NL2KB: Resolving Vocabulary Gap between Natural Language and Knowledge Base in Knowledge Base Construction and Retrieval

Sheng-Lun Wei, Yen-Pin Chiu, Hen-Hsen Huang, and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
{weisl, ypchiu, hhhuang}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

Abstract

Words to express relations in natural language (NL) statements may be different from those to represent properties in knowledge bases (KB). The vocabulary gap becomes barriers for knowledge base construction and retrieval. With the demo system called **NL2KB** in this paper, users can browse which properties in KB side may be mapped to for a given relational pattern in NL side. Besides, they can retrieve the sets of relational patterns in NL side for a given property in KB side. We describe how the mapping is established in detail. Although the mined patterns are used for Chinese knowledge base applications, the methodology can be extended to other languages.

1 Introduction

Knowledge bases (KBs) such as YAGO (Suchanek et al., 2007) and DBpedia (Lehmann et al., 2014) are useful resources in various applications such as question answering (Yih et al., 2015). KBs contain rich information of entities and their properties. A fact in a KB is usually represented as the form (*entity*1, *property*, *entity*2). Most KBs rely on manpower for editing and maintenance, so it is challenging to keep them up-to-date. Frank et al. (2012) point out the latency issue in knowledge base update. How to construct and update the knowledge base automatically is indispensable.

Mining facts from natural language (NL) statements and introducing them to knowledge base becomes a trend. In the sentence "蜜雪兒歐巴馬嫁給巴拉克奧巴馬" (Michelle Obama is married to Barack Obama), there are the two entities, i.e., 蜜雪兒歐巴馬 (Michelle Obama) and 巴拉克奧巴馬 (Barack Obama), and a relation 嫁給 (is married to) between them. In DBpedia, the relation 嫁給 (is married to) is represented as the property <spouse>. In other words, 嫁給 (is married to) in NL side is an NL relational pattern of the property <spouse> in KB side.

The vocabulary gap not only affects knowledge base construction, but also knowledge retrieval applications such as question answering. English relational patterns like PATTY (Nakashole et al., 2012) show efficacy on related applications (Dutta et al., 2015). In this work, we present a system for Chinese relation extraction and release a collection of human-verified Chinese relational patterns as a resource. We also demonstrate the applications of relational patterns on the demo website.

This paper is organized as follows. Section 2 surveys the related work. Section 3 describes the methodology. Section 4 shows and discusses the results. Section 5 demonstrates the **NL2KB** system.

2 Related Work

Information extraction (IE) models like ReVerb (Fader et al., 2011) automatically extract information from unstructured or semi-structured documents. Given an English sentence, ReVerb identifies two arguments and their relation in the form of (*argument*1, *relation*, *argument*2). PATTY (Nakashole et al., 2012) is a taxonomy system of relational patterns in English. From Wikipedia and the New York Times, 127,811 relational patterns are mined to describe 225 DBpedia properties, and 43,124 relational patterns are mined to describe 25 YAGO's properties. However, the coverage is still an issue.

Most open IE systems are developed for English, and few are for other languages. ZORE (Qiu et al., 2014) is a model that extracts relations from Chinese articles and presents them in the format of ReVerb style. However, this system does not deal with vocabulary mapping between NL and KB sides.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 282–286, Osaka, Japan, December 11-17 2016.

3 Method

In this paper, we extract relational patterns from the Chinese Wikipedia corpus and map them to the properties defined in DBpedia. In other words, the mapping between NL and KB is established. The DBpedia dataset used in our system was released on 8th May, 2014, and the dump of Chinese Wikipedia was released on 25th March, 2015. Figure 1 shows an overview of Chinese pattern extraction.

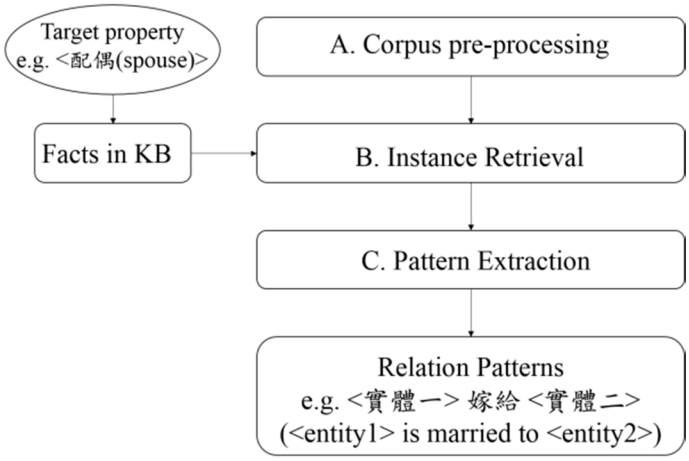


Figure 1: System overview.

