Evaluating *Roget's* Thesauri

Alistair Kennedy
School of Information Technology
and Engineering
University of Ottawa
Ottawa, Ontario, Canada
akennedy@site.uottawa.ca

Stan Szpakowicz
School of Information Technology
and Engineering
University of Ottawa
Ottawa, Ontario, Canada
and
Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland
szpak@site.uottawa.ca

Abstract

Roget's Thesaurus has gone through many revisions since it was first published 150 years ago. But how do these revisions affect *Roget's* usefulness for NLP? We examine the differences in content between the 1911 and 1987 versions of *Roget's*, and we test both versions with each other and *WordNet* on problems such as synonym identification and word relatedness. We also present a novel method for measuring sentence relatedness that can be implemented in either version of *Roget's* or in *WordNet*. Although the 1987 version of the Thesaurus is better, we show that the 1911 version performs surprisingly well and that often the differences between the versions of *Roget's* and *WordNet* are not statistically significant. We hope that this work will encourage others to use the 1911 *Roget's* Thesaurus in NLP tasks.

1 Introduction

Roget's Thesaurus, first introduced over 150 years ago, has gone through many revisions to reach its current state. We compare two versions, the 1987 and 1911 editions of the Thesaurus with each other and with *WordNet* 3.0. *Roget's* Thesaurus has a unique structure, quite different from *WordNet*, of which the NLP community has yet to take full advantage. In this paper we demonstrate that although the 1911 version of the Thesaurus is very old, it can give results comparable to systems that use *WordNet* or newer versions of *Roget's* Thesaurus.

The main motivation for working with the 1911 Thesaurus instead of newer versions is that it is in the public domain, along with related NLP-oriented software packages. For applications that call for an NLP-friendly thesaurus, *WordNet* has become the de-facto standard. Although *WordNet* is a fine resources, we believe that ignoring other thesauri is a serious oversight. We show on three applications how useful the 1911 Thesaurus is. We ran the well-established tasks of determining semantic relatedness of pairs of terms and identifying synonyms (Jarmasz and Szpakowicz, 2004). We also proposed a new method of representing the meaning of sentences or other short texts using either *WordNet* or *Roget's* Thesaurus, and tested it on the data set provided by Li et al. (2006). We hope that this work will encourage others to use *Roget's* Thesaurus in their own NLP tasks.

Previous research on the 1987 version of *Roget's* Thesaurus includes work of Jarmasz and Szpakowicz (2004). They propose a method of determining semantic relatedness between pairs of terms. Terms that appear closer together in the Thesaurus get higher weights than those farther apart. The experiments aimed at identifying synonyms using a modified version of the proposed semantic similarity function. Similar experiments were carried out using *WordNet* in combination with a variety of semantic relatedness functions. *Roget's* Thesaurus was found generally to outperform *WordNet* on these problems. We have run similar experiments using the 1911Thesaurus.

Lexical chains have also been developed using the 1987 *Roget's* Thesaurus (Jarmasz and Szpakowicz, 2003). The procedure maps words in a text to the Head (a *Roget's* concept) from which they are most likely to come. Although we did not experiment

Proceedings of ACL-08: HLT, pages 416–424,
Columbus, Ohio, USA, June 2008.

with lexical chains here, they were an inspiration for our sentence relatedness function.

Roget's Thesaurus does not explicitly label the relations between its terms, as *WordNet* does. Instead, it groups terms together with implied relations. Kennedy and Szpakowicz (2007) show how disambiguating one of these relations, hypernymy, can help improve the semantic similarity functions in (Jarmasz and Szpakowicz, 2004). These hypernym relations were also put towards solving analogy questions.

This is not the first time the 1911 version of *Roget's* Thesaurus has been used in NLP research. Cassidy (2000) used it to build the semantic network FACTOTUM. This required significant (manual) restructuring, so FACTOTUM cannot really be considered a true version of *Roget's* Thesaurus.

The 1987 data come from *Penguin's Roget's Thesaurus* (Kirkpatrick, 1987). The 1911 version is available from Project Gutenberg[1]. We use *WordNet* 3.0, the latest version (Fellbaum, 1998). In the experiments we present here, we worked with an interface to *Roget's* Thesaurus implemented in Java 5.0[2]. It is built around a large index which stores the location in the thesaurus of each word or phrase; the system individually indexes all words within each phrase, as well as the phrase itself. This was shown to improve results in a few applications, which we will discuss later in the paper.

