Lessons learned from tagging clinical Hungarian

György Orosz[1,2], Attila Novák[1,2], and Gábor Prószéky[1,2]

[1] Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
50/a Práter street, 1083 Budapest, Hungary
[2] MTA-PPKE Hungarian Language Technology Research Group
50/a Práter street, 1083 Budapest, Hungary
http://nlpg.itk.ppke.hu
{oroszgy,novak.attila,proszeky}@itk.ppke.hu

Abstract. As more and more textual resources from the medical domain are getting accessible, automatic analysis of clinical notes becomes possible. Since part-of-speech tagging is a fundamental part of any text processing chain, tagging tasks must be performed with high accuracy. While there are numerous studies on tagging medical English, we are not aware of any previous research examining the same field for Hungarian. This paper presents methods and resources which can be used for annotating medical Hungarian and investigates their application to tagging clinical records. Our research relies on a baseline setting, whose performance was improved incrementally by eliminating its most common errors. The extension of the lexicon used raised the overall accuracy significantly, while other domain adaptation methods were only partially successful. The presented enhancements corrected almost half of the errors. However, further analysis of errors suggest that abbreviations should be handled at a higher level of processing.

Keywords: medical text processing, PoS tagging, morphological disambiguation, domain adaptation, clinical notes

1 Introduction

Hospitals produce a huge amount of clinical notes that have solely been used for archiving purposes and have generally been inaccessible to researchers. However, nowadays medical resources are becoming available, enabling computer scientist to support medical researchers. As natural language processing (NLP) algorithms are getting more and more accurate, their usage can not just cut costs but can also boost medical research. PoS tagging is a fundamental task of computational linguistics: labeling words with their part-of-speech is essential for further processing algorithms. While tagging of general texts is well-known and considered to be solved, most of the commonly used methods usually fail on medical texts.

English has been the main target of many NLP applications up to the present time, thus less-resourced languages, which are usually morphologically complex often fell beyond the scope. Similarly, there are just a few studies attempting to

annotate non-English medical texts. Thus, the processing of Hungarian clinical records has a very little literature. Moreover, there is not any research on tagging such texts. Therefore, this study aims to investigate how existing techniques can be used for the morphological tagging of Hungarian clinical records presenting possible pitfalls of a medical parsing chain.

This paper is structured as follows. The background of our research is described in the next section. Then a corpus is presented which has been created for development and evaluation purposes. In Section 4, we detail the baseline morphological disambiguation setting used, which is commonly employed for Hungarian. Afterwards, we present the most frequent errors made by the baseline tagger and we describe and evaluate the enhancements that were carried out on the text processing chain. Finally, Section 6 provides the final conclusions.

2 Parsing of biomedical texts

2.1 Biomedical tagging

Parsing of biomedical texts has an extensive literature, since there are numerous resources accessible. In contrast, much less manually annotated corpora of clinical texts are available. Most of the work in this field has been done for English and only a few attempts have been published for morphologically rich languages (e. g. [19, 26]).

A general approach for biomedical PoS tagging is to employ supervised learning algorithms, which require manually annotated data. In the case of tagging biomedical texts, domain-specific corpora are used either alone [23, 28, 32] or in conjunction with a (sub)corpus of general English [6, 13, 18] as training data. While using texts only from the target domain yields acceptable performance [23, 28, 32], several experiments have shown that accuracy further increases with incorporating annotated sentences from the general domain as well [1, 6]. A general observation is that the more data is used from the reference domain, the higher accuracy can be achieved (e. g. [24]). On the contrary, Hahn and Wermter argue for training learners only on general corpora [15] (for German). Further on, there are studies on selecting training data (e. g. [17]) that increase the accuracy. What is more, there are taggers (such as [5]) which learn from several domains in a parallel fashion, thus the model selection decision is delayed for the decoding process.

Using target specific lexicons is another way of adapting taggers, as they can improve tagging performance [6, 27]. Some of these studies extend existing PoS dictionaries [9], while others build new ones specific to the target domain [32]. All of the experiments using such resources yield significantly reduced error rates.

