Czech Text Document Corpus v 2.0

Pavel Král[1,2]**, Ladislav Lenc**[1,2]

[1]Dept. of Computer Science & Engineering
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic

[2]New Technologies for the Information Society
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic

Abstract

This paper introduces "Czech Text Document Corpus v 2.0", a collection of text documents for automatic document classification in Czech language. It is composed of the text documents provided by the Czech News Agency and is freely available for research purposes at `http://ctdc.kiv.zcu.cz/`. This corpus was created in order to facilitate a straightforward comparison of the document classification approaches on Czech data. It is particularly dedicated to evaluation of multi-label document classification approaches, because one document is usually labelled with more than one label. Besides the information about the document classes, the corpus is also annotated at the morphological layer. This paper further shows the results of selected state-of-the-art methods on this corpus to offer the possibility of an easy comparison with these approaches.

Keywords: corpus, Czech, document classification, multi-label, text

1. Introduction

Automatic classification (or categorization) of text documents is very important for information organization and storage because of the significant increase of the amount of electronic documents and the rapid growth of the Internet.

Many efficient approaches have been proposed. They are usually based on supervised machine learning. The documents are projected into the so-called vector space model, basically using the words as features for various classification algorithms. The approaches differ in the used methods, however the common point is that all of them need an annotated document corpus to train the parameters.

A sufficient number of the corpora in several languages, particularly in English, is freely available. However, to the best of our knowledge, the Czech one is missing.

The main goal of this paper consists in presenting a corpus of Czech text documents. It is composed of real newspaper articles provided by the Czech News Agency (ČTK)[1] and is available for research purposes for free. It is created for a straightforward comparison of the document classification approaches in Czech. One document is usually labelled with more than one label, therefore this corpus is usually used for evaluation of multi-label document classification. Besides the information about the document classes, the corpus is also morphologically annotated. The morphological annotation has been done fully automatically.

Another research contribution of this paper represents the reported results of selected state-of-the-art methods on this corpus to offer the possibility of an easy comparison with these approaches.

The paper structure is as follows. The following section presents other text corpora for document classification freely available for research purposes. Section 3. details our corpus. Section 4. presents the results of the selected state-of-the art methods on this dataset. The last section concludes the paper.

2. Other Text Corpora

Some important existing text classification corpora in several languages are described below.

2.1. Reuters-21578

Reuters-21578[2] corpus is a collection of 21,578 documents. The training part is composed of 7769 documents, while 3019 documents are reserved for testing. The number of possible categories is 90 and the average label/document number is 1.23. This dataset is the most frequently used benchmark for English.

2.2. RCV1-V2

RCV1-V2[3] (Lewis et al., 2004) is another text classification test collection which is freely available for research purposes. It contains about 800,000 manually categorized newswire English stories from Reuters, Ltd. RCV1 contains English documents, while RCV2 is composed of text documents in French, German, Italian, Spanish and others. This dataset is also widely used as a benchmarking corpus for English and the languages mentioned above.

2.3. Other Corpora

For other corpora dedicated for text categorization, you can visit for instance `http://mulan.sourceforge.net/datasets-mlc.html`.

3. Corpus Description

3.1. General Information

The main part (for training and testing) of the Czech Text Document Corpus v 2.0 is composed of 11,955 real newspaper articles provided by the Czech News Agency. We provide also a development set which is intended to be used for tuning of the hyper-parameters of the created models. This set contains 2735 additional articles.

---

[1]http://www.ctk.eu/

[2]http://www.daviddlewis.com/resources/testcollections/reuters21578/
[3]https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual,-

The documents belong to different categories (classes) such as weather, politics, sport, culture, etc. Each document is associated with one or more labels (classes). It is thus beneficial to use it for multi-label document classification scenarios. The multi-label classification task is considerably more important than the single-label classification because it usually corresponds better to the needs of the current applications.

The total category number is 60[4] out of which 37 most frequent ones are used for classification. The reason of this reduction is to keep only the classes with the sufficient number of occurrences to train the models. The corpus was annotated by professional journalists from the Czech News Agency. All documents are further automatically morphologically annotated using UDPipe tool.

