Hluchý, L., Bieliková, M., Paralič, J. (Eds.) : WIKT 2014, pp. 18–21

Towards Understanding Multilingual Search Query Intent

Michal Laclavík[1,2], Marek Ciglan[2], Štefan Dlugolinský[2]
Sam Steingold[1], and Alex Dorman[1]

[1] Magnetic Media Online, New York, USA,
[2] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia

Abstract. In this paper we describe an Entity Search and Query Understanding experiment on Slovak Language. In our previous work we participated in the ERD Challenge focusing on recognizing entities in search queries, here we try to extend this approach to other languages, while experimenting with Slovak Language. Wikipedia is used as Knowledge Base providing entities such as people, places or locations to be recognized in and answered for user search queries.

Keywords: entity search, Slovak language, query understanding, Wikipedia

1 Introduction

In this paper we describe the Slovak language extension of our contribution [1] to Entity Search by participating at *2014 Entity Recognition and Disambiguation Challenge*[3] [2]. We have participated in the *Short Track* of the challenge, which focused on recognizing mentions of entities in a search queries, disambiguating them, and mapping them to the entities in a given knowledge base - subset of Freebase[4] containing of more than 2 million of entities.

For Slovak language, we have taken the Slovak Wikipedia containing more than 190,000 articles (concepts or entities) as a knowledge base for entity recognition and disambiguation and we discuss our result in applying our original ERD approach on this Slovak data.

In the ERD, our system was evaluated as the 4^{th} best with F1 score of 65.57%[5]. We believe that our system has some unique features. In our research [3], we try to address Entity Search, Query Understanding or Question Answering problems by combing efforts from typical information retrieval models, semantic web, information extraction and complex networks.

We are developing Entity Search related applications like Query Categorization or Enterprise Search, where some of these efforts have been developed and tested. Participation in the ERD challenge [2] helped us enhance and introduce new techniques, where we combine approaches from these fields as described in

[3] http://web-ngram.research.microsoft.com/erd2014/
[4] http://www.freebase.com/
[5] http://tinyurl.com/ShortTrackERD14

the ERD paper[1] and extended here to Slovak and, potentially, to any language where sufficient Wikipedia data is available.

Motivation for Magnetic (Magnetic Media Online[6]) and IISAS (Institute of Informatics, Slovak academy of Sciences[7]) to participate in the ERD Challenge comes from our effort to build a scalable Query categorization (QC) system based on Wikipedia corpus instead of the entire web. Magnetic needs to understand query intend to perform well the business of Search Retargeting, a form of targeted online advertising. In the domain of search retargeting, the audiences are modeled based on the search queries users conduct on the websites they visited. Search retargeting focuses on displaying advertisements to users who conducted searches for specific keywords or categories in the past. For this domain, QC is the essential technique for user modeling and better user targeting. Magnetic tries to address foreign languages, thus support for multilingual query categorization is essential. Entity search, which was focus of the ERD challenge [2] and also this paper is the first step in our QC approach.

In addition to QC, IISAS motivation comes also from the VENIS project[8], where we tried to solve Enterprise Search by incorporating both structured data (database items) and unstructured data (emails, documents) to model entities in an enterprise - especially in SMEs. We would like to address also Multilingualism in our Enterprise Search.

The main contributions of this paper in addition to the ERD approach[1] applied on English are the following:

- Selecting and annotating 100 user queries as an evaluation dataset.
- Creating two version of Slovak Wikipedia index with and without special diacritic characters.
- Evaluating search query results on both indexes.

2 Experimenting on Slovak Wikipedia

In the past we have already experimented with Slovak Wikipedia[4], we have parsed the data and tried to use it for search or Named Entity Recognition. Now we used the same approach as for ERD, but we have created an index from Slovak Wikipedia. In our ERD solution we did not use any special language-dependent NLP to disambiguate entities or map them on search queries. So here we tested if this approach can really be applied to such languages as Slovak.

Soon after the first tests we discovered the following problems: not enough alternative names for entities; user queries with and without special characters.

We did not tackle the first problem by any means, but we can add additional alternative names from Wikipage infoboxes for example. Concerning high number of queries without the special diacritic characters, we have created two indexes with and without diacritics characters and experimented with them.

[6] http://www.magnetic.com/
[7] http://ikt.ui.sav.sk/
[8] http://www.venis-project.eu/

We have selected 100 queries from Magnetic search data[9], which are available online for future research. We have filtered out queries containing profanity or sexually explicit content. We have manually annotated these queries with concepts from Slovak Wikipedia. Slovak Wikipedia does not contain all desired information compared to the English one. For example, concepts like some popular TV series or Social Security office wikipage were missing, which somewhat limits the returned search results and can have an impact on lower coverage of entity search or query categorization.