Concerning tagging algorithms, researchers tend to prefer already existing applications, such as the OpenNLP toolkit[3], which is the basis of the cTakes system [28]; while Brill's method [3] and TnT [2] are widely used (e.g. [15, 28,

[3] `http://opennlp.apache.org/`

24]) as well. There are other HMM-based solutions which have been shown to perform well [1, 6, 9, 15, 23, 26, 27] on such texts. Besides, a number of experiments have revealed [13, 27, 32] that domain-specific OOV words are primarily responsible for a reduced tagging performance. Thus successful methods employ either guessing algorithms [1, 9, 26, 27, 32] or broad coverage lexicons (as detailed above). Beyond supervised algorithms, other approaches were also shown to be effective: Miller et al. [18] use semi-supervised methods; Dwinedi and Sukhadeve build a tagger system based only on rules [10]; while Ruch et al. propose a hybrid system [27]. Further on, domain adaptation methods (such as EasyAdapt [8] or ClinAdapt [13]) also perform well. However, they need an appropriate amount of manually annotated data from the target domain, which limits their applicability.

2.2 Tagging general Hungarian

For Hungarian, tagging experiments generally rely on the Szeged Corpus [7] (SZC), since this is the only contemporary linguistic resource that is manually annotated with morphological tags. It contains about 1.2 million words from six different genres, but does not involve texts from the biomedical domain. The original annotation of the corpus uses the MSD scheme proposed by the MULTEXT-East project [11]. Besides this, other morphosyntactic coding systems are commonly used as well. One of them is the system employed by the HuMor morphological analyzer [25], whose labels are composed of morpheme tags. Another annotation scheme is named KR, which is the default of morphdb.hu [35], a freely available Hungarian morphological resource. Although the Szeged Corpus contains only MSD codes, there are also automatically converted variants of it that use the latter schemes.

For agglutinating languages such as Hungarian, labeling a word only with its part-of-speech tag is not satisfactory (as described in [21]), since further parsing methods require full morphosyntactic labels and lemmata as well. Consequently, tagger tools must perform full morphological disambiguation which also involves lemmatization. For Hungarian, such tools are the following.

PurePos [21] an open-source full morphological disambiguator system which is able to incorporate the knowledge of a morphological analyzer (MA), thus providing state-of-the-art accuracy above 98%. The system is based on statistical trigram-tagging algorithms, but it is extended to employ language specific rule-based components effectively.

`magyarlanc` **[36]** a freely available[4] language processing chain for morphological and dependency parsing of Hungarian that contains several language specific components. Its morphological disambiguator module is based on the Stanford tagger [34] and incorporates a MA based on morphdb.hu.

[4] It is available only without the source code.

2.3 Processing clinical Hungarian

There are only a few studies on processing Hungarian medical records. Siklósi et al. [31, 30] presented a system that is able to correct spelling errors in clinical notes. A resolution method for clinical abbreviations was also presented by them [29], in which they used pattern matching methods on domain-specific lexicons. Recently, Orosz et al. introduced [22] a partly unsupervised algorithm for segmenting tokens and sentences in clinical texts: their approach is a combination of collocation extraction algorithms and rule-based methods.

As far as we know, no study exists either investigated possible approaches or established a proper method for tagging clinical Hungarian. Therefore we aim to examine special properties of clinical notes first, then to develop a disambiguation methodology. The experiments described below use methods that rely on an error analysis of the baseline system (in Section 4), while also incorporate ideas from previous studies (cf. Section 2.1).

3 The clinical corpus

First of all, special properties of clinical texts need to be considered. Such records are created in a special environment, thus they differ from general Hungarian in several respects. These attributes are the following (cf. [22, 29, 31]): *a*) notes contain a lot of erroneously spelled words, *b*) sentences generally lack punctuation marks and sentence initial capitalization, *c*) measurements are frequent and have plenty of different (erroneous) forms, *d*) a lot of (non-standard) abbreviations occur in such texts, *e*) and numerous medical terms are used that originate from Latin.

Since there was no corpus of clinical records available that was manually annotated with morphological analyses, a new one was created for testing purposes. This corpus contains about 600 sentences, which were extracted from the notes of 24 different clinics. First, the textual parts of the records were identified (as described in [31]), then the paragraphs to be processed were selected randomly. Then manual sentence boundary segmentation, tokenization and normalization was performed, which were aided by methods detailed in [22]. Manual spelling correction was carried out by using suggestions provided by the system of Siklósi et al. [30]. Finally, morphological disambiguation was performed: the initial annotation was provided by PurePos, then its output was checked manually.

