Aligning WordNet Synsets and Wikipedia Articles

Samuel Fernando and **Mark Stevenson**
Department of Computer Science, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
{s.fernando, m.stevenson}@dcs.shef.ac.uk

Abstract

This paper examines the problem of finding articles in Wikipedia to match noun synsets in WordNet. The motivation is that these articles enrich the synsets with much more information than is already present in WordNet. Two methods are used. The first is title matching, following redirects and disambiguation links. The second is information retrieval over the set of articles. The methods are evaluated over a random sample set of 200 noun synsets which were manually annotated. With 10 candidate articles retrieved for each noun synset, the methods achieve recall of 93%. The manually annotated data set and the automatically generated candidate article sets are available online for research purposes.

Introduction

This paper presents methods for combining key information from two widely used knowledge bases. The first is WordNet (Fellbaum 1998) which contains synsets which encapsulate information about different senses of commonly used words. Specifically this paper focuses on the 82115 noun synsets. For example one synset contains the words 'car, auto, automobile, machine, motorcar'. Each noun synset contains a short descriptive gloss e.g. 'a motor vehicle with four wheels; usually propelled by an internal combustion engine'. Each synset is linked to other synsets by hypernymy (is-a) and meronymy (part-of) relations e.g. 'car' is-a 'vehicle', 'accelerator' part-of 'car' etc.

The second source of information is Wikipedia which contains 3.19 million encyclopedic articles on a huge range of subjects. Articles contain links to other articles within the text. Also present in Wikipedia is tag information such as categories which group articles under some common theme.

The aim here is to find for each synset in WordNet an article which is about the same concept as the synset, or decide that none exists (for the 'car' synset the match would be the 'Automobile' article). Aligning the two resources in this way allows both to be enriched. Each matched synset benefits from much more descriptive information, along with topical or thematic links to other articles (e.g. 'Passenger', 'Transport'). Each matched article is enriched with new relations to other synsets (and therefore articles) by hypernym and meronym relations.

WordNet has previously been used for many language processing tasks such as word sense disambiguation (Banerjee and Pedersen 2002). By enriching WordNet in this way it is hoped that performance in these tasks can also be improved.

This paper presents methods which find for each noun synset a small (but high recall) set of candidate articles which may be possible matches for the synset. This dramatically reduces the search space with only a small drop in recall. The methods were evaluated on a manually annotated set of 200 noun synsets. Analysis of this set shows that over 60% of noun synsets have a close matching article, providing a good source of information to enrich the synsets.

Background

There has been some previous work on linking Wikipedia articles with WordNet synsets (Ruiz-Casado, Alfonseca, and Castells 2005). This used text similarity metrics (cosine similarity with TF-IDF weighting) to find the best synsets for each article (the converse of the approach in this paper). Experiments were performed using the much smaller Simple English Wikipedia, achieving accuracy of 91%. Other work includes combining the two resources by linking synsets to Wikipedia categories (Ponzetto and Navigli 2009) and using IR approaches for enriching WordNet using the Web (Agirre et al. 2001).

Methods for retrieving candidate articles

Two methods were used to retrieve candidate articles from Wikipedia for each noun synset. With all experiments the Wikipedia snapshot of 28th November 2009 was used. The first was title matching using the words in the synset. The second was information retrieval using queries formed from words in the synset.

Title matching

Each noun synset S contains several synonymous words. For each word w within the synset, a search is carried out in Wikipedia. Three approaches are used to build the set of candidate articles C.

1) In the basic approach only pages A whose title matches w will be added to the candidate article set C. For exam-

ple the article 'Automobile' for the word 'automobile' in the synset.

2) Additionally to 1) we can search for redirects from w to some other page B, adding B to C. For example the word 'car' redirects to 'Automobile'.

3) Additionally to 1) and 2) we can add all disambiguation links where A or B are disambiguation pages. For example the 'Car' disambiguation page links to the 'Automobile' article, a movie and song with the title 'Cars' and many other pages.