We have also found out that 35% of queries (35/101) were missing special characters. This means that building an index without special characters is important, otherwise 35% of queries would stay unanswered. On the other hand this can bring some decrease of precision, where "dieta" would be same basic form for "diéta" (diet) as well as "dieťa" (child). However, results where special characters were removed were much better as described in next section.

3 Evaluation

In this section we discuss results on a sample of 100 Slovak queries - the data with annotations which we have prepared for this paper.

In the ERD Challenge [2], the evaluation focused solely on F1, because it was easier to identify borderline cases with zero retrieved or annotated entities for a query. However, we wanted to get an idea about Precision and Recall while developing the system, so we have calculated Macro Precision and Macro Recall, where there was no problem with borderline cases. In addition to these, we have also calculated Macro F1 and two types of Micro F1. Micro F1 calculated in the same way as defined by the ERD organizers, which we refer to as *Micro F1 Set* and *Micro F1*, which considered each returned entity as correct or incorrect independently of the defined interpretation sets. We have applied the same technique to the Slovak dataset evaluation. For more details see also ERD paper [1] and ERD guidelines[10].

Additionally, we computed the novel information-theoretic Proficiency metric [5], which measures the share of information content of the annotated dataset captured by the categorization. Its value is 1 for the perfect categorization and 0 for a categorization which is independent (in the sense of Probability Theory) from the annotation.

As one can see, there is always a few percent gap between Micro F1 and Micro F1 set. On our Slovak dataset it is even broader than on English one. The results for all applied measures are summarized in Table 1. In the Table 1, we list evaluations for Slovak query dataset with two different indexes - with special characters "SK" row and without special characters "SK ASCII" row, where all special characters were converted to its ASCII equivalent. We can see that improvement with ASCII index is very significant. Nevertheless, while on SK dataset we have achieved only about 47% F1 set, on the English TREC dataset

[9] http://ikt.ui.sav.sk/research/ERD/

[10] http://web-ngram.research.microsoft.com/erd2014/Docs/Detail%20Rules.pdf

Table 1. Results on Slovak queries dataset compared with beta TREC data[1] and ERD results[1]. New Proficiency metric is also reported.

	Macro Precision	Macro Recall	Macro F1	Micro F1	Micro F1 Set	Proficiency
SK ASCII	0.6067	0.5094	0.5538	0.5871	0.4686	0.4818
SK	0.4646	0.4340	0.4488	0.5005	0.3660	0.4333
EN TREC	0.7222	0.7761	0.7482	0.7968	0.7674	0.7650
EN ERD	-	-	-	-	0.6557	-

(see [1]) we have achieved a higher F1 set of 77% or 66% on ERD evaluation. ERD results (the "ERD" row) are available only for the F1 set since this results were evaluated by ERD organizers and they provided only one measure - F1.

4 Conclusion

In this paper we have shown that our Entity Search approach [1] used for the ERD challenge [2] can be applied also to other languages. The results are still worse than on English, but can be improved. Slovak Wikipedia is also one of the smaller Wikipedias and it is likely that better results can be achieved on Wikipedias with more then 1 million of pages.

In the future we would like to enhance our approach with new sources of alternative names, and also applying this approach on other European languages.

Acknowledgments. This work is supported by Magnetic, and also by project VENIS FP7-284984, VEGA 2/0185/13 and CLAN APVV-0809-11.

References

1. Michal Laclavik, Marek Ciglan, Alex Dorman, Stefan Dlugolinsky, Sam Steingold, and Martin Seleng. 2014. A search based approach to entity recognition: magnetic and IISAS team at ERD challenge. In Proceedings of the first international workshop on Entity recognition & disambiguation (ERD '14). ACM, New York, NY, USA, 63–68. DOI=10.1145/2633211.2634352
2. David Carmel, Ming-Wei Chang, Evgeniy Gabrilovich, Bo-June (Paul) Hsu, Kuansan Wang. 2014. ERD 2014: Entity Recognition and Disambiguation Challenge. SIGIR Forum, 2014 (forthcoming), ACM.
3. Michal Laclavík, Marek Ciglan. 2013. Towards entity search: Research roadmap. In WIKT 2013 proceedings, 2013, p. 161-166. ISBN 978-80-8143-128-9.
4. Michal Laclavík, Štefan Dlugolinský, Michal Blanárik. 2013. Experimenting with Slovak Wikipedia as a Source for Language Technologies. In Proceedings of SLOVKO 2013, pages 160-165, 2013,
5. Sam Steingold, Michal Laclavík. 2014. An Information Theoretic Metric for Multi-Class Categorization. In preparation[11].

[11] `https://github.com/Magnetic/proficiency-metric`