Extracting Part-Whole Relations from Online Encyclopedia

Fei Xia, Cungen Cao

Extracting Part-Whole Relations from Online Encyclopedia

Fei Xia [1,2], Cungen Cao [1]

[1]Key Laboratory of Intelligent Information Processing, Institute of Computer Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China

xiafei.1986@163.com,cgcao@ict.ac.cn

Abstract. Automatic discovery of part-whole relations is a fundamental problem in the area of information extraction. In this paper, we present an unsupervised approach to learning lexical patterns from online encyclopedia for extracting part-whole relations. The only input is a few part-whole instances. To tackle the term recognition problem, terms from the domain of the seeds are extracted taking use of the semantic information contained in the online encyclopedia. Instead of collecting sentences that contain relation instances from the seeds, we introduce a novel process to select sentences that may indicate part-whole relations. Patterns are produced from these sentences with terms replaced by *Part* and *Whole* tags. A similarity measurement based on a new edit distance is used and an algorithm is described to cluster similar patterns. We rank the pattern clusters according to their frequencies, and patterns from the top-k clusters are chosen to be applied to identify the new part-whole relations. Experimental results show that our method can extract abundant part-whole relations and achieve a preferable precision compared to the other state-of-the-art approaches.

Keywords: part-whole relations; lexico-syntactical patterns; online encyclopedia; edit distance; clustering

1 Introduction

Part-whole relations, also known as meronymy, are fundamental semantic relations that exist in many semantic networks, such as WordNet and HowNet. Those semantic networks play a key role in many natural language processing (NLP) systems like information retrieval, and automatic question answering. One of the traditional approaches to automatically extracting part-whole relations is pattern-based, which identifies the relation from patterns like "Y consists of X" that indicates Part-Whole(X, Y) (means X is a part of Y). Those patterns are either designed manually by experts [1], which is very low efficient, or learned from sentences that contain the

given part-whole relation instances [2], which needs massive corpora such as the Web.

In recent years, online encyclopedias have drawn attention to many researchers and become an ideal source for semantic information extraction, since they have both broad coverage and high accuracy. BaiduBaike is one of the largest Chinese Wikipedia-like online encyclopedias, and it contains more than 7 million entries by March 2014. Each entry has a corresponding web page which describes that entry in detail (Fig 1a). More importantly, tags (also called folksonomy) labeled by editors and the related entries are listed in the bottom of that entry's page (Fig 1b), which are of great help for semantic relation extraction since they contain semantic information of that entry.

(a) *The top of the page of "banana"* (b) *Tags and related entries of "banana"*

Fig. 1. *The corresponding web page of the entry "banana"*

This paper presents an approach to learning lexical patterns from BaiduBaike to extract part-whole relations. The only input are some part-whole instance pairs, and we firstly extract domain terms of both parts and wholes using their related entries and tags. Sentences from pages of domain terms are labeled and those contain both part-terms and whole-terms are selected and transferred into patterns. A similarity formula based on edit distance is used to cluster patterns. Pattern clusters are ranked according to their frequencies, and patterns from the top-k clusters are chosen to be applied to identify the new part-whole relations.

The paper is organized as follows. Section 2 presents an overview of some related work. The details of our procedure are described in section 3. In section 4, the experimental design and results are presented. Finally, section 5 summarizes the work, and proposes future work.

2 Related Work

Several taxonomies of part-whole relations exist since researchers believe part-whole relations "should be treated as a collection of relations, not as a single relation" [3]. One of the most widely accepted taxonomies is developed by Winston et al. [4], and they mentioned six types of part-whole relations based on the way that the parts contribute to the structure of the wholes: component-integral, member-collection, stuff-object, place-area, portion-mass, and feature-activity.

