Query Expansion for Cross-Language Question Re-Ranking

Muhammad Mahbubur Rahman Sorami Hisamoto Kevin Duh
mahbubur@jhu.edu,s@89.io,kevinduh@cs.jhu.edu
Johns Hopkins University
Baltimore, MD, USA

ABSTRACT

Community question-answering (CQA) platforms have become very popular forums for asking and answering questions daily. While these forums are rich repositories of community knowledge, they present challenges for finding relevant answers and similar questions, due to the open-ended nature of informal discussions. Further, if the platform allows questions and answers in multiple languages, we are faced with the additional challenge of matching cross-lingual information. In this work, we focus on the cross-language question re-ranking shared task, which aims to find existing questions that may be written in different languages. Our contribution is an exploration of query expansion techniques for this problem. We investigate expansions based on Word Embeddings, DBpedia concepts linking, and Hypernym, and show that they outperform existing state-of-the-art methods.

KEYWORDS

Query Expansion, Cross-Language Information Retrieval, Community Question-Answering, DBpedia Concept Linking

1 INTRODUCTION

Due to the huge popularity of community question-answering (CQA) platforms, such as Quora and Stack Overflow, it has captured the attention of researchers as an area with social impact. Users who ask questions receive quick and useful answers from these community platforms. Since these platforms are open-ended having crowd-source nature, it is a challenging task to find relevant questions and answers. To get the best use of these community knowledge repositories, it is very important to find the most relevant questions and retrieve their relevant answers to a new question. The informal writing makes this a challenge.

To deal with the need of real applications in CQA, we focus on the task of *question re-ranking*. As defined by SemEval-2016, given (i) a new question and (ii) a large collection of existing questions in the CQA platform, *rerank all existing questions by similarity to the new question*, with the hope that they may provide the answers to the new question [15]. This addresses the challenge that there exists many ways to ask the same question, and an existing question-answer pair may already satisfy the information need.

If a CQA platform supports text entry in multiple languages, this becomes a type of cross-language information retrieval (CLIR) problem. A machine translation (MT) model may be used to translate one language into another language prior to indexing and search, and translation errors may lead to degradation in either precision or recall. For example, Fig 1 shows a question in English, Arabic, and MT English from [5], who extends the SemEval-2016 task to cross-language (CL) settings: the collection of existing questions are in English, and the user queries are simulated as new questions written in Arabic.[1] We can see that there are various translation errors in the MT English compared to the original English question.

Original English:
can i sponsor my kids even if i am a husband sponsor of my new husband here in Doha?we just got married here in DOHA ;July 2012..and he treated my kids as his own and we're planning to get them next year by march?/who will sponsor them ..Me or my husband?
Human translated Arabic:
الدوحة؛ في للتو تزوجنا لقد الدوحة؟ في منا الجديد زوجي كفيلة كنت إذا حتى أولادي كفالة ستطيع أ هل
سيكون من/مارس؟ شهر بحلول القادم العام في لاستقدامهم ونخطط كأطفاله، أطفالي يعامل وهو..2012 يوليو
زوجي؟ أم أنا.. كفيلهم
MT English:
you can't ensure that my children, even if you're capable of my husband's new here in Doha? I've been married for us just in Doha; July 2012.. and is treated as a young kid, my children, and we plan to recruit them in the month of March next year, with the advent of ?/ will be their sponsor. I or my husband?

Figure 1: English-Arabic-MT Question

In order to address the complexity on question re-ranking in a CLIR setting for CQA platforms, we explore different query expansion (QE) techniques. Our hypothesis is that mis-translations are often different nuances of related concepts, so by expansion with similar terms, we may recover the underlying terms needed for matching. We investigate Word Embedding, DBpedia, and Hypernym knowledge graph to expand query in a question-question re-ranking task. To the best of our knowledge, we first propose QE techniques for CL question re-ranking on CQA platforms. Our QE work flow is given in Fig 2. We adopt a query translation approach, followed by QE: Given an initial query (e.g. MT English), we expand each term with information from outside resources, then match against the existing questions which are indexed as English documents in a search server like ElasticSearch.

We develop baseline and aggregated systems using QE methods and evaluate our approaches on the CL extension of the SemEval-2016 dataset [5, 15]. The evaluation results show that our QE systems achieve significant improvement over existing methods on CL question re-ranking.

2 RELATED WORK

Aiming to retrieve information in a language different from the query language, a wide range of research has been done in CLIR [2, 11, 13, 16, 19, 20]. To improve the search performance of a CLIR system, researchers have been giving more importance on QE techniques, such as, using of external lexical database [14], co-occurrence analysis [23] and pseudo-relevance feedback [10, 24]. Zamani et al. [1], Kuzi et al. [9] and Diaz et al. [6] presented QE techniques using Word Embedding. A different approach using external knowledge base were developed by Xiong et al. [22] and Zhang et al. [25] to expand queries in CLIR.

[1]In the original SemEval-2016 dataset, both new and existing questions were in English, but for the CL extension in [5], the new questions were replaced with their manual Arabic translations, and Arabic-to-English MT results were added to simulate a CLIR setup.

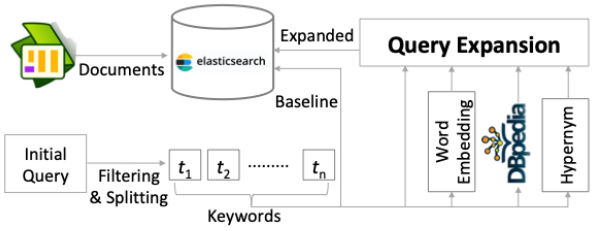

Figure 2: Query Expansion Work Flow
[Query = Question and Documents = Existing Questions]

Although, CQA [3, 4, 8, 12, 21] is a popular research area, there has not been much work done in CL CQA task. The CL CQA heavily depends on CL question ranking. SemEval-2016 introduced a shared task on CQA in English and Arabic [15]. They also had a subtask question-question similarity ranking [7]. A later work was done by Martino et al. [5] in CL question re-ranking using a CL tree kernel. There is still a big scope of improving the performance of community question re-ranking task. To the best of our knowledge, no research has been done on QE for community question re-ranking task.

3 APPROACH

3.1 Query Expansion

Query Expansion is a technique of improving the search performance in IR by augmenting the initial query using different methods. To achieve a better search performance, it is important to expand the initial query to retrieve more relevant documents. Specially it is very useful CL settings since the initial query may miss important clue to retrieve more relevant documents in a different language than the original query. To expand any query we use different expansion techniques and take the union of expanded terms. For any query Q the expanded query is given in the equation 1.

$$QE^Q = kw^Q \bigcup qe^Q_{we} \bigcup qe^Q_{db} \bigcup qe^Q_h \tag{1}$$

Where $\bigcup$ is a union operator. QE^Q is the union of all expanded queries along with the original keyword query kw^Q. And qe^Q_{we}, qe^Q_{db} and qe^Q_h are QE using Word Embedding, DBpedia concepts linking and Hypernym respectively.

3.1.1 Word Embedding. Word embedding is a type of word representation that transforms a text into a numerical vector of real numbers where semantic similar words are grouped together. By the characteristics of word embedding, two words having semantic similar meaning share a fair amount of surrounding context. It implies that search engine should return relevant documents to the query term using the similar word. In a CL setting, due to MT, the translated text may miss the original word, and produce a different word of similar meaning, the indexed documents may not have the exact keyword. Hence the retrieval engine won't return documents of out of vocabulary.

From these intuitions and inspired by [1, 9], we use a pre-trained word embedding GloVe vectors [17] based on Wikipedia and Gigaword corpus to get the most similar words for each of the query terms from the original query. The dimension of the vector is 100 and it contains 400k vocab. For a query Q having query terms q_1, q_2 ..... q_n, let $t_1^{q_i}$ and $t_2^{q_i}$ are the 2 most similar terms for query term q_i obtained from the pre-trained word embedding model, then QE using word embedding is computed by the equation 2.

$$qe^Q_{we} = \bigcup_{i=1}^{n} \{t_1^{q_i}, t_2^{q_i}\} \tag{2}$$

As an example, a query term "travel" from a query, *Im likely to travel in the month of june... just wanna know some good places to visit....*, is expanded using "travelers" and "trips" where "travelers" and "trips" are 2 most similar terms for the query term "travel".

