ITAT 2013: Workshops, Posters, and Tutorials, pp. 2–4
ISBN 978-1490952086, © 2013 J. Hlaváčová

Special domain data mining through DBpedia on the example of Biology

Jaroslava Hlaváčová

ÚFAL MFF UK
hlava@ufal.mff.cuni.cz

Abstract: Wikipedia is not only a large encyclopedia, but lately also a source of linguistic data for various applications. Individual language versions allow to get the parallel data in multiple languages. Inclusion of Wikipedia articles into categories can be used to filter the language data according to a domain.

In our project, we needed a large number of parallel data for training systems of machine translation in the field of biomedicine. One of the sources was Wikipedia. To select the data from the given domain we used the results of the DBpedia project, which extracts structured information from the Wikipedia articles and makes them available to users in RDF format.

In this paper we describe the process of data extraction and the problems that we had to deal with, because the open source project like Wikipedia, to which anyone can contribute, is not very reliable concerning consistency.

1 Introduction — machine translation within the Khresmoi project

Khresmoi[1] is the European project developing a multilingual multimodal search and access system for biomedical information and documents. There are 12 partners from 9 European countries. The languages involved are English, Czech, German and French. The Czech part is responsible for machine translations between English and one of the other languages.

Machine translation is processed by means of statistical methods. For achieving good results, big amounts of language data are needed. They are used especially for training the system and afterwards for evaluations. The whole process of machine translation is nicely described in Czech in [3].

There are two types of data needed for the statistical machine translation task:

- parallel data — the same text in two languages, aligned on the sentence level
- monolingual data — for creating language model that is needed for the correct sentence creation in the target language

Both types of data is necessary to collect and preprocess. There are sets of data already prepared for various purposes, but for every special task it is usually necessary to collect more data or special sort of data.

In our case it was the need for data from the special domain — namely biomedicine. In the following text we will call them in-domain data. Apart from existing in-domain databases and registers we decided to extract in-domain data from a large general source — Wikipedia, especially its superstructure DBpedia.

2 DBpedia as a source of linguistic data

DBpedia[2] [2, 1] is a large multi-lingual knowledge base of structured information extracted from Wikipedia articles. The data is stored in RDF format putting together different entities, categories, languages. The data in DBpedia are divided into two datasets:

- **Canonicalized** — data having an equivalent in English.
- **Localized** — data from non-English Wikipedias.

As English was a central target language, we used the canonicalized data sets for our experiments.

The DBpedia has its own ontology, which is however not complete and in its recent shape is not possible to use for our purpose, namely the biomedical domain. Nevertheless, there are files in DBpedia (skos_categories_XX.ttl, where XX stands for abbreviation of a language (en for English, cs for Czech, fr for French, de for German).) putting together names of articles and their Wikipedia categories. The relations between the categories use the SKOS[3] vocabulary, namely the link *skos:broader* indicating that one category is more general (broader) than the other. We used this relation for extracting chains of Wikipedia subcategories for all the languages mentioned above. As the top category, we used the category *Biology*, as it appeared that all the medical categories are transitively subcategories of the category *Biology*.

3 Wikipedia categories and their relations

The categories are assigned to Wikipedia articles by their authors. Thus, the assignments are to a considerable extent subjective which has the troublesome consequence: the system of Wikipedia subcategories is not properly ordered. There are cycles, which means that one category

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

[2] http://dbpedia.org/About

[3] http://www.w3.org/2004/02/skos

might be transitively subcategory of itself. We present an example from the Czech category with the name *Endemité* (*Endemic*). There are two paths from the top category *Biology* leading to that category. The number at the beginning of each path represents number of levels from the top category:

44 Biologie Život Evoluce Strom_života Eukaryota Opisthokonta Živočichové Strunatci Obratlovci Čtyřnožci Synapsida Savci Placentálové Primáti Hominidé Člověk Lidé Profese Inženýrství Teorie_systémů Ekonomie Ekonomika Služby Zdravotnictví Lékařství Lékařské_obory Biomedicínské_inženýrství Lékařská_diagnostika Klinické_příznaky Psychologické_jevy Psychické_procesy Myšlení Abstraktní_vztahy Systematika Systémy Sluneční_soustava Planety_sluneční_soustavy Země Vědy_o_Zemi Geografie Geografické_disciplíny Fyzická_geografie Biogeografie Endemité

2 Biologie Endemité

In French, there is only one path of the length 7 leading to the category *Endémique*, English category Endemism (as the Wikipedia counterpart of the Czech category name) is not a subcategory of *Biology*, German category of that name does not exist. From this small example, we can get an idea of the extent of the inconsistency within Wikipedia categories.

There are even the cycles leading to the top level category *Biologie* in Czech and French (but not in English and German). They have the same length, but it is only an accident, as we can directly see from the paths — the individual levels do not correspond between the languages:

Czech (36) *Biologie* Život Evoluce Strom_života Eukaryota Opisthokonta Živočichové Strunatci Obratlovci Čtyřnožci Synapsida Savci Placentálové Primáti Hominidé Člověk Lidé Profese Inženýrství Teorie_systémů Ekonomie Ekonomika Služby Zdravotnictví Lékařství Lékařské_obory Biomedicínské_inženýrství Lékařská_diagnostika Klinické_příznaky Psychologické_jevy Psychické_procesy Myšlení Znalosti Věda Přírodní_vědy *Biologie*

French (36) *Biologie* Discipline_de_la_biologie Zoologie Animal Phylogénie_des_animaux Vertebrata Gnathostome Tétrapode Mammalia Eutheria Epitheria Boreoeutheria Euarchontoglires Euarchonta Primate Haplorrhini Simiiforme Catarrhini Hominoidea Hominidé Homininae Hominini Humain Sciences_humaines_et_sociales Économie Branche_de_l'économie Économie_publique Administration_publique Service_public Travail_social Éducation Association_ou_organisme_lié_à_l'éducation Académie Discipline_académique Sciences_naturelles *Biologie*

We present some more statistics about the cycles in the category systems of individual languages — see table 1. In all the languages except English, the shortest cycles are only 2 levels long, similarly as in the previous example with Czech *Endemité*. In English, the shortest cycles have 8 levels.

