CNU System in NTCIR-11 IMine Task

Global Semantic Expansion for Hierarchical Query Intent Identification

Wei Song, Wenbin Xu, Lizheng Liu, Hanshi Wang
College of Information Engineering
Capital Normal University, China
{wsong, wbxu, lzliu, hswang}@cnu.edu.cn

ABSTRACT

Understanding user intent is important for interactive and personalized information retrieval. For ambiguous queries, user intent space actually forms a hierarchical top down architecture: from senses to subtopics, rather than a flat structure. This paper presents the CNU system in NTCIR-11 IMine task. Our method constructs the hierarchical structure by exploiting global semantic representation and expansion. The highlights include: 1) We use word semantic vectors and propose a query dependent semantic composition for representing query aspect phrases. Our target is to alleviate the term-mismatch and data sparseness problems, which shallow lexical matching and co-occurrence based local semantics are ineffective to overcome. 2) We expand query subtopics by introducing new words according to global semantic relatedness and cluster these words for query sense induction. The evaluation results on and post NTCIR-11 show that our method could mine query subtopics and senses effectively.

Team Name

CNU

Subtask

Subtopic Mining

Keywords

Query Intent, Query Sense, Word Embeddings, Semantic Expansion, Query Subtopic

1. INTRODUCTION

Inferring query intent is one of the most important tasks for information retrieval. Web queries are often ambiguous and multi-faceted. This brings in great challenges for providing most relevant information to users. If the system could mine all potential query intents and construct a proper structure to organize them, it could provide more sophisticated services to deal with such difficult queries. For example, we could diversify the search results [9] or provide query summarizations [11] according to multiple query aspects. We could also represent potential query intents to users in an interactive way, and make decision with the help from user feedbacks.

In recent years, query intent mining has gained much attention. Much work has been done using information from different resources such as query logs or search results [1, 2, 3, 8]. However, most systems provide a flat list of phrases to represent query subtopics. Actually, query intent are hierarchical rather than flat. This is obvious for ambiguous queries. The top layer covers different senses of an ambiguous query, which usually refer to different meanings in reality. The lower layer covers various facets related to the meaning. For example, "apple" is an ambiguous query. It has several meanings referring to different objects in reality. Subtopics reveal different aspects of a specific object. A flat list structure is not enough to represent the information need space for a query. For example, "apple diet" and "apple notebook sale" are intents for different meanings but are mixed together. While "apple price" could be a subtopic for multiple meanings but it is unable to distinct in a flat list.

In addition, in most architectures of existing systems, clustering text fragments is an important component [1, 2, 12]. Due to the shortness of text fragments, either extracted from query logs or documents, data sparseness and term-mismatch problem result in great challenges for clustering. It is necessary to exploit new semantic representations to overcome these problems.

This paper introduces our system in NTCIR-11 IMine task which induces senses and subtopics of an ambiguous query automatically. Specially, we make use of global semantic vector representations and expand query subtopics to bring in global semantic information. The global information is useful for bridging local subtopics which are difficult to be judged as similar using local semantic representations. Our main contributions and preliminary findings include:

- We apply word semantic vectors and propose a sematic composition method to represent query aspect phrases. The experimental results show that such representations are more effective than traditional semantic representations for clustering query reformulations to infer query subtopics.
- We induce query senses by clustering expanded words which are related to query subtopics in terms of global semantics. The experiments prove that global semantic expansion is effective for query sense induction.

2. NTCIR-11 IMINE TASK

NTCIR-11 organizes the IMine task which is short for search Intent Mining [6]. IMine task consists of two subtasks: Subtopic Mining and Document Ranking. We attend the subtopic mining subtask on Chinese queries. The subtopic mining subtask this year is different from before. It requires participants to submit a two-level hierarchy of

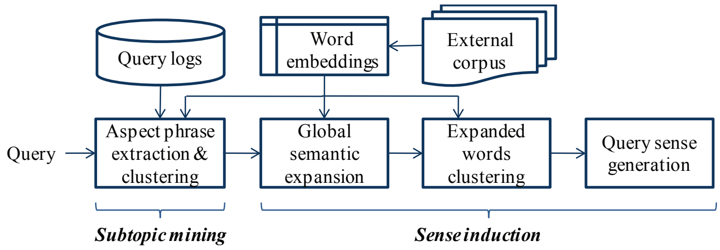

Figure 1: The flowchart of proposed system.

sub-intents for query topics. No doubt, it advances the research on query intent understanding deeper. The hierarchical structure presents the intent structure more precisely, especially for ambiguous queries.

Without losing generality, we call the top level subtopics of this hierarchy as **query senses**, and call the second level subtopics as **query subtopics** which now are used to describe the intent for specific query senses.

3. PROPOSED METHOD

3.1 Terminology

We introduce the basic terminologies that will be used throughout this paper.

- Query: the user issued keywords which represent (vague) user information need, noted as q.
- Query reformulation: a short text span which contains the query as a substring. For a query, we would extract a set of query reformulations $R = \{r\}$, where each $r \in R$ is a reformation of query q. For simplicity, we only consider the text spans that begin with or end with the query q.
- Query aspect phrase: the text surround the original query in a query reformulation form an aspect phrase. By this definition, a query reformulation r could be represented as $r = q + a$ or $r = a + q$, where a is a query aspect phrase.

For example, if the query is "apple", we could collect a set of user reformulations such as "apple notebook" and "apple diet" from query logs or other resources. And "notebook" and "diet" are viewed as aspect phrases for query "apple".

3.2 Framework

The framework of proposed method is shown in figure 1. From a big picture, our method has 2 phases: subtopic mining and sense induction.

In the first phase, we first extract aspect phrases from resources such as query logs, and then cluster them into groups, each of which forms a query subtopic. In the second phase, we expand the subtopics with new words based on global semantic similarity, and then cluster the words. Finally, query subtopics are assigned to word clusters. The word clusters which no query subtopic is assigned to are discarded. The remaining ones are used to represent query senses.

In both phases, distributed semantic vectors are used to represent words and phrases. The vectors for words are learned offline from an external large corpus using a neural networks based model [7]. We propose a query dependent linear semantic composition approach to compute the vectors for phrases online. Global semantic expansion for query subtopics and word clustering are also based on the semantic representation. Next, we first introduce the semantic representations for words and phrases and then introduce other modules in our framework.

3.3 Semantic Representations

3.3.1 *Word Embeddings*

Recently, deep learning is hot for machine learning and natural language processing. An important product is to represent the words by distributed continuous vectors which are called word embeddings. One of the advantages is that the continuous vector representations could be used to measure the semantic relatedness between words which vary in lexical surface strings, such as "teacher" and "professor".

Various models have been developed for learning word embeddings efficiently on large scale corpus. Notice that this procedure is offline, but the learned embeddings could be used conveniently for online applications. In order to overcome the term-match problems when processing short texts, we borrow the idea of using continuous vector representations and attempt to represent query subtopics in this way.

3.3.2 *Aspect Phrase Embeddings*

Aspect phrases are usually multi-word expressions. Given the word embedding of individual words, we compute the semantic vectors of phrases by using a linear composition. The general form is shown in equation 1, where $vec(w)$ is used to represent the vector of word w, and $phrVec(w_1, ..., w_n)$ is used to represent the vector representation of a phrase with n words, and α_i is a weighting parameter for the ith word in the phrases, subject to $\sum_i \alpha_i = 1$. In this way, a phrase is represented using a vector which has the same dimensions with a word vector. This is convenient to compute similarities between phrases with different length.

$$phrVec(w_1, ..., w_n) = \sum_{i=1}^{n} \alpha_i vec(w_i) \tag{1}$$

Obviously, words in a phrases are not equally important to indicate its meaning. Therefore, we aim to assign more weights to words to keep the composite embeddings close to its real position in vector space. Huang proposed a method to weight words using IDF [5]. The IDF scheme reveals the importance of a term from a global view. However, here we focus on clustering query reformulations, so that it is more reasonable to consider the relatedness to the original query. Here, we use a simple metric to measure the importance of a word in a phrase. We assume that the more times a word co-occurs with the query, the more important it is. Formally, the frequencies of the query reformulations are $RF = \{(r_i, f_i)\}$, where f_i is the frequency of query reformulation r_i. The weight of word w is:

$$score(w) \propto \sum_{i=1}^{|RF|} f_i \cdot indicator(w, r_i) \tag{2}$$

where

$$indicator(w, r_i) = \begin{cases} 1, & r_i \text{ contains } w; \\ 0, & \text{otherwise.} \end{cases}$$

3.4 Query Subtopic Mining

Before inducing query senses, we first mine a flat list of query subtopics. These subtopics provide fine-grained descriptions about different aspects of query senses and are fundamental for query sense induction.

Actually, any existing subtopic mining algorithms could be used. Here, we first extract aspect phrases from query reformulations in query logs or text fragments in search results. Then, a clustering method is applied here to cluster aspect phrases into groups. If the size of one result cluster is smaller than 3, this cluster is considered as an outlier and discarded. The remaining query aspect phrase clusters represent query subtopics.

Cosine similarity is used to measure the similarity between two semantic vectors of query aspect phrases for clustering, which is defined as:

$$cosine(u, v) = \frac{u \cdot v}{|u||v|} \tag{3}$$

where u and v are two vectors with the same number of dimensions. We use KMeans algorithm for clustering. Actually, other clustering algorithms could be applied as well.

According to the report of NTCIR 9, the average number of subtopics is about 10 [10]. Considering that duplications may exist in automatic generated subtopics, we set $K = 20$ to insure that most true query subtopics could be collected. The initial centroids of KMeans are selected randomly. We run KMeans several times and choose the result with the largest sum of intra cluster similarities.

3.5 Query Sense Induction

Now we have a set of fine-grained subtopics. We aim to cluster these query subtopics further to induce query senses. However, the aspect phrase of these subtopics may have different lexical representations. It is difficult to measure the similarity between them based on lexical matching. Exploiting the co-occurrence information is an option [1] but also faces the data sparseness problem. In order to alleviate these problems, we propose a global semantic expansion based method for incorporating more global information.

3.5.1 Global Semantic Expansion

For each query subtopic, we view every aspect phrase a as a seed, and expand a list of most similar words based on the learned word embeddings. Our motivation first borrows the idea of query expansion by bringing in more related words about different query senses to alleviate the data sparseness problem. To select proper words to be expanded, we further make use of globally learned word embeddings. Because word embeddings are learned on large scale of text data so that the representations should model the relatedness between words more accurately than local co-occurrences.

Therefore we select additional words according to the cosine similarities between the distributed vectors of all words in vocabulary and the query aspect phrase. The set of expanded words is noted as

$$\{w | \text{if } cosine(vec(w), phrVec(a)) > T \ \& \ rank(w) < N\} \tag{4}$$

Where T is a threshold that the similarity must be larger than it, and we choose the words whose ranks are among the top N closest to the seed in order to reduce noises. The values of T and N are set to 0.5 and 20 respectively after several trials.

3.5.2 Word Clustering

Given the expanded words, we group them into clusters using a clustering algorithm. Again the consine similarity is used here. The generated clusters are viewed as candidate query senses.

We compare two clustering algorithms: KMeans and Affinity Propagation (AP) algorithms [4]. For KMeans we set K equals to 5 according to the output requirement of IMine task. AP algorithm could determine the number of clusters automatically. We set the preference values uniformly to the median value of pair similarities. If the number is equal or smaller than 5, retain the clustering result. Otherwise, we retain the top 5 clusters after query sense generation to satisfy the task requirement.

