The importance of cross-lingual information for matching Wikipedia with the Cyc ontology. *

Aleksander Smywiński-Pohl[1,2] and Krzysztof Wróbel[1,2]

[1] Chair in Computational Linguistics, Jagiellonian University,
ul. Łojasiewicza 4, 30-348 Kraków, Poland,
http://www.klk.uj.edu.pl

[2] Faculty of Computer Science, Electronics and Telecommunications,
AGH University of Science and Technology,
al. Mickiewicza 30, Kraków, Poland,
http://www.dsp.agh.edu.pl

Abstract. In this paper we try to answer the question how cross-lingual evidence may improve matching between different classification schemas. We concentrate specifically on the task of mapping between Wikipedia categories and Cyc terms as well as the classification of Wikipedia articles to the Cyc taxonomy and show how this process may be improved by consuming the evidence that is available in different editions of Wikipedia. The results show that the performance of the mapping procedure may be improved from 0.6 to 4.9 percentage points, depending on the number of external Wikipedia editions and the given task.

Keywords: Ontology, ontology mapping, classification, multilingual data, Wikipedia, Cyc

1 Approach

To answer the question how the additional Wikipedia editions influence the performance of the mapping between Wikipedia and Cyc (cf. [2]) we have defined the following tasks: 1) mapping of the Wikipedia categories to Cyc terms; 2) classification of the Wikipedia articles to the Cyc ontology based on the first sentences. In each case the decision of selecting the corresponding Cyc term requires disambiguation of some English expressions against the Cyc ontology. This decision is based on the contextual data that are available for each Wikipedia article and category. Consulting of the supplementary Wikipedia editions extends the context available when making the decision and in general should improve the performance of the corresponding algorithms.

In case of the category mapping (based on the identification of plural head nouns in category names cf. [3]), when an English category is mapped, the corresponding Dutch, German, etc. categories are inspected. Then the parent and

* This work was partly supported by Stuctured Dynamics LLC, partly by the Polish National Center for Research and Development under LIDER/37/69/L-3/11/NCBR/2012 grant and partly by the Faculty of Management and Social Communication, Jagiellonian University in Krakow.

child categories as well as articles of the corresponding categories in the other editions are looked up in a reverse interlingual mapping index and if there is an English Wikipedia page, that was not present in the original context, it is included in the new, extended context. Then a support value used to disambiguate the category is computed against the extended context.

In case of article classification (based on the first sentence parsing, cf. [1]) the supplementary Wikipedia editions provide additional categories for the classified article, that are used to verify the disambiguation decision. The manner of operation is similar to that from the previous task – the corresponding articles in other Wikipedia editions are consulted, their categories are translated back to English and these new categories are included in the extended context.

2 Results

There was a small improvement (F1 increased from 86.8% to 78.4%) in the performance of the category mapping when the English Wikipedia is supported by three other Wikipedias (`de`,`nl`,`sv`). However providing the algorithm with more data from other Wikipedia editions, increased the computation time, but did not further improve the results.

On the other hand the influence of the additional Wikipedia editions in the task of the classification of the articles into the Cyc ontology was much stronger. Not only the additional Wikipedia editions improved the recall, but also the precision. The maximum precision was achieved for 5 and 6 additional Wikipedias (96.6% compared to 95.8% for the sole English Wikipedia). The F1 was the largest for the 8 additional Wikipedias resulting in an increase from 66.2% to 71.1%.

The overall conclusion from the results is that the influence of the supplementary Wikipedias is task dependant and in general the extra time necessary to pre-process the data and the increase of the computation time may not be justified. However the task of articles classification shows also that such supplementary data may be very valuable and may increase both the precision and the recall of the results.

References

1. Gangemi, A., Nuzzolese, A.G., Presutti, V., Draicchio, F., Musetti, A., Ciancarini, P.: Automatic typing of DBpedia entities. In: The Semantic Web–ISWC 2012, pp. 65–81. Springer (2012)
2. Pohl, A.: Classifying the Wikipedia Articles into the OpenCyc Taxonomy. In: Rizzo, G., Mendes, P., Charton, E., Hellmann, S., Kalyanpur, A. (eds.) Proceedings of the Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference. pp. 5–16 (2012)
3. Suchanek, F., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706. ACM (2007)