Wikipedia as the Premiere Source for Targeted Hypernym Discovery

Tomáš Kliegr[1], Vojtěch Svátek[1], Krishna Chandramouli[2], Jan Nemrava[1] and Ebroul Izquierdo[2]

[1] University of Economics, Prague, Department of Information and Knowledge Engineering, Winston Churchill sq. 4, Prague 3, 130 67, Czech Republic
[2] Queen Mary University of London, Multimedia and Vision Research Group, Mile End Road, London, E1 4NS, UK

Abstract. Targeted Hypernym Discovery (THD) applies lexico-syntactic (Hearst) patterns on a suitable corpus with the intent to extract one hypernym at a time. Using Wikipedia as the corpus in THD has recently yielded promising results in a number of tasks. We investigate the reasons that make Wikipedia articles such an easy target for lexicosyntactic patterns, and suggest that it is primarily the adherence of its contributors to Wikipedia's Manual of Style. We propose the hypothesis that extractable patterns are more likely to appear in articles covering popular topics, since these receive more attention including the adherence to the rules from the manual. However, two preliminary experiments carried out with 131 and 100 Wikipedia articles do not support this hypothesis.

1 Introduction

Most research in the field of hypernym discovery has been so far focused on non-statistical approaches, particularly on lexico-syntactic patterns (Hearst patterns) first introduced in [4]. Lexico-syntactic patterns were in the past primarily used on larger text corpora with the intent to discover all word-hypernym pairs in the collection. The extracted pairs were then used e.g. for taxonomy induction [10] or ontology learning [2]. This effort was, however, undermined by the low performance of lexico-syntactic patterns in the task of extracting *all* relations from a *generic* corpus. On this task, the state-of-the-art algorithm of Snow [9] achieves an F-measure of 36 %.

However, applying lexico-syntactic patterns on a *suitable document* with the intent to extract *one hypernym* at a time can achieve extraction accuracy close to 90% [6]. We refer to this approach as Targeted Hypernym Discovery (THD). Particularly, using THD in conjunction with Wikipedia has been found to bring promising results in a number of applications – ranging from named entity recognition [5] through query refinement [1] to image classification [6].

In this paper, we investigate the reasons why Wikipedia, particularly the first sentences of its articles, is such an easy target for lexicosyntactic patterns. Based on our prior experience with THD [1,6] we suggest the hypothesis that articles covering popular topics are more likely to contain an extractable hypernym (at the beginning of the text) than less popular ones.

Paper organization. Section 2 gives an overview of existing research; Section 3 discusses the role of Wikipedia in THD and proposes a hypothesis. The experimental setting and the experiments are described in Section 4. Conclusions presented in Section 5 summarize our findings and outline future work.

2 Related Research

Recently there has been a resurgence of interest in lexicosyntactic patterns [4], which can be in part attributed to the new possibilities given by the large amounts of textual content freely available on the web.

With respect to targeted hypernym discovery, [5] noticed that the Wikipedia encyclopedia can be used to improve the accuracy of Named Entity Recognition (NER) systems since it contains many articles defining named entities. For a given named entity extracted from text, the algorithm of [5] automatically found a Wikipedia entry for this entity and applied simple lexico-syntactic patterns to extract a hypernym from the first sentence of the article. The authors do not report on the number of correct and incorrect hypernyms and conclude that while they achieved an improvement on the NER task by using the extracted hypernyms as features of Conditional-Random-Fields (CRF) NER tagger, they believe that the hypernyms extracted from Wikipedia are too fine-grained for the classical NER task.[3]

In our recent work [1] we similarly tried to exploit hypernyms contained in Wikipedia for improving the performance of image retrieval through query refinement. In this research we tried to address some of the issues encountered in [5], particularly, we used a more sophisticated extraction grammar, relaxed the requirement for strict match between the article title and the named entity name by utilizing string similarity functions and increased the scope of extraction from the first sentence to the lead (introductory) section of the article. Preliminary results showed an improvement of image retrieval precision by 27%.