Several properties of the corpus created differ from general ones. Beside characteristics described above, the corpus contains numerous *x* tokens which denote multiplication and are labeled as numerals. Latin words and abbreviations are analyzed regarding their meaning: e. g. *o.* denotes *szem* 'eye', thus it is tagged with N.NOM. Further on, names of medicines are labeled as singular nouns. Finally, as missing sentence final punctuation marks were not recovered in the test corpus, these are not tagged either.

The corpus was split into a development and a test set (see Table 1). The first part was employed for development purposes, while the methods detailed

Table 1. Size of the clinical corpus created

	Sentences	Tokens
Development set	240	2230
Test set	333	3155

below were evaluated against the second part. Evaluation was carried out by calculating per-word accuracy omitting punctuation marks.

4 The baseline setting and the analysis of its errors

Hereunder we introduce the baseline tagging chain. First we describe its components, then the performance of the tagger is evaluated by detailing the most common error types. Concerning the parts of the chain we follow the work of Orosz et al. [21]. Thus *(morphosyntactic tag, lemma)* pairs represent the analyses of HuMor, which are then disambiguated by PurePos. However, the output of the MA is extended with the new analyses of x in order to fit the corpus to be tagged.

This baseline text processing chain produced 86.61% token accuracy on the development set, which is remarkably lower than tagging results for general Hungarian using the same components (96–98% [20]). Measuring the ratio of the correctly tagged sentences revealed that less than the third (28.33%) of the words were tagged correctly. This amount indicates that the models used by the baseline algorithm are weak for such a task. Therefore, errors made by the baseline algorithm are investigated first to reveal how the performance could be improved.

Table 2. Distribution of errors caused by the baseline algorithm – dev. set

Class	Frequency
Abbreviations and acronyms	49.17%
Out-of-vocabulary words	27.27%
Domain specific PoS of wordforms	14.88%
Other	0.06%

Table 2 shows that the top error class is the mistagged abbreviations and acronyms. A reason for the high number of such errors is that most of these tokens are unknown to the tagger. Moreover, abbreviations usually refer to medical terms that originate from Latin.

Another frequent error type is caused by the out-of-vocabulary (OOV) words. This observation is in accordance with the PoS tagging results for medical English (as described above). Similarly, in the case of Hungarian, most of the OOV tokens are specific to the clinical domain and often originate from Latin. However, several inflected forms of such terms also exist in clinical notes due to

agglutination. Therefore, listing only medical terms and their analyses could not be a proper solution. This problem demands for complex algorithms.

Furthermore, the domain-specific usage of general words leads the tagger astray as well. Frequently, participles are labeled as verbs such as *javasolt* 'suggested' or *felírt* 'written'. In addition, numerous mistakes are due to the lexical ambiguity that is present in Hungarian (such as *szembe* which can refer to 'into an eye' or 'in the face of').

Our investigation shows that most of the baseline system's errors are made up of three categories. We can use the categorization above to enhance its performance by eliminating the typical sources of errors.

5 Incremental improvements

Based on the observations above, systematic changes were carried out to improve the tagging accuracy of the chain. First, the processes of lexicon extension and algorithmic modifications are described, then an investigation is presented aiming to find the optimal training data. Each enhancement is evaluated against the test corpus. Table 3 contains the part-of-speech tagging, lemmatization and the whole morphological tagging performance of each system.

Table 3. Evaluation of the enhancements – test set

ID	Method	PoS tagging	Lemmatization	Morph. disambig.
0	Baseline system	90.57%	93.54%	88.09%
1	0 + Lexicon extension	93.89%	96.24%	92.41%
2	1 + Handling abbreviations	**94.81%**	**97.60%**	**93.73%**
3	2 + Training data selection	94.25%	97.36%	93.29%

5.1 Extending the lexicon of the morphological analyzer

Supervised tagging algorithms commonly use augmented lexicons in order to reduce the number of out-of-vocabulary words (see Section 2.1). In the case of Hungarian, this must be performed at the level of the MA. Here we describe the process which was carried out to extend the lexicon of the HuMor analyzer.

