developing a large representative corpus of 50,000 sentences taken from clinical notes. Previous reports of annotated corpus of clinical notes have been small and they do not represent the whole domain of clinical notes. The sentences included in this corpus have been selected from a very large raw corpus of ten thousand documents. These ten thousand documents are sampled from an internal repository of more than 700,000 documents taken from multiple health care providers. Each of the documents is de-identified to remove any PHI data. Using the Penn Treebank tagging guidelines with a bit of modifications, we annotate this corpus manually with an average inter-annotator agreement of more than 98%. The goal is to create a parts of speech annotated corpus in the clinical domain that is comparable to the Penn Treebank and also represents the totality of the contemporary text as used in the clinical domain. We also report the output of the TnT tagger trained on the initial 21,000 annotated sentences reaching a preliminary accuracy of above 96%.]

1 Introduction

Automated parts of speech (PoS) annotation have been an active field of research for more than 40 years now. Obviously, there are quite a few of tools already available with an impressive accuracy returns (Toutanova et al, 2003; Shen et al., 2007; Spoustov´a et al., 2009; Søgaard, 2010). This is true in the general domain text such as news reports or general domain articles. But when it comes to a niche area like clinical domain, no automated parts of speech taggers are readily available nor has there been any report of any such large corpus developed that meet the standards as set out in the general domain. Interest has grown now as NLP is sought after in the clinical domain, particularly for the task of information extraction from clinical notes.

There have been previous attempts for creating PoS annotated corpus in the clinical domain (Tateisi et al., 2004; Pakhomov et al, 2006; Albright et al., 2013). All of these corpora are relatively small and the PoS taggers trained on them have not been shown to reach above 96% in the clinical domain. Attempts at adapting a general domain PoS tagger to work better for clinical domain include Easy Adapt (Daumé H., 2007) and ClinAdapt (Ferraro et al., 2013). But none of these two adaptation methods enhance the accuracy levels to more than 95%.

Given that the text in clinical notes is radically different from what appears in the general domain, the general domain English PoS tagger models do not perform well on the clinical text. Our experiments with three such general domain taggers, namely Charniak (Charniak and Johnson, 2005), Stanford (Klein and Manning, 2003) and OpenNLP, yielded not more than 95% accuracy. This motivated us to take a radical step of developing a fresh parts of speech annotated corpus comparable to the Penn Treebank. Well, we are aware that it is going to take a lot of money, time and effort. But we also believe that it is necessary if we need better NLP tools for this domain.

2 The Representative Corpus

To ensure that we have a representative corpus, we sampled a corpus of more than 750,000 documents from 119 providers (hospitals and specialty clinics). The biggest challenge was to take a representative sample of documents from various specialties and different work types. Thanks to the metadata information available in our internal repository, this was solved in a rather easier way although we did need to look for information on classification and sub-classification of the domain manually.

LAW VIII - The 8th Linguistic Annotation Workshop, pages 87–92,
Dublin, Ireland, August 23-24 2014.

2.1 Sampling Task

Out of these 750,000 documents, we selected only ten hospitals for document sampling as they were large providers with a greater number of note count and provided a diversity of the specialty doctors/providers dictating the clinical notes. These ten hospitals amounted to a total of 237,110 documents written by 508 doctors of roughly 97 clinical specialties. A summary of this is given in Appendix A.

2.2 Sentence Clustering

We have used 237,110 documents for the process of selecting the sample sentences undergoing the PoS annotation. All of these documents were classified into different categories based on their work types (operative notes, admission notes, discharge summaries etc.), service line (cardiology, oncology, medicine, ambulatory etc.), section headers (History of Present Illness, Chief Complaints, Physical Examination, Laboratory Data etc.). Based on this classification, we have selected a sample of 10,000 documents fairly representing the 237,110 documents selected in the first phase.

These 10,000 documents were parsed using the Charniak's full syntactic parser (Charniak & Johnson, 2005). After some modifications, the Charniak parser on clinical data gives an accuracy of about 95% at the PoS level. A graph based string similarity algorithm was used to find similar sentences from these 10,000 documents. A summary of what it yielded is as follows:

Total Number of Sentences:	704,271
Total Number of Unique Sentences:	365,518
Total Number of Unique Patterns:	234,909

The unique patterns were clustered together using a hierarchical clustering algorithm. Patterns were grouped together by calculating the Euclidean distance with a threshold similarity of 80 or more. Following this method, we got a total of 3,768 patterns that represented all of the unique patterns. We call them pattern heads.

By giving a proportional weightage to each of these pattern heads as per their occurrence in the unique patterns, we derived a total of 56,632 sentences. While no two sentences selected are same, about 41% of the patterns in the sample corpus have a frequency of more than 1.