3.1.1. Statistical Information

Table 1 shows the statistical information about the corpus[5]. It shows for instance that lemmatization decreases the vocabulary size from 150,899 to 82,986 which represents the reduction by 45%. Another interesting observation is the distribution of the POS tags in this corpus.

Figure 1 illustrates the distribution of the documents depending on the number of labels. It shows that the maximal number of categories associated with one document is eight, the majority of documents has two categories and the average label number is 2.55.

Figure 2 shows the distribution of the document lengths (in word tokens). The documents are relatively long and the longest documents are composed of more than 7000 word tokens. Another interesting information is that the most documents (about 2000) contain at most 50 words. The average document size is 277 words.

3.1.2. Download

This dataset is licensed under the Attribution-NonCommercial-ShareAlike 3.0 Unported License[6]. Therefore it is freely available for research purposes, however any commercial use is strictly excluded. This corpus is possible to download at http://ctdc.kiv.zcu.cz/.

Unit name	Number	Unit name	Number
Document	11,955	Word	3,505,965
Category	60	Unique word	150,899
Cat. classif.	37	Unique lemma	82,986
Noun	894,951	Punct	553,099
Adjective	369,172	Adposition	340,785
Verb	287,253	Numeral	265,430
Pronoun	258,988	Adverb	144,791
Coord. conj.	100,611	Determiner	84,681
Pronoun	74,340	Aux. verb	70,810
Subord. conj.	41,487	Particle	12,383
Symbol	2420	Interjection	142
Other	4126		

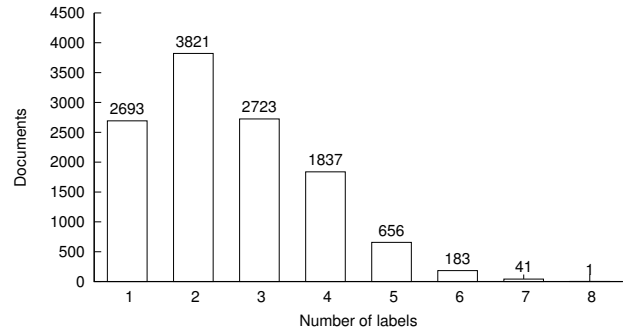Table 1: Corpus statistical information

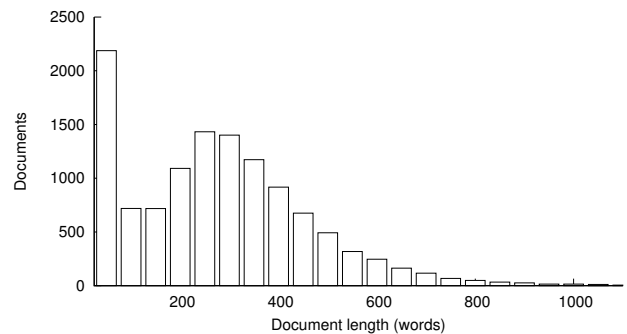Figure 1: Distribution of documents depending on the number of labels

Figure 2: Distribution of the document lengths

3.2. Morphological Annotation

As already mentioned, we used UDPipe tool (Straka and Straková, 2017)[7] for automatic morphological analysis of the corpus. This tool provides en efficient pipeline for sentence segmentation, tokenization, POS tagging, lemmatization and dependency parsing. It also contains models for 50 languages of universal dependencies (UD) 2.0. This system can be further used with data in CoNLL-U format[8].

According to the authors, the accuracy of the lemmatizer and of the POS tagger are both about 98% on the UD version 2 of the Prague Dependency Treebank (PDT) 3.0[9]. The performance of the syntactic parsing is represented by labelled attachment score (LAS) which is about 83%. This system uses 17 part-of-speech categories drawn from the revised version of the Google universal POS tags.

3.3. Technical Details

Text documents are stored in the individual text files using UTF-8 encoding. Each filename is composed of the serial number and the list of the categories abbreviations separated by the underscore symbol and the *.txt* suffix. Serial numbers are composed of five digits and the numerical series starts from the value one.

For instance the file *00046_kul_nab_mag.txt* represents the document file number 46 annotated by the categories *kul* (culture), *nab* (religion) and *mag* (magazine selection). The content of the document, i.e. the word tokens, is stored in one line. The tokens are separated by the space symbols.

---

[4] This list is reported in Table 3 in Section 8..

[5] Development set is excluded from all analyses reported in this section.

[6] http://creativecommons.org/licenses/by-nc-sa/3.0/

[7] http://ufal.mff.cuni.cz/udpipe

[8] http://universaldependencies.org/format.html

[9] http://ufal.mff.cuni.cz/pdt3.0/

Every text document was further automatically mophologically analyzed. This analysis includes lemmatization, POS tagging and syntactic parsing. The fully annotated files are stored in *.conll* files. We also provide the lemmatized form, file with suffix *.lemma*, and appropriate POS-tags, see *.pos* files. The tokenized version of the documents is also available in *.tok* files.

3.4. Evaluation Protocol

All following experiments use the five-fold cross validation procedure, where 20% of the corpus is reserved for testing and the remaining part for training of the models. The development set was used to tune the hyper-parameters of the models.

For evaluation of the multi-label document classification, it is used the standard Precision (P), Recall (R) and F-measure (F_m) metrics (Powers, 2011). For evaluation of the single-label classification, the authors use the standard accuracy metric.

The confidence interval of the experimental results is 0.6% at a confidence level of 0.95 (Press et al., 1996).

4. Experiments

The two following sections present the results of the selected classification algorithms on this dataset. The first section deals with multi-label document classification while the second one describes the classification score of single-label classification task.

4.1. Multi-label Document Classification

The first reported approach (Hrala and Král, 2013a) uses Bag of Words (BoW) to create the features. Non-significant words are removed using Part of speech (POS) filtering and for feature selection, the mutual information method is used. In the original paper, the authors show the results of three classifiers, namely Naive Bayes, Maximum entropy (ME) and Support vector machine with three traditional multi-label classification approaches. Only the best classification accuracy obtained by ME classifier is reported in this paper.

The second method (Brychcín and Král, 2014) proposes novel unsupervised features using an unsupervised stemmer, latent Dirichlet allocation and semantic spaces (HAL and COALS). These features are integrated with word features to improve classification results. Multi-label classification scenario is realized using a set of binary classifiers. Maximum entropy model is used for classification.

Neural networks are very popular in natural language processing field today and they outperform many state-of-the-art approaches with only very simple preprocessing. The following approach (Lenc and Král, 2017) uses two different feed-forward neural networks, namely multi-layer perceptron (MLP) and convolutional neural networks (CNN) to achieve new state-of-the art results on this corpus. The authors use thresholding to realize multi-label document classification task.

The results of the above described approaches are illustrated in Table 2.

Method	P	R	F_m
ME (Hrala and Král, 2013a)	-	-	76.8
words+ME (Brychcín and Král, 2014)	88.1	72.7	79.7
unsup+ME (Brychcín and Král, 2014)	89.0	75.6	81.7
MLP (Lenc and Král, 2017)	83.7	83.6	83.9
CNN (Lenc and Král, 2017)	86.4	82.8	84.7

Table 2: Multi-label document classification results of the different approaches

4.2. Single-label Document Classification

The authors (Hrala and Král, 2013b) evaluate five feature selection methods and three classifiers on this corpus. Lemmatization and POS tagging are used for a precise representation of the Czech documents. It was demonstrated that POS-tag filtering is very important, while the lemmatization plays only a marginal role for classification. In this work, only the first document class was considered for classification and the authors consider it as the main document category. The best classification accuracy was obtained by SVM classifier and is 91.2%.

5. Conclusions

This paper introduced a novel collection of Czech text documents. This corpus is composed of real newspaper articles provided by the Czech News Agency and is available for research purposes for free. It was created to facilitate a straightforward comparison of the document classification approaches in Czech language.