The primary source for the extension process was a spelling dictionary of medical terms [12] that contained about 90000 entries. Beside this, a freely available list of medicines [14] of about 38000 items was used as well. Since neither of these resources contained any morphological information concerning these words, such analyses were created. For this, we followed an iterative process which included both human work and automatic algorithms. The steps of our workflow were the following: 1) a set of wordforms was prepared and analyzed automatically (detailed below); 2) the analyses were checked and corrected manually; 3) the training sets of the supervised learning methods were extended

with the results of step 2). Before each iteration, compounds of known items were selected to be processed first. This enhancement reduced the time spent on manual correction and granted the consistency of the database created. In the end, approximately 41000 new entries were added to the lexicon of the HuMor analyzer.

Since latinate words can either be written as pronounced in Hungarian[5] or can appear with Latin spelling, having both variants is necessary. Most of the entries in the dictionary had both the Hungarian and Latin spelling variants, but this was not always the case. Language identification of the words was carried out to distinguish Hungarian terms from the ones that have Greek, Latin, English or French spelling. For this, TextCat [4] was involved in the iterative process to decide whether a word is Hungarian or not. If it was necessary, missing Hungarian spelling variants were produced and were added semi-automatically to the lexicon.

As for the calculation of the morphological analyses, the guesser algorithm of PurePos was employed. Separate modules were employed for each language, thus language-specific training sets were maintained for them as well. In Hungarian, the inflection paradigm depends on vowel harmony and the ending of the word as it is pronounced, thus the pronunciation of foreign words had to be calculated first. This could be carried out using simple hand-written rules most of which implement Latin grapheme-to-phoneme correspondences.

The lexicon extension process above reduced the OOV word ratio from 34.57% to 26.19% (development set), and resulted in an accuracy of 92.41% (test set). Since the medical dictionary [12] contained abbreviated words as well, this process could also decrease the number of mistagged abbreviations.

5.2 Dealing with acronyms and abbreviations

Despite the changes in Section 5.1, numerous errors made by the enhanced tagger were still connected to abbreviations. Thus we first examined erroneous tags of abbreviated terms, then developed methods aiming to improve the performance of the disambiguation chain.

A detailed error analysis revealed that some of the erroneous tags of abbreviated terms were due to the over-generating nature of HuMor, which could be reduced by a filtering method. For words with full stops an analysis was considered to be false if its lemma was not an abbreviation. This modification increased the overall accuracy significantly, reducing the number of errors by 9.20% on the development set (cf. “Filtering” in Table 5).

Another typical error type was the erroneous tagging of unknown acronyms. Since PurePos did not employ features that could deal with such cases, these tokens were left to the guesser. However, acronyms should have been tagged as singular nouns. Thus a pattern matching component relying on surface features could fix their tagging (see “Acronyms” in Table 5).

[5] An example is the Latin word *dysplasia* [displa:zia] can be spelled as *diszplázia* in Hungarian.

The rest of the errors were mainly connected to those abbreviations that were both unknown to the analyzer and had not been seen previously. For this, the distribution of the labels of abbreviations in the development data is compared to that of the Szeged Corpus (see Table 4 below). While there are several common properties between the two columns (such as the ratio of adverbs), discrepancies occur even more often. One of them is the ratio of adjectives, which is significantly higher in the medical domain than in general Hungarian. Comparing the values, it must be noted that 10.85% of the tokens are abbreviated in the development set, while the same ratio is only 0.37% in the Szeged Corpus.

Table 4. Morphosyntactic tag frequencies of abbreviations – dev. set

Tag	Clinical texts	Szeged Corpus
N.NOM	67.37%	78.18%
A.NOM	19.07%	3.96%
CONJ	1.27%	0.50%
ADV	10.17%	11.86%
Other	2.12%	5.50%

Since the noun tag was the most frequent amongst abbreviations, a plausible method was to assign N.NOM to all of these tokens (cf. "UnkN" in Table 5) and to keep the original wordforms as lemmata. This baseline method resulted in a surprisingly high error rate reduction of 31.54%.

Another approach was to model the analyses of abbreviations with data observed in Table 4. The first experiment ("UnkUni") employed the uniform distribution of such labels present in the development set. Thus all the tags (A.NOM, A.PRO, ADV, CONJ, N.NOM, V.3SG, V.PST_PTCL) were used with equal probability as a sort of guessing algorithm.