Appendix B shows example pairs of sentences having the same tag pattern and Appendix C shows example pairs of sentences having similar pattern.

The final selected candidate sentences also contained quite a few junk sentences (which came of course from the clinical notes themselves) or some very frequent smaller patterns (e.g. date patterns), we manually removed them to get a total of 49,278 sentences with a total word count of 491,690 and an average per sentence word count of 9.97. The greatest number of token for a sentence was found to be 221 in the sampled corpus (while the same in the original, actual corpus is 395).

3 Annotation Method

As against the common practice of semi-automatic method of annotating text, we purposely chose to annotate the text from scratch. It has been reported that tools do affect the decisions of the annotators (Marcus et al., 1993). We asked the annotators to use simple notepad and for each of the tokens they had to key in the appropriate PoS label. Tokenization and sentence boundary detection were automatically done before it went to the annotators.

As against the common practice of engaging annotators with a medical background and training them into linguistic annotation (Pakhomov et al., 2006; Fan et al., 2011), we purposely chose to engage linguists and train them into medical language. The annotators were all graduate level researchers in linguistics and had a deep knowledge of theoretical syntax. As next step in linguistics analysis after PoS tagging is syntactic parsing or chunking, the linguists were also motivated to learn about the goals of this task i.e. we informed them about our interests in developing a chunker and a parser afterwards. This information helps the annotators to think in terms of making syntactic tree while assigning a PoS tag. For example, there is always confusion among the tag pairs IN/RP, VBN/JJ and so on. But if one can try drawing a syntactic tree, the confusion gets cleared. While training annotators with medical background in linguistics for the task of PoS tagging may seem rather easy, the same cannot be said for syntactic tree formation. Besides, the linguists always had the choice of consulting medical experts

(medical coders, medical transcriptionists with more than 5 years of experience) in case any phrase had to be explained in terms of its meaning.

Training sessions were held for linguists for first 15 days during which differences were brought to fore and a consensus was reached. This period was strictly for training purposes and text annotated during this period was validated more than thrice before getting included in the final corpus. After this training period, an inter-annotator agreement round was run with 10,000 sentences distributed to four annotators in turn. Each file was annotated by at least two annotators. The differences were then compared and arbitrated by a third annotator who discussed the conflicting cases with the initial annotators and brought a consensus among them.

Inter-annotator agreement at the start of this phase was 93% to 95%. This after a month increased to a consistent 97% to 100%. We are at the end of this phase and the accuracy is consistently close to 99%. Also of note is the fact that apart from the initial 5 days of face-to-face training session, the annotators never sit together and they work remotely from the convenience of their location and have a flexible time. We also ensured that they do not work long hours at a stretch doing this job as we know that this is a tedious job and cannot be done in a hurry. For a full-time annotator, the target goal was annotation of 1600 word per day (8 hours) and for the part-time annotators, it was half of that. They were always encouraged to come up with any issues for a weekly discussion on the conflicting or confusing cases.

For the later phases of annotation process, it is ensured that each annotation is validated by at least one other annotator. If disagreements arise, arbitration is done by involving a third annotator following a discussion.

As the text might contain tokenization errors, sentence boundary detection errors and other grammatical or typographical mistakes, the annotators are asked to document them in a separate spreadsheet. The sentences themselves are sacrosanct to the annotators and they can at the most make changes in separating the hyphenated words if they are not properly hyphenated by the tokenizer and document this change.

4 Annotation Guideline

Barring a couple of new tags, the annotation guideline largely follows the Penn Treebank PoS annotation guidelines (Santorini, B., 1990) and takes inputs from various other guidelines such as the Penn Treebank II parsing guidelines (Bies et al., 1995) and MiPACQ guidelines (Warner et al., 2012). A new tag that we have added on top of the Penn tagset looks for marking a difference between the expletive “it” and the pronominal “it” as it helps in tasks like anaphora resolution. The new tag for the expletive use of ‘it’ is given as “EXP”. The tagset contains a total of 41 tags. The other four tags are HYPH, AFX, GW and XX. These tags are well described in the MiPACQ guideline.

As we have also seen the PoS labels given to the Penn Treebank data, we find that we are differing in assigning the tag to some of the words. For example, for the temporal expressions like “today”, “yesterday” and “tomorrow”, the tag in Penn corpus is invariably NN while we make a difference in their adverbial use and nominal use and assign the tag accordingly as “RB” or “NN”.

5 Initial Training Results

After 4 months of annotation, we achieved a total of 21 thousand sentences annotated. For an experimental run, we trained a tagger to test how far we can go with this data. We implemented a modified version of the TnT (Trigram and Tag) (Brants, 2000) algorithm to train a PoS tagger. This tagger was given an input of 17,586 sentences containing a total of 158,330 words and was tested against 3,924 sentences containing a total of 38,143 words.

