Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 129–138
Marseille, 24 June 2022

Building a Biomedical Full-Text Part-of-Speech Corpus Semi-Automatically

Nicholas Elder, Robert E. Mercer, Sudipta Singha Roy
The University of Western Ontario
London, Ontario, Canada
nelder@uwo.ca, mercer@csd.uwo.ca, ssinghar@uwo.ca

Abstract

This paper presents a method for semi-automatically building a corpus of full-text English-language biomedical articles annotated with part-of-speech tags. The outcomes are a semi-automatic procedure to create a large silver standard corpus of 5 million sentences drawn from a large corpus of full-text biomedical articles annotated for part-of-speech, and a robust, easy-to-use software tool that assists the investigation of differences in two tagged datasets. The method to build the corpus uses two part-of-speech taggers designed to tag biomedical abstracts followed by a human dispute settlement when the two taggers differ on the tagging of a token. The dispute resolution aspect is facilitated by the software tool which organizes and presents the disputed tags. The corpus and all of the software that has been implemented for this study are made publicly available.

Keywords: semi-automatic corpus annotation, biomedical document annotation, part-of-speech

1. Introduction

Training and evaluating machine learning Natural Language Processing (NLP) systems require benchmark corpora annotated for the NLP task being learned. Manually curated gold standard corpora, the language resources that are typically used to train and test such systems, are unfortunately, costly to produce especially in domains requiring specialized knowledge to understand the text.

Our goal is to provide a large corpus of biomedical text annotated with part-of-speech (POS) using the Penn Treebank Tagset to facilitate the training of a deep learning model. Our current corpus, which we call BioPOSTAg, drawn from full-text biomedical articles, has 5 million sentences and we continue to work toward a corpus containing 35 million sentences. Due to the size of this corpus, no completely manual annotation is possible. An alternative to a gold standard annotated corpus is a silver standard corpus (Rebholz-Schuhmann et al., 2010). Therefore we have decided on a silver standard approach. The silver standard was first proposed to be generated in a fully automatic way (Rebholz-Schuhmann et al., 2011) using annotation systems and some method to harmonize their resulting annotations. Researchers continue with this practice (Sousa et al., 2019), while others incorporate some manual annotations (Eckart and Gärtner, 2016). Because our building of the silver standard corpus uses only two automated annotators, we need to have some human intervention to make decisions when the annotators disagree. Since this human intervention is added, the process that is described herewith is termed semi-automatic.

The BioPOSTAg corpus is evaluated by comparing the performance of a model trained on the silver standard corpus versus the same model trained on a human-annotated gold standard corpus on a POS tagging task. We have chosen the CRAFT corpus (Verspoor et al., 2012) as the gold standard training and test sets for this comparison.

Our contributions can be summarized as follows: a semi-automatic procedure to create a large silver standard corpus; a large corpus of complete biomedical articles annotated for part-of-speech has been built and is made available to the research community; and a robust, easy-to-use software tool that assists the investigation of differences in two annotated datasets facilitating the human dispute resolution aspect of the semi-automated procedure.

2. Background

Part-of-speech (POS) tagging assigns a POS to each token in a text. Modern POS taggers are trained using some form of machine learning. Training requires an annotated corpus. Training of a deep learning model requires a corpus with a large number of samples, in this case sentences with the tokens annotated for POS. The manually tagged gold standard corpora that have been built, e.g., the GENIA corpus (1997 abstracts) (Kim et al., 2003) and the CRAFT corpus (97 full-text papers) (Verspoor et al., 2012) are reasonably small. Having larger tagged corpora may be beneficial. In addition, while part-of-speech tagging in the biomedical literature genre has long been a topic of research (Kim et al., 2003; Tateisi and Tsujii, 2004), the early focus has been on POS tagging of article abstracts. POS tagging of complete article texts provides some subtle differences due to sentence structure and other writing and content issues (Cohen et al., 2010).

Complete biomedical article datasets are becoming available to the research community, so having machine methods that work with full papers is both feasible and critically important given the large amount of literature produced in this socially significant research field. Because manual annotation is costly, especially in the biomedical domain since it requires specialized knowledge, large annotated corpora of full text biomedical articles do not currently exist. The focus of the current study is the semi-automatic curation of a sufficiently large silver standard corpus of complete article texts annotated with POS tags that might boost the performance of deep learning trained POS taggers.

To provide the automatic aspect of this silver standard corpus curation task, this study uses two top ranked biomedical POS taggers: the popular Genia (Kim et al., 2003; Tateisi and Tsujii, 2004) and a variant of MedPost (Smith et al., 2004) that we call PostMed so as not to confuse it with the original but to pay homage to the original name and work. Genia was trained on a corpus of 1999 manually annotated MEDLINE abstracts. In addition to POS tagging, Genia's other abilities (named-entity tagging and chunk parsing) are not used in this study. MedPost is a POS tagger, but also of importance for this study, it can work with .nxml files: interpreting the xml tags, breaking the file into sentences, and performing tokenization. It was designed to work with MEDLINE abstracts, so a wrapper was provided by the second author giving PostMed, the modified version that works with full article texts (e.g., figures and tables are removed). These POS taggers have achieved over 98% and 97% accuracy, respectively, on MEDLINE citations.

