A New Czech Pipeline in Sketch Engine

Vlasta Ohlídalová and Miloš Jakubíček

Lexical Computing, Brno, Czechia
Faculty of Informatics, Masaryk University, Brno, Czechia

Abstract. This paper introduces a new Czech pipeline that is now available in Sketch Engine. It describes the tools used for this pipeline and for some of them, we add details of how they were altered in recent years. The most complex part discusses adjustment of the training data used for Czech language – the DESAM corpus – and its effect on accuracy of the POS tagging performed by `RFTagger`.

Keywords: Morphological analysis, corpora annotation.

1 Introduction

For people working with language corpora (linguists, lexicographers, terminologists, translators, ...), POS tagging and lemmatization is the most basic feature that they are utilizing daily. Sketch Engine, as one of the leading corpus managers, does indeed offer those for Czech corpora. However, the quality is not always as good as one would expect. For that reason, years after the previous version, a new version of Czech pipeline was introduced this year. The updated pipeline is based on a more recent version of the Majka [10] morphological database and uses the `RFTagger` [7] instead of the `desamb` tagger [9][1].

The core part of this contribution consists of semi-manual changes to the training data (the DESAM corpus [6]) so that the corpus matches the current version of the Majka morphological database.

2 Majka

Majka is a Czech morphological database containing 3,393,080 wordform + lemma + tag triplets, which are made of 903,888 distinct word forms and 46,000 lemmas. It comes together with a fast (about 1 million words per second) a fast morphological analyzer that queries the database.

A project of building a new Czech dictionary [3] has been running for the last year. The first stage – choosing the lemmas that should be included in the dictionary – is almost finished, so we have used this opportunity to compare the items in the dictionary with the Majka lexicon.

There are altogether 61,676 lemmas approved for the dictionary at the moment (including 407 MWEs). Out of these, approximately 4.5K were not

[1] The POS tagging evaluation comparing desamb and `RFTagger` can be found in [1].

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, pp. 101–107, 2024.

found in Majka lexicon, which makes around 7.4 % of all headwords. A part of it (13.5 %) were abbreviations, either written all with capital letters (10 %) or with a dot (3.5 %).

Without these, 3,950 headwords were left as candidates for new words that could be added to the lexicon. Table 1 shows some of the most common words according to frequency list from csTenTen23 [2].

Table 1: Selected words that were accepted in the Czech dictionary project, but they are missing in the Majka lexicon (ordered by frequency).

online	834843
info	164701
Wi-Fi	114165
blog	108700
on-line	104178
vs	98442
iPhone	82655
nej	69862
CO2	68886
fitness	67087
wellness	62956
off	60052
play-off	47187
D1	44251
e-shop	42226
elektro	35803
profi	32209
naživo	32056
wifi	31959
dovolatel	31883
live	31123
make-up	30680
insolvenční	30104

Some of these words can be added to Majka semi-automatically by finding the most similar already existing entries based on their suffix and then use the same patter for the new word. If there are more options, checking frequency of all generated word forms will show us what is the most probable one.

For example, the word "dovolatel" ends with known Czech suffix *-tel* (the suffix expresses that the derived noun denotates the subject of the action of the base verb [8]) and therefore it is very likely that it will be declined in the same manner as other words with this suffix (krotitel, podnikatel, ...).

A manual check is needed at the moment, though.

3 Desam

DESAM [6] is a disambiguated corpus of Czech texts originating from newspapers and scientific magazines. A lemma and morphological tag is specified for each token in the corpus. However, although manually disambiguated, the DESAM corpus is far from perfect. Some tokens were originally not assigned a tag, so the noun tag `k1` was added to all of them (most of those words are proper nouns that are unknown to the morphological analyzer; however, there are exception such as typos, rare words that the morphological analyzer does not contain or words written in non-standard manner). In the original version, there were 16,697 tokens with this tag and no further specification (1.7 % of all tokens).

Further on, even words that are properly disambiguated do not always match the lexicon entries available in Majka. This is because later changes in Majka were never properly entered in DESAM. Modifications in Majka mostly mean deleting one of possible POS interpretations for words where linguistic agreement isn't high enough and therefore annotation wasn't consistent.

