PRIS at TREC 2011 Entity Track: Related Entity Finding and Entity List Completion

Zhanyi Wang, Wenlong Lv, Heng Li, Wenyuan Zhou, Li Zhang,
Xiao Mo, Liaoming Zhou, Weiran Xu, Guang Chen, Jun Guo
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing, China
wangzhanyi@gmail.com

Abstract. The group of PRIS focuses on both tasks in Entity Track this year, Related Entity Finding (REF) and Entity List Completion (ELC). This paper reports the approaches to the two tasks. In REF, three points are improved based the method of last two year: building entity lexicons including more information, introducing a distance algorithm between keywords and entities to entity ranking, allocating homepages in a deeper and more reasonable way. The Entity Activation Force (EAF) and Affinity Measure are used in ELC task for completing and reordering the entity list. The evaluation of experimental results shows that the performance is better than previous ones significantly.

1. Introduction

As the aims of last two years, Entity track is to evaluate entity-related searches on Web data[1]. The track of this year includes two main tasks, Related Entity Finding (REF) and Entity List Completion (ELC). REF is defined 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 [1]. The key changes introduced to the 2010 edition of the REF task are as follows[2]:

• Only primary homepages are accepted, i.e., relevance is binary.

• For each answer, a (single) supporting document is required.

• Wikipedia pages are (still) not accepted as entity homepages, but they can be supporting documents.

• Target type is not limited anymore to the four high-level entity types (person, organization, location, product). The target type is extracted from the narrative and is always given in singular form.

[1] http://ilps.science.uva.nl/trec-entity/guidelines/guidelines-2010/
[2] http://bit.ly/entity2011-guidelines

• Groups that generate results using Web Search Engines are required to submit an obligatory run, using the Lemur ClueWeb Online Query Service.

The ELC task is defined as follows: Given an information need and a list of known relevant entity homepages, return a list of relevant entity URIs from a specific collection of Linked Open Data. ELC addresses essentially the same task as REF does: finding entities that are engaged in a specific relation with an input entity. There are two main differences to REF. First, entities are not represented by their homepages, but by a unique URI (from a specific collection, a sample from the Linked Open Data cloud). Second, a number of entity homepages (i.e., ClueWeb docIDs) are made available as part of the topic definition, as examples of known relevant answers.

In REF, the framework is similar as the description of last year's report [2]. This year we pay more attention to entity extraction and homepage ranking. Aiming at them, we improve our framework in three points: building entity lexicons including more information, introducing a distance algorithm between keywords and entities to entity ranking, allocating homepages in a deeper and more reasonable way. In ELC, we introduce a new statistic named Entity Activation Force (EAF). It is used to compute affinity measure between two entities for completing and reordering the entity list.

The report is organized as follows. Section 2 describes our methods of entity extraction. Section 3 introduces the distance algorithm to entity ranking. Section 4 proposes the improvement of allocating homepages. The key arithmetic for ELC is presented in Section 5. Submitted runs show in Section 6 and Section 7 gives the conclusion and future work.

2. Entity Extraction

This year we exact entities in various ways as before. It is described as the figure below:

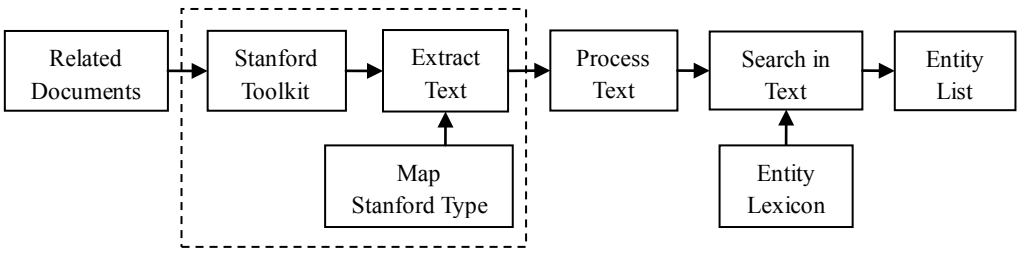


Figure 1. Flowchart of Automatic Entity Extraction in Task REF.

As the figure shown, some related documents retrieved by search engine. In our runs, one of them uses Lemur ClueWeb Online Query Service to abide by the guideline. The others are developed with the help of Google. Stanford toolkit analyzes three types of entities: person, organization and location. Other types did not apply in this step, so the frame is dotted.

