Disambiguating Hypernym Relations for *Roget's Thesaurus*

Alistair Kennedy and Stan Szpakowicz
School of Information Technology and Engineering
University of Ottawa, Ottawa, Canada

Outline

- What is *Roget's Thesaurus*?
- Relationships in *Roget's Thesaurus*
- How to disambiguate relations?
  - Resources
  - Techniques for identifying hypernyms
- Evaluating the *Thesaurus*
  - Manual evaluation
  - Evaluation through applications

What is *Roget's Thesaurus*?

- Created in 1852 by Dr. Peter Mark Roget
- *Roget's Thesaurus* contains many English words and phrases organized into a hierarchy
  - We work with the 1987 Penguin version
  - Multiple word senses
- Thesaurus entries for:
  - nouns, verbs, adjectives and adverbs
- Words that are grouped together are closely semantically related --
  - -- but how exactly?
    - synonymy, hypernymy, hyponymy, meronymy...

Roget's Hierarchy

- Class
- Section
- Sub-Section
- Head Group
- Head
- Part of Speech (POS)
- Paragraph
- Semicolon Group
- Words and Phrases

Sample Head

[149] Revolution: sudden or violent change

N. revolution, full circle, circuit, (315) rotation; radical change, organic change; tabula rasa, clean slate, clean sweep, (550) obliteration; sudden change, catastrophe, peripeteia, surprise, coup d'etat, (508) lack of expectation; transilience, leap, plunge, jerk, start, throe, (318) spasm; shift, swing, switch, switch over, landslide; violent change, bouleversement, upset, overthrow, subversion, inversion, (221) overturning; convulsion, shake-up, upheaval, reorganizat ion, eruption, explosion, cataclysm, (176) outbreak; avalanche, landslip, crash, debacle, (309) descent, (165) havoc; revulsion, rebellion, counter-revolution, (148) reversion, (738) revolt; total change, sea change, metamorphosis, abolition, nullification, (752) abrogation, (752) deposal.

revolutionist, abolitionist, radical, revolutionary, Marxist, Red, (738) revolter; seditionist, (738) agitator; anarchist, (168) destroyer; idealist, (654) reformer.

ADJ. revolutionary (126) new; innovating, radical, thoroughgoing,out-and-out, root and branch, (54) complete; cataclysmic, catastrophic, seismic, earth-shaking, world-shaking, (176) violent; seditious, subversive, Marxist, red, (738) disobedient; anarchistic, (165) destructive.

VB. revolutionize, subvert, overturn, (221) invert; switch over, (603) tergiversate; uproot, eradicate, make a clean sweep, (550) obliterate, (165) demolish; break with the past, remodel, restructure, reorganize, refashion, (126) modernize; change the face of, change beyond recognition, metamorphose, (147) transform.

Comparison of *WordNet* and *Roget's Thesaurus*

- *WordNet* is built around a hypernym hierarchy of nouns and verbs
  - Nouns, verbs, adjectives and adverbs are separated
  - Semantic relations are explicitly labeled
  - Each term has a definition
- *Roget's Thesaurus* has a fixed 8-level hierarchy
  - Nouns, verbs, adjectives and adverbs are all contained in the same Head
  - Semantic relations are implied
  - Words are not explicitly defined

Explore the Problem: Manually Identify Relations

- Manually label relations in a sample of approximately 1% of noun semicolon groups
- Since a semicolon group can have many terms in it, it can also have many different relations (percentages do not add up to 100%)
- Only the top three relations are worth pursuing

Synonym	29.9%
Hypernym	55.5%
Coordinate Terms	56.0%
Meronymy	2.7%
Antonymy	0.3%
Causal	1.6%
Unknown	4.7%

Explore the Problem: Count Noun Hypernyms using *WordNet*

	Total
POS	57478
Paragraph	45481
Semicolon Group	15106

- About 80% of all relations found at the POS level can be found at the Paragraph level
- When disambiguating relations, select terms in the same *Roget's* paragraph

