Mitigating Vocabulary Mismatch on Multi-domain Corpus using Word Embeddings and Thesaurus

Nagesh Yadav, Alessandro Dibari, Miao Wei, John Segrave-Daly, Conor Cullen, Denisa Moga, Jillian Scalvini, Ciaran Hennessy, Morten Kristiansen and Omar O'Sullivan

International Business Machines, U.S.A.

nyadav@ie.ibm.com, adibari@us.ibm.com, miao.wei@ibm.com, john.segrave@ie.ibm.com, conor.cullen@ie.ibm.com, denimoga@ie.ibm.com, jscalvin@us.ibm.com, ciaran.hennessy@ie.ibm.com, morten@ie.ibm.com, oosulli@ie.ibm.com

Keywords: Information Retrieval, Query Expansion, Word Embedding, Thesaurus.

Abstract: Query expansion is an extensively researched topic in the field of information retrieval that helps to bridge the vocabulary mismatch problem, i.e., the way users express concepts differs from the way they appear in the corpus. In this paper, we propose a query-expansion technique for searching a corpus that contains a mix of terminology from several domains - some of which have well-curated thesauri and some of which do not. An iterative fusion technique is proposed that exploits thesauri for those domains that have them, and word embeddings for those that do not. For our experiments, we have used a corpus of Medicaid healthcare policies that contain a mix of terminology from medical and insurance domains. The Unified Medical Language System (UMLS) thesaurus was used to expand medical concepts and a word embeddings model was used to expand non-medical concepts. The technique was evaluated against elastic search using no expansion. The results show 8% improvement in recall and 12% improvement in mean average precision.

1 INTRODUCTION

When searching for information, users often express their search queries in a way that differs significantly from how that information is presented in the corpus being searched. For example: a user may search for medical insurance information such as *'Flu jab coverage for children'*, and be interested in results (i.e., snippet from a document) like *'Influenza immunizations are a benefit for both child and adult members...'*. Here, the desired result matches little or none of the vocabulary used in the query. Some of these differences are straightforward – e.g.,in medicine, the terms 'Flu' and 'Influenza' are synonymous. Others are contextual – e.g., in Insurance, the terms 'coverage' and 'benefits' are semantically similar (in contrast to Project Management, where they have distinctly different meanings). This is a standard problem faced in Information Retrieval (IR) systems referred to as vocabulary mismatch problem (Xu and Croft, 2017).

Query expansion is a commonly-used approach to mitigating the vocabulary mismatch problem (Carpineto and Romano, 2012)(Crimp and Trotman, 2018)(Kuzi et al., 2016). IR systems commonly 'expand' users' searches to include synonyms, and semantically-related terms, such as those in the Flu example above. Many domains have well-established thesauri that enable this type of query expansion. The NIH Unified Medical Language System (UMLS) metathesaurus is a well-curated example applicable to Medical and Healthcare domain (Bodenreider, 2004). However, thesauri alone are not sufficient to bridge the vocabulary mismatch gap. First, they often have gaps in their vocabulary – e.g., due to infrequent updates. Secondly, to look up expansions in a thesaurus, an IR system must first match n-grams in the users' query to concepts in that thesaurus ('concept detection'). Differences between the user's representation of that concept in the query (e.g., 'flu jab') and the thesaurus representation ('Influenza virus vaccine') can cause detection to fail and thus, no expansions to be identified. Finally, many domains lack high-quality thesauri, or those available are not well-suited to Information Retrieval tasks. e.g., in Insurance, there are no widely-adopted thesauri to indicate that terms like 'PAR' and 'prior authorization' are synonymous, while 'covered', 'benefit' and 'available' share similar meanings. In this paper, we address the aforementioned problems by combining thesauri and word embeddings, as both have useful properties for addressing these problems - e.g., by representing

