Graded Relevance Ranking for Synonym Discovery

Andrew Yates
Information Retrieval Lab
Department of Computer Science
Georgetown University
andrew@ir.cs.georgetown.edu

Nazli Goharian
Information Retrieval Lab
Department of Computer Science
Georgetown University
nazli@ir.cs.georgetown.edu

Ophir Frieder
Information Retrieval Lab
Department of Computer Science
Georgetown University
ophir@ir.cs.georgetown.edu

ABSTRACT
Interest in domain-specific search is steadfastly increasing, yielding a growing need for domain-specific synonym discovery. Existing synonym discovery methods perform poorly when faced with the realistic task of identifying a target term's synonyms from among many candidates. We approach domain-specific synonym discovery as a graded relevance ranking problem in which a target term's synonym candidates are ranked by their quality. In this scenario a human editor uses each ranked list of synonym candidates to build a domain-specific thesaurus. We evaluate our method for graded relevance ranking of synonym candidates and find that it outperforms existing methods.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – Thesauruses

General Terms
Algorithms, Experimentation, Measurement

Keywords
Synonym discovery; thesaurus construction; domain-specific thesaurus construction; domain-specific search

1. INTRODUCTION
Interest in domain-specific search, such as legal search (e-discovery) and medical search, is intensifying, promoting the need for automatic synonym discovery methods. For example, in the legal domain, domain-specific synonym discovery is important because an entity subject to e-discovery may use internal terms or acronyms that are synonyms only within the context of that entity. Contrary to prior synonym discovery efforts, we use graded relevance ranking to discover and rank a target term's potential synonyms. We do so without requiring difficult-to-obtain resources used by previously proposed methods, such as query logs [1] or parallel corpora [7] (i.e., a corpus available in two languages). These existing approaches are commonly based on distributional similarity, that is, the similarity of contexts in which terms appear [2, 5] or based on Web search results [6], which are too general for domain-specific synonym discovery. Our method favorably compares against the distributional similarity methods [2, 5].

A commonly used framework to evaluate synonym discovery methods uses synonym questions from TOEFL (Test of English as a Foreign Language). In this evaluation framework, the method is required to choose a target term's synonym from among four choices with only one correct synonym. While existing methods perform well when used to answer TOEFL questions with only one synonym (correct choice), existing methods falter when identifying a target term's synonyms from among many choices.

In a preliminary experiment, we created TOEFL-style questions from our medical domain dataset (described in section 3). These questions had 1,000 synonym candidates, rather than four, and we found that [2, 5] answered only 7% of the questions correctly when required to choose a target term's synonym from among all 1,000 candidates. We hypothesize that by treating domain-specific synonym discovery as a graded relevance ranking problem, we can improve the quality of synonym discovery. A target term's most general synonyms should be ranked above terms that may only be synonyms within limited contexts. By ranking high-quality candidates above low-quality candidates, the manual effort of a human editor is minimized.

2. METHODOLOGY
Baselines: We compare against two baselines. We implemented the best-performing method of Terra & Clarke [5]. This method calculates the pointwise mutual information (PMI) between the 16-term windows that a given pair of terms appear in. When used to rank synonym candidates, this method ranks them by the PMI between a target term and each candidate. We refer to this method as *Terra & Clarke.* We also implemented Hagiwara's method as described in [2], as well as an improved version we conceived based on the method, which performed better than the original method. Hagiwara's method creates a feature vector for each pair of terms and uses SVM to identify pairs of synonyms. The vector for terms w_i and w_j contains the feature $PMI(w_i, c_k) + PMI(w_j, c_k)$ for each context c_k. Contexts are based on the dependency structures identified in the text by the RASP3 parser[1]. The improved version of the method represents each term w_i as a vector containing the feature $PMI(w_i, c_k)$ for each context c_k and it identifies synonyms by calculating the cosine similarity between the target term's vector and the candidate terms' vectors. A target term's synonym candidates are ranked by their cosine similarity to the target term. We refer to this method as *Hagiwara (Improved).* Due to space limitations we do not report results for Hagiwara's original method; it consistently performed worse than *Hagiwara (Improved)* in both our preliminary TOEFL-style evaluation (section 1) and in our ranking evaluation (section 3).

Linear model: Our proposed method is a linear regression with three contextual features and one string similarity feature. We hypothesize that combining different types of features will result in a more robust synonym discovery method.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

[1] http://www.informatics.sussex.ac.uk/research/groups/nlp/rasp/

We construct one feature vector for each pair of terms <w_i, w_j>. Our features are: **(1)** the number of distinct contexts that both w_i and w_j appear in, normalized by the minimum number either term appears in. As with Hagiwara's method, contexts correspond to dependency structures identified by the RASP3 parser. **(2)** the number of sentences both w_i and w_j appear in, normalized by the minimum number either one appears in. **(3)** the cosine similarity between w_i and w_j as calculated by *Hagiwara (Improved)* - a feature that weights contexts by their PMI measures in contrast to the first feature (distinct context) that assigns them equal weights. **(4)** the word distance between terms w_i and w_j. Note that terms may contain multiple words (e.g., "sore_throat"). The word distance is calculated by finding the Levenshtein distance between two terms when words are treated as single units (instead of treating characters as single units as is normally done).