Unlike the method of last year in this section, in order to get accurate and normative entities, we extract the snippets about the entity instead of the entity itself. If the type that recognized by Stanford toolkit is satisfactory, we reserve the text around the entity for further extraction using lexicon.

We create an entity lexicon with richer types. Besides three basic types, more than twenty kinds of lexicon are created. The source data is Wikipedia. Wikipedia has rich entities and category labels. According to topics, the map of topic and entity is set up manually in advance. A topic maybe corresponds to one or more labels. For example, "manufacturer" ~"manufacturer", "author"~"writer", "film"~"movie", etc. The type of person and organization maps some new labels besides rules proposed by University of Amsterdam [3]. Related Wikipedia pages are collected and the titles are taken as entity names simply.

The process of entity extraction in ELC task this year is similar to REF. The topic is the same as REF of last year. We adjust the scale of document for extraction, refine the Wikipedia lexicon and filter entities that are too short.

3. The Distance Algorithm between Keywords and Entities

Because the Document-Centered Model (DCM) showed very good performance, we still adopt it as the basic retrieval model. To improve the accuracy, we propose a distance algorithm between keywords and candidate entities. The algorithm is described as follows.

Step1. Several keywords (1~5 generally) are picked up manually from narrative field in the topic. According to Word Activation Force (WAF) and affinity measure [4], more keywords are expanded by in British National Corpus (BNC). We take top 5 words in the affinity list of a keyword and make up the keyword set K. Given a topic t in topic set T and a candidate entity e in the whole set E, entities and keywords generate pairs set $\{(e,k)|e \in E, k \in K\}$.

Step2. Distance d between each k and e are counted in all relative documents C. The distance d is defined as c+1. c is the word count between k and the nearest e. The distance score of an entity is

$$S_{Distance} = \max_{k \in K, m \in C} \left(1 - \frac{d(e,k,m)}{l_m} \right)$$

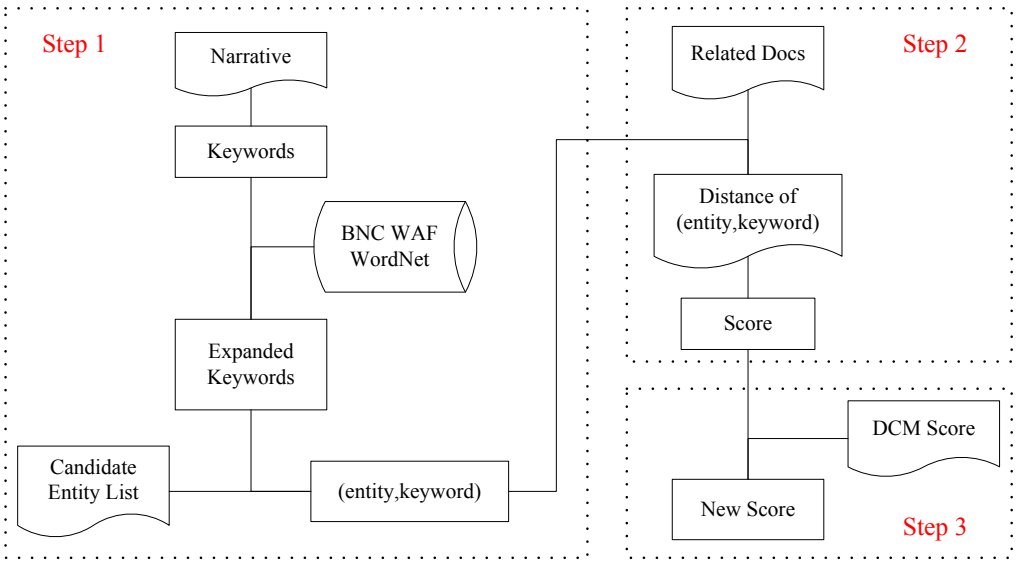
where $d(e,k,m)$ is d in the document m. l_m is the length of m.

Step3. Now an entity has two values. One is from DCM results. The other is from step 2. Then a new score is merged as

$$S_{new} = \alpha \cdot S_{DCM} + (1-\alpha) \cdot S_{Distance}.$$

Note that this method only ranks the entities in DCM result. If an entity is not in DCM result, but appears in the distance result, it should be ignore. On the other hand, if an entity is not in the distance result, the final score keeps the original DCM score. That means the parameter α is set as 1 in both cases. In generally, 0<α<1.

