First Step Towards Enhancement of Searching Within Medical Curriculum in Czech Language using Morphological Analysis

Matěj Karolyi, Jakub Ščavnický and Martin Komenda
Institute of Biostatistics and Analysis, Faculty of Medicine, Masaryk University, Brno, Czech Republic

Keywords: Morphological Analysis, Medical Curriculum, OPTIMED, Search Engine.

Abstract: This paper is focused on natural language processing techniques and morphological analysis of medical and healthcare curriculum corpus in the Czech language. We show an overview of basic steps that should lead to the improvement of search engine of an existing curriculum management system. The main goals of this initial phase are: to understand the morphological analysis and the currently used morphological analyser, to explore the possibilities of analyses and to deduce if these are suitable for the purposes of enhancing the search engine. As results, we explain how the morphological analysis helped us to reduce the density of unique words describing the curriculum and what are the key features of morphological analysis of the Czech language.

1 INTRODUCTION

Many innovative trends and ideas have appeared in medical and healthcare curriculum development, management and mapping during the last decade (Dent, Harden and Hunt 2017, Komenda, Víta, et al. 2015, Brauer and Ferguson 2015). However, the crucial goal is still the same: to better understand what students learn and what teachers really teach. A curriculum usually consists of a set of compulsory, compulsory-optional and optional courses, where a suitable combination of theoretically focused topics together with a clinical teaching base are taught. Curriculum designers have designed their study programmes using various modules and components (e.g. sequence blocks, events or competencies) and descriptive attributes (e.g. category, keywords or assessment form). As an illustration, the full metadata description of the General Medicine study programme at Masaryk University in the Czech Republic covers in total approximately 2,500 standard pages, i.e. pages containing 1,800 characters each (Komenda 2015). For a global in-depth inspection and analysis of such a huge amount of data records, an accurate search engine built on a given curriculum is needed.

1.1 OPTIMED Platform

For the purposes of optimising and harmonising the medical and healthcare curriculum, an original curriculum management system (CurrMS) called OPTIMED has been developed (Komenda, Schwarz, et al. 2015). The OPTIMED platform covers a detailed description of formal metadata associated with the so-called building blocks: medical sections, disciplines, courses, learning units, and interconnections to the learning outcomes. Some of these building blocks are associated with printed and e-learning study materials. The organisation of these metadata is mostly designed in the parametric form supplemented with textual attributes in order to allow further pre-processing and analyses. OPTIMED is based on a modular design that divides the platform into four independent modules providing separate functionalities according to its use in practice: (i) Learning outcome register, (ii) Learning unit register, (iii) Browser, (iv) Reporting toolbox.

The Browser module is a full-text search engine, which is designed to answer user search queries and to provide the required results for the broad academic community (students, teachers, guarantors, curriculum designers and faculty management).

.

Karolyi, M., Ščavnický, J. and Komenda, M.
First Step Towards Enhancement of Searching Within Medical Curriculum in Czech Language using Morphological Analysis.
DOI: 10.5220/0006757902880293
In *Proceedings of the 10th International Conference on Computer Supported Education (CSEDU 2018)*, pages 288-293
ISBN: 978-989-758-291-2

1.2 Focus and Goals of the Study

This paper focuses on natural language processing techniques used to process a medical and healthcare curriculum corpus and provides the summary of basic steps that should lead to the improvement of search engine in the OPTIMED platform. The main motivation is enhancement of already existing search engine. Thanks to the enhancement by morphological analysis, we will be able to return results that are even more accurate. Because we are at the very start of our efforts, we have defined our primary areas of interest as follows: to understand the morphological analyser and morphological analysis itself; to study the analyser's interface and general work with it; to explore its possibilities; to analyse its outputs and to evaluate if it is suitable for our purposes; and finally, to design a procedure for its integration into the existing platform. We have identified the following research questions, which are focused on how to perform morphological transformations of a medical text: (i) Can we adopt the morphological analyser to enhance the medical curriculum management system (CurrMS)? (ii) Are we able to identify features of morphological analysis that make the CurrMS search engine more effective (more accurate and faster)? (iii) Can we implement specific morphological analysis features as an extension of the CurrMS search engine?

