Adapting the TTL Romanian POS Tagger to the Biomedical Domain

Maria Mitrofan
Research Institute for AI
"Mihai Drăgănescu"
Romanian Academy
13 "Calea 13 Septembrie",
Bucharest 050711, Romania
maria@racai.ro

Radu Ion
Research Institute for AI
"Mihai Drăgănescu"
Romanian Academy
13 "Calea 13 Septembrie",
Bucharest 050711, Romania
radu@racai.ro

Abstract

This paper presents the adaptation of the Hidden Markov Models-based TTL part-of-speech tagger to the biomedical domain. TTL is a text processing platform that performs sentence splitting, tokenization, POS tagging, chunking and Named Entity Recognition (NER) for a number of languages, including Romanian. The POS tagging accuracy obtained by the TTL POS tagger exceeds 97% when TTL's baseline model is updated with training information from a Romanian biomedical corpus. This corpus is developed in the context of the CoRoLa (a reference corpus for the contemporary Romanian language) project. Informative description and statistics of the Romanian biomedical corpus are also provided.

1 Introduction

Natural Language Processing (NLP) is one of the key technologies that can be employed to extract valuable information from unstructured text (e.g. discharge summaries, clinical notes, medical reference books, research papers, medical blog posts) and transform it into a desired form to support activities related to the healthcare domain.

NLP technologies have been adapted to the biomedical domain and applied on a vast amount of clinical data to enhance the research process and to extract relevant information from textual data. For example, clinical notes have been used for identifying cardiovascular risk factors (Abdulrahman and Meystre, 2015), electronic medical records have been used for detecting diabetes mellitus (Chung-Il et al., 2017). Jackson et al. (2017) applied NLP to extract symptoms of severe mental illness from clinical text. NLP tools have been proven to be an efficient way to enhance the identification on Alzheimer's disease (Shibata et al., 2016) and even the Human Genome project used NLP techniques in order to explore the relationships between biomedical literature and genes sequences (Yandell and W. H. Majoros, 2002).

A typical NLP pipeline consists in sentence delimitation, tokenization, part-of-speech (POS) tagging, lemmatization and parsing. More advanced NLP pipelines will perform NER and/or word sense disambiguation.

POS tagging (the process of labeling a token with a part of speech tag) is one of the initial pipelined components and it is an important step that performs morphosyntactic disambiguation. Therefore, the quality of the POS tagging is very important because cascading errors generated in POS tagging processes affect the overall performance of NLP pipelines. Consequently, it is very important that a POS tagger performs as optimally as possible.

The accuracy of a POS tagger is expected to be high (e.g. at least 97% for English) when the tagged text is similar, domain-wise, to the tagger's training data, but when the tagger is used on texts belonging to significantly different domains than the ones the tagger was trained on (e.g. train on newspaper articles and test on biomedical documents), its performance can degrade significantly. Ferraro et al. (2013) showed that the accuracy of state-of-the-art English POS taggers trained on news texts plummeted from 97% to 85% when POS tagging has been applied to clinical narratives, mainly because biomedical texts have different linguistic characteristics. Therefore the target domain adaptation of the POS tagger is needed.

Ferraro et al. (2013) note that there are multiple POS tagger domain adaptation techniques, out of which the simplest one is what they call "source-target labeled data aggregation" which refers to the

Proceedings of the Biomedical NLP Workshop associated with RANLP 2017, pages 8–14,
Varna, Bulgaria, 8 September 2017.
https://doi.org/10.26615/978-954-452-044-1_002

training of a POS tagging model based on a labeled corpus obtained from both the source and the target domains. A simplified version of this approach is the approach we follow here in order to adapt the baseline model of the trigram HMM-based TTL POS tagger (Ion, 2007) to Romanian biomedical POS tagging.