This corpus is particularly intended to evaluate multi-label document classification approaches, because one document is usually associated with more than one label. Besides the information about the document classes, the corpus is automatically annotated at morphological layer. This paper further shows the results of the selected state-of-the-art algorithms on this corpus to offer the possibility of a straightforward comparison with the future research.

We plan to submit this corpus to be a part of the Language Research Infrastructure of LINDAT/CLARIN project[10].

6. Acknowledgements

7. Language Resource References

Brychcín, T. and Král, P. (2014). Novel unsupervised features for Czech multi-label document classi cation. In *13th Mexican International Conference on Artificial Intelligence (MICAI 2014)*, pages 70–79, Tuxtla Gutierrez, Chiapas, Mexic, 16-22 November. Springer.

Hrala, M. and Král, P. (2013a). Multi-label document classification in Czech. In *16th International conference on Text, Speech and Dialogue (TSD 2013)*, pages 343–351, Pilsen, Czech Republic, 1-5 September. Springer.

---

[10]https://lindat.mff.cuni.cz/en

Hrala, M. and Král, P. (2013b). Evaluation of the Document Classification Approaches. In *8th International Conference on Computer Recognition Systems (CORES 2013)*, pages 877–885, Milkow, Poland, 27-29 May. Springer.

Lenc, L. and Král, P. (2017). Deep neural networks for Czech multi-label document classification. *arXiv preprint arXiv:1701.03849*.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1996). *Numerical recipes in C*, volume 2. Citeseer.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

8. Appendix

Abbr.	Category in Czech	English translation
aut	Automobilový průmysl	Automobile industry
bos	Bohemika	Czech Rep. from abroad
bsk	Sklářský průmysl	Glass industry
bua	Burzy akciové	Stock exchanges
buk	Burzy komoditní	Commodity exchanges
bup	Burzy peněžní	Currency exchanges
bur	Burzy	Exchanges
cen	-	-
che	Chemický a farmaceutický průmysl	Chemical and pharmaceutical industry
den	Zpravodajské deníky	News schedules
dpr	Doprava	Transport
dre	Dřevozpracující průmysl	Woodworking industry
efm	Firmy	Companies
ekl	Životní prostředí	Environment
eko	Ekologie	Ecology
ene	Energie	Energy
eur	Evropská unie - zprávy	European union - news
fin	Finanční služby	Financial services
for	Parlamenty a vlády	Parliaments and governments
fot	Fotbal - zprávy	Soccer
hok	Hokej - zprávy	Ice hockey
hut	Hutnictví	Metallurgy
kat	Neštěstí a katastrofy	Accidents and disasters
kul	Kultura	Culture
mag	Magazínový výběr	Magazine selection
mak	Makroekonomika	Macroeconomics
med	Média a reklama	Media and advertising
met	Počasí	Weather
mix	Mix	Mix
mot	Motorismus	Motoring
nab	Náboženství	Religion
obo	Obchod	Trade
odb	Práce a odbory	Labour and Trade Unions
pit	Telekomunikace a IT	Telecommunications & IT
pla	Plány zpravodajství ČTK	Events news
pod	Politika ČR	Czech Republic Politics
pol	Politika	Politics
prg	Pragensie	Prague issues
prm	Lehký průmysl	Light industry
ptr	Potravinářství	Food industry
reg	Region	Region
sko	Školství	Educational system
slo	Slovenika	Slovakia from abroad
slz	Služby	Services
sop	Sociální problematika	Social problems
spc	-	-
spl	Životní styl	Life style
spo	Sportovní zpravodajství	Sports
sta	Stavebnictví a reality	Building industries and property
str	Strojírenství	Mechanical engineering
sur	Suroviny	Raw materials
tlk	Telekomunikace	Telecommunications
tok	Textil	Textile
tur	Cestovní ruch	Tourism
vat	Věda a technika	Science and technology
zah	Zahraniční	Foreign
zak	Kriminalita a právo	Criminality and law
zbr	Zbraně	Arms
zdr	Zdravotnictví	Health service
zem	Zemědělství	Agriculture

Table 3: List of the categories