Automatic acquisition of part-whole relations from unlabeled corpora was first presented in [5]. A few part-whole instance pairs as initial seeds were used to learn

lexical patterns, and then they extracted the corresponding "parts" of six "whole" instances by these patterns from the North American News Corpus (NANC), which achieves an accuracy of 55%. The low accuracy is due to the noisy patterns they used that tend to indicate both part-whole and non-part-whole relations. Patterns extracted in [2] are more reliable since they were learned from the web by using 503 part-whole seeds derived from a special thesaurus. The Espresso algorithm in [6] uses a novel measurement of pattern reliability based on point-wise mutual information to rank all patterns and keep the top-k ones. Both of the latter two approaches achieve higher precision.

Instead of generating more reliable patterns, some researchers spent more effort on verifying the new extracted part-whole instances. Algorithms developed in [7] improved the performance by using the part-whole relations from WordNet to train a decision-tree classifier which was used to predict the previously unseen instance pairs, and they provided a superior precision rate (83%) and recall rate (98%). However, the supervised approach they used requires extensive manual work and relies heavily on external tools like word-sense disambiguation. Other verification methods include heuristic rules [8] and graph models [9].

With the rapid growth of online encyclopedias, more and more work emerged to extract semantic relations from them. [10] first introduced an algorithm to extract part-whole relationships from the Wikipedia. Sentences containing two terms that share part-whole relation in WordNet were collected and transferred to lexical patterns. Similar patterns were generalized using the edit-distance algorithm and were used to extract new part-whole relations from the Wikipedia. The precision they reported is around 60% with different thresholds in the generalization step.

3 Our Method

Our method to learn part-whole relations from online encyclopedia consists of three steps:

(1) Domain Terms Extraction: terms from domains of part and whole instances are extracted using the semantic information in online encyclopedia.
(2) Pattern Learning: sentences from web pages in the online encyclopedia are analyzed and transferred to patterns if both part-terms and whole-terms are found. Patterns are clustered if they are similar.
(3) Identification of New Part-Whole Relations: those patterns with high frequency learned in step 2 are used to discover new part-whole relation instances.

The details of all the steps are described in the following sections.

3.1 *Domain Terms Extraction*

One of the critical problems in relation extraction is term recognition, which lies in both pattern learning and new relation identification phrases. Some approaches recognize terms with the help of specialized thesauri [2] or online dictionaries [7, 10, 11].

An online encyclopedia can be treated as a dictionary since it consists of tremendous entries and most of them can be seen as terms. The aim of this step is to extract as many domain terms as possible by giving some entries as seeds. We'll go through the related entries of seeds and pick all the relevant ones. To do this, several relevance measurements are defined as follows.

Tags-based relevance

The tags are a collection of keywords that refer to the category of that entry or other useful information. For example, the tags of the entry 香蕉 (banana) consists of 农产品 (agricultural product), 水果 (fruit), 食品 (food), 绿色植物 (greenery) and so on. The more tags that two entries have in common, the more relevant they are considered to be. Based on this hypothesis, we use a tags-based relevance measurement as follows.

$$\text{relevance}_1(\text{e}_i, \text{e}_j) = \frac{|\text{Tags}(\text{e}_i) \cap \text{Tags}(\text{e}_j)|}{|\text{Tags}(\text{e}_i) \cup \text{Tags}(\text{e}_j)|} \tag{1}$$

Where Tags(e_i) means all the tags of the entry e_i.

Related-entries-based relevance

As depicted in Fig 1b, many related entries are listed in the bottom of an entry's page. We use *related*($e_i \rightarrow e_j$) to denote that e_j is one of the related entries of e_i. Clearly, two entries e_i and e_j are relevant if *related*($e_i \rightarrow e_j$) or *related*($e_j \rightarrow e_i$).

$$\text{relevance}_2(\text{e}_i, \text{e}_j) = \begin{cases} 1, if\ related(e_i \rightarrow e_j) \text{ and } related(e_j \rightarrow e_i) \\ 0.5, if\ related(e_i \rightarrow e_j) \text{ or } related(e_j \rightarrow e_i) \\ 0, otherwise \end{cases} \tag{2}$$

On the other side, the more related entries that two entries have in common, the more relevant they are considered to be. So, another relevance measurement based on related entries is defined.

$$\text{relevance}_3(\text{e}_\text{i}, \text{e}_\text{j}) = \frac{|RelatedEntries(e_i) \cap RelatedEntries(e_j)|}{|RelatedEntries(e_i) \cup RelatedEntries(e_j)|} \tag{3}$$

Where RelatedEntries(e_i) means all the related entries of the entry e_i.