Information retrieval

The Terrier IR system (Ounis et al. 2007) was used for information retrieval over the Wikipedia data set indexing each page as a document in the collection. Queries were formed using various combinations of the information in the synset. These include:

- Lemmas (L) e.g. (car, automobile etc.)
- Gloss (G) e.g (a motor vehicle with four wheels).
- Lemmas in related synsets (RL) e.g. (vehicle, accelerator etc.)

Stopwords were excluded and the remaining terms stemmed. Terrier offers a variety of weighting models for retrieval. The one used here was the widely used TF-IDF model.

Manually annotated data set

A random sample of 200 noun synsets was annotated for article matches (referred to as the 200NS set). Two annotators (both native English speakers) independently searched for matching articles and marked each synset with the correct article. Candidate articles retrieved using the methods from the previous section were presented to the annotators as suggestions but the annotators were free to look elsewhere in Wikipedia for articles if necessary. Analysis of the synsets showed sometimes that even if no exact match could be found sometimes the synset was described in part of an article or was closely related to an article, resulting in the categories listed here:

- Matches *article* - This indicates the article is a match for the synset, exclusively describing the same concept as the synset. For example the article 'Poaching is the process of gently simmering food in liquid' matches the synset 'poaching, cooking in simmering liquid'.
- Related to *article* - This indicates that no matching article could be found, but there is an article that is directly related to the synset. If more than one article meets this requirement, the most closely related is chosen. For example, the synset 'bath powder, a fine powder for spreading on the body' is marked as a hyponym of the article 'Powder, a dry, bulk solid composed of a large number of very fine particles'.
- Part-of *article* - The synset corresponds to part of the article, but not the whole. If more than one article meets this requirement, the most strongly related is chosen. An example is the synset 'tenon, a projection at the end of a piece of wood' which is described in part of the article about 'Mortise and tenon'.
- No related article found.

The initial inter-annotator agreement was 86%. The annotators discussed and resolved the disagreements to give the final 200NS data set. The distribution of categories for is shown in Table 1.

Category	Synsets
Match	126 (63%)
Related	36 (18%)
Part-of	11 (5.5%)
Not found	27 (13.5%)
Total	200 (100%)

Table 1: Grouping of synsets into different categories.

The majority of the synsets (63%) have a good matching article in Wikipedia. These provide a straightforward way of enriching the synsets. Many of the remaining synsets (27.5%) are directly related to some article in Wikipedia.

For the 13.5% of synsets where no article was found, there were two possible cases. Most (23/27) were where the synset defines a term we would expect to see defined in a dictionary, but not in an encylopedia. An example is the synset 'dumpiness, a short or stout physique'. This would not be an appropriate candidate for an encyclopedic article. The remainder (4/27) are synsets which would be something we might expect to find in Wikipedia, but cannot be found, for example 'vegetable sheep, cushion-forming New Zealand herb' is a New Zealand herb but no reference could be found in Wikipedia. There will be other relatively obscure terms like this in WordNet which are not yet included in Wikipedia.

Experiments

The question explored in this section is how good the recall of the candidate article retrieval process is i.e. where a related article exists for a synset how often does this appear within the candidate articles. Since we are only considering recall performance at this stage any synsets from the part-of, related, or matching categories will be considered a positive example since an article with some kind of relation to the synset was found.

From the 200 synsets, 173 were marked as related to the article in some way (whether an exact match, related, or part-of an article). The experiments here test what proportion of these matching articles were retrieved in the candidate article set using the following methods.

First are the set of methods using title matching. Here articles are retrieved which match any of the lemmas representing each synset.

- Title (T) only - Use only articles whose title matches one of the lemmas.
- Title (T), redirects (R) - As above but also follow any redirects.

- Title (T), redirects (R), all disambiguation links (D) - Add all links from disambiguation pages.

Next are the methods which use the Terrier IR system. The different queries are described below.

- Lemmas (L) only.
- Lemmas (L), gloss (G).
- Lemmas (L), all related lemmas (RL).
- Lemmas (L), gloss (G), related lemmas (RL).
- Lemmas (L), gloss (G), all related lemmas (RL), all glosses of related lemmas (RG).

