

Are Human-Input Seeds Good Enough for Entity Set Expansion? Seeds Rewriting 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 is an important task for open information extraction, which refers to expanding a given partial seed set to a more complete set that belongs to the same semantic class. Many previous researches have proved that the quality of seeds can influence expansion performance a lot since human-input seeds may be ambiguous, sparse etc. In this paper, we propose a novel method which can generate new, high-quality seeds and replace original, poor-quality ones. In our method, we leverage Wikipedia as a semantic knowledge to measure semantic relatedness and ambiguity of each seed. Moreover, to avoid the sparseness of the seed, we use web resources to measure its population. Then new seeds are generated to replace original, poor-quality seeds. Experimental results show that new seed sets generated by our method can improve entity expansion performance by up to average 9.1% over original seed sets.

Keywords: information extraction, seed rewrite, semantic knowledge.

1 Introduction

Entity Set expansion refers to the problem of expanding a given partial (3~5) set of seed entities to a more complete set which belongs to the same semantic category. For example, a person may give a few elements like “Gold”, “Mercury” and “Xenon” as seeds; the entity set expansion system should discover other elements such as “Silver”, “Oxygen” etc. based on the given seeds.

These collections of entities are used in many commercial and research applications. For instance, question answering systems can use the expansion tools to handle List questions [1]. And search engines collect large sets of entities to better interpret queries.[2]

Several researches have been proposed for solving this problem, like [3][4][5]. These methods generally include two components: 1) find candidates which may have the same semantics with the given seeds; 2) measure the similarity between each

* Corresponding author.

candidate and the given seeds. Candidates with higher similarity score will be extracted as results. A typical method starts from several seeds (usually 3-5), then it employ distributional features [6][7] or context patterns [8][9] to find entities of the same category in external data sources such as large corpora of text or query logs.

Table 1. Seeds greatly influences entity set expansion quality

Concept	MAX	MIN	AVG
California Counties	1.000	0.103	0.859
Countries	0.885	0.008	0.667
Elements	0.991	0.026	0.784
F1 Drivers	0.959	0.000	0.456
Roman Emperors	0.804	0.109	0.503
U.S. States	1.000	0.640	0.908

However, since the seeds are provided by human as will, they may be poor-quality and have problems such as being ambiguous or sparse etc.. Taking the three seeds we mentioned at the beginning of this section as instance, seed “Mercury” is ambiguous because it appears as instance of two completely unrelated semantic class *planets* and *elements* with almost equally probability. Seed “Xenon” is so sparse in common corpus that there is a high probability we find very few useful templates by using it.

To study the impaction of seeds, we employ a state-of-art set expansion system [10] 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 select 3 entities as seeds. Table 1 shows the maximum, minimum and average expansion performance of these seeds sets measured by R-precision. We see there is a variation as much as 35% between the max and average performance, confirming that the quality of seeds truly has great influence on the expansion performance. Other studies have come to similar conclusion [2]. Furthermore, previous studies have shown that human editors generally provide very bad seeds [2]. So generating high-quality seeds is very important for entity set expansion.

To avoid seed ambiguity and sparseness, we propose a novel method for generating better seeds in this paper. First, we link original seeds to Wikipedia articles. Second, we measure the quality of seeds by the following three factors and decide which seeds should be replaced: 1) **Semantic Relatedness**. High-quality seeds should have high semantic relatedness among each other; 2) **Ambiguity**. Good seeds should have less ambiguity; 3) **Population**. High-quality seeds shouldn’t be sparse. Lastly, we generate new seeds which have high quality by using Wikipedia. In detail, we adopt a three-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 original seeds and find out poor-quality seeds. Third, we generate new, high-quality seeds using the category structure and semantic knowledge of Wikipedia.

Specifically, our contributions are:

- We believe the quality of seeds has great influence on entity expansion performance. We identify three factors to measure seed quality. And we present three algorithms to measure these factors respectively.

- We propose a novel method to find out poor-quality seeds and generate high-quality seeds to replace them. The new generated seeds will be used for expanding entities in Web data. Experimental results on data from different domains show that our method can effectively generate high-quality seeds and improve the entity set expansion performance.

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. Section 6 concludes this paper and discusses the future work.

2 Problem Statement and Related Work

As mentioned in last section, the problem of seeds rewriting for entity set expansion can be defined as follows:

For a semantic category C , given M entities belong to C , the seed rewriting system should find out which K entities from M given ones have poor quality and generate K new, high-quality seeds to replace them. In this paper we make $M = 3$, which is also used in [10]. Since M is small and there may be high-quality seeds in the original ones, we just replace the most poor-quality seed which means $K = 1$.