The algorithm is implemented by the following flowchart.



Figrue 2. Flowchart of distance algorithm between keywords and entities.

4. Homepage Identification and Ranking

The main problem is how to identify a homepage for an entity from large amounts of data. We define some patterns and values as Table 1:

Table 1. Homepage patterns and values.

Homepage pattern	Pattern value	Example
consist of entity name directly	10	www.entity.(com|net|org|edu)
include entity name	2	www…entity…
in the title field	2	<title>…entity…</title>
Top 1 in search engine results	3	-

If a URL satisfies one or more patterns in the table, it may be a homepage of an entity. According to the importance, the patterns are given appropriate values. Entity name is generalized here. Specifically, the styles are listed in Table 2(take Michael Jackson as an example).

Last year, if a webpage was taken for a homepage, it replaced entities directly. But the rank of homepages should not be the same as the order of entities, because it is affected by the relation between homepages and entities. This time we combine the two parts as

$$S(Q,H) = \beta \cdot S(Q,E) + (1-\beta) \cdot S_h$$

Table 2. Different patterns for entity names.

Entity name pattern	Example
Capitalize each word	MichaelJackson
Uppercase	MICHAELJACKSON
Lowercase	mickaeljackson
Uppercase abbreviation	MJ
Uppercase abbreviation	mj
A part of words	Michael
Words linked with connector	Michael-Jackson, MICHAEL_JACKSON, mickael jackson

$S(Q,E)$ is the value generated from DCM. S_h is the normalized pattern value. β is a variable parameter. In experiments, it is set as 0.3.

In ELC task, homepages are in the Sindice dataset. We take entities as keywords and analyse the searching results in the system. The process of identifying homepages is shown as Figure 3.

By our observation on the dataset, the DBpedia pages are often considered as homepages. For DBpedia pages, the homepage value is written as

$$S_{Dbpeida} = \frac{N_{match}}{N_{total}} \cdot \frac{L_{match}}{L_{total}}.$$

For an entity e, N_{total} is the number of words in e. N_{match} is the number of matched words. L_{total} is the character length of e. L_{match} is the length of matched characters. For entities that are not in a DBpedia page, the first result in the relevant documents is taken as the homepage.

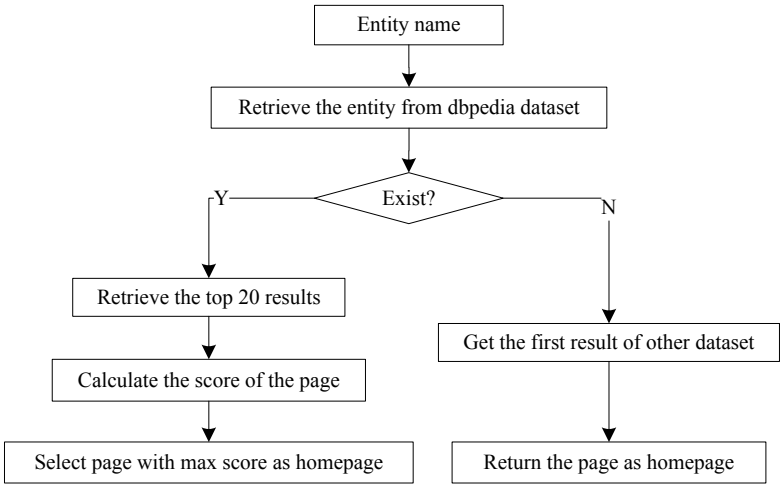


Figure 3. Flowchart of homepage identification in ELC.

5. Entity Activation Force in ELC

The task of ELC refers more to analyzing relations among entities. The entity list is completed and reordered by the relations. So besides the methods in REF task, a novel algorithm for the establishment of entity associations is introduced. Entity Activation Force (EAF) is developed

from the activation force statistics which proposed by J. Guo et.al [4]. The main idea of EAF is to weight the links between words and entities. Here, words are considered as the context of entities. We aim at measuring the activation force from an entity to their context words and calculating the affinity of entities. Finally, the entities whose affinities are high to the given entities are helped to improve the entity list.