Without giving any extra features apart from the ones mentioned in Brants, we got a total of 2,621 sentences and 36,234 words annotated correctly. That is the TnT out-of-the-box accuracy was 95.00% as against the Charniak out-of-the-box accuracy of 91.36%.

We also compared the same test data against the Charniak parser (without the resource of tag dictionary and the rules). We find that the current tagger was actually performing better. Results improved by 0.33% if we modified the algorithm to handle unknown words using suffixes from the medical domain. These suffixes were collected specifically from the medical domain and were such for which a single tag could be given.

We also experimented with another method for improvements. This included using a dictionary of unambiguous words (words having single tags invariably, for adjectives and verbs only) and resetting the emission probability to 1 for them. These two improvement techniques combined enhanced the results by 1.24% to push the accuracy to 96.24%.

Given that a fraction of our corpus is giving us 95% accuracy which is at par with or better than reported anywhere else for PoS tagging task in the domain of clinical NLP, we believe that the results should only improve once we increase the training data and apply the improvement techniques available in the book.

6 Conclusion

There is a paucity of good and large enough annotated corpus in the domain of clinical NLP. The existing corpora are small although extensive analysis has been done on them. Our effort through this project is to fill the gap of having a large corpus comparable to the Penn Treebank.

In this paper we described an ongoing effort to create a sample corpus of clinical notes across most of the sub-domains and including all the different types of linguistic styles in this domain. We have also used a novel method for creation of a representative corpus which can be said to represent the whole of the clinical text in current practice across providers within United States.

As compared to semi-automated methods of annotation practiced even in big corpus like the Penn Treebank, we are following a fully manual process of annotation where the annotators are only given contextual information and no other help or props are provided apart from the guidelines to fasten the annotation process. We obtain an inter-annotator agreement of 98.93 and we believe that this is the best approach to go for this task.

Using the basic TnT algorithm we also train a tagger using 30% of our data (17,500 sentences) annotated in the initial 3 months of the project and achieve a baseline accuracy of 95%. We expect that our accuracy should improve to more than 98% once we train the same algorithm on all the 50,000 annotated sentences.

After the completion of the project, we may release this corpus for research use.

Acknowledgements

We acknowledge the contribution of the linguists at JNU, New Delhi namely, Oinam Nganthoibi, Gayetri Thakur, Shiladitya Bhattacharya and Srishti Singh. Thanks also to Suhas Nair and Hiral Dawe for their help in making us understand the nuances of the clinical text. We would also like to thank Prof. Pushpak Bhattacharya and Dr. Girish Nath Jha for their advice.

References

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann, Marcinkiewicz and Britta Schasberger. 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. TR, University of Pennsylvania

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler IV, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, Wayne Ward, Martha Palmer, Guergana K Savova. 2013. *Towards Comprehensive Syntactic and Semantic Annotations of the Clinical Narrative*. J. Am. Med. Inform Assoc. 20:922–930

Thorsten Brants. 2000. TnT: A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Conference on Applied Natural Language Processing. Pp. 224-231.

Hal Daumé III. 2007. *Frustratingly Easy Domain Adaptation*. In: Proceedings of 45th Ann Meeting of the Assoc Computational Linguistics, 2007; 45:256–63.

Eugene Charniak and Mark Johnson. 2005. *Coarse-to-Fine n-best Parsing and MaxEnt Discriminative Reranking*. ACL'05.

Jung-wei Fan, Rashmi Prasad, Romme M. Yabut, Richard M. Loomis, Daniel S. Zisook, John E. Mattison and Yang Huang. 2011. *Part-of-Speech Tagging for Clinical Text: Wall or Bridge between Institutions?* In: AMIA Annu Symp Proc 2011 22; 2011:382-91.

Jeffrey P Ferraro, Hal Daume III, Scott L DuVall, Wendy W. Chapman, Henk Harkema and Peter J Haug. 2013. *Improving Performance of Natural Language Processing Part-of-Speech Tagging on Clinical Narratives through Domain Adaptation*. Journal of the American Medical Informatics Association. 2013; 20:931–939.

Dan Klein and Christopher D. Manning. 2003. *Accurate Unlexicalized Parsing*. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational Linguistics 19: 313–330

Serguei V. Pakhomov, Anni Coden and Christopher G. Chute. 2006. *Developing a Corpus of Clinical Notes Manually Annotated for Part of Speech*. International Journal of Medical Informatics. 75(6):418-429

Beatrice Santorini. 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Lui Shen, Giorgio Satta and Arvind Joshi, 2007. *Guided Learning for Bidirectional Sequence Classification*. In: ACL 2007.