Unlike other silver standard corpora building which use techniques developed for the type of data that is represented in the silver standard corpus, the techniques that we are using have been trained on MEDLINE abstracts whereas the data that we are annotating with our semi-automatic method are full-text articles. Full-text articles contain language that is not found in abstracts, such as references to figures and tables. So, the use of these two taggers could be considered akin to cross-domain tagging but obtaining good performance may not be as difficult as sometimes is the case with cross-domain tagging. Our hypothesis was that the outputs of these two part of speech taggers would perform reasonably well on this new type of data, that the number of differences would be manageable, and that human intervention would be able to enhance the final outcome. The second part of our hypothesis, that the number of differences would be manageable was overly optimistic. As a result, we developed a software tool, a data viewer, whose purpose was to organize these differences along different dimensions thereby facilitating our viewing of the differences in various ways.

3. Related Work

Research related to this study falls into three categories: corpora annotated for POS, POS taggers, and studies of the performance on full-text articles of taggers trained on article abstracts. Some small corpora annotated for POS based on clinical notes (Pakhomov et al., 2006) and on patient records (Huseth and Rost, 2007) have been built, the latter one being annotated semi-automatically. Because few biomedical corpora with POS annotations exist, methods such as cross-training have been used to circumvent this paucity of data, but the resulting performance tends to be low (Barrett and Weber-Jahnke, 2014). Adding a biomedical domain-specific corpus has been shown to improve results (Coden et al., 2005). MedPost (Smith et al., 2004) uses a lexicon that enumerates permitted POS tags for the most frequently occurring 10,000 words in MEDLINE to improve its performance (Smith et al., 2006). And, some improvements with cross-trained taggers have been reported by introducing specialized lexicons to address the problems associated with unknown words (Miller et al., 2007). It has been demonstrated (Tateisi et al., 2006) that because biomedicine has subdomains, performance drops when taggers are required to tag a subdomain that differs from the training subdomain. And, some results show excellent performance by off-the-shelf POS taggers (TnT (Brants, 2000)) for tagging clinical reports (Hahn and Wermter, 2004). Other POS taggers have been developed for the biomedical domain, some being better performers than others. dTagger (Divita et al., 2006), trained and tested on the MedPost corpus, performs with 95.1% accuracy. TcT (Barrett and Weber-Jahnke, 2014) performs with 96.7% accuracy on the MedPost corpus. These last two taggers have not been used in the present study because they are no longer available.

When using taggers that have been trained on biomedical article abstracts, it is important to know how well they scale up when they are used to tag full-text articles. Results suggest a 7-8 percentage point drop between testing the taggers on abstracts and testing them on full-text journal articles (Verspoor et al., 2012).

4. Data Set and Curated Corpus

The dataset used in this study is the complete article dataset that was first made available by The National Center for Biotechnology Information (PubMed Central) in 2009. It consists of the full text of articles published in 288 biomedical journals. Our goal is to build a corpus annotated for part-of-speech from the full set comprising approximately 35 million sentences. The current BioPOSTAg corpus[1] has been built from a set of 49 biomedical journals. The corpus comprises approximately 5 million sentences. These articles were POS tagged by Genia and PostMed.

The corpus is part-of-speech tagged using the biomedical update (Warner et al., 2012) of the Penn Treebank Tagset (Marcus et al., 1993). The updated tagset consists of the original 36 part-of-speech tags and 12 other tags for punctuation and currency symbols together with 4 additional tags added in the biomedical update. Tagging guidelines (Santorini, 1995; Warner et al., 2004; Warner et al., 2012) were consulted. The MedPost (and hence, PostMed) tagset used here is the original Penn Treebank Tagset. The Genia tagger uses the enlarged tagset.

[1] https://github.com/nelder/Biomedical-POS-Tagger/

5. Building the Corpus

The first step in the building of the corpus is to generate the tagging of Genia and PostMed. PostMed is used first to preprocess the .nxml files as described previously and to generate its tagged output. Genia then takes the tokenized output of PostMed and performs its tagging. These files can then be compared to discover the POS differences. We now direct our discussion in the next sections to how the POS differences are resolved with human intervention.

5.1. Part of Speech Tagging Difference

Because we were using only two POS taggers, our goal to produce a silver standard corpus could not use a scheme such as voting to decide a tagging outcome when the tags from the two taggers differed. So, we opted to have some human intervention to make decisions when this situation arose. Due to the volume of data and frequency of mismatch, it was not feasible to manually verify the tagged text produced by each of these taggers. As such we developed a software data viewer, using which, as humans, we could navigate and compare the outputs of these two taggers to identify where they disagreed. Implicit in this approach is the assumption that when Genia and PostMed specify the same tag for a particular word, then they are correct. While this might not be strictly true (see Section 6 for details), this assumption has seemed not to be deleterious. We harnessed the discord between these two taggers by assuming that one was correct and the other was incorrect. Our main focus thus, was the part of speech tagging difference (POSDiff). To illustrate, Genia and PostMed assigned the following tags:

Committee for Animal Research
NNP IN NNP NNP (Genia)
NN IN NN NN (PostMed)

POSDiff instances exist, one for each of the words: Committee, Animal, and Research. The POSDiff allowed us to group like errors in an attempt to provide human solutions to classes of problems as opposed to individual instances of tagging errors. Our method to find and correct errors is described later.

5.2. POSDiffs Discovered

With all of our tools in hand we began the process of building a better corpus by analyzing the POSDiffs between Genia and PostMed tagged data for our 49 journal corpus. We discovered that 5% of POSDiffs account for 81% of the disagreements. This means that a small handful of the POSDiffs disproportionally are responsible for the tagging errors which also means that solutions to these POSDiffs would be highly valuable for overall corpus quality. We also noted that across our 5 million sentence corpus there were a total of 496 POSDiffs. As Table 1 outlines, the top 25 most frequent POSDiffs accounted for 81.38% of the disagreements. The full list of POSDiffs is included in https://github.com/nelder/Biomedical-POS-Tagger/ as a csv file.

