Choosing Better Seeds for Entity Set Expansion by Leveraging Wikipedia Semantic Knowledge

Zhenyu Qi, Kang Liu, and Jun Zhao

National Laboratory of Pattern Recognition(NLPR)
Institute of Automation Chinese Academy of Sciences
100190 Beijing, China
{zyqi,kliu,jzhao}@nlpr.ia.ac.cn

Abstract. Entity Set Expansion, which refers to expanding a human-input seed set to a more complete set which belongs to the same semantic category, is an important task for open information extraction. Because human-input seeds may be ambiguous, sparse etc., the quality of seeds has a great influence on expansion performance, which has been proved by many previous researches. To improve seeds quality, this paper proposes a novel method which can choose better seeds from original input ones. In our method, we leverage Wikipedia semantic knowledge to measure semantic relatedness and ambiguity of each seed. Moreover, to avoid the sparseness of the seed, we use web corpus to measure its population. Lastly, we use a linear model to combine these factors to determine the final selection. Experimental results show that new seed sets chosen by our method can improve expansion performance by up to average 13.4% over random selected seed sets.

Keywords: information extraction, seed set refinement, semantic knowledge.

1 Introduction

Entity Set expansion refers to the problem of expanding a small (3~5) set of seed entities to a more complete set which belongs to the same category. For example, a person may give a few well-known US presidents like “Washington”, “Lincoln” and “Roosevelt” as seeds; the entity set expansion system should discover other US presidents such as “Obama”, “Bush” etc. based on the given seeds.

These collections of entities can be used in various applications [1]. For instance, search engines collect large sets of entities to better interpret queries. Question answering systems can use the expansion tools to deal with List questions [2].

There are several researches for solving this problem, like [3][4]. These methods always include two important components: 1) find the candidates which may have the same semantics with the given seeds; 2) estimate the average similarity between each candidate and the given seeds. The item with higher similarity score will be extracted as results. A typical method begins with several seeds (usually 3-5), then it employ context patterns [5][6] or distributional features [7][8] to find entities of the same category in external data sources such as large corpora of text or query logs.

C.-L. Liu, C. Zhang, and L. Wang (Eds.): CCPR 2012, CCIS 321, pp. 655–662, 2012.

Table 1. Seeds greatly influences entity set expansion quality

Concept	MAX	MIN	AVG
California Counties	1.000	0.345	0.753
Countries	0.889	0.008	0.676
Elements	0.983	0.026	0.714
F1 Drivers	0.927	0.000	0.478
Roman Emperors	0.804	0.087	0.529
U.S. States	1.000	0.880	0.967

To study the effect of seeds, we employ a state-of–art set expansion system [9] to evaluate the performance of different seeds. We use 6 benchmark concepts described in Section 5. For each concept we do 10 trials. In each trial, we randomly choose 3 entities as seeds. Table 1 shows the maximum, minimum and average expansion performance of these seeds sets measured by R-precision. We can see that the quality of seeds has a great influence on the expansion performance. Furthermore, previous studies have shown that human editors generally provide very bad seeds [1]. So finding better seeds is very important for entity set expansion.

We believe that there are three factors (semantic relatedness, ambiguity and sparseness) which would affect the entity expansion performance. For example, "Lincoln" is a seed for "US President", but it can also mean a kind of car or a battleship, this ambiguity of seed may bring in mistake. Also, the less a seed appears in the corpus, the harder we can find it. So sparseness impacts the performance too.

In this paper we propose a novel method for selecting better seeds. We select seeds under the following three rules: 1) the selected seeds should have high semantic relatedness among each other; 2) the seeds should have less ambiguity; 3) the seeds shouldn't be sparse. To accomplish this, we leverage Wikipedia semantic knowledge to measure the semantic relatedness and ambiguity. And we use the seed frequency in the web corpus to measure population. Lastly, we use a linear model to combine these factors to determine the final selection. In detail, we adopt a two-phase strategy: First, we propose a disambiguation algorithm to identify the articles in Wikipedia which describe the original seeds. Second, by using the semantic knowledge contained in these articles, we measure the quality of the seeds, and choose the better ones.

Specifically, our contributions are:

- We believe the quality of the input seeds has huge influence on the performance of entity expansion. And we consider three factors to measure seed quality. We also present three algorithms to measure these factors.
- We propose a measuring method to choose better seeds which considers three factors mentioned above at the same time. Experimental results on data from different domains show that our method can effectively figure out seeds and improve the entity set expansion quality.