Anders Søgaard. 2010. *Simple Semi-Supervised Training of Part-of-Speech Taggers*. In: Proceedings of the ACL 2010 Conference Short Papers. 205–208

Drahomira J. Spoustova, Jan Hajic, Jan Raab and Miroslav Spousta. 2009. *Semi-Supervised Training for the Averaged Perceptron POS Tagger*. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). 763–771

Yuka Tateisi and Jun'ichi Tsujii. 2004. *Part-of-Speech Annotation of Biology Research Abstracts*. In: the Proceedings of 4th International Conference on Language Resource and Evaluation (LREC2004). IV. Lisbon, Portugal, pp. 1267-1270

Kristina Toutanova, Dan Klein, Christopher D. Manning and Yoram Singer. 2003. *Feature-rich Part-of-Speech Tag-ging with a Cyclic Dependency Network*. In: NAACL 73. 252–259

Colin Warner, Arrick Lanfranchi, Tim O'Gorman, Amanda Howard, Kevin Gould and Michael Regan. 2012. *Bracketing Biomedical Text: An Addendum to Penn Treebank II Guidelines*. http://clear.colorado.edu/compsem/documents/treebank_guidelines.pdf (accessed 11 May, 2014)

Appendix A: Summary of the Medical Sub-Domains Included in the Sample Corpus of Clinical Notes

Domains	Count	Sub-Specialties	Doctor Count	Top Level Domains	Note Count	Sub-Specialties	Doctor Count
Family Medicine	12515	9	58	Pathology	4901	1	6
Vascular and Thoracic Surgery	2447	7	11	Obstetrics	3771	1	7
IM_Cardiology	11555	6	40	IM_After Hours Care	3072	1	5
IM_Pulmonology	6563	6	21	Urology	2976	1	13
Emergency Medicine	12742	5	28	IM_Neurology	2796	1	12
Oncology	8325	4	11	IM_Hematology	1475	1	1
IM_Nephrology	5684	4	24	IM_General Medicine	1457	1	9
Unclassified	1941	4	4	IM_Pediatrics	1326	1	13
IM_Infectious Diseases	376	4	11	Anesthesiology	1211	1	1
Hospitalist	17767	3	12	IM_Oncology	1138	1	4
IM_Internal Medicine General	15751	3	56	Psychiatry	827	1	5

Surgery	9287	5	33	Neurosurgery	729	1	5
Otorhinolaryngology	605	3	6	IM_Physician Assistant	437	1	4
Radiology	84635	2	16	Podiatry	345	1	10
IM_Gastroenterology	6592	2	23	Opthalmology	321	1	3
IM_Physical Medicnie and Rehabilitation	5495	2	6	Nurse Practitioner	314	1	4
Orthopedics	5480	2	19	IM_Pain Management	305	1	2
Obstetrics & Gynecology	1243	2	16	IM_Occupational Medicine	82	1	1
IM_Geriatrics	103	2	2	IM_Rheumatology	77	1	2
IM_Hospice Care and Palliative Medicine	153	2	2	IM_Endocrinology	31	1	2

Appendix B: Example of Sentences having the same pattern

Sentence	Another Sentence With Same Pattern
ALLERGIES : He is allergic to procaine .	ALLERGIES : HE IS ALLERGIC TO IODINE .
ABDOMEN : Soft with no tenderness .	Abdomen : Soft with no organomegaly .
He had an unknown syncopalepisode .	He underwent a third cardiopulmonary resuscitation .
There was no significant ST depression .	There was no distal pedal edema .
There was no associated mass shift .	There was no apparent air leak .
The sheath was removed from the sling material .	The patient was resuscitated in the emergency room .
The patient was intubated in the emergency room .	The patient was placed on a CPAP mask .

Appendix C: Example of Sentence Header and Similar Patterns

Header Sentence	Similar Sentence
The patient was admitted into the hospital under observation .	The patient was hospitalized for this in 04/12 .
Sodium is 131 , potassium is 3.9 , chloride is 104 , bicarbonate is 23 , glucose is 174 , BUN is 12 , and creatinine is 0.82 .	Total protein is 7.4 , albumin is 4.8 , total bili 0.3 , alkphos is 99 , AST is 53 , ALT is 112 , serum osmo is 271 .
LUNGS : Lung sounds reveal still scattered wheezes .	LUNGS : Lung reveals some scattered wheezes .
ALLERGIES : He is allergic to sulfa medications .	ALLERGIES : He has no allergies to medications .
Pleasant Caucasian gentleman in no acute distress.	She is in no apparent distress.
Left L5-S1 stenosis with associated left S1 radiculopathy.	Left hip impingement syndrome with probable labral tear.
Lab work today shows the following hemoglobin 11.7 , white cell count 9.8 , platelet count is 59.	Shows hemoglobin is stable , WBC count is stable , and platelet count is stable.