5.2.1. Decision Making

With all of the information now in hand we began to look through the POSDiffs from most common to least common and apply human judgement to correct each of the POSDiffs. For each of the POSDiffs we assessed a random sample of instances from the most frequent words within each POSDiff to develop an understanding of the cause. We took into consideration the pattern, whether it be each example looking consistent or more sporadic to decide when to direct more energy into looking at additional examples. For each of the 13 most frequent POSDiffs listed in Table 1 we created a decision procedure which selected between the taggers. The encoded procedure (which is machine interpretable) indicates whether either of the taggers is globally correct for a given POSDiff. If not, it will indicate the preferred tagger and a procedure of specific interventions to apply before using the default preferred tagger. These interventions pattern match either words or word patterns and apply an intervention. These interventions can be a specific POS tag, a tagger to use, or a context specific procedure. For instance, “positive = mix : PRIOR_WORD_TAG@JJ|NN? postmed,genia”. In this case the word “positive” is tagged using PostMed’s tag when the tag on the prior word is either a JJ or NN, otherwise it uses Genia’s tag. These decision procedures now exist for 70% of the POSDiff instances that occurred and as such we’ve eliminated many of the disagreements between the two taggers that were originally present with these procedures. The remaining 30% were eliminated by choosing the Genia tagger as providing the correct tag.

5.2.2. Sample Decision Procedures: Globally Correct Tagger

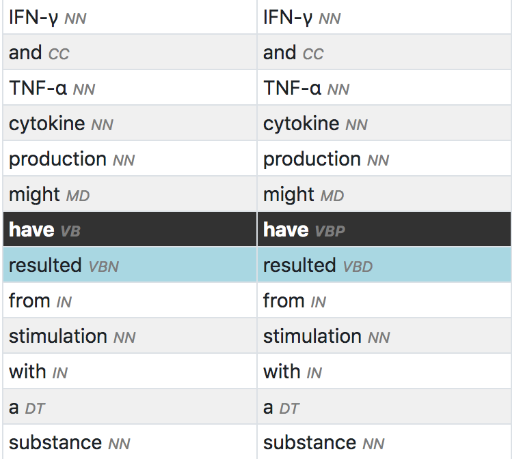
For Genia tagging VB (Verb, base form) and PostMed tagging VBP (Verb, non-3rd person singular present), we determined that Genia was tagging correctly in the vast majority of the sampled cases we examined. In all cases the syntactic structure involved a modal verb, then base case verb, followed up by the participle form of the verb. The issue was that PostMed was tensing the base form of the verb and then making a mistake on the main verb following this incorrectly tensed verb. An example is outlined in Figure 1, where this particular POSDiff is highlighted in black and other POSDiffs present in that selected sentence are highlighted in blue. Given the consistent cause we saw across the 10 sampled cases we assigned Genia to be the correct tagger globally for this POSDiff.

5.2.3. Sample Decision Procedures: Word Specific Solution

For Genia tagging NN (Noun, singular or mass) and PostMed tagging JJ (Adjective), we noted that neither tagger was exclusively correct. This tagging error was

Table 1: POSDiffs discovered (subset of most frequent 25).

POSDiff	Instances	Freq. (%)	Cumulative Freq. (%)	Unique Words	Instances per Word
G:NNP | P:NN	572,633	15.60%	15.60%	65607	9
G:JJ | P:NN	430,673	11.73%	27.33%	42387	10
G:VB | P:VBP	338,190	9.21%	36.54%	2312	146
G:VBN | P:JJ	270,197	7.36%	43.90%	4666	58
G:VBG | P:JJ	162,882	4.44%	48.34%	3727	44
G:NN | P:JJ	156,541	4.26%	52.60%	8360	19
G:NN | P:SYM	142,748	3.89%	56.49%	9	15861
G:VBG | P:NN	120,278	3.28%	59.77%	4290	28
G:VBN | P:VBD	91,012	2.48%	62.25%	1968	46
G:DT | P:PRP	87,779	2.39%	64.64%	22	3990
G:NNS | P:VBZ	63,313	1.72%	66.36%	2755	23
G:NNS | P:NN	54,028	1.47%	67.83%	4667	12
G:VBD | P:VBN	53,115	1.45%	69.28%	1762	30
G:RB | P:WRB	48,486	1.32%	70.60%	5	9697
G:NN | P:VBP	46,106	1.26%	71.86%	2090	22
G:FW | P:NN	46,018	1.25%	73.11%	1458	32
G:RBR | P:RB	44,569	1.21%	74.32%	22	2026
G:NNP | P:JJ	41,995	1.14%	75.46%	2586	16
G:CD | P:NN	36,809	1%	76.46%	7496	5
G:JJ | P:RB	35,246	0.96%	77.42%	2763	13
G:JJ | P:DT	32,357	0.88%	78.30%	11	2942
G:NNP | P:NNS	31,441	0.86%	79.16%	4178	8
G:VBD | P:JJ	28,888	0.79%	79.95%	1816	16
G:NN | P:NNS	27,213	0.74%	80.69%	2494	11
G:VBZ | P:NNS	25,422	0.69%	81.38%	2150	12



