A Treebank for the Healthcare Domain

Oinam Nganthoibi
Computational Linguist
ezDI Inc. Kentucky
oinam.n@ezdi.us

Diwakar Mishra
Computational Linguist
ezDI Inc. Kentucky
diwakar.m@ezdi.us

Pinal Patel
Team Lead, Research
ezDI Inc. Kentucky
pinal.p@ezdi.us

Narayan Choudhary
Lecturer cum Junior Research Officer
CIIL, Mysore
choudharynarayan@gmail.com

Hitesh Desai
Research Engineer
ezDI Inc. Kentucky
hitesh.d@ezdi.us

Abstract

This paper presents a treebank for the healthcare domain developed at ezDI. The treebank is created from a wide array of clinical health record documents across hospitals. The data has been de-identified and annotated for constituent syntactic structure. The treebank contains a total of 52053 sentences that have been sampled for subdomains as well as linguistic variations. The paper outlines the sampling process followed to ensure a better domain representation in the corpus, the annotation process and challenges, and corpus statistics. The Penn Treebank tagset and guidelines were largely followed, but there were many syntactic contexts that warranted adaptation of the guidelines. The treebank created was used to re-train the Berkeley parser and the Stanford parser. These parsers were also trained with the GENIA treebank for comparative quality assessment. Our treebank yielded greater accuracy on both parsers. Berkeley parser performed better on our treebank with an average F1 measure of 91 across 5-folds. This was a significant jump from the out-of-the-box F1 score of 70 on Berkeley parser's default grammar.

1 Introduction

There is severe paucity of data in healthcare due to the confidentiality regulations entailed. However, the importance of domain specific training data cannot be denied. It is a well acknowledged fact that systems trained on the general domain do not perform well in highly specialized domains like healthcare (Jiang et al., 2015; Zhang et al., 2015; Ferraro et al., 2013). The research is further hindered for tasks that require a large volume of annotated data such as syntactic parsing.

Parsing is one of the complex natural language processing (NLP) tasks. Its complexity is inherited from syntax. Syntactic annotation is based on phrase structure grammar which posits a universal framework based on well-formedness conditions (Chomsky, 1965, 1993, 1995). However, these frameworks are modeled on formal language and therefore fail to account for ungrammaticality or variations in style. A universal syntactic framework even for a well-studied language like English is not established due to these reasons. Clinical healthcare data is an apposite example. It is populated with ungrammatical fragments and domain specific idiosyncrasies that cannot be accounted by standard grammatical rules. Therefore, the annotation task involves a high level of complexity and subjectivity. This paper showcases specific examples that justified adoption of new rules that are not postulated under the Penn Treebank guidelines (Bies et al., 1995). This is domain specific annotation. This approach has been rewarding. The Berkeley parser (Petrov et al., 2006) trained on this domain specific treebank gave a high F1 score of 91.58 using ParsEval (Harrison et al., 1991) method of evaluation. This is a remarkable improvement from the F1 of 70 that was attained on the parser's default grammar model.

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 144–155
Santa Fe, New Mexico, USA, August 25-26, 2018.

2 Related work

There are various types of corpora available in the field of clinical/medical NLP research. Good examples of raw text corpora include Stockholm corpus (Dalianis et al., 2012) which contains over a million patient record documents; Mayo Clinic clinical notes, referred in Wu et al. (2012) which contains 51 million documents, and a public repository of medical documents available at ClinicalTrials.gov (Hao et al., 2014). Zweigenbauma et al. (2001) also created a balanced raw text corpus annotated with meta-information to represent the medical domain sub-language.

Some corpora are annotated for part-of-speech (PoS), such as GENIA corpus (Tateisi and Tsujii, 2004) and the corpus of Pakhomov et al., (2006) while others are annotated and trained for named entity recognition (NER) such as Orgen et al. (2007), who have created a medical NER evaluation corpus that contains 160 clinical notes, 1556 annotations of 658 concept codes from SNOMED CT. Wang (2009) also reports training an NER system on a corpus of Intensive Care Service documents, containing >15000 medical entities of 11 types.