To obtain good POS tagging results with the source-target labeled data aggregation domain adaptation method, high-quality training data in both the source and the target domains is vital for a good performance of the POS tagger. Consequently most of the work concentrates on building high-quality training corpora, which are typically hand-made and slow to produce and which are, for this reason, very hard to find. In general, the lack of sufficient data in biomedical domain remains a barrier for biomedical NLP, especially for under-resources languages. Even though at the international level biomedical resources have been developed (e.g. HIMERA - a collection of historical medical documents manually annotated at semantic level with information relevant for public health, BMC - a corpus which contains full medical articles provided by BioMed Central, GENIA - a collection of 2000 biomedical abstracts annotated at syntactic and semantic level), at a national level (at least in Romania) it is very difficult to obtain texts for specialized corpora in the biomedical domain due to copyright laws and lack of biomedical literature published in Romanian language that is readily available in electronic format.

Efforts to improve the availability of Romainian biomedical training data for POS tagging are currently carried on. The most important is the CoRoLa project which was started in 2012 by the Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăgănescu" (RACAI) and the Institute for Computer Science in Iași. It aims to create a reference corpus of the contemporary Romanian language (CoRoLa) (Mititelu et al., 2014), which will be useful for different types of NLP tasks, including POS tagging.

In what follows, we will briefly review related word in POS tagging domain adaptation for the biomedical domain (Section 2), we will introduce the Romanian biomedical corpus that we used to adapt TTL to the biomedical domain (Section 3), we will briefly describe TTL (Section 4) and we will present our initial experiment in Romanian biomedical POS tagging (Section 5). The paper ends with our concluding remarks (Section 7).

2 Related Work

Domain adaptation received significant attention from the NLP research community and multiple approaches have been developed to improve the tagging accuracy and to reduce the errors caused by out-of-vocabulary words. A very common approach used for domain adaptation is to combine both the source and the target training data to train a new model. This method was used by Coden et al. (2005) when an HMM POS tagger was trained on both news and a medical corpus of clinical notes. After this experiment they reported an accuracy of almost 93% when the tagger was tested on the medical test set, compared to a little over 87% when the tagger was trained on the news corpus and tested on the medical test set.

For the GENIA POS tagger, Tsuruoka et al. (2005) presented several experimental results for domain adaptation on GENIA, PennBioIE and Wall Street Journal (WSJ) corpora. POS tagging performances has been evaluated for seven different combinations of the corpora as the training data. When the tagger was trained on WSJ corpus (without the distinction between nouns and proper nouns) and tested also on a test set extracted from WSJ corpus (in-domain testing), the accuracy was 97.20%, but when the tagger was applied on test sets extracted from biomedical corpora (out-of-domain testing), the accuracy dropped significantly: 91.55% on GENIA and 90.51% on PennBioIE. On the other hand, when the GENIA tagger was trained both on WSJ and GENIA corpora, it achieved an accuracy of 98.32% on the GENIA test set and an accuracy of 96.96% on the WSJ test set (and a lower accuracy on PennBioIE test set, 91.98%). This shows that domain adaptation is worth doing even though in-domain accuracy may drop a little.

cTAKES tagger is an example of a biomedical tagger that demonstrates the variability of the biomedical domain. This tagger was trained with Mayo Clinic's notes and tested on a set of clinical notes from Kaiser Permanente Southern California (KPSC) on which it obtained an accuracy of 88.1%. Moreover, the cTAKES tagger tested on set of clinical notes from the University of Pittsburg Medical Center (UPMC) achieved an accuracy of 88.3%. This is to show that POS tagging in the biomedical domain is more difficult than, e.g.

news POS tagging, mainly because of the extensive lexicon of the domain.

Finally, we present two experiments demonstrating that good accuracy can be obtained with in-domain biomedical data even with small training sets. Smith et al. (2004) trained the MedPost tagger on 5,716 manually tagged sentences taken from Medline abstracts within the Genomics domain and achieved an accuracy of 97.43% on 1000 sentence test set extracted also from MEDLINE abstracts. In order to train the Brill tagger on biomedical domain, Campbell and Johnson (2001) tagged by hand 100,000 words from a corpus of discharge summaries, 90% of the hand tagged corpus was used to train the tagger an the remaining 10% was used to test the tagger. This process was repeated ten times and achieved an accuracy of 96.9%, each time using a different 10% as the test set.