2 METHODS

A morphological analysis is not a trivial problem in the Czech language. There are some sophisticated tools – the so-called morphological analysers – that can be used for these purposes. One of them, Ajka (Sedláček), is based on finite state machines; it is well documented and ready to use via an online user interface reachable by the web browser. Moreover, an interface for PERL and Python is implemented by the tool provider. Based on Ajka, Majka is another morphological analyser for the Czech language. Majka is a fast analyser which assigns basic word forms and their corresponding grammatical tags to one word on the input. This assignment is done only in case when a given word on input is present in the Majka's vocabulary. This vocabulary can be extended on demand. Majka is available in two forms: (i) as a command line tool for batch processing of text files, (ii) as a library in the form of calling functions in the C language. Despite the fact that Majka is built on Ajka and therefore gives more less the same results, it is a more flexible, faster and independent implementation based on a finite state machine (Šmerk and Rychlý 2009).

2.1 Majka Morphological Analyser

We decided to use the Majka morphological analyser to enhance the full-text search within the OPTIMED curriculum management system. The main reason for selecting Majka was the availability of its binaries, the vocabulary for assigning basic word forms and tags to analysed words and the possibility of simple integration in CurrMS using the command line. Furthermore, the vocabulary we use was enriched by a set of medical terms by courtesy of its authors.

Calling the Majka analyser on a word input requires the specification of a vocabulary file. The vocabulary files are essential for the finite state machine, which is the computational engine of whole analyser. There are several types of vocabularies: (i) w-lt: from word to basic forms and tags, (ii) l-wt: from basic form to all word forms and tags, (iii) lt-w: from lemma and tag to word. All vocabularies in their basic form can be downloaded from the developer's website (Šmerk 2007). We chose the right type of vocabulary depending on the desired action from Majka.

The basic command form using the w-lt vocabulary is as follows:

```
$ echo robot|
           majka -f majka.w-lt
```

This command passes the word "robot" to Majka, which subsequently returns output (a list of relevant words with tags) using the w-lt type of vocabulary for the Czech language. The –f option enables the user to specify the desired dictionary file. By default, Majka is case-sensitive to names and titles; therefore, one needs to add the option `-i` in order to correctly analyse words in lowercase.

2.2 Attributes of Search Analysis

Majka's output for a particular word is a list of relevant basic word forms found in the vocabulary with corresponding grammatical tags, where each word and its tag is divided by a colon.

For example, analysis of the word 'robot' returns the following output:

```
robot:k1gInSc1

robot:k1gInSc4
robot:k1gMnSc1
robota:k1gFnPc2
```

Czech is a fusional language, which among others means that each tag consists of several pairs of strings which represent morphological features such as the part of speech (pars orationis), gender (genus grammaticum), number (numerus grammaticus), paradigm (paradigma), case (casus garmmaticus) or tense (tempus grammaticum). Some of these properties are not available to all words, so tags might be of different lengths. The first character of a pair identifies the property, whereas the second character is its value. In the first line of the example with word “robot”, the grammatical tag k1gInSc1 is returned. Meanings of each part of the tag are as follows:

- k1 – Character ‘k’ represents the part of speech. In this case, ‘1’ stands for noun. In the Czech language, we recognize 10 different parts of speech, see Table 1.
- gI – Character ‘g’ stands for the gender, while ‘I’ specifies the masculine.
- nS – Character ‘n’ represents the number, while ‘S’ stands for singular.
- c1 – Character ‘c’ stands for the case of a particular word. In this case, ‘1’ specifies the first paradigm. In the Czech language, we recognize 7 different cases.

In summary, the first morphological result tells us that one of the basic forms of the word “robot” is indeed the word “robot” itself, which is a noun with a masculine gender, in singular form, and in nominative.

2.3 Integration to the Search Engine

The strong point of a fully integrated Majka is that all functions can be called on demand. Thanks to its application programming interface (API), functions can be called and their returned outputs can be processed by an existing PHP application, the OPTIMED CurrMS. This application is based on the Symfony framework (SensioLabs) and has been designed for the development of web-based applications and systems; the whole application is used for curriculum development and management.

Our long-term goal is to refine the existing full-text search engine by a morphological analysis, with the aim of obtaining more relevant results to queries of users. We have created two pieces of code especially for the purposes of work with Majka.

The first piece of code is the CallMajka service: it takes any string as the input and, for each word, returns the most relevant basic form and a corresponding tag based on a specific decision algorithm. This service is created only as part of the application responsible for the entire preparation of data for Majka and also for all post-processing of data that are produced by Majka.

The second piece of code is the CallMajkaCommand, which is a PHP command that calls Majka on a single input word. The command is designed to be the only entry point of Majka’s output to the application. All actions which somehow communicate with the Majka’s interface must use a functionality implemented within this command.

