Extracting Semantic Relations for Mining of Social Data

Shinichi Nagano[1], Masumi Inaba[2], and Takahiro Kawamura[1]

[1] Corporate R&D Center, Toshiba Corporation, Japan
[2] Platform Solution Business Division, Toshiba Solution Corporation, Japan

Abstract. This paper proposes a novel method that extracts semantic relations from social data in order to acquire ontologies that are used for mining social data. A set of nouns are iteratively extracted from documents in a bootstrapping manner, and then a semantic relation between a noun pair is identified by a clustering procedure. The main feature is exploitation of the co-occurrence of a verb and a noun in a sentence, considering that a verb plays an important role in expressing the meaning of a sentence. The paper presents a preliminary study to clarify problems in order to achieve practical performance.

1 Introduction

The Social Web is communication infrastructure allowing people to argue, collaborate, or cooperate with other individuals or communities. Numerous opinions and experiences are mentioned on the Social Web. Thus, enterprises are eager to utilize social data for their advertising, marketing, or product planning[10, 17]. Mining social data involves digging up fragments of information scattered on the Web, and then discovering knowledge[15, 16]. Since semantic technologies have yet to catch up with the explosive growth in the publishing of data on the Social Web, mining the social data is still a challenging issue.

It is well known that ontologies are fundamental resources for semantic technologies, and their development from social data is being pursued around the world. Among the projects associated with this work are YAGO[1], DBpedia[3]. They are often provided in a machine readable, reusable, and extensible form, which makes it possible to develop an ontology for a certain purpose without building up from scratch. However, local ontologies dependent on their particular domains are often desired because generic ontologies are insufficient for certain purposes.

This paper addresses the issue of semantic relation extraction from documents on the Social Web. Most of the conventional methods are based on discovery of lexico-syntactic patterns that are dependency paths indicating a semantic relation between nouns[5]. Applicability of such methods is limited to term extraction from sentences matching the patterns and thus the methods need explore a large amount of documents.

The paper proposes a novel method that extracts terms in weak semantic relations from documents in a predetermined domain, yielding a hierarchical

form of the terms. A weak semantic relation means that terms in the relation might not exist on dependency paths. The main feature is exploitation of the co-occurrence of a verb and a noun in a sentence, considering that a verb plays an important role in expressing the meaning of a sentence. For instance, people often mention, on the Social Web, their daily experiences and opinions, which are characterized by verb-noun pairs such as *where to go*, *what to eat*, and *what to watch*. Prompted by this observation, the method identifies nouns appearing together with particular verbs at a high frequency. This is a hybrid method involving a bootstrapping procedure and a clustering procedure. The bootstrapping procedure is a process that initially extracts patterns from given verb-noun pairs and then alternately extracts pairs and patterns. Iterative application of the procedure yields a set of nouns. The clustering procedure identifies weak semantic relations for each pair of nouns obtained by the bootstrapping procedure, resulting in the hierarchical form of the nouns.

The remainder of this paper is organized as follows. Related work is discussed in Section 2. Next, Section 3 illustrates the proposed method. Then, Section 4 presents a preliminary evaluation of the method. Finally, Section 5 gives the conclusion and refers to future work.

2 Related Work

Many previous work on automatic extraction of terms in semantic relations has been based on a key insight of Hearst[4]. He noticed that the presence of certain lexico-syntactic patterns may indicate a particular semantic relation between two nouns. For instance, linking two noun phrases (NPs) via the constructions "Such NP_Y as NP_X" often implies that NP_X is a hyponym of NP_Y. Initially, a small number of handcrafted patterns like these were used to try to automatically label such semantic relations[7, 8, 11–14]. Following this approach, supervised learning algorithms were devised to obtain a large number of useful lexico-syntactic patterns. Snow et al.[5] proposed a generic method that formalizes lexico-syntactic patterns with dependency paths as features for prediction of hypernyms. Suchanek et al.[1, 2] applied a supervised learning algorithm to fact extraction from Wikipedia yielding a common ontology as social semantic knowledge.

Generic pattern classification is one of the most significant issues in semantic relation extraction. Generic patterns[13] have broad but noisy coverage. Difficulties in using these patterns have been a major impediment for supervised algorithms, resulting in either very low precision or recall. Espresso[9] is a novel unsupervised method that automatically excludes generic patterns and separates correct and incorrect noun pairs, giving higher reliance to patterns retrieved from correct pairs. Saeger et al.[6] proposed a weakly supervised method that classifies generic patterns with the use of semantic word classes acquired through learning class-dependent patterns.