For example, suppose we want to find out all elements. And we already know some of them such as “Gold”, “Mercury”, “Xenon”. The seed generation method should be able to find out the one that should be replaced (suppose it is “Mercury”) and generate new high-quality seed to replace it (suppose it is “Oxygen”).

A similar problem is “better seeds selection”. The key point of that problem is to choose K -best seeds from given M ones. Previous studies have proved that methods which solve that problem can also improve the expansion performance. A prominent work about better seeds selection is proposed by Vyas et al [2]. 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 remove the error-prone seeds and return the remaining seeds as results.

Those methods have some limitations: 1) They can only choose relatively better seeds from original ones but cannot generate new, high-quality seeds. If unfortunately original seeds are not high-quality, they can only get a poor performance. 2) In many situations, it is hard to get enough original seeds for selection since seeds are provided by human as will.

To overcome these deficiencies, we propose a novel method to resolve the problem of seeds rewriting. Generally a three-stage strategy is designed to generate new, high-quality seeds: First, we link original seeds to Wikipedia articles which describe them. Second, we present three factors to measure seed quality and propose three algorithms to measure these factors respectively. Then we attempt three ways to decide which

seed should be replaced. Lastly, we present a method to generate new high-quality seed and replace the old one. To accomplish this, we use Wikipedia semantic knowledge and web corpus frequency. 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 because of its large amount of concepts, up-to-date information, and rich semantic knowledge. The English version Wikipedia contains more than 4 million 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 some characters of Wikipedia which will be used for our task.

3.1 Wikipedia Articles

In Wikipedia, an article is usually used to describe a single entity. Figure 1 is a snapshot of part of the article “Mercury (element)”. The red boxes in the Figure markup links to other articles. Previous study shows that an article in Wikipedia has 34 links out to other articles and receives another 34 links from them on average [11].



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 [11]. 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:

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

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 “Mercury” might refer to a kind of element, a planet, a roman god 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:

$$\text{Prob}(\langle a, d \rangle) = \frac{\text{count}(\langle a, d \rangle)}{\sum_{d \in D} \text{count}(a \rightarrow d')} \quad (2)$$

where $\text{count}(\langle a, d \rangle)$ is the number of times that anchor a links to article d . 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

In Wikipedia, each article has several “Category Labels”, which means it belongs to the category. Figure 2 shows the category labels of the article “Mercury (element)”.

Categories: [Chemical elements](#) | [Mercury \(element\)](#) | [Occupational safety and health](#) | [Endocrine disruptors](#) | [Transition metals](#) | [Post-transition metals](#) | [Coolants](#)
[Nuclear reactor coolants](#) | [Neurotoxins](#) | [Native element minerals](#)

Fig. 2. A Snapshot of Category Labels of A Wikipedia Article

A category label usually indicates a semantic class. So articles that belong to the same category label may belong to a same semantic class with high probability. For instance, “Mercury (element)” has the label “Chemical elements” etc. Oxygen, Gold and many other elements all have category label “Chemical elements”. So they have a high probability to belong to the same semantic category. We will show how to use these labels to generate new seeds in section 4.

4 Seeds Rewriting by Leveraging Wikipedia Semantic Knowledge

In this section, we introduce our method in detail and show how to generate new, high-quality seeds by leveraging Wikipedia semantic knowledge. Totally, our system is comprised of three major components: the Linker, the Measure and the Generator. Figure 3 is a schematic diagram of our seeds rewriting system.

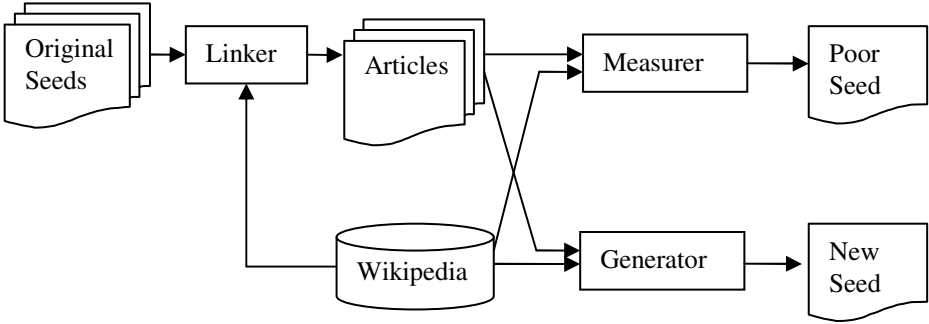


Fig. 3. Flow Chart of the Seeds Rewriting System

The Linker links every seed s in original seed set S to the article d which describes it in Wikipedia. This can be seen as a procedure of disambiguation. The Measurer measures the quality of every seed s from the following three factors: semantic relatedness, population and ambiguity and decide which seed should be replaced based on those three factors. The Generator generates new seeds using Wikipedia category structure and semantic knowledge and returns the most high-quality one as result.

