# Building Large Scale Relation KB from Text

Junfeng Pan, Haofen Wang, and Yong Yu

APEX Data & Knowledge Management Lab
Shanghai Jiao Tong University
800 Dongchuan Rd., Shanghai 200240, China
{panjf, whfcarter, yyu}@apex.sjtu.edu.cn

**Abstract.** Recently more and more structured data in form of RDF triples have been published and integrated into Linked Open Data (LOD). While the current LOD contains hundreds of data sources with billions of triples, it has a small number of distinct relations compared with the large number of entities. On the other hand, Web pages are growing rapidly, which results in much larger number of textual contents to be exploited. With the popularity and wide adoption of open information extraction technology, extracting entities and relations among them from text at the Web scale is possible. In this paper, we present an approach to extract the subject individuals and the object counterparts for the relations from text and determine the most appropriate domain and range as well as the most confident dependency path patterns for the given relation based on the EM algorithm. As a preliminary results, we built a knowledge base for relations extracted from Chinese encyclopedias. The experimental results show the effectiveness of our approach to extract relations with reasonable domain, range and path pattern restrictions as well as high-quality triples.

**Keywords:** Linked Data, Relation Extraction, Expectation Maximization.

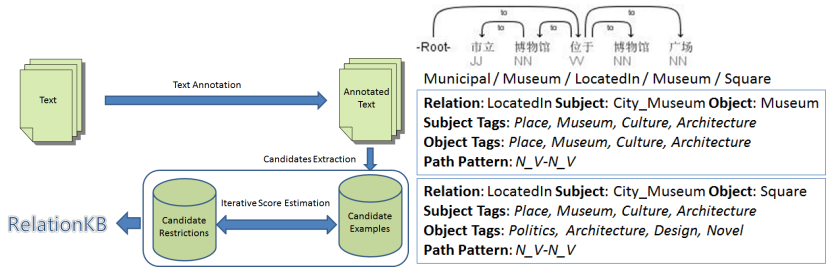
## 1 Introduction

In recent years, many knowledge bases, such as DBpedia[1], YAGO[6], Zhishi.me[5], have been published on the Web in the form of linked data, which are very useful for both human reading and machine consumption. Comparing with the number of different entities in these knowledge bases, there are only a few distinct relations. Furthermore, these knowledge bases only extract data from structured or semi-structured data sources without considering implicit knowledge from unstructured text, which is in a huge and increasing amount on the Web. On the other hand, open information extraction, such as Machine Reading[4] and Never-Ending Language Learning[2], focuses on extracting entities and their relations from text at the Web scale.

In this paper, we are motivated to build a knowledge base of relations extracted from text by leveraging open information extraction techniques. Moreover, for each relation, we extract not only subject-object examples but also high level restrictions such as the domains and ranges from text. Both information are quite useful to describe relations, which can be used for further natural language processing training or high-quality ontologies for additional extraction. To extract such information, we adapt a novel algorithm based on Expectation Maximization (EM). And an experimental relation knowledge base is built to show the effectiveness of our algorithm.

## 2 Building the Relation Knowledge Base

As shown in Figure 1(a), the process of building relation knowledge base has three main steps: text annotation, candidates extraction and iterative score estimation.




(a) Architecture Overview (b) Text Annotation / Candidate Extraction

**Fig. 1.** Building Relation KB

The text annotation step includes tokenization (word segmentation for Chinese), POS tagging and dependency parsing on top of the raw input text. The annotated text is used as the input for the candidate extraction step.

According to the concept of dependency grammar, we can directly extract candidate examples from the annotated text. We identify the relations in the sentences according to POS tags, and for each relation, we enumerate nouns in its subtrees as subject or object candidates. We also extract candidate restrictions in the candidates extraction step. For each candidate example, we tag the head nouns of the subjects and objects to form the candidate domain and range of the relations. We also extract candidate path patterns from subject-relation-object paths on the dependency trees. Figure 1(b) shows two candidates, including examples and restrictions, are extracted from an annotated Chinese sentence means *the Municipal Museum is Located in the Museum Square*.

In the score estimation step, we group the candidate examples and restrictions by relation, and process each relation independently. For each relation $v$, we group the candidates from one sentence into one candidate group. Suppose there are $n$ candidate groups for $v$ and $m_i$ candidates in group $c_i$. In one group, each candidate $c_{ij}$ contains the extracted candidate subject $s_{ij}$ with tags $T(s_{ij})$, object $o_{ij}$ with tags $T(o_{ij})$, and path pattern $p_{ij}$. Figure 1(b) can be seen as a two-candidate group for relation $LocatedIn$.