3.1 Corpus Pre-processing

We discard all non-text information from the Chinese Wikipedia corpus such as html tags, xml tags, and cited tags, and perform sentence segmentation. Three punctuations, i.e., period, question mark, and exclamation mark, are regarded as sentence delimiters. After segmentations, we index each sentence into a search engine based on Solr[1] in order to do instance retrieval in the following step.

3.2 Alias Expansion

People may refer to an entity in different ways. For example, 貝拉克奧巴馬 (Barack Obama) is also called 巴拉克歐巴馬 (Barack Obama) and 巴拉克海珊歐巴馬二世 (Barack Hussein Obama II). We construct an alias dictionary for entities by collecting redirect pages from Wikipedia. The alias dictionary consisting of 1,317,829 entities is consulted for entity expansion to retrieve more instances from the corpus.

3.3 Instance Retrieval

If a sentence contains two entities and these two entities are connected with a property, we regard this sentence is an instance of the property. For each fact in DBpedia, we search the instances that describe the same fact in Chinese Wikipedia and extract relational patterns from these instances. All the sentences that contain the entity pair in the fact are retrieved. Figure 2 considers the target property "spouse" as an example to describe the process of instance retrieval.

3.4 Pattern Extraction

The instances retrieved by the method specified in Section 3.3 have some similar manifestations that are valuable to extract relational patterns from them. Figure 3 shows the process of pattern extraction in detail. First, Stanford toolkit[2] is performed to generate the dependency parse tree of each instance. Then, we find the shortest path between the two entities in the dependency tree, and regard the words in the shortest path as a relational pattern. Figure 4 shows the shortest path from 李雪主 (Ri Sol-ju) to 金正恩 (Kim Jong-un) is 李雪主 (Ri Sol-ju) => 嫁给 (is married to) =>金正恩 (Kim Jong-un). Thus, we regard (<*entity*1>, 嫁給 (is married to), <*entity*2>) as a relational pattern of the property <spouse>.

