Mining Implicit Entities in Queries

Wei Li, Wenjie Li, Qin Lu

Department of Computing
The Hong Kong Polytechnic University
{cswli, cswjli, csluqin}@comp.polyu.edu.hk

Abstract

Entities are pivotal in describing events and objects, and also very important in Document Summarization. In general only explicit entities which can be extracted by a Named Entity Recognizer are used in real applications. However, implicit entities hidden behind the phrases or words, e.g. entity <country> referred by the phrase "cross border", are proved to be helpful in Document Summarization. In our experiment, we extract the implicit entities from the web resources.

1. Introduction

Entities, such as <person> and <location>, are pivotal in describing events and objects. Commonly, they function as agents and patients, occurrence time and locations of events; or, describe attributes or aliases of objects. Their semantic implication, or to say the transferred meaning, can deviate greatly from its original meaning. For example, "Great Wall" is a famous resort in China. It represents a location rather than the object "brick wall". The task of query-oriented summarization locates and organizes the events and the objects which are most related to a query from a set of relevant document set. It would make the response seeking process more effective if the summarization system could recognize the entities both explicit and implicit in queries.

In the DUC 2005 query-oriented multi-document summarization (QMS) task, the PolyU group focuses on entity and entity-based pattern matching. They simply make use of five types of explicit entities in queries which can be recognized by a Named Entity Recognizer, namely GATE[1]. Also the query type[2] is determined as one of these five types. The submitted system achieves quite competitive performance (ranking 2nd in the ROUGE evaluation) (Li, Li, Li, Chen and Wu, 2005). Nevertheless, experiments reveal that many entities are actually hidden behind the queries rather than explicitly exposed in the surface text. Look at the following examples[3].

[e1]. Identify and describe types of organized crime that *crosses borders*…

[e2]. Also identify the *perpetrator*s involved with each type of crime…

The above examples present two kinds of implicit entities, namely *associated entities* and *referred entities* respectively. In [e1], the phrase "cross border" often co-occurred with the entity <location> (in particular <country>), such as "<location> and <location> cross borders". We then call the co-occurred entity <location> as the associated entity of the phrase "cross border". In [e2], the phrase "perpetrator" implies the entity <person>, which is a name of the set that a particular element "perpetrator" belongs to. We regard the entity <person> as the referred entity.

The aim of the work reported in this paper is to mine the referred and associated entities implicit in queries by making use of various information resources, such as the knowledge base WordNet or web resources. The evaluations on the DUC2005 data set show that the approaches by enhancing the system with implicit entities outperform the ones using explicit entities alone.

This paper is organized as follows. Section 2 introduces the previous work in document summarization and entity-based methods. Section 3 and 4 describes our approach and schemes of implicit entity mining. Section 5 presents the experiment results. Finally, we conclude the paper and discuss the future work in the last section.

2. Related Work

Entity-based approach has already been used in document summarization (Barzilay and Lapata, 2005), answer extraction and question classification (Li., Roth and Small, 2004). (Barzilay and Lapata, 2005) propose an entity-based method to improve the coherence of a generated summary, assuming that the coherence of a summary is highly correlated with the entity distribution in a document. Our method makes a similar assumption that entities are more thematic and play an important role in document summarization.

WordNet has been widely used in query expansion as a knowledge base. Query classification also lends the knowledge of word relation from WordNet. Similar to (Li., Roth and Small, 2004), we use the hypernym and hyponym information provided in the WordNet.

Utilizing web resources is a hot topic recent years. Compared with the knowledge base (e.g. WordNet) and corpus (e.g. TREC corpus), web resources contains much more abundant and various information. For example, the phrase "cross border" only appear 3 times in the AQUAINT (1998 NYU)[4] corpus, while 92 times in the

[1] Free downloadable from http://gate.ac.uk/.
[2] Query type indicates the entity that the query is looking for. For example, for the query "*Who* criticized World Bank…", the query type is <person>.
[3] The examples are extracted from the query of Topic D301i in DUC2005 evaluation.
[4] TREC QA corpus.

first 100 searching result[5] of AltaVista. In fact, much previous work has paid attention to the web resources. For example, (Hutchinson, 2004) proposed a discourse marker leaning method based on web resources. Meanwhile, (Li, Li, Lu,and Wong, 2005) learns the user preference from the internet.

3. Implicit Entity Mining

Assume a query Q includes a set of phrases (ph) and a phrase is composed of a set of words (w), i.e. $Q=\{ph_i, i=1,\ldots,m\}$, $ph=\{w_i, i=1,\ldots, n\}$. The referred and associated entities are extracted from the query as long as they fall into one of the five types, i.e. $E=\{e_i, i=1,\ldots,5\}=\{person, organization, location, time, number\}$

3.1. Referred Entity (RE) Extraction

Referred entities are identified with WordNet. The referred entity of a phrase (ph) indicated by $RE(ph)$ is mined as follows.