Take the forward context as an example, EAF is defined as

$$eaf_{ew} = g \cdot \frac{(f_{ew} / f_e) \cdot (f_{ew} / f_w)}{d_{ew}^2}.$$

In this formula, eaf_{ew} means the EAF from an entity e to words w. Correspondingly, the EAF from words to the entity is also existed. f_e is the entity frequency in all the training data. f_w is the word frequency. f_{ew} is the co-occurrence amount of e and w in a certain distance scale. d_{ew} is the average distance between the entity and the word. In the type of webpage or plain text, distance minus one equals the number of words between them. g is a parameter. Note that the form is pretty similar to the law of universal gravitation. So we name it a kind of "Force". Because the EAF network is a directed graph, the EAF matrix is asymmetric.

Moreover, the entity affinity measure is proposed. The affinity between two entities e_i and e_j is defined as

$$A_{ij}^{eaf} = \left[\frac{1}{|K_{ij}|} \sum_{k \in K_{ij}} OR\left(eaf_{ki}, eaf_{kj}\right) \cdot \frac{1}{|L_{ij}|} \sum_{k \in L_{ij}} OR\left(eaf_{il}, eaf_{jl}\right) \right]^{\frac{1}{2}}$$

where $K_{ij}=\{k|eaf_{ki}>0 \text{ or } eaf_{kj}>0\}$, $L_{ij}=\{l|eaf_{il}>0 \text{ or } eaf_{jl}>0\}$, $OR(x,y)=min(x,y)/max(x,y)$.

The entity list can be modified by the entity affinity measure. For example, the 37th query is to find people that appeared on the Tavis Smiley show in December 2008. If we are sure that "chris_cillizza" is a correct answer, we give more weight to the entities "thomas_friedman" and "malcolm_gladwell" whose affinity to "chris_cillizza" are very high. Before we fuse this algorithm, the two entities are just 14th and 30th in our REF list.

6. Experiments and Results

Table 3 is REF result of all our runs. PRISREF1 doesn't include the distance algorithm. PRISREF2 uses snippets instead of the full text of relevant documents. PRISREF3 includes all the above algorithms. PRISREF4 adjusts parameters of DCM.

Table 4 shows our REF results in recent three years. Hp_ret means the percentage of homepages that we find. No_hp_topic is the percentage of topics whose homepages we don't find. All the metrics goes better significantly year by year.

ELC results are not yet published at the time of writing.

Table 3. REF runs of PRIS in 2011.

Run_id	MAP	Num_rel_ret	P@5	P@10
PRISREF1	0.2509	310	0.4280	0.3340
PRISREF2	0.2329	300	0.4080	0.3000
PRISREF3	0.2450	310	0.4160	0.3180
PRISREF4	0.2448	326	0.4440	0.3260

Table 4. REF runs of PRIS from 2009-2011.

Year	MAP	NDCG	P@10	Hp_ret	No_hp_topic
'11	0.2509	-	0.3340	0.4512	0.0400
'10	0.1607	0.2846	0.2489	0.4005	0.1702
'09	-	0.0892	0.0150	0.0180	0.3500

7. Conclusions

This paper reported our approach to the task of Entity Track 2011. We proposed some mining and ranking methods for entities and homepages. We extracted entities in a deeper way. We merged a distance algorithm for entity ranking in the model. We got a breakthrough in homepage identification and ranking. In ELC task, we introduced a novel algorithm for completing and optimizing the entity list. In the future, we will pay more attention to the ELC task and the study of entity associations.

Acknowledge

This work was supported by NSFC (No.60905017).

References

[1] K. Balog, P. Serdyukov and A. P. Vries. Overview of the TREC 2010 Entity Track. In the Eighteenth Text REtrieval Conference Proceedings (TREC 2010).

[2] Z. Wang, C. Tang, et. al. PRIS at TREC 2010: Related Entity Finding Task of Entity Track. In the Eighteenth Text REtrieval Conference Proceedings (TREC 2010).

[3] R. Kaptein, M. Koolen, and J. Kamps. Result diversity and entity ranking experiments: Text, anchors, links, and wikipedia. In the Eighteenth Text REtrieval Conference Proceedings (TREC 2009). NIST Special Publication, 2010.

[4] J. Guo, H. Guo, and Z. Wang. An Activation Force-based Affinity Measure for Analyzing Complex Networks. Sci. Rep. 1, 113; DOI:10.1038/srep00113 (2011).