3 Corpus Structure

In order to perform domain adaptation we have developed a domain-specific training corpus, because sub-domain languages present distinct linguistic features, usually not found in general language, in this case Romanian language.

The process of collecting the texts was not an easy task, firstly because of the intellectual property restrictions and secondly because in general, biomedical literature is published in English and not in the Romanian language. At the end of this process the Romanian medical corpus contained texts from different sources such as medical books published at the Romanian Academy Publishing House and Polirom publishing house, free medical online resources, medical blogs, online courses made for medical students.

The biomedical corpus has evolved from a collection of texts extracted from different biomedical sub-domains such as: cardiology, endocrinology, diabetes, oncology, surgery, genetics, nephrology, neurology, psychiatry etc. The textual resources available in the corpus were initially available in different formats such as .doc and unprotected .pdf and they had to be converted into a raw text format in order to be annotated by our processing tools (Tufiş et al., 2008). The conversion of the files involved a boilerplate removal step in which footers, headers, page numbers, figures, tables, footnotes, etc. have been removed. For this step we used the tool designed by (Moruz and Scutelnicu, 2014). In order to improve the linguistic annotation we considered only texts with correct diacritical characters, encoded in UTF-8.

The Romanian biomedical corpus used for domain adaptation of the TTL POS tagger contains about 206,020 sentences and 4,390,707 million tokens (words and punctuation) distributed in more than nine medical sub-domains (see above) extracted from academic books and journals and one which contains information from different free medical online resources such as medical blogs and Romanian medical publications.

The resources extracted from online sources have not been grouped into medical categories because most of them belong to more than one medical category and medical expertise was needed in order to fulfill this task. Furthermore the POS tagging step is not affected by this lack of classification. All the texts were split into tokens, POS tagged and lemmatized with the baseline model of TTL (see Section 5).

Table 1 shows some statistics of the automatically POS tagged biomedical corpus: we counted all tokens (words plus punctuation), words (functional words and content words), unique lemmas and sentences. Content words also included abbreviations because these represent an important feature of the biomedical texts. The punctuation count is obtained by subtracting the words count from the tokens count (Table 2). From a statistical point of view, the corpus is balanced in terms of tokens per sentence, content words per sentence and punctuation per sentence (Table 2 and Table 3) when comparing sub-domains.

Table 3 shows that the texts obtained from online resources contain the highest use of content words per sentence; at the other end the texts from endocrinology domain use the lowest number of content words. An interesting fact is that the average number of punctuation per sentence contained in the texts extracted from online sources remains in compliance with the average number of punctuation used in academic medical literature.

In Table 4 the distribution of content words is presented among the POS tags types. While online resources texts make use of more nouns and less adjectives, the other medical sub-domains use less nouns and more adjectives. A characteristic specific to the biomedical domain, which it is also shown in table 4 is represented by the high use of the total nouns and adjectives.

# tokens, punctuation included	4,390,707
# words	3,750,242
# unique lemmas	101,348
# sentences	206,020
average tokens per sentence	21.31
average words per sentence	18.20
average punctuation per sentence	3.10

Table 1: Statistics over the Romanian medical corpus.

4 Tokenizing, Tagging and Lemmatizing (TTL) Platform

TTL is a Perl module supporting Romanian, English, French and Bulgarian, with the following functionalities: sentence splitting, tokenization, POS tagging, lemmatization, chunking and Named Entity Recognition (NER).

TTL's tokenizer takes two input parameters (the code of the language and the sentence) and returns a list of tokens. Moreover the tokenization procedure is language independent and identifies clitics, contractions and multiword expressions (MWEs), provided that language-dependent resources exist (i.e. list of MWEs and affix words that should be split).

The POS tagger is a heavily-improved reimplementation of the Hidden Markov Models (HMM) tagger presented in Brants (2000). It uses the tiered tagging technology (Tufiș, 1999; Ceaușu, 2006) for a more accurate POS labeling with a large tagset: the MSD tagset [1]. The Romanian MSD tagset has 736 labels and the general purpose Romanian language POS tagging accuracy is over 98% with this tagset (Tufiș, 1999).