4.1 Linking Seeds to Wikipedia Articles

In order to use the semantic knowledge of Wikipedia articles, the Linker needs to find out the exact articles which describe the seeds. These articles should have high semantic relatedness among themselves because they all describe instances of the same concept. Moreover, the probability of the articles being linked to should also be considered. So we design the following method to solve this problem:

For every seed s_i in S , we use it as an anchor a_i , then we can get an article set A_i includes all articles that a_i links to. So for three input seeds we get three article sets $\{A_1, A_2, A_3\}$. For every possible article group $G : \{A_{1_i}, A_{2_j}, A_{3_k}\}$, we use formula 3 to compute its confidence and choose the group which gets the highest score as result.

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

Here $\text{Relatedness}(G)$ is the average relatedness of each two articles in G , which is computed by using formula (1). $\text{Probability}(G)$ is the conduct of the probabilities of the articles in G which are computed by formula (2). And for $\text{Category}(G)$, 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 $\{s_1, s_2, s_3\}$ to their related articles $\{A_{1_i}, A_{2_j}, A_{3_k}\}$.

When used as an anchor a_i , every seed s_i in S links to about 10 articles in Wikipedia so we can see that the computational complexity is acceptable.

4.2 Seeds Quality Measuring

For every seed s , after linking it to article a in Wikipedia, the Measurer measures its quality from the following three factors: semantic relatedness, population and ambiguity. Then the seed with worst quality is found out and replaced.

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. The higher semantic relatedness a seed has with a concept, the better it can represent the concept. So we should replace seeds with low semantic relatedness. Since 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:

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

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 replace seeds with low population.

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

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

where $count(s)$ of each seed s is the number of web pages returned when we use s as query searching by the Bing API. 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 “Mercury” may refers to the element “Mercury (element)”, or a planet “Mercury (planet)”, or the roman god “Mercury (roman god)”. So seed “Mercury” can results in errors during expansion for the concept “Element”. 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):

$$Amb(s) = Pr ob(< s, a >) \quad (6)$$

where a is the article which describes s in Wikipedia.

For the three original seeds, we measure their quality from the above three factors. In order to find out influence of different factors, we attempt three ways to decide which seed to be replaced. In each way, we measure one factor. A detailed analysis is shown in Section 5.

4.3 New Seed Generation

The Generator generates new, high-quality seeds. It extracts candidate new seeds using the category structure of Wikipedia and then measures their quality and returns the one with highest quality. A schematic diagram is shown in Figure 4.

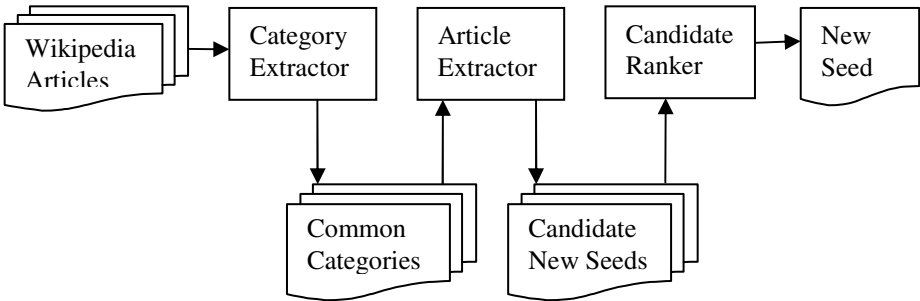


Fig. 4. Flow chart of the new seed generation procedure

Given the three original articles, the Category Extractor gathers their common category labels for further processing. If there is no common category for all three articles, it gathers common category labels for every two articles. The Article Extractor collects all articles belong to the common category labels and using their title as candidate new seeds. The Candidate Ranker uses the following combined formula to measure the quality of every candidate new seed from the three factors discussed in section 4.2. Lastly the Generator returns the one with highest score as the new seed.

$$\text{Qua}(s) = \text{Rel}(s) + \text{Pop}(s) + \text{Amb}(s) \quad (7)$$

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.20110722¹. To deal with Wikipedia data, we use the Wikipedia miner² toolkit.

To expand the seeds, we employ the expansion algorithm described in [10]. We use R-precision to evaluate the expansion performance. It is also used by [1].

$$R - precision(L) = \frac{RM}{N} \quad (8)$$

Where L is a ranked list of extracted mentions; N is the size of the gold standard set; RM is the num of right mentions in the list.

5.2 Linking Method Evaluation

To evaluate our linking algorithm proposed in section4.1, we make 500 trials for every list. Table 2 shows the linking result. By using our combined disambiguation algorithm, 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 seed generation method, we do 10 trials for each concept. In each trial, we randomly choose 3 entities as input seeds. Then we use the method described in section 4 to generate a new, high-quality seed to replace the most poor-quality original seed.

Table 3 shows the overall performance. Column3~5 show the expansion performance after using the new seed to replace the most poor-quality seed measured by one single factor. As comparison, Column 2 shows the performance for the original input seeds. We can see that by replacing the original poor-quality seed with the new high-quality seed generated by our method, the average expansion performance can get obviously improved by up to 9.1% in R-precision.

¹ <http://download.wikipedia.com/enwiki/>

² <http://wikipedia-miner.cms.waikato.ac.nz/>

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

Concept	Original	Ambiguity	Population	Relatedness
California Counties	0.859	0.976	0.926	0.964
Countries	0.667	0.701	0.652	0.670
Elements	0.784	0.894	0.981	0.889
F1 Drivers	0.456	0.569	0.628	0.549
Roman Emperors	0.503	0.509	0.486	0.554
U.S. States	0.908	0.902	0.858	0.824
Average	0.696	0.759	0.755	0.742

5.4 Detailed Analysis

Experimental results in Table 3 show that the expansion performance differs a lot among various concepts. Experimental results published by former study also show this phenomenon [1]. This can be ascribed to the difference in the natures of concepts. Some concepts are more common (such as “Countries” or “U.S. States”) or have less ambiguity (such as “Elements”). Entities belonging to them are easier to be found so we get better expansion performance.

We also conclude that the three factors have different impact on the performance of entity set expansion. We get the best performance when replacing the original seed which performs worst measured by ambiguity. This suggests that the ambiguity is the most important factor in measuring the quality of seeds. We see population has almost the same influence. This phenomenon supports our hypothesis that the ambiguity and population are both very important for seeds. As comparison, Semantic Relatedness has less influence.

6 Conclusions and Future Work

In this paper, we propose a novel method for seeds rewriting for entity set expansion. For every input seed, we measure its semantic relatedness, ambiguity and population and decide which one to be replaced. Then we generate new high-quality seed by leveraging Wikipedia semantic knowledge and replace the old one. Experimental results show that our method can improve expansion performance by up to average 9.1% over original input seed sets.

For future work, we plan to use other semantic knowledge provided by Wikipedia like category hierarchy and structural description of entities to help generating 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 Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA06030300) and the Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation. We thank the anonymous reviewers for their insightful comments.

References

1. 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)
2. 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)
3. Marco, P., Patrick, P.: Entity Extraction via Ensemble Semantics. In: Proceedings of EMNLP 2009, pp. 238–247. ACL, Singapore (2009)
4. Luis, S., Valentiin, J.: More Like These: Growing Entity Classes from Seeds. In: Proceedings of CIKM 2007, pp. 959–962. ACM, Portugal (2007)
5. Richard, W., William, C.: Automatic Set Instance Extraction using the Web. In: Proceedings of ACL/AFNLP 2009, pp. 441–449. ACL, Singapore (2009)
6. Patrick, P., Eric, C., Arkady, B., Ana-Maria, P., Vishnu, V.: Web-Scale Distributional Similarity and Entity Set Expansion. In: Proceedings of EMNLP 2009, Singapore, pp. 938–947 (2009)
7. Yeye, H., Dong, X.: SEISA Set Expansion by Iterative Similarity Aggregation. In: Proceedings of WWW 2011, pp. 427–436. ACM, India (2011)
8. Marius, P.: Weakly-supervised discovery of named entities using web search queries. In: Proceedings of CIKM 2007, pp. 683–690. ACM, Portugal (2007)
9. 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)
10. Richard, W., William, C.: Language-Independent Set Expansion of Named Entities using the Web. In: Proceedings of ICDM 2007, USA, pp. 342–350. IEEE Computer Society (2007)
11. David, M., Ian, H.W.: Learning to link with Wikipedia. In: Proceedings of CIKM 2008, pp. 509–518. ACM, USA (2008)