The remainder of the paper is organized as follows. Section 2 states the impact of seed set and reviews related work. In Section 3, we introduce Wikipedia as a semantic knowledge base. Section 4 introduces the three factors in measuring seed quality and describes our proposed method in detail. Experimental results are discussed in Section 5. We conclude with some discussion and future work in Section 6.

2 Problem Statement and Related Work

As mentioned in [1], the problem of choosing better seeds for entity set expansion can be defined as follows:

For a semantic concept C, given M entities belong to C, the seed-choosing system should choose the K-best entities from the M given ones which can get best expansion performance. In this paper we make K=3, which is also used in [9].

For example, suppose we want to find out all countries. And we already know some of them such as “China”, “USA”, “Russia”, “Germany”, “Canada” and “Japan”. The seed-choosing method should be able to find out which 3 of these 6 entities should be used as seeds for entity set expansion.

A prominent work about better seeds selection is proposed by Vyas et al [1]. They measure every seed according to the following three factors: 1) Prototypicality, which weighs the degree of a seed’s representation of the concept; 2) Ambiguity, which measures the polysemy of a seed; 3) Coverage, which measures the degree of the amount of semantic space which the seeds share in common with the concept. Then, they design three methods, each of which deals with one factor.

However, this method has some limitations: First, they only use seed contexts as features. However, because the seeds may be ambiguous, the context of an ambiguous seed is ambiguous too. So the use of context can’t avoid bringing in mistakes. Second, they don’t consider the seeds’ population. So they may choose seeds which appear so few times in the corpus that can’t show any statistic typicality.

To overcome these deficiencies, we propose to measure seed quality from the following three factors: Semantic Relatedness, Population and Ambiguity. In detail, we first find out the Wikipedia articles which the seeds related to. Then, we use Wikipedia as the background knowledge to measure the three factors. In the following sections, we will show our method in detail.

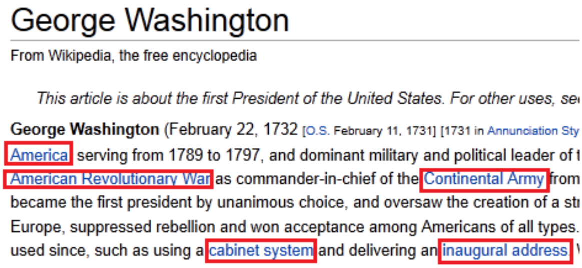
3 Wikipedia as a Semantic Knowledge Base

Wikipedia is the largest encyclopedia in the world and surpasses other knowledge bases in its large amount of concepts, up-to-date content, and rich semantic information. The English version Wikipedia contains more than 3,900,000 articles and new articles are added very quickly.

Because of its large scale and abundant of semantic information, Wikipedia has been widely used in Information Retrieval and Nature Language Processing. In the following subsections, we will introduce the characters of Wikipedia.

3.1 Wikipedia Articles

Wikipedia uses an article to describe a single entity. Figure 1 is a snapshot of part of the article “George Washington”. The red boxes in the Figure markup links to other articles. In general, an article in Wikipedia has average 34 links out to other articles and receives average another 34 links from them [10].



George Washington

From Wikipedia, the free encyclopedia

This article is about the first President of the United States. For other uses, se

George Washington (February 22, 1732 [O.S. February 11, 1731] [1731 in Annunciation Sty
America serving from 1789 to 1797, and dominant military and political leader of t
American Revolutionary War as commander-in-chief of the Continental Army from
became the first president by unanimous choice, and oversaw the creation of a st
Europe, suppressed rebellion and won acceptance among Americans of all types.
used since, such as using a cabinet system and delivering an inaugural address.

Fig. 1. A Snapshot of A Typical Wikipedia Article

These links can also be used to measure the semantic relatedness between Wikipedia entities. In this paper, we adopt the method described in [10]. Based on the idea that the higher semantic relatedness two entities share, the more common links they have, this method measures semantic relatedness as follows:

$$sr(a,b) = 1 - \frac{\log(\max(|A||B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \tag{1}$$

Where *a* and *b* are the two entities of interest, *A* and *B* are sets of all entities that link to a and b respectively, and *W* is the entire Wikipedia.

3.2 Wikipedia Anchors

In Wikipedia, anchors refer to the terms or phrases in articles texts to which links are attached. Texts in red boxes in Figure 1 are examples of anchors.

Anchors have a tendency to link to multiple articles in Wikipedia. For example, anchor “Apple” might refer to the IT Company Apple Inc, a kind of fruit, a film and so on. Suppose article set D is consisted of the articles that anchor *a* links to, we can calculate the probability that *a* links to article *d* which belong to *D* as follows:

$$\Pr ob(<a,d>) = \frac{count(<a,d>)}{\sum_{d' \in D} count(a \rightarrow d')} \tag{2}$$

In section 4, we will discuss how to use this property to calculate the ambiguity of a seed in detail.

