Automatic mapping of Wikipedia categories into OpenCyc types*

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

[1] Jagiellonian University, Faculty of Management and Social Communication
[2] AGH University of Science and Technology, Faculty of Computer Science, Electronics and Telecommunications

Abstract. The aim of the research presented in the article is the mapping between the English Wikipedia categories and OpenCyc types. The mapping algorithm is heuristic and it takes into account structural similarities between the categories and the corresponding types. The achieved mapping precision ranges from 82 to 92 % (depending on the evaluation scheme), recall from 67 to 76%. The results of the algorithm and its code are available at `http://cycloped.io`.

1 Approach

The aim of this research is automatic mapping of Wikipedia categories into OpenCyc [1] types. Although Wikipedia category system is hierarchical in nature, it is more like a thesaurus than a classification scheme [5], since it lacks any clear-defined hierarchical structure [4]. By mapping the categories into OpenCyc types we will be able to levarage the well defined structure of that ontology in Wikipedia-related information extraction tasks.

The automatic mapping of categories is divided into three stages. In the first stage the categories are pre-processed, in order to filter-out the uninteresting categories. In the second stage for each category a set of candidate mappings is generated and in the last stage disambiguation is performed by comparing the context of the category with the contexts of the candidate types. As such it is similar to the method employed in YAGO for mapping the categories into WordNet synsets [3].

The disambiguation is based on structural similarities between the OpenCyc ontology and Wikipedia category system treated as a taxonomy. The primary means for structuring Wikipedia is the *inclusion* relation that holds between categories and articles as well as categories themselves. In the first case, if the article represents an entity, the inclusion in a category might be approximated by *instantiation* relation, while in the second case the inclusion of category might be approximated by *specialization* relation. *Instantiation* and *specialization* are strictly defined in OpenCyc and are the primary means for structuring its contents. Checking if inclusion of articles and categories in the category that is being

* This work was supported by the Faculty of Management and Social Communication, Jagiellonian University in Krakow.

mapped has a corresponding instantiation and specialization assertions stated in OpenCyc provides evidence for validity of a given candidate mapping.

2 Results

Out of 616 thousand of categories with plural noun-heads we were able to assign some corresponding type to 484 thousand categories (78.6%). We have manually validated 600 mappings in order to assess the quality of the category mapping algorithm. We assumed that there is up to one valid OpenCyc type for each Wikipedia category. We have not assigned any type if the category was ambiguous or should be filtered out as administrative. In cases the algorithm assigned some types to such categories, they were treated as false positives. For the other categories we have either accepted the mapping provided by the algorithm or manually assigned the correct mapping in cases when the algorithm's decision was invalid.

We measured the performance of the algorithm using standard information retrieval measures of precision and recall, employing two evaluation scenarios. In the first one strict equivalence between the results obtained by the algorithm and the reference mapping was required and in the second, we have extended the set of true positives, by including results that were either specializations or generalizations of the terms defined in the reference set. In the first scenario we have obtained 82.5% precision, 67.5% recall and 74.2% F_1 and in the second we have obtained 92.9% precision, 76.1% recall and 83.6% F_1.

The results of the algorithm and the source code are available at `http://cyclo-ped.io`. We plan to extend the mapping and classification into other natural languages, as well as automatically extend the OpenCyc taxonomy. Although the results of the automatic mapping are worse than manually established correspondence from our past efforts [2], the achieved coverage is much better. Moreover the algorithms allow for providing new mappings when Wikipedia grows, making it very useful for converting it into computable knowledge base.

References

1. Lenat, D.B.: CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM 38(11), 33–38 (1995)
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: Williamson, C., Zurko, M.E., Patel-Schneider, P., Shenoy, P. (eds.) Proceedings of the 16th international conference on World Wide Web. pp. 697–706. ACM (2007)
4. Suchecki, K., Salah, A.A.A., Gao, C., Scharnhorst, A.: Evolution of Wikipedia's Category Structure. Advances in Complex Systems 15(supp01) (2012)
5. Voss, J.: Collaborative thesaurus tagging the Wikipedia way. arXiv preprint cs/0604036 (2006)