Evaluation Set for Slovak News Information Retrieval

Daniel Hládek, Jan Staš, Jozef Juhár
Department of Electronics and Multimedia Communications
Technical University of Košice, Slovak Republic
daniel.hladek@tuke.sk, jan.stas@tuke.sk, jozef.juhar@tuke.s

Abstract

This work proposes an information retrieval evaluation set for the Slovak language. A set of 80 queries written in the natural language is given together with the set of relevant documents. The document set contains 3980 newspaper articles sorted into 6 categories. Each document in the result set is manually annotated for relevancy with its corresponding query. The evaluation set is mostly compatible with the Cranfield test collection using the same methodology for queries and annotation of relevancy. In addition to that it provides annotation for document title, author, publication date and category that can be used for evaluation of automatic document clustering and categorization.

Keywords: information retrieval evaluation, Cranfield test collection, Slovak language

1. Introduction

In order to accelerate research of information retrieval (IR) techniques for less resourced languages we present an evaluation benchmark consisting of a document set, queries and relevant documents. This paper proposes a method to evaluate information retrieval system for a morphologically rich language - Slovak. Usually it is hard to evaluate a design of an IR system for a language different from English because of the lack of properly annotated databases, especially if number of speakers of the language is lower. Importance of IR for non-English languages rise together with number and size of national Internets and amount of human-entered textual data in business and government databases (Korra et al., 2011; Lazarinis et al., 2009).

Most problems in IR are language independent, such as document representations, clustering or classification. On the other hand, there are several issues that are bound to a specific language and have to be solved in order to build a successful IR system. This contribution aims to fill this gap and helps to evaluate techniques that are adapted to a non-English language. Slovak language has lesser number of language-specific resources available that are necessary to build an IR system, such as WordNet, morphological analysis tools or vocabularies. Languages with a rich morphology and free word order usually require specialized methods for stem identification, word and sentence boundary detection or different features for chunking or named entity recognition.

2. Previous Works

Most of the research in the field of information retrieval is focused on the English language. There is a large amount of evaluation benchmarks for IR in English. The most basic is the Cranfield collection based on work (Cleverdon, 1967). It contains a set of information needs from a database of abstracts. TREC (Simpson et al., 2014) and CLEF (Suominen et al., 2014) are the biggest series of evaluation campaigns focused on various tasks of IR. Multi-lingual (Peters et al., 2012) and cross-lingual IR are gaining a lot of attention, but most of the current evaluation databases contain just couple of the most commonly used languages such as Chinese or French. There is a proposal for Czech (Straková and Pecina, 2010) which has similar properties than Slovak and is evaluated using CLEF 2007 Ad-Hoc Track (Nunzio et al., 2007).

3. Linguistic Issues of IR in Slovak

Before implementation of IR system for the Slovak language, or other similar Slavic language with rich morphology and arbitrary order of words in sentence, the following specific issues have to be taken into account:

- stemming or lemmatization,
- multi-word expressions and named entities,
- synonyms and homonyms.

The main problem specific to the Slovak language is identification of indexing terms in a text document. In the first step it is necessary to perform morphological analysis to identify the original basic morphological form. Morphological analysis of the Slovak language, dealing with unsupervised identification of word suffix and identification of morphological form using hidden Markov model was proposed in paper (Hládek et al., 2015). Similar approach can be taken to identify stem of a word or word lemma according to context. Lemma identification was used in IR system for Czech (Straková and Pecina, 2010). The other approach to identify stem of a word is to use rule-based system (e.g. Hunspell[1]), as it was presented in (Wilhelm-Stein et al., 2013).

Our previous research in the field of multi-word expressions in Slovak is presented in (Staš et al., 2013).

4. The Document Set

The previously submitted work - Slovak Categorized News Corpus (Hládek et al., 2014) has been selected as the document set. Only minor adjustments have been made. Some duplicate documents have been removed. Automatic morphological and named entity annotations from the previous

[1] https://hunspell.github.io/

```
i 877 pravda : Dvanáste víťazstvo obhajcu titulu Šport 27.12.2003 14:00

Jeden z najlepších rozohrávačov v súťaži dosiahol za 24 bodov , 12 asistencií
a 11 doskokov piaty triple-double v sezóne ( 55. v kariére ) a navyše v záverečných
73 sekundách premenil šesť trestných hodov za sebou .

Obhajcovia titulu zo San Antonia sú nezastaviteľní , keď ďalšou obeťou
Spurs bolo tentoraz Orlando .

k 12. víťazstvu v sérii doviedol domácich tradične Tim Duncan , ktorý k
27 bodom pridal aj 16 doskokov a päť blokov .

Zaujímavý záver mal zápas v Utahu .
--endtext
```

Figure 1: Example document in the document set

version of the database were removed (because the documents remain the same, these annotations are still accessible from the first version of the database[2]). The whole database is now stored in a single structured file.
Items in the header are separated by tab character. Each document consists of document heading with the following document meta-information:

- unique document ID;
- author name;
- document title;
- document category;
- publication date.