Alnazzawi et al. (2014), GENIA corpus version 3.0 (Kim et al., 2003) and CLEF corpus (Roberts et al., 2009) are examples of semantically annotated corpora. The source of GENIA corpus is 2000 research abstracts from MEDLINE database and is limited to specific type of documents while Alnazzawi's (2014) corpus is limited to covering only congestive heart and renal failure. BioScope corpus (Vinczer et al., 2008) is annotated for uncertainty, negation and their scope; THYME corpus (Styler et al., 2014) is annotated for temporal ordering using THYME-TimeML guidelines, an extension of ISO-TimeML; and Chapman et al. (2012) have annotated a corpus of 180 clinical reports for all anaphora-antecedent pairs. Xia and Yetisgen-Yildiz (2012) describe 3 clinical domain corpora, in which, the first corpus is annotated at the sentence level; the second corpus is annotated at the document level, for presence of pneumonia and infection score in X-ray reports; and the third corpus is annotated for pneumonia detection per patient in ICU reports. Some researchers have combined parse trees and multiword entities for specific tasks such as multiword entity recognition (Finkel and Manning, 2009) and entity relation identification (Shi et al., 2007). Cohen et al. (2005) list and classify six publically available biomedical corpora, namely, PDG, Wisconsin, GENIA, MEDSTRACT, Yapex and GENETAG, according to various corpus design features and characteristics.

Apart of these, work such as Pathak et al. (2015), have customized Unified Medical Language System (UMLS) thesaurus for concept unique identifier (CUI) detection as part of disorder detection.

In literature closely related to the work presented here, there are treebanks (syntactically annotated corpora) that use customized or original Penn Treebank guidelines (Bies et al., 1995). Albright et al. (2013) have annotated 13091 sentences of MiPECQ corpus for syntactic structure, predicate-argument structure and UMLS based semantic information. Fan et al. (2013) have customized Penn parsing guidelines to handle ill-formed sentences and have annotated 1100 sentences for syntactic structure. A subset of GENIA corpus, 500 abstracts, is also annotated for syntactic structure (Tateisi et al., 2005) using GENIA corpus manual (Kim et al., 2006). This is further extended to 1999 abstracts (GENIA project website). These three treebanks have sentences annotated for constituency structure. There are also treebanks annotated with dependency structure such as The Prague Dependency Treebank (Hajic, 1998).

As evident from the work listed above, there has not been any attempt of corpus creation to the expanse of the project presented here. Our corpus exceeds in quantity with 52053 sentences covering a variety of sentence structures from various document types and sources. Our work also differs in the corpus sampling process. MiPACQ corpus consists of randomly selected clinical notes and pathology notes from Mayo Clinic related to colon cancer. GENIA corpus is a set of abstracts from MEDLINE database that contain specific keywords. We have followed a sampling procedure that takes into consideration sentence patterns and domain representation. Our corpus sampling method covers the clinical domain on a large scale by giving representation to a variety of hospitals, specialties and document types. A more detailed comparison of corpus structure between our work and the GENIA Treebank (biomedical domain) and Wall Street Journal section of Penn Treebank (general domain) is shown in Section 4.

3 Creation of the Treebank

The task of treebank creation can be divided into two major parts - data sampling and annotation/bracketing. Section 3.1 describes how the data was sampled from clinical documents of different hospitals and specialty clinics. Section 3.2 discusses special cases of annotation that are peculiar to this domain.

3.1 Data Sampling

The current corpus has been assembled over time from different databases. The first set was extracted from an internal database of 237,100 documents from 10 hospitals in the US from the year 2012-2013. These hospitals were selected due to the fact that they were large establishments housing variety of specialties and therefore a good resource for different types of documents. These documents were classified into different work types, service lines and section heads, based on which, 10,000 representative documents were manually selected. A graph-based string similarity algorithm was used to find similar sentences which resulted in a collection of unique patterns. A sentence clustering algorithm was then used to narrow them down into pattern heads that were representative of all the unique patterns. The final corpus was selected by giving proportional weight to each pattern head. A detailed discussion of the methodology is found in Choudhary et al., (2014). This set was created for the development of a part-of-speech (PoS) tagger. 38,000 sentences from this dataset were used as the base for this parsing project as well. The Table 1 below shows the sub-domains included in this dataset.

IM_After Hours Care	IM_Endocrinology	Pathology	IM_Oncology
Vascular and Thoracic Surgery	Emergency Medicine	IM_Occupational Medicine	IM_Internal Medicine General
Obstetrics	Psychiatry	Anesthesiology	Neurosurgery
IM_Pain Management	Family Medicine	Urology	Opthalmology
IM_Physical Medicine and Rehabilitation	IM_General Medicine	IM_Physician Assistant	Nurse Practitioner
IM_Nephrology	IM_Hematology	IM_Pediatrics	Otorhinolaryngology
IM_Gastroenterology	IM_Neurology	IM_Geriatrics	Radiology
IM_Infectious Diseases	IM_Rheumatology	Hospitalist	Orthopedics
Obstetrics & Gynecology	IM_Cardiology	Oncology	Unclassified
Podiatry	Surgery		

Table 1: Subdomains included in Dataset 1 from Database 1

The second dataset was sampled from a different database of 3 hospitals in the US containing 1,473 documents dated April to September, 2016. This database was used to update the corpus with current clinical data. The need to update arose from the observation that the style of electronic health record documentation has changed significantly between 2012 and 2016. It is evident by the relative proportion of S nodes (well-formed sentences/clauses) and FRAG nodes (fragments) between the two datasets. Dataset 1 contains 29435 S nodes and 15118 FRAG nodes (S-FRAG ratio of 1:0.513), while Dataset 2 contains 2786 S nodes and 12102 FRAG nodes (S-FRAG ratio of 1:4.343).