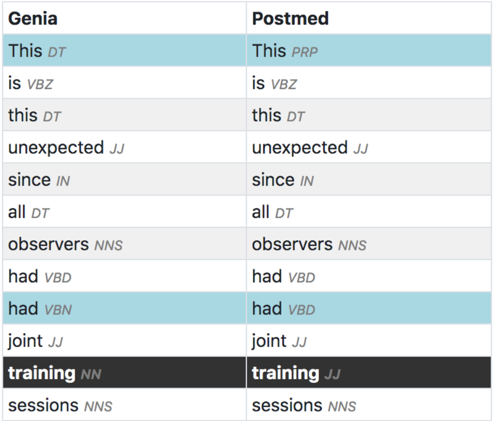
IFN-γ NN	IFN-γ NN
and CC	and CC
TNF-α NN	TNF-α NN
cytokine NN	cytokine NN
production NN	production NN
might MD	might MD
have VB	**have** VBP
resulted VBN	resulted VBD
from IN	from IN
stimulation NN	stimulation NN
with IN	with IN
a DT	a DT
substance NN	substance NN

Figure 1: Example of a POSDiff that can be corrected globally

related to noun compounds: the use of a noun as a noun premodifier in English. In this case the noun acts as an adjective though is in fact a noun. In this case we sampled the more frequent words and assigned correct taggers on a word by word basis. We also noted that words ending in "ing" were in some cases (Manning, 2011) to be tagged as VBG (Verb, gerund or present participle) and as such we overrode both taggers and used our own tag. In this process we worked with a random sample of 5 examples for 10 different words. We noted that Genia was correct more often than PostMed and as such assigned it as the tagger to side with for less frequent words we were not able to assign a solution to. An example of this tagging error is illustrated in Figure 2, where the POSDiff of interest is highlighted in black and other POSDiffs present in that selected sentence are highlighted in blue. In this case our decision procedure for the correct tag is based upon the word within the POSDiff.

5.2.4. Sample Decision Procedures: Context Specific Solution

For Genia tagging NNS (Noun, plural) and PostMed tagging NN (Noun, singular or mass), we noted that there were cases in which both taggers were correct. This POSDiff was caused by tags for irregular plural forms of nouns. We selected correct taggers for 12 of the most common words, set a tag override to NNP (Noun, proper) for one word, but had a more complex pattern necessary for the word *bacteria*. After examin-



Genia	Postmed
This DT	This PRP
is VBZ	is VBZ
this DT	this DT
unexpected JJ	unexpected JJ
since IN	since IN
all DT	all DT
observers NNS	observers NNS
had VBD	had VBD
had VBN	had VBD
joint JJ	joint JJ
training NN	training JJ
sessions NNS	sessions NNS

Figure 2: Example of a word specific POSDiff

ing 5 samples for the word *bacteria* we concluded that if the following word after *bacteria* was either a NNP or NN we would use the PostMed tag, and otherwise use the Genia tag. These contextually based decision procedures were used in a number of other instances to handle complex errors. Genia was selected as the default tagger for words which were not captured by our rules.

5.2.5. Decision Procedure Language

In order to encode the decision procedure model we were building for each POSDiff we developed a machine interpretable language which was quick for us to type. This language was later interpreted by software when it was understanding the decisions we had made for each POSDiff so that we could build the new corpus. An example of this language was previously seen in Section 5.2.1.

5.3. The BioPOSTAg Corpus

The current BioPOSTAg corpus consists of 119,348,590 words, 4,790,737 sentences, part-of-speech annotated with the biomedical update (Warner et al., 2012) of the Penn Treebank Tagset (Marcus et al., 1993). It is publicly available at https://github.com/nelder/Biomedical-POS-Tagger/.

5.4. The Data Viewer

5.4.1. Comparison of Taggers

To construct this set of POSDiffs we built software which processed the tagged output from Genia and PostMed. The corpora these taggers had annotated was full-text data from 49 biomedical journals, as mentioned previously. We then kept track of each instance of a POSDiff, the particular word on which it occurred, and an address to the original article which would allow us to view the context in which this POSDiff occurred. In the example shown in abbreviated form in Figure 3 we can see the case where Genia tagged AFX and PostMed labeled JJ. This POSDiff occurred 12 times, 8 times on the word “non”. We also can see the address of each instance of this difference in the form of a file path to the Genia and PostMed tagged journal papers including the line and word number (FILEPATH — line_number / word_number).

5.4.2. Complementary POSDiffs

Having collected this information we also considered the significance of the concept of a complementary part of speech difference (POSDiff-C). So far we have considered Genia saying tag A, and PostMed saying tag B to be entirely distinct from PostMed saying tag A and Genia saying tag B. While this was a valid assumption to make in pursuit of grouping likely similar errors together under each POSDiff (combining POSDiff & POSDiff-C likely would just create more complex decision criteria to pick the correct tagger later on) we may want to consider this data elsewhere in our assessment. As such we were interested in seeing the cardinality in terms of frequency of occurrence of each POSDiff versus its POSDiff-C. Within each POSDiff we also wanted to understand if particular words appeared in both POSDiff and POSDiff-C. If for example there was a case that for the word “web” Genia said common noun and PostMed said adjective as well as there existing cases where Genia said adjective and PostMed said common noun, then the decision criteria for selecting between taggers in these cases would need to be more nuanced. Otherwise if there were not many of these cases we could likely select with more basic criteria. The significance of this information was better understood as the decision making model is put together.

5.4.3. Context to POSDiff

We also constructed a tool to enable us to understand the window of context surrounding each POSDiff occurrence for a given POSDiff. By understanding the preceding and following words and POS tags around each instance, we were able to get a better understanding of the cause of each error. This information aided in our construction of a model for addressing the POSDiffs.

5.4.4. Data Explorer Tool

Having generated a large dataset of POSDiffs as well as a complementary data set around the number of occurrences of POSDiff-Cs we developed a viewing framework to enable easy traversal of this information. Using a HTML/CSS front-end, we were able to leverage libraries like JQuery and Bootstrap Data Tables to expedite our development process.