In [6] we performed additional experiments with an enhanced version of THD and introduced Semantic Concept Mapping, which exploits WordNet similarity measures in an effort to bridge the gap between the often too fine-grained hypernyms extracted from Wikipedia and a custom set of classes to which entities appearing in text need to be classified. The fact that the accuracy of hypernym discovery was 88% while the accuracy of mapping entities from text to 10 general Wordnet concepts was only 55% partly supports the conclusion of [5].

Apart from the free text, the (semi)structured information contained in Wikipedia articles—the infoboxes and the categories to which the article is assigned—are another prospective source of hypernyms. Further research will probably focus on fusing information retrieved from both the free text and the structured part of Wikipedia articles.

[3] I.e. classification to the PER, LOC, ORG and MISC categories.

3 Wikipedia as the Source for Hypernym Discovery

WordNet is usually considered to be a gold-standard dataset for training and testing hypernym discovery algorithms [9]. Its structured nature and general coverage make it a good choice for general disambiguation tasks. However, we noticed in our previous work [8] that WordNet is less useful for analysis of text with high proportion of named entities.

Various free-text corpora have been used to overcome the problem of WordNet sparsity. State-of-the-art approaches based on mining patterns from text already achieve a higher F-measure than WordNet (with human judgment as the ground truth). Interestingly, Wikipedia as a free and comprehensive source of information plays a vital role in these efforts; one of the best results [9] were achieved by extending the TREC corpus with articles from Wikipedia.

Inspired by the promising results of [9], we used Wikipedia as the sole source of knowledge in our Targeted Hypernym Discovery tool. Unlike standard hypernym discovery, THD only discovers hypernyms for the given 'hypernym query'. It first uses the Wikipedia search API to find the most relevant articles for the query and then calls the NLP components available in the GATE framework [3] to extract the hypernym from the first 'is a' pattern that appears in each article. Empirical results suggest that this approach may be quite effective.

Experiments carried out in [6] hint that targeted hypernym discovery could successfully extract hypernyms from more than 90% of Wikipedia articles describing named entities[4], if the extraction grammar used in [6], which performed at 88% only using variations of the verb *to be*, were extended to cover some less frequent lexical constructions. Only about 6% of articles do not contain a hypernym extractable by a lexicosyntactic pattern.

Upon a manual inspection of the results we observed that, with a few exceptions, the first match of an ideal Hearst pattern in the article provided an informative and specific hypernym. Indeed, we found the vast majority of Wikipedia articles to open with clear definitions following the pattern "XYZ is a ... *detailed hypernym.*" Exceptions included cases such as "XYZ is a *cross* between a A and B". In this case, the word *cross* matches the lexicosyntactic pattern "XYZ is a ?" but cannot be accepted as a useful hypernym for XYZ.

We consider this rigidity in opening sentences surprising. Such a clearness and uniformity of articulation could be expected from an expert-created encyclopedia or thesaurus but not from a resource collaboratively created by unpaid volunteers whose only training comes, in general, from reading the Wikipedia guidelines. The Wikipedia's *Manual of Style*[5] has, indeed, a special section on first sentences, which instructs authors "to put the article title as the subject of the first sentence". A special article on the lead (introductory) section[6] states that the first paragraph "needs to unambiguously define the topic for the reader".

[4] In the assessment of 129 articles obtained through the Wikipedia 'random article' link, 102 articles (79%) describe named entities.

[5] `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style#First_sentences`

[6] `http://en.wikipedia.org/wiki/Wikipedia:Lead_section`

Remarkably, Wikipedia contributors seem to follow these guidelines and even exceed them by refraining from the use of a more varied vocabulary when writing the opening definitions. For example, instead of "Diego Armando Maradona is a former Argentine football *player*"[7], the article title could start e.g. by "Diego Armando Maradona, a former Argentine football player, played in four World cups..." or even worse "D.A. Maradona was a *backbone* of Argentine football...".

