Building Disambiguation System for Compound Noun Analysis Based on Lexical Conceptual Structure

Koichi Takeuchi, Kyo Kageura and Teruo Koyama
National Institute of Informatics (NII)
2-1-2, Hitotsubashi, Chiyodaku, Tokyo,
101-8430, Japan
{koichi, kyo, t_koyama}@nii.ac.jp

Abstract

In this paper, we propose a principled approach for disambiguating relations between constituent words of compound nouns whose heads are deverbal nouns, using the framework of lexical conceptual structure. The aim of this research is to reveal the complete set of lexical factor and disambiguation rules needed for application. The results of experiment for Japanese deverbal compounds and nominalization of English compounds show that our approach is highly promising.

1 Introduction

In recent lexical semantics, theoretically oriented approach has been proposed and making steady progress. In order to apply these lexical semantic theories to a practical natural language processing (NLP) system, however, we need to deal with the following two problems:

(1) linguistic theories tend to give only a framework and fragments of descriptions. In NLP applications, it is necessary to have a complete set of lexical factors to achieve sufficient coverage.

(2) linguistic theories are descriptive, while in NLP applications it is necessary to use them for processings such as disambiguation.

These are essential for constructing a practical NLP system based on some linguistic theories. This cannot be solved in the theoretical research in lexical semantics but should be approached from the application point of view, because the same theoretical framework may have to be modified or emphasis should be shifted to suit for particular applications; this may in turn contribute to the theoretical research.

In this paper, we propose a principled method for disambiguating the relations between constituent elements of compound nouns whose heads are deverbal nouns, using the theoretical framework of lexical conceptual structure (LCS). We developed the framework on the basis of Japanese data, but the experimental results for English also shows that the same approach is promising. We show how the lexical factors and disambiguation mechanisms are related for compound noun analysis.

2 Compound Noun Analysis

2.1 Previous work

The existing work on compound noun analyses takes either the statistical approach or the semantic approach. The former is more concerned with contextual aspects of compounding, while the latter with lexical aspects.

Statistical techniques (Lauer, 1995; Buckeridge and Sutcliffe, 2002) are useful for broad-coverage sallow analysis when training methods are available.

On the other hand, semantic approaches explore types of relations between constituents in compounds (Isabelle, 1984; Levi, 1978; Iida et al., 1984). Some of the approaches (Frabre, 1996; Takahashi, 2002) are based on the framework of Generative Lexicon (GL) (Pustejovsky, 1995). Semantic approaches based on GL are especially well designed but they did not still show the complete lexical factors needed for the analysis model. These are essential for not only the extendibility of the semantic approaches but also estimation for lexical semantics from the view of the application.

2.2 Our approach

In this paper we try to establish a semantic framework of the analysis of compounds whose heads are deverbal[1] nouns. Deverbal compounds constitute a major part of compound nouns and to develop a method to deal with deverbal compounds is an essential element of compound noun analyzer. We focus on compounds with only two constituents words, more composed compounds are basically constructed by the recursive binary rules. For the analysis of deverbal compounds, we propose a method based on LCS. As LCS gives a clear framework for describing verb semantics, the lexicon of deverbal nouns can be constructed consistently and is thus extendable to a large scale, which is another advantage of the LCS-based method.

3 Framework of Compound Noun Analysis and TLCS

The framework of LCS (Hale and Keyser, 1990; Rappaport and Levin, 1988; Jackendoff, 1990; Levin and Hovav, 1995; Kageyama, 1996; Sugioka, 1997) has shown that semantic decomposition based on the LCS framework can systematically explain the word formation as well as the syntax structure. However existing LCS frameworks cannot be applied to the analysis of compounds straightforwardly because they do not give restriction rules nor extensive semantic predicates for LCS. Therefore we construct an original LCS, called TLCS,[2] based on the LCS framework with a clear set of LCS types and basic predicates. We use the acronym "TLCS" to avoid the confusion with other LCS-based schemes.