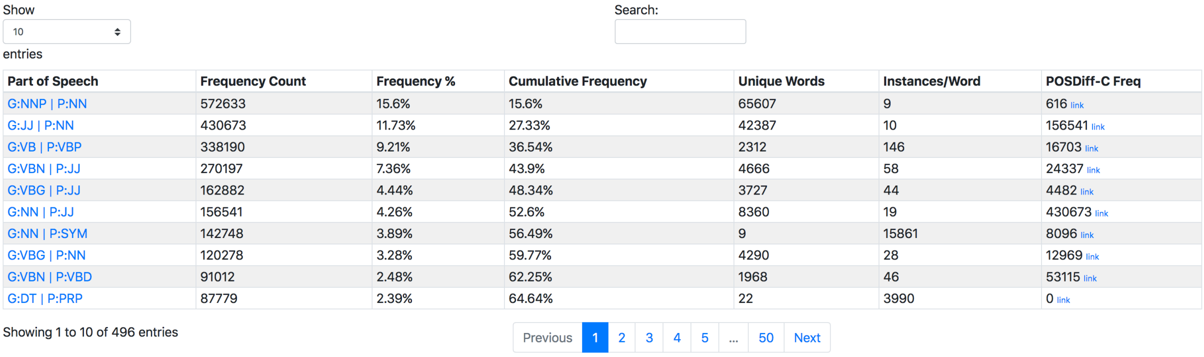
As illustrated in Figure 4 our table library made it easy to sort the information by any attribute and traverse our large data set. The first view enables a top level look at all POSDiffs. Clicking on any of the particular POSDiffs reveals information about the words on which a particular POSDiff occurred. This page also allows us to collect notes on which of the taggers was correct. This notes field will serve as the basis of our decision making model. Figure 5 reveals the frequency of each

```
"G:AFX|P:JJ": {
    "pos_frequency": 12,
    "words": [
        [
            "non",
            "Acta_Vet_Scand/
Acta_Vet_Scand-42-1-2202332.nxml.genia.tagged|31/29",
            "Acta_Vet_Scand/
Acta_Vet_Scand-42-1-2202332.nxml.postmed.tagged|31/29"
        ],
        [
            "non",
            "Acta_Vet_Scand/
Acta_Vet_Scand-43-2-1764189.nxml.genia.tagged|99/15",
            "Acta_Vet_Scand/
Acta_Vet_Scand-43-2-1764189.nxml.postmed.tagged|99/15"
        ],

.. abbreviated ..

    ],
    "words_freq": [
        [
            "non",
            8
        ],
        [
            "anti",
            4
        ]
    ],
    "words_freq_alpha": {
        "anti": 4,
        "non": 8
    }
},
```

Figure 3: View of the database containing the POSDiff information



Show

10

entries

Search:

Part of Speech	Frequency Count	Frequency %	Cumulative Frequency	Unique Words	Instances/Word	POSDiff-C Freq
G:NNP | P:NN	572633	15.6%	15.6%	65607	9	616 link
G:JJ | P:NN	430673	11.73%	27.33%	42387	10	156541 link
G:VB | P:VBP	338190	9.21%	36.54%	2312	146	16703 link
G:VBN | P:JJ	270197	7.36%	43.9%	4666	58	24337 link
G:VBG | P:JJ	162882	4.44%	48.34%	3727	44	4482 link
G:NN | P:JJ	156541	4.26%	52.6%	8360	19	430673 link
G:NN | P:SYM	142748	3.89%	56.49%	9	15861	8096 link
G:VBG | P:NN	120278	3.28%	59.77%	4290	28	12969 link
G:VBN | P:VBD	91012	2.48%	62.25%	1968	46	53115 link
G:DT | P:PRP	87779	2.39%	64.64%	22	3990	0 link

Showing 1 to 10 of 496 entries

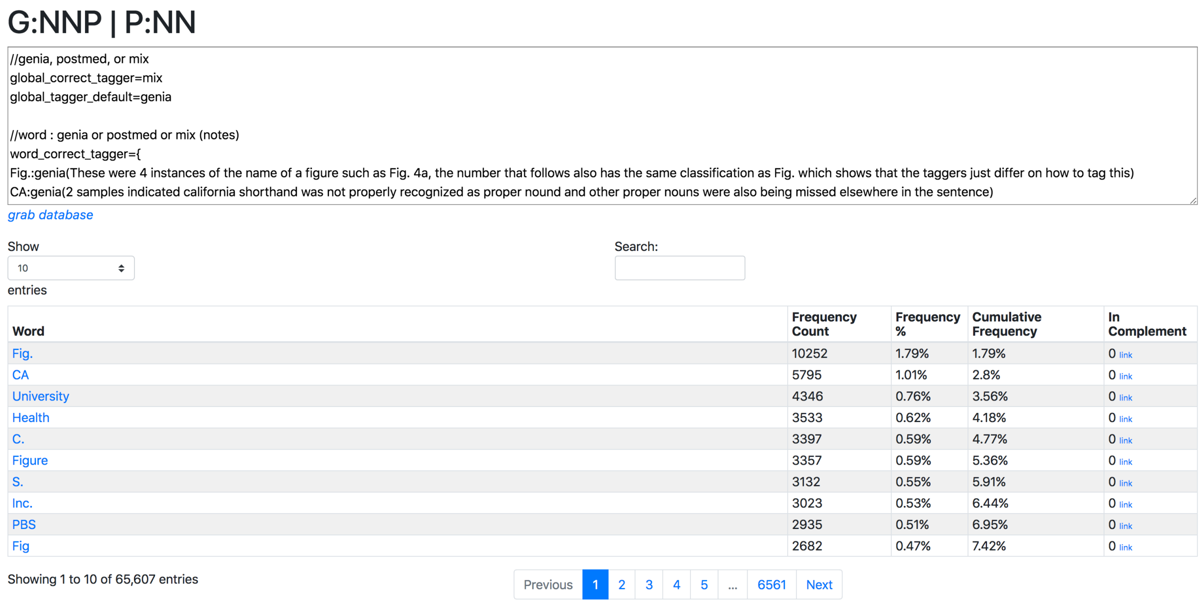
Figure 4: Summary of the POSDiffs provided by the data viewer

word within this POSDiff as well as information about the POSDiff-C.