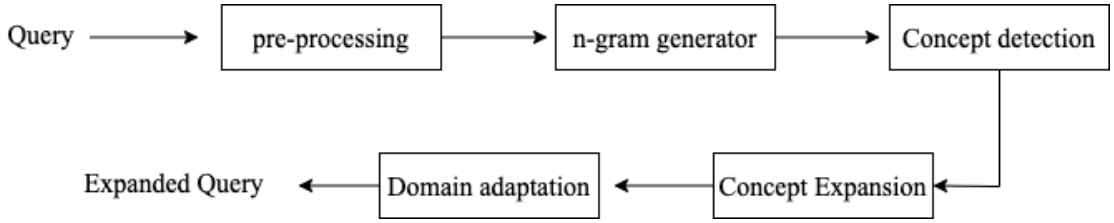

Figure 1: Overview of our approach.

words as learned vectors (distributed representations), proximity can be used as a rough approximation of semantic similarity.

2 APPROACH

2.1 Overview

In this paper we propose an iterative query expansion technique, an overview of which is depicted in this pipeline (figure 1). First, the user query is pre-processed, including stop-word removal and case normalization. An n-gram generator takes the normalized query and creates a list of all possible n-grams, starting from the largest and iteratively reducing to individual words. Each of these candidate n-grams is passed through a concept expansion module to get a list of candidate expansions. This module is described in section 2.3 and exploits a combination of thesauri and word embeddings. Some of these candidate expansions are noisy - terms that appear in the thesaurus, but are not relevant to the corpus in question. For example, medical research terms that never appear in an insurance policy corpus. To address this, all expansions go through 'domain adaptation' where they are cross-checked against the corpus linguistics using word embeddings, as described in section 2.4. A query is then generated and submitted to the search engine.

2.2 Concept Detection

The Concept Detection module is responsible for mapping a candidate n-gram to a known concept in the domain thesaurus, if such a match exists. In our application, UMLS is used as a thesaurus of known medical/healthcare concepts. As depicted in figure 2, concept detection is first attempted by looking-up the candidate n-gram in the thesaurus. If matched, the concept is retrieved from the thesaurus and passed onto to the concept expansion module. In our application, these concepts are represented by Concept Unique Identifiers (CUIs) from the UMLS thesaurus. If the n-gram cannot be matched by a simple lookup, a second attempt is made - this time exploiting word embeddings. Potential synonyms and other semantically-related terms for the n-gram are derived from a word embedding model. Since word embeddings provide only a rough approximation of semantic similarity, a high similarity threshold is used to mitigate the risk of noisy/irrelevant terms. (Optimal values for this threshold are determined experimentally - e.g., for the Medicaid IR system in our case). At this point, a second attempt at concept detection can be made by looking-up these embedding-derived terms in the thesaurus. If matched, the concept is retrieved from the thesaurus as before and passed onto the concept expansion module.

To illustrate the concept detection process - suppose the user query is *Panoramic x-ray coverage*. Within this query is the bi-gram *panoramic x-ray*. This bi-gram is looked up in the UMLS thesaurus, where no matching concept can be found. Semantically-related terms for this bi-gram are derived from a word embedding model, yielding the new term *panoramic radiography*. The embedding-derived term is again looked up in the UMLS thesaurus, and this time a matching concept is found (e.g., a UMLS CUI). Once the concepts have been detected, the thesaurus can then be exploited to expand them, as described in the following sections.

After concept detection, a single concept can still be ambiguous. More formally, a concept (initially represented only by its ngram) is ambiguous when after concept detection phase there are more than one Concept Unique Identifier (CUI) matched in UMLS against that term (n-gram). For our specific application wherein we are interested in medical and healthcare entities, we are exploiting UMLS Semantic Groups to solve these ambiguities in most cases. Finally, when concept detection is complete, a numerical confidence score is assigned to every identified concept. This number is a function of the longest n-grams matched, with respect to the given user query.

$$C = f(NG_{max}, UQ) \tag{1}$$

where NG_{max} is the number of terms in the longest concept and UQ is the number of terms in the original user query.