The document body consists of one sentence per line and end of the document marker. Example of a document with meta-information header is depicted in Fig. 1.

5. Document Set Preparation

Process of the document set preparation is described in (Hládek et al., 2014). The document set contains 3980 text documents. The document set preparation can be briefly summarized as sequence of the following steps.

1. A custom web-crawling agent has been used to gather a set of raw HTML pages.
2. These pages have been parsed to extract raw text and meta-data information.
3. The tokenizer tool[3] has been used for word and sentence boundary detection in raw text.
4. Finally, documents were analyzed for possible multiple occurrence using control sums of paragraphs in text and calculation of unique paragraph ratio. Documents with low ratio of unique paragraphs were removed from the database.

Documents are sorted into 6 categories:

1. Economy and Business;
2. Culture;
3. Sport;
4. Domestic News;
5. World News;
6. Health Care.

6. The Query and Result Set

After the document was prepared it is possible to write a set of queries to the database. One query correspond to one *information need* (Cole, 2011) and is written in natural language, as it would be produced by a person doing search in the database of newspaper articles. Information need is seen as request for information written in natural language, describing the needed information in detail.

Typical issues in the Slovak natural language processing tasks are free word order, rich morphology and insufficient language resources (Hládek et al., 2014). The proposed information needs are written in a way that improvement in language specific preprocessing tasks will improve overall precision-recall values (Korra et al., 2011).

It is necessary to ensure that at least one document relevant to information need exists in the document set. The first step of the query set construction is selection of a set of keywords. Each keyword has been reformulated as an information need. For each information need a keyword stem and alternative forms of a keywords were searched in the document set and each matching document was evaluated for relevancy with the corresponding information need. Result of this process of manual search is the set of information needs and ranked relevant documents.

[2] http://nlp.web.tuke.sk/categorizednews
[3] http://nlp.web.tuke.sk/tokenizer/

Relevancy	Id	Number of documents
Complete answer	1	54
Highly relevant answer	2	764
Useful answer	3	163
Minimal or historic answer	4	116

Table 1: Answers Summary

Query	Document	Relevancy
79	27510	3
79	28825	3
79	30466	2
80	26932	2
80	30511	2
80	28522	3

Table 2: Result set example

Number of information needs	80
Number of relevant documents	1097
Size of the document set	3980
Average result set size	13,71

Table 3: Query and result set characteristics

The approach of the Cranfield test collection (Cleverdon, 1967) is strictly followed for relevancy annotation as it is written in the database documentation[4]:

> *The qrels are in three columns: the first is the query number, the second is the relevant document number, and the third is the relevancy code.*

The codes are defined by Cleverdon as follows (Cleverdon, 1967):

1. *References which are a complete answer to the question.*
2. *References of a high degree of relevance, the lack of which either would have made the research impracticable or would have resulted in a considerable amount of extra work.*
3. *References which were useful, either as general background to the work or as suggesting methods of tackling certain aspects of the work.*
4. *References of minimum interest, for example, those that have been included from an historical viewpoint.*
5. *References of no interest.*

The evaluation set consists of information need (a query on the database in natural language), set of relevant documents and document relevancy for each relevant document. Example information needs in the evaluation set is summarized in Tab. 5.

Each information need has a set of relevant documents. Document relevancy is expressed by integer number 1-4, where 1 is the most relevant and 4 is of minimum interest. Example of result set is displayed in Tab. 2. Rest of the documents in the document set is considered equally irrelevant. As it is described in (Manning et al., 2008), the minimal number of information needs that can be considered sufficient is 80. Answers present in the result set and their relevancy are summarized in Table 1.

[4]http://ir.dcs.gla.ac.uk/resources/test_collections/cran/

7. Conclusion

In the future work more manual annotations will be added to the document set such as named entities, document keywords and word chunk annotation making it even more useful for evaluating various natural language processing techniques. The proposed set is useful as Slovak language information retrieval evaluation task with or without result ranking, document clustering and categorization evaluation. Slovak language information retrieval research will support of development of existing speech and language technologies. The proposed evaluation corpus adds another language resource for Slovak that can be used for research in IR and can be interesting challenge for non-Slovak researchers. It supports development of multi-lingual systems by creation of another language evaluation set that can be used for IR evaluation.

8. Acknowledgements

The research presented in this paper was partially supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the project VEGA 1/0075/15 (50%) and the Research and Development Operational Programme funded by the ERDF under the project implementation University Science Park TECHNICOM for Innovation Applications supported by Knowledge Technology, ITMS code: 26220220182 (50%).