Lemmatization is achieved after the POS tagging process is complete. TTL lemmatizer uses a large human-validated Romanian inflected lexicon, currently holding 1,152,506 entries. For the out-of-dictionary words, the TTL lemmatizer selects the most probable lemma provided by a five-gram letter Markov Model-based guesser (see Ion (2007) for details).

Chunking is another functionality of the TTL platform and it is based on a set of regular expressions applied on sequences of POS tags. The TTL chunker recognizes nominal, verbal, adjectival, adverbial and prepositional phrases.

[1] http://nl.ijs.si/ME/V4/msd/html/

5 Adapting TTL to the Biomedical Domain

As already stated in the Introduction, we attempted to adapt the baseline model of the Romanian TTL POS tagger to the biomedical domain by following the "source-target labeled data aggregation" paradigm. In our case, we have updated the baseline model's parameters by training on a sample of the Romanian biomedical corpus, for reasons to be explained below.

It is a well-known fact that the performance of a POS tagger depends crucially on the quality of the labeled corpus on which it trains. Thus, the baseline model for Romanian POS tagging that TTL uses is based on training on news (some "Adevărul" and "România Liberă" issues, 98,194 tokens) and fiction (Orwell's "1984", 118,357 tokens) corpora whose POS labeling was carefully checked by trained linguists, word by word (Tufiș, 2000).

Our initial experiment in biomedical POS tagging domain adaptation focused on experimentally verifying the assumption that we can get good results with an in-domain corpus whose POS labeling *is semi-automatically corrected*. That is, what results do we get if the biomedical corpus that is used to adapt TTL to the domain is not checked word for word but is corrected using some semi-automatic procedures (to be described below) whose output is checked by the trained linguist.

Since we could not hope to manually check 4.4M tokens as our Romanian biomedical corpus has (nor did we want to commit to such a task), we performed a random sampling of that corpus in order to obtain reasonable-sized train and test corpora. We concluded that, with our resources, we could check around 600K tokens, which, according to the English domain adaptation literature cited above, is a reasonable size. Thus, after splitting our sample into train and test sets, the train set contained 545,977 tokens (words and punctuation) and the test set contained 60,520 tokens, which is about 10% of the part we selected. The selection was done randomly, but enforcing the following conditions:

- We have sentences of all lengths from the Romanian biomedical corpus (short, average and long);
- All sentences have Romanian diacritics in

	Sentences	Tokens	Content words	Punctuation
Online resources	52,708	1,146,052	772,564	151,189
Cardiology	35,505	754,394	418,619	110,850
Surgery	51,367	989,335	550,037	156,140
Diabetes	33,538	775,017	411,393	114,123
Oncology	22,746	523,568	281,331	78,693
Endocrinology	10,156	202,341	112,826	29,470
Total	206,020	4,390,707	2,546,770	640,465

Table 2: Statistics on medical domains

	Sentences	Tokens	Content	Punctuation
Online resources	52,708	21.74	14.65	2.86
Cardiology	35,505	21.24	11.79	3.21
Surgery	51,367	19.26	10.65	3.03
Diabetes	33,538	23.10	12.26	4.44
Oncology	22,746	23.01	12.36	3.45
Endocrinology	10,156	19.92	11.10	2.90
Total	206,020	21.31	12.34	3.10

Table 3: The average number of tokens, content words and punctuation per sentence by biomedical subdomain

place and are written using the Romanian Academy Romanian writing reform (i.e. using 'â' instead of 'î' inside words);

- There are no duplicate sentences.

Both the train and the test sets were automatically POS tagged with the TTL's baseline model. The test set was manually checked, word by word, by a trained linguist. The manual correction procedure involved reading each sentence from the test set, word by word, and making sure that the POS labellings are correct (the test set had to be thoroughly checked because the POS tagger performance was going to be measured against it).