Beside this, a better method was to use maximum likelihood estimation for calculating a priori probabilities ("UnkMLE"). In this case, relative frequency estimates were calculated for all the above tags used. While the latter approaches could increase the overall performance, none of them managed to reach the accuracy of the "UnkN" method (cf. Table 5).

Table 5. Comparison of the approaches aiming to handle acronyms and abbreviations – dev. set

ID	Method	Morph. disambig.
0	Medical lexicon	90.11%
1	0 + Filtering	91.02%
2	1 + Acronyms	91.41%
3	2 + UnkN	**94.12%**
4	2 + UnkUni	92.82%
5	2 + UnkMLE	94.01%

5.3 Choosing the proper training data

Since many studies showed (cf. Section 2.1) that the training data used significantly affects the result of the annotation chain, we investigated the usage of sub-corpora available in the Szeged Corpus. Several properties of the corpus were examined (cf. Table 6) in order to find the training dataset that fits best for tagging clinical Hungarian. Measurements regarding the development set were calculated manually where it was necessary.

Table 6. Properties of training corpora

Corpus	Avg. sent. length	Abbrev. ratio	Unknown ratio	Perplexity	
				Words	Tags
Szeged Corpus	16.82	0.37%	**1.78%**	2318.02	22.56
Fiction	12.30	0.10%	2.44%	995.57	32.57
Compositions	13.22	0.14%	2.29%	1335.90	30.78
Computer	20.75	0.14%	2.34%	854.11	22.89
Newspaper	21.05	0.20%	2.10%	1284.89	**22.08**
Law	23.64	1.43%	2.74%	**824.42**	29.79
Short business news	23.28	0.91%	2.50%	859.33	27.88
Development set	9.29	10.85%	–	–	–

First of all, an important attribute of a corpus is the length of its sentences. Texts having shorter sentences tend to have simpler grammatical structure, while longer sentences are grammatically more complex. Further on, clinical texts have a vast amount of abbreviations, thus the ratio of abbreviations is also relevant during the comparison.

Furthermore, the accuracy of a tagging system is strongly related to the ratio of unknown words, thus these proportions were calculated for the development set using the vocabulary of each training corpus (see Table 6). This ratio could function as a similarity metric, but entropy based measures work better [16] in such scenarios. We use perplexity, which is calculated here as follows: trigram models of word and tag sequences are trained on each corpus using Kneser-Ney smoothing, then all of them are evaluated against the development set[6].

Measurements show that there is no such part of the Szeged Corpus which has as much abbreviated terms as clinical texts have. Likewise, sentences written by clinicians are significantly shorter than the ones in general Hungarian. Neither the calculations above, nor the ratio of unknown words suggest that we should use sub-corpora for training. However, the perplexity scores contradict: sentences from the law domain have the most phrases in common with clinical notes, while news texts have the most similar grammatical structures.

Therefore, all sub-corpora were involved in the evaluation, which was carried out by employing all of the enhancements described in previous sections. Results

[6] The SRILM toolkit [33] was employed for the calculations.

showed that training on news texts resulted in the highest accuracy. However, it was not able to outperform the usage of the whole corpus.

Table 7. Evaluation of the tagger using the subcorpora as training data – test set

Corpus	Morph. disambiguation accuracy
Szeged Corpus	**93.73%**
Fiction	92.01%
Compositions	91.97%
Computer	92.73%
Newspaper	**93.29%**
Law	92.17%
Short business news	92.69%

6 Conclusion

In this study, resources and methodologies were introduced which enabled us to investigate morphological tagging of clinical Hungarian. First, a test corpus was created and was compared in detail with a general Hungarian corpus. This corpus also allowed for the evaluation of numerous tagging approaches. These experiments were based on the PurePos tagger tool and the HuMor morphological analyzer. Errors made by the baseline morphological disambiguation chain were investigated, then several enhancements were carried out aiming at correcting the most common mistakes of the baseline algorithm. Amongst others, we extended the lexicon of the morphological analyzer and introduced several methods to handle the errors caused by abbreviations.