We can see that the ratio of cycles to all biological subcategories is very high. It suggests that almost one half of categories may be reached via more than one path from the top category of *Biology*. Only German has significantly less cycles. We might only guess the reason, there might be better checking team for the German Wikipedia.

The longest paths are usually cycles, but it is not always so. For instance in German, there are paths of the length 16, that are not cycles. Also in Czech, among the 5 longest paths, there are only four cycles. The fifth path leads to a category unambiguously.

The examples demonstrate that there is not possible to use the category structure for parallel mapping between the languages. Every languages has its own category system, they are not related. It even happens that articles with the same meaning are incorporated in different categories for different languages.

Table 1: Number of cycles in Wikipedia categories for individual languages. (Cycles means number of cycles, All subcat. is number of transitive subcategories of *Biology*, the column Longest presents the length of the longest cycle.)

	Cycles	All subcat.	Ratio	Longest
Czech	56 061	113 376	49,45%	54
English	374 357	782 325	47,85%	166
French	170 000	344 359	49,37%	62
German	1 219	6 186	19,71%	12

To avoid cycles during the processing the data is not difficult. The more problematic is the scope of the transitive subcategories. Table 2 shows that the subcategories cover almost all the Wikipedia categories, especially in case of Czech and French. The German Wikipedia again appears to be mantained more carefully.

Table 2: Ratio of in-domain categories to all the categories for different languages.

	All categories	In-domain subcateg.	Ratio in-domain/all
Czech	58 329	57 315	98,26%
English	865 900	407 968	47,11%
French	206 324	174 359	84,51%
German	144 876	4 967	3,43%

It was the reason why we tried to use the German in-domain categories as a basis. In DBpedia, there are files (interlanguage_links_same_as_XX.ttl) mapping names of

all titles, including categories, among all the languages, where such a mapping appears in Wikipedia. The relation sameAs is used to link pairs of titles between two languages. As the relation is symmetric and for our purposes, English is always one member of the language pair, we could use only the file for English (namely interlanguage_links_same_as_en.ttl).

Resulting number of categories in other languages is shown in table 3. The result is not satisfactory, the number of in-domain categories for other languages is about one third of the number of German ones, which seems to be too few. When we collected all the titles of Wikipedia articles from those categories, we missed a lot of relevant terms.

Table 3: In-domain categories based on Germann

		Ratio to German
Czech	1 174	0.24
English	1 981	0.40
French	1 496	0.30
German	4 967	1

The reason was simple — the system of subcategories does not match among the languages. Moreover, the same terms are often put into a different category in different languages. For instance the article *Plodová voda Amniotic fluid* belongs only to one Czech category *Těhotenství Human pregnancy*, which is not a category for the German Wikipedia. That is why this term did not appear in the result.

Our findings confirm the way how the Wikipedia is created and mantained. There is no (or not satisfactory) coordination among the languages involved.

4 Combination with other sources

We had to find another way how to extract the in-domain data from Wikipedia.

For every language, we used other DBpedia source files for selection all titles belonging to in-domain categories acquired through German in-domain categories. Then, we used files interlanguage_links_same_as_XX.ttl providing translations of all Wikipedia titles and made translations for all pairs among our four languages. It did not help much, there were still missing many useful terms.

We decided to take all the terms acquired so far, find all categories they belong to, and add all the rest titles from those categories. We got again into trouble with inconsistency of categories and had to adopt a limit of at least two terms in a category to be accepted as in-domain. Thus, we took titles of every category that contained at least two terms selected as in-domains in previous steps.

The last decision was to add data from external source, namely MeSH. MeSH is the abbreviation for Medical Subject Headings[4]. It is a vocabulary thesaurus mantained by the U.S. National Library of Medicine. It has translations into many languages and is used for indexing medical articles all over the world. We used the list of MeSH terms in all languages the same way as the last step described above; we tried to find all categories that included at least two MeSH terms. Then, we copied all the terms from those categories into the final lists.

The last step was building in-domain dictionaries with English. The final table 4 presents number of in-domain term pairs. We made a small manual evaluation of in-domainness for the Czech-English pair. We randomly selected 200 pairs and manually checked those belonging to the domain of biomedicine — they were only 14. However, we did not evaluate personal names that constitute almost 50% of the selection. A next evaluation should probably exclude the personal names. We will make a similar evaluation for other languages.

The result is not very impressive. Nevertheless, our selections present reasonably big and consistent in-domain dictionaries that can be used as a basis in further processing toward using in statistical machine translation.

Table 4: Sizes of final dictionaries

	Number of terms
Czech-English	69 598
French-English	379 830
German-English	310 203

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 257528 (KHRESMOI).

References

[1] Pablo N. Mendes, Max Jakob and Christian Bizer. DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2012, 21–27 May 2012, Istanbul, Turkey.

[2] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, (7):154– 165.

[3] Bojar Ondřej: Čeština a strojový překlad. Studies in Computational and Theoretical Linguistics. Praha, ÚFAL 2012

[4] http://www.nlm.nih.gov/mesh/