The 1,473 documents from these three hospitals were categorized according to their work types. It contained 535 documents of 25 work types from hospital A, 93 documents of 20 work types from hospital B, and 845 documents of 14 work types from hospital C. These work types were grouped into three broader categories – Admission, Progress and Discharge. So, for example, worktypes "History and Physical" and "ER Physician Document" were kept under the 'Admission' category; "Preprocedure Checklist" and "Anesthesia postoperative Note" were kept in 'Progress' category, and so on. Then, a certain number of documents were manually selected from each of the three categories, keeping in mind a balanced ratio of the original work types. The following Table 2 shows the number of documents selected from each hospital from each category.

Hospitals	Admission	Progress	Discharge	Total
A	8 / **36**	42 / **457**	10 / **42**	60 / **535**
B	4 / **16**	20 / **57**	8 / **20**	32 / **93**
C	6 / **26**	50 / **775**	8 / **44**	64 / **845**
All Hospitals	18 / **78**	112 / **1289**	26 / **106**	156 / **1473**

Table 2: Number and type of documents from each category included in Dataset 2 from Database 2

After a simple algorithm to remove duplicate sentences, this process resulted in a dataset of 19,011 unique sentences. 12,000 sentences were eventually selected from this source.

The third sampling stage was done on the basis of 'rare syntactic pattern'. Low frequency patterns were extracted from the corpus compiled so far. These patterns were identified based on grammatical categories, keywords and subject to human judgment in the background of extended interaction with the domain. These patterns were then converted to regular expressions, which was used to extract similar sentences from Database 1 and 2. For example, sentences with wh-questions have a low distribution in clinical texts and were therefore left out during the sampling methods employed so far. These were added. Low frequency closed grammatical categories like prepositions were also added to the corpus. This method contributed to around 2,000 sentences. The Table 3 below is a summary of the corpus creation in the three steps.

Source	No. of Hospitals	Dated	Method	Dataset
Database 1	10	2012-13	sub-domain selection + sentence clustering	(1) 38,000 sents
Database 2	3	2016	sub-domain selection + duplicate removal	(2) 12,000 sents
Database 1 + 2	13	2012-16	rare syntactic pattern + regular expression	(3) 20,53 sents
Total corpus				**52053 sents**

Table 3: Source databases, methods and resulting datasets

3.2 Challenges in annotation

The corpus was annotated for phrase structure following a customized version of the Penn Treebank guidelines (Bies et al., 1995). Null elements and function tags have not been incorporated at this point. This section discusses the data structures where deliberate and novel guidelines were adopted.

3.2.1 Binary branching vs Tertiary branching

Binary branching was adopted in post phrase structure syntactic theories viz. Government and Binding (Chomsky, 1993) and Minimalist Program (Chomsky, 1995). In the binary structure, the relations between parts of the sentences are expressed through hierarchy. This hierarchy is also crucial for the linear ordering of the sentence. However, there were many instances where binary branching could not be adopted. For example, there is no hierarchy in conjunction and therefore conjunctive phrases/clauses/sentences or multiple elements inside the NP were kept in multiple branches, which extended beyond 3 tokens or more. For example, Figure (1) has four phrase branches because there are four different fragments conjoined by commas and a conjunctive word. Each branch has a composite meaning that does not have a hierarchical relationship with the others. The same is true for elements inside the noun phrase (NP). In clinical texts, an NP can have a token span of up to 5 and more. There is, again, no hierarchical relationship between NP internal elements. Figure (2) shows a typical NP in the clinical domain which contains numerals and symbols as part of the NP.