2 Content comparison of the 1911 and 1987 Thesauri

Although the 1987 and 1911 Thesauri are very similar in structure, there are a few differences, among them, the number of levels and the number of parts-of-speech represented. For example, the 1911 version contains some pronouns as well as more sections dedicated to phrases.

There are nine levels in *Roget's* Thesaurus hierarchy, from Class down to Word. We show them in Table 1 along with the counts of instances of each level. An example of a Class in the 1911 Thesaurus is "Words Expressing Abstract Relations", a Section in that Class is "Quantity" with a Subsection "Comparative Quantity". Heads can be thought of as the heart of the Thesaurus because it is at this level that the lexical material, organized into approximately a thousand concepts, resides. Head Groups often pair up opposites, for example Head #1 "Existence" and Head #2 "Nonexistence" are found in the same Head Group in both versions of the Thesaurus. Terms in the Thesaurus may be labelled with cross-references to other words in different Heads. We did not use these references in our experiments.

Hierarchy	**1911**	**1987**
Class	8	8
Section	39	39
Subsection	97	95
Head Group	625	596
Head	1044	990
Part-of-speech	3934	3220
Paragraph	10244	6443
Semicolon Group	43196	59915
Total Words	98924	225124
Unique Words	59768	100470

Table 1: Frequencies of each level of the hierarchy in the 1911 and 1987 Thesauri.

The part-of-speech level is a little confusing, since clearly no such grouping contains an exhaustive list of all nouns, all verbs etc. We will write "POS" to indicate a structure in *Roget's* and "part-of-speech" to indicate the word category in general. The four main parts-of-speech represented in a POS are nouns, verbs, adjectives and adverbs. Interjections are also included in both the 1911 and 1987 thesauri; they are usually phrases followed by an exclamation mark, such as "for God's sake!" and "pshaw!". The Paragraph and Semicolon Group are not given names, but can often be represented by the first word.

The 1911 version also contains phrases (mostly quotations), prefixes and pronouns. There are only three prefixes – "tri-", "tris-", "laevo-" – and six pronouns – "he", "him", "his", "she", "her", "hers".

Table 2 shows the frequency of paragraphs, semicolon groups and both total and unique words in a given type of POS. Many terms occur both in the 1911 and 1987 Thesauri, but many more are unique to either. Surprisingly, quite a few 1911 terms do not appear in the 1987 data, as shown in Table 3; many of them may have been considered obsolete and thus dropped from the 1987 version. For example "ingrafted" appears in the same semicolon group as

[1] http://www.gutenberg.org/ebooks/22

[2] http://rogets.site.uottawa.ca/

POS	Paragraph		Semicolon Grp	
	1911	1987	1911	1987
Noun	4495	2884	19215	31174
Verb	2402	1499	10838	13958
Adjective	2080	1501	9097	12893
Adverb	594	499	2028	1825
Interjection	108	60	149	65
Phrase	561	0	1865	0
	Total Word		**Unique Words**	
	1911	1987	1911	1987
Noun	46308	114473	29793	56187
Verb	25295	55724	15150	24616
Adjective	20447	48802	12739	21614
Adverb	4039	5720	3016	4144
Interjection	598	405	484	383
Phrase	2228	0	2038	0

Table 2: Frequencies of paragraphs, semicolon groups, total words and unique words by their part of speech; we omitted prefixes and pronouns.

POS	Both	Only 1911	Only 1987
All	35343	24425	65127
N.	18685	11108	37502
Vb.	8618	6532	15998
Adj.	8584	4155	13030
Adv.	1684	1332	2460
Int.	68	416	315
Phr.	0	2038	0

Table 3: Frequencies of terms in either the 1911 or 1987 Thesaurus, and in both; we omitted prefixes and pronouns.

"implanted" in the older but not the newer version. Some mismatches may be due to small changes in spelling, for example, "Nirvana" is capitalized in the 1911 version, but not in the 1987 version.