3.1 Relation between a noun and a deverbal noun

The relations between the words in deverbal compounds can be divided into two: (i) the modifier becomes an internal argument (Grimshaw, 1990) of the deverbal head, and (ii) the modifier functions as an adjunct. The disambiguation of these two relations is an essential element in compound nouns analysis. For example, take the following two Japanese compounds.

kikai	sousa
machine	operate
(machine	operation)

[1] In the case of English the equivalent is nominalizations, but for simplicity we use deverbal compounds.

[2] 'T' in 'TLCS' denotes the initial of terminology as well as the first character of the first author's name.

kikai hon'yaku
machine translate
(machine translation)

The modifier 'kikai' is the internal argument of the deverbal head in the former, while it is the adjunct in the latter. In English compound case, as you see in English translation in example, it is the same[3] as Japanese case. This disambiguation level is simple but basic for expanding detailed analysis of relations.[4]

3.2 Compound noun analysis using TLCS

We assume that the relation can be determined by the combination of the TLCS on the side of deverbal heads and the consistent categorization of modifier nouns on the basis of their behavior vis-à-vis a few canonical TLCS types of deverbal heads.

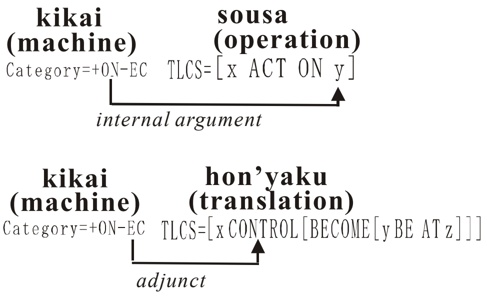

Figure 1: Example of disambiguation of relations using TLCS types of deverbal heads and categorization of modifier nouns.

Figure 1 shows examples of disambiguating relations using the TLCS types of deverbal heads. The description in square brackets denotes the TLCS for the deverbal heads 'soursa' (operate) and 'hon'yaku' (translate). In TLCSes, the words written in capital letters are semantics predicates, 'x' denotes the external argument, and 'y' and 'z' denote the internal arguments.

Our approach consists of three elements: (1) categorization for deverbals, (2) for nouns and (3) restriction rules for identifying relations. In the next sections, we will sketch them briefly.

4 TLCS

Based on the existing work on LCS (Kageyama, 1996; Sugioka, 1997), we established a TLCS, i.e. a set of original predicates and basic structure types that can describe the semantic structure of deverbal nouns for compound noun analyzer. The following list is for Japanese deverbals, but the same basic structure is applied for English verbs.

Table 1: List of TLCS

1 [x ACT ON y]
enzan (calculate), sousa (operate)

2 [x CONTROL[BECOME
[y BE AT z]]]
kioku (memorize), hon'yaku (translate)

3 [x CONTROL[BECOME
[y NOT BE AT z]]]
shahei(shield), yokushi (deter)

4 [x CONTROL [y MOVE TO z]]
densou (transmit),
dempan (propagate)

5 [x=y CONTROL[BECOME
[y BE AT z]]]
kaifuku (recover), shuuryou (close)

6 [BECOME[y BE AT z]]
houwa (become satulated),
bumpu (be distributed)

[3] It is necessary to be given the verbal root of the nominalization like 'operate' from 'operation'.

[4] Sugioka (1997) shows the approach to deal with detailed relations such as 'method' and 'cause' in Japanese compounds based on extended LCS.

7 [y MOVE TO z]
idou (move), sen’i (transmit)
8 [x CONTROL [y BE AT z]]
iji (maintain), hogo (protect)
9 [x CONTROL[BECOME
[x BE WITH y]]]
ninshiki (recognize), yosoku (predict)
10 [y BE AT z]
sonzai (exist), ichi (locate)
11 [x ACT]
kaigi (hold a meeting),
gyouretsu (queue)
12 [x CONTROL[BECOME
[[FILLED]y BE AT z]]]
shomei (sign-name)

Table 1 shows the current complete set of TLCS types we established. The numbers attached to each TLCS type in Table 1 will be used throughout the paper refer to specific TLCS types. Examples of Japanese deverbal nouns are also given as well in Table 1. In the Table, the capital letters (such as ‘ACT’ and ‘BE’) are semantic predicates, which are 11 types. ‘x’ denotes an external argument and ‘y’ and ‘z’ denote an internal argument (see Kageyama (1996) and Levin and Hovav (1995)).

