Word Sense Disambiguation in Roget's Thesaurus Using WordNet

Vivi Nastase and **Stan Szpakowicz**
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada
{vnastase,szpak}@site.uottawa.ca

Abstract

We describe a simple method of disambiguating word senses in *Roget's Thesaurus* using information about the sense of the word in *WordNet*. We present a few variations on this method, compare their performance and discuss the results. We explain why this type of disambiguation can be useful.

1 Introduction

The work presented in this paper is part of a larger research project. We learn the assignment of semantic relations to head nouns and modifiers in base noun phrases extracted automatically from *SemCor* and other texts ((Larrick, 1961), manually from (Levi, 1978)). In the absence of morphological and syntactic indicators of such relations, we rely on lexical resources for the semantic characterization of words. We want to compare the performance of different lexical resources available for this task. We have previously used *WordNet*. Turning to another lexical resource, *Roget's Thesaurus*, raises the problem of finding the proper senses of words. This is the problem we discuss here.

We have created two data sets for our machine-learning experiments. *DataSet1* contains 600 base noun phrases extracted from (Larrick, 1961) and (Levi, 1978). *DataSet2* contains base noun phrases extracted automatically from *SemCor*. Open-class words in *SemCor* are annotated with *WordNet* senses.

We are looking for a simple and fast algorithm that will help us annotate *DataSet2* with *Roget's* senses, using information about word senses from *WordNet*. We use *DataSet1* to test our word-sense disambiguation algorithm. We will then apply the algorithm to the semi-automatic annotation of *DataSet2*.

2 The algorithm

Previous work on selecting the right sense from *Roget's* for words in context includes (Yarowsky, 1995). The author used information from the context of words – keywords in the vicinity of the target word – that is indicative of the semantics of polysemous words. We want to see what results we can get by using as little context information as possible; for *DataSet1* the only context information is the base noun phrase in which the word appears. We will compare results obtained when no context information is available with results obtained when the base noun phrase serves as context. We will first consider head nouns in the context of their base noun phrase, and then head nouns and modifiers that mutually play the role of context.

Kwong (1998) has worked on aligning lexical resources. She used *WordNet* as the intermediary in passing from a word sense in *LDOCE* to a word sense in *Roget's*. She showed that it is possible to determine the sense of a word in *Roget's* by manually applying a simple algorithm that uses synsets, hypernym sets and coordinate terms from *WordNet* to a small set of words (36 divided into 3 test groups, according to the number of possible senses). Kwong experimented only with nouns.

Word-sense disambiguation that we propose handles nouns, adjectives and adverbs from base noun phrases. The algorithm uses the following information:

- from *Roget's* – all the paragraphs. A paragraph in *Roget's* is a group of words with the same part of speech, split into semicolon groups. Words in paragraphs are known to be related, although no explicit semantic link exists between them. Words in a semicolon group are known to be most closely related.
- from *WordNet* – all the information at one-link distance from a certain word (we call this a *WordNet mini-net*):
  - nouns – the word's synonyms, hyponyms, hypernyms, meronyms and holonyms;
  - adjectives, adverbs – the synonyms (the other sets are left empty);
  - words that are derived from another word w – the information pertaining to the word w, according to its part of speech.

Here is an example of a derived word:

sense 1 of adjective *cosmic*, its synset,

and the synonyms, hyponyms, hypernyms, meronyms and holonyms of the word *cosmos*.

```
cosmic
   Pertains to noun cosmos (Sense 1)
   => universe, existence, nature, creation,
      world, cosmos, macrocosm
     => natural object

Synonyms/Hypernyms (Ordered by Frequency) of
noun cosmos
Sense 1
universe, existence, nature, creation, world,
cosmos, macrocosm
       => natural object

Hyponyms of noun cosmos
Sense 1
universe, existence, nature, creation, world,
cosmos, macrocosm
       => natural order

Meronyms of noun cosmos
Sense 1
universe, existence, nature, creation, world,
cosmos, macrocosm
          HAS PART: celestial body,
                    heavenly body
(cosmos has no holonyms)
```

These sets are represented as lists, and we group them in a structure that represents a mini-net. It will be used in step 3 of the algorithm presented in Section 3. The word *cosmic* is represented as follows:

```
[[cosmic],
 [universe,existence,nature,creation,
  world,cosmos,macrocosm],
 [natural order],
 [natural object],
 [celestial body, heavenly body],
 []]
```

We will look for the occurrences of the word alone, and in phrases. Stemming was not done. In *DataSet1*, the words were represented by their root form. *SemCor* also contains information about the root form of all the open-class words annotated with *WordNet* senses.

