Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition

Jana Straková and **Milan Straka** and **Jan Hajič**
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{strakova,straka,hajic}@ufal.mff.cuni.cz

Abstract

We present two recently released open-source taggers: NameTag is a free software for named entity recognition (NER) which achieves state-of-the-art performance on Czech; MorphoDiTa (Morphological Dictionary and Tagger) performs morphological analysis (with lemmatization), morphological generation, tagging and tokenization with state-of-the-art results for Czech and a throughput around 10-200K words per second. The taggers can be trained for any language for which annotated data exist, but they are specifically designed to be efficient for inflective languages, Both tools are free software under LGPL license and are distributed along with trained linguistic models which are free for non-commercial use under the CC BY-NC-SA license. The releases include standalone tools, C++ libraries with Java, Python and Perl bindings and web services.

1 Introduction

Morphological analysis, part-of-speech tagging and named entity recognition are one of the most important components of computational linguistic applications. They usually represent initial steps of language processing. It is no wonder then that they have received a great deal of attention in the computational linguistics community and in some respect, these tasks can even be considered very close to being "solved".

However, despite the fact that there is a considerable number of POS taggers available for English and other languages with a large number of active users, we lacked a POS tagger and NE recognizer which would

- be well suited and trainable for languages with very rich morphology and thus a large tagset of possibly several thousand plausible combinations of morphologically related attribute values,
- provide excellent, preferably state-of-the-art results for Czech,
- be distributed along with trained linguistic models for Czech,
- allow the user to train custom models for any language,
- be extremely efficient in terms of RAM and disc usage to be used commercially,
- offer a full end-to-end solution for users with little computational linguistics background,
- be distributed as a library without additional dependencies,
- offer API in many programming languages,
- be open-source, free software.

Following these requirements, we have developed a morphological dictionary and tagger software, which is described and evaluated in Section 3; and a named entity recognizer, which is described and evaluated in Section 4. The software performance and resource usage are described in Section 5 and the release and licensing condition information is given in Section 6. We conclude the paper in Section 7.

2 Related Work

2.1 POS Tagging

In English, the task of POS tagging has been in the center of computational linguists' attention for decades (Kucera and Francis, 1967), with renewed interest after significant improvements achieved by (Collins, 2002). The recent state-of-the-art for English POS supervised tagging without external data for training is by (Shen et al., 2007) and there are many available taggers, such as well-known Brill tagger (Brill, 1992), TnT tagger (Brants, 2000) and many others.

Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 13–18,
Baltimore, Maryland USA, June 23-24, 2014. ©2014 Association for Computational Linguistics

In Czech, the POS tagging research has been carried out mostly by Czech speaking linguistic community and the current state-of-the-art was reported by (Spoustová et al., 2009) in Morče research project[1]. Based on this project, two taggers were released: Morče tagger (released as part of COMPOST[2] containing morphological analyzer, tagger and trained models, available to registered users only) and Featurama[3] (source code only, no trained models publicly available).

2.2 Named Entity Recognition

For English, many NE datasets and shared tasks exist, e.g. CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), MUC7 (Chinchor, 1998). These shared tasks and the associated freely available NE annotated corpora allowed wide and successful research in NE recognition in English. For example, the systems which published high scores on the CoNLL-2003 task include (Suzuki and Isozaki, 2008), (Ando and Zhang, 2005) and to our knowledge, the best currently known results on this dataset were published by (Ratinov and Roth, 2009). One should also mention a well-known and widely used Stanford parser (Finkel et al., 2005).

In Czech, the referential corpus for NE recognition is called the Czech Named Entity Corpus[4] (Ševčíková et al., 2007) and we describe its' properties further in Section 4.2. The development of the Czech NE recognition research is easy to follow: started by a pilot project by (Ševčíková et al., 2007), the results were improved by (Kravalová and Žabokrtský, 2009), (Konkol and Konopík, 2011) and (Konkol and Konopík, 2013). The current state-of-the-art results for CNEC are reported by (Straková et al., 2013). So far, there was no freely available Czech NE recognizer.

3 MorphoDiTa: Morphological Dictionary and Tagger

3.1 Morphological Dictionary Methodology

The morphological dictionary is specially designed for inflective languages with large number of suffixes (endings) and we propose an effective method for handling rich morphology.

In inflective languages,[5] words take endings (suffixes) to mark linguistic cases, grammatical number, gender etc. Therefore, many forms may be related to one lemma. For example, the lemma "zelený" ("green" in Czech) can appear as "zelený", "zelenější", "zelenému" etc. – there are several tens of forms for this type of adjective. Corpus-wise, there are 168K unique forms and 72K lemmas in a corpus of 2M words (Prague Dependency Treebank 2.5 (Bejček et al., 2012)) in Czech. It is therefore crucial to handle the endings effectively and to reduce the processing costs where regularities are found.

Given a resource with forms, lemmas and tags,[6] MorphoDiTa estimates regular patterns based on common form endings and automatically clusters them into morphological "templates" without linguistic knowledge about the language. We now describe the method for template set creation.

During template set creation, MorphoDiTa takes lemmas one by one. For each lemma, it collects all corresponding forms and builds a trie (De La Briandais, 1959; Knuth, 1997). Trie is a tree structure in which one character corresponds to a node and all descendants of a node share the same prefix. The procedure then finds a suitable common ancestor in the trie (common prefix or stem). The heuristics is "such a node whose subtree has depth at most N and at the same time has the maximal number of ancestors with one child". Intuitively, this means we want to select a long prefix (stem) – hence "maximal number of ancestors" but at the same time, the linguistic endings are not too long (at most N). Having selected a common prefix, all the endings (including their corresponding tags) in its subtree define a template. A rich trie with many subtrees may be split into multiple templates. For example, a simple trie for noun "hrad" ("castle" in Czech) with one template, and also two lemmas sharing two templates are shown in Fig. 1. When processing the next lemma and its corresponding forms, either new template is created, or the templates are reused if the set of endings is the same. Larger N leads to longer endings and larger number of classes, and smaller N leads to short endings and less classes.[7]

[1] http://ufal.mff.cuni.cz/morce/index.php

[2] http://ufal.mff.cuni.cz/compost/

[3] http://sourceforge.net/projects/featurama/

[4] http://ufal.mff.cuni.cz/cnec/

[5] In the following, we describe features of a core group of inflective languages, such as Slavic languages of all types. Sometimes, the word "inflective" is used also for agglutinative languages such as Turkish, Hungarian or Finnish; we believe our tools are suitable for these, too, but we have not tested them on this group yet.

[6] In Czech, the resource used was Morfflex CZ by Jan Hajič: http://ufal.mff.cuni.cz/morfflex.

[7] Our morphological dictionary representation cannot be replaced with a minimized finite state automaton with marked

The number of templates determines the efficiency of dictionary encoding. When too few templates are used, many are needed to represent a lemma. When too many are used, the representation of the templates themselves is large.

The morphological dictionary is then saved in binary form and the software offers a higher level access: given a form, morphological analysis lists all possible lemma-tag pairs; given a lemma-tag pair, MorphoDiTa generates the respective form. The analysis function is then used in tagging, which we describe in the next section.

The heuristics described above does not require linguistic knowledge about the language and handles linguistic regularities very well. The major advantage is a significant data compression leading to efficient resource usage: in our setting, the original morphology dictionary, the Czech Morfflex, contains 120M form-tag pairs derived from 1M unique lemmas, using 3 922 different tags, of total size 6.7GB.[8] Using the proposed heuristics with $N = 8$, there are 7 080 templates created, such that the whole dictionary is encoded using 3M template instances. The resulting binary form of the dictionary uses 2MB, which is 3 000 times smaller than the original dictionary.

In order to look up a word form in the dictionary, we split it into a prefix and an ending for all ending lengths from 1 to N. We then find templates associated with both the prefix and the ending. For each such template, we return the lemma corresponding to the prefix and the tag corresponding to the ending. The result is a set of lemma-tag pairs found during this procedure. This algorithm can be implemented efficiently – our implementation performs 500k word form lookups per second in the Czech morphological dictionary.

3.2 POS Tagger Methodology

The POS tagger is an offspring of Morče and Featurama research projects based on (Spoustová et al., 2009). For each form in the text, the morphological dictionary suggests all possible lemma-tag candidates and these lemma-tag pairs are disambiguated by the tagger. The tagger is implemented as supervised, rich feature averaged perceptron (Collins, 2002) and the classification features are adopted from (Spoustová et al., 2009).

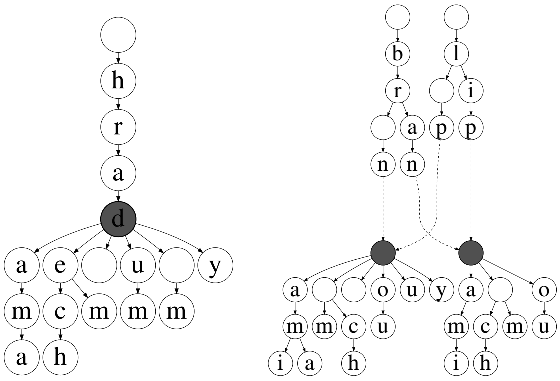

Figure 1: A simple trie for noun "hrad" (castle in Czech), and two lemmas sharing templates.

Czech language was trained on the training part of the Prague Dependency Treebank 2.5 (Bejček et al., 2012). The English language was trained on the standard training portion (Sections 0-18) of the Wall Street Journal part of the Penn Treebank (Marcus et al., 1993). In both cases, the system was tuned on the development set (Sections 19-21 in PTB/WSJ in English) and tested on the testing section (Sections 22-24 in PTB/WSJ in English).

3.3 POS Tagger Evaluation

An evaluation of POS taggers, which do not use external data, is shown in Table 1 for Czech and in Table 2 for English. MorphoDiTa reaches state-of-the-art results for Czech and nearly state-of-the-art results for English. The results are very similar for the three Czech systems, Morče, Featurama and MorphoDiTa, because in all three cases, they are implementations of (Spoustová et al., 2009). However, MorphoDiTa is the first end-to-end application released under a free license.

Due to rich morphosyntactic complexity of the Czech language and the positional tagging scheme proposed by (Hajič, 2004), there are 3 922 plausible tags in Czech (although only 1 571 unique tags actually appear in training data).

However, in many applications, only the first two tagging positions, which correspond to POS and sub-POS,[9] are actually needed for further processing, which greatly reduces the complexity of the task, leaving only 67 possible tags (64 in training data), although some morphological information, such as case, is lost.

lemmas, because the process of minimization cannot capture templates containing word forms (or their prefixes) of multiple lemmas.

[8] Which compresses to 454MB using `gzip -9`.

[9] Sub-POS is detailed set of POS labels, which includes basic properties such as the type of pronouns, conjunctions, adjectives, also some tense and active/passive/mood information for verbs, etc.

Tagger	Task	Accuracy
Morče	tag	95.67%
Featurama	tag	95.66%
MorphoDiTa	tag	95.75%
MorphoDiTa	lemma	97.80%
MorphoDiTa	lemma+tag	95.03%
MorphoDiTa	tag-first two pos.	99.18%

Table 1: Evaluation of Czech POS taggers.

Tagger	Accuracy
Morče (Spoustová et al., 2009)	97.23%
(Shen et al., 2007)	97.33%
MorphoDiTa	97.27%

Table 2: Evaluation of the English taggers.

An example of a full 15-position tag and the restricted 2-position tag for an adjective "zelený" is "`AAIS1----1A----`" and "`AA`", respectively. The first two positions are in fact quite similar to what the Penn-style tags encode (for English). MorphoDiTa therefore also offers models trained on such a restricted tagging scheme. The tagger evaluation for the 2-position, restricted tags is given in the last row of Table 1.

4 NameTag: Named Entity Recognizer

4.1 NER Methodology

The NE recognizer is an implementation of a research project by (Straková et al., 2013). The recognizer is based on a Maximum Entropy Markov Model. First, maximum entropy model predicts, for each word in a sentence, the full probability distribution of its classes and positions with respect to an entity. Consequently, a global optimization via dynamic programming determines the optimal combination of classes and named entities chunks (lengths). The classification features utilize morphological analysis, two-stage prediction, word clustering and gazetteers and are described in (Straková et al., 2013).