5. Categorization of Modifier Nouns

The essential underlying assumption of the categorization is this: If the LCS (and TLCS) represents can contribute to explaining phenomena related to the argument structure in a principled way, then, correspondingly, there should be some general and principled categorization of nouns. On this account, nouns in the modifier position of compounds categorized according to TLCS as well as the property of them.

5.1 Categorization by the accusativity of modifiers

In Japanese compounds, there is the modifier without its accusative. This is an adjectival stem and it does not appear with inflections. Therefore, the modifier is always the adjunct in the compounds. So we introduce the distinction of `-ACC' (unaccusativity) and `+ACC' (accusativity). For example, `kimitsu' (secrecy) and `kioku' (memory) are `+ACC', and `sougo' (mutual-ity) and `kinou' (inductiv-e/ity) are `-ACC'.

In English, modifiers categorized in unaccusativity correspond to the following two types: the one is an adjective modifier such as ‘recursive’ or ‘semantic’, and the other is a word with both characteristics of adjectival and nominal usually behave as adjectival in compound nouns such as ‘serial’ and ‘polynomial’.

5.2 Categorization by the basic components of TLCS

From the preliminary examination, we have found that some TLCS types can be formed into the groups that correspond to modifier categories in Table 2. For example, TLCS **2**, **3** and **4** form the group that corresponds to the modifier category `EC'. This means that TLCS types in the group are regarded as the same nature from the view of the relation to the modifier category.

In order to categorize nouns, we check whether they appear in sentences as an object of the verb whose TLCS has each of these specific components. If a noun does not appear as the object of each component, the noun is categorized as a negative category denoted by ‘-’. If it does, ‘+’ is assigned. Below are examples of modifier nouns categorized as negative or positive in terms of each of these TLCS components.[5]

ON ‘koshou’ (fault) and ‘seinou’ (performance) are ‘+ON’, and ‘heikou’ (parallel) and ‘rensa’ (chain) are ‘-ON’.

EC ‘imi’ (semantic) and ‘kairo’ (circuit) are ‘+EC’, and ‘kikai’ (machine) and ‘densou’ (transmission) are ‘-EC’.

IC ‘fuka’ (load) and `jisoku' (flux) are `+IC', and `kakusan' (diffusion) and `senkei' (linear) are `-IC'.

UA `jiki' (magnetic) and `joutai' (state) are `+UA', and `junjo' (order) and `heikou' (parallel) are `-UA'.

6 Disambiguation Rules for Compound Noun Analysis

The noun categories introduced in section 5 can be used for disambiguating the intra-term relations in deverbal compounds with various deverbal heads that take different TLCS types. The range of application of the noun categorizations with respect to TLCS types is summarized in Table 2. The number in the TLCS column corresponds to the number given in Table 1.

The procedure of our compound noun analyzer follows in accordance with table 2.

Step 1. If the modifier has the category ‘-ACC’, then declare the relation as adjunct and terminate. If not, go to next.

Step 2. If the TLCS of the deverbal head is **10**, **11**, or **12** in Table 1, then declare the relation as adjunct and terminate. If not, go to next.

Step 3. The analyzer determines the relation from the interaction of lexical meanings between a deverbal head and a modifier noun. In the case of ‘-ON’, ‘-EC’, ‘-IC’ or ‘-UA’, declare the relation as adjunct and terminate. If not, go to next. It is the advantage of our approach to realize such a disambiguation based on semantic restriction.

Step 4. Declare the relation as internal argument and terminate.

Table 2: Disambiguation rule of combination of modifier nouns and TLCS of deverbals

relation type	Modifier category	TLCS
Adjunct	-ACC	any
	Any	10,11,12
	-ON	1
	-EC	2,3,4
	-IC	5
	-UA	6,7
Int. argument	Other combinations	

With these rules and categories of nouns, we can analyze the relations between words in compounds with deverbal heads. For example, when the modifier `kikai' (machine) is categorized as ‘-EC’ but ‘+ON’, the modifier in *kikai-hon’yaku* (machine translation) is analyzed as adjunct (that means `translation by a machine'), and the modifier in *kikai-sousa* (machine operation) is analyzed as internal argument (that means `operation of a

[5] ‘ON’ stands for the predicate ‘ON’ in the ‘ACT ON’ of TLCS. ‘EC’ and ‘IC’ stands for ‘external controllability’ and ‘internal controllability’. ‘UA’ stands for ‘unaccusativity’.

Table 3: Statistics of disambiguation rules applied to the correct analysis

process	Relation type	modifier category	TLCS	freq. in Japanese	freq. in English
Step1	Adjunct	-ACC	Any	263	87 (54+33)
Step2		Any	10,11,12	88	1
Step3		-ON	1	95	10
		-EC	2,3,4	186	14
		-IC	5,	26	1
		-UA	6,7	59	0
Total of step2 and step3				454	26
Step4	int. arg.	other combinations		498	82
		Total		1215	195

machine'), both correctly.[6]

7 Experiments and Results

We applied the method to 1223 two-constituent compound nouns with deverbal heads in Japanese. 809 of them are taken from a dictionary of technical terms, and 414 from news articles in a newspaper. We also applied the method to 200 compound nouns of technical terms in English. In the experiment, we assumed that input words are segmented in Japanese and the root verbs of head nouns are given in English.

According to the manual evaluation of the experiment, 99.3% (1215 words) of the results were correct in Japanese, and 98% (195 words) were correct in English. The performance is very high.

Table 3 shows the details of how the rules are applied to disambiguating the relations between constituent words in the deverbal compounds. The numbers in blankets show that 87 cases analyzed as unaccusative modifier in English consist of 54 adjectivals and 33 modifier nouns. These results indicate that our set of lexical factors has the enough to disambiguate the relationships we assumed.

8 Diagnosis and Discussion

All in all, our approach can be available both Japanese and English deverbal nouns. Comparing with the results between Japanese compounds and English compounds, the factor '-ACC' looks effective to disambiguate relations, but the degree of its effectiveness are different between them. In adjunct relations of English, '-ACC', i.e., step1 takes major part of analysis comparing

[6] Since our approach checks only an adjunct relation, there is possibility that the relation of internal argument categorized by our approach becomes an adjunct relation in further steps of analyses that are contextual and background knowledge. For example 'dog hunting' would be categorized as internal argument at the current framework. We assume that other knowledge (ex. context or background knowledge) also contributes to final decision mechanism of relations in compound nouns.

with step2 and step3, while most of the adjunct relations are analyzed using step2 and step3 in Japanese compounds. The reason is that the most of modifiers in English indicate adjective function by using adjectival modifiers and analyzed as adjunct relations by '-ACC' of step1, while most of Japanese modifiers take no inflection then the relation can be analyzed effectively by step2 and step3 taking the advantage of TLCS.

The adjectival modifiers in English (that are 54 cases in Table 3) can be analyzed as adjunct relation using just part-of-speech (POS) of modifiers. Our categorization '-ACC' is effective for disambiguation of them because '-ACC' not only takes into account POS information, but also it can deal with other 33 modifiers that have both POSes of noun and adjective (ex. 'serial' or 'polynomial').

We found that a small number of modifier nouns deviate from our assumptions. The most typical case is that our analysis model fails in a word with multiple semantics. For example, 'right justify' is misunderstood as internal argument relation because of ambiguity of the word 'right' which has both meanings as 'keep to the right' and as 'human rights'. We consider dealing with them as each different words like 'right_1', 'right_2' in future work.

Our approach is easy to be extended to the framework of GL taking a methodology like Fabre (1996). However her approach does not look easy to disambiguate the relations like 'machine operation' and 'machine translation' because root verbs that are 'operate' and 'translate' have the same types of 'agent' and 'theme' in argument structure. On this account, the disambiguation level of our approach is a little deeper than Fabre's approach though our current model only deals with deverbal compounds.