An additional page for each word provides links to view the particular source for each instance of a POSDiff which is displayed on a page as illustrated in Figure 6. Note this particular POSDiff is highlighted in black and other POSDiffs present in that selected sentence are highlighted in blue. Each word is followed by the POS tag it received from each of the taggers. Source data can be viewed at the bottom of this page.

Other views of the database have been presented earlier in Figures 1 and 2.

The software tool organizes and displays the differences in the tagging provided in two files. The tool



G:NNP | P:NN

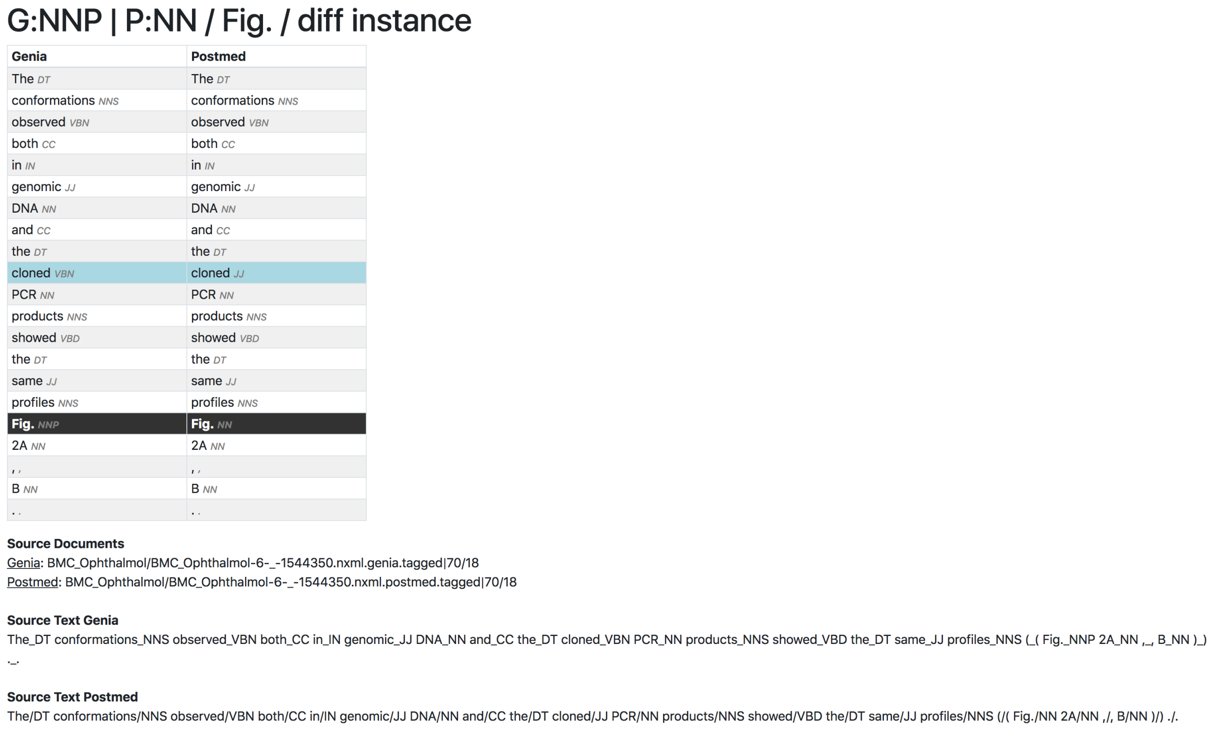
```
//genia, postmed, or mix
global_correct_tagger=mix
global_tagger_default=genia

//word : genia or postmed or mix (notes)
word_correct_tagger={
Fig.:genia(These were 4 instances of the name of a figure such as Fig. 4a, the number that follows also has the same classification as Fig. which shows that the taggers just differ on how to tag this)
CA:genia(2 samples indicated california shorthand was not properly recognized as proper nound and other proper nouns were also being missed elsewhere in the sentence)
```

Word	Frequency Count	Frequency %	Cumulative Frequency	In Complement
Fig.	10252	1.79%	1.79%	0 link
CA	5795	1.01%	2.8%	0 link
University	4346	0.76%	3.56%	0 link
Health	3533	0.62%	4.18%	0 link
C.	3397	0.59%	4.77%	0 link
Figure	3357	0.59%	5.36%	0 link
S.	3132	0.55%	5.91%	0 link
Inc.	3023	0.53%	6.44%	0 link
PBS	2935	0.51%	6.95%	0 link
Fig	2682	0.47%	7.42%	0 link

Showing 1 to 10 of 65,607 entries

Figure 5: Example of a POSDiff view provided by the data viewer



G:NNP | P:NN / Fig. / diff instance

Genia	Postmed
The DT	The DT
conformations NNS	conformations NNS
observed VBN	observed VBN
both CC	both CC
in IN	in IN
genomic JJ	genomic JJ
DNA NN	DNA NN
and CC	and CC
the DT	the DT
cloned VBN	cloned JJ
PCR NN	PCR NN
products NNS	products NNS
showed VBD	showed VBD
the DT	the DT
same JJ	same JJ
profiles NNS	profiles NNS
Fig. NNP	**Fig.** NN
2A NN	2A NN
, ,	, ,
B NN	B NN
. .	. .