3.3 Wikipedia Category Labels

Each Wikipedia article has several “Category Labels”, which means it belongs to the category. A Label usually indicates a semantic class. For example, “George

Washington" has the label "American planer" etc. We will use these labels when we link seeds to articles in section 4.1.

4 Choosing Better Seeds by Leveraging Wikipedia Semantic Knowledge

In this section, we introduce our method in detail and show how to leverage Wikipedia semantic knowledge for better seeds selection. Totally, there are two steps: 1) given an original seed set *S*, we link every seed *s* in *S* to the article *d* which describes it in Wikipedia. This can be seen as a process of disambiguation. 2) For every seed *s*, we measure its quality from the following three factors: semantic relatedness, population and ambiguity.

4.1 Linking Seeds to Wikipedia Articles

For every seed s_i in *S*, we use it as an anchor a_i, then we get an article set *Ai* including all articles that a_i links to. So for *M* input seeds we get *M* article sets $\{A1, A2, ..., AM\}$. Then we need to find out the exact articles which describe the seeds.

For every article group $G:\{A1_i, A2_j, ..., AM_k\}$, we use the following formula to compute its confidence:

$$\text{Conf}(G) = Relatedness(G) + Probability(G) + Category(G) \qquad (3)$$

Relatedness(G) is the average relatedness of each two articles in G, which is computed by using formula (1). *Probability(G)* is the conduct of the probabilities of the articles in G which are computed by formula (2). And for *Category(G)*, if there exists common category label for all the articles, it is set to be 1, if not it is 0. Finally we choose the group that has highest confidence and link seeds $\{s_1, s_2, ..., s_M\}$ to their related articles $\{A1_i, A2_j, ..., AM_k\}$.

4.2 Measuring Seed Quality

For every seed *s*, after linking it to article *a* in Wikipedia, we measure its quality from the following three factors: semantic relatedness, population and ambiguity.

4.2.1 Semantic Relatedness

The first factor which affects the quality of expansion is the semantic relatedness between a seed and the target concept. Because target concept is unknown, we approximate the semantic relatedness of a seed as the average semantic relatedness of this seed and all other original given seeds. We should choose seeds with high semantic relatedness.

$$Rel(a) = \frac{\sum_{b \in S, b \neq a} sr(a,b)}{(M-1)} \tag{4}$$

Where *S* is the given seed set, *a* is a seed and *M* is the size of *S*.

4.2.2 Population

The second factor which determines the quality of a seed is population. Some entities are sparser than other ones. If we use sparse entities as seeds, we may learn fewer templates which may lead to poor expansion performance. So we should choose seeds with high population.

In this paper, we use the following formula to calculate the population of a seed:

$$Pop(s) = \frac{count(s)}{MAX_{s' \in S}[count(s')]} \tag{5}$$

Where count(*s*) of each seed *s* is the number of web pages contains it in the web corpus. *S* refers to the given seeds set.

4.2.3 Ambiguity

The third factor which determines the quality of a seed is ambiguity. As the former example shows, the seed "Lincoln" may refers to the president "Abraham Lincoln", or a luxury brand of car "Lincoln", or the aircraft carrier of USA "Lincoln". So seed "Lincoln" can results in errors during expansion for the concept "US President". In this paper we define ambiguity as the probability that a seed link to the target article.

To calculate the ambiguity of a seed, we use the method described in formula (2):

$$Amb(s) = \Pr ob(< s, a >) \tag{6}$$

Where *a* is the article which describes *s* in Wikipedia.

At last, we combine the three factors and use the following formula to measure the quality of every seed. Then we choose the 3 seeds with highest Qua(*s*) as new seeds.

$$\text{Qua}(s) = \alpha * Rel(s) + \beta * Pop(s) + (1 - \alpha - \beta) * Amb(s) \tag{7}$$

In this paper, we make $\alpha = \beta = 1/3$, so the three factors weight equally.

5 Experiments

In this section, we analyze the experimental results of our methods. First, we explain our data set and evaluation criteria. Then we discuss the performance of our method.

5.1 Experimental Setup

For evaluating our algorithm, we use 6 lists of named entities chosen from Wikipedia "List of" pages as the gold standard which is the same as [1]. The lists are: *CA counties, Countries, F1 Drivers, Elements, US States and Roman Emperors*. Each list represents a single concept. We use English Wikipedia Ver.20110722. And we use Wikipedia miner toolkit downloaded from http://wikipedia-miner.cms.waikato.ac.nz/.