We need a score function $f_E(s_{ij}, o_{ij}|v)$ to judge whether the example $(s_{ij}, o_{ij})$ is a credible subject-object pair for $v$. It is reasonable that the credibility of the example is positively correlated to the credibility of its tags and path pattern so we use Equation (1) to estimate the score. Here $N_w(t)$ is the number of words having tag $t$. And we also introduce three functions $f_D(t_d|v)$, $f_R(t_r|v)$, $f_P(p|v)$ to indicate the credibility of the tag $t_d$ in the domain, the tag $t_r$ in the range, and the path pattern $p$ for $v$.

$$f_E(s_{ij}, o_{ij}|v) = f_P(p_{ij}|v)\Big(\sum_{t_d \in T(s_{ij})} \frac{f_D(t_d|v)}{N_w(t_d)}\Big)\Big(\sum_{t_r \in T(o_{ij})} \frac{f_R(t_r|v)}{N_w(t_r)}\Big) \tag{1}$$

Intuitively, $f_D(t_d|v)$ should be the real count of $t_d$ over the maximum possible count. We assume that each candidate group is worth 1 count (which implies the maximum possible count is $n$) and allocate it for each candidate in one group according to $f_E(s_{ij}, o_{ij}|v)$. The way to compute $f_R$ and $f_P$ is almost the same as $f_D$. Details are given in Equation (2).

$$f_L(t|v) = \frac{1}{n} \sum_{cond} \frac{f_E(s_{ij}, o_{ij}|v)}{\sum_{k=1}^{m_i} f_E(s_{ik}, o_{ik}|v)} \tag{2}$$
$$(L, t, cond) \in \{(D, t_d, t_d \in T(s_{ij})), (R, t_r, t_r \in T(o_{ij})), (P, p, p = p_{ij})\}$$

According to Equation (1) and (2), the scores of the examples and the restrictions are interdependent. It is natural to design an EM[3] algorithm to estimate them simultaneously:

1. Initialize $f_E(s_{ij}, o_{ij}|v) = \frac{1}{m_i}$.
2. Update $f_D$, $f_R$, $f_P$, $f_E$ iteratively until convergence:
   (a) **E-step:** For every tag and path pattern, update $f_D(t_d|v)$, $f_R(t_r|v)$, $f_P(p|v)$ using Equation (2).
   (b) **M-step:** For every example, update $f_E(s_{ij}, o_{ij}|v)$ using Equation (1).

Finally we can build the relation knowledge base by the examples (subject-object pairs) and the restrictions (tags for domain and range as well as path patterns) for each relation according to the scores of these candidates.

## 3 Preliminary Results on Chinese Encyclopedias and Zhishi.me

We built an experimental relation knowledge base from the abstract text of all the entries in three online Chinese encyclopedias, Wikipedia in Chinese, Baidu Baike and Hudong Baike. There are 2,517,826 pieces of text and 20,637,524 sentences in total.

In addition to the text, we use the property `zhishi:category` at a Chinese Linked Open Data, Zhishi.me which extracts the categories for the entities from the above three online encyclopedias. We used them to tag the extracted entities. There are a total of 985 tags and 3,109,448 words that have at least one of these tags.

Our experimental relation knowledge base contains 7,097 relations from the text in total. For one relation, there are 107.46 tags in domain, 103.13 tags in range, 9.68 path patterns and 276.03 subject-object pairs in average.

We sampled 40 relations for human evaluation. For each relation, the domain, the range and the examples can be seen as three ranked lists according to the scores of the elements in them. We use the average precision (average of the precision value obtained for the top k elements, each time a relevant/correct element is retrieved) as the metric. The mean average precision is 0.788 for the domain (top 5), 0.865 for the range (top 5), and 0.657 for the examples (top 15). The average precision distribution on the sample relations is shown in Figure 2.

We have found errors are mainly caused by the following reasons: (1) Some words has wrong tags; (2) Some relation phrases are not actually relations (maybe just modifiers to nouns) in the sentence; (3) Some extracted noun phrases are incomplete or linked

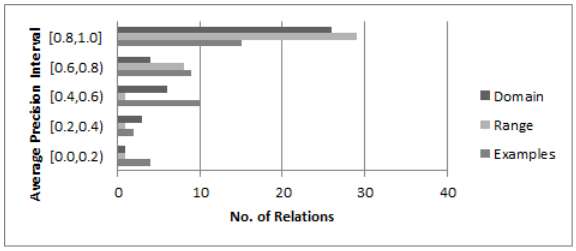



**Fig. 2.** The Average Precision Distribution on the Sample Relations

to wrong entities. These kinds of errors indicate how to improve the quality of our relation knowledge base in the future (e.g. to improve tag cleaning, relation identification, entity linking and disambiguation, etc).

We also provide a simple web site for users to browse our experimental relation knowledge base at `http://relationkb.apexlab.org`. There are two services in the web site, a lookup service and a SPARQL query service.

## 4 Conclusion and Future Work

In this paper, we leverage open information extraction and propose an EM algorithm to build a knowledge base which contains the examples and restrictions of the relations from text. Also we give some preliminary results in Chinese to show effectiveness of our algorithm. In the future, we are planning to use advanced method to identify the relations and better entity linking algorithm to improve the quality of the relation knowledge base. In additional, it would be better if we added some structures about the relations, such as relation clusters or relation hierarchies. Finally we are also planning to use our relation knowledge to populate more and more linked data from text and update the relation knowledge base itself when populating.

## References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: a nucleus for a web of open data. In: ISWC/ASWC (2007)
2. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E.R.H., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI (2010)
3. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) (1977)
4. Etzioni, O., Banko, M., Cafarella, M.J.: Machine reading. In: AAAI Spring Symposium (2007)
5. Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., Yu, Y.: Zhishi.me: weaving chinese linking open data. In: ISWC (2011)
6. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW (2007)