Experimenting with Slovak Wikipedia as a Source for Language Technologies

Michal Laclavík[1], Štefan Dlugolinský[1], and Michal Blanárik[2]

[1] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Faculty of Informatics and Information Technologies, Slovak University of Technology, Bratislava, Slovakia

Abstract. In this paper we discuss how Slovak Wikipedia can be used for Natural Language Processing tasks on Slovak texts. We briefly discuss similar work done in past and also provide several experiments on Slovak Wikipedia related to Entity Search or Named Entity Recognition. In the paper we also try to motivate future research on Slovak Wikipedia, since valuable data can be gathered for building and improving Language Technologies.

1 Introduction

Wikipedia is a well known source of human knowledge created and maintained by crowd. It contains a variety of human maintained information on many topics of knowledge as well as facts, relations on entities such as people, organizations or locations. English Wikipedia was used in many ways [1] to create NLP tools including "wikifiers" such as Wikipedia Miner[1], Illinois Wikifier [2] or DBPedia Spotlight[2] [3]. Tools like these identify Wikipedia entities in a text. They are based on Wikipedia downloadable archives[3] or DBPedia[4], which contains structured information in a from of RDF graphs. Slovak Wikipedia was not explored much so far for NLP tasks, however some early work exists on this topic.

In this paper we discuss two experiments focused on Slovak Wikipedia parsing, entity search, as well as named entity recognition. Experiments are provided to show the potential of the Wikipedia as a text corpus with additional information rather then showing results of ready to use NLP tools.

Wikipedia is not composed only of articles but it also includes links representing relations among entities mentioned in the articles. Links contain anchor texts representing alternative names (inflected forms, abbreviations) or properties of addressed entities, which can be used in NLP tasks. In addition, DBPedia includes structured information now available in 111 languages including Slovak. Tools such as DBPedia Spotlight can be built [4] for multiple languages including Slovak with limited number of NLP tools.

2 Wikipedia Parsing, Indexing and Search

In this chapter, we describe our first experiment, which was aimed on extraction of Wikipedia links and their anchor texts. We have used a dump of the Slovak Wikipedia from the 30th April 2013. It contained 310,571 articles (including redirects to other articles).

[1] `http://wikipedia-miner.cms.waikato.ac.nz`
[2] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki`
[3] `http://en.wikipedia.org/wiki/Wikipedia:Database_download`
[4] `http://dbpedia.org`

The processing of the articles was performed by a MapReduce application on a Hadoop cluster in two stages: parsing and indexing. The size of Slovak Wikipedia is manageable on single machine, but we have used MapReduce approach since ready made tools for text processing are available. We have used and customized such tools for Wikipedia parsing and indexing.

The parsing stage has been performed within a map phase of the MapReduce application. Articles from the dump file have been processed one by one by multiple mappers. There have been links pointing to other articles (i.e. outlinks) extracted from each parsed article together with anchor texts and additional outlink data. Additional outlink data included:

- title of an article on which the outlink was pointing (extracted from URL)
- title of an article which contained the outlink
- title of the outlink
- path in article structure, where the out-link was found (built up of titles and subtitles of article sections).

If there was a redirect processed instead of a standard Wikipedia article, we have treated it as an outlink and its title became the outlink's anchor text. There have been totally 4,212,467 outlinks extracted in the parsing stage and emitted by mappers to one reducer in the second (indexing) stage.

Emitted outlinks have been combined by their URLs (URLs of Wikipedia articles) and converted to inlinks, so we got for each referenced Wikipedia article all its inlinks. We have further processed only those articles, which belong to the main Wikipedia namespace (all encyclopedia articles, lists, disambiguation pages, and encyclopedia redirects) and indexed them using Solr[5].

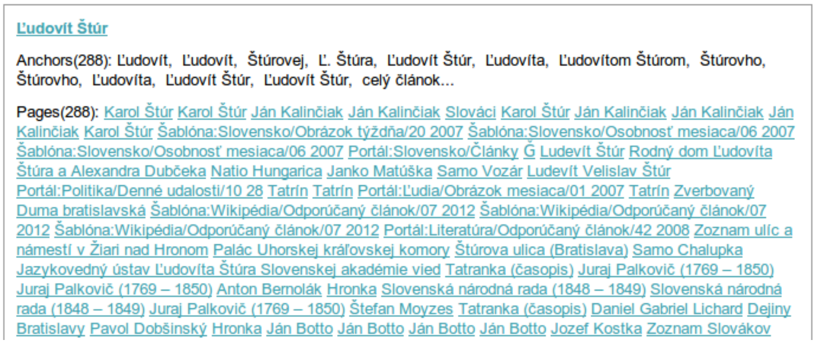
There have been 696,874 articles with 3,977,843 inlinks indexed. The number of indexed articles compared to the number of parsed articles was higher because there were many non-existing articles referenced in the Slovak Wikipedia, which is about 55%. The average number of inlinks per article was 5.71. If we consider only those articles, which were referenced more than one time, then there have been in average 13.60 inlinks per article and median was 3 inlinks per article.