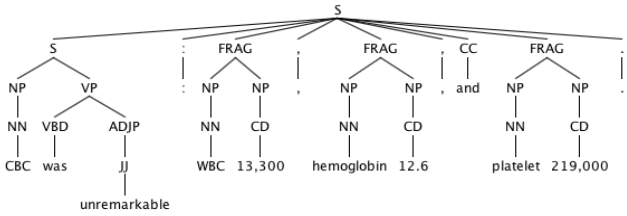

Figure 1: Multiple branch at the clause level

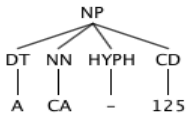

Figure 2: Tertiary branch inside an NP

Tertiary branching is also adopted for moved elements such as sentential adverbs and prepositional adjuncts that are topicalized. Other such data include section heads and list symbols that appear at the front of the phrase as shown in Figure (3)

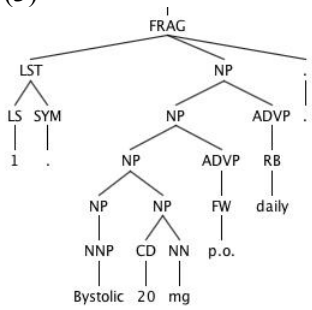

Figure 3: Tertiary branch in list items

This shows that multiple branching is present at the highest clausal/sentential level as well as at the lowest phrase internal elements. Given that there is no theoretical limit to conjunction or NP internal elements (especially within this domain), there is no limit to the number of branches as well.

3.2.2 Ambiguous categories

Clinical text contains Latin abbreviations indicating manner of medical dosage or manner of action. We adopted a principle to annotate the abbreviation based on the syntactic category of its translation or the full-form. For example, 'q. 8 h' stands for 'every 8 hours' and therefore annotated as an NP. 'IV' is tagged as JJ (adjectival token) and does not have a maximal projection when it functions as an adjective in a phrase like 'IV (intravenous) fluid'. However, it has a maximal projection ADVP (adverbial phrase) when it modifies a verb as in Figure (4). Beyond abbreviations, clinical texts also contain phrases like 'x 3' which stands for 'times 3' in the context of the test results such as 'test is negative/positive x 3' or in the context of a patient's condition as in 'the patient is oriented x 3'. Such phrases that have an adverbial flavor but do not explicitly function as adverbs are kept under NP. Such NPs are however post-modifiers of the preceding category and therefore they also form a maximal projection of their own as shown in Figure (5).

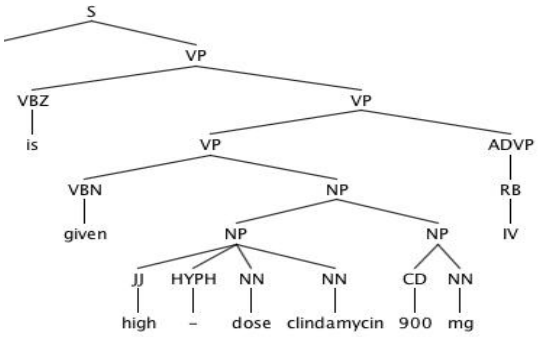

Figure 4: 'IV' forms an ADVP node

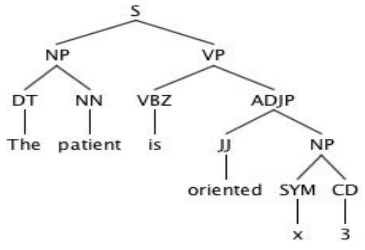

Figure 5: 'x 3' labelled as NP

Another form of ambiguity in category arises in clinical texts due to a practice of omitting the head of the phrase. This creates a mismatch between the rightmost PoS tag (the head of the phrase) and the maximal category. This violates the 'projection principle' which states that 'lexical structure must be represented categorically at every syntactic level' (Chomsky, 1986). However, this mismatch is deliberately maintained in our annotation for accuracy at the phrase level. For example, in Figure (6) 'celiac' stands for 'celiac artery' but the token 'artery' is absent. So, rightmost tag is JJ but the phrase label is kept as an NP.

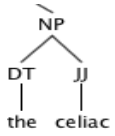

Figure 6: Mismatch between PoS and phrase labels

3.2.3 Multi-level NPs

Clinical data contains instances of multiple NPs modifying one another. We used right C-adjunction to account for these kinds of data. C-adjunction is a syntactic operation in which an element is added to the constituent of a category X by moving the element and adjoining it to a mother node above category X. Multi-level NP is peculiar to, as well as widely distributed in this domain. It is found mostly in the documentation of medical dosages. Each modifier such as the duration, manner or quantity is adjoined to the first NP as shown in Figure (7) below.