3 RESULTS

We were able to integrate and start using Majka easily, thanks to the well-prepared application programming interface, a robust documentation and an on-demand created vocabulary. The vocabulary was created from the whole content of the OPTIMED platform. Therefore, we were sure that the vocabulary used for our further analysis contained the most relevant words depending on the current content.

During our initial phase of Majka integration into the web-based application, we created several analysis outputs. Their main purpose was to show the power of morphological analyser for our needs. We examined the distribution of occurring word types (parts of speech) on the input, counts of most frequent words and the analyser’s ability to reduce the number of unique words of the corpus.

3.1 Dataset Transformation

The process of data transformation uses the CallMajka service, which removes specific punctuation marks (~`!@#$%^&*()_={}[]:;'<>,.?/) using regular expression (regex), puts text to lowercase and then splits it to separate words. Afterwards, the CallMajka service calls Majka on each word by the command line. Finally, one basic word form with a corresponding tag for each input word is selected algorithmically. For the purpose of enhancing the medical query search, we agreed that substantives, adjectives and verbs in singular form and in nominative are most informative. Therefore, based on these conditions, this algorithm decides which basic form is the most relevant one. If the vocabulary does not recognise the input word, the algorithm labels the word as unclassified. The input dataset that we used for the analysis contained more than 860,000 words.

Using the CallMajka service, we acquired an output dataset of a similar length but containing transformed words. A mapping dictionary that consisted of each unique word, an algorithmically selected basic word form, corresponding tag and its frequency in dataset was the side product of this transformation.

For example, the sentence *"radioterapie tvoří významnou součást léčby nádorových onemocnění"* on the input is transformed to an output which has the following form: *"radioterapie tvořit významný součást léčba nádorový onemocnění"*. All words from the input are also present in the output (in same places as in the original sentence) but they have different word forms, which suits us more.

3.2 Output Structure

Using the dictionary output of dataset transformation, we analysed the distribution of parts of speech using the Majka tags. The division to presented categories is created by Majka by default (Jakubícek, Kovár and Šmerk 2011). Corresponding results are shown in Table 1.

Table 1: Distribution of word types in output dataset.

Part of speech	Dictionary occurrence (%)	Dataset occurrence (%)
Unclassified	0.48013	0.15948
Interjections	0.00049	0.00113
Nouns	0.25877	0.42941
Adjectives	0.20702	0.19262
Pronouns	0.00227	0.01647
Numerals	0.00107	0.00176
Verbs	0.03725	0.05091
Adverbs	0.00979	0.01573
Prepositions	0.00026	0.01702
Conjunctions	0.00042	0.00518
Particles	0.00023	0.00140
Other	0.00232	0.10889

It was obvious that the "unclassified" category occupied a relatively large part of the distribution. This result meant that the used vocabulary failed to identify (i.e. did not contain) many of the input words. However, by identifying these words and with the help of Majka's authors, we were able to add them to the vocabulary iteratively. As we have already mentioned earlier, we were mainly interested in substantives, adjectives and verbs because they carry the most information. Due to the identification of these words and their further usage, we were able to enhance the search engine. Furthermore, we were able to reduce the density of unique words in output by almost 28% in comparison to input dataset.

3.3 Most Frequent Words

In order to classify the most frequent words that are relevant for our purposes, both input and output datasets were cropped to subsets of words that were at least three characters long. This action mainly involved removing interjections, prepositions, conjunctions and particles, i.e words that were not informative enough. Afterwards, 30 most frequently occurring words and corresponding total frequencies were selected from both datasets. Results are shown in Tables 2 and 3 below.

Table 2: Top 30 words and their frequencies from the input dataset.

Czech word	English translation	Frequency
vyšetření	examination	3,264
při	at	3,003
onemocnění	illness	2,652
syndrom	syndrome	2,550
pro	for	2,463
jsou	are	2,205
základní	basic	2,166
poruchy	disorders	2,064
nebo	or	1,992
jejich	their	1,935
cvičení	exercise	1,631
nádory	tumours	1,619
zkouška	exam	1,506
forma	form	1,499
léčba	treatment	1,438
terapie	therapy	1,408
infekce	infection	1,372
lékařství	medicine	1,316
závěrečná	final	1,303
jako	as	1,294
nemoci	diseases	1,285
poranění	injury	1,169
přednáška	lecture	1,146
popsat	describe	1,109
diagnostika	diagnosis	1,103
buňky	cells	1,082
příznaky	symptoms	1,058
metody	methods	1,043
systému	system	1,016
význam	importance	1,011

Table 3: Top 30 words and their frequencies from the output dataset.