Source Documents
Genia: BMC_Ophthalmol/BMC_Ophthalmol-6-_-1544350.nxml.genia.tagged|70/18
Postmed: BMC_Ophthalmol/BMC_Ophthalmol-6-_-1544350.nxml.postmed.tagged|70/18

Source Text Genia
The_DT conformations_NNS observed_VBN both_CC in_IN genomic_JJ DNA_NN and_CC the_DT cloned_VBN PCR_NN products_NNS showed_VBD the_DT same_JJ profiles_NNS (_(Fig._NNP 2A_NN ,_, B_NN)_) ._.

Source Text Postmed
The/DT conformations/NNS observed/VBN both/CC in/IN genomic/JJ DNA/NN and/CC the/DT cloned/JJ PCR/NN products/NNS showed/VBD the/DT same/JJ profiles/NNS (/(Fig./NN 2A/NN ,/, B/NN)/) ./.

Figure 6: Example of a POSDiff, the document that it occurs in, and the Part-of-Speech tagging by Genia and PostMed

is very versatile. It was initially designed to compare the output given by two part-of-speech taggers but it is easily convertible to comparing any two files, so it can be used for human analysis of the differences between a machine tagged output and gold standard tags.

6. Evaluating the Quality and Effectiveness of the Corpus

Much interest in having POS taggers for biomedical text (Kim et al., 2003; Smith et al., 2004; Nguyen and Verspoor, 2019) and to have full-text corpora (Verspoor et al., 2012) to train from is evident. An in-depth manual study of a representative portion of the full-text silver standard corpus that we have developed here to determine the quality of the corpus is our ultimate goal and is our intention in future work. In the meantime, we have provided two evaluations of the silver standard corpus. First, we evaluate on a small sample, the percentage of correct tags provided by the Genia and PostMed taggers. In addition, we are interested in our assumption that the two taggers provide the correct tag when they agree. We have chosen a representative portion of the CRAFT corpus (Verspoor et al., 2012) for this test. The second evaluation method is to compare a model trained on the silver standard corpus compared to the same model trained on a human-annotated gold standard corpus on the downstream task of interest, i.e., POS tagging. We have chosen the CRAFT corpus (Verspoor et al., 2012) as the gold standard training and test sets for this comparison. There is no overlap between the papers in the CRAFT corpus and the papers used to build the silver standard corpus.

For the first test, we have chosen one paper from the CRAFT corpus consisting of approximately 8,700 tokens. With this subset of tokens the Genia tagger correctly predicts 87% and PostMed predicts 84%. These scores are approximately 10 percentage points below their scores when tagging abstracts. Of course, the human intervention that we have described previously improves this performance. When these two taggers agree, they disagree with the CRAFT corpus tag on about 1% of the tags. While this seems high (and higher than we expected), approximately half of these disagreements are between the JJ and NN tags when the word is used as a modifier. However, as we discuss below, this mistagging (since the human intervention does not correct these mistags) does not seem to be deleterious.

Second, to evaluate the effectiveness of this silver corpus, we have conducted two experiments to provide the comparison. In the first experiment, we have done a 5-fold cross-validation by training a third party BioRoBERTa-based POS-tagger (Trevett, 2021) with the training data portion of the CRAFT dataset and tested it against the test set portion of the CRAFT dataset. This experiment achieves an average 97.89% test set accuracy with a standard deviation of 0.04. In the second experiment, the same model is trained with the silver standard corpus and tested against the same five test set portions of the CRAFT corpus that were set aside in the 5-fold cross-validation evaluation. It achieves an average accuracy of 98.09% with a standard deviation of 0.05. The silver standard trained model outperforms the gold standard trained model used in the first experiment by a noteworthy, for this level of accuracy, 0.2 percentage point improvement. This performance gain is statistically significant, $p < 0.0001$. This evaluation is summarized in Table 2.

Tagger trained on:	Accuracy (mean and s.d.)
the CRAFT dataset	97.89 ± 0.04
the BioPOSTAg dataset	98.09 ± 0.05

Table 2: Evaluation on the CRAFT dataset of a third party BioRoBERTa-based POS-tagger (Trevett, 2021) trained on the CRAFT dataset and the BioPOSTAg dataset, mean and standard deviation from 5-fold cross-validations, $p < 0.0001$

We provide the following information about the model and the training. The original BERT-based model (Trevett, 2021) consists of a BERT-based embedding layer followed by a linear layer to predict the POS tag of the input sentence. For the two biomedical text based experiments, we fine-tuned a BioRoBERTa embedding layer. The learning rate was initialized to 0.01 and it was decayed by 80% after any epoch if the validation accuracy decreased. The model was fine-tuned for 20 epochs in both experiments.

7. Conclusions

Our goal to provide a large silver standard corpus of biomedical text annotated with part-of-speech using the Penn Treebank Tagset to facilitate the training of deep learning models has been partially fulfilled. Our current corpus, drawn from full-text biomedical articles, has 4,790,737 sentences comprised of 119,348,590 tokens annotated for part-of-speech, and we continue to work toward a corpus containing 35 million sentences. The corpus is available online at elder.ca/research/biomed_pos_corpus.txt. In addition to this language resource, we have also designed, implemented, and made available a robust, easy-to-use software tool that assists the investigation of differences in two tagged datasets. It is available at https://github.com/nelder/Biomedical-POS-Tagger/.