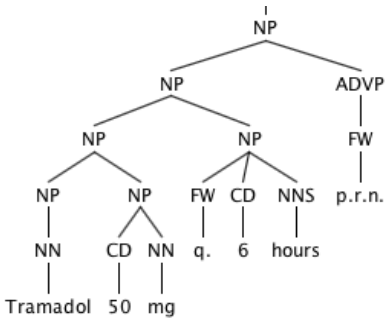

Figure 7: Multi-level NP

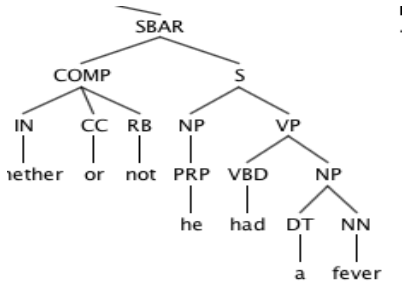

Figure 8: Multi-word complementizer under COMP

3.2.4 Multi-word Complementizer

Complementizers/subordinators can be multi-words. This is a phenomenon not peculiar to the clinical domain but nevertheless inadequately addressed in theoretical syntax or annotation literature. Some examples of multi-word complementizers are 'Even if', 'Whether or not', 'So that', 'As if', 'If and when', 'Should if' etc. To handle such words, we introduced the phrase label COMP which stands for Complemtizer Phrase, a commonly used in generative syntax. This phrase layer is necessary for accurate representation of syntactic objects and syntactic relations. Projection principle (Chomsky, 1986) allows only one head to project. Without the COMP layer, it would appear that both lexical categories in a multi-word complementizer are projecting to be the head of the SBAR. The COMP layer enables only one head to project at the phrase level.

3.2.5 FRAG

FRAG is the label used for fragmented sentences/clauses that arise due to transcription errors, grammatical errors or shorthand documentation. Its abundant occurrence in clinical health data creates much unwanted variation within the domain itself. The fragments however fall within identifiable patterns as follows:

- Isolated phrases: These are instances of medical dosage, description of patient status etc. written in shorthand. It can be any phrase, although, the majority present in this domain are NPs. (Figure 9)
- Copula dropped sentences: These are sentences where the copula ‘is’ or ‘are’ are missing. (Figure 10)
- Subject-less sentences: Existential subjects such as ‘It is’ as well as nominal subjects like ‘He/She/Patient’ are omitted from these sentences. (Figure 11)
- Irregular conjunctive phrases: These are sentences where two syntactically different objects are conjoined using punctuations. (Figure 12)
- Template data: These are sentences where the left token denotes a disease/condition and the right token gives value such as ‘present/absent’, ‘yes/no’ etc. (Figure 13)
- Incomplete sentences: These sentences are incomplete due to line break or transcription error. (Figure 14)

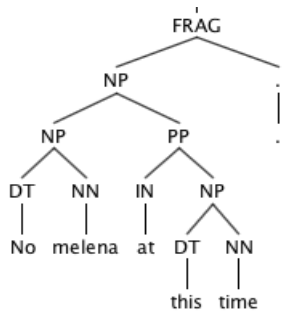

Figure 9: Isolated phrases

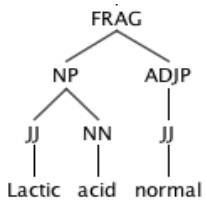

Figure 10: Copula drop

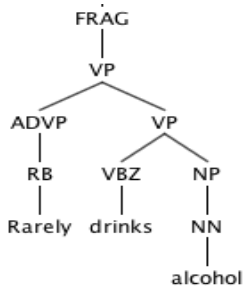

Figure 11: Subject less sentence

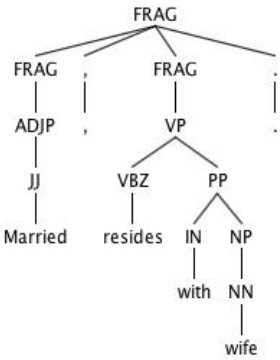

Figure 12: Irregular conjunction

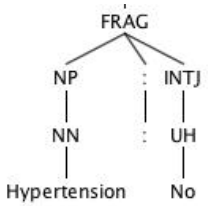

Figure 13: Template

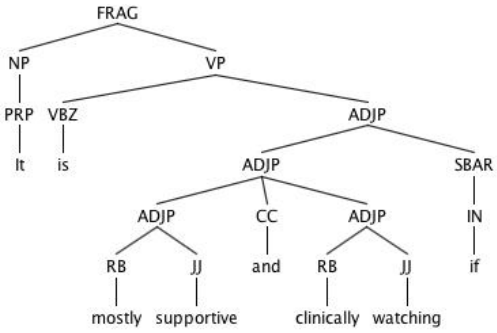

Figure 14: Incomplete sentences

These types of FRAGs can occur at the top sentential level as well as deep within the clause.

3.2.6 XXP

Missing data in this case does not mean ellipses (which are a part of syntactic transformation rules). These are data missing due to the de-identification process or incomplete transcription. In such cases, XXP is used to represent a placeholder node which can be computed in relation to other categories in the tree.