[r1] If the phrase is indexed by WordNet, then the referred entities of the phrase can be inferred from the hypernyms. That is

If $O = Hyper(ph) \cap E \neq \phi$, then $RE(ph) = O$.

For example, for the word "perpetrator", its *direct hypernym* in WordNet is,

wrongdoer, offender (a person who transgresses moral or civil law)

Then, *RE*("*perpetrator*")=*Hyper*("*perpetrator*") = {*person*}

[r2] If the phrase has not been indexed by WordNet, then the referred entities of the phrase are the union of the referred entities of each non-stop word included in the phrase. That is

If $O = Hyper(ph) \cap E = \phi$

Then $RE(ph) = \bigcup_{i=1,\ldots,n} Hyper(w_i) \cap E, w_i \in ph$

The above *Hyper*() is the hypernym set of a term[6] in WordNet. The referred entities of the query Q are the union of the referred entities of each phrase included in the query Q.

$$RE(Q) = \bigcup_{i=1,\ldots,m} RE(ph_i)$$

3.2. Associated Entity (AE) Extraction

Web provides sufficient data to overcome the problem of resource shortage in Question Answering [4]. The associated entities of a phrase are extracted based on the co-occurrences of the phrases and the entities gathered from the website "www.altavista.com". For each phrase, we process the first 100 results retrieved by Altavista. Here is an example.

Table 1. The co-occurrence of the phrase "cross border" and the entity <location>.

	Cross border	¬Cross border
Location	18	74
¬location	3	5

The phrase "cross border" often co-occurs with entity <location>, as illustrated in Table 1. It occurred 21 times in the first 100 results returned by AltaVista, in which the entity <location> occurres 18 times.

Assume a set of entities $\{e_i, i=1,\ldots,v\}$ are extracted from the website with the phrase (ph). Two schemes are considered to weigh the importance of the associated entities of phrase.

[s1]. Threshold Scheme. The entities are equally weighted if the co-occurrences of the phrase and the entity are more than 2.

[s2]. Probability Scheme. each entity is assigned with a different weight which is the probability $P(e|Q)$.

$$P(e_i \mid ph) = \frac{\#(e_i, ph)}{\#(ph)}$$

$$P(e_i \mid Q) = \frac{\sum_{j=1}^{j=n} P(e_i \mid ph_j)}{\sum_{k=1}^{k=5}\sum_{j=1}^{j=n} P(e_k \mid ph_j)}$$

3.3. Associated Entity Pattern (PER) Extraction

In associated entity extraction, we find that some entity patterns also co-occur frequently with certain phrases. For example, the pattern "<location>… <location>" often co-occur with the phrase "cross border" in a sentence. So we extract these associated entity patterns in the same way as we do in associated entity extraction. If the co-occurrences are less than 3, the pattern is ignored.

4. Implicit Entity Extraction Scheme

Implicit entities are extracted based on phrases (or terms). We first decompose the query into a set of phrases. In our work, the terms are extracted from a query by using an English parser MINIPAR[7]. MINIPAR decomposes a query into a set of <term, relation, term> triples. The following criteria are used to collect terms which are then used to mine the implicit entities:

[c1] The term cannot be a stop-word;

[c2] The category of a term is noun (N), noun phrase (NN), adjective (A) or Verb (V).

[c3] The relation between terms can be noun-noun ("nn"), subject-verb ("subj"), verb-object ("obj"), or adjective-noun ("mod").

For example, for the Topic D301i query of DUC 2005:

> Q: International Organized Crime
>
> Identify and describe types of organized crime that crosses borders or involves more than one country.

[5] Searching with the keywords "cross border".

[6] Here a term can be a phrase or a word.

[7] Free downloadable from http://www.cs.ualberta.ca/~lindek/minipar.htm.

Name the countries involved. Also identify the perpetrators involved with

Terms extracted from the query are then:

International organized crime; describe type cross border; involve country one country; involve country involve name; involve country identify perpetrator; involve perpetrator include individual; individual organization

Afterwards, the implicit entities are extracted based on these phrases. The system flowchart is shown in Figure 1. The referred entities and the associated entities are extracted form WordNet and web resources, respectively.