[1] http://lucene.apache.org/solr/
[2] http://stanfordnlp.github.io/CoreNLP/

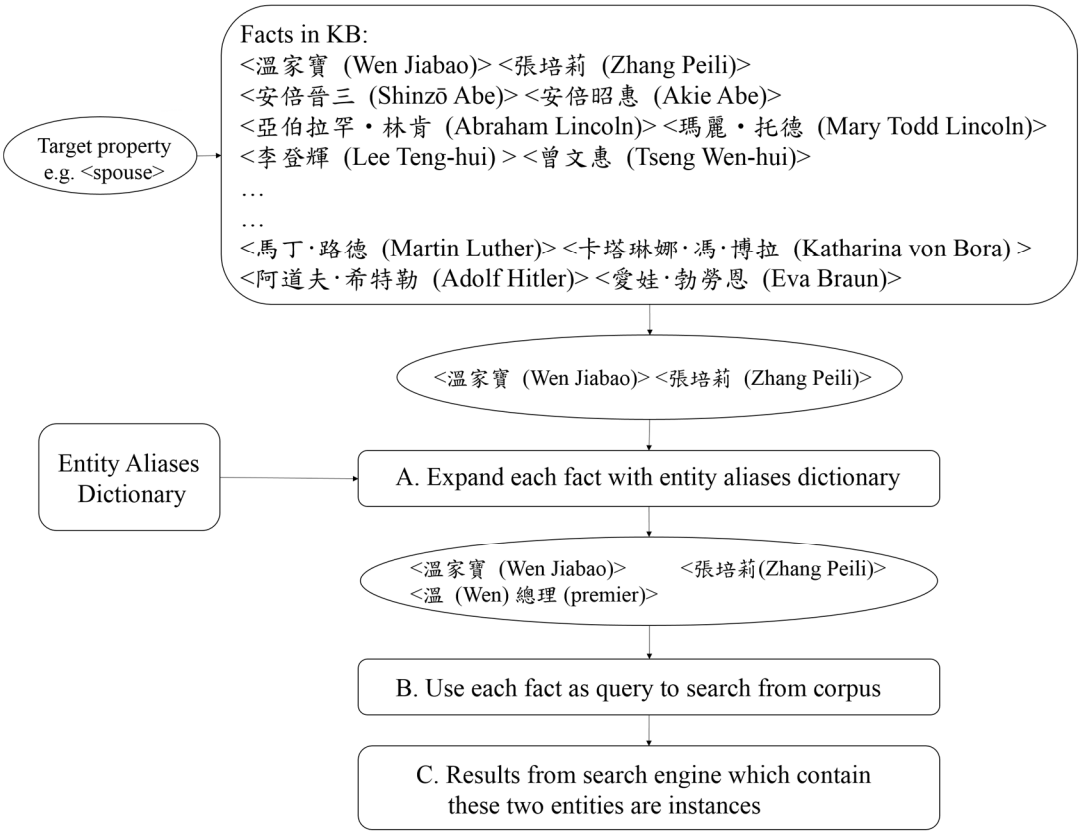


Figure 2: Instance retrieval.

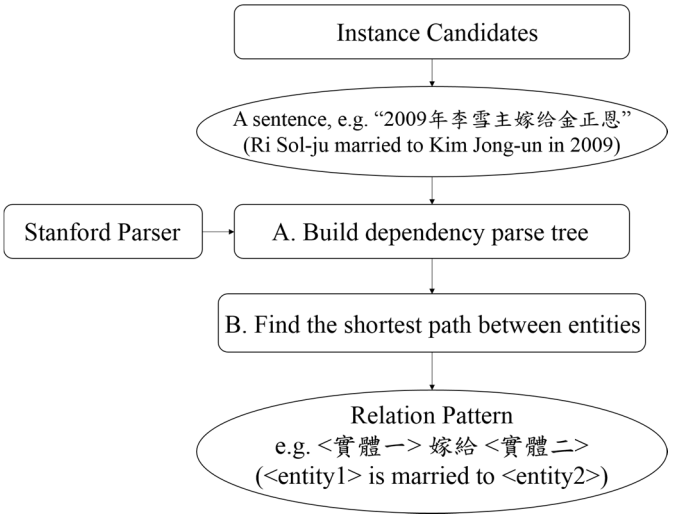


Figure 3: System for pattern extraction.

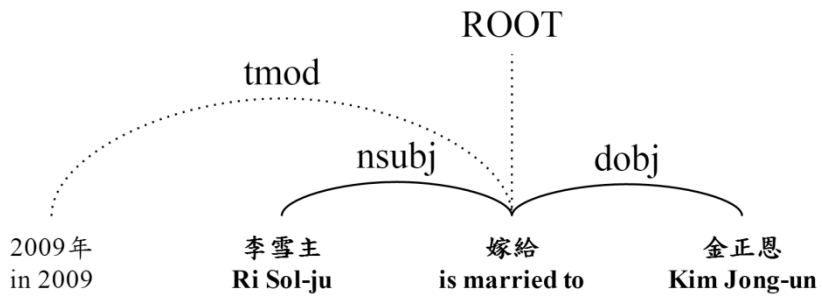


Figure 4: Dependency parse tree for a Chinese example.

4 Experiments and Analysis

There are 2,614 properties that contain at least 10 facts found in DBpedia. We exclude the properties <subdivisionType>, <subdivisionName>, and the properties related to <time zone>. A total of 2,608 properties remain as our target. We extract relational patterns for all of them. A minimum support threshold is set to 5 for each pattern, and the top 15 patterns for each property are selected. Finally, a total of 7,139 relational patterns covering 1,087 properties are collected.

To evaluate the performance of our method, each relational pattern is verified by three annotators, and the majority is taken as ground-truth. The Fleiss' kappa among the annotators is 0.52 (moderate agreement). P@5, P@10, and P@15 are 0.6, 0.597, and 0.587, respectively. The relational patterns can be downloaded from the website– http://nlg.csie.ntu.edu.tw/nlpresource/nl2kb/.

We also evaluate our relational patterns based on their part of speech (POS) tags. We focus on nouns and verbs. The results are shown in Table 1. "Verb" means the relational pattern consists of a single verb such as (<*entity*1>, 加盟 (join), <*entity*2>). "Noun" means the relational pattern consists of a single noun such as (<*entity*1>, 妻子 (wife), <*entity*2>). "Partial Verb" means the relational pattern consists of multiple words and contains a verb like (<*entity*1>, 運動員 (athlete) 效力 (play for), <*entity*2>). "Partial Noun" means the relational pattern consists of multiple words and contains a noun such as (<*entity*1>, 電視劇 (TV show) 主演 (starring), <*entity*2>). Obviously, the relational patterns containing verbs are more accurate than the noun-based patterns.

