Context-based Quasi-Synonym Extraction

Van Dang, Xiaobing Xue and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
{vdang, xuexb, croft}@cs.umass.edu

ABSTRACT

Near- or quasi-synonyms can have an important role in the task of Information Retrieval since they may help to address the problem of vocabulary mismatch between queries and documents. One current approach to generating quasi-synonyms uses a general similarity measure to score the synonymy of two words based on their context vectors. In contrast, we compare two simple measures that take into account more directly the contextual evidence that supports a synonymy relation. Experimental results using the Google n-gram collection show that our methods produce better synonyms than existing approaches.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: General.

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Quasi-synonym extraction, similarity measure, vocabulary mismatch.

1. INTRODUCTION

Quasi-synonym extraction can have various applications in IR. Primarily, quasi-synonyms can be used for query modification to address the problem of vocabulary mismatch [6, 7]. Intuitively, words that can be used in the same context have a high chance to have similar meanings. Based on this, many methods have been proposed focusing on two aspects – what type of context to consider and what similarity measure to employ.

Grefenstette [1] and Lin [5] proposed using syntactic analysis of contexts. Since this approach can involve natural language processing techniques such as POS tagging, parsing, and morphology analysis, it is impractical for very large corpora.

Jing and Croft [4] represented the context of a word by the words that co-occurred with that word in text passages over the entire corpus. Quasi-synonyms were words with similar contexts. Other researchers [2, 3, 7] represented contexts as the distribution of words surrounding a given word, then used the KL divergence between two distributions as a measure of quasi-synonymy for two words. KL divergence is a popular measure yet has drawbacks for this application. For example, consider two context words "carries" and "points" in the following sentences: (1) "He **carries** a *gun* in the bag", (2) "He **carries** a *pistol* in the bag", (3) "He **points** his *gun* at us", and (4) "He **points** his *pistol* at us". Obviously, "*points*" is a better context than "*carries*" for determining that "*gun*" and "*pistol*" should be synonyms. Not only is how often a context word co-occurs with a term important, but also how many *different* terms a context word co-occurs with (henceforth, the *quality* of contexts) – one can "*carry*" many things, but not "*point*" that many. KL divergence, however, fails to capture the second factor.

In this paper, we propose two simple methods addressing the quality of contexts. The methods described in [2, 3, 7] are the most similar to ours, but there are two major differences: (i) while they consider only unigram on the left or the right of a given word as the context, we consider n-grams of variable length from one to four; and (ii) our formulas take into account the quality of contexts.

2. THE METHOD

We define the contexts of a word to be the n-grams around it, and use two approaches to compute the probability that a word is a quasi-synonym of another word ($P(w_2|w_1)$), as described below.

2.1 Formula 1

$$P(w_2 \mid w_1) = \frac{P(w_1 w_2)}{\sum_{w_i} P(w_i w_1)} = \frac{\sum_{c \in C} P(w_1 w_2 c)}{\sum_{w_i} P(w_i w_1)}$$

where C is the set of contexts that w_1 and w_2 have in common and $P(c) = \frac{1}{Z} \times \frac{1}{W}$ is the prior capturing the quality of a context. W is the number of distinct words going with the context c and Z is the normalization factor. We assume that given the context c, w_1 and w_2 are independent of each other. Thus,

$$P(w_2 \mid w_1) \propto \sum_{c \in C} P(w_1 \mid c) P(w_2 \mid c) P(c) \quad (1)$$

The idea is that the more contexts two words have in common, the more similar they are; and among those shared contexts, contexts of better quality contribute more to the similarity.

2.2 Formula 2

Since (1) only rewards two words if they share a context but does not penalize them when they disagree on a context (which means one word has a context that the other does not), we develop a second formula,

$$P(w_2 \mid w_1) \propto \frac{1}{\log_2(N)} \sum_{c \in C'} P(c)$$

where *P(c)* is the same as in (1) and *N* is the number of contexts that w_1 and w_2 disagree. *S*hared contexts in *C* that share their head word are put together to form the group *C'*. Head words are obtained by sorting all words extracted from all contexts in *C* descending according to their frequency in *C*; contexts that share the first head word in the list form a group C_1, the one with highest *P(c)* of which is put into *C'*. Contexts in C_1 are removed from *C* and the process is iterated until we go through the head word list. The idea of using *C'* is to avoid overcounting. For example, three contexts "who **drives** a <*car*>", "is **driving** a <*car*>", "I **drive** my <*car*>" for the word "*car*" should be counted once instead of three times.