Although this previous work has been successful in identifying pairs in handcrafted or learned patterns from a large number (typically more than a million)

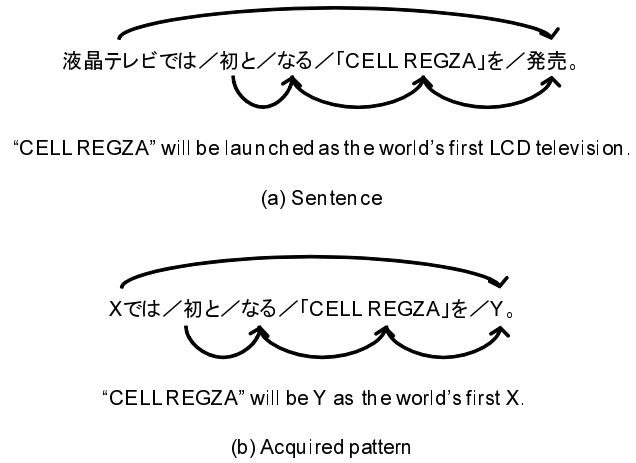

Fig. 1. Example of pattern acquisition

of documents, it is strictly limited to discovery from patterns, i.e., discovery is only possible in the case of that two nouns exist on a dependency path corresponding to a pattern. Our method is a novel one that identifies noun pairs in weak semantic relations, clustering nouns obtained by verb-noun patterns, instead of directly acquiring noun pairs from patterns. It is possible to obtain a noun pair that does not co-occur on a dependency path, and thus our method is widely applicable.

3 Semantic Relation Extraction

3.1 Problem Definition

A weak semantic relation is defined as a relationship between two nouns that appear together with particular verbs. As mentioned in Section 1, a verb often plays an important role in expressing the meaning of a sentence. If two nouns frequently appear with particular verbs, then it often implies that they are in a certain semantic relation. For instance, people often mention, on the Social Web, their daily experiences and opinions, which are characterized by verb-noun pairs such as *where to go*, *what to eat*, and *what to watch*. Clustering nouns appearing with eat could yield a hierarchy of nouns representing food names. The definition is derived from this observation.

A pattern is defined in this paper as a lexico-syntactic pattern, which is a set of dependency paths indicating a semantic relation between a verb and a noun.

3.2 Proposed Method

We outline the proposed method. Given a set of documents in a predetermined domain and a minimum set of verb-noun pairs in a designated semantic relation,

the method retrieves nouns in the relation from the documents and then yields the hierarchical form of obtained nouns. Note that the given relation is used for retrieving verb-noun pairs whereas our objective is to obtain nouns in a hypernym-hyponym relation.

The method consists of two procedures: bootstrapping and clustering. The bootstrapping procedure consists of pattern acquisition and term acquisition steps. The pattern acquisition step explores the documents to find a syntactic tree in such a way that the given verb-noun pair appears on the tree. Replacing the verb and noun with two variables, it acquires the tree as the relation extraction pattern. On the other hand, the term acquisition step finds the sentence that has the same syntactic structure as the pattern except for the variables. Identifying the terms corresponding to the variables, it newly acquires a verb-noun pair. This procedure performs both steps alternately and iteratively. As a result, it can acquire nouns in an incremental manner. An example of pattern acquisition is illustrated in Fig 1.

Following the bootstrapping procedure, the clustering procedure finds verbs co-occurring with each noun of a pair. Let $V(NP)$ be the set of verbs that appears with a noun phase NP. For two nouns NP_1 and NP_2, it determines whether or not the pair is in a hypernym-hyponym relation, comparing a subset relation between $V(NP_1)$ and $V(NP_2)$ as follows:

- If $V(NP_1) \subset V(NP_2)$ holds, then it is determined that NP_1 is a hyponym of NP_2 (NP_2 is a hypernym of NP_1).
- If $V(NP_2) \subset V(NP_1)$ holds, then it is determined that NP_2 is a hyponym of NP_1 (NP_1 is a hypernym of NP_2).
- Otherwise, it is determined that NP_1 is determined to be a synonym of NP_2.

4 Preliminary Evaluation

4.1 Overview of Evaluation

We apply the proposed method to a set of three hundred Web documents, which are blog posts mentioned on food products. We first label pairs of nouns in a weak semantic relation for the documents, and then conduct random sampling of the pairs. Our experiment starts at applying the bootstrapping procedure to a set of the sampled pairs as seed pairs. After five iteration of the procedure, the clustering procedure extracts a hypernym-hyponym relation for the food products, which is composed of a product category and a product name. Precision and recall of acquired term pairs are used as evaluation metrics. For analyzing Japanese sentences, JUMAN[3] and KNP[4] are used for a part-of-speech tagger and a dependency analyzer, respectively. Note that dictionaries for the tagger and the analyzer are not compiled for processing the input documents.