Finally, we developed a relevance measurement using the above three measurements.

$$\text{relevance}(\text{e}_i, \text{e}_j) = \sum_{k=1}^{3} \theta_k \times \text{relevance}_k(\text{e}_i, \text{e}_j) \tag{4}$$

Where θ_k are tunable parameters, and $\theta_1 + \theta_2 + \theta_3 = 1$.

Giving a set of entries E from a certain domain Domain(E), the probability of a new entry e belongs to Domain(E) is estimated by the following formula:

$$\text{prob}\big(\text{e} \in \text{Domain(E)}\big) = \frac{\sum_{\text{e}_i \in E} \text{relevance(e,e}_i)}{|E|} \quad (5)$$

Therefore, the iterative algorithm to extract domain terms with a set of seeds *E* is described in Algorithm 1:

Algorithm 1. Extracting domain terms

Input: domain entries seeds *E*, threshold λ_1

Output: domain terms *V*

```
1.  add all entries in E to V
2.  S=E
3.  S'=NULL
4.  Do Begin
5.    foreach entry e ∈ RelatedEntries(S)
6.      if prob(e ∈ Domain(S)) ≥ λ1
7.        add e to S'
8.        add e to V
9.    S=S'
10.   S'=NULL
11. Repeat (4)~(10) Until S is empty
12. End
```

We use this algorithm to extract two set of terms, one of them from the domain of part instances in the seeds, and the other from the domain of wholes. We name them *DV(p)* and *DV(w)*, respectively. So the output of this step is *DV(p)* and *DV(w)*.

3.2 *Pattern Learning*

The goal of this step is to extract patterns that indicate a part-whole relation. Traditionally, patterns are learned from those sentences that contain both part and whole instances from the seeds. However, this needs a massive corpus that contains lots of duplicated pieces of text that provides cues for relations between two terms. Compared to the whole Web, an online encyclopedia is a small corpus and the content is not so abundant. So with a small set of seeds, only a few sentences can be extracted from the online encyclopedia.

We collect the sentences that imply part-whole patterns in a different way. After the previous step, two set of terms are build. Sentences that contain terms from *DV(p)* and terms from *DV(w)* are selected and patterns are produced from them. The sentence selection process works in the following way:

(1) Go through all sentences from the web pages of entries from *DV(p)* and *DV(w)*.
(2) Sentences are segmented and part-of-speech tagged using ictclas [12].
(3) If an entry from *DV(p)* appears in the sentence, replace it by the tag *Part*. A list of *Part* tags appearing in a coordinate structure will be replaced with only one *Part*.
(4) If an entry from *DV(w)* appears in the sentence, replace it by the tag *Whole*. A list of *Whole* tags appearing in a coordinate structure will be replaced with only one *Whole*.

(5) Select all sentences that contain both *Part* and *Whole* tags.

After this process, the selected sentences are transferred into patterns; for example, sentence (s1c) contains an entry 苹果 (apple) from *DV(w)*, and two entries 维生素C (vitamin C) and 果胶 (pectin) in a coordinate structure from *DV(p)*:

(s1c) 苹果/n 富含/v 维生素/n C/x 和/c 果胶

(s1e) Apples are high in vitamin C and pectin

Therefor pattern (p1c) produced from (s1c) is:

(p1c) ***Whole*** 富含/v ***Part***

(p1e) ***Whole*** are high in ***Part***.

Usually, a reliability measurement will be chosen to rank those produced patterns, and the top-k most reliable patterns are selected. Instead of measuring a single pattern, we firstly cluster similar patterns and then pattern-clusters are ranked according to their frequency, and then the patterns in the top-k most frequent clusters are selected.

We use a similarity measurement based on edit distance to judge whether two patterns are similar:

$$\mathrm{Sim}(\mathrm{p_i}, \mathrm{p_j}) = 1 - \frac{\mathrm{EditDist}(\mathrm{p_i}, \mathrm{p_j})}{\max(|\mathrm{p_i}|, \mathrm{p_j}|)} \tag{6}$$

Where EditDist(p_i, p_j) is the edit distance between two patterns p_i and p_j, $|p_i|$ is the length of p_i which is defined as the number of the words it contains. A pattern p with length m can be expressed as $p(1 \ldots m)$, and the k-th word in p is expressed as $p[k]$. The edit distance between two patterns p_i and p_j is defined as the minimum number of changes (word insertion, deletion and replacement) to transfer p_i to p_j:

$$\begin{aligned} &\mathrm{EditDist}(p_i, p_j) = \mathrm{EditDist}\left(p_i(1 \ldots m), p_j(1 \ldots n)\right) = \\ &min\begin{pmatrix} \mathrm{EditDist}\left(p_i(1 \ldots m-1), p_j(1 \ldots n)\right) + 1, \\ \mathrm{EditDist}\left(p_i(1 \ldots m), p_j(1 \ldots n-1)\right) + 1, \\ \mathrm{EditDist}(p_i(1 \ldots m-1), p_j(1 \ldots n-1)) + diff(p_i[m], p_j[n]), \end{pmatrix} \end{aligned} \tag{7}$$

Where m is the length of p_i and n is the length of p_j, and

$$\mathrm{diff}(\mathrm{p_i}[\mathrm{m}], \mathrm{p_j}[\mathrm{n}]) = \begin{cases} 0, \text{if } \mathrm{p_i}[\mathrm{m}] = \mathrm{p_j}[\mathrm{n}] \text{ or they are synonymous} \\ 0.5, \text{if } \mathrm{POS}(\mathrm{p_i}[\mathrm{m}]) = \mathrm{POS}(\mathrm{p_j}[\mathrm{n}]) \\ 1, \text{otherwise} \end{cases} \tag{8}$$

Where POS($p_i[m]$) means the part-of-speech tag of p_i, and we use a Chinese synonymy dictionary "TongYiCiCiLin" (Extension Version) [13] to judge whether two words are synonymous or not.

The algorithm to cluster similar patterns is described in Algorithm 2.

Algorithm 2. Clustering patterns

Input: set of patterns P={p_1, p_2, …, p_n}, threshold λ_2

Output: clusters of patterns C={c_1, c_2, …}

1. C is initialized as {c_1, c_2, …, c_n} where c_i={p_i}
2. Do Begin
3. calculate the similarities of each pair of clusters
4. take the two clusters with the biggest similarity, c_i and c_j, the max similarity is *max*
5. if max < λ_2
6. return
7. else
8. merge c_i and c_j to a new cluster c'
9. recalculate the similarities between c' and all the other clusters
10. Repeat (2)~(9)
11. End

The similarity between two clusters is calculated using the average-linkage method; that is, their similarity is computed as the average similarity between every pair of patterns from the two clusters:

$$\mathrm{Sim}(c_1, c_2) = \frac{\sum_{p_i \in c_1} \mathrm{Sim}(p_i, p_j)}{|c_1| \times |c_2|} \quad (9)$$

We rank those generalized patterns according to their frequencies which are computed as the sum of the frequencies of the patterns before generalization. Finally, patterns in the top-k clusters are chosen as the output of this step.

3.3 *Identification of New Part-Whole Relations*

We apply the patterns learned in the previous step to the online encyclopedia corpus to identify new part-whole relations. The corpus is split into individual sentences and preprocessed with word segmentation and part-of-speech tagging. Once a sentence matches a pattern, substrings corresponding to the *Part* and *Whole* tags are extracted, and then terms are recognized using *DV(p)* and *DV(w)*.

Since part instances are always showed in coordinate structures, one sentence may contain several pairs of part-whole relations. The output of this step is a list of pairs of candidate part-whole relations.

4 Experimental Results

We evaluate the performance of our method in extracting stuff-object relation from the BaiduBaike corpus. Our method also can be applied to other types of part-whole relations and we plan to do it in the future work. The five pairs of part-whole relation seeds used in the experiment are listed in Table 1. They are also entries in the online encyclopedia.

Table 1. *Part-whole relation seeds*

Whole	*Part*
苹果 (apple)	维生素C (vitamin c)
香蕉 (banana)	钾 (potassium)
橘子 (orange)	柠檬酸 (citric acid)
栗子 (chestnut)	蛋白质 (protein)
龙眼 (longan)	葡萄糖 (glucose)

Two set of domain terms are extracted using Algorithm 1 with $\theta_1 = 0.5, \theta_2 = 0.3, \theta_3 = 0.2, \lambda_1 = 0.25$. The size of *DV(p)* is 2070, and the size of *DV(w)* is 2556. We randomly chose 200 terms from each set, and the precisions all exceed 95%, which shows the efficiency of our domain terms extraction algorithm.

After the sentence selection process, 11003 sentences that contain both terms from *DV(p)* and *DV(w)* are extracted and 9235 distinct patterns are produced, which indicates that online encyclopedia really contain few duplicated sentences. Similar patterns are clustered by Algorithm 2 with $\lambda_2 = 0.5$, and 330 patterns from top-15 clusters are chosen to be applied to extract new part-whole relations. Table 2 shows some of the patterns learned by our approach.

Table 2. *Samples of part-whole patterns*

Part-whole pattern	Example
Whole 含有 ***Part*** ***Whole*** contains ***Part***	桑枝 含有 鞣质 ramulus mori contains tannin
Whole 所含的 ***Part*** ***Part*** contained in ***Whole***	葱 所含的 大蒜素 allicin contained in oniions
Whole 中的 Part 含量 ***Part*** content of ***Whole***	玉米 中的 蛋白质 含量 protein content of corns
Whole 含 Part 量 ***Part*** in ***Whole***	樱桃 含 铁 量 iron in cherrys
Part 的 主要 来源 是 ***Whole*** the main source of ***Part*** is ***Whole***	维生素B6 的 主要 来源 是 瘦肉 the main source of vitamin B6 is lean
食用 ***Whole*** 可以 补充 ***Part*** eating ***Whole*** can supplement ***Part***	食用 芒果 可以 补充 维生素C eating mangoes can supplement vitamin C

In the new-relation identification step, 5184 sentences match at least one of the patterns, 9868 pairs of candidate part-whole relations are extracted. We calculate the precision of each pattern manually according to the following equation:

$$\mathrm{P} = \frac{\mathrm{Cnt(correct-extracted)}}{\mathrm{Cnt(all-extracted)}} \times 100\% \qquad (10)$$

Where Cnt(all-extracted) means the number of all the part-whole relations extracted by that pattern and Cnt(correct-extracted) means the number of the correct ones among them. Table 3 shows the results of those patterns listed in the above table:

Table 3. *Relation extraction results*

Part-whole pattern	Cnt(all-extracted)	precision
Whole 含有 ***Part***	6776	82.1%
Whole 所含的 ***Part***	788	84.6%
Whole 中的 Part 含量	698	87.7%
Whole 含 Part 量	169	95.9%
Part 的 主要 来源 是 ***Whole***	35	91.4%
食用 ***Whole*** 可以 补充 ***Part***	13	100%

Over 85.8% relations are extracted by the above patterns and 80% of them are identified by the first one. We have got a preferable precision compared to other reported algorithms, such as Girju's [7] 83%.

Two types of common errors are identified in our experimental results:

1) Unable to recognize some of the multiword terms. For example, the part-whole pair extracted from the following sentence (s2c) is Part-Whole(酸, 黄瓜) (Part-Whole(acid, cucumber)), while the correct part instance should be 丙醇二酸 (tartronic acid). This accounts for over 50% of the errors.
   (s2c) 黄瓜/n 中/f 含有/v 丙醇/n 二/m 酸/a
   (s2e) Cucumbers contain tartronic acid
2) Some errors are due to the lack of anaphora resolution algorithm. For example, sentence (s3c) returns the relation Part-Whole(胡萝卜素, 蔬菜) (Part-Whole(carotene, vegetable)), but the whole instance has to be the specific vegetable it referred to, which may be contained in the previous sentence.
   (s3c) 这些/rz 蔬菜/n 富含/v 胡萝卜素/n
   (s3e) These vegetables are high in carotene

To improve the performance of our approach, external tools like term recognition and anaphora resolution will be used in the future work.

5 Conclusions and Future Work

In this paper, we present an unsupervised approach to learning lexical patterns from online encyclopedia to extract part-whole relations. The only input is 5 pairs of part-whole instances. The major contributions of this paper include:

1) An algorithm to extract domain terms taking use of the semantic information contained in online encyclopedia is proposed, these terms are of great help for term recognition in relation identification;
2) A novel method to collect sentences that may indicate part-whole relation is described, compared to those approaches used in previous work, which set a high requirement for the corpus;
3) A new method to select reliable patterns. Similar patterns are clustered and pattern clusters are ranked according to their frequencies, those patterns from the top-k clusters are chose to be applied to identify the new part-whole relations.

Experimental results show that our method can extract abundant part-whole relations and achieve a preferable precision compared to the other approaches.

Future work focuses on improving the accuracy of our approach and applying the method to other relations.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant No.91224006, 61173063, 61035004, 61203284, 30973716 and National Social Science Foundation of China under grant No.10AYY003.

References

1. Hearst, M.A. *Automatic acquisition of hyponyms from large text corpora*. in *Proceedings of the 14th conference on Computational linguistics-Volume 2*. 1992. Association for Computational Linguistics.
2. Van Hage, W.R., H. Kolb, and G. Schreiber, *A method for learning part-whole relations*, in *The Semantic Web-ISWC 2006*. 2006, Springer. p. 723-735.
3. Iris, M.A. *Problems of the part-whole relation*. in *Relational models of the lexicon*. 1989. Cambridge University Press.
4. Winston, M.E., R. Chaffin, and D. Herrmann, *A taxonomy of part‐whole relations.* Cognitive science, 1987. **11**(4): p. 417-444.
5. Berland, M. and E. Charniak. *Finding parts in very large corpora*. in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. 1999. Association for Computational Linguistics.
6. Pantel, P. and M. Pennacchiotti. *Espresso: Leveraging generic patterns for automatically harvesting semantic relations*. in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. 2006. Association for Computational Linguistics.
7. Girju, R., A. Badulescu, and D. Moldovan, *Automatic discovery of part-whole relations.* Computational Linguistics, 2006. **32**(1): p. 83-135.
8. Cao, X., et al. *Extracting Part-Whole Relations from Unstructured Chinese Corpus*. in *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on*. 2008. IEEE.
9. Wu, J., B. Luo, and C. Cao, *Acquisition and verification of mereological knowledge from Web page texts.* JOURNAL-EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY, 2006. **32**(11): p. 1310.
10. Ruiz-Casado, M., E. Alfonseca, and P. Castells, *Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia.* Data & Knowledge Engineering, 2007. **61**(3): p. 484-499.
11. Hearst, M.A., *Automated discovery of WordNet relations.* WordNet: an electronic lexical database, 1998: p. 131-151.
12. Zhang, H.P. and Q. Liu, *ICTCLAS.* Institute of Computing Technology, Chinese Academy of Sciences: http://www. ict. ac. cn/freeware/003_ictclas. asp, 2002.
13. Mei, J., *Chinese Synonym Thesaurus*. 1983, Shanghai: Shanghai Lexicology Press.