Automatically Expanding the Lexicon of *Roget's Thesaurus*

Alistair Kennedy

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

Abstract. In recent years much research has been conducted on building Thesauri and enhancing them with new terms and relationships. I propose to build and evaluate a system for automatically updating the lexicon of *Roget's Thesaurus*. *Roget's* has been shown to lend itself well to many Natural Language Processing tasks. One of the factors limiting *Roget's* use is that the only publicly available version of *Roget's* is from 1911 and is sorely in need of an updated lexicon.

1 Introduction

Thesauri are valued resources for the Natural Language Processing (NLP) community, and have played a role in many applications including building lexical chains and text summarization. *WordNet* [1] has become the default thesaurus that NLP researchers turns to. It is important for NLP researchers to remember that *WordNet* represents just one of many ways of organizing the English lexicon and is not necessarily the best system available for a given NLP task. The 1911 version of *Roget's Thesaurus* (available through Project Gutenberg[1]) was recently released in a Java package called Open *Roget's Thesaurus*[2]. The goal of my thesis is to create an accurate system for automatically updating *Roget's Thesaurus* with new words.

Roget's is a hierarchical thesaurus consisting of nine levels from top to bottom: *Class* → *Section*→ *Subsection*→ *Head Group*→ *Head*→ *Part of Speech (POS)*→ *Paragraph*→ *Semicolon Group (SG)*→ *Words*. Words always appear in the lowest, 9th level, of the hierarchy. I will denote the set of words contained within one of these levels as a *Roget's-grouping*. SGs contain the closest thing to synonyms, though their grouping tends to be looser than synsets in *WordNet*.

2 The Methodology

This project is planned in three stages. The first is to identify pairs of closely related words using corpus based measures of semantic relatedness, such as Lin [2].

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

[2] rogets.site.uottawa.ca/

A. Farzindar and V. Keselj (Eds.): Canadian AI 2010, LNAI 6085, pp. 410–411, 2010.

Using a variety of these measures as features for a Machine Learning classifier I will determine which pairs of words are likely to appear in the same *Roget's-grouping* (specifically the same POS, Paragraph or SG).

The second step is to use these pairs of related words to determine the correct location in the *Thesaurus* to place a new word. Probabilities that pairs of words belong in the same *Roget's-grouping* can be used to determine the probability that a new word should be placed into a particular *Roget's-grouping.*

The last step, evaluation, can be done both manually and automatically. For manual evaluation an annotator could be given a *Roget's-grouping* and asked to identify the new words. If humans have difficulty in identifying which words are new additions then I can deem the additions to be as good as human additions. For automatic evaluation there are a number of applications that can be used to compare the original and updated *Roget's.* These tasks include measuring semantic distance between word or sentences [3] and ranking sentences as a component of a text summarization application [4].

3 Progress so Far

At this stage I have implemented a prototype system, of the first two steps, to place words into a *Roget's-grouping.* I performed evaluation of this prototype by removing a set of words from *Roget's* and attempted to place the words back into the *Thesaurus.* The early results show a relatively good precision for adding new terms at the Paragraph level, however the results are lower at the SG level. I hope to improve these results by experimenting with more semantic distance measures and Machine Learning classifiers. The above mentioned applications for automatic evaluation of *Roget's* have already been implemented [3,4]. As I have not yet produced an updated version of *Roget's* no manual or automatic evaluation has yet been carried out.

Acknowledgments

My research is supported by the Natural Sciences and Engineering Research Council of Canada and the University of Ottawa. I would also like to thank my thesis supervisor Dr. Stan Szpakowicz for his guidance.

References

1. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press, Cambridge (1998)
2. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th international conference on Computational linguistics, Morristown, NJ, USA, pp. 768–774. Association for Computational Linguistics (1998)
3. Kennedy, A., Szpakowicz, S.: Evaluating Roget's thesauri. In: Proceedings of ACL 2008: HLT, pp. 416–424. Association for Computational Linguistics (2008)
4. Copeck, T., Kennedy, A., Scaiano, M., Inkpen, D., Szpakowicz, S.: Summarizing with Roget's and with FrameNet. In: The Second Text Analysis Conference (2009)