Czech word	English translation	Frequency
být	be	11,320
syndrom	syndrome	3,384
vyšetření	examination	3,383
porucha	disorder	3,013

Table 4: Top 30 words and their frequencies from the output dataset (cont.).

pře	to argue	3,010
léčba	treatment	2,957
onemocnění	illness	2,940
základní	basic	2,862
nádor	tumour	2,702
klinický	clinical	2,686
nemoc	disease	2,672
pro	for	2,463
systém	system	2,337
buňka	cell	2,140
protein	protein	2,100
forma	form	2,002
nebo	or	1,992
pacient	patient	1,980
student	student	1,967
který	which	1,965
jejich	their	1,935
faktor	factor	1,900
popsat	describe	1,879
infekce	infection	1,858
terapie	therapy	1,727
receptor	receptor	1,721
cvičení	exercise	1,715
diagnostika	diagnostics	1,714
znát	know	1,604
jednotlivý	individual	1,583

4 DISCUSSION

We consider our work to be very promising because we were able to answer all stated research questions, to achieve meaningful results and to complete successfully the first step towards the integration of morphological analyser into the search engine of the OPTIMED curriculum management system. During this phase, we closely cooperated with authors of the Majka morphological analyser and we have extended the analyser's vocabulary by medical terms from the OPTIMED database. Moreover, we have started to analyse the OPTIMED curriculum corpus (input dataset). The input dataset was transformed using the Majka analyser and analysed with a corresponding mapping dictionary.

Our subsequent analysis of word types (parts of speech) showed that we reduced the density of unique words in the output by almost 28%. Results of the frequency analysis proved that the majority of basic forms of transformed words come from the medical background. This means that the transformation process and the selected algorithm can indeed map the curriculum terminology according to our expectations.

Therefore, we plan to implement Majka fully as a useful and effective tool for the OPTIMED search engine improvement. Our concept is as follows: nowadays, our users can search the OPTIMED content by entering a particular search query; as a result, they get a list of learning units containing the entered keyword or phrase. The implementation of Majka will reduce the searched content and increase the accuracy of returned results at the same time. In order to provide a fast OPTIMED search engine, curriculum data will be pre-computed; but in the access time, only the entered query will be computed. During our future work will be defined concrete process of search data preparation and results retrieval using the morphological analysis as one of enhancements.

ACKNOWLEDGEMENTS

The authors were supported from the following grant projects: (i) MEDCIN – Medical Curriculum Innovations – Reg. No. 2015-1-CZ01-KA203-013935; this project has been funded from the European Commission ERASMUS+ programme; (ii) OPTIMED – implementation of new features for effective searching – Reg. No. MUNI/FR/1568/2016; this project has been funded from the specific research and rector's programme of the Masaryk University.

REFERENCES

Brauer, D.G. and Ferguson, K.J. 2015. The integrated curriculum in medical education: AMEE Guide No. 96. *Medical teacher*, 37(4), pp.312–322.

Dent, J., Harden, R.M. and Hunt, D. 2017. *A practical guide for medical teachers*. Elsevier Health Sciences.

Jakubícek, M., Kovár, V. and Šmerk, P. 2011. Czech morphological tagset revisited. *Proceedings of Recent Advances in Slavonic Natural Language Processing*, 20, pp.29–42.

Komenda, M., Schwarz, D., Vaitsis, C., Zary, N., Štěrba, J. and Dušek, L. 2015. OPTIMED Platform: Curriculum Harmonisation System for Medical and Healthcare Education. *Studies in Health Technology and Informatics*, 210, pp.511–515.

Komenda, M., Víta, M., Vaitsis, C., Schwarz, D., Pokorná, A., Zary, N. and Dušek, L. 2015. Curriculum Mapping with Academic Analytics in Medical and Healthcare Education. *PloS one*, 10(12).

Sedláček, R. Ajka - morphological analyser of Czech. Available from: https://nlp.fi.muni.cz/projekty/ajka/index.html [Accessed November 5, 2017].

SensioLabs Symfony, High Performance PHP Framework for Web Development. Available from: http://symfony.com/ [Accessed November 13, 2017].

Šmerk, P. 2007. Free natural language morphology for Czech, Slovak, Polish, Swedish, German, French, Italian, English, Portuguese, Catalan, Welsh, Spanish, Galician, Asturian and Russian. Available from: https://nlp.fi.muni.cz/czech-morphology-analyser/ [Accessed November 14, 2017].

Šmerk, P. and Rychlý, P. 2009. Majka – rychlý morfologický analyzátor. *Masaryk University*. Available from: https://www.muni.cz/en/research/publications/935762 [Accessed November 5, 2017].