We have created a searchable corpus of Wikipedia entities (articles) represented by their titles, links from the other pages and anchor texts. User can query the corpus using Solr web interface[6] or REST web services.

In Figure 1, we can see a screenshot of results for "Ľudovít Štúr" article. We can see anchor texts as well as other pages (inlinks) referencing this article. In addition, other metadata is present in the corpus, which can serve as a resource for creating training sets for NLP methods such as lemmatization, stemming or named entity recognition. Similar possibilities are discussed in chapter 3.

[5] http://lucene.apache.org/solr/

[6] http://147.213.75.180:8080/stevo/skwikislovco/browse



Ľudovít Štúr

Anchors(288): Ľudovít, Ľudovít, Štúrovej, Ľ. Štúra, Ľudovít Štúr, Ľudovíta, Ľudovítom Štúrom, Štúrovho, Štúrovho, Ľudovíta, Ľudovít Štúr, Ľudovít Štúr, celý článok...

Pages(288): Karol Štúr Karol Štúr Ján Kalinčiak Ján Kalinčiak Slováci Karol Štúr Ján Kalinčiak Ján Kalinčiak Ján Kalinčiak Karol Štúr Šablóna:Slovensko/Obrázok týždňa/20 2007 Šablóna:Slovensko/Osobnosť mesiaca/06 2007 Šablóna:Slovensko/Osobnosť mesiaca/06 2007 Portál:Slovensko/Články Ğ Ludevít Štúr Rodný dom Ľudovíta Štúra a Alexandra Dubčeka Natio Hungarica Janko Matúška Samo Vozár Ludevít Velislav Štúr Portál:Politika/Denné udalosti/10 28 Tatrín Tatrín Portál:Ľudia/Obrázok mesiaca/01 2007 Tatrín Zverbovaný Duma bratislavská Šablóna:Wikipédia/Odporúčaný článok/07 2012 Šablóna:Wikipédia/Odporúčaný článok/07 2012 Šablóna:Wikipédia/Odporúčaný článok/07 2012 Portál:Literatúra/Odporúčaný článok/42 2008 Zoznam ulíc a námestí v Žiari nad Hronom Palác Uhorskej kráľovskej komory Štúrova ulica (Bratislava) Samo Chalupka Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied Tatranka (časopis) Juraj Palkovič (1769 – 1850) Juraj Palkovič (1769 – 1850) Anton Bernolák Hronka Slovenská národná rada (1848 – 1849) Slovenská národná rada (1848 – 1849) Juraj Palkovič (1769 – 1850) Štefan Moyzes Tatranka (časopis) Daniel Gabriel Lichard Dejiny Bratislavy Pavol Dobšinský Hronka Ján Botto Ján Botto Ján Botto Ján Botto Jozef Kostka Zoznam Slovákov

Fig. 1. Screenshot from Solr search interface

3 Named Entity Recognition

The second experiment was focused on named entity recognition. We have taken an XML dump of the Slovak Wikipedia from 21st February 2013 and focused on two types of named entities: person names and locations. We have exploited anchor texts of links, which referred to these kinds of entities and collected their inflected forms. More detailed information about links can be found in Table 1.

Links	9,935,074
Links with inflected forms	549,740

Table 1. Basic information about Slovak version of Wikipedia

3.1 Person and Location Extraction

We have used two extraction methods for person names in a basic form. The first one relied on a specific markup pattern typical only for mentioning person: "[[*person_name*]] (*[[*date_of_birth*]])", e.g. "[[Ľudovít Štúr]] (* [[1815]])". The second method was based on infobox information fields related to person only: date of birth and name. There have been 16,454 and 11,404 person names in a basic form extracted with the first method and the second method respectively. The total number of unique person names in a basic form was 22,511. The list of names was complemented by their inflected forms discovered in anchor texts of links. The final list of person names contained 42,500 names.

Extraction of location entities was similar to the second method used for person names. The only difference was that for locations, we have used information fields related to geographic coordinates instead of date of birth and name fields. There have been 37,121 location names in a basic form extracted and complemented by their inflected forms discovered in anchor texts of links. All together we have discovered 37,603 different location names.