The lexical data in Project Gutenberg's 1911 *Roget's* appear to have been somewhat added to. For example, the citation "Go ahead, make my day!" from the 1971 movie *Dirty Harry* appears twice (in Heads #715-Defiance and #761-Prohibition) within the *Phrase* POS. It is not clear to what extent new terms have been added to the original 1911 *Roget's* Thesaurus, or what the criteria for adding such new elements could have been.

In the end, there are many differences between the 1987 and 1911 *Roget's* Thesauri, primarily in content rather than in structure. The 1987 Thesaurus is largely an expansion of the 1911 version, with three POSs (phrases, pronouns and prefixes) removed.

3 Comparison on applications

In this section we consider how the two versions of *Roget's* Thesaurus and *WordNet* perform in three applications – measuring word relatedness, synonym identification, and sentence relatedness.

3.1 Word relatedness

Relatedness can be measured by the closeness of the words or phrases – henceforth referred to as *terms* – in the structure of the thesaurus. Two terms in the same semicolon group score 16, in the same paragraph – 14, and so on (Jarmasz and Szpakowicz, 2004). The score is 0 if the terms appear in different classes, or if either is missing. Pairs of terms get higher scores for being closer together. When there are multiple senses of two terms A and B, we want to select senses $a \in A$ and $b \in B$ that maximize the relatedness score. We define a distance function:

$$semDist(A,B) = \max_{a \in A, b \in B} 2 * (depth(lca(a,b)))$$

lca is the *lowest common ancestor* and *depth* is the depth in the *Roget's* hierarchy; a Class has depth 0, Section 1, ..., Semicolon Group 8. If we think of the function as counting edges between concepts in the *Roget's* hierarchy, then it could also be written as:

$$semDist(A,B) = \max_{a \in A, b \in B} 16 - edgesBetween(a,b)$$

We do not count links between words in the same semicolon group, so in effect these methods find distances between semicolon groups, that is to say, these two functions will give the same results.

The 1911 and 1987 Thesauri were compared with *WordNet* 3.0 on the three data sets containing pairs of words with manually assigned similarity scores: 30 pairs (Miller and Charles, 1991), 65 pairs (Rubenstein and Goodenough, 1965) and 353 pairs[3] (Finkelstein et al., 2001). We assume that all terms are nouns, so that we can have a fair comparison of the two Thesauri with *WordNet*. We measure the correlation with Pearson's Correlation Coefficient.

[3] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html

Year	Miller & Charles	Rubenstein & Goodenough	Finkelstein et. al
Index words and phrase			
1911	0.7846	0.7313	0.3449
1987	0.7984	0.7865	0.4214
Index phrase only			
1911	0.7090	0.7168	0.3373
1987	0.7471	0.7777	0.3924

Table 4: Pearson's coefficient values when not breaking / breaking phrases up.

A preliminary experiment set out to determine whether there is any advantage to indexing the words in a phrase separately, for example, whether the phrase "change of direction" should be indexed only as a whole, or as all of "change", "of", "direction" and "change of direction". The outcome of this experiment appears in Table 4. There is a clear improvement: breaking phrases up gives superior results on all three data sets, for both versions of *Roget's*. In the remaining experiments, we have each word in a phrase indexed.

We compare the results for the 1911 and 1987 *Roget's* Thesauri with a variety of *WordNet*-based semantic relatedness measures – see Table 5. We consider 10 measures, noted in the table as J&C (Jiang and Conrath, 1997), Resnik (Resnik, 1995), Lin (Lin, 1998), W&P (Wu and Palmer, 1994), L&C (Leacock and Chodorow, 1998), H&SO (Hirst and St-Onge, 1998), Path (counts edges between synsets), Lesk (Banerjee and Pedersen, 2002), and finally Vector and Vector Pair (Patwardhan, 2003). The latter two work with large vectors of co-occurring terms from a corpus, so *WordNet* is only part of the system. We used Pedersen's Semantic Distance software package (Pedersen et al., 2004).

The results suggest that neither version of *Roget's* is best for these data sets. In fact, the Vector method is superior on all three sets, and the Lesk algorithm performs very closely to *Roget's* 1987. Even on the largest set (Finkelstein et al., 2001), however, the differences between *Roget's* Thesaurus and the Vector method are not statistically significant at the $p < 0.05$ level for either thesaurus on a two-tailed test[4]. The difference between the 1911 Thesaurus and Vector *would* be statistically significant at $p < 0.07$.

[4]http://faculty.vassar.edu/lowry/rdiff.html

Method	Miller & Charles	Rubenstein & Goodenough	Finkelstein et. al
1911	0.7846	0.7313	0.3449
1987	0.7984	0.7865	0.4214
J&C	0.4735	0.5755	0.2273
Resnik	0.8060	0.8224	0.3531
Lin	0.7388	0.7264	0.2932
W&P	0.7641	0.7973	0.2676
L&C	0.7792	0.8387	0.3094
H&SO	0.6668	0.7258	0.3548
Path	0.7550	0.7842	0.3744
Lesk	0.7954	0.7780	0.4220
Vector	0.8645	0.7929	0.4621
Vct Pair	0.5101	0.5810	0.3722

Table 5: Pearson's coefficient values for three data sets on a variety of relatedness functions.

On the (Miller and Charles, 1991) and (Rubenstein and Goodenough, 1965) data sets the best system did not show a statistically significant improvement over the 1911 or 1987 *Roget's* Thesauri, even at $p < 0.1$ for a two-tailed test. These data sets are too small for a meaningful comparison of systems with close correlation scores.

3.2 Synonym identification

In this problem we take a term q and we seek the correct synonym s from a set C. There are two steps. We used the system from (Jarmasz and Szpakowicz, 2004) for identifying synonyms with *Roget's*. First we find a set of terms $B \subseteq C$ with the maximum relatedness between q and each term $x \in C$:

$$B = \{x \mid \operatorname*{argmax}_{x \in C} semDist(x, q)\}$$

Next, we take the set of terms $A \subseteq B$ where each $a \in A$ has the maximum number of shortest paths between a and q.

$$A = \{x \mid \operatorname*{argmax}_{x \in B} numberShortestPaths(x, q)\}$$

If $s \in A$ and $|A| = 1$, the correct synonym has been selected. Often the sets A and B will contain just one item. If $s \in A$ and $|A| > 1$, there is a tie. If $s \notin A$ then the selected synonyms are incorrect. If a multi-word phrase $c \in C$ of length n is not found,

ESL						
Method	Yes	Tie	No	QNF	ANF	ONF
1911	27	3	20	0	3	3
1987	36	6	8	0	0	1
J&C	30	4	16	4	4	10
Resnik	26	6	18	4	4	10
Lin	31	5	14	4	4	10
W&P	31	6	13	4	4	10
L&C	29	11	10	4	4	10
H&SO	34	4	12	0	0	0
Path	30	11	9	4	4	10
Lesk	38	0	12	0	0	0
Vector	39	0	11	0	0	0
VctPair	40	0	10	0	0	0
TOEFL						
1911	52	3	25	10	5	25
1987	59	7	14	4	4	17
J&C	34	37	9	33	31	90
Resnik	37	37	6	33	31	90
Lin	33	41	6	33	31	90
W&P	39	36	5	33	31	90
L&C	38	36	6	33	31	90
H&SO	60	16	4	1	0	1
Path	38	36	6	33	31	90
Lesk	70	1	9	1	0	1
Vector	69	1	10	1	0	1
VctPair	65	2	13	1	0	1
RDWP						
1911	157	13	130	57	13	76
1987	198	17	85	22	5	17
J&C	100	146	54	62	58	150
Resnik	114	114	72	62	58	150
Lin	94	160	46	62	58	150
W&P	147	87	66	62	58	150
L&C	149	93	58	62	58	150
H&SO	170	82	48	4	6	5
Path	148	96	56	62	58	150
Lesk	220	7	73	4	6	5
Vector	216	7	73	4	6	5
VctPair	187	10	103	4	6	5

Table 6: Synonym selection experiments.

it is replaced by each of its words $c_1, c_2..., c_n$, and each of these words is considered in turn. The c_i that is closest to q is chosen to represent c. When searching for a word in *Roget's* or *WordNet*, we look for all forms of the word.

The results of these experiments appear in Table 6. "Yes" indicates correct answers, "No" – incorrect answers, and "Tie" is for ties. QNF stands for "Question word Not Found", ANF for "Answer word Not Found" and ONF for "Other word Not Found". We used three data sets for this application: 80 questions taken from the Test of English as a Foreign Language (TOEFL) (Landauer and Dumais, 1997), 50 questions – from the English as a Second Language test (ESL) (Turney, 2001) and 300 questions – from the Reader's Digest Word Power Game (RDWP) (Lewis, 2000 and 2001).

Lesk and the Vector-based systems perform better than all others, including *Roget's* 1911 and 1987. Even so, both versions of *Roget's* Thesaurus performed well, and were never worse than the worst *WordNet* systems. In fact, six of the ten *WordNet*-based methods are consistently worse than the 1911 Thesaurus. Since the two Vector-based systems make use of additional data beyond *WordNet*, Lesk is the only completely *WordNet*-based system to outperform *Roget's* 1987. One advantage of *Roget's* Thesaurus is that both versions generally have fewer missing terms than *WordNet*, though Lesk, Hirst & St-Onge and the two vector based methods had fewer missing terms than *Roget's*. This may be because the other *WordNet* methods will only work for nouns and verbs.

3.3 Sentence relatedness

Our final experiment concerns sentence relatedness. We worked with a data set from (Li et al., 2006)[5]. They took a subset of the term pairs from (Rubenstein and Goodenough, 1965) and chose sentences to represent these terms; the sentences are definitions from the Collins Cobuild dictionary (Sinclair, 2001). Thirty people were then asked to assign relatedness scores to these sentences, and the average of these similarities was taken for each sentence.

Other methods of determining sentence semantic relatedness expand term relatedness functions to

[5]http://www.docm.mmu.ac.uk/STAFF/D.McLean/SentenceResults.htm

create a sentence relatedness function (Islam and Inkpen, 2007; Mihalcea et al., 2006). We propose to approach the task by exploiting in other ways the commonalities in the structure of *Roget's* Thesaurus and of *WordNet*. We use the OpenNLP toolkit[6] for segmentation and part-of-speech tagging.

We use a method of sentence representation that involves mapping the sentence into weighted concepts in either *Roget's* or *WordNet*. We mean a concept in *Roget's* to be either a Class, Section, ..., Semicolon Group, while a concept in *WordNet* is any synset. Essentially a concept is a grouping of words from either resource. Concepts are weighted by two criteria. The first is how frequently words from the sentence appear in these concepts. The second is the depth (or specificity) of the concept itself.

3.3.1 Weighting based on word frequency

Each word and punctuation mark w in a sentence is given a score of 1. (Naturally, only open-category words will be found in the thesaurus.) If w has n word senses $w_1, ..., w_n$, each sense gets a score of $1/n$, so that $1/n$ is added to each concept in the *Roget's* hierarchy (semicolon group, paragraph, ..., class) or *WordNet* hierarchy that contains w_i. We weight concepts in this way simply because, unable to determine which sense is correct, we assume that all senses are equally probable. Each concept in *Roget's* Thesaurus and *WordNet* gets the sum of the scores of the concepts below it in its hierarchy.

We will define the scores recursively for a concept c in a sentence s and sub-concepts c_i. For example, in *Roget's* if the concept c were a Class, then each c_i would be a Section. Likewise, in *WordNet* if c were a synset, then each c_i would be a hyponym synset of c. Obviously if c is a word sense w_i (a word in either a synset or a Semicolon Group), then there can be no sub-concepts c_i. When $c = w_i$, the score for c is the sum of all occurrences of the word w in sentence s divided by the number of senses of the word w.

$$score(c, s) = \begin{cases} \frac{instancesOf(w,s)}{sensesOf(w)} & \text{if} \quad c = w_i \\ \sum_{c_i \in c} score(c_i, s) & \text{otherwise} \end{cases}$$

See Table 7 for an example of how this sentence representation works. The sentence "A gem is a jewel or stone that is used in jewellery." is represented using the 1911 *Roget's*. A concept is identified by a name and a series of up to 9 numbers that indicate where in the thesaurus it appears. The first number represents the Class, the second the Section, ..., the ninth the word. We only show concepts with weights greater than 1.0. Words not in the thesaurus keep a weight of 1.0, but this weight will not increase the weight of any concepts in *Roget's* or *WordNet*. Apart from the function words "or", "in", "that" and "a" and the period, only the word "jewellery" had a weight above 1.0. The categories labelled 6, 6.2 and 6.2.2 are the only ancestors of the word "use" that ended up with the weights above 1.0. The words "gem", "is", "jewel", "stone" and "used" all contributed weight to the categories shown in Table 7, and to some categories with weights lower than 1.0, but no sense of the words themselves had a weight greater than 1.0.

It is worth noting that this method only relies on the hierarchies in *Roget's* and *WordNet*. We do not take advantage of other *WordNet* relations such as hyponymy, nor do we use any cross-reference links that exist in *Roget's* Thesaurus. Including such relations might improve our sentence relatedness system, but that has been left for future work.

3.3.2 Weighting based on specificity

To determine sentence relatedness, one could, for example, flatten the structures like those in Table 7 into vectors and measure their closeness by some vector distance function such as cosine similarity. There is a problem with this, though. A concept inherits the weights of all its sub-concepts, so the concepts that appear closer to the root of the tree will far outweigh others. Some sort of weighting function should be used to re-adjust the weights of particular concepts. Were this an Information Retrieval task, weighting schemes such as *tf.idf* for each concept could apply, but for sentence relatedness we propose an *ad hoc* weighting scheme based on assumptions about which concepts are most important to sentence representation. This weighting scheme is the second element of our sentence relatedness function.

We weight a concept in *Roget's* and in *WordNet* by how many words in a sentence give weight to it. We need to re-weight it based on how specific it is. Clearly, concepts near the leaves of the hierarchy are more specific than those close to the root of the hierarchy. We define specificity as the distance in levels between a given word and each concept found above

[6] http://opennlp.sourceforge.net

Identifier	Concept	Weight
6	Words Relating to the Voluntary Powers - Individual Volition	2.125169028274
6.2	Prospective Volition	1.504066255252
6.2.2	Subservience to Ends	1.128154077172
8	Words Relating to the Sentiment and Moral Powers	3.13220884041
8.2	Personal Affections	1.861744448402
8.2.2	Discriminative Affections	1.636503978149
8.2.2.2	Ornament/Jewelry/Blemish [Head Group]	1.452380952380
8.2.2.2.886	Jewelry [Head]	1.452380952380
8.2.2.2.886.1	Jewelry [Noun]	1.452380952380
8.2.2.2.886.1.1	jewel [Paragraph]	1.452380952380
8.2.2.2.886.1.1.1	jewel [Semicolon Group]	1.166666666666
8.2.2.2.886.1.1.1.3	jewellery [Word Sense]	1.0
or	-	1.0
in	-	1.0
that	-	1.0
a	-	2.0
.	-	1.0

Table 7: "A gem is a jewel or stone that is used in jewellery." as represented using *Roget's* 1911.

it in the hierarchy. In *Roget's* Thesaurus there are exactly 9 levels from the term to the class. In *WordNet* there will be as many levels as a word has ancestors up the hypernymy chain. In *Roget's*, a term has specificity 1, a Semicolon Group 2, a Paragraph 3, ..., a Class 9. In *WordNet*, the specificity of a word is 1, its synset – 2, the synset's hypernym – 3, *its* hypernym – 4, and so on. Words not found in the Thesaurus or in *WordNet* get specificity 1.

We seek a function that, given s, assigns to all concepts of specificity s a weight progressively larger than to their neighbours. The weights in this function should be assigned based on specificity, so that all concepts of the same specificity receive the same score. Weights will differ depending on a combination of specificity and how frequently words that signal the concepts appear in a sentence. The weight of concepts with specificity s should be the highest, of those with specificity $s \pm 1$ – lower, of those with specificity $s \pm 2$ lower still, and so on. In order to achieve this effect, we weight the concepts using a normal distribution, where the mean is s:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-s)^2}{2\sigma^2}\right)}$$

Since the Head is often considered the main category in *Roget's*, we expect a specificity of 5 to be best, but we decided to test the values 1 through 9 as a possible setting for specificity. We do not claim that this weighting scheme is optimal; other weighting schemes might do better. For the purpose of comparing the 1911 and 1987 Thesauri and *WordNet*, however, this method appears sufficient.

With this weighting scheme, we determine the distance between two sentences using cosine similarity:

$$cosSim(A, B) = \frac{\sum a_i * b_i}{\sqrt{\sum a_i^2} * \sqrt{\sum b_i^2}}$$

For this problem we used the MIT Java *WordNet* Interface version 1.1.1[7].

3.3.3 Sentence similarity results

We used this method of representation for *Roget's* of 1911 and of 1987, as well as for *WordNet* 3.0 – see Figure 1. For comparison, we also implemented a baseline method that we refer to as Simple: we built vectors out of words and their count.

It can be seen in Figure 1 that each system is superior for at least one of the nine specificities. The Simple method is best at a specificity of 1, 8 and 9, *Roget's* Thesaurus 1911 is best at 6, *Roget's* Thesaurus 1987 is best at 4, 5 and 7, and *WordNet* is best at 2 and 3. The systems based on *Roget's* and *WordNet* more or less followed a bell-shaped curve, with the curves of the 1911 and 1987 Thesauri following each other fairly closely and peaking close together. *WordNet* clearly peaked first and then fell the farthest.

[7] http://www.mit.edu/~markaf/projects/wordnet/

The best correlation result for the 1987 *Roget's* Thesaurus is 0.8725 when the mean is 4, the POS. The maximum correlation for the 1911 Thesaurus is 0.8367, where the mean is 5, the Head. The maximum for *WordNet* is 0.8506, where the mean is 3, or the first hypernym synset. This suggests that the POS and Head are most important for representing text in *Roget's* Thesaurus, while the first hypernym is most important for representing text using *WordNet*. For the Simple method, we found a more modest correlation of 0.6969.

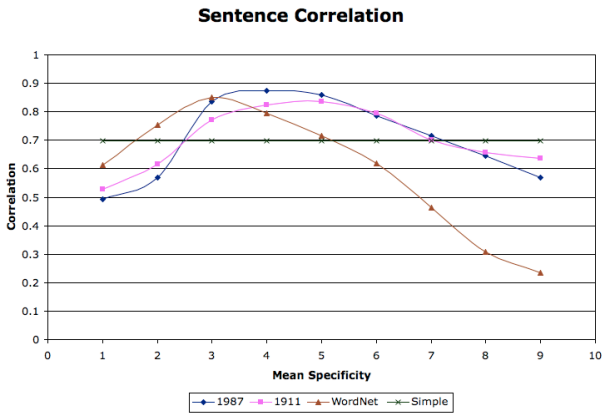

Figure 1: Correlation data for all four systems.

Several other methods have given very good scores on this data set. For the system in (Li et al., 2006), where this data set was first introduced, a correlation of 0.816 with the human annotators was achieved. The mean of all human annotators had a score of 0.825, with a standard deviation of 0.072. In (Islam and Inkpen, 2007), an even better system was proposed, with a correlation of 0.853.

Selecting the mean that gives the best correlation could be considered as training on test data. However, were we simply to have selected a value somewhere in the middle of the graph, as was our original intuition, it would have given an unfair advantage to either version of *Roget's* Thesaurus over *WordNet*. Our system shows good results for both versions of *Roget's* Thesauri and *WordNet*. The 1987 Thesaurus once again performs better than the 1911 version and than *WordNet*. Much like (Miller and Charles, 1991), the data set used here is not large enough to determine if any system's improvement is statistically significant.

4 Conclusion and future work

The 1987 version of *Roget's* Thesaurus performed better than the 1911 version on all our tests, but we did not find the differences to be statistically significant. It is particularly interesting that the 1911 Thesaurus performed as well as it did, given that it is almost 100 years old. On problems such as semantic word relatedness, the 1911 Thesaurus performance was fairly close to that of the 1987 Thesaurus, and was comparable to many *WordNet*-based measures. For problems of identifying synonyms both versions of *Roget's* Thesaurus performed relatively well compared to most *WordNet*-based methods.

We have presented a new method of sentence representation that attempts to leverage the structure found in *Roget's* Thesaurus and similar lexical ontologies (among them *WordNet*). We have shown that given this style of text representation both versions of *Roget's* Thesaurus work comparably to *WordNet*. All three perform fairly well compared to the baseline Simple method. Once again, the 1987 version is superior to the 1911 version, but the 1911 version still works quite well.

We hope to investigate further the representation of sentences and other short texts using *Roget's* Thesaurus. These kinds of measurements can help with problems such as identifying relevant sentences for extractive text summarization, or possibly paraphrase identification (Dolan et al., 2004). Another – longer-term – direction of future work could be merging *Roget's* Thesaurus with *WordNet*.