3. EVALUATION

3.1 Experimental Setup

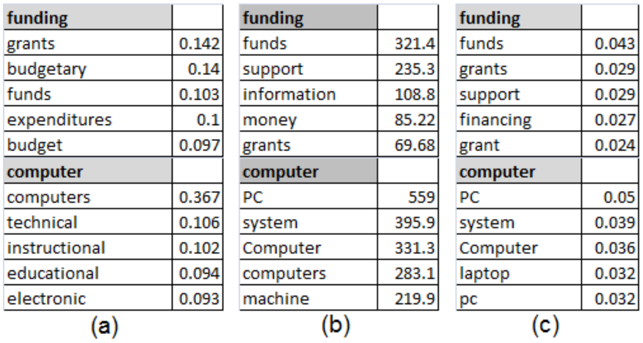
Since it is hard to evaluate extracted synonyms automatically, we used manual evaluation in this study. We manually picked 50 words from TREC queries that we think would have reasonable impact on search effectiveness. Then we used the K-L method [2, 3, 7] and our two proposed methods to find synonyms for these 50 words from the Google n-gram collection. The top ten synonyms for each word were recorded and evaluated manually. Each candidate pair was judged on a scale where 3 means "definitely a synonym", 1 means "definitely not a synonym" and 2 means something in between. The average DCG (Discounted Cumulative Gain) score is used to compare methods.

In [7], both the left and right unigram contexts were used. In our case, however, we found that right contexts were not helpful and we only show the results using left contexts. In our two methods, we considered left (L), right (R) and left-right (LR – left and right at the same time) contexts. Specifically, we used ***n***-gram with ***n*** up to 4 for *L* and *R*, and ***n*** up to 2 for *LR*.

We evaluated each method with and without stop word removal. The Porter Stemmer was used in constructing the group *C'*.

3.2 Results

Fig. 1 shows top 5 synonyms for "*funding*" and "*computer*" extracted by three systems.



funding	
grants	0.142
budgetary	0.14
funds	0.103
expenditures	0.1
budget	0.097
computer	
computers	0.367
technical	0.106
instructional	0.102
educational	0.094
electronic	0.093

(a)

funding	
funds	321.4
support	235.3
information	108.8
money	85.22
grants	69.68
computer	
PC	559
system	395.9
Computer	331.3
computers	283.1
machine	219.9

(b)

funding	
funds	0.043
grants	0.029
support	0.029
financing	0.027
grant	0.024
computer	
PC	0.05
system	0.039
Computer	0.036
laptop	0.032
pc	0.032

(c)

Figure 1- Top-5 synonyms for the word "*funding*" and "*computer*" extracted by three systems. (a) KL (b) Formula 1 (c) Formula 2. Scores are not normalized.

DCG score of each system averaged over 50 words is shown in Table 1. We use the term "KL divergence" to indicate the method described in [2, 3, 7].

Table 1. Performance comparison among three systems in terms of Average DCG.

System	Average DCG	
	w stop word	w/o stop word
KL Divergence	2.6387	5.3925
Formula 1	4.0595	4.3076
Formula 2	5.8181	5.8601

Table 1 shows that both of our methods outperform KL divergence, showing that the prior works well for controlling the quality of contexts. We can also see that our second method is significantly better than the first one. This is due to formula 1 over-rewarding some candidate pairs. An example of this is given in Table 2, which shows the top 5 shared contexts (those with highest values of *P(c)*) of two candidate pairs "*funding – information*" and "*funding – financing"*. Though "*funding – information*" shares many contexts, most of them are related to "source-" while "*funding – financing*" shares contexts with different head words. Therefore, counting all shared contexts as in Formula 1 results in "*information*" being ranked higher than "*financing*".

Table 2. Top 5 shared contexts of two candidate pairs.

funding – information	funding – financing
sources of <WRD> from within	has secured <WRD> of $
source of <WRD> that enables	on state <WRD> of religious
resources and <WRD> sources .	Series A <WRD> round of
source of <WRD> to allow	analize the <WRD> bundles ,
sources of <WRD>)	benefits and <WRD> solutions ,

4. CONCLUSION AND FUTURE WORK

In this paper, we propose two simple methods to extract quasi-synonyms from text corpora. Both new methods take into account the quality of contexts, which previous research using KL divergence does not. The experimental results show that the new methods outperform the KL divergence approach.

In future work, we will evaluate our methods with more words and use the extracted synonyms in an IR system via a query modification technique. This will help us both to evaluate the quality of synonyms automatically and study the effect of synonyms on retrieval effectiveness.

5. ACKNOWLEDGMENT

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #IIS-0711348, in part by NSF CLUE IIS-0844226, and in part by NSF grant #IIS-0534383. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In Proceedings of SIGIR, pages 89-97, 1992.

[2] F. Pereira, N. Tishby and L. Lee. "Distributional Clustering of English Words". In Proceedings of ACL, pages 183-190, 1993.

[3] I. Dagan, F. Pereira and L. Lee. Similarity-Based Estimation of Word Cooccurrence Probabilities. In Proceedings of ACL, pages 272-278, 1994.

[4] Y. Jing and W. B. Croft. An Association Thesaurus for Information Retrieval. In Proceedings of RIAO, pages 146-160, 1994.

[5] D. Lin. Automatic retrieval and clustering of similar words. In COLING-ACL, pages 768-774, 1998.

[6] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst., pages 79-112, 2000.

[7] X. Wang and C. Zhai. Mining term association patterns from search logs for effective query reformulation. In Proceedings of CIKM, pages 479-488, 2008.