Utilizing SVM variants instead of a linear regression did not perform well, which may be caused by the high ratio of negative examples to positive examples. For brevity we forgo those results.

3. EVALUATION

Dataset: We used the MedSyn synonym list [8], a list of synonyms in the medical side effect domain, as our ground truth for synonyms. We used the UMLS metathesaurus' semantic network[2] to obtain hyponyms, hypernyms, and related terms for each synonym in MedSyn. We removed synonyms from the list that either did not occur or occurred only once in our corpus, as well as terms that had fewer than three synonyms, hyponyms, or hypernyms in total in our corpus. This removal process did not include related terms as a criterion because the other relationships were much less common. The remaining 1,384 synonyms were split into a training set (384 terms) for development (e.g., choosing features) and a testing set (1,000 terms) for evaluating the methods. Approximately 50% of the synonyms were phrases treated as a single term (e.g., "sore_throat") and the remaining 50% were single terms (e.g., "headache").

Our corpus, which was used by methods to discover synonyms, consisted of 400,000 posts crawled from the Breastcancer.org[3] and FORCE[4] forums. A complete list of the URLs crawled is available at http://ir.cs.georgetown.edu/data/www13_synonyms. While this dataset is focused on the medical side effect domain, our methods do not take advantage of any domain knowledge, and thus, should be applicable to any domain. We stemmed both the synonym list and corpus with the Porter stemmer [4].

Results: We evaluated each method's ability to rank a target term's synonyms by quality when provided with a list of non-synonym candidates and synonyms of varying quality (i.e., synonyms, related terms, hyponyms, and hypernyms). We use NDCG [3] to evaluate the ranked lists. A target term's synonyms were given a relevance score of 4, its related terms were given a 3, hyponyms a 2, hypernyms a 1, and non-synonym candidates (i.e., other terms) a 0; these scores correspond to the estimated likelihood that synonym candidates with a given relationship to the target term would be a synonym in some domains.

For each synonym in our testing set, we used each method to rank the term's synonym candidates (synonyms, related terms, hypernyms, and hyponyms) and *n* non-synonym candidates. We created an average NDCG from the ranked lists' NDCGs. We increased *n* to simulate progressively more realistic conditions in which a method is required to choose a target term's synonyms from among many non-synonyms. We used five-fold cross-validation with the supervised method (*Regression*).

The results are shown in Figure 1. Our method, *Regression*, consistently outperforms both *Terra and Clarke* and *Hagiwara (Improved)*. At *n*=20 *Regression* performs 8% better than the other two methods. As *n* increases and the task is made more realistic, *Regression* consistently outperforms the other methods. At *n*=1000 *Regression* outperforms *Terra and Clarke* by 9.5% and *Hagiwara (Improved)* by 35%.

4. CONCLUSION

We introduced synonym discovery as a ranking problem in which a target term's synonym candidates are ranked by their quality. Results demonstrate that our proposed method outperforms the existing methods. The fact that NDCG remains relatively high even at *n*=1000 supports our hypotheses that synonym discovery can be treated as a ranking problem and that a mixture of features can be used to create a robust synonym discovery method.

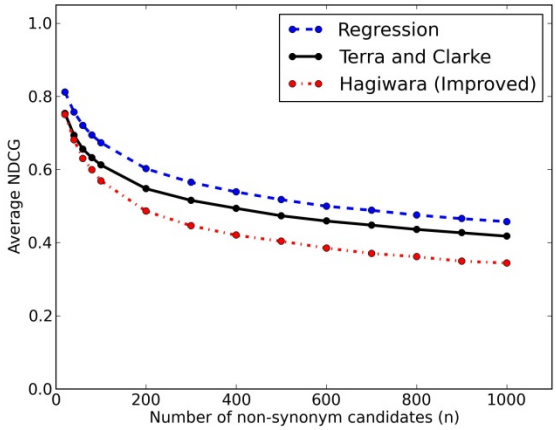

Figure 1: Ranking results

5. REFERENCES

[1] Grigonytė, G. et al. Paraphrase alignment for synonym evidence discovery. *COLING '10* (Aug. 2010), 403–411.

[2] Hagiwara, M. A Supervised Learning Approach to Automatic Synonym Identification based on Distributional Features. *HLT-SRWS '08* (2008).

[3] Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*. 20, 4 (Oct. 2002), 422–446.

[4] Porter, M.F. 1997. An algorithm for suffix stripping. *Readings in information retrieval*. 313–316.

[5] Terra, E. and Clarke, C.L.A. Frequency estimates for statistical word similarity measures. *HLT-NAACL '03* (2003).

[6] Turney, P.D. Mining the Web for Synonyms□: PMI-IR versus LSA on TOEFL PMI-IR. *EMCL '01* (2001).

[7] Wei, X. et al. Search with Synonyms□: Problems and Solutions. *COLING '10* (2010).

[8] Yates, A. and Goharian, N. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. *ECIR'13* (2013).

[2] http://www.nlm.nih.gov/research/umls/

[3] http://community.breastcancer.org/

[4] http://www.facingourrisk.org/messageboard/index.php