The discrepancy between the current version of Majka and DESAM was solved semi-automatically by these steps:

1. If there is only one tag offered by the current version of Majka, it is considered correct. For example, for the word *sotva* (barely), three different POS tags were used in DESAM (adverb, a conjunction and particle). As the distinction isn't clear, the only available POS tag is now adverb, and all occurrences in DESAM were changed to adverb.
2. If more than one possible tag is offered in Majka, a table was created listing the options. In some cases, I could delete the ones that were clearly incorrect in the specific context and it wasn't necessary to check each case. For example, the word už (already) was tagged as particle (`k9`) in the original corpus. Now, the choice is between three tags: už (`k6eAd1tT`), už (`k5eAaImRp2nSrD`), už (`k5eAaPmRp2nS`), but the two later actually refer to lemma *úžit* (to make something narrow). In this case, we can choose the adverb tag without a manual check.
3. All other tokens found in Majka (but with multiple possible tags) were annotated manually.
4. For out-of-vocabulary tokens, the tag `k1` (noun) stayed in place.

After the aforementioned changes, 15,669 tokens were left with `k1` (noun) tag[2].

[2] This number should not be compared with the previously stated number of tokens with this tag (16,697), because MWEs were not taken into account. As described in the next section, 3,593 new tokens submerged by splitting MWEs. Altogether, we have therefore decreased the number of tokens with tag `k1` from 20,290 to 15,669.

To provide a better idea about the main changes that were performed in the corpus, Table 2 shows the tokens that were changed the most often in DESAM. For example, the token *i* (also) was originally tagged mostly as particle (in 70 %) and as conjunction in the remaining 30 %. While the particle interpretations is possible, it is very rare and by no means should cover 70 % of the cases. It is also not a straightforward task to distinguish those categories, so the only possible tag for *i* is conjunction now.

Table 2: The most frequently changed tokens in DESAM.

frequency	word	original tag	new tag
508	ještě	`k9`	`k6eAd1`
575	a	`k9`	`k8xC`
579	tedy	`k9`	`k8xC`
625	totiž	`k9`	`k8xC`
654	proto	`k8xC`	`k6eAd1xC`
686	tak	`k9`	`k6eAd1tMtQxCxD`
705	jako	`k9`	`k8xS`
778	jak	`k8xS`	`k6eAd1yR`
822	až	`k9`	`k8xS`
1,225	také	`k9`	`k6eAd1`
2,203	však	`k9`	`k8xC`
3,662	i	`k9`	`k8xC`

3.1 Multiword expressions

In the original data, some multi-word expressions were treated as one token. While justified in some cases, the annotation was not consistent and the same expressions sometimes appeared as one token and other times they were separated into multiple tokens. Also, this is not a desired state, as the current tools we are using would never recognize such MWEs as a single token.

To put it into numbers, there were 2,733 MWEs in the corpus. Once separated by spaces, it would make 6,326 tokens. Most MWEs were proper nouns (names of people or organizations), but there were others such as: *Na rozdíl od*, *Z hlediska*, *Mimo jiné*, *V důsledku*, *Ve srovnání s*, *V souvislosti s*, *V porovnání s*.

3.2 The results of Desam & Majka unification

As a simple mean to measure the effect of the modifications, `RFTagger`[3] was trained with each version of DESAM. A comparison of accuracy achieved for each version is provided in Table 3. The last column depicts results of `RFTagger` after various modifications described in [5]. In short, the work consists in altering tagset by:

- adding new attributes to specific words/group of words that linguistically differ from the rest, but the information is not considered in the current tagset (e.g. proper nouns or the word *být* that is mostly used as auxiliary verb in Czech texts).
- deleting attributes, but in the way that would not delete any information that cannot be obtained by the lexicon (e.g. change order of the attributes, delete the information about degree of adjectives and adverbs, that would be added back after training from the lexicon).

Table 3: Error rates of `RFTagger` trained on three different versions of DESAM. We consider the number "error `kgncp`" as the most important, as only attributes that are linguistically well-defined and cannot be simply taken from a dictionary are taken into account.

	original	original %	Majka	Majka %	changed	changed %
error	84,385	8.616	73,077	7.462	67,066	6.848
error in kgncp	**81,078**	**8.279**	**69,710**	**7.118**	**61,632**	**6.293**
k	20,775	2.121	9,634	0.984	7,696	0.786
c	38,058	3.886	37,976	3.878	32,414	3.31
g	19,500	1.991	19,284	1.969	18,859	1.926
n	11,547	1.179	11,013	1.125	10,366	1.058
p	15	0.002	12	0.001	13	0.001

Table 3 shows that the difference in the error rate of `RFTagger` trained on the original data and the version after matching DESAM to the current version of Majka is predominantly in the POS attribute, which is not surprising, as this is mostly what was altered in this step. The last version shows slightly better numbers in all categories, the biggest difference being the case.

The error rate of the tagger used in the new pipeline is therefore 6.293 % when being measured on the `kgncp` attributes (i.e. part-of-speech, grammatical case, number, gender and person) only.

[3] This tagger was chosen for simplicity reasons, as it is easy to use and very fast, making it possible to do various experiments in limited time span and evaluate them using the 10-fold cross-evaluation.

4 Known issues & future work

- Lemma disambiguation – lemmas in Czech language are rarely ambiguous when POS tag is taken into account. However, if such case occurs (for example for word form *karet*, there are two possible lemmas – karta (card) or kareta (loggerhead sea turtle), there is currently no disambiguation performed.
- Guesser for unknown words – the suggested solution for incorrectly guessed lemmas was drafted a long time ago (see [4]), but so far it was not implemented into the pipeline.

5 Conclusions

In this paper, we described a new version of the Czech pipeline used in Sketch Engine for corpus processing. The new pipeline uses a tagset that follows the most recent version of the Majka morphological database and uses the `RFTagger` trained on a compatible corpus (DESAM) for disambiguation. Previous experiments evaluating the `RFTagger` showed an accuracy of 93.7 % when measured on the `kgncp` attributes only.

Acknowledgements. This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2023062.

References

1. Jakubíček, M.: Rule-Based Parsing of Morphologically Rich Languages [online]. Disertační práce, Masarykova univerzita, Fakulta informatiky, Brno (2017 [cit 2024-11-17]), `https://is.muni.cz/th/h1xfz/`, sUPERVISOR : Aleš Horák
2. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. 7th International Corpus Linguistics Conference CL 2013 (07 2013)
3. Kovařík, F., Kovář, V., Blahuš, M.: On Rapid Annotation of Czech Headwords: Analysing the First Tasks of Czech Dictionary Express. In: Lexicography and Semantics, Proceedings of the XXI EURALEX International Congress (2024)
4. Kovář, V., Jakubíček, M.: DMoG: A Data-Based Morphological Guesser. In: RASLAN (2018)
5. Ohlídalová, V.: Improvements of the tagset used for automatic morphological analysis of Czech [online]. Master's thesis, Masaryk University, Faculty of ArtsBrno (2023 [cit 2024-04-24]), `https://theses.cz/id/hftnho/`, supervisor: RNDr. Miloš Jakubíček, Ph.D.
6. Pala, K., Rychlý, P., Smrz, P.: DESAM - Annotated Corpus for Czech. In: Conference on Current Trends in Theory and Practice of Informatics (1997)
7. Schmid, H., Laws, F.: Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In: Scott, D., Uszkoreit, H. (eds.) Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). pp. 777–784. Coling 2008 Organizing Committee, Manchester, UK (Aug 2008), `https://aclanthology.org/C08-1098`

8. Šimandl, J.: Slovník afixů užívaných v češtině. Karolinum (2016)
9. Šmerk, P.: Unsupervised Learning of Rules for Morphological Disambiguation. Lecture Notes in Computer Science **3206** (2004)
10. Šmerk, P.: Fast Morphological Analysis of Czech. In: RASLAN (2009), https://api.semanticscholar.org/CorpusID:3550809