For the train set, to speed up the correction process, we adopted the following semi-automatic approach:

- We extracted the list of unknown words with all their inflected forms (7,816 unique word forms) and checked their POS labellings, adding alternate analyses where it was necessary (e.g. adding a noun analysis for an existing adjective analysis);

- Noticing that TTL does not (usually) assign the wrong POS to a word (e.g. if a word is a noun, TTL will recognize it as such but, for unknown words, it may give the wrong gender or case), we automatically replaced the POS labels of all unknown words in the train set with the corresponding POS labels from the curated unknown list. We were thus able to automatically fix 26,184 occurrences of unknown words in the train set;

- We built a TTL POS tagging model only from the train set and re-tagged the train set with it (we call this a 'biased evaluation'). We then inspected manually all the differences in POS labeling between the original tagging and the biased tagging. Some more (about 2% of the train set) inconsistencies were fixed this way;

- We also corrected every error that we saw in the train set, *but without going through it, word by word*.

Tables 5 and 6 present the TTL POS tagger accuracy on the biomedical test set. From Table 5 we see that general POS tagging accuracy degrades a little and this can be explained by the fact that the biomedical train set is not yet fully correct when it comes to POS labeling.

The baseline TTL model is trained over texts that were corrected *at word-level* by trained linguists while our biomedical train set was mostly automatically corrected with only a small part being manually validated. That the biomedical train

	Nouns	Verbs	Adjectives	Adverbs	Abbreviations
Online resources	477,208	137,729	120,382	24,717	12,528
Cardiology	224,758	70,684	102,039	14,717	6,421
Surgery	284,549	100,567	138,777	19,695	6,449
Diabetes	222,905	82,638	81,327	17,559	6,964
Oncology	153,955	53,357	58,350	9,528	6,141
Endocrinology	59,776	21,074	25,262	4,601	2,113
Total	1,423,151	466,049	526,137	90,817	40,616

Table 4: POS statistics for content words in each biomedical sub-domain.

	Errors	Accuracy
Baseline model	1,068	98.23%
Biomedical model	1,310	97.83%

Table 5: Overall TTL accuracy on the test set

	Errors	Percent
Baseline model	486	45.50%
Biomedical model	448	34.19%

Table 6: Errors on biomedical terminology

set still contains general language POS annotation errors becomes evident when the most frequent errors (on the test set) are identified (which *are not produced* by the baseline model):

- Verb 'a fi' (English 'to be') can occur as an auxiliary ('a fi' plus past participle) or main (61 errors);
- Verb 'a avea' (English 'to have') can also occur as an auxiliary (when forming the present perfect tense) or main (15 errors).

Table 6 shows the benefit of doing domain adaptation, *even with a minimally corrected in-domain corpus*: the percentage of errors relating to biomedical terminology (i.e. nouns, main verbs, adjectives and adverbs that are specific to the domain) is smaller when we use the adapted POS tagging model. At this point, if the degradation in general-purpose POS tagging is acceptable (0.4% in our case) the much lower error rate (11.31% in our case) in biomedical terminology POS tagging could be of help in applications such as biomedical NER.

6 The Availability of the Data

After the train set and the test set will be checked in detail ("word by word") both of them will be freely available for download[2] and non-commercial use. Special use-cases require license permissions from the author.

The biomedical corpus will be available in the context of the CoRoLa project copyright agreement signed with the publishing houses and with the editorial offices representatives. The whole corpus will be available to the public through KorAP platform (Banskiand et al., 2013), but will not be downloadable. The KorAP platform allows multiple linguistic types of searches in the corpus. However, all the results of the interrogation of the corpus outside the scope of the copyright restrictions will be downloadable.

7 Conclusions and Future Work

This paper presents a newly created text corpus aimed at providing support for NLP on biomedical text and an initial experiment about the adaptation of the TTL POS tagger to the biomedical domain. Currently our text corpus is still under development, but the available data and the biomedical TTL POS tagger can already be considered important resources in order to perform more advanced NLP tasks in the Romanian biomedical domain. To the best of our knowledge, the Romanian biomedical corpus is the first of its kind.

Our initial experiment was promising in the sense that, with minimal POS labeling correction efforts, we were able to improve the accuracy of the tagger where it matters most for other biomedical applications using POS tagging: the biomedical terminology. Thus, the error rate of biomedical terminology was reduced by 11.31%. We plan to fully validate the biomedical train set, with the help of trained linguists, and repeat the experiments to ensure that we obtain comparable (with the baseline) general language POS tagging accuracy (over 98% accuracy) while lowering even

[2] http://slp.racai.ro/index.php/resources/

more the error rate on biomedical terminology.

Compared to other corpora used for domain adaptation, our biomedical train set is larger (545,977 tokens) than most of the POS train sets. Another important characteristic of the biomedical corpus used for the adaptation of the TTL POS tagger is the variability of its lexicon: it contains words from five major biomedical sub-domains and a collection of texts extracted from online sources. Thus, we think that any POS tagger trained on it will perform better on a wider range of Romanian biomedical texts.

The train and test sets will also be annotated with biomedical named entities and parsed with our Romanian Universal Dependencies parser developed in the SSPR project (Mititelu et al., 2016). Thus, we will have a Romanian biomedical corpus that can be used as training data for other useful NLP tasks such as biomedical terminology identification, biomedical NER, biomedical text mining, etc.

References

Khalifa Abdulrahman and Stéphane Meystre. 2015. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes.

P. Banskiand, J. Bingel, N.Diewald, E. Frick, M. Hanl, M. Kupietz, P. Pezik, C. Schnober, and A. Witt. 2013. The new corpus analysis platform at ids mannheim.

T. Brants. 2000. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*.

D. Campbell and S. Johnson. 2001. Comparing syntactic complexity in medical and non-medical corpora.

Alexandru Ceauşu. 2006. Maximum entropy tiered tagging. In *Proceedings of the 11th ESSLLI student session.*.

Wi Chung-Il, E.and Voge G. Sohn S.and Rolfes M. C.and Seabright A.and Ryu, and H Liu. 2017. Application of a natural language processing algorithm to asthma ascertainment: An automated chart review.

A. R. Coden, S. V. Pakhomov, R. K. Ando, P. H. Duffy, and C. G. Chute. 2005. Domain-specific language models and lexicons for tagging.

J.P. Ferraro, H. Daumé III, S. L. DuVall, W. W. Chapman, H. Harkema, and P. J. Haug. 2013. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation.

Radu Ion. 2007. *Word Sense Disambiguation Methods Applied to English and Romanian (in Romanian)*. Ph.D. thesis, Romanian Academy.

R. G. Jackson, R. Patel, N. Jayatilleke, A. Kolliakou, M. Ball, G. Gorrell, and R. Stewart. 2017. Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (cris-code) project.

V. Barbu Mititelu, E. Irimia, and D. Tufiş. 2014. Corola – the reference corpus of contemporary romanian language. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation - LREC*. pages 1235–1239.

Verginica Barbu Mititelu, Radu Ion, Radu Simionescu, Andrei Scutelnicu, and Elena Irimia. 2016. Improving parsing using morpho-syntactic and semantic information, in revista romana de interactiune om-calculator.

Alex Moruz and Andrei Scutelnicu. 2014. An automatic system for improving boilerplate removal for romanian texts. In *Proceedings of the 10th International Conference "Linguistic resources and Tools for Processing the Romanian Language*.

D. Shibata, S. Wakamiya, E. Aramaki, and A. Kinoshita. 2016. Detecting japanese patients with alzheimer's disease based on word category frequencies.

L. Smith, T. Rindflesch, and W. J. Wilbur. 2004. Medpost: a part of speech tagger for biomedical text.

Y. Tsuruoka, Y.Tateishi, J. D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. I. Tsujii. 2005. *Developing a robust part-of-speech tagger for biomedical text.*.

D. Tufiş, R. Ion, A. Ceauşu, and D. Ştefănescu. 2008. In proceedings of the 6th language resources and evaluation conference-lrec.

Dan Tufiş. 2000. Using a large set of eagles-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In *Proceedings of LREC*.

Dan Tufiș. 1999. *Tiered tagging and combined language models classifiers.*. Springer.

M. D. Yandell and W. H W. H. Majoros. 2002. Genomics and natural language processing.