The baseline setup labeled every eighth token erroneously. Although this tagging chain is commonly used for parsing general Hungarian, it resulted in mistagged medical sentences in two thirds of the cases. In contrast, our enhancements raised the ceiling of the tagging accuracy to 93.79% by eliminating almost half (47.36%) of the mistakes. Deeper investigation revealed that this error reduction rate was mainly due to the usage of the extended lexicon, which significantly decreased the number of the out-of-vocabulary tokens. While this research did not manage to find decent training data for tagging clinical Hungarian, it showed that neither part of the Szeged Corpus was able to outperform the whole as a training corpus. Finally, results of tagging abbreviations suggest that abbreviated terms should not be tagged directly. They should be resolved first or should be labeled with a uniform tag.

The main limitation of this research is the corpus used. It contains a few hundred sentences, which is only enough to reveal the main pitfalls of the tagging method. Furthermore, most of the domain adaptation methods rely on target-specific corpora that have several thousands of sentences. Taking these into consideration, further investigation should involve more manually annotated data from the medical domain.

In sum, commonly used methodologies alone fail to tag Hungarian clinical texts with a satisfactory accuracy. One of the main problems is that such algorithms are not able to deal with the tagging of abbreviations. However, our results suggests that the usage of an extended lexicon considerably increases the accuracy of an HMM tagger.

Acknowledgement

We would like to thank Nóra Wenszky for their comments on preliminary versions of this paper. This work was partially supported by TÁMOP – 4.2.1.B – 11/2/KMR-2011-0002 and TÁMOP – 4.2.2/B – 10/1–2010–0014.

References

1. Neil Barrett and Jens Weber-Jahnke. A Token Centric Part-of-Speech Tagger for Biomedical Text. In Mor Peleg, Nada Lavrač, and Carlo Combi, editors, *Artificial Intelligence in Medicine*, volume 6747 of *Lecture Notes in Computer Science*, pages 317–326. Springer Berlin Heidelberg, 2011.
2. Thorsten Brants. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 224–231. Universität des Saarlandes, Computational Linguistics, Association for Computational Linguistics, 2000.
3. Eric Brill. A simple rule-based part of speech tagger. *Proceedings of the Third conference on Applied Natural Language Processing*, 28(4):152–155, 1992.
4. William B. Cavnar and John M. Trenkle. N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US, 1994.
5. Jinho D. Choi and Martha Palmer. Fast and Robust Part-of-Speech Tagging Using Dynamic Model Selection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 363–367. Association for Computational Linguistics, The Association for Computer Linguistics, 2012.
6. Anni Coden, Sergey V. Pakhomov, Rie Kubota Ando, Patrick H. Duffy, and Christopher G. Chute. Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, 38(6):422–430, 2005.
7. Dóra Csendes, János Csirik, and Tibor Gyimóthy. *The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus*, volume 3206 of *Lecture Notes in Computer Science*, pages 19–23. Springer, 2004.
8. Hal Daumé III. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
9. Guy Divita, Allen C. Browne, and Russell Loane. dTagger: a POS tagger. In *AMIA Annual Symposium Proceedings*, pages 200–203. American Medical Informatics Association, 2006.
10. Sanjay Kumar Dwivedi and Pramod P Sukhadeve. Rule-based Part-of-speech Tagger for Homoeopathy Clinical Realm. *IJCSI International Journal of Computer Science*, 8(4):350–354, July 2011.