The recognizer is available either as a run-time implementation with trained linguistic models for Czech, or as a package which allows custom models to be trained using any NE-annotated data.

4.2 Czech Named Entity Corpus

For training the recognizer, Czech Named Entity Corpus(Ševčíková et al., 2007) was used. In this corpus, Czech entities are classified into a two-level hierarchy classification: a fine-grained set of 42 classes or a more coarse classification of 7 super-classes. Like other authors, we report the evaluation on both hierarchy levels.

System	F-measure (42 classes)	F-measure (7 classes)
(Ševčíková et al., 2007)	62.00	68.00
(Kravalová et al., 2009)	68.00	71.00
(Konkol and Konopík, 2013)	NA	79.00
(Straková et al., 2013)	79.23	82.82
NameTag CNEC 1.1	77.88	81.01
NameTag CNEC 2.0	77.22	80.30

Table 3: Evaluation of the Czech NE recognizers.

Corpus	Words / sec	RAM	Model size
CNEC 1.1	40K	54MB	3MB
CNEC 2.0	45K	65MB	4MB

Table 4: Evaluation of the NE recognizer tagger throughput, RAM and model size.

Czech Named Entity Corpus annotation allows ambiguous labels, that is, one entity can be labeled with two classes; however, NameTag predicts exactly one label per named entity, just like the previous work does (Straková et al., 2013).

Furthermore, CNEC also allows embedded entities, which is also somewhat problematic. NameTag always predicts only the outer-most entity (the embedding entity), although it is penalized by the evaluation score which includes correct prediction of the nested entities.

4.3 NER Evaluation

For comparison with previous work, we report results for the first version of the Czech Named Entity Corpus (CNEC 1.1). The linguistic models released with NameTag are trained on the most current version of the Czech Named Entity Corpus (CNEC 2.0), which has been recently released. We report our results for both CNEC 1.1 and CNEC 2.0 in Table 3.

5 Software Performance

We designed MorphoDiTa and NameTag as lightweight, efficient software with low resource usage.

Depending on the morphosyntactic complexity of the language and the selected tagging scheme, the MorphoDiTa tagger has a throughput around 10-200K words per second on 2.9GHz Pentium computer with 4GB RAM. Table 4 shows the system word throughput, allocated RAM and model size on such a machine for NameTag and Table 5 shows these parameters for MorphoDiTa.

Task	System	Words / sec	RAM	Model size
Czech tag	Morče (Spoustová et al., 2009)	1K	902MB	178MB
Czech tag	Featurama	2K	747MB	210MB
Czech tag	MorphoDiTa	10K	52MB	16MB
Czech tag–first two pos.	MorphoDiTa	200K	15MB	2MB
English Penn style	Morče (Spoustová et al., 2009)	3K	268MB	42MB
English Penn style	Featurama	10K	195MB	49MB
English Penn style	MorphoDiTa	50K	30MB	6MB

Table 5: Evaluation of the POS tagger throughput, RAM and model size.

	MorphoDiTa	NameTag
Binaries and source code	`https://github.com/ufal/morphodita`	`https://github.com/ufal/nametag`
Project website	`http://ufal.mff.cuni.cz/morphodita`	`http://ufal.mff.cuni.cz/nametag`
Demo	`http://lindat.mff.cuni.cz/services/morphodita/`	`http://lindat.mff.cuni.cz/services/nametag/`
Web services	`http://lindat.mff.cuni.cz/services`	
Language models	`http://lindat.mff.cuni.cz`	

Table 6: Web links to MorphoDiTa and NameTag downloads.

6 Release

Both MorphoDiTa and NameTag are free software under LGPL and their respective linguistic models are free for non-commercial use and distributed under CC BY-NC-SA license, although for some models the original data used to create the model may impose additional licensing conditions. Both MorphoDiTa and NameTag can be used as:

- a standalone tool,
- C++ library with Java, Python, Perl bindings,
- a web service, which does not require any installation at the user's machine whatsoever,
- an on-line demo.

MorphoDiTa and NameTag are platform independent and do not require any additional libraries. Web services and demo for the Czech and English languages are also available.

Table 6 lists the web links to all resources. The pre-compiled binaries and source code are available on GitHub, the language models are available from the LINDAT/CLARIN infrastructure and the documentation can be found at the respective project websites.

7 Conclusion

We released two efficient, light-weight POS- and NE taggers (especially efficient for inflective languages), which are available to a wide audience as an open-source, free software with rich API and also as an end-to-end application. The taggers reach state-of-the-art results for Czech and are distributed with the models. We are currently working on more language releases (Slovak, Polish and Arabic). We are also aware that the creation of the dictionary relies on the existence of a resource annotated with forms, lemmas and tags, which may not be readily available. Therefore, our future work includes developing a guesser for analyzing previously unseen but valid word forms in inflective languages, using only data annotated with disambiguated POS tags. We hope the release for Czech will prove useful for broad audience, for example for shared tasks which include Czech language data.

Acknowledgments

This work has been partially supported and has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education of the Czech Republic (project LM2010013). This research was also partially supported by SVV project number 260 104. We are grateful to the reviewers for comments which helped us to improve the paper.

References

Rie Kubota Ando and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 1–9. Association for Computational Linguistics.

Eduard Bejček, Jarmila Panevová, Jan Popelka, Pavel Straňák, Magda Ševčíková, Jan Štěpánek, and Zdeněk Žabokrtský. 2012. Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In

Martin Kay and Christian Boitet, editors, *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 231–246, Mumbai, India. IIT Bombay, Coling 2012 Organizing Committee.

Thorsten Brants. 2000. TnT: A Statistical Part-of-speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eric Brill. 1992. A Simple Rule-based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, pages 152–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nancy A. Chinchor. 1998. Proceedings of the Seventh Message Understanding Conference (MUC-7) Named Entity Task Definition. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, page 21 pages, April.

Michael Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics, July.

Rene De La Briandais. 1959. File Searching Using Variable Length Keys. In *Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference*, IRE-AIEE-ACM '59 (Western), pages 295–298, New York, NY, USA. ACM.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370. Association for Computational Linguistics.

J. Hajič. 2004. *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Karolinum Press.

Donald Knuth, 1997. *The Art of Computer Programming, Volume 3: Sorting and Searching, Third Edition*, chapter Section 6.3: Digital Searching, pages 492–512. Addison-Wesley.

Michal Konkol and Miloslav Konopík. 2011. Maximum Entropy Named Entity Recognition for Czech Language. In *Text, Speech and Dialogue*, volume 6836 of *Lecture Notes in Computer Science*, pages 203–210. Springer Berlin Heidelberg.

Michal Konkol and Miloslav Konopík. 2013. CRF-Based Czech Named Entity Recognizer and Consolidation of Czech NER Research. In Ivan Habernal and Vclav Matouek, editors, *Text, Speech, and Dialogue*, volume 8082 of *Lecture Notes in Computer Science*, pages 153–160. Springer Berlin Heidelberg.

Jana Kravalová and Zdeněk Žabokrtský. 2009. Czech named entity corpus and SVM-based recognizer. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, NEWS '09, pages 194–201. Association for Computational Linguistics.

H. Kucera and W. N. Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Libin Shen, Giorgio Satta, and Aravind Joshi. 2007. Guided Learning for Bidirectional Sequence Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 760–767, Prague, Czech Republic, June. Association for Computational Linguistics.

Drahomíra "johanka" Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-Supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece, March. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajič. 2013. A New State-of-The-Art Czech Named Entity Recognizer. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech and Dialogue: 16th International Conference, TSD 2013. Proceedings*, volume 8082 of *Lecture Notes in Computer Science*, pages 68–75, Berlin / Heidelberg. Západočeská univerzita v Plzni, Springer Verlag.

Jun Suzuki and Hideki Isozaki. 2008. Semi-Supervised Sequential Labeling and Segmentation using Giga-word Scale Unlabeled Data. *Computational Linguistics*, (June):665–673.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named entities in Czech: annotating data and developing NE tagger. In *Proceedings of the 10th international conference on Text, speech and dialogue*, TSD'07, pages 188–195. Springer-Verlag.