However, when working on papers [1,6] we got the impression that hypernym discovery from shorter, less elaborate Wikipedia articles, which often describe uncommon entities, tends to be less successful than that from long articles on popular topics. If this empirical observation proved to be true, it would have implications on the predicted accuracy of extraction, depending on the kind of named entities expected.

In the rest of this paper, we try to empirically evaluate this hypothesis.

4 Experimental Setup

Two experiments were conducted in order to explore if there is a correlation between the popularity of Wikipedia article and the successfulness of hypernym discovery from this article. The experimental setup particularly included a) determining a measure of article popularity, b) determining the dataset and c) choosing a hypernym extraction algorithm. It should be noted that these choices were influenced by our previous work [1,6], which evaluated the usefulness of hypernyms discovered from Wikipedia for image retrieval.

Measure of Popularity. The popularity of Wikipedia articles can be measured in several ways. Experiment 1 takes the viewpoint of Wikipedia contributors and uses the number of inbound links from other Wikipidia articles. In turn, Experiment 2 takes the viewpoint of Wikipedia readers and uses the number of hits the article receives as the measure of popularity.

Datasets. The dataset and results[8] obtained in [1] were used as a basis of Experiment 1 presented in section 4.1. In [1] we used Wikipedia to discover possible hypernyms for a set of likely real-world queries from a specific domain. While multiple hypernyms from articles of varying popularity were extracted, in [1] we used background knowledge to filter out articles, which were not from the target domain. In contrast, Experiment 1 uses the full set of retrieved article-hypernym pairs.

The article-hypernym pairs extracted in our previous work [6] (which followed after [1]) form the dataset of Experiment 2. In [6] we established the name *Targeted Hypernym Discovery* for the task addressed by the algorithm introduced in [1] and used THD as a part of a larger framework, this time focused on entity classification. Nevertheless, we further evaluated our THD implementation on a

[7] Opening sentence of Wikipedia article on Maradona as of the time of writing.

[8] The results also included the relevancy of each article to the query, which is at the time of writing no longer available in Wikipedia's search results.

dataset consisting of 100 randomly selected Wikipedia articles. This experiment aimed at evaluating the hypernym extraction from the Wikipedia articles under the condition that an article defining the given entity is available. Article titles were used as queries for hypernym and the articles as the corpus. First sections (only these are processed by our THD algorithm) of the 100 articles contained approximately 100.000 words and 50.000 noun pairs.

Extraction Algorithm. The implementation of THD used in the experiments was built on top of the GATE NLP text engineering framework, which was used for shallow NLP parsing (particularly sentence splitting and part-of-speech tagging) of the document. The resulting tags were stored in annotations. These were further processed by a JAPE grammar engine. The extraction grammar matched several variation of the "is a" pattern, followed by a rather loosely defined sequence of unimportant words and finally by the desired hypernym. A detailed description of the THD implementation used in the experiments is presented in [1,6].

4.1 Experiment 1: Influence of Article Popularity (Links)

This experiment aimed to evaluate the influence of article popularity as measured by the number of inbound links from Wikipedia articles on the performance of THD. The underlying rationale is that the higher the number of linking articles, the higher the chance that other contributors would intervene if an article did not comply with the guidelines or its opening section was poorly/unclearly written.

The hypernym discovery implementation used [1] utilized Wikipedia's search interface to retrieve articles. Wikipedia MediaWiki search engine can use article popularity as measured by the number of articles that link to it as one of the ranking factors in addition to text-based relevance.[9] However, since we are only interested in article popularity, we try to mitigate the influence of text-based relevance by only involving articles whose title contains the entity for which the hypernym is sought, assuming that the part of relevance coming from the textual similarity between the article and the query is the same for all the relevant (returned) articles (up to a certain threshold). Manual inspection of the results showed that this technique was effective and the relevance measure generally reflected the relative popularity of the article subject in our dataset.