[3] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html
[4] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html

Table 1. Evaluation results

#Seed pairs	10	20	30	40	50
Precision	0.806	0.383	0.231	0.288	0.343
Recall	0.004	0.006	0.009	0.012	0.018

4.2 Evaluation Results

We summarize the evaluation results in Table 1. Each metric value represents an average over evaluation results for three distinctive sets of seed pairs. The given seed pairs have an influence on precision of the proposed method. Indeed, as shown in Table 1, precision value is 0.806 in case that the number of seed pairs is 10, and it is within a range of 0.2 to 0.4 in other cases. Investigating term pairs incorrectly acquired, we found that most of the pairs were extracted by generic patterns, which are high-frequency ones. Since a generic pattern could express different kinds of semantic relations, it should be classified into an adequate pattern class according to a context in a document. On the other hand, a number of false negative pairs resulted in low recall. It is primarily because most of the compound nouns and adjectives were not correctly extracted by dependency analysis. It should be integrated to identify compound nouns according to adjacency of terms.

5 Conclusion

The paper have proposed a novel method for semantic relation extraction and presented a preliminary evaluation. Improvement of the recall is a subject for future work.

References

1. F.M. Suchanek, G. Kasneci, G. Weikum, Yago - A large ontology from Wikipedia and WordNet, Journal of Web Semantics, vol.6, no.3, pp.203-217, 2008.
2. F. M. Suchanek, G. Ifrim, G. Weikum, Combining linguistic and statistical analysis to extract relations from web documents, Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining, 2006.
3. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, DBpedia - a crystallization point for the web of data, Journal of Web Semantics, vol.7, no.3, pp.154-165, 2009.
4. M. Hearst, Automatic acquisition of hyponyms from large text corpora, Proc. of the 14th International Conference on Computational Linguistics, 1992.
5. R. Snow, D. Jurafsky, A.Y. Ng, Learning syntactic patterns for automatic-hypernym discovery, Advanced in Neural Information Processing Systems, pp.1297-1304, 2005.
6. S.D. Saeger, K. Torisawa, J. Kazama, K. Kuroda, M. Murata, Proc. of the IEEE International Conference on Data Mining, pp.764-769, 2009.

7. A. Akbik, J. Brob, Wanderlust: extracting semantic relations from natural language text using dependency grammar patterns, Proc. of the Workshop on Semantic Search, pp.6-15, 2009.
8. P. Pantel, D. Ravichandran, Automatically labeling semantic classes, Proc. of North American Chapter of the Association for Computational Linguistics, 2004.
9. P. Pantel, M. Pennacchiotti, Espresso: leveraging generic patterns for automatically harvesting semantic relations, Proc. of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp.113-120, 2006.
10. R. Feldman, M. Fresko, J. Godenberg, O. Netzer, L.H. Ungar, Extracting product comparisons from discussion boards, Proc. of the IEEE International Conference on Data Mining, pp.469-474, 2007.
11. B. Rosenfeld, R. Feldman, Self-supervised relation extraction from the web, Knowledge and Information Systems, vol.17, no.1, pp.17-33, 2008.
12. M. Banko, O. Etzioni, The tradeoffs between open and traditional relation extraction, Proc. of the 46th Annual Meeting of the Association for Computational Linguistics, pp.28-36, 2008.
13. J.R. Curran, T. Murphy, B. Scholz, Minimising semantic drift with mutual exclusion bootstrapping, Proc. of the 10th Conference of the Pacific Association for Computational Linguistics, pp.172-180, 2007.
14. E. Agichtein, L. Gravano, Snowball: extracting relations from large plain-text collections, Proc. of the 5th ACM conference on Digital Libraries, pp.85-94, 2000.
15. K. Khan, B.B. Baharudin, A. Khan, F. e-Malik, Mining opinion from text documents: a survey, Proc. of the 3rd IEEE International Conference on Digital Ecosystems and Technologies, pp.217-222, 2009.
16. B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, vol.2, nos.1-2, pp.1-135, 2008
17. T. Kawamura, S. Nagano, M. Inaba, Y. Mizoguchi, WOM Scouter: mobile service for reputation extraction from weblogs, International Journal of Metadata, Semantics and Ontologies, vol.3, no.2, pp.132-141, 2008.