3.1.2 DBpedia. DBpedia [2] is a structured open knowledge base derived from the Wikipedia. The current version of DBpedia has 4.58 million English concepts and 38.3 millions of concepts in 125 different languages. Motivated by the work presented in [22] to associate a term with entities from Freebase knowledge graph, we use DBpedia to extract concepts and linked each query term with DBpedia concepts. Given a query Q having query terms q_1, q_2 ..... q_n, we retrieve DBpedia concepts for each of the query terms q_i. Each of the returned concepts is associated with different types and properties that reflect the concept. In the expansion term selection, we choose a simple but useful strategy. We select a property called $dct : subject$ that links a concept with relevant subjects. We use the relevant subjects to expand the query term, q_i. The intuition is that the concepts associated with query terms are able to capture more relevant documents. The QE is computed by the equation 3.

$$qe^Q_{db} = \bigcup_{i=1}^{n} \bigcup_{j=1}^{k} \{subject : q_i^j\} \tag{3}$$

Where k is the number of subjects for a query term q_i and q_i^j is a subject of a concept associated with the query term q_i.

As an example, a query term "travel" from a query, *Im likely to travel in the month of june... just wanna know some good places to visit....*, is expanded using "Tourism", "Tourist activities" and "Transport culture" where "Tourism", "Tourist activities" and "Transport culture" are the subjects of a concept associated with "Travel".

3.1.3 Hypernym. Hyponym is a specific word or phrase of a more generic term. The generic term is called hypernym. Due to MT, the translated term may have a different form of the original term. We use a publicly available hypernym knowledge graph[3] developed by [18] to extract hyponym labels with a high confidence score for each of the query terms and include them in the QE. The motivation is to retrieve more relevant documents which may contain hyponym terms of an original query term. For a query Q having q_1, q_2 ...q_i... q_n query terms, we expand Q using the equation 4.

$$qe^Q_h = \bigcup_{i=1}^{n} \bigcup_{j=1}^{k} \{hyponym : q_i^j \mid cs(q_i^j, q_i) >= 0.75\} \tag{4}$$

Where, k is the number of hyponym labels for a query term q_i, q_i^j is a hyponym label and $cs(q_i^j, q_i)$ is a function that gives a confidence score for q_i^j with respect to q_i.

[2] https://wiki.dbpedia.org/
[3] http://webisa.webdatacommons.org/

As an example, a query term "travel" from a query, *Im likely to travel in the month of june... just wanna know some good places to visit....*, is expanded using "operating expense", "related expense" and "personal expense" where "operating expense", "related expense" and "personal expense" are hyponym labels for "travel".

3.2 Search and Ranking

To index documents, search queries and rank retrieved documents, we use Elasticsearch [4], a Lucene based distributed, RESTful search and analytics engine. We use a built-in English analyzer, which is used to converting documents into tokens to the inverted index for searching. The same analyzer is applied to the query string at the search time. As a ranking or scoring algorithm, we use BM25 similarity algorithm. We also configure the similarity algorithm by hyper-parameter tuning.

4 DATASET

To evaluate our systems, we use SemEval-2016 Task 3 CQA dataset (Subtask B) [15] and MT version of human translated Arabic questions [5]. In [5], a new question is written in Arabic, and the goal is to retrieve similar questions in English; an MT system was available for translating the new Arabic question into English. In this research, the new question is considered as a query and a set of the first ten existing questions are considered as documents. The task is to re-rank the documents using different QE techniques. The dataset has 267 train, 50 dev and 70 test queries; there are 10 existing questions to be re-ranked for each new question, leading to 2670 train, 500 dev and 700 test "documents". In this research, we use only the dev and test datasets. To setup a CL environment, we choose MT version of 50 dev and 70 test questions from [5] and consider them as machine translated queries.

5 EXPERIMENTAL SETUP

The datasets explained in the previous section were used for our experiments in the cross-language question re-ranking task. Initially we indexed all existing questions, which are considered as documents in this research, using the approach described in the Search and Ranking section. We observed that both English and machine translated queries might have punctuation and common words. To build a more clean queries, we filtered out punctuation and common English words from the initial queries. Then we split each query into words which are considered as keyword query, our baseline system. We experimented in two scenarios: a) English query, the original SemEval-2016 task B, and b) Machine translated query, where Arabic version of the English queries are translated back into English. The second scenario is the CLIR setting we focus here, while the first scenario provides comparison results in the monolingual setting.

We configured 18 systems for English and MT queries, by combining each of the four basic systems: (a) Keyword, (b) Word Embedding, (c) DBpedia and (d) Hypernym. The system (a) is the baseline system whereas (b), (c) and (d) are systems based on QE using Word Embedding, DBpedia and Hypernym knowledge graph respectively. The average query lengths for baseline systems are 18.67(EN) and 19.96(MT) words. And the average word additions are Word Embedding: 30.98(EN) and 35.34(MT); DBpedia: 25.81(EN) and 31.32(MT); Hypernym: 21.58(EN) and 29.04(MT); Best system: 78.3(EN) and 95.6(MT). The combination of 18 systems are given in table 1. All the systems are experimented in two settings - **Dev** and **Test**. The search ranking scores are calculated for all 18 systems using BM25 algorithm. We tuned BM25 hyper-parameters, k1 and b on **Dev** set to get the optimized values where k1 controls non-linear term frequency normalization and b controls to what degree document length normalizes tf values. The score for each query is calculated based on 10 existing documents to re-rank them.

[4]https://www.elastic.co/products/elasticsearch

System	QR	MAP		Δ
		Dev	Test	
1. Keyword(KW) (Baseline)	EN	72.60	71.43	00.00
2. Word Embedding(WE)	EN	64.40	63.86	-07.57
3. DBpedia(DB)	EN	41.00	45.29	-26.14
4. Hypernym(HN)	EN	21.00	27.86	-34.57
5. 1 + 2 (KW+WE)	EN	80.20	79.86	+08.43
6. 1 + 3 (KW+DB)	EN	76.00	75.29	+03.86
7. 1 + 4 (KW+HN)	EN	75.20	76.00	+04.57
8. 2 + 3 + 4 (WE+DB+HN)	EN	72.20	75.86	+04.43
9. 1 + 2 + 3 + 4 (Best)	EN	**84.00**	**82.00**	+10.57
10. UH-PRHLT(SemEval[7, 15])	EN	75.90	76.70	-
11. SVM + TK [5]	EN	73.02	77.41	-
12. Keyword(KW) (Baseline)	MT	72.20	67.57	00.00
13. Word Embedding(WE)	MT	64.40	63.43	-04.14
14. DBpedia(DB)	MT	43.20	45.71	-21.86
15. Hypernym(HN)	MT	26.80	32.71	-34.86
16. 12 + 13 (KW+WE)	MT	79.20	75.57	+08.00
17. 12 + 14 (KW+DB)	MT	75.40	71.43	+03.86
18. 12 + 15 (KW+HN)	MT	76.40	72.29	+04.72
19. 13 + 14 + 15 (WE+DB+HM)	MT	77.60	73.14	+05.57
20. 12 + 13 + 14 + 15(Best)	MT	**84.00**	**78.29**	+10.72
21. SVM+TK([5])	MT	72.94	76.67	-

Table 1: MAP scores for various QE on English (EN) questions (monolingual setup) and MT questions (CLIR setup).

6 RESULTS AND ANALYSIS

Table 1 compares the MAP scores on the dev and test sets. The "System" column compares QE, Keyword baseline without QE, and previously-published state-of-the-art methods in this task (UH-PRHLT, SVM+TK); the QR column shows whether the query was English (monolingual setup) or Machine-translated English (cross-language setup). The Δ column displays the difference between test MAP against the baseline.

The results on the original English queries are shown in rows 1 to 11. Row 1 is the baseline system using keyword query without any QE. Rows 2 to 4 show scores for QE using Word Embedding, DBpedia and Hypernym respectively. We observe that each of the individual systems from rows 2 to 4 has a negative Δ score. That means QE using any single approach doesn't beat the baseline system. Combination of baseline and any QE method are shown in rows 5 to 7 where each of the combinations has better performance

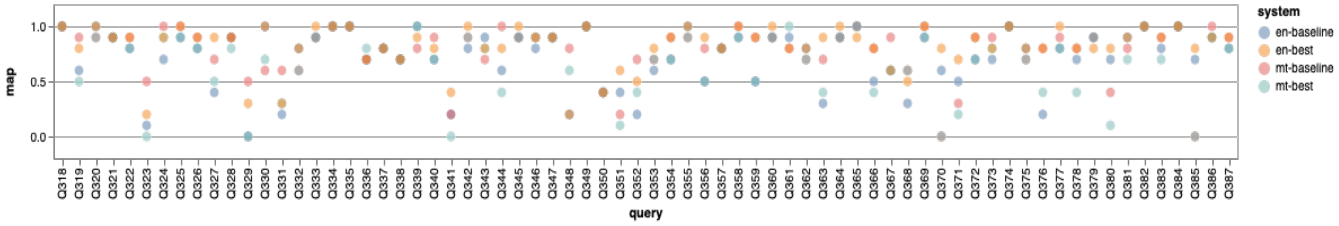

Figure 3: MAP scores over queries

than the baseline system. Among them, combination of word embedding and the baseline system has the highest Δ score which is +08.43. Row 8 shows QE using a combination of Word Embedding, DBpedia and Hypernym, which also beats the baseline system by +04.43 Δ score. An aggregation of all systems from rows 1 to 4 is shown in row 9, which is the best performing system. The best system beats our baseline system by +10.57 Δ score which is a substantial improvement over the baseline system.

The best performing systems from the SemEval-2016 Task 3 [7, 15] and the state-of-the-art system [5] in the community question re-ranking task are shown in row 10 and 11 respectively. Our best system outperforms them by +05.30 and +04.59 MAP scores respectively which is an effective improvement in the community question re-ranking task.

The results on MT queries are presented in rows 12 to 21 where the baseline system is shown in row 12. Individual QE using Word Embedding, DBpedia and Hypernym are displayed in rows 13 to 15. We find a similar pattern between these and the QE in English that is negative Δ score which implies individual QE technique also doesn't perform well for MT queries. Union of the baseline system and any other expansion method from rows 13 to 15 are given in rows 16 to 18. Similar to the English query, we also achieve positive Δ scores for each of them where expansion using Word Embedding has the highest Δ score +08.00.

Row 19 shows a combination of QE methods which also beats the baseline by +05.57 MAP score. The best system, union of baseline and all QE given in row 20, improves the performance by +10.72 Δ score. Our best system in MT setting also outperforms the state-of-the-art system given in row 21 by +01.62 MAP score. The significant improvement compared to the baseline and the state-of-the-art, implies that our QE approaches are also strong to MT.

In the comparison of baseline systems in English and MT (row 1 and 12), we notice that the MT baseline system has a lower MAP score by -3.86. We also observe that the MT best system degrades the MAP score by -03.71 than that of English. We assume the reason behind these low map scores for MT systems is the effect of the output of machine translation. We see that individual QE using DBpedia and Hypernym have slightly better performance in MT than English by +00.42 (diff. between row 14 and row 3) and +04.85 (diff. between row 15 and row 4) map scores respectively.

Most importantly, we find that our best systems (English and MT), outperform the baseline systems and state-of-the-art results in community question re-ranking task. These indicate that our QE methods are robust and effective in both monolingual and CL settings. Figure 3 shows MAP scores for each query of baseline and best systems for both English and MT. One interesting observation is that we get MAP score 0 for 5 out of 70 test queries, and all these are for either MT baseline or MT best system. Only query 329 has a 0 MAP for English baseline system along with MT best system.

7 CONCLUSION

We investigate different query expansion techniques for improving cross-language question re-ranking in community question answering platforms. Our techniques, though simple, outperform current state-of-the-art on SemEval-2016 Task 3 and its CLIR extension. As a future work, we plan to improve methods for candidate terms selection for each of the different query expansions types.

ACKNOWLEDGMENTS

This work is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract #FA8650-17-C-9115. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] Saeid Balaneshin-kordan and Alexander Kotov. 2017. Embedding-based Query Expansion for Weighted Sequential Dependence Retrieval Model. In *Proceedings of SIGIR*. ACM, 1213–1216.
[2] Lisa Ballesteros and Bruce Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *Proceedings of DEXA*. Springer, 791–801.
[3] Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of WWW*. ACM, 467–476.
[4] David Carmel, Avihai Mejer, Yuval Pinter, and Idan Szpektor. 2014. Improving term weighting for community question answering search using syntactic analysis. In *Proceedings of CIKM*. ACM, 351–360.
[5] Giovanni Da San Martino, Salvatore Romeo, Alberto Barroón-Cedeño, Shafiq Joty, Lluís Maàrquez, Alessandro Moschitti, and Preslav Nakov. 2017. Cross-Language Question Re-Ranking. In *Proceedings of SIGIR*. ACM, 1145–1148.
[6] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query expansion with locally-trained word embeddings. *arXiv preprint arXiv:1605.07891* (2016).
[7] Marc Franco-Salvador, Sudipta Kar, Thamar Solorio, and Paolo Rosso. 2018. UH-PRHLT at SemEval-2016 Task 3: Combining lexical and semantic-based features for community question answering. *arXiv preprint arXiv:1807.11584* (2018), 1–8.
[8] Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. 2012. Question-answer topic model for question retrieval in community question answering. In *Proceedings of CIKM*. ACM, 2471–2474.
[9] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query expansion using word embeddings. In *Proceedings of CIKM*. ACM, 1929–1932.
[10] Victor Lavrenko and W Bruce Croft. 2017. Relevance-based language models. In *SIGIR Forum*, Vol. 51. ACM, 260–267.

[11] Bo Li and Ping Cheng. 2018. Learning Neural Representation for CLIR with Adversarial Framework. In *Proceedings of EMNLP*. 1861–1870.

[12] Baichuan Li, Tan Jin, Michael R Lyu, Irwin King, and Barley Mak. 2012. Analyzing and predicting question quality in community question answering services. In *Proceedings of WWW*. ACM, 775–782.

[13] Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised Cross-Lingual Information Retrieval Using Monolingual Data Only. In *Proceedings of SIGIR*. ACM, 1253–1256.

[14] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[15] Preslav Nakov, LluÃŋs MÃărquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed A. Freihat, James Glass, and Bilal Randeree. 2016. SemEval-2016 Task 3: Community Question Answering. *Proceedings of SemEval* (2016), 525–545.

[16] Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan and Claypool Publishers.

[17] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP*. 1532–1543.

[18] Julian Seitner, Christian Bizer, Kai Eckert, Stefano Faralli, Robert Meusel, Heiko Paulheim, and Simone Paolo Ponzetto. 2016. A Large DataBase of Hypernymy Relations Extracted from the Web. In *LREC*.

[19] Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. 2002. Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. In *Proceedings of CICLing*. Springer, 415–424.

[20] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In *Proceedings of SIGIR*. ACM, 363–372.

[21] Yang Xiang, Qingcai Chen, Xiaolong Wang, and Yang Qin. 2017. Answer Selection in Community Question Answering via Attentive Neural Networks. *Proceedings of Signal Processing Letters* 24, 4 (2017), 505–509.

[22] Chenyan Xiong and Jamie Callan. 2015. Query expansion with Freebase. In *Proceedings of ICTIR*. ACM, 111–120.

[23] Jinxi Xu and W Bruce Croft. 2017. Quary expansion using local and global document analysis. In *SIGIR forum*, Vol. 51. ACM, 168–175.

[24] Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM*. ACM, 403–410.

[25] Lei Zhang, Michael Färber, and Achim Rettinger. 2016. Xknowsearch!: exploiting knowledge bases for entity-based cross-lingual information retrieval. In *Proceedings of CIKM*. ACM, 2425–2428.