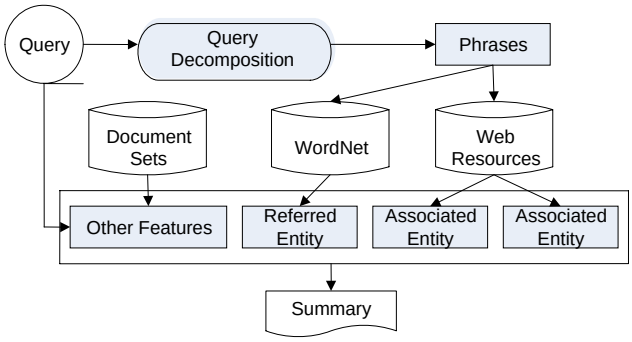

Figure 1. System flowchart

5. Evaluation

With the document sets and the manually created model summaries provided by the DUC 2005, we evaluate the impact of implicit entity mining with ROUGEs. The DUC 2005 provides fifty clusters. Each cluster includes one query and a set of documents. From the fifty queries, 361 phrases are extracted, 62 phrases are found to refer or associate to the entities, and 14 phrases contain the associated entity patterns. Five types of features are used in the experiments: term-based features[8] (F_1), explicit entities and query type (F_2), referred entities and associated entities with threshold scheme (F_3), referred entities and associated entities with probability scheme (F_4), and associated entity patterns (F_5). We use the following function to score each sentence and select the top sentences into the summary. The sentence longer than 50 words or shorter than 9 words are all discarded.

$Score(s) = \sum_{i=1}^{i=5} \lambda_i \alpha_{F_i}$, λ_i are the weights of the features and the default value is 1.0. α_F is scoring methods of each feature. For each type of feature, we assume that the query Q has x features and the sentence S has y features.

$$\alpha_F = \frac{1}{x} \sum_{j=1}^{j=x} \sum_{k=1}^{k=y} I(f_j^Q, f_k^S)$$

$$I(f_j^Q, f_k^S) = \begin{cases} 1 & f_j^Q = f_k^S \\ 0 & f_j^Q \neq f_k^S \end{cases} \quad F \neq F_4;$$

$$I(f_j^Q, f_k^S) = \begin{cases} P(f_j^Q \mid Q) & f_j^Q = f_k^S \\ 0 & f_j^Q \neq f_k^S \end{cases} \quad F = F_4.$$

Table 2. Experimental results[9]

Exp	Features	ROUGE-2	ROUGE-SU4
1	F1,2	0.06844	0.12654
2	F1,3	0.06964	0.12744
3	F1,4	0.06744	0.12508
4	F1,2*,3	0.06917	0.12739
5	F1,2*,4	0.06857	0.12662
6	F1,2*,3*,5*	0.06917	0.12739

As shown in Table 2, the improvement is achieved by considering the implicit entities in addition to explicit entities (Exp2 vs. Exp1). Furthermore, the threshold scheme outperforms the probability scheme (Exp2 vs. Exp3). However, the associated entity pattern makes no contribution. When looking closely, we find that only 3.8% of phrases associate the patterns.

Note that in the experiment we assign some features 0.5, e.g. Exp4, 5 and 6. The comparison between weight 0.5 and 1.0 are listed in Table3.

Table 3. Comparison between different weight assignments.

Exp	Features	ROUGE-2	ROUGE-SU4
1	F1,2*,3	0.06917	0.12739
2	F1,2,3	0.06844	0.012646
3	F1,2*,3*,5*	0.06917	0.12739
4	F1,2,3,5	0.06744	0.12508

Feature 2, 3, 4 and 5 are all entity-related features. Assigning excessive weight on these features will impose negative impact on the results (Exp1 vs. Exp2; Exp3 vs. Exp4).

6. Conclusion and Future Work

In this paper we focus on implicit entity mining in queries. WordNet and web resources are used for mining. When apply our method in the DUC 2005 summarization task, the evaluations show that the implicit entities in queries are helpful in query-based summarization.

In the future, we plan to further investigate on the logical relation between entities and apply them in summarization.

7. Acknowledgment

The work presented in this paper is supported by Hong Kong Research Grant Council (RGC) (project number CERG PolyU5181/03E).

8. References

Wenjie Li, Wei Li, Baoli Li, Qing Chen, Mingli Wu (2005). The Hong Kong Polytechnic University at DUC

[8] Here a term can be a word or an entity.

[9] The weights of the features with "*" are 0.5.

2005. To appear in the notebook of the DUC 2005 workshop.

Li X., Roth D., and Small K (2004). The Role of Semantic Information in Learning Question Classifiers. Proceedings of the International Joint Conference on Natural Language Processing.

Regina Barzilay, Mirella Lapata (2005). Modeling local coherence: an entity-based approach. In Proceedings of the 43rd Annual Meeting of the Association for. Computational Linguistics (ACL'05), pages 141–148, University of Michigan.

Ben Hutchinson (2004). Mining the web for discourse markers. In the Proceedings of the Fourth Interna-tional Conference on Language Resources and Evaluation Pages 407-410.

Wei Li, Wenjie Li, Qin Lu, and Kam-Fai Wong (2005). A Preliminary Work on Classifying Time Granularities of Temporal Questions. The 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005), Oct. 10-15, 2005 at Jeju Island, Korea, LNAI 3651, pp. 414 – 425.