As test hypernym queries we used the surnames of ten top-rated NHL goalkeepers: Nabokov, Brodeur, Lundqvist, Luongo, Leclaire, Giguere, Miller, Bryzgalov, Turco and Kiprusoff. We already used this test set in our earlier work [1], where we found these words to provide enough ambiguity, as each represents a surname of several important persons from different fields, in addition to other meanings such as names of jobs, places or companies.

For each query, we downloaded the first section of all returned articles from Wikipedia up to a relevancy threshold of 50%. Redirects were followed but disambiguation articles and articles where the query term was not in the title were discarded. The resulting collection contained 131 documents (articles).

[9] http://www.mediawiki.org/wiki/Extension:Lucene-search

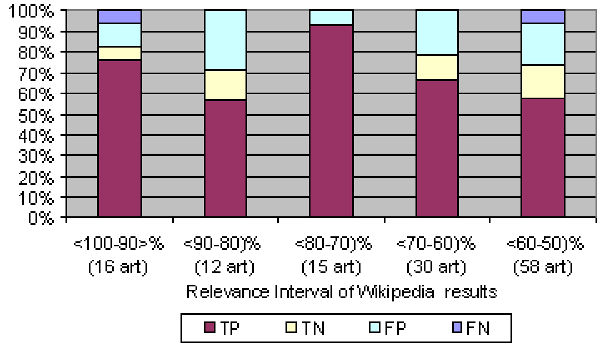

Fig. 1. Impact of article popularity on THD performance

Two human annotators annotated each of the documents. They were instructed to only mark the first hypernym per document (as does the used THD algorithm), regardless of how much more general this hypernym would be than the query. The annotation was not restricted to the context of ice hockey; hypernyms expressing all conceivable meanings of the original query were considered. The annotators agreed on 96% of annotations, which formed the ground truth.

With only five rules in our grammar we were able to discover 95 out of 124 hypernyms on which annotators agreed. Figure 1 shows the distribution of the THD outcome depending on article popularity. If the word marked as hypernym by the algorithm matched the ground truth then it was considered as a true positive (TP); false negative (FN) was the case when no hypernym was found but there was one according to the ground truth, and false positive (FP) in the opposite case. The case when both annotations were present but did not match was counted as two errors—FP as well as FN error—as this introduces a false hypernym and misses the true one.

We performed a statistical test to explore the significance of association between the article popularity (given by the respective relevance interval) and the successfulness of THD (1 for TP – correct hypernym extracted, 0 otherwise). The test used was one-sided Kendall's Tau B [7], which makes adjustment for ties and is also suitable for binary variables. A value of zero indicates the absence of association, while -1 or 1 mark perfect negative/positive association. Since the one-tailed Kendall's Tau B between the success of the extraction in these two groups is equal to 0.1281 (p-value of 0.055), we cannot reject the null hypothesis at a 5% significance level that there is not correspondence between article popularity and the successfulness of hypernym extraction.

It should be noted that the experimental results may be skewed by the narrow character of the dataset and by the residual influence of article-query relevance.

4.2 Experiment 2: Influence of Article Popularity (Hits)

Another way of measuring the popularity of an article is the number of hits (views) it receives from the general public. Again, since anyone can edit Wikipedia, the higher the number of hits, the higher the chance that a random user would intervene if the article was poorly written.

This experiment was carried out with the sample of 100 articles describing named entities, which were randomly selected using the Wikipedia's random article link in [6]. The ground truth was established in a similar manner as in Experiment 1, but hypernyms were extracted for the topic of the article and not for a query. Additionally, articles were only annotated by one annotator[10]. The system failed to extract the correct hypernym from 14 articles: a Hearst-like pattern was not present in 8 articles.[11]. In 6 cases a Hearst-like pattern was present but was not matched by the grammar.

We evaluated whether the inability of the system to extract a hypernym is dependent on the number of hits each of the 100 articles obtained during a one-month period.[12] The range of hits was between 1 (for *Kielpino Kartuskie*) and 25.253 (for *Dead Space (video game)*), with the median value being 237. The result of extraction was marked as either 1 (success) or 0 (all other cases). The different kinds of error were alone too rare to be tested separately.