8. Future Work

As stated earlier, the goal is to completely annotate 35 million sentences drawn from 288 biomedical journals with POS tags. These journals represent both experimental and clinical research. Having a corpus comprised of writing styles across a wide variety of journals will facilitate having a more robust deep learning trained POS tagger.

When correcting the POSDiffs, some decisions were made for purposes of expediency. A more careful analysis of the word specific and context specific solutions needs to be carried out. As part of its functionality, the data viewer captures both the language that describes how modifications to the corpus are to be carried out by the associated software and notes discussing the rationale for these modifications. With these sources of information, the corpus can be easily modified after careful consideration of the discussion.

To enhance our understanding of quality of the corpus beyond the small study reported above, an in-depth manual study of a representative portion of the full-text silver standard corpus to provide measures of the quality of the corpus will be done.

9. Bibliographical References

Barrett, N. and Weber-Jahnke, J. (2014). A token centric part-of-speech tagger for biomedical text. *Artificial Intelligence in Medicine*, 61(1):11–20.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231.

Coden, A. R., Pakhomov, S. V., Ando, R. K., Duffy, P. H., and Chuteb, C. G. (2005). Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, 38(6):422–430.

Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., and Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11:492.

Divita, G., Browne, A. C., and Loane, R. (2006). dTagger: A POS tagger. In *AMIA Annual Symposium Proceedings*, pages 200–203.

Eckart, K. and Gärtner, M. (2016). Creating silver standard annotations for a corpus of non-standard data. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 90–96.

Hahn, U. and Wermter, J. (2004). High-performance tagging on medical texts. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 973–979.

Huseth, O. and Rost, T. B. (2007). Developing an annotated corpus of patient histories from the primary care health record. In *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 165–173.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 171–189.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Miller, J. E., Torii, M., and K., V.-S. (2007). Adaptation of POS tagging for multiple biomedical domains. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, (BioNLP'07)*, pages 179–180.

Nguyen, D. Q. and Verspoor, K. (2019). From POS tagging to dependency parsing for biomedical event extraction. *BMC Bioinformatics*, 20:72.

Pakhomov, S. V., Coden, A., and Chute, C. G. (2006). Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, 75(6):418–429. Recent Advances in Natural Language Processing for Biomedical Applications Special Issue.

Rebholz-Schuhmann, D., Jimeno Yepes, A. J., van Mulligen, E. M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Tomanek, K., Beisswanger, E., and Hahn, U. (2010). The CALBC silver standard corpus for biomedical named entities — a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 568–573.

Rebholz-Schuhmann, D., Jimeno Yepes, A., Li, C., Kafkas, S., Lewin, I., Kang, N., Corbett, P., Milward, D., Buyko, E., Beisswanger, E., Hornbostel, K., Kouznetsov, A., Witte, R., Laurila, J. B., Baker, C. J. O., Kuo, C.-J., Clematide, S., Rinaldi, F., Farkas, R., Móra, G., Hara, K., Furlong, L. I., Rautschka, M., Neves, M. L., Pascual-Montano, A., Wei, Q., Collier, N., Chowdhury, M. F. M., Lavelli, A., Berlanga, R., Morante, R., Van Asch, V., Daelemans, W., Marina, J. L., van Mulligen, E., Kors, J., and Hahn, U. (2011). Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *Journal of Biomedical Semantics*, 2(S11).

Santorini, B. (1995). Part-of-speech tagging guidelines for the Penn Treebank Project (3rd Revision, 2nd printing). Technical report, University of Pennsylvania, February. Reprint of original June 1990 report updated and slightly reformatted by Robert MacIntyre.

Smith, L., Rindflesch, T., and Wilbur, W. J. (2004). MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*, 20(14):2320–2321.

Smith, L., Rindflesch, T., and Wilbur, W. (2006). The importance of the lexicon in tagging biological text. *Natural Language Engineering*, 12(4):335–351.

Sousa, D., Lamurias, A., and Couto, F. M. (2019). A silver standard corpus of human phenotype-gene relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1487–1492.

Tateisi, Y. and Tsujii, J. (2004). Part-of-speech annotation of biology research abstracts. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1267–1270.

Tateisi, Y., Tsuruoka, Y., and Tsujii, J. (2006). Subdomain adaptation of a POS tagger with a small corpus. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 136–137.

Trevett, B. (2021). Fine-tuning pretrained transformers for POS tagging. https://github.com/bentrevett/pytorch-pos-tagging/blob/master/2_transformer.ipynb. Accessed: 2021-12-30.

Verspoor, K., Cohen, K. B., Lanfranchi, A., Warner, C., Johnson, H. L., Christophe, R., Choi, J. D., Funk, C., Malenkiy, Y., Eckert, M., Xue, N., Baumgartner Jr, W. A., Bada, M., Palmer, M., and Hunter, L. E. (2012). A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13:207.

Warner, C., Bies, A., Brisson, C., and Mott, J. (2004). Addendum to the Penn Treebank II style bracketing guidelines: Biomedical treebank annotation. Technical report, University of Pennsylvania Linguistic Data Consortium, November.

Warner, C., Lanfranchi, A., O'Gorman, T., Howard, A., Gould, K., and Regan, M. (2012). Bracketing biomedical text: An addendum to Penn Treebank II guidelines. Technical report, Institute of Cognitive Science, University of Colorado at Boulder, January.