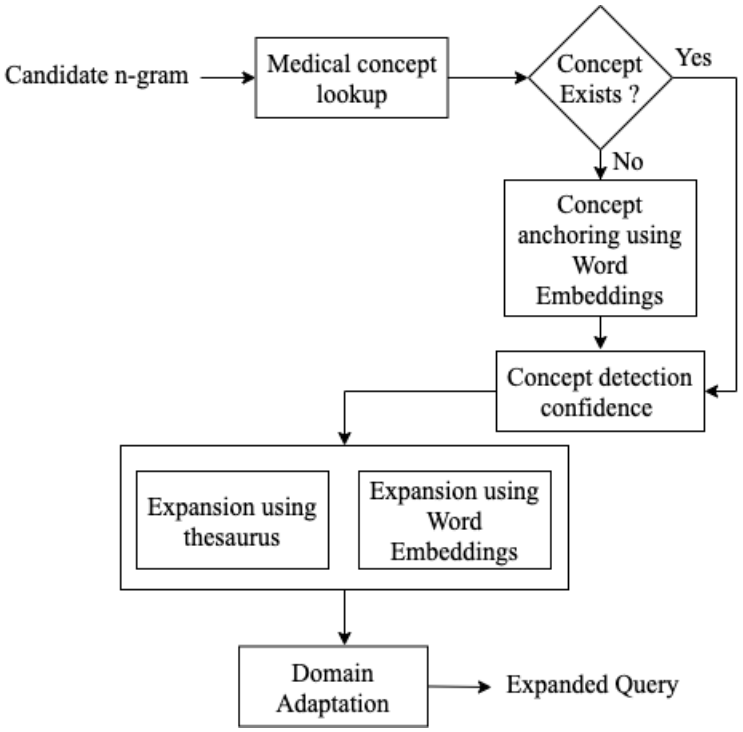

Figure 2: Query Expansion Flow.

2.3 Concept Expansion

The Concept Expansion module is responsible for 'expanding' the terms in a user query with synonymous or closely semantically-related terms. This is straightforward for simple, single-concept queries that the concept detection module already matched in a thesaurus. e.g., in our system, the CUI identified for a concept is used to look up variant and synonymous terms in the UMLS thesaurus.

However, queries that contain a mix of concepts from different domains are less straightforward, particularly where some of those domains have a thesaurus and others do not. To address this, we combine expansions from both thesaurus and corpus-specific word embeddings, using the confidence score described in section 2.2 to determine the relative contribution of each. The most trustworthy information source is the thesaurus, which often has decades of use and curation (e.g., UMLS). So when most of the terms expressed in a user query can be mapped to concepts in a thesaurus, then the confidence score is high. In this case, expansions are primarily contributed by the thesaurus, and less emphasis is placed on word-embedding based expansions. In contrast, when most of the concepts in a user query cannot be found in the thesaurus, then the confidence score will be low. In this case, expansions are primarily contributed by a corpus-specific word embedding model, with only a few (or none) contributed by the thesaurus. This word embedding is learned from the corpus being searched, to avoid introducing additional vocabulary gaps - i.e., gaps between the vocabulary on which the word embedding was trained and that of the corpus being searched. Expansions are obtained from this model using the same semantic similarity threshold as before. In our application, Word2Vec was used to train this word embedding model from a corpus of Medicaid policy documents (Le and Mikolov, 2014). Once obtained, all expansions are sent for domain adaptation before being submitted to the search engine.

2.4 Domain Adaptation

The Domain adaptation module reduces noise in the search results by ensuring that expansions are only used when they relate to the actual terminology used in the corpus. Some thesaurus-based expansions may be removed at this stage, if they are found not to be relevant for the particular corpus being searched. For each expansion resulting from the previous module, we calculate its cosine similarity with the original user query term it was derived from. This ensures that the set of expansions that are closer to the corpus (measured in terms of semantic distance) and

Table 1: Search metrics.