The test used was the same as in Experiment 1 – Kendall's Tau B. The value of the Kendall's Tau B in our experiment was -0.037 (p-value of 0.320). This result hints that there is not a statistically significant correspondence between article hit-based popularity and the successfulness of hypernym extraction.

5 Conclusions and Further Work

In the paper we reviewed the recent work on Targeted Hypernym Discovery (THD) from Wikipedia and analyzed the reasons contributing to its success. Unlike existing approaches to hypernym discovery, THD selects a suitable document and extracts the most likely hypernym from it. The latter task is particularly interesting, since it seems that very good results are achieved by only considering the first hypernym matching the grammar from the article.

Our suggested explanation is that this can be partly attributed to the Wikipedia authoring guidelines. We conjectured that for articles covering less popular topics these guidelines are less rigidly applied, which may result in a worse performance of hypernym discovery algorithms.

[10] Exp. 1 showed that annotations by two humans do not significantly differ.

[11] Interestingly, the hypernym was a part of the article name in 6 cases, in 1 case there was no hypernym. The last case has interesting history. At the time of the preparation of the camera ready version of this paper, Wikipedia editors corrected the first sentence of this entry on R. E. Holz from "Richard E Holz, ... an American brass band composer,.." to now extractable "Richard E Holz was an American brass band composer..."

[12] during May 2008, using the http://stats.grok.se/en tool.

Our preliminary experimental results carried out altogether on 231 Wikipedia documents do not, however, support this hypothesis. Nevertheless, it should be noted that the value of the test in Experiment 1 was very close to the critical value for $\alpha = 5\%$. Since Experiment 1 was conducted on a sample from a specific domain and the article popularity was inferred from search relevance results, which might have introduced additional error, a larger scale experiment is thus indispensable to give the final answer.

6 Acknowledgements

The work has been partially supported by the European Commission under the IST research network of excellence K-SPACE of the 6th Framework programme, the work of Tomas Kliegr is also supported by grant GACR 201/08/0802 of the Czech Grant Agency.

References

1. Krishna Chandramouli, Tomáš Kliegr, Jan Nemrava, Vojtěch Svátek, and Ebroul Isquierdo. Query refinement and user relevance feedback for contextualized image retrieval. In *VIE 08: Proceedings of the 5th International Conference on Visual Information Engineering*. IET, 2008.
2. P. Cimiano and J. Voelker. Text2onto - a framework for ontology learning and data-driven change discovery. In *NLDB'05: Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems*, 2005.
3. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. A framework and graphical development environment for robust NLP tools and applications. In *ACL 2002*, 2002.
4. M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *14th International Conference on Computational Linguistics*, pages 539–545, 1992.
5. Jun'ichi Kazama and Kentaro Torisawa. Exploiting Wikipedia as external knowledge for named entity recognition. In *EMNLP-CoNLL'07: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707, 2007.
6. Tomáš Kliegr, Krishna Chandramouli, Jan Nemrava, Vojtěch Svátek, and Ebroul Izquierdo. Combining captions and visual analysis for image concept classification. In *MDM/KDD'08: Proceedings of the 9h International Workshop on Multimedia Data Mining*. ACM, 2008. To appear.
7. James E. De Muth. Basic statistics and pharmaceutical statistical applications, 2nd edn. *Journal Of The Royal Statistical Society Series A*, 1999.
8. Jan Nemrava. Refining search queries using wordnet glosses. In Helena Sofia Pinto and Martin Labsky, editors, *Poster and Demo Proceedings of EKAW 2006*, 2006.
9. R. Snow, D. Jurafsky, and A. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, number 17, pages 1297–1304, Cambridge, MA, 2005. MIT Press.
10. R. Snow, D. Jurafsky, and A. Ng. Semantic taxonomy induction from heterogenous evidence. In *COLING/ACL 06*, pages 801–808, Sydney, Australia, 2006. ACM.