Resources for Disambiguating Hypernym Relations

- Existing lexical resources and ontologies
  - WordNet and Open Cyc
- Dictionaries
  - LDOCE and Wiktionary
- Corpora for identifying relations in text
  - BNC and Waterloo MultiText System

Identify Relations in Lexical Resources

- It is fairly simple to extract relations from WordNet
  - 53,404 relations extracted
- In OpenCyc, we worked with the “genls” hierarchy
  - Cyc regularly uses multiple inheritance
  - It is not really a lexical resource
    - It does not try to map every word to a concept, but uses words to describe concepts
  - Depth of 4
    - No significant advantage to going to a depth of 5
  - 1,608 relations extracted

Identify Relations From Dictionaries

- Two patterns, modified from Nakamura and Nagao (1988), are used to extract hypernyms
  - *{determiner} {adjective}* key noun*
  - *{determiner} {adjective}* function noun of key noun*
- Examples
  - LDOCE: biopic *is a(n)* film
  - LDOCE: police court *is a(n)* court of law
  - Wiktionary: concerto *is a(n)* piece of music
- 5,153 relations extracted from LDOCE
- 4,483 relations extracted from Wiktionary

Identify Hypernyms in Text

- Six patterns used by Hearst (1992)
  - *such NP as {NP, }* {(and | or)} NP*
  - *NP such as {NP, }* {(and | or)} NP*
  - *NP {NP, }* or other NP*
  - *NP {NP, }* and other NP*
  - *NP {,} including {NP, }* {(and | or)} NP*
  - *NP {,} especial ly {NP, }* {(and | or)} NP*
- The BNC
  - 1,332 relationships discovered

Waterloo MultiText Corpus

- Hearst's patterns were edited to search for hypernyms of specific words
  - such *NP* as *football*
  - *Protestant* and other *NP*
- This search was done for 26,430 words/phrases
  - for 15,443 words/phrases at least one query result was returned
- 11,392 relations ultimately were imported into *Roget's Thesaurus*

Co-occurring Hypernyms

- Some hypernym pairs appeared in more than one resource
  - 1 resource: 61581
  - 2 resources: 5839
  - 3 resources: 1102
  - 4 resources: 171
  - 5 resources: 21
  - 6 resources: 3
- Probability of error for each resource r
  - $P_e(r)$=1 - accuracy_for_the_resource

Overall Accuracy

- Error for each hypernym h
  - $P_e(h) = P_e(r1) * \ldots * P_e(r6)$
    for each resource in which *h* is found
- Accuracy P(h) for each hypernym pair is
  - $P(h) = 1-P_e(h)$
- The average accuracy across all hypernyms: 73%
- Most accurate hypernyms, examples:
  - drill *is a(n)* tool
  - crow *is a(n)* bird
  - cactus *is a(n)* plant

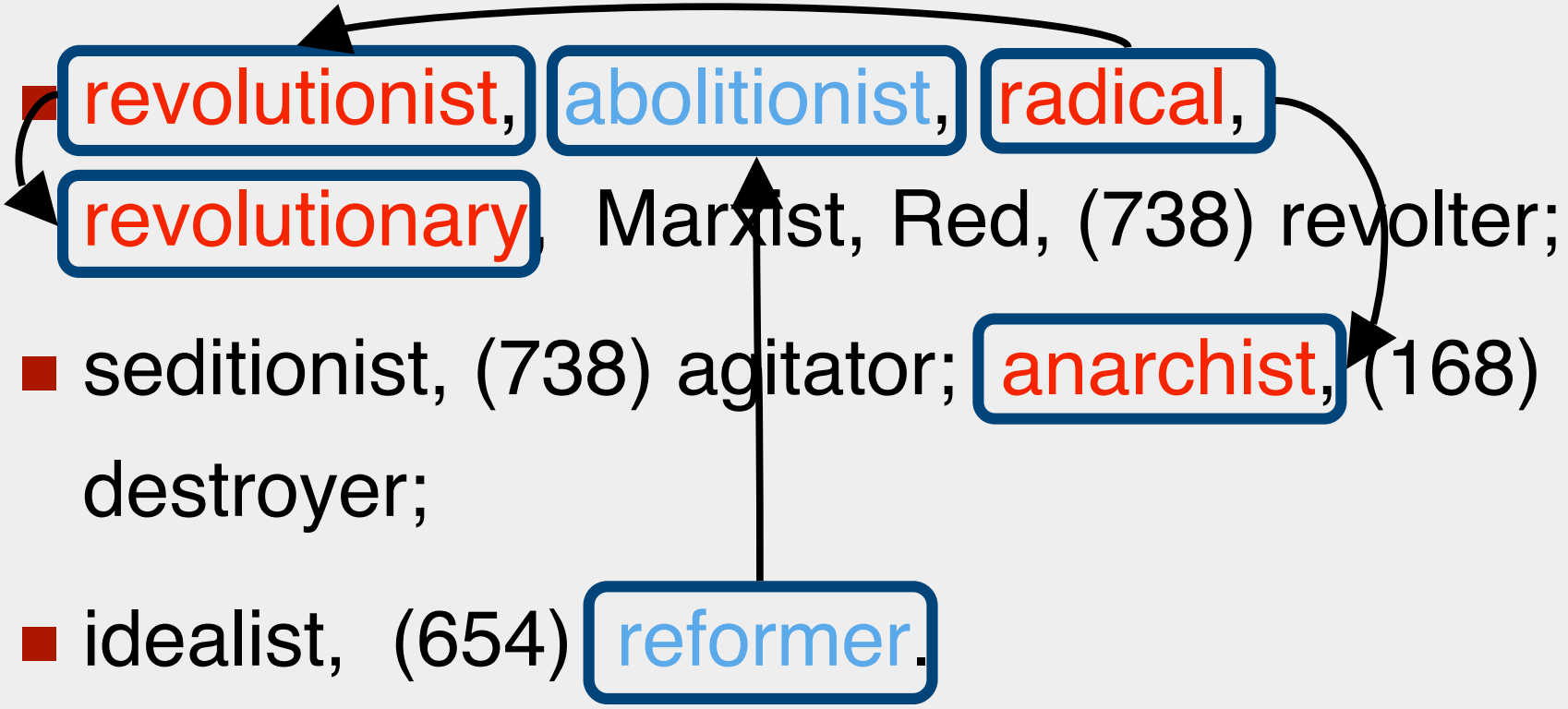
Remove Redundant Hypernym Links and Cycles

- There were 68,717 unique hypernyms and 92,675 total hypernyms
- Redundant hypernym links were direct links between two terms that were already indirectly linked as hypernyms
  - homogeneity *is a* uniformity *is a* sameness
  - redundant hypernym: homogeneity *is a* sameness
- The redundant links — 30,068 of them — were removed

Remove Cycles

- In a cycle, a term becomes its own hypernym
  - cycle: place *is a* rank *is a* position *is a* place
  - remove the hypernym with the lowest accuracy
    - place *is a* rank: 0.735 ←
    - rank *is a* position: 0.94223
    - position *is a* place: 0.782
- place *is a* rank is removed, leaving:
  - rank *is a* position *is a* place
- 3,756 cycles were removed
- 58,851 total relationships remained

Hypernym Examples



Manual Evaluation

Five judges received 200 samples per resource

- Labeled each sample as true or false hypernym
- (One judge completed part of the evaluation)

Resource	*Ave*	*Kappa*	*Count*	P	R	F
BNC	*.663*	*.436*	*1,332*	.663	.017	.034
CYC	*.865*	*.379*	*1,608*	.865	.021	.041
LDOCE	*.782*	*.267*	*5,153*	.782	.067	.123
WMT	*.536*	*.371*	*11,392*	.536	.147	.231
Wiki Hyp	*.726*	*.168*	*4,483*	.726	.057	.107
WordNet	*.735*	*.106*	*53,404*	.735	.690	.712

Why the low scores in WordNet?

- Some hypernyms could as well be synonyms
  - accused *is a* defendant
- Three hypernyms were always listed as incorrect
  - cup *is a* beverage
    - 7. (2) cup -- (a punch served in a pitcher instead of a punch bowl)
  - round *is a* line
    - 3. (1) beat, round -- (a regular route for a sentry or policeman; "in the old days a policeman walked a beat and knew all his people by name")
  - round *is a* whole
    - 1. (6) round, unit of ammunition, one shot -- (a charge of ammunition for a single shot)

Evaluation Through Applications

- Evaluation on a few applications:
  - term pair similarity
  - synonyms in a set of candidates
  - SAT analogies (inconclusive)
- These applications use a new semantic similarity function
- The old function (Jarmasz & Szpakowicz 2003) determined semantic similarity based on the levels of granularity in the thesaurus

Roget's Semantic Distance

- Similarity scores for two words in the same...
  - Thesaurus 0
  - Class 2
  - Section 4
  - Sub-Section 6
  - Head Group 8
  - Head 10
  - Part of Speech 12
  - Paragraphs 14
  - Semicolon Group 16
- This function is then modified using hypernyms

Enhanced Similarity Score

- To the original similarity score, add
  - 4 if connected by 1 hypernym/hyponym link
  - 3 if connected by 2 hypernym/hyponym links
  - 2 if connected by 3 hypernym/hyponym links
  - 1 if connected by 4 hypernym/hyponym links
- For each word with no hypernym/hyponym links, subtract 1 from the score
  - This is because words with no hypernyms or hyponyms is likely not as closely related to the other words in the paragraph
- Add 2 to the score to give a range of 0 to 22

Semantic Similarity

Data Set	Original	Enhanced
Miller & Charles (1991)	.773	.836
Rubenstein & Goodenough (1965)	.781	.838
Finkelstein et al. (2001)	.411	.435

Pearson product-moment correlation coefficient

Best Synonym

Command
- **Mastery**
- Observation
- Love
- Awareness

	Original	Enhanced
ESL	38/50 (3 ties)	42/50 (0 ties)
TOEFL	58/80 (5 ties)	58/80 (5 ties)
RDWP	201/300 (23 ties)	205/300 (13 ties)

SAT Analogies

- A is to B as C is to D
  - ostrich bird
    - **lion cat**
    - goose flock
    - ewe sheep
    - cub bear
    - primate monkey
- Use semantic similarity to discover analogies

$$dist = \left|semDist(A,B) - semDist(C,D)\right| + \frac{1}{semDist(A,C) + semDist(B,D) + 1}$$

- Hypernym matching

$$dist = dist - \left(8 - \left|hypernymDist(A,B) - hypernymDist(C,D)\right|\right)$$

Analogy Results

	Correct	Incorrect	Ties	Omit
Original	124	226	14	10
Original with Hypernyms	**130**	**220**	**14**	**10**
Enhanced	129	231	4	10
Enhanced with Hypernyms	**130**	**230**	**4**	**10**

Improves from 33.2% to 34.8%

Best system: 56% (Turney 2006)

Analogy, only hypernyms

	Correct	Incorrect	Ties
Original	7	15	2
Original with Hypernyms	13	9	2
Enhanced	13	10	1
Enhanced with Hypernyms	**14**	**9**	**1**

Improves from 29.2% to 58.3%

Conclusion

- People are not very good at agreeing on what is / is not a hypernym
  - (even *WordNet*...)
- Testing through applications shows that adding hypernym relationships can be useful for the *Thesaurus*
  - Semantic similarity
  - Showed improvement on a few applications

Future Work

- Incorporate new terms/phrases into *Roget's Thesaurus*
  - 1911 and 1987 versions
- Apply *Roget's Thesaurus* to other tasks
  - sentence relatedness
- Identify other relationships in the *Thesaurus*

References

- Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora..
- Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity.
- Jarmasz, M. and Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity.
- Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: the concept revisited.
- Turney, P. (2006). Similarity of semantic relations.
- Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge
- Lewis, M., editor (2000-2001). Readers Digest.
- Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL