Extracting Semantic Taxonomies of Nouns from a Korean MRD Using a Small Bootstrapping Thesaurus and a Machine Learning Approach*

SeonHwa Choi and HyukRo Park

Dept. of Computer Science, Chonnam National University,
300 Youngbong-Dong, Puk-Ku Gwangju, 500-757, Korea
csh123@dreamwiz.com, hyukro@chonnam.ac.kr

Abstract. Most approaches for extracting hypernyms of a noun from the definition in an MRD rely on the lexico-syntactic patterns compiled by human experts. Not only these methods require high cost for compiling lexico-syntatic patterns but also it is very difficult for human experts to compile a set of lexical-syntactic patterns with a broad-coverage, because in natural languages there are various different expressions which represent the same concept. To alleviate these problems, this paper proposes a new method for extracting hypernyms of a noun from an MRD. In proposed approach, we use only syntactic(part-of-speech) patterns instead of lexico-syntactic patterns in identifying hypernyms to reduce the number of patterns while keeping their coverage broad. Our experiment shows that the classification accuracy of the proposed method is 92.37% which is significantly much better than those of previous approaches.

1 Introduction

The importance of broad-coverage lexical/semantic knowledge-bases has been stressed more than ever before as the natural language processing (NLP) systems became large and applied to wide variety of application domains. These lexical/semantic knowledge-bases include such as lexicons, thesauri and ontologies and machine-readable dictionaries. A lexical/semantic knowledge-base contains a list of terms, their semantic definitions and some of the relationships that exist between terms such as synonym, antonym and hypernym. Among the various relationships between terms, many researchers believe that the taxonomy relationship is especially useful because it can be utilized in various inference processes found in machine translation, information retrieval, word sense disambiguation and so on.

The taxonomy relationship is usually represented in thesauri as the broad-term (BT) narrow-term (NT) relations. However, because building broad-coverage thesauri is a very costly and time-consuming job, they are not readily available and often too general to be applied to a specific domain.

The work presented here is an attempt to alleviate this problem by devising a method for constructing taxonomy relations automatically from a machine readable dictionary (MRD). We use semantic hierarchies of nouns in a small thesaurus and a definition of a noun in a Korean MRD.

* This study was financially supported by special research fund of Chonnam National University in 2004.

A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 1–9, 2005.

Most of the previous approaches for extracting hypernyms of a noun from the definition in an MRD rely on the lexico-syntactic patterns compiled by human experts. Not only these methods require high cost for compiling lexico-syntactic patterns but also it is very difficult for human experts to compile a set of lexical-syntactic patterns with a broad-coverage because, in natural languages, there are various different expressions which represent the same concept. Accordingly the applicable scope of a set of lexico-syntactic patterns compiled by human is very limited.

To overcome the drawbacks of human-compiled lexico-syntactic patterns, we use part-of-speech (POS) patterns only and try to induce these patterns automatically using a small bootstrapping thesaurus and machine learning methods.

The rest of the paper is organized as follows. We introduce the related works to in section 2. Section 3 deals with the problem of features selection. In section 4, our problem is formally defined as a machine learning method and discuss implementation details. Section 5 is devoted to experimenal result. Finally, we come to the conclusion of this paper in section 6.

2 Related Work

[3] introduced a method for the automatic acquisition of the hyponymy lexical relation from unrestricted text, and gives several examples of lexico-syntactic patterns for hyponymy that can be used to detect these relationships including those used here, along with an algorithm for identifying new patterns. Her approach is complementary to statistically based approaches that find semantic relations between terms, in that hers requires a single specially expressed instance of a relation while the others require a statistically significant number of generally expressed relations. The hyponym-hypernym pairs found by Hearst's algorithm include some that she describes as "context and point-of-view dependent", such as "Washington/nationalist" and "aircraft/target". [4] was somewhat less sensitive to this kind of problem since only the most common hypernym of an entire cluster of nouns is reported, so much of the noise is filtered. [3] tried to discover new patterns for hyponymy by hand, nevertheless it is a costly and time-consuming job. In the case of [3] and [4], since the hierarchy is learned from text, it get to be domain-specific different from a general-purpose resource such as WordNet.

[2] proposed a method that combines a set of unsupervised algorithms in order to accurately build large taxonomies from any MRD, and a system that 1)performs fully automatic extraction of taxonomic links from MRD entries and 2) ranks the extracted relations in a way that selective manual refinement is allowed. In this project, he introduced the idea of the hyponym-hypernym relationship appears between the entry word and the genus term. Thus, usually a dictionary definition is written to employ a genus term combined with differentia which distinguishes the word being defined from other words with the same genus term. He finds the genus term by simple heuristic defined using several examples of lexico-syntactic patterns for hyponymy.

[1] presented the method to extract semantic information from standard dictionary definitions. His automated mechanism for finding the genus terms is based on the observation that the genus term from verb and noun definitions is typically the head of the defining phrase. The syntax of the verb phrase used in verb definitions makes it possible to locate its head with a simple heuristic: the head is the single verb follow-

ing the word *to*. He asserted that heads are bounded on the left and right by specific lexical defined by human intuition, and the substring after eliminating boundary words from definitions is regarded as a head.

By the similar idea to [2], [10] introduced six kinds of rule extracting a hypernym from Korean MRD according to a structure of a dictionary definition. In this work, she proposed that only a subset of the possible instances of the hypernym relation will appear in a particular form, and she divides a definition sentence into a head term combined with differentia and a functional term. For extracting a hypernym, she analyzes a definition of a noun by word list and the position of words, and then searches a pattern coinciding with the lexico-syntactic patterns made by human intuition in the definition of any noun, and then extracts a hypernym using an appropriate rule among 6 rules. For example, rule 2 states that if a word X occurs in front of a lexical pattern "*leul bu-leu-deon i-leum* (*the name to call*)",then X is extracted as a hypernym of the entry word.

Several approaches[11][12][13] have been researched for building a semantic hierarchy of Korean nouns adopting the method of [2].

3 Features for Hypernym Identification

Machine learning approaches require an example to be represented as a feature vector. How an example is represented or what features are used to represent the example has profound impact on the performance of the machine learning algorithms. This section deals with the problems of feature selection with respect to characteristics of Korean for successful identification of hypernyms.

Location of a Word. In Korean, a head word usually appears after its modifying words. Therefore a head word has tendency to be located at the end of a sentence. In the definition sentences in a Korean MRD, this tendency becomes much stronger. In the training examples, we found that 11% of the hypernyms appeared at the start, 81% of them appeared at the end and 7% appeared at the middle of a definition sentence. Thus, the location of a noun in a definition sentences is an important feature for determining whether the word is a hypernym or not.

POS of a Function Word Attached to a Noun. Korean is an agglutinative language in which a word-phrase is generally a composition of a content word and some number of function words. A function word denotes the grammatical relationship between word-phrases, while a content word contains the central meaning of the word-phrase.

In the definition sentences, the function words which attached to hypernyms are confined to a small number of POSs. For example, nominalization endings, objective case postpositions come frequently after hypernyms but dative postpositions or locative postpositions never appear after hypernyms. A functional word is appropriate feature for identifying hypernyms.

Context of a Noun. The context in which a word appears is valuable information and a wide variety of applications such as word clustering or word sense disambiguation make use of it. Like in many other applications, context of a noun is important in deciding hyperhyms too because hypernyms mainly appear in some limited context.

Although lexico-syntactic patterns can represent more specific contexts, building set of lexco-syntactic patterns requires enormous training data. So we confined ourselves only to syntactic patterns in which hypernyms appear.

We limited the context of a noun to be 4 word-phrases appearing around the noun. Because the relations between word-phrases are represented by the function words of these word-phrases, the context of a noun includes only POSs of the function words of the its neighboring word-phrases. When a word-phrase has more than a functional morpheme, a representative functional morpheme is selected by an algorithm proposed by [8].

When a noun appears at the start or at the end of a sentence, it does not have right or left context respectively. In this case, two treatments are possible. The simplest approach is to treat the missing context as don't care terms. On the other hand, we could extend the range of available context to compensate the missing context. For example, the context of a noun at the start of a sentence includes 4 POSs of function words in its right-side neighboring word-phrases.

4 Learning Classification Rules

Decision tree learning is one of the most widely used and a practical methods for inductive inference such as ID3, ASSISTANT, and C4.5[14]. Because decision tree learning is a method for approximating discrete-valued functions that is robust to noisy data, it has therefore been applied to various classification problems successfully.

Our problem is to determine for each noun in definition sentences of a word whether it is a hypernym of the word or not. Thus our problem can be modeled as two-category classification problem. This observation leads us to use a decision tree learning algorithm C4.5.

Our learning problem can be formally defined as followings:

- Task T: determining whether a noun is a hypernym of an entry word or not .
- Performance measure P: percentage of nouns correctly classified.
- Training examples E: a set of nouns appearing in the definition sentences of the MRD with their feature vectors and target values.

To collect training examples, we used a Korean MRD provided by Korean TermBank Project[15] and a Korean thesaurus compiled by Electronic Communication Research Institute. The dictionary contains approximately 220,000 nouns with their definition sentences while the thesaurus has approximately 120,000 nouns and taxonomy relations between them. The fact that 46% of nouns in the dictionary are missing from the thesaurus illustrates the necessity of this research i.e. to extend a thesaurus using an MRD.

Using the thesaurus and the MRD, we found that 107,000 nouns in the thesaurus have their hypernyms in the definition sentences in the MRD. We used 70% of these nouns as training data and the remaining 30% of them as evaluation data. For each training pair of hypernym/hyponym nouns, we build a triple in the form of (hyponym definition-sentences hypernym) as follows.

ga-gyeong	[a-leum-da-un gyeong-chi *(a beautiful scene)*]	gyeong-chi
hyponym	definition sentence	hypernym

Morphological analysis and Part-Of-Speech tagging are applied to the definition sentences. After that, each noun appearing in the definition sentences is converted into a feature vector using features mentioned in section 3 along with a target value (i.e. whether this noun is a hypernym of the entry word or not).

Table 1 shows some of the training examples. In this table, the attribute *IsHypernym* which can have a value either Y or N is a target value for given noun. Hence the purpose of learning is to build a classifier which will predict this value for a noun unseen from the training examples.

In Table 1, *Location* denotes the location of a noun in a definition sentence. 0 indicates that the noun appears at the start of the sentence, 1 denotes at the middle of the sentence, and 2 denotes at the end of a sentence respectively. *FW of a hypernym* is the POS of a function word attachted to the noun and *context1,...,context4* denote the POSs of function words appearing to the right/left of the noun. "*" denotes a don't care condition. The meanings of POS tags are list in Appendix A.

Table 1. Some of training examples.

Noun	*Location*	*FW of a hypernym*	*context1*	*context2*	*context3*	*context4*	*IsHypernym*
N1	1	jc	ecx	exm	nq	*	Y
N2	2	*	exm	ecx	jc	nq	Y
N3	2	*	exm	jc	nca	exm	Y
N4	1	exm	jc	jc	ecx	m	N
N5	1	jc	jc	ecx	m	jca	N
N6	1	jc	ecx	m	jca	exm	Y
N7	2	*	exm	exm	jca	exm	Y
N8	1	*	nc	jca	exm	jc	N
N9	1	jca	nc	nc	nc	jc	Y
N10	2	exn	a	nca	jc	nca	Y
..	..	..	..	..	..	..	..

Fig. 1 shows a part of decision tree learned by C4.5 algorithm. From this tree, we can easily find that the most discriminating attribute is *Location* while the least one is *Context*.

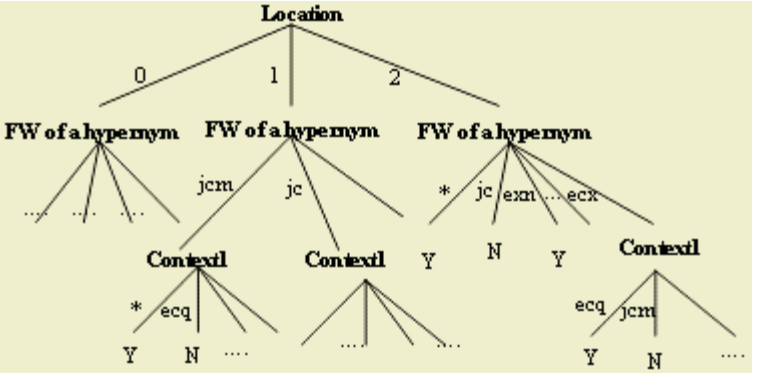

Fig. 1. A learned decision tree for task T.

5 Experiment

To evaluate the proposed method, we measure classification accuracy as well as precision, recall, and F-measure which are defined as followings respectively.

$$classification\ accuracy = \frac{a+d}{a+b+c+d}$$

$$precision = \frac{a}{a+b}$$

$$recall = \frac{a}{a+c}$$

$$F-Measure = \frac{2 * precision * recall}{precision + recall}$$

Table 2. Contingency table for evaluating a binary classifier.

	Yes is correct	No is correct
Yes was assigned	a	b
No was assigned	c	d

Table 3 shows the performance of the proposed approach. We conducted two experiments using 2 different definitions for the context of a word as stated in section 3. In the experiment denoted A in table 3, the context of a word is defined as 4 POSs of the function words, 2 of them immediately proceeding and 2 of them immediately following the word. In the experiment denoted B, when the word appears at the beginning of a sentence or at the end of a sentence, we used only right or left context of the word respectively. Our experiement shows that the performance of B is slightly better than that of A.

Table 3. Evaluation result.

	Classification accuracy	Precesion	Recall	F-Measure
A	91.91%	95.62%	92.55%	94.06%
B	92.37%	93.67%	95.23%	94.44%

Table 4 compares the classification accuracy of the proposed method with those of the previous works. Our method outperforms the performance of the previous works reported in the literature[10] by 3.51%.

Because the the performance of the previous works are measured with small data in a restricted domain, we reimplemented one of the those previous works[10] to compare the performances using same data. The result is shown in Table 4 under the column marked D. Column C is the performance of the [10] reported in the literature. This result shows that as the heuristic rules in [10] are dependent on lexical information, if the document collection is changed or the application domain is changed, the performance of the method degrades seriously.

Table 4. Evaluation result.

	Proposed		M.S.Kim 95[11]	Y.J.Moon 96[10]		Y.M.Choi 98[13]
	A	B		C	D	
classification accuracy	91.91%	92.37%	88.40%	88.40%	68.81%	89.40%

6 Conclusion

There have been several works to build a noun taxonomy from an MRD. However, most of them relied on the lexico-syntactic patterns compiled by human experts. Not only these methods require high cost for compiling lexico-syntactic patterns but also it is very difficult for human experts to compile a set of lexical-syntactic patterns with a broad-coverage, because in natural languages there are various different expressions which represent the same concept. Accordingly the applicable scope of a set of lexico-syntactic patterns compiled by human is very limited.

This paper has proposed a new method for extracting hypernyms of a noun from an MRD. In proposed approach, we use only syntactic patterns instead of lexico-syntactic patterns in identifying hypernyms to reduce the number of patterns while keeping their coverage broad. We also adopted a machine learning method to collect the patterns automatically.

Our experiment shows that the classification accuracy of the proposed method is 92.37% which is significantly much better than those of previous approaches. Throughout our research, we have found that machine learning approaches to the problems of identifying hypernyms from an MRD could be a competitive alternative to the methods using human-compiled lexico-syntactic patterns.