9 Conclusion

In this paper, we propose a principled approach for disambiguating relations between constituent words of compound nouns whose heads are deverbal nouns, using the framework of lexical conceptual structure we call TLCS. We apply our approach to 1223 two-constituent compound nouns with deverbal heads in Japanese, and 200 compound nouns with nominalized heads in English. The method analyzed 99.3% (1215) of the Japanese compounds and 98 % (195) of the English compounds correctly.

The experimental results show that LCS-based lexical factors can catch the decision mechanism of adjunct/internal argument relations in deverbal compounds. This means that our approach trying to clarify set of lexical factors from the view of practical application model is highly promising.

As a next step of our research, detailed analysis of adjunct relations should be needed. Sugioka (1997) showed extended LCS framework can explain detailed adjunct relations, but her work does not show how the lexical factor of modifier noun contributes to disambiguation of the relations. On this point of view, the categorization of modifier nouns we proposed here is important to be applied to extended theory based on LCS.

The other important direction of our research is an analysis of noun-noun compounds. Some of the approaches take advantage of noun-verb relation (Fabre, 1996; Koyama, 2001). Fabre applies the framework of GL to analysis of relations between nouns through arguments of the related verbs in the qualia structure. Since our approach deals with noun-verb relations, it is high perspective that our approach can be expanded to analysis of noun-noun compounds straightforwardly.

Acknowledgments

We thank Prof. T. Kageyama, for giving us important suggestions on the nature of LCS. This work was supported in part by Japan Society for the Promotion of Science (grant no. 14780313).

References

M. Lauer. 1995. Designing Statistical Language Learners: Experiments on Noun, Compounds. Ph.D. thesis, Department of Computing, Macquarie University.

A. M. Buckeridge and R. F. E. Sutcliffe. 2002. Disambiguation Noun Compounds with Latent Semantic Indexing. Proceedings of the 2nd International Workshop on Computational Terminology (COMPUTERM2002), pages 71-77.

P. Isabelle. 1984. Another Look at Nominal Compounds. In Proceedings of COLING-84, pages 509-516.

J. N. Levi. 1978. The Syntax and Semantics of Complex Nominals. Academic Press.

H. Iida, K. Ogura, and H. Nomura. 1984. Analysis of Semantic Relations and Processing for Compound Nouns in English. In Proceedings of Information Processing Society of Japan, SIG Notes, NL, 46-4, (in Japanese), pages 1-8.

C. Fabre. 1996. Interpretation of Nominal Compounds: Combining Domain-Independent and Domain-Specific Information. COLING-96, pages 364-369.

S. Takahashi, S. Lee, K. Mogi, M. Kobayashi and S. Sato. 2002. Building Dynamic Lexical Model Based on the Framework of the Generative Lexicon. Proceedings of Information Processing Society of Japan, SIG Notes, 2002-NL-150, (in Japanese) pages 125-132.

J. Grimshaw. 1990. Argument Structure. MIT Press.

J. Pustejovsky. 1995. The Generative Lexicon. MIT Press.

K. Hale and S. Keyser. 1990. A View from the Middle Lexicon (Lexicon Project Working Papers 10). MIT.

M. Rappaport and B. Levin. 1988. What to do with theta-roles. In W. Wilkins, editor, Thematic Relations (Syntax and Semantics 21), pages 7-36. Academic Press.

R. Jackendoff. 1990. Semantic Structures. MIT Press.

T. Kageyama. 1996. Verb Semantics. Kurosio Publishers, (in Japanese).

Y. Sugioka. 1997. Projection of Arguments and Adjuncts in Compounds. In Grant-in-Aid for COE Research Report (1) (No. 08CE1001), (in Japanese), pages 185-220.

K. Takeuchi, K. Uchiyama, M. Yoshioka, K. Kageura, and T. Koyama. 2001. Categorising Deverbal Nouns Based on Lexical Conceptual Structure for Analysing Japanese Compounds. In Proceedings of the IEEE SMC 2001 Conference, pages 904-909.

T. Koyama. 2001. Structural Analysis of Japanese Compound Terms by Inserting Verbs, NII Journal, No.2, (in Japanese), pages 39-44.