11. Tomaž Erjavec. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1):131–142, March 2012.
12. Pál Fábián and Péter Magasi. *Orvosi helyesírási szótár*. Akadémiai Kiadó, Budapest, 1992.
13. Jeffrey P. Ferraro, Hal III Daumé, Scott L. DuVall, Wendy W. Chapman, Henk Harkema, and Peter J Haug. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association*, pages 931–939, 2013.
14. Országos Gyógyszerészeti Intézet Főigazgatóság. Forgalomba hozatali engedéllyel rendelkező allopátiás és homeopátiás készítmények. `http://www.ogyi.hu/generalt_listak/tk_lista.csv`. Online; accessed 20-December-2012.
15. Udo Hahn and Joachim Wermter. Tagging Medical Documents with High Accuracy. In Chengqi Zhang, Hans W. Guesgen, and Wai-Kiang Yeap, editors, *PRICAI 2004: Trends in Artificial Intelligence*, volume 3157 of *Lecture Notes in Computer Science*, pages 852–861. Springer, 2004.
16. Adam Kilgarriff and Tony Rose. Measures for corpus similarity and homogeneity. In Nancy Ide and Atro Voutilainen, editors, *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, pages 46–52. Association for Computational Linguistics, 1998.
17. Kaihong Liu, Wendy Chapman, Rebecca Hwa, and Rebecca S Crowley. Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. *Journal of the American Medical Informatics Association*, 14(5):641–650, 2007.
18. John Miller, Manabu Torii, and K Vijay-Shanker. Building Domain-Specific Taggers without Annotated (Domain) Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1103–1111, 2007.
19. Michel Oleynik, Percy Nohama, Pindaro Secco Cancian, and Stefan Schulz. Performance analysis of a POS tagger applied to discharge summaries in Portuguese. *Studies in health technology and informatics*, 160(2):959–963, 2009.
20. György Orosz and Attila Novák. PurePos – an open source morphological disambiguator. In Bernadette Sharp and Michael Zock, editors, *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*, pages 53–63, Wroclaw, 2012.
21. György Orosz and Attila Novák. PurePos 2.0: a hybrid tool for morphological disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, pages 539–545, Hissar, Bulgaria, 2013.
22. György Orosz, Attila Novák, and Gábor Prószéky. Hybrid text segmentation for Hungarian clinical records. In *Lecture Notes in Artificial Intelligence. Advances in Artificial Intelligence and Its Applications*, pages 306–317. Springer, Berlin Heidelberg, 2013.
23. Serguei V. S. Pakhomov, Anni Coden, and Christopher G. Chute. Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, 75(6):418–429, 2006.
24. John Pestian, Lukasz Itert, and Wlodzislaw Duch. Development of a Pediatric Text-Corpus for Part-of-Speech Tagging. In Mieczyslaw A. Klopotek, Slawomir T. Wierzchon, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 219–226. Springer, 2004.

25. Gábor Prószéky. Industrial Applications of Unification Morphology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, ANLC '94, pages 213–214, Morristown, NJ, USA, October 1994. Association for Computational Linguistics.
26. Thomas Brox Røst, Ola Huseth, Øystein Nytrø, and Anders Grimsmo. Lessons from Developing an Annotated Corpus of Patient Histories. *Journal of Computing Science and Engineering*, 2(2):162–179, 2008.
27. Patrick Ruch, Robert Baud, Pierrette Bouillon, and Gilbert Robert. Minimal commitment and full lexical disambiguation: Balancing rules and hidden markov models. In *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning*, pages 111–114. Association for Computational Linguistics, 2000.
28. Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin Kipper Schuler, and Christopher G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.
29. Borbála Siklósi and Attila Novák. *Detection and Expansion of Abbreviations in Hungarian Clinical Notes*, volume 8265 of *Lecture Notes in Artificial Intelligence*, pages 318–328. Springer-Verlag, Heidelberg, 2013.
30. Borbála Siklósi, Attila Novák, and Gábor Prószéky. Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. In Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, and Bianca Truthe, editors, *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 248–259. Springer Berlin Heidelberg, 2013.
31. Borbála Siklósi, György Orosz, Attila Novák, and Gábor Prószéky. Automatic structuring and correction suggestion system for Hungarian clinical records. In Guy De Pauw, Gilles-Maurice de Schryver, Mike L. Forcada, Francis M. Tyers, and Peter Waiganjo Wagacha, editors, *8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*, pages 29–34, Istanbul, 2012.
32. Lawrence H. Smith, Thomas C. Rindflesch, and W. John Wilbur. The importance of the lexicon in tagging biological text. *Natural Language Engineering*, 12(4):335–351, 2006.
33. Andreas Stolcke. SRILM – an extensible language modeling toolkit. In John H. L. Hansen and Bryan L. Pellom, editors, *Proceedings International Conference on Spoken Language Processing*, pages 257–286. ISCA, November 2002.
34. Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In Marti Hearst and Mari Ostendorf, editors, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Edmonton, Canada, 2003. Association for Computational Linguistics.
35. Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of the Fifth conference on International Language Resources and Evaluation*, pages 1670–1673, Genoa, 2006.
36. János Zsibrita, Veronika Vincze, and Richárd Farkas. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of Recent Advances in Natural Language Provessing 2013*, pages 763–771, Hissar, Bulgaria, 2013. Association for Computational Linguistics.