For each property, we search all instances of its facts. The more facts for a property, the more instances we retrieve. We divide our relational patterns into three groups, i.e., "Frequent", "Medium", and "Infrequent", by the number of facts. "Frequent" covers properties containing at least 1,000 facts such as <starring>, <author>, and <spouse>. "Medium" covers properties contain at least 100 facts and less than 999 facts such as <education>, <currency>, and <mother>. "Infrequent" covers properties containing at least 10 facts and less than 99 facts. Table 2 shows the results. For each group, the top 5 patterns always outperform the top 10 and top 15 ones. The group "Frequent" has the best performances, while "Infrequent" has the lowest ones. In other words, the more the facts, the more the reliable patterns.

POS Tags	# Patterns	P@15
Verb	1,311	**0.709**
Partial Verb	1,305	**0.641**
Noun	3,897	0.547
Partial Noun	1,718	0.575
All	7,139	0.587

Table 1: Performances in different POS tags.

	Frequent	Medium	Infrequent	All
# Patterns	2,333	3,481	1,325	7,139
P@5	**0.671**	0.602	0.534	0.600
P@10	**0.652**	0.596	0.523	0.597
P@15	**0.636**	0.581	0.515	0.587

Table 2: Performances in numbers of facts.

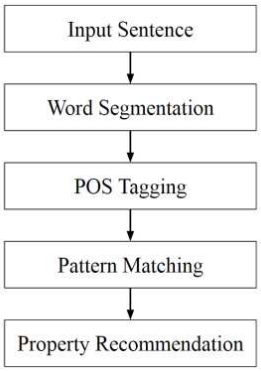
5 A Demo System

We demonstrate an application of our relational patterns on our website: http://nlg.csie.ntu.edu.tw/nlpresource/nl2kb/. Given a sentence in Chinese, our system will extract all the possible properties to which the relation in the sentence is mapped. As shown in Figure 5, the input sentence is first word segmented and POS tagged by the Stanford toolkit. Then pattern matching is applied to identify relations in the sentence, and the possible KB properties of each relation are recommended. We measure the score of each property by multiplying its support value and its confidence value. Finally, we show the results ranked by the scores.

Three functions shown as follows are demonstrated:

(1) Select a property and find all its relational patterns along with their support and confidence.
(2) Select a relational pattern and find all its properties along with their support and confidence.

(3) Enter a sentence and find which properties it contains. That is a fundamental task for knowledge base construction and retrieval.



1989 年 (in 1989) 江澤民 (Jiang Zemin) 當選/VV (was elected) 中共(CPC) 總書記 (general secretary)

1989 年/NT 江澤民/NR 當選/VV 中共/NR 總書記/NN

Property	Relational pattern	Support	Confidence
incumbent	<Entity1> 當選 <Enity2>	65	0.289
office	<Entity1> 當選 <Enity2>	47	0.209
party	<Entity1> 總書記 <Enity2>	39	0.250
leader1Name	<Entity1> 總書記 <Enity2>	36	0.231

Figure 5: The workflow for our demo system.

6 Conclusion

In this study, we create a Chinese relational pattern resource based on properties in the DBpedia knowledge base. We propose a system that extracts relational patterns by using the syntactic information. A total of 7,139 relational patterns that cover 1,087 properties are extracted and verified. We release the human-verified Chinese relational patterns as a resource (http://nlg.csie.ntu.edu.tw/nlpresource/nl2kb/), which can be utilized in various tasks such as knowledge base acceleration and question-answering. Although our system is designed for mining Chinese relational patterns, the methodology can be extended to other languages.

7 Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-102-2221-E-002-103-MY3 and MOST-105-2221-E-002-154-MY3, and National Taiwan University under grant NTU-ERP-104R890858.

References

Arnab Dutta, Christian Meilicke, and Heiner Stuckenschmidt. 2015. Enriching Structured Knowledge with Open Information. In *Proceeding of the 24th International Conference on World Wide Web* (*WWW*), pages 267-277.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 1535–1545.

John R. Frank, Max Kleiman-Weiner, Daniel A. Roberts, Feng Niu, Ce Zhang, Christopher Ré, and Ian Soboroff. 2012. Building an Entity-Centric Stream Filtering Test Collection for TREC 2012. In *Proceedings of the Twenty-First Text REtrieval Conference*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer. 2012. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* 1:1–29.

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: A Taxonomy of Relational Patterns with Semantic Types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1135–1145.

Likun Qiu and Yue Zhang. 2014. ZORE: A Syntax-based System for Chinese Open Relation Extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 1870–1880.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web* (*WWW*), pages 697–706.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1321–1331.