3.2 Experiment and Evaluation

We have trained Named Entity Recognition using Apache OpenNLP toolkit. The model was trained on data obtained from Wikipedia. Person names and locations have been tagged for algorithm that performs training of the model.

Training file for recognition of person's names	
Number of sentences	184,602
Number of tagged names	91,915
Training file for recognition of location	
Number of sentences	40,579
Number of tagged locations	38,538

Table 2. Training set data

The models for recognition of person names and locations were trained with 500 iterations on training files made in advance. In Table 3 we summarize achieved results applied on training data. The models perform pretty well on these files, but also it has been shown on that parameter cutoff, which determines how many times a specific feature must occur to be added to the model, should be set on higher value for the person recognition than for location recognition.

	Cutoff	Precision	Recall	F-Measure
Evaluation of trained model for persons recognition	3	0.901	0.617	0.733
	5	0.896	0.839	0.867
Evaluation of trained model for locations recognition	3	0.998	0.995	0.997

Table 3. Testing NER on trained data

In order to evaluate trained models in real use, we had manually tagged person names in 10 sample articles and locations in other 3 sample articles, both from sme.sk web site. Person names recognition was evaluated on model trained with parameter cutoff set on 5, because this model performed better in previous test. In Table 4 we show the results that have been achieved by applying trained models on sample articles. The models recognized entities with high precision but with poor recall. This behavior is similar that can be achieved with gazetteer, because trained models mainly recognized entities, which were present in training data set, but recognition of new entities was quite poor.

	Precision	Recall	F-Measure
Evaluation of trained model for persons recognition	0.891	0.372	0.517
Evaluation of trained model for locations recognition	1.0	0.292	0.433

Table 4. NER performance

NER on Slovak text have several big challenges. The one of big challenges is having several inflected forms of same named entity. In order to identify entities correctly, we need to group identified inflected forms of named entities. To cope with this challenge we have analyzed suffixes of extracted named entities. This list is the outcome of analysis of difference between lemma (base form) and inflected form of names previously obtained from Wikipedia. In this list is only that suffix which occurred more than 0.2 percent of all the suffixes were detected.

Suffix	Number of occurrences	Suffix	Number of occurrences	Suffix	Number of occurrences
a	5,543	**ov**	130	**ea**	55
om	4,406	**ová**	127	**s**	54
ovi	1,323	**ova**	103	**e**	53
m	779	**ovho**	88	**ových**	52
ho	589	**mu**	76	**ovom**	44
ou	541	**ove**	71	**ému**	41
ej	325	**ovou**	62	**o**	41
ovej	192	**vi**	59	**ovo**	41
ého	189	**eho**	56	**i**	40
us	143	**ovu**	56	Other:	1,345

Table 5. List of most occurred suffixes in people names and locations

Merging inflected forms of named entities was based on Levenshtein distance between two words with addition of obtained list of suffixes. This algorithm performed pretty well in case of merging word with same lemma but it failed in case of change in lemma of word during process of inflection.

4 Conclusions

In this article, we have discussed possibilities of using Slovak Wikipedia for Natural Language Processing. We have conducted several experiments analysing Wikipedia for the task of named entity recognition or statistics on word suffixes of selected named entity types. Experiments do not provide ready to use solutions for named entity recognition, lemmatization, stemming or other NLP tasks, but show the use pattern of growing Wikipedia resource. Our intent was to show that Slovak Wikipedia can serve as decent source for Language Technology evaluation and training supporting various NLP tasks.

Acknowledgements

This work is supported by projects: VEGA 2/0185/13 and CLAN APVV-0809-11. This publication is also the result of the project implementation ITMS: 26240220072 supported by Operational Programme Research & Development funded by the ERDF.

References

[1] Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716-754, DOI=10.1016/j.ijhcs.2009.05.004.

[2] Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1375–1384, Association for Computational Linguistics, Portland, Oregon, URL: `http://dl.acm.org/citation.cfm?id=2002472.2002642`.

[3] Mendes, P. N., Jakob, M., García-Silva, A., Bizer, Ch. (2011). DBpedia spotlight: shedding light on the web of documents. In Ghidini, Ch., Ngonga Ngomo, A.-C., Lindstaedt, S. and Pellegrini, T., editors, *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11)*, pages 1–8, ACM, New York, Accessible at: DOI=10.1145/2063518.2063519, `http://doi.acm.org/10.1145/2063518.2063519`.

[4] DBPedia, DBPedia Spotlight Internationalization. (2013). URL: `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Internationalization`.