TongKey at Entity Track TREC 2011:

Related Entity Finding

Zhengcai Pan
Department of Computer Science and Technology
Shnu University, Shanghai, 201100
stevepzc@gmail.com

Haiguang Chen
Tongkey Network Technology Co,Ltd
Room 2205, No 9116, Humin Road Shanghai, 201100
hr@shtknet.com

ABSTRACT

This paper presents our work done for the TREC 2011 Entity track. A retrieval model was proposed for the task of related entity finding. This model consists of several parts: In order to get more accurate document collection, query analysis method was utilized to format the narrative of each query. Then, our dataset was generated by using ClueWeb09 API[2]. Moreover, we employed the NER tools and text pattern recognition to extract entities from this processed dataset. In particular, the types of target entities are not so well-defined as last year. Therefore, a specific classifier trained by employing Wikipedia titles and category was utilized to identify the categories of target entities. To find related entity homepages and supporting documents, a set of feature-based methods were applied.

1. INTRODUCTION

In entity track of TREC 2011, the main task of the Related Entity Finding (REF) task is elaborated as follows:

Given an **input entity**, by its name and homepage, the **type of the target entity**, as well as the **nature of their relation**, described in free text, **find related entities** that are of target type, standing in the required relation to the input entity.

In this paper, our retrieval system was proposed and we submitted two runs through exploiting more sophisticated techniques to complete our task. In particular, the probability model is used to train the classifier for detecting the types of target entities. Authority pages were mined deeply and more entity filtering rules were designed to shift entities that had been extracted. For entity homepage finding, some feature-based methods which had been proven effective were utilized.

2. APPROACHES

2.1 System Overview

We complete our experimental system model as pipeline architecture by devising from IR questions answering (QA) framework. Our framework includes four major components: query generator method, dataset processing, entity extraction method, entity ranking algorithm, homepage and supporting document finding. The following figure displays the outline of the individual components of our framework for REF task.

[1]http://nlp.stanford.edu/software/CRF-NER.shtml

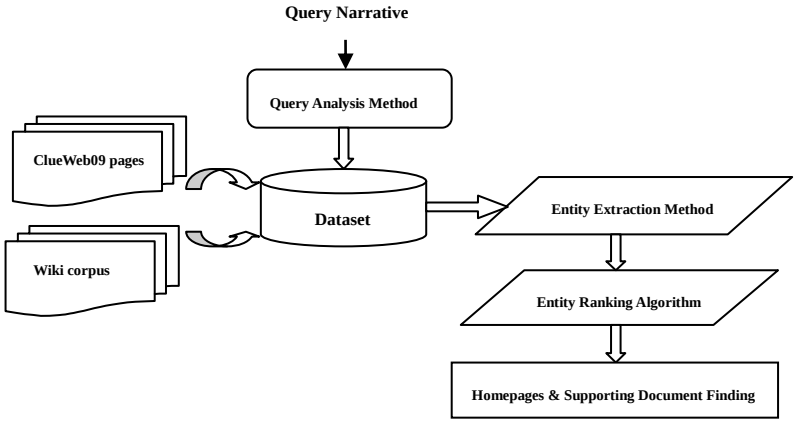

Figure 1: Framework of Our Retrieval System

2.2 System Framework

2.2.1 Query Analysis & Dataset Processing

To improve the accuracy of target-entity, we applied a set of strict strategies to analyze queries syntactically and semantically. A stop words list [3] was used to exclude stop words and punctuation from queries. Moreover, the plural of individual noun or verb in a query is processed to singular. In particular, wordnet[3] was employed to form our query according to the keywords extracted from the given query. For instance, we used "*recording company sell kingston Trio song*" instead of "*what recording companies now sell the kingston Trio's songs?*".

We used top 15 pages returned by the ClueWeb09 API[2] and source entity's Wikipedia pages as our dataset. Then, we processed the dataset by excluding some unnecessary information such as html tags.

2.2.2 Entity Extraction & Entity Ranking

Stanford NER tools are widely used to recognize entities with type of person, organization or location. Because the type of target-entity is not clearly provided for each query this year, it is necessary to divide the types of all the target entities into categories. In TongKeyEN1, we used appropriate Wikipedia titles and categories to train classification model for the given types of the target entities into three categories. In TongKeyEN2, we manually select the category for each target entity. At the same time, text pattern recognition was used to be enhanced for entity extraction.

On the entity ranking step, we refer to the ranking algorithm on[1] to compute the ranking scores of the candidate entities such as tf-idf, multiply keywords, association rules and so on. In particular, The factors of tables and lists are further considered to extract candidate entities from Wikipedia pages.

2.2.3 Entity Homepages and Supporting Documents Finding

For homepages finding, we proposed a novel method to detect homepages for the target entites. We got top 5 pages with ClueWeb09 API[2] by target-entity names, and 5 pages by the combination of input entity and target entity names. Thus, 10 pages were collected as homepage set. Then, we calculate weights according to multiply features such URL format, page content and so on. The page with highest weight was selected to be homepage for corresponding entity.

