Disambiguating Biomedical Acronyms using EMIM

Nut Limsopatham, Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis
{nutli,rodrygo,craigm,ounis}@dcs.gla.ac.uk

School of Computing Science
University of Glasgow
G12 8QQ, Glasgow, UK

ABSTRACT

Expanding a query with acronyms or their corresponding 'long-forms' has not been shown to provide consistent improvements in the biomedical IR literature. The major open issue with expanding acronyms in a query is their inherent ambiguity, as an acronym can refer to multiple long-forms. At the same time, a long-form identified in a query can be expanded with its acronym(s); however, some of these may be also ambiguous and lead to poor retrieval performance. In this work, we propose the use of the EMIM (Expected Mutual Information Measure) between a long-form and its abbreviated acronym to measure ambiguity. We experiment with expanding both acronyms and long-forms identified in the queries from the adhoc task of the TREC 2004 Genomics track. Our preliminary analysis shows the potential of both acronym and long-form expansions for biomedical IR.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Experimentation, Performance

Keywords: Biomedical Retrieval, Acronym Expansion

1. INTRODUCTION

The frequent use of (non-standardised) acronyms is one of the major problems in biomedical information retrieval [1]. In particular, the terms of a 'long-form' that has been abbreviated with an acronym in a document will have lower term frequencies in that document, and hence the document will less likely be retrieved for a query with that long-form. Moreover, acronyms may have different meanings in different documents. For example, the acronym "AD" is often used in the biomedical context to refer to "Alzheimer's disease". However, according to the ADAM database of biomedical acronyms [5], "AD" can refer to 35 unique long-forms. Therefore, if a query containing "AD" is expanded with all of the corresponding long-forms, the other 34 unrelated long-forms would result in irrelevant documents being retrieved. Conversely, if another query containing "Alzheimer's disease" is expanded with "AD", the other 34 meanings of "AD" would result in irrelevant documents being retrieved as well.

These difficulties may explain the inconsistent conclusions in the literature for acronym expansion. For instance, Stokes et al. [3] used a pseudo-relevance feedback strategy to expand the query acronyms using the long-forms found in the retrieved documents. However, they reported that their approach was insufficient to deal with the ambiguity of acronyms. Similarly, Zhou et al. [6] used the ADAM acronym database to expand long-forms in the query with the corresponding acronyms as query concepts for their concept-based retrieval framework. Still, they encountered problems with the ambiguity of acronyms and gene names. In contrast, Büttcher et al. [1] were able to successfully improve retrieval performance by expanding acronyms in queries with their corresponding long-forms. However, their approach is based on a simple heuristic, which ignores the strength of the relationship between an acronym and its long-form.

To cope with acronym ambiguity, we propose to infer the ambiguity of acronyms during the expansion process. In particular, our proposed approach, described in Section 2, expands both the acronyms and long-forms in a query, and uses EMIM to measure the ambiguity of the acronyms. In Section 3, we evaluate the proposed approach in the context of the adhoc task of the TREC 2004 Genomics track. Finally, conclusions are given in Section 4.

2. ACRONYM DISAMBIGUATION

Our approach comprises three steps. Firstly, following Schwartz and Hearst [2], we build a dictionary of acronym and long-form pairs automatically from the target corpus used in our investigations, as described in the next section. Table 1 shows the statistics of the generated dictionary.

total # of acronyms	89,052
total # of long-forms	300,905
avg. # of acronyms per long-form	0.2959
avg. # of long-forms per acronym	3.3790

Table 1: Statistics of the expansion dictionary.

Next, we use this dictionary to identify 'triggers' in a query. A trigger can be either an acronym or a long-form. For each trigger, we match 'candidate expansions' (either long-forms or acronyms, respectively) in the dictionary. As there may be multiple candidate expansions for a given trigger, we propose to weight different candidates based on their probability of co-occurring with the trigger in the target corpus. It is intuitive that the more an acronym and a long-form co-occur, the more likely that the acronym refers to the long-form exclusively, and hence can safely be used as an alternative for the long-form. In particular, van Rijsbergen [4] proposed to derive the level of dependence between terms from the distribution of co-occurrences in a document set, which can be measured by EMIM (Expected Mutual

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

Candidate Expansion		Our approach	Min-occur [1]
Acronyms	Long-forms		
✔	✗	38	23
✗	✔	24	15
✔	✔	42	33

Table 2: Topics affected by acronym expansion.

Information Measure). Therefore, we use EMIM to measure the co-occurrence between a trigger and its candidate expansion in the corpus. EMIM is calculated as:

$$\text{EMIM}(tr, ce) = \log \frac{\text{P}(tr, ce)}{\text{P}(tr)\,\text{P}(ce)} \quad (1)$$

where tr is a trigger, and ce is a candidate expansion. P is the maximum likelihood estimation function, while $\text{P}(tr, ce)$ is the joint probability of tr and ce, estimated as the fraction of documents where they co-occur.

Finally, the calculated EMIM of each trigger and candidate expansion pair is integrated into the retrieval score of a document for a query as follows:

$$score(d, Q) = \sum_{t \in Q} score(d, t) \quad (2)$$
$$+ \lambda \cdot \sum_{\langle tr, ce \rangle \in matches(Q)} \text{EMIM}(tr, ce) \sum_{t' \in ce} score(d, t')$$

where λ is a parameter to weight the score of the expanded terms, $matches(Q)$ uses the dictionary to calculate a set of pairs $\langle tr, ce \rangle$, such that tr is a trigger in the query Q and ce is a corresponding (acronym or long-form) candidate expansion, and t' is a term of ce.

3. EXPERIMENTAL RESULTS

We evaluate our proposed acronym expansion approach using the 50 title-only topics from the adhoc task of the TREC 2004 Genomics track.[1] This task uses a corpus of 4.6M MEDLINE abstracts. We index this corpus using the Terrier[2] information retrieval platform, with Porter's stemming and removing stopwords. For retrieval, we use the Divergence From Randomness DLH weighting model. DLH is a parameter-free model, hence no training is required.

In Figure 1, we show the retrieval performance, in terms of mean average precision (MAP), of different acronym expansion approaches as we vary the expansion weight λ in Equation (2). In particular, we evaluate our approach using EMIM to add acronyms, long-forms, or both to the queries. As a baseline, we consider the approach of Büttcher et al. [1]—henceforth referred to as 'min-occur'—which expands the queries with long-forms that co-occur in at least five documents with the trigger acronyms. Additionally, as the min-occur approach only expands acronyms, we extend it to also expand long-forms with acronyms. Table 2 shows the number of queries impacted by the expansion approaches—in all cases, our approach expands more queries. In addition to the min-occur baseline, we consider a simple baseline that performs no expansion (i.e. λ=0).

From Figure 1, we first observe that our approach can substantially outperform the no-expansion and the min-occur expansion baselines for an appropriate setting of λ. Indeed, for acronym expansion, improvements are observed for λ in the range [0, 0.06], and are as high as 0.91% compared to no-expansion. In contrast, min-occur achieves an improvement of 0.58% in the same scenario. For long-form expansion, the ideal range for our approach is [0, 0.03], with gains up to 4.48% compared to no-expansion. In comparison, min-occur improves by 1.50 and 2.08% when adding long-forms or both acronyms and long-forms, respectively. Finally, while min-occur performs effectively for a wide range of λ values (particularly when adding long-forms), it can only improve over no-expansion by 2.08% in its best setting when expanding both acronyms and long-forms. In turn, for the same scenario, although our approach has a comparatively narrower range of effective λ values, its potential improvement is as high as 5.20%. Moreover, the effective range of λ values is stable across the three variants of our approach, which shows that it can be easily tuned in a deployment scenario.

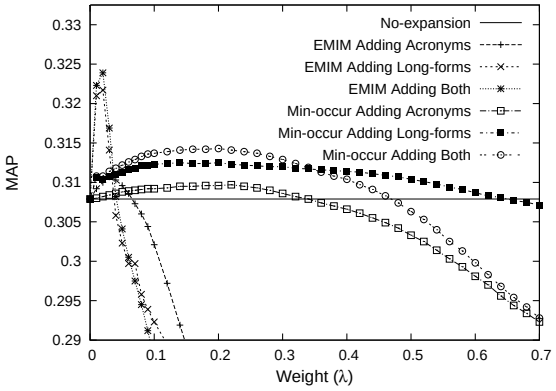

Figure 1: Acronym expansion effectiveness for a range of expansion weights (λ).

4. CONCLUSIONS

We have proposed to improve acronym expansion for biomedical IR by disambiguating candidate expansions using EMIM. Our results show that the proposed approach can potentially outperform existing approaches in the literature, without requiring other techniques (e.g. synonym expansion, pseudo-relevance feedback) or external resources (e.g. ADAM). In the future, we plan to further investigate alternative mechanisms for estimating the probability of co-occurrence of acronyms and long-forms.

5. REFERENCES

[1] S. Büttcher, C. L. A. Clarke, and G. V. Cormack. Domain-specific synonym expansion and validation for biomedical information retrieval (MultiText experiments for TREC 2004). In *Proc. of TREC*, 2004.

[2] A. S. Schwartz and M. A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proc. of PSB*, pages 451–462, 2003.

[3] N. Stokes, Y. Li, L. Cavedon, and J. Zobel. Exploring abbreviation expansion for genomic information retrieval. In *Proc. of ALTW*, pages 100–108, 2007.

[4] C. J. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *J. Documentation*, 33(2):106–199, 1977.

[5] W. Zhou, V. Torvik, and N. Smalheiser. ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*, 22(22):2813–2818, 2006.

[6] W. Zhou, C. Yu, and W. Meng. A system for finding biological entities that satisfy certain conditions from texts. In *Proc. of CIKM*, pages 1281–1290, 2008.

[1] We do not consider the TREC 2005 Genomics collection, as its topics include both acronyms and long-forms.

[2] `http://terrier.org`