Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion

Sneha Shankar Narayan
(snehas@cs.umass.edu)

Background

- Using a combination of thesaurus based approaches to expand queries in order to improve the performance of the retrieval system.

Thesaurus – Construction Methods

- Wordnet based thesaurus

$$sim_{path}(w1, w2) = max[-\log(\frac{N_p}{2D})]$$

  - Np : Number of nodes in path from w1 to w2
  - D: Maximum depth of taxonomy

Thesaurus – Construction Methods

- Co-occurence based thesaurus

$$I(a,b) = \log \frac{P(a,b)}{P(a)P(b)}$$

  - P(a), P(b): Estimated by counting the number of occurrences of words a and b.
- Combined approach
  - Average of the two similarities

Thesaurus - examples

- Wordnet based thesaurus
  - “Beethoven's birthplace” - expanded to “Beethoven birthplace origin sources composer”
- Co-occurence based thesaurus
  - “Best retirement country” - expanded to “best retirement country two pension project”

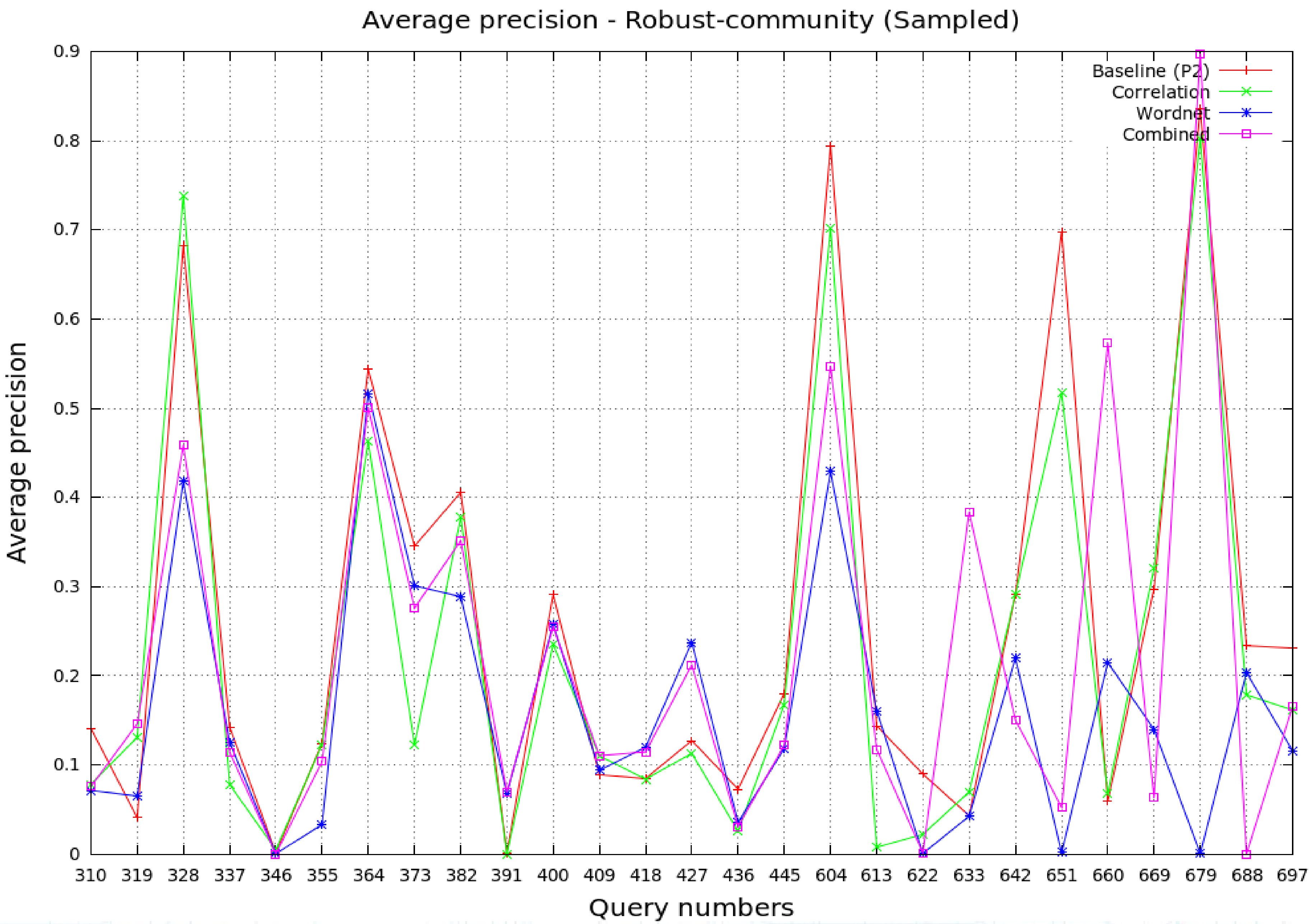
Evaluation Methodology

- Similar terms to query terms are determined for each of the approaches.
- Original query is concatenated with the determined terms and the experiments in P2 are repeated again. System used: Galago
- Ranked lists for the expanded queries are compared against the P2 baseline.

Results – Mean Average Precisions

	Baseline (P2)	Wordnet thesaurus	Co-occurence thesaurus	Combined approach
Books	0.3647	0.1954 (-46%)	0.2906(-20.3%)	0.2047 (-43.84%)
Robust-class	0.3910	0.2812 (-28%)	0.3811 (-2 %)	0.2879 (-26.3%)
Robust-community	0.2295	0.1549 (-32%)	0.1843 (-19.69 %)	0.1582 (-31.06%)

Results - Graphs



Observations

- Comparing average precisions show that the combined approach is not as bad as the MAP values show it to be.
- Main issue: Query words are considered as unique entities and expanded, expanding each word may not give the same connotation as all of the original query words put together.
  - Example: “brazil forest industry” becomes “brazil forest industry woods timberland manufactures”
- Words used while calculating similarities is restricted to the words in the top 50 ranked documents of the original query which may result in bad expansions.
- Future work: Perhaps the co-occurence can be tried with considering two of the query words at a time.

References

- "Combining multiple evidence from different types of Thesaurus for query expansion" by Rila Mandala, Takenobu Tokunaga, Hozumi Tanaka.