3 The Algorithm

The idea of the algorithm is to consider all *Roget's* paragraphs in which a word appears (with the appropriate part of speech), and select the paragraph that captures best the semantics of the word. Words in paragraphs from *Roget's* are semantically related, but the relations are not specified. We try to match the information about the sense of the word in *WordNet* with each of these paragraphs, and the one that scores best in this matching becomes the corresponding sense in *Roget's*.

Let $\mathcal{P}$ be the set of paragraphs in our copy of *Roget's Thesaurus* (6380 paragraphs in total).

Let $\mathcal{W}$ be the set of words from *DataSet1*, without sense duplicates (the same word can appear several times, if every time it is associated with a different *WordNet* sense).

For each $w_i \in \mathcal{W}$:

1. Let $\mathcal{S}n, \mathcal{H}o, \mathcal{H}r, \mathcal{M}r, \mathcal{H}l$ be the synset, set of hyponyms, hypernyms, meronyms and holonyms, in this order, extracted from *WordNet*;
2. Let $\mathcal{S}(\mathcal{P}) \subseteq \mathcal{P}$ be a set of paragraphs, such that $\forall p_j \in \mathcal{S}(\mathcal{P}) \quad w_i \in p_j$;
3. For each $p_j \in \mathcal{S}(\mathcal{P})$, compute the score as a 5-tuple: $[|p_j \cap \mathcal{S}n|, |p_j \cap \mathcal{H}o|, |p_j \cap \mathcal{H}r|, |p_j \cap \mathcal{M}r|, |p_j \cap \mathcal{H}l|]$;
4. Arrange the results in lexicographic order of the scores;
5. Choose the sense that scored best.

This is the basic algorithm. We will experiment with variations on steps 3 and 4.

A paragraph in *Roget's* can contain single words and phrases that we treat as sets of words. In step 3, the intersection of a paragraph p from *Roget's* with a set of words from *WordNet* will be performed in two ways (WordNetSet is one of $\mathcal{S}n, \mathcal{H}o, \mathcal{H}r, \mathcal{M}r, \mathcal{H}l$):

- $x \in p \cap WordNetSet \Leftrightarrow$
  $x \in p \wedge x \in WordNetSet$
- $x \in p \cap WordNetSet \Leftrightarrow$
  $\exists phrase \in p \wedge x \in phrase \wedge x \in WordNetSet$

In step 4, the scores are ordered in two ways ($Score_i$ is the score associated with paragraph p_i, and $Score_j$ with paragraph p_j; recall that a score is a 5-tuple) :

- **lexicographic**
  $Score_i > Score_j \Leftrightarrow \exists k, 1 \leq k \leq 5$

  $$\begin{cases} s_{il} = s_{jl}, l < k \\ s_{ik} > s_{jk} \end{cases} \quad s_{ix} \in Score_i, s_{jx} \in Score_j$$
- **sum**

  $$Score_i \geq Score_j \Leftrightarrow \sum_{s_k \in Score_i} s_k \geq \sum_{s_l \in Score_j} s_l$$

For words w that in *WordNet* pertain to (or are derived from) another word w_p, the synset usually contains just the word w. The algorithm will then find the same score [1,0,0,0,0] for all the paragraphs that contain w. We have therefore decided to use the word w_p instead, and its disambiguated *Roget's* sense, for the semantic characterization of w.

Here is an example run of the algorithm for the word *cosmic* – adjective.

cosmic pertains to noun *cosmos*, sense 1.
The mini-net for Sense 1 of noun *cosmos*:

$$\begin{array}{ll} \mathcal{S}n = & \{universe, existence, nature, creation, \\ & world, cosmos, macrocosm\} \\ \mathcal{H}o = & \{natural\ order\} \\ \mathcal{H}r = & \{natural\ object\} \\ \mathcal{M}r = & \{celestial\ body,\ heavenly\ body\} \\ \mathcal{H}l = & \{\} \end{array}$$

The subset of paragraphs $\mathcal{S}(\mathcal{P})$ from *Roget's* such that $cosmos \in \mathcal{S}(\mathcal{P})$:

$$\begin{array}{ll} \mathcal{S}(\mathcal{P}) = \{ & \{whole, wholeness, integrality, ...\}\} \\ & \{arrangement, reduction\ to\ order, ...\}, \\ & \{universe, omneity, whole, world, ...\}, \end{array}$$

The lexicographically ordered list of scores:

$$\begin{array}{ll} \{ & \{universe, 27377, [5,0,0,0,0]\}, \\ & \{whole, 4804, [3,0,0,0,0]\}, \\ & \{arrangement, 5720, [2,0,0,0,0]\}\} \end{array}$$

A more complex example – for the noun *rain* (in the noun phrase *autumnal rain*).
The mini-net for Sense 1 of noun *rain*:

$$\begin{array}{ll} \mathcal{S}n = & \{rain, rainfall\} \\ \mathcal{H}o = & \{monsoon, downpour, cloudburst, \\ & deluge, waterspout, torrent, soaker \\ & drizzle, mizzle, shower, rain\ shower\} \\ \mathcal{H}r = & \{precipitation, downfall\} \\ \mathcal{M}r = & \{raindrop\} \\ \mathcal{H}l = & \{\} \end{array}$$

The subset of paragraphs $\mathcal{S}(\mathcal{P})$ from *Roget's* such that $rain \in \mathcal{S}(\mathcal{P})$:

$$\begin{array}{ll} \mathcal{S}(\mathcal{P}) = \{ & \{storm, turmoil, turbulence, ...\} \\ & \{descent, declension, declination, ...\} \\ & \{water, H2O, heavy\ water, ...\} \\ & \{weather, the\ element, fair\ weather, ...\} \\ & \{moisture, humidity, sap, juice, ...\}, \\ & \{moistening, humidification, ...\} \\ & \{rain, rainfall, moisture, ...\}\} \end{array}$$

The lexicographically ordered list of scores:

$$\begin{array}{ll} \{ & \{rain, 29270, [2,8,1,0,0]\} \\ & \{moisture, 28722, [2,1,0,1,0]\} \\ & \{descent, 26458, [1,2,1,0,0]\} \\ & \{moistening, 28743, [1,2,0,0,0]\} \\ & \{storm, 14216, [1,2,0,0,0]\} \\ & \{water, 28576, [1,1,0,0,0]\} \\ & \{weather, 28660, [1,0,0,0,0]\}\} \end{array}$$

4 Results and Discussion

We have conducted six experiments, using variations on the main algorithm, presented in Section 3.

In four of these experiments we did not look for occurrences of words in combinations, even if the correct sense of some of such words only occurs in phrases. (For example, *[sugar] cane* – this sense of *cane* only appears in combination with its modifier *sugar*). We made this decision because a paragraph in *Roget's* usually contains numerous phrases. Let w be a word taken from one of the five *WordNet* sets we consider. If w appears in many phrases in a paragraph in *Roget's*, this paragraph's score may become unduly high.

We also tried looking for words in phrases, to confirm our hypothesis that looking for words alone gives a better result. The results presented in Table 1 suggest that looking for words in combinations decreases the precision of the algorithm – the number of senses correctly marked decreases – but it increases the recall – some senses missed before are now found. We also note an increase of the number of senses correctly marked but at a tie with others, and of second choices.

Another variation is in comparing scores. We did lexicographic ordering of the lists of results, and also compared the sums of the elements of these lists.

While studying *Roget's Thesaurus*, we made a useful observation. If we look for the sense of a word w that appears as the first word in a *Roget's* paragraph, the paragraph overlaps mostly with w's hyponym set in *WordNet*. We have therefore chosen the ordering of the information from *WordNet* presented in step 1 of the algorithm: the synset is considered most important, then hyponyms, hypernyms, meronyms and finally holonyms. Lexicographic order is more discriminating than just a sum of individual scores, and our experiment variations show this.