3.5.3 Query Sense Generation

We assign query subtopics to word clusters. The assignment is done by computing the similarity between the centroid vector of a subtopic and the centroid of every word cluster. Each subtopic is assigned to the cluster which has the biggest similarity between them. The subtopics belonging to the same query sense are ranked according to the similarity [1]. The clusters, which no subtopic is assigned to, are discarded. The remaining ones are used to represent query senses. The query senses are ranked according to the accumulative number of query reformulations they have. Naturally, a hierarchical structure forms.

4. EXPERIMENTAL SETTINGS

4.1 Research Questions

We seek to answer the following research questions:

- Whether word vectors are better representations comparing to traditional ones for clustering short texts?
- Whether global semantic expansion benefits the query sense induction?

4.2 Dataset

4.2.1 Dataset for Subtopic Mining

As discussed before, the first step for subtopic mining is to get a set of query reformulations. The query reformulations we used are provided by the organizer, including query suggestions from major search engines, query dimensions from search result pages and related queries extracted from Sogou query logs[2]. Because we are interested in the above research questions, we didn't use data from other resources for better coverage.

4.2.2 Dataset for Learning Word Embeddings

We crawled a collection of webpages covering various topics from Baidu Baike[3], the largest Chinese knowledge base. We used the Word2vec software[4], which is an implementation based on [7], to learn word embeddings. It provides two particular models for efficiently inducing word embeddings:

[1]The scores are also used as the weights of each subtopics in our submitted results. The subtopics we submitted in fact are not ranked globally.

[2]`http://www.thuir.cn/imine/`

[3]`http://baike.baidu.com/`

[4]`https://code.google.com/p/word2vec/`

the Skip-gram and CBOW model. We employed the CBOW model for estimating word embeddings, because it has comparative performance comparing to Skip-gram model but is faster. The number of word vector dimensions is 300.

4.3 Runs

We submitted 4 runs for evaluation. The difference between 4 runs include: 1) clustering metrics: similarity between word vectors (WordVec), or similarity based on co-occurrence (Cooc) [1]. 2) whether using semantic expansion: Expansion or NoExpansion. 3) word clustering algorithms: KMeans or AP. According to this, the 4 runs are described as follows.

- Run1: WordVec-Expansion-KMeans.
- Run2: WordVec-Expansion-AP.
- Run3: WordVec-NoExpansion-KMeans.
- Run4: Cooc-NoExpansion-Cooc

4.4 Evaluation Metrics

NTCIR 11 provides 3 separate scores $HScore$, $FScore$, $SScore$. $HScore$ is to measure the accuracy of query sense identification. $FScore$ is to measure the ranking quality of query senses. $SScore$ is to measure the ranking quality of all query subtopics from all query senses. Finally, $H-Measure$ is used to combine the 3 scores. Please refer to the overview paper [6] to get the detail definitions of these evaluation metrics.

5. EXPERIMENTAL RESULTS

Table 1 shows the official results of our 4 Runs and the best scores of other systems. We can see Run2 gets the best performance among all participants on $HScore$, and Run4 gets the best performance on $FScore$. As we have mentioned earlier, because our subtopics' weights are oriented to query senses but not the original query, it is not proper to use the weights for global ranking. This is one of the reasons that our runs get very poor performance on $SScore$. Another reason might be that we only use 20 subtopics, while the organizer evaluates the performance on top 50. After evaluation and based on the standard results, we rerank 20 subtopics of each query using the strategy used in [12]. The results are shown in Table 2. Run*-20 represents the results of reranked 20 subtopics, while Run*-50 represents the results of full filled 50 positions by copying the top 20 topics repeatedly. Run*-50 is meaningless in practice, just as a reference here.