References

1. Martin S. Chodorow, Roy J. Byrd, George E. Heidorn.: Extracting Semantic Hierarchies From A Large On-Line Dictionary. In Proceedings of the 23rd Conference of the Association for Computational Linguistics (1985)
2. Rigau G., Rodriguez H., Agirre E.: Building Accurate Semantic Taxonomies from Mololingual MRDs. In Proceedings of the 36th Conference of the Association for Computational Linguistics (1998)
3. Marti A. Hearst.: Automatic acquisition of hyonyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics (1992)
4. Sharon A. Caraballo.: Automatic construction of a hypernym-labled noun hierarchy from text. In Proceedings of the 37th Conference of the Association for Computational Linguistics (1999).
5. Fernando Pereira, Naftali Thishby, Lillian Lee.: Distributional clustering of English words. In Proceedings of the 31th Conference of the Association for Computational Linguistics (1993)
6. Brian Roark, Eugen Charniak.: Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In Proceedings of the 36th Conference of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (1998)
7. Tom M. Mitchell.: Machine Learning. Carnegie Mellon University. McGraw-Hill (1997).
8. SeonHwa Choi, HyukRo Park.: A New Method for Inducing Korean Dependency Grammars reflecting the Characteristics of Korean Dependency Relations. In Proceedings of the 3rd Conterence on East-Asian Language Processing and Internet Information Technology (2003)
9. YooJin Moon, YeongTak Kim.:The Automatic Extraction of Hypernym in Korean. In Preceedings of Korea Information Science Society Vol. 21, NO. 2 (1994) 613-616
10. YooJin Mon.: The Design and Implementation of WordNet for Korean Nouns. In Proceedings of Korea Information Science Society (1996)

11. MinSoo Kim, TaeYeon Kim, BongNam Noh.: The Automatic Extraction of Hypernyms and the Development of WordNet Prototype for Korean Nouns using Koran MRD. In Proceedings of Korea Information Processing Society (1995)
12. PyongOk Jo, MiJeong An, CheolYung Ock, SooDong Lee.: A Semantic Hierarchy of Korean Nouns using the Definitions of Words in a Dictionary. In Proceedings of Korea Cognition Society (1999)
13. YuMi Choi and SaKong Chul.: Development of the Algorithm for the Automatic Extraction of Broad Term. In Proceedings of Korea Information Management Society (1998) 227-230
14. Quinlan J. R.: C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufman (1993) http://www.rulequest.com/Personal/
15. KORTERM.: KAIST language resources http://www.korterm.or.kr/

Appendix A: POS Tag Set

Table 5. POS tag set.

CATEGORY		TAG	DESCRIPTION
noun	common	nn	common noun
		nca	active common noun
		ncs	statove common noun
		nct	time common noun
	proper	nq	proper noun
	bound	nb	bound noun
		nbu	unit bound noun
	numeral	nn	numeral
	pronoun	npp	personal pronoun
		npd	demonstrative pronoun
predicate	verb	pv	verb
	adjective	pa	adjective
		pad	demonstrative adjective
	auxiliary	px	auxiliary verb
modification	adnoun	m	adnoun
		md	demonstrative adnoun
		mn	numeral adnoun
	adverb	a	general adverb
		ajs	sentence conjunctive adverb
		ajw	word conjunctive adverb
		ad	demonstrative adverb
independence	interjection	ii	interjection
particle	case	jc	case
		jca	adverbial case particle
		jcm	adnominal case particle
		jj	conjunctive case particle
		jcv	vocative case particle
	auxiliary	jx	auxiliary
	predicative	jcp	predicative particle

Table 5. (continued).

CATEGORY		TAG	DESCRIPTION
ending	Prefinal	efp	prefinal ending
	conjunctive	ecq	coordinate conjunctive ending
		ecs	subordinate conjunctive ending
		ecx	auxiliary conjunctive ending
	transform	exn	nominalizing ending
		exm	adnominalizing ending
		exa	adverbalizing ending
	final	ef	final ending
affix	prefix	xf	prefix
	suffix	xn	suffix
		xpv	verb-derivational suffix
		xpa	adjective-derivational suffix