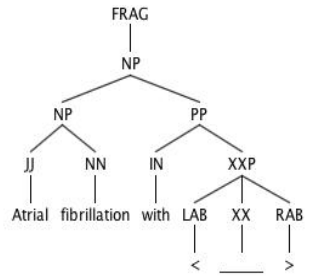

Figure 15: 'XX' and 'XXP' denoting missing elements

3.3 Inter-annotator agreement

The corpus was annotated by four linguistics students and reviewed by 2 in-house linguists. The students were first trained on clinical language annotation with 1000 sentences for one month to establish familiarity with the domain. Each annotator was then provided with sets of 50 PoS tagged sentences. PoS tagged sentences were provided to facilitate a better understanding of the meaning of the sentence. The annotated file was reviewed by two expert linguists. The annotators were instructed to follow the Penn Treebank guidelines (Bies et al., 1995). Given the anomalistic nature of the data as described in the previous sections, there were significant disagreements among the annotators. Weekly discussions were held to resolve the ambiguities and doubts. New rules were adopted based on these discussions. The annotator-reviewer disagreement is an indicator of the complexity of the task. Table 4 shows the inter-annotator agreement among the annotators and the two reviewers. The agreement between the four annotators was calculated using ParsEval (Harrison et al., 1991) F1 score on 500 sentences from each annotator. The inter-annotator agreement (IAA) between the two reviewers responsible for finalizing the corpus was 98.2%, based on 2000 sentences. Table 5 shows a comparative study with other treebanks. The IAA for GENIA corpus is reported to be 96.7% and 97.4% for two annotators respectively, as measured on 108 sentences, compared against 'gold standard' (Tateisi et al., 2005). The IAA for MiPACQ treebank is reported to be 0.926 (92.6%) (Albright et al., 2013).

	Reviewer1	**Reviewer2**	**Reviewer1- Reviewer2**
Annotator1	86.53	86.22	98.07
Annotator2	88.88	88.29	98.72
Annotator3	86.31	85.39	98.11
Annotator4	87.16	87.84	97.92

Table 4: Inter-annotator agreement calculated on 500 sentences from each annotator

	Method	**Evaluated sentences**	**IAA**
ezDI treebank	ParsEval F1	2000	98.2%
GENIA	Manual comparison against gold standard	108	96.7%, 97.4%
MiPACQ	EvalB F1	8% of total data	92.6%

Table 5: Comparison of inter-annotator agreement for various treebanks

4 Corpus Statistics

A treebank can be interpreted as a set of context free grammar (CFG) rules in the form of 'A → B C', where A is the higher node that branches into two lower nodes viz. B and C. For instance, our treebank

has 5580 unique CFG rules which is an indication of the size of the grammar model. A detailed examination of the nodes and labels and the relative frequency of these rules is also an indicator of the data structure contained in the corpus.

Table 6 is a comparison of the percentage of non-terminal nodes (phrase and clause labels) in our treebank, GENIA Treebank and Wall Street Journal (WSJ) section of the Penn Treebank. The most notable difference is the proportion of FRAG which is more than 100 times higher than the other two treebanks. This shows that ill-formedness in clinical texts is a norm rather than an exception. Other significant differences are in lower proportion of the WH-phrases (WHNP, WHPP, WHADVP) and PP (prepositional phrase), and higher proportion of LST (list marker), and UCP (unlike coordinated phrase). There is also a higher proportion of ROOT node in our treebank, which signifies less average sentence length. The table also shows that there is a high frequency of NP nodes in our treebank on par with the other two treebanks. In our case however, this may be an indication of the number of hierarchical noun phrases created due to multiple adjunctions of the type discussed in Section 3.2.3 (Figure 7).

Node name	ezDI tree-bank	GENIA	WSJ (PTB)		Node name	ezDI tree-bank	GENIA	WSJ (PTB)
-NONE-	--	--	7.15		ROOT/S1	10.23	4.55	4.44
ADJP	2.60	2.91	1.62		RRC	--	0.00024	0.005
ADVP	3.04	2.06	2.50		S	8.24	8.899	11.16
COMP	0.016	--	--		SBAR	2.02	2.18	3.41
CONJP	0.02	0.17	0.03		SBARQ	0.026	0.0017	0.026
FRAG	6.55	0.03	0.06		SINV	0.0019	0.012	0.233
INTJ	0.05	--	0.01		SQ	0.047	0.004	0.04
LST	0.167	0.061	0.006		UCP	0.112	0.059	0.053
NAC	0.002	--	0.049		VP	16.71	13.47	16.31
NP	40.97	47.96	39.09		WHADJP	--	0.0007	0.0059
NX	--	--	0.15		WHADVP	0.078	0.114	0.294
PP	8.64	15.19	10.64		WHNP	0.249	0.689	1.012
PRN	--	1.28	0.27		WHPP	0.009	0.073	0.043
PRT	0.13	0.009	0.29		XXP	0.036	--	--
QP	--	0.23	1.03					

Table 6: Comparative percentage of non-terminal nodes in three treebanks

5 Discussion

As expected, the out-of-the-box performance of open source parsers did not perform well when tested on our treebank. Berkeley parser (Petrov et al., 2006) gave an F1 score of 70 on its default grammar. For a comparative study, Stanford parser (Manning et al., 2014) and Berkeley parser (Petrov et al., 2006) were trained with our treebank. Training took place in two stages to assess the quality of the treebank and the parser. We also did a comparative study with Penn Treebank (PTB) style version of GENIA corpus distributed by McClosky (2009). The GENIA Treebank consisted of 18541 sentences while our treebank consisted of 52053 sentences. Both the corpora were divided into 20% test data and 80% training data. The performance of the parsers was evaluated with ParsEval (Harrison et al., 1991) method across 5 folds. The training results show that our treebank performs better on both the Stanford parser and Berkeley parser. The F1 score of 85.32 and 91.58 are also significantly higher than the original Berkeley score of 70.

Treebank → Parser ↓	GENIA TB	ezDI TB 30k	ezDI TB 52k
Test Sentences	3708	6011	10411
Stanford F1	83.36	87.65	85.32
Berkeley F1	87.38	92.69	91.58

Table 7: Comparison of treebanks trained on different parsers. Results in F1 score using ParsEval

Albright et al. (2013) have shown the importance of in-domain annotation. Even a small amount of in-domain annotated data can enhance the performance of different NLP components down the pipeline. They also suggest that more divergent data will result in less improvement. The same is also reflected in our training results. The accuracy on both parsers came down by an average of 2 percent when the corpus size increased from 30+k to 52+k. As noted in Section 3.1, more variety of sentences were added in Datasets 2 and 3, which were not part of Dataset 1. An open source parser trained with this treebank is being currently used to enhance the performance of systems like information extraction which ultimately improves the performance of the end products like computer assisted coding (CAC) and clinical document improvement (CDI). The treebank is not publicly available, but a parser trained with this treebank will be made available as a part of a clinical NLP service

6 Conclusion

This paper showed the processes entailed in the development of a representative treebank for clinical healthcare. An elaborate and meticulous data sampling is an important first step towards creating a treebank. Next, the annotation has to be domain specific in the sense that grammatical principles have to be adapted to tackle the variety of linguistic structures present in the domain. This also means that the resulting structures should follow a pattern that is not far removed from theoretical principles of phrase structure rules. This will create a grammar that can generate domain specific structures. Finally, the need for domain specific treebank is validated by the high performance of Stanford and Berkeley parsers. Further research should focus on automation of the annotation task and optimal use of the parser for various NLP tasks.

Acknowledgements

We acknowledge the contribution of the linguists at JNU, New Delhi, namely, Srishti Singh, Arushi Uniyal, Sakshi Kalra and Azzam Obaid. We also acknowledge the help of Dr. Binni Shah and Disha Dave in understanding domain specific concepts and expressions. We would also like to thank Prof. Pushpak Bhattacharya and Prof. Girish Nath Jha for their advice.

References

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950-966.

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. *Bracketing guidelines for Treebank II Style Penn Treebank project.* University of Pennsylvania. Retrieved, February 2015, from https://catalog.ldc.upenn.edu/docs/LDC99T42/prsguid1.pdf.

Christopher D. Manning, Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*:55-60.

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 and Guergana K. Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922-930.

David McClosky. 2009. Self-trained biomedical parsing. Retrieved March 8, 2018, from https://nlp.stanford.edu/~mcclosky/biomedical.html.

GENIA Project website. Retrieved March 8, 2018, from http://www.geniaproject.org/genia-corpus/treebank.

Fei Xia and Meliha Yetisgen-Yildiz. 2012. Clinical Corpus Annotation: Challenges and Strategies. *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) under LREC-2012*.

Hercules Dalianis, Martin Hassel, Aron Henriksson and Maria Skeppstedt. 2012. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. *Proceedings of Swedish Language Technology Conference*:17-18.

Jeffrey P Ferraro, Hal Daume III, Scott L Du Vall, 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*, 20:931–939.

Jenny Rose Finkel and Christopher D. Manning. 2009. Joint Parsing and Named Entity Recognition. *Proceedings of Human Language Technology: 2009 Conference of the North American Chapter of the Association of Computational Linguistics*:326-334.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi and Jun'ichi Tsujii. 2003. GENIA corpus – A semantically annotated corpus for bio-text mining. *Bioinformatics*, 19(1):i180-i182.

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi and Jun'ichi Tsujii. 2006. *GENIA Corpus Manual - Encoding schemes for the corpus and annotation*. Technical Report (TR-NLP-UT-2006-1). Tsujii Laboratory, University of Tokyo.