We also plan to study methods of automatically updating the 1911 *Roget's* Thesaurus with modern words. Some work has been done on adding new terms and relations to *WordNet* (Snow et al., 2006) and FACTOTUM (O'Hara and Wiebe, 2003). Similar methods could be used for identifying related terms and assigning them to a correct semicolon group or paragraph.

Acknowledgments

Our research is supported by the Natural Sciences and Engineering Research Council of Canada and the University of Ottawa. We thank Dr. Diana Inkpen, Anna Kazantseva and Oana Frunza for many useful comments on the paper.

References

S. Banerjee and T. Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proc. CICLing 2002*, pages 136–145.

P. Cassidy. 2000. An investigation of the semantic relations in the roget's thesaurus: Preliminary results. In *Proc. CICLing 2000*, pages 181–204.

B. Dolan, C. Quirk, and C. Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proc. COLING 2004*, pages 350–356, Morristown, NJ.

C. Fellbaum. 1998. A semantic network of english verbs. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 69–104. MIT Press, Cambridge, MA.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2001. Placing search in context: the concept revisited. In *Proc. 10th International Conf. on World Wide Web*, pages 406–414, New York, NY, USA. ACM Press.

G. Hirst and D. St-Onge. 1998. Lexical chains as representation of context for the detection and correction malapropisms. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–322. MIT Press, Cambridge, MA.

A. Islam and D. Inkpen. 2007. Semantic similarity of short texts. In *Proc. RANLP 2007*, pages 291–297, September.

M. Jarmasz and S. Szpakowicz. 2003. Not as easy as it seems: Automating the construction of lexical chains using roget's thesaurus. In *Proc. 16th Canadian Conf. on Artificial Intelligence*, pages 544–549.

M. Jarmasz and S. Szpakowicz. 2004. Roget's thesaurus and semantic similarity. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003, Current Issues in Linguistic Theory*, volume 260, pages 111–120. John Benjamins.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. 10th International Conf. on Research on Computational Linguistics*, pages 19–33.

A. Kennedy and S. Szpakowicz. 2007. Disambiguating hypernym relations for roget's thesaurus. In *Proc. TSD 2007*, pages 66–75.

B. Kirkpatrick, editor. 1987. *Roget's Thesaurus of English Words and Phrases*. Penguin, Harmondsworth, Middlesex, England.

T. Landauer and S. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

C. Leacock and M. Chodorow. 1998. Combining local context and wordnet sense similiarity for word sense disambiguation. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–284. MIT Press, Cambridge, MA.

M. Lewis, editor. 2000 and 2001. *Readers Digest, 158(932, 934, 935, 936, 937, 938, 939, 940), 159(944, 948)*. Readers Digest Magazines Canada Limited.

Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

R. Mihalcea, C. Corley, and C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proc. 21st National Conf. on Artificial Intelligence*, pages 775–780. AAAI Press.

G. A. Miller and W. G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Process*, 6(1):1–28.

T. P. O'Hara and J. Wiebe. 2003. Classifying functional relations in factotum via wordnet hypernym associations. In *Proc. CICLing 2003)*, pages 347–359.

S. Patwardhan. 2003. Incorporating dictionary and corpus information into a vector measure of semantic relatedness. Master's thesis, University of Minnesota, Duluth, August.

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proc. of the 19th National Conference on Artificial Intelligence.*, pages 1024–1025.

P. Resnik. 1995. Using information content to evaluate semantic similarity. In *Proc. 14th International Joint Conf. on Artificial Intelligence*, pages 448–453.

H. Rubenstein and J. B. Goodenough. 1965. Contextual correlates of synonymy. *Communication of the ACM*, 8(10):627–633.

J. Sinclair. 2001. *Collins Cobuild English Dictionary for Advanced Learners*. Harper Collins Pub.

R. Snow, D. Jurafsky, and A. Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proc COLING/ACL 2006*, pages 801–808.

P. Turney. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proc. 12th European Conf. on Machine Learning*, pages 491–502.

Z. Wu and M. Palmer. 1994. Verb semantics and lexical selection. In *Proc. 32nd Annual Meeting of the ACL*, pages 133–138, New Mexico State University, Las Cruces, New Mexico.