Metric	Without expansion	With expansion	% improvement
Mean Average Precision	0.63	0.71	12
Precision(@Rank 10)	0.49	0.53	8
Recall(@Rank 10)	0.62	0.67	8
NDCG	0.79	0.83	5

also closer to the original user query are ranked higher than those which are further apart.

3 RESULTS AND DISCUSSION

The proposed technique was tested in IBM Policy Insights, an application that helps Medicaid Fraud, Waste and Abuse (FWA) investigators find policy information relevant to their investigations.

For Ground Truth, 3 subject matter expert (SME) investigators identified a corpus of 345 Medicaid policy documents. They also defined 80 queries that were representative of those typically used in their investigations, ranging from single words to phrase queries. Every document in the corpus was labelled as either *relevant* or *not-relevant* for each query. The corpus, queries and relevancy labels were used as ground truth for measuring document retrieval performance. To evaluate the performance, the following metrics were used, Precision, Recall, Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR) (Buttcher et al., 2016).

The underlying search engine used was Watson Discovery Service (WDS) (Turner, 2016), using its out-of-the-box elastic search capabilities only. A WDS collection was created for the corpus of 345 policy documents. UMLS was used as a thesaurus. A word embeddings model was trained on this corpus, using word2vec (Le and Mikolov, 2014). To evaluate baseline document retrieval performance, all 80 queries were run against WDS' out-of-the-box Natural Language Query (NLQ) API and the above performance metrics calculated for the returned documents. For our experiment, the same setup was used, with the 80 queries passed through our proposed query expansion pipeline in advance of being sent to the NLQ search engine. The same performance metrics were again calculated for the returned documents and these were compared to the baseline performance.

Table 1 summarizes the comparison of key metrics. From the average performance over 80 queries, it can be concluded that recall and precision are both improved by our proposed query expansion pipeline. We also observed that queries containing healthcare concepts matched by our thesaurus (as described in section 2.2) showed more significant improvements than those that did not. This is expected, due to the nature of our corpus and available thesauri i.e., UMLS contains high-quality human-curated knowledge. We conclude that combining domain knowledge from human-curated thesauri and automatically-learned word embeddings enhances document retrieval performance on a corpus containing a mix of terminology from several domains.

4 FUTURE WORK

The proposed technique was tested on a corpus of healthcare policy documents. Further testing needs to be done on different domains to assess the generalization of the proposed technique. Specifically, the approach needs to be tested on domains with a less detailed human-annotated thesaurus (such as insurance sector). Furthermore, metadata of detected concepts such as semantic groups will be exploited in future iterations for concept disambiguation.

REFERENCES

Bodenreider, O. (2004). The unified medical language system: integrating biomedical terminology. In *Nucleic acids research*, volume 32, pages D267–D270. Oxford University Press.

Buttcher, S., Clarke, C. L., and Cormack, G. V. (2016). Information retrieval: Implementing and evaluating search engines. MIT Press.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. In *ACM Comput. Surv.*, volume 44, pages 1:1–1:50, New York, NY, USA. ACM.

Crimp, R. and Trotman, A. (2018). Refining query expansion terms using query context. In *Proceedings of the 23rd Australasian Document Computing Symposium*, ADCS '18, pages 12:1–12:4, New York, NY, USA. ACM.

Kuzi, S., Shtok, A., and Kurland, O. (2016). Query expansion using word embeddings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1929–1932, New York, NY, USA. ACM.

Le, Q. and Mikolov, T. (2014). Distributed representations

of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Turner, E. (2016). Watson discovery service: understand your data at scale with less effort. In *IBM Developer Works, https://www.ibm.com/blogs/watson/2016/12/watson-discovery-service-understand-data-scale-less-effort*.

Xu, J. and Croft, W. B. (2017). Query expansion using local and global document analysis. In *ACM SIGIR forum*, volume 51, pages 168–175. ACM.