Jung-wei Fan, Elly W. Yang, Min Jiang, Rashmi Prasad, Richard M. Loomis, Daniel S. Zisook, Josh C. Denny, Hua Xu and Yang Huang. 2013. Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. *Journal of the American Medical Informatics Association*, 20(6):1168-1177.

Kevin Bretonnel Cohen, Philip V. Ogren, Lynne Fox and Lawrence Hunter. 2005. Corpus design for biomedical natural language processing. *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*:38-45.

Min Jiang, Yang Huang, Jung-Wei Fan, Buzhou Tang, Joshua C. Denny and Hua Xu. 2015 Parsing clinical text: how good are the state-of-the-art parsers? *BMC Medical Informatics and Decision Making*, 15(1):S2.

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

Narayan Choudhary, Parth Pathak, Pinal Patel, Vishal Panchal. 2014. Annotating a Large Representative Corpus of Clinical Notes for Parts of Speech. *Proceedings of 8th Linguistic Annotation Workshop*:87-92

Noam Chomsky. 1965. *Aspects of the theory of syntax*. MIT Press, Cambridge, Massachusetts.

Noam Chomsky. 1993. *Lectures on government and binding: the Pisa lectures*. 7th edition (1st edition, 1981). Mouton de Gruyter, Berlin and New York.

Noam Chomsky. 1995. *The minimalist program*. MIT Press, Cambridge, Massachusetts and London.

Noha Alnazzawi, Paul Thompson and Sophia Ananiadou. 2014. Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature. *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis*:69-74.

Parth Pathak, Pinal Patel, Vishal Panchal, Sagar Soni, Kinjal Dani, Amrish Patel, and Narayan Choudhary. 2015. *ezDI: A supervised NLP system for clinical narrative analysis*. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015):412-416.

Philip Harrison, Steven Abney, Ezra Black, Dan Flickinger, Claudia Gdaniec, Ralph Grishman, Donald Hindle, Robert Ingria, Mitch Marcus, Beatrice Santorini, Tomek Strzalkowski. 1991. Evaluating syntax performance of parser/grammars. *Proceedings of the Natural Language Processing Systems Evaluation Workshop, Berkeley (Rome Laboratory Technical Report, RL-TR-91-362)*.

Philip V. Orgen, Guergana K. Savova and Christopher G. Chute. 2007. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. *Proceedings of the 12th World Congress on Health (Medical) Informatics*:3143-3150.

Pierre Zweigenbauma, Pierre Jacquemarta, Natalia Grabara and Benoît Habert. 2001. Building a Text Corpus for Representing the Variety of Medical Language. *Studies in health technology and informatics*, 84(1):290-294.

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.

Slav Petrov, Leon Barrett, Romain Thibaux and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. *Proceedings of the 21st International conference on computational linguistics and the 44th annual meeting of the Association for Computational Linguistics*:443–440.

Stephen T. Wu, Hongfang Liu, Dingcheng Li, Cui Tao, Mark A. Musen, Christopher G. Chute and Nigam H. Shah. 2012. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. *Journal of the American Medical Informatics Association*, 19(e1):e149-e156.

Tianyong Hao, Alexander Rusanov, Mary Regina Boland and Chunhua Weng. 2014. Clustering clinical trials with similar eligibility criteria features. *Journal of Biomedical Informatics,* 52:112-120.

Veronika Vinczer, György Szarvas, Richárd Farkas, György Móra and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*:38-45.

Wendy W. Chapman, Guergana K. Savova, Jiaping Zheng, Melissa Tharp and Rebecca Crowley. 2012. Anaphoric reference in clinical reports: Characteristics of an annotated corpus. *Journal of Biomedical Informatics*, 45(3):507-521.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen and Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova and James. 2012. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143-154.

Yefeng Wang. 2009. Annotating and recognising named entities in clinical notes. *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*:18-26.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta and Jun'ichi Tsujii. 2005. Syntax annotation for the GENIA corpus. *Companion Volume to the Proceedings of Second international joint conference on natural language processing*:220-225.

Yuka Tateisi and Jun-ichi Tsujii. 2004. Part-of-Speech Annotation of Biology Research Abstracts. *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*:1267-1270.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda and Satoshi Nakamura. 2015. Ckylark: A more robust PCFG-LA parser. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*:41-45.

Zhongmin Shi, Anoop Sarkar and Fred Popowich. 2007. Simultaneous Identification of Biomedical Named-Entity and Functional Relations Using Statistical Parsing Techniques. *Proceedings of Human Language Technology: 2007 Conference of the North American Chapter of the Association of Computational Linguistics*:161-164.