[2]http://lemurproject.org/clueweb09.php/

[3]http://wordnet.princeton.edu/

Target entity supporting document is the page from where the entity extracted. For multiple documents containing the same target entity, the homepage of the source entity and its Wikipedia page was given a prior consideration.

3. EXPERIMENTAL RESULTS

On the REF task, we submit two runs for the official evaluation:

- TongKeyEN1: Run using the methods described in the previous sections automatically;
- TongKeyEN2: Run using the previous methods with manually select category for target entities and keywords in the entity ranking step;

	num_rel_ret	map	R-prec	P5	P10	P20	P100
TongKeyEN1	163	0.1209	0.1972	0.2760	0.2260	0.1420	0.0326
TongKeyEN2	183	0.1266	0.1984	0.2760	0.2260	0.1490	0.0366

Table 1: Results for num_rel_ret, map, R-prec, P@5, P@10, P@20, P@100

Table 1 lists the results of our runs for REF task. Almost all the indicators in TongKeyEN1 is lower than TongKeyEN2 such as num_rel_ret, map, R-prec, P@20 and P@100. The problem might be caused by classification model and strict text pattern recognition. Therefore, the type classifier needs to be improved for better performance.

Figure 2 shows R-precision scores of our runs and best runs of TREC 2011 for each topic. From this figure, we can see our system returns highest scores for 12 topics while 10 topics got zero. It may because we depend too much on NER tools and the text recognition pattern is not able to cover most cases.

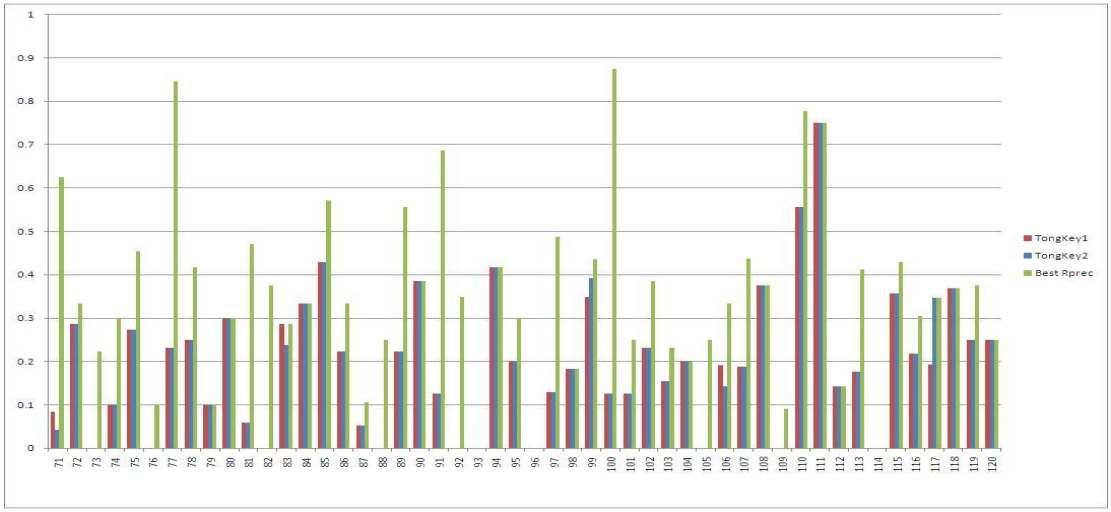

Figure 2: R-precision for each topic

4. CONCLUSIONS

According to the evaluation, our system get high scores on location category and some organization category topics, while other topics get low scores even zero. This proves our system is sensitive to entity type. After the compare of TongKeyEN1 and TongKeyEN2 evaluation results, keywords of the narrative are proved to be useful for entity ranking. Our system returns zero related homepage for 10

topics, it may be caused by limited dataset and too much depending on NER tools. In-depth study and analysis need to be done in the future.

ACKNOWLEDGEMENT

We thank to Dong Wang and Tongkey laboratory, this work had been received a lot of help from them.

REFERENCES

[1] Dong Wang, Qing Wu, Haiguang Chen, Junyu Niu. A Multiple-Stage Framework for Related Entity Finding: FDWIM at TREC 2010 Entity Track. In :Gaithersburg, MD, 2010.
[2] Y. Fang, L. Si, Z. Yu, Y. Xian, and Y. Xu. Entity Retrieval with Hierarchical Relevance Model, Exploiting the Structure of Tables and Learning Homepage Classifiers. In : Gaithersburg, MD, 2009.
[3] Silva, C., Ribeiro, B. The importance of stop word removal on recall values in text categorization. In Proceedings of the International Joint Conference, Neural Networks, 2003.