Table 2 shows the recall for the title matching and information retrieval experiments with varying numbers of articles retrieved.

Articles	1	5	10	20
Title matching				
T	48.55	53.18	53.18	53.18
T,R	61.85	67.05	67.05	67.05
T,R,D	**63.01**	**71.1**	**71.68**	**72.25**
Information retrieval				
L	43.35	65.32	74.57	78.61
L,G	**55.49**	**80.35**	85.55	87.86
L,RL	41.62	69.36	78.03	83.82
L,G,RL	51.45	79.77	**86.13**	**88.44**
L,G,RL,RG	37.57	65.32	72.83	78.03
Title matching & IR Combined				
T,R,L,G	69.94	**91.33**	**93.06**	**93.06**
T,R,D,L,G	69.36	87.86	89.6	91.33
T,R,L,G,RL	**70.52**	90.17	92.49	**93.06**
T,R,D,L,G,RL	**70.52**	86.13	89.6	91.33

Table 2: Recall (%) against number of articles combining title matching & IR methods.

For the title matching methods, adding the redirects gave a clear boost to recall performance. Using the disambiguation links also improves performance slightly. With the IR methods using the lemma, gloss and related lemmas gives the best performance, slightly better than using lemma and gloss alone.

The final set of experiments combines retrieved articles from both the title matching and IR methods. The articles from the title method are used first followed by the articles from the IR method. The best performing title matching methods (T,R and T,R,D) are combined with the best IR methods (L,G) and (L,G,RL). The results from this show that using the IR articles gives a bigger boost to recall with fewer additional articles than using the disambiguation links. This is most likely due to the fact that the disambiguation links will not be necessarily ranked in order of similarity to the synset, which is the case with the IR articles. The results from the T,R,L,G and the T,R,L,G,RL are very close, converging to the same recall performance (93.06%) after 20 articles. However the T,R,L,G reaches this level quicker, after only 10 articles.

Summary and future work

This paper presented methods for finding a high recall set of candidate articles from Wikipedia matching noun synsets in WordNet. Evaluation was performed over a set of 200 noun synsets which had been manually annotated. The best results used a combination of title matching using redirects and information retrieval using the lemma words and gloss in each synset, achieving 93.06% recall with 10 candidate articles for each synset. This provides a useful resource for future work in the area, reducing the search space for selection methods from the set of 3.19 million articles to just 10 articles with only a small recall penalty.

Analysis of the manually annotated set shows that 63% of synsets had a clear matching article in Wikipedia, providing a potentially rich source of new information. A further 27.5% of the synsets had a closely related article. Most of the remaining 13.5% were terms we would expect to find in a dictionary but not an encyclopedia.

Previous work (Ruiz-Casado, Alfonseca, and Castells 2005) performed a similar task finding noun synsets to match articles in Simple English Wikipedia. This paper presented the converse of this task, over the much larger full English Wikipedia. The recall of 93.06% achieved here compares favourably with the results from this previous work although further methods are needed to select the best article from the candidate article set.

Both the manually annotated data set and the full set of noun synsets with 20 candidate articles retrieved by the best performing method (T+R+L+G) are available online `http://www.dcs.shef.ac.uk/~samf/resources.html`. Future work could extend the annotated data, possibly with crowd sourcing techniques such as Mechanical Turk.

References

Agirre, E.; Ansa, O.; Martinez, D.; and Hovy, E. 2001. Enriching WordNet concepts with topic signatures. In *Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*.

Banerjee, S., and Pedersen, T. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. *Lecture notes in computer science* 136–145.

Fellbaum, C., ed. 1998. *Combining local context and WordNet similarity for word sense identification.* MIT Press. 265–283.

Ounis, I.; Lioma, C.; Macdonald, C.; and Plachouras, V. 2007. Research Directions in Terrier. *Novatica/UPGRADE Special Issue on Web Information Access*.

Ponzetto, S. P., and Navigli, R. 2009. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proc. of IJCAI-09*, 2083–2088.

Ruiz-Casado, M.; Alfonseca, E.; and Castells, P. 2005. Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. *Proc. of AWIC 2005*.