9. Bibliographical References

Cleverdon, C. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6):173–194.

Cole, C. (2011). A theory of information need for information retrieval that connects information to knowledge. *Journal of the American Society for Information Science and Technology*, 62(7):1216–1231.

Hládek, D., Staš, J., and Juhár, J. (2014). The Slovak Categorized News Corpus. *LREC 2014 - Ninth International Conference on Language Resources and Evaluation*, pages 1705–1708. WOS:000355611003050.

Hládek, D., Staš, J., and Juhár, J. (2015). Morphological analysis of the Slovak language. *Advances in Electrical and Electronic Engineering*, 13(4):289–294.

Korra, R., Sujatha, P., Chetana, S., and Kumar, M. (2011). Performance evaluation of Multilingual Information Retrieval (MLIR) system over Information Retrieval (IR) system. pages 722–727. DOI: 10.1109/ICRTIT.2011.5972453.

Question type	Question Subject	Count	Question Examples
Any information	Random	14	'Všetky informácie o '
Where	Place	23	'Kde na': 2, 'Ktoré krajiny': 2, 'V ktorých': 2, 'Kde všade': 2, 'Kde je': 2, 'Kde zvyčajne': 1, 'Kde sa': 7, 'Ktorá krajina': 1, 'Kde žijú': 1, 'Kam utekajú': 1, 'Kde bol': 1, 'V ktorej': 1,
What, How	Type, Way, Method	19	'Aké sú': 6, 'Ako prebieha': 3, 'Aká je': 3, 'Aké boli': 2, 'Ako prebiehal': 1, 'Ako vykonávali': 1, 'Ako sa': 1, 'Aké problémy': 1,
Who, What	Person, Organization	18	'Kto je': 5, 'Kto každý': 2, 'Kto vlastní': 1, 'Ktorí politici': 1, 'Komu bola': 1, 'Kto sú': 1, 'Kto najviac': 1, 'S kým': 1, 'Ktorí športovci': 1, 'Kto hral': 1, 'Kto sa': 1, 'Kto bol': 1 ,'Aké zdravotné': 1,
How big	Number, dimension	4	'Aký veľký': 1, 'Výsledky v': 1, 'Aké je': 1, 'Koľko lekárov': 1,
When	Time	3	'V ktorom': 1,'Kedy bolo': 1, 'Kedy sa': 1,

Table 4: Information need types in the database

Information Need	Translation
Kto je Robert Fico	Who is Robert Fico
Kto hral na Australian Open	Who played at Australian Open
Aké je HDP Slovenskej republiky	How big is GDP (Gross Domestic Product) of the Slovak Republic
Kde sa nachádza slovenský závod Volkswagen	Where is the Slovak Volkswagen factory

Table 5: Information need examples

Lazarinis, F., Vilares, J., Tait, J., and Efthimiadis, E. (2009). Current research issues and trends in non-English Web searching. *Information Retrieval*, 12(3):230–250.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, July.

Nunzio, G. M. D., Ferro, N., Mandl, T., and Peters, C. (2007). CLEF 2007: Ad Hoc Track Overview. In Carol Peters, et al., editors, *Advances in Multilingual and Multimodal Information Retrieval*, number 5152 in Lecture Notes in Computer Science, pages 13–32. Springer Berlin Heidelberg, September. DOI: 10.1007/978-3-540-85760-0_2.

Peters, C., Braschler, M., and Clough, P. (2012). *Multilingual information retrieval: From research to practice*. Multilingual Information Retrieval: From Research to Practice. DOI: 10.1007/978-3-642-23008-0.

Simpson, M. S., Voorhees, E. M., and Hersh, W. (2014). Overview of the TREC 2014 Clinical Decision Support Track. Technical report, November.

Staš, J., Hládek, D., Juhár, J., and Ološtiak, M. (2013). Automatic extraction of multiword units from Slovak text corpora. *Conference Proc. of 7th International Conference on NLP, Corpus Linguistics, E-Learning, SLOVKO 2013, At Bratislava, Slovakia*.

Straková, J. and Pecina, P. (2010). Czech information retrieval with syntax-based language models. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Suominen, H., Schreck, T., Leroy, G., Hochheiser, H., Goeuriot, L., Kelly, L., Mowery, D. L., Nualart, J., Ferraro, G., and Keim, D. (2014). Task 1 of the CLEF eHealth Evaluation Lab 2014 : Visual-Interactive Search and Exploration of eHealth Data. pages 1–30.

Wilhelm-Stein, T., Schürer, B., and Eibl, M. (2013). Identifying the most suitable stemmer for the CHiC multilingual ad-hoc task. *CEUR Workshop Proceedings*, 1179.