Runs	HScore	FScore	SScore	H-Measure
Run1	0.5353	0.5867	0.2045	0.1739
Run2	**0.5789**	0.5569	0.1932	0.1748
Run3	0.4611	0.6073	0.1910	0.1407
Run4	0.5086	0.4708	0.1626	0.1189
BEST	0.5436	**0.7191**	**0.6718**	**0.3360**

Table 1: The official results of different Runs on IMine Chinese topics. BEST refers to the best scores of other systems.

Runs	SScore	H-Measure
Run1-20	0.3195	0.2425
Run2-20	0.3180	**0.2532**
Run3-20	0.3062	0.2106
Run4-20	0.2703	0.1884
Run1-50	0.5642	0.3080
Run2-50	0.5627	**0.3240**
Run3-50	0.5560	0.2507
Run4-50	0.4953	0.2457

Table 2: The post-evaluation results of Sscore and H-Measure on IMine Chinese topics.

Analysis. From the results, we can get several preliminary observations. First, word embedding based query aspect phrase representation and similarity computation (Run1, Run2, Run3) are more effective comparing to co-occurrence based strategy (Run4) for query subtopic mining. Second, global semantic expansion (Run1, Run2) could have better $HScore$ but worse $FScore$ for query sense induction. This indicates global semantic expansion is very useful for clustering subtopics into higher semantic levels. The reason may be that the new expanded words alleviate the data sparseness problem. These words could bridge the semantic gap between query aspect phrases and help distinguishing different query senses. Third, for query sense induction, clustering algorithms which could determine the number of clusters automatically (such as AP) are preferred. It is easy to understand, since different queries have various number of senses.

By comparing the golden answers and our system outputs, we find that global semantic expansion based methods tend to group subtopics into more general senses. For example, for the query "小米", Run2 found 2 query senses: one is related to "小米手机(Xiaomi mobile phone)", the other is related to food "小米(millet)". But the golden answers separate the first sense into more specific ones, including "小米手机", " 小米公司", "小米论坛", "小米官网". In our opinion, it is difficult to say which granularity is better.

6. CONCLUSIONS AND FUTURE WORK

We have exploited global semantic representations for subtopic mining and query sense induction. The empirical results on NTCIR-11 IMine task confirm the advantage of the proposed method: The continuous word vectors and linear semantic composition based phrase representations are useful for measuring the similarity between query reformulations and help improving the performance of short text clustering for query subtopic mining. The global semantic expansion for query subtopics improves the performance of query sense induction, especially on $HScore$. This indicates that global semantic representation is more effective than traditional representations based on local information for overcoming the term-mismatch and data sparseness problems. In addition, we find that determining the granularity of query senses is not easy to reach an agreement. More practical evaluation strategy may be necessary.

Acknowledgments

This research is supported by NSFC under Grant No. 61303105 and No. 61402304, by MOE Project of Humanities and Social Sciences (No. 14YJAZH046).

7. REFERENCES

[1] V. Dang, X. Xue, and W. B. Croft. Inferring query aspects from reformulations using clustering. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2117–2120, New York, NY, USA, 2011. ACM.

[2] A. Di Marco and R. Navigli. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754, 2013.

[3] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 475–484, New York, NY, USA, 2011. ACM.

[4] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[5] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

[6] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. Kato, H. Ohshima, and K. Zhou. Overview of the ntcir-11 imine task. *Proceedings of NTCIR-11*, 2014.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[8] F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*, pages 1171–1172. ACM, 2010.

[9] R. L. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 595–604. ACM, 2011.

[10] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the ntcir-9 intent task. *Proceedings of NTCIR-9*, pages 82–105, 2011.

[11] W. Song, Q. Yu, Z. Xu, T. Liu, S. Li, and J.-R. Wen. Multi-aspect query summarization by composite query. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 325–334, New York, NY, USA, 2012. ACM.

[12] W. Song, Y. Zhang, H. Gao, T. Liu, and S. Li. Hitscir system in ntcir-9 subtopic mining task.