To expand the seeds, we employ the algorithm proposed in [9]. We use R-precision to evaluate the expansion performance, which is also used by [1].

5.2 Linking Method Evaluation

To evaluate the linking algorithm proposed in section4.1, we make 500 trials for every list. Table 2 shows the linking result. By using the combined disambiguation method, we get 94% linking precision which can meet the need for further processing.

Table 2. Linking precision analysis over six gold standard entity types

Concept	Relatedness	Probability	Category	Combine
California Counties	0.886	0.842	0.916	0.910
Countries	0.270	0.274	0.856	0.946
Elements	0.980	0.960	1.000	1.000
F1 Drivers	0.454	0.402	0.354	0.902
Roman Emperors	0.760	0.752	0.392	0.880
U.S. States	0.136	0.142	0.938	1.000
Average	0.581	0.563	0.726	**0.940**

5.3 Overall Performance

For evaluating the effectiveness of our method, we do 10 trials for each concept. In each trial, we randomly choose 6 entities as input seeds. Then we use the method described in section 4 to select 3 seeds as the new seeds set.

Table 3. Overall R-precision analysis over six gold standard entity types

Concept	Relatedness	Population	Ambiguity	Combined	Random
California Counties	1.000	1.000	1.000	1.000	0.753
Countries	0.795	0.738	0.807	0.881	0.676
Elements	0.974	0.350	0.966	0.974	0.714
F1 Drivers	0.108	0.721	0.108	0.133	0.478
Roman Emperors	0.246	0.652	0.587	0.680	0.529
U.S. States	1.000	1.000	1.000	1.000	0.967
Average	0.687	0.744	0.745	**0.778**	0.686

Table 3 shows the overall expansion performance. Column2~4 show results using seeds chosen by only one factor. Column 5 shows the result using seeds chosen by all three factors. Column 6 shows the result for random seeds from input ones. We see our method using all three factors get the best result and improve the performance by 13.4%. This proves the effectiveness of our method.

We also see factors have different impact on various concepts. This can be ascribed to the difference in the natures of concepts. So we should consider all the three factors.

6 Conclusions and Future Work

In this paper, we propose a novel method for selecting better seeds for entity set expansion. For every input seed, we measure its semantic relatedness, ambiguity and population. Then we choose the top three seeds. Experimental results show that our seeds can improve expansion performance by up to average 13.4% over random seeds.

For future work, we plan to use other semantic knowledge provided by Wikipedia like category hierarchy to help finding better seeds.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 61070106), the National Basic Research Program of China (No. 2012CB316300), the Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation.

References

1. Vishnu, V., Patrick, P., Eric, C.: Helping editors choose better seed sets for entity set. In: Proceedings of CIKM 2009, pp. 225–234. ACM, Hong Kong (2009)
2. Richard, W., Nico, S., William, C., Eric, N.: Automatic Set Expansion for List Question Answering. In: Proceedings of EMNLP 2008, pp. 947–954. ACL, USA (2008)
3. Marco, P., Patrick, P.: Entity Extraction via Ensemble Semantics. In: Proceedings of EMNLP 2009, pp. 238–247. ACL, Singapore (2009)
4. Richard, W., William, C.: Automatic Set Instance Extraction using the Web. In: Proceedings of ACL/AFNLP 2009, pp. 441–449. ACL, Singapore (2009)
5. Marius, P.: Weakly-supervised discovery of named entities using web search queries. In: Proceedings of CIKM 2007, pp. 683–690. ACM, Portugal (2007)
6. Richard, W., William, C.: Iterative set expansion of named entities using the web. In: Proceedings of ICDM 2008, pp. 1091–1096. IEEE Computer Society, Italy (2008)
7. Patrick, P., Eric, C., Arkady, B., Ana-Maria, P., Vishnu, V.: Web-Scale Distributional Similarity and Entity Set Expansion. In: Proceedings of EMNLP 2009, pp. 938–947 (2009)
8. Yeye, H., Dong, X.: C.: SEISA Set Expansion by Iterative Similarity Aggregation. In: Proceedings of WWW 2011, pp. 427–436. ACM, India (2011)
9. Richard, W., William, C.: Language-Independent Set Expansion of Named Entities using the Web. In: Proceedings of ICDM 2007, pp. 342–350. IEEE Computer Society, USA (2007)
10. David, M., Ian, H.W.: Learning to link with Wikipedia. In: Proceedings of CIKM 2008, pp. 509–518. ACM, USA (2008)