We also introduced a bit of context. In one of the experiments we used the modifiers as a context for the head nouns. In another experiment modifiers and head nouns served as a context for one another. We consider another set $\mathcal{C}x$ containing the word and its context, and the score is a 6-tuple

Table 1: *WordNet-Roget's* sense correspondence for the 908 unique senses

	W1L	W1S	WCL	WCS	W1LH	W1LB
Words that do not appear in *Roget's*	28	28	22	22	28	28
Senses that do not appear in *Roget's*	16	16	18	18	16	16
Words for which all *Roget's* senses are a tie	75	70	44	40	70	67
Senses correctly marked	500	490	485	451	502	504
Senses marked correct but at a tie with others	82	125	89	140	82	82
Senses for which the second choice is the right one	175	153	208	204	180	181
Senses missed	32	26	42	33	30	30

Table 2: Comparison of results

	W1L	W1S	WCL	WCS	W1LH	W1LB
Correct sense identified	56.81%	55.68%	54.49%	50.67%	57.04%	57.27%
The correct sense is first or second sense indicated	86.02%	87.27%	88.26%	89.73%	86.81%	87.15%
Average number of senses	7.55	7.55	28.01	28.01	7.55	7.55

computed as follows (p is a paragraph from *Roget's*):

$$[|p \cap \mathcal{C}x|, |p \cap \mathcal{S}n|, |p \cap \mathcal{H}o|, |p \cap \mathcal{H}r|, |p \cap \mathcal{M}r|, |p \cap \mathcal{H}l|]$$

The context will be considered more important than the synset in our list of scores, because, if we find the exact base noun phrase in a *Roget's* paragraph, the words will have the senses we are looking for.

The results are presented for the following variants of the algorithm, and of the computation of the score:

- **W1L** Look for words alone; order the lists of results lexicographically.
- **W1S** Look for words alone; order the sum of the elements of the result lists.
- **WCL** Look for words in combinations; order the lists of results lexicographically.
- **WCS** Look for words in combinations; order the sum of elements of the result lists.
- **W1LH** Look for words alone; order the lists of results lexicographically, treating the modifier as the context for the head noun .
- **W1LB** Look for words alone; order the lists of results lexicographically, assuming that the head noun and the modifier have each other as context.

We have manually assigned senses from *Roget's* to the words in *DataSet1*, using information about the word senses in *WordNet* and the corresponding base noun phrases. From the 600 modifier-noun pairs in *DataSet1*, we extracted 1200 words, among them 908 unique senses. The results presented in Table 1 are computed by comparison with the manually annotated data.

Here are examples of senses from *DataSet1*, with their associated base noun phrases:

- Words that do not appear in *Roget's Thesaurus*: *lacy* - adjective (lacy pattern), *weekend* - noun (weekend boredom), *flounder* - adjective (flounder fish)
- Senses that do not appear in *Roget's Thesaurus*: *laser* (laser printer), *Tuesday* (Tuesday night)
- Words for which all senses are a tie: *bathing* (bathing suit), *laugh* (laugh wrinkles), *pet, spray* (pet spray)
- Senses correctly marked: *cloud, storm* (storm cloud), *weather* (weather report), *moist* (moist air)
- Senses marked correctly but at a tie with others: *tiny* (tiny clouds), *rope* (giant rope), *report* (weather report), *space* (open space)
- Senses for which the second choice is the right one: *purified* (purified water), *plan* (group plan), *aquatic* (aquatic mammal)
- Senses missed: *wire* (electric wire), *lightning* (lightning rod), *heavy* (heavy storm), *point* (sharp point)

There are differences in the numbers of words that do not appear in *Roget's* for the different variants of the algorithm. This is because sometimes a word appears in *Roget's* only in combinations (which are not found when we look for words alone), for example:

cane in *sugar cane*. There are cases when a word is found in a combination, but its sense is not found, for example: *Tuesday* (day of the week).

In Table 2 we show some statistical results. The percentages shown are computed using as a base the number of unique senses that appear in *Roget's*, according to that particular variation of the algorithm.

We have computed the accuracy separately for the 719 nouns in our data set. We obtain a precision average of 55.18% (word senses correctly and unambiguously marked) and 66.06% (first sense chosen is the correct one, ties may occur). Kwong reports an average of 72.77% senses accurately mapped. The average number of possible senses in *Roget's* for the words in our corpus is 7.55 when we look for the word alone, and 28.01 when we accept occurrences of words in combinations. The average number of senses for the words used by Kwong is 7.14.

Even though our algorithm is very simple and operates in a relative "semantic void", one would have liked the accuracy to be higher than 51-57%. We looked at our results, and we have observed the following phenomena:

- some *Roget's* paragraphs are quite similar, and they get the same score. For the adjective *tiny*, for example, the two paragraphs in which it appears received the same score. The paragraph keywords are *small* and *little-small*.
- *WordNet* is oriented toward words, whereas *Roget's* is oriented toward phrases. Words that appear in many phrases in a paragraph, change the score in favour of this paragraph, and the score obtained can be misleading.

We can also note that we have set a fairly high standard for a rather simple algorithm. In the process of manually assigning correct senses from *Roget's*, we consider very carefully the sense of the word not only in the context of its base noun phrase, but also within paragraphs to which it could belong. If, as we have repeatedly observed, paragraphs contained words that were indeed very close in meaning, we chose to weigh complete paragraphs against each other. Sometimes our eventual choice rated second, or even lower, on the ordered list of scored paragraphs.

5 Conclusions

We needed a simple and fast algorithm to help us annotate a list of base noun phrases with senses from *Roget's Thesaurus*. The present study helped eliminate two variations of our simple algorithm – the ones that look for the word in phrases. They are more computationally expensive, and the gain of at most 2.5% in recall – comparing W1LB (best precision) and WCS (best recall) – with a loss of 6.6% in precision does not justify using these two methods.

We want to perform supervised machine learning on base noun phrases annotated with semantic relations, while the individual words are described by information coming from different lexical resources. The correct *Roget's* sense can be selected from the first two senses indicated by our simple algorithm with a recall of 86.81% (when the average number of senses in *DataSet1* is 7.5). This significantly reduces the amount of manual labour necessary to annotate our corpora with *Roget's* senses.

In our experiment, the words were annotated with *WordNet* senses. This algorithm and Kwong's paper could serve as proof that there is a correspondence between *WordNet* and *Roget's* senses. We can use these two resources in a word-sense disambiguation algorithm for untagged data; they should reinforce each other's choice of the correct sense of a word in context.

The uses of our disambiguation mechanism could be numerous. *Roget's* hierarchy is very homogeneous. This could be a great help in testing the similarity of words, as shown by McHale (1998). When the similarity of words is decided by edge counting, McHale shows that the homogeneous organization of *Roget's Thesaurus* gives better results than *WordNet*.

Another use would be to find related words that have different parts of speech, or using *Roget's Thesaurus* to add pertainyms to *WordNet*. This is because words in *Roget's* with different parts of speech are grouped together under the same headwords.

We have shown that, with a very simple algorithm, the correct sense from *Roget's* can be selected from the first two senses indicated with an average recall of 86.81%, with very little context information. We find this a very promising result.

6 Acknowledgements

Pearson Education licensed to us the 1987 *Roget's Thesaurus* for research purposes. Partial funding for this work comes from the Natural Sciences and Engineering Research Council of Canada.

References

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT press.

Betty Kirkpatrick. 1987. *Penguin Authorized Roget's Thesaurus of English Words and Phrases*. Penguin Books, London & New York edition.

Oi Yee Kwong. 1998. Aligning WordNet with Additional Lexical Resources. In *COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 73–79.

N. Larrick. 1961. *Junior Science Book of Rain, Hail, Sleet and Snow*. Garrard Publishing Company, Champaign, IL.

J. N. Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

Michael McHale. 1998. A Comparison of WordNet and Roget's Taxonomy for Measuring Semantic Similarity. In *COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.

George A. Miller. 1990. Five Papers on WordNet. *Special issue of International Journal of Lexicography 3(4)*.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA.