Deverbal Compound Noun Analysis Based on Lexical Conceptual Structure

Koichi Takeuchi **Kyo Kageura** **Teruo Koyama**
Human and Social Information Research Division
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyodaku, Tokyo 101-8430, Japan
{koichi,kyo,t_koyama}@nii.ac.jp

Abstract

This paper proposes a principled approach for analysis of semantic relations between constituents in compound nouns based on lexical semantic structure. One of the difficulties of compound noun analysis is that the mechanisms governing the decision system of semantic relations and the representation method of semantic relations associated with lexical and contextual meaning are not obvious. The aim of our research is to clarify how lexical semantics contribute to the relations in compound nouns since such nouns are very productive and are supposed to be governed by systematic mechanisms. The results of applying our approach to the analysis of noun-deverbal compounds in Japanese and English show that lexical conceptual structure contributes to the restrictional rules in compounds.

1 Introduction

The difficulty of compound noun analysis is that the effective way of describing the semantic relations in compounds has not been identified. The description should not remain just a kind of categorization. Rather, it should take into account the construction of the analysis model.

The previous work proposed semantic approaches based on semantic categories (Levi, 1978; Isabelle, 1984; Iida et al., 1984) had proposed detailed analysis of relations between constituents in compound nouns. Some of approaches (Fabre, 1996; Johnston and Busa, 1998) take the framework of Generative Lexicon (GL) (Pustejovsky, 1995). Semantic approaches are especially well designed but they should still clarify the complete lexical factors needed for analysis model.

Probabilistic approaches (Lauer, 1995; Lapata, 2002) have been proposed to disambiguate semantic relations between constituents in compounds. Their experimental results show a high performance, but only for shallow analysis of compounds using semantically tagged corpora. To be fully effective, they also need to incorporate factors that are effective in disambiguating semantic relations. It is therefore necessary to clarify what kinds of factors are related to the mechanisms that govern the relations in compounds.

Against this background, we have carried out a research which aims at clarifying how lexical semantics contribute to, independently of languages, the relations in compound nouns. This paper proposes a principled approach for the analysis of semantic relations between constituents in compound nouns based on the theoretical framework of lexical conceptual structure (LCS), and shows that the framework originally developed on the basis of Japanese compound noun data works well for both Japanese and English compound nouns.

2 The Basic Framework

2.1 The Relation between Modifier and Deverbal Head

The relation between constituents in deverbal compounds[1] can first be divided into two: (i) the modifier becomes an internal argument (Grimshaw, 1990) and (ii) the modifier functions as an adjunct. We as-

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

sume these two kinds of relations are the target of our analysis model because argument/adjunct relations are basic but extensible to more detailed semantic relations by assuming more complex semantic system. Besides these relations related to argument structure of verbs are the boundary between syntax and semantics, then our approach must be extendable to be incorporated into sytactic analysis.

2.2 LCS-based Disambiguation Model

We assume that the discrimination between argument and adjunct relations can be done by the combination of the LCS (we call 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 shows examples of disambiguating relations using TLCS for the deverbal heads ‘sousa’ (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 (see Section 3).

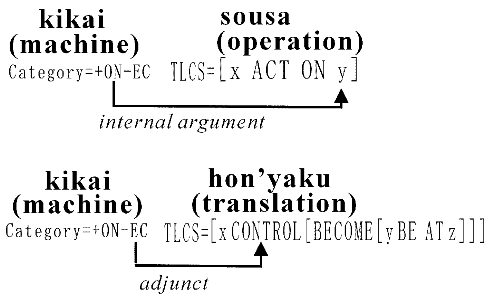

Figure 1: Disambiguation of relations between noun and deverbal head

The approach we propose consists of three elements: categorization of deverbals and nominalizations, categorization of modifier noun and restriction rules for identifying relations.

3 TLCS

The framework of LCS (Hale and Keyser, 1990; Rappaport and Levin, 1988; Jackendoff, 1990; Kageyama, 1996) 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 extensive semantic predicates for LCS. Therefore we construct an original LCS, called TLCS, 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.

Table 1 shows the current complete set of TLCSes types we elaborated.[2] The following list is for Japanese deverbals, but the same LCS types are applied for nominalizations in English.[3]

Table 1: List of TLCS types

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 saturated)
bumpu (be distributed)
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)

The number attached to each TLCS type in Table 1 will be used throughout the paper refer to specific TLCS types. In Table 1, 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 (Grimshaw, 1990)). [4]

[2] Basicaly these 12 types are set by the combination of argument structure and aspect analysis that is telic or atelic. After applying all the combination, we arrange the TLCS patterns by deleting patterns that does not appear and subcategorizing certain patterns.

[3] At the moment, there are about 500 deverbals in Japanese and 40 nominalizations in English.

[4] In this paper, we limit the types of arguments are three, i.e. x (Agent), y (Theme) and z (Goal).

4 Categorization of Modifier Noun

4.1 Categorization by the Accusativity of Modifiers

In Japanese compounds, some of modifiers can not take an accusative case. 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' (unaccusative) and '+ACC' (accusative).

ACC 'kimitsu' (secrecy) and 'kioku' (memory) are '+ACC', and 'sougo' (mutual-ity) and 'kinou' (inductiv-e/ity) are '-ACC'. In English, they correspond to adjective modifier such as '-ent' of 'recurrent' or '-al' of 'serial'.

4.2 Categorization by the Basic Components of TLCS

If, as argued by some theoretical linguists, the LCS representation can contribute to explaining these phenomena related to the arguments and aspect structure consistently, and if the combination of LCS and noun categorization can explain properly these phenomena related to argumet/adjunct, then there should be a level of consistent noun categorization which matches the LCS on the side of deverbals. We used the predicates of some TLCS types to explore the noun categorizations.

In the preliminary examination, we have found that some TLCS types can be formed into the groups that correspond to modifier categories in Table 2.

Below are examples of modifier nouns categorized as negative or positive in terms of each of these TLCS groups.

ON 'koshou' (fault) and 'seinou' (performance) are '+ON', and 'heikou' (parallel) and 'rensa' (chain) are '-ON'. ('ON' stands for the predicate in 'ACT ON'.)

EC 'imi' (semantic) and 'kairo' (circuit) are '+EC', and 'kikai' (machine) and 'densou' (transmission) are '-EC'. ('EC' stands for an External argument Controls an internal argument'.)

AL 'fuka' (load) and 'jisoku' (flux) are '+AL', and 'kakusan' (diffusion) and 'senkei' (linearly) are '-AL'. ('AL' stands for alternation verbs.)

UA 'jiki' (magnetic) and 'joutai' (state) are '+UA', and 'junjo' (order) and 'heikou' (parallel) are '-UA'. ('UA' stands for UnAccusative verbs.)

5 Procedure of Compound Noun Analysis

The noun categories introduced in Section 4 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 groups is summarized in Table 2. The number in the TLCS column corresponds to the number given in Table 1.

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','-AL' or '-UA', declare the relation as adjunct and terminate. If not, go to next.

Step 4 Declare the relation as internal argument and terminate.

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 machine'), both correctly.

6 Experiments and Evaluations

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 (Aiso, 1993), and 414 from news articles in a newspaper. We also applied the method to 200 compound nouns of technical terms (Aiso, 1993) in English. They are extracted randomly.

According to the manual evaluation of the experiment, 99.3% (1215/ 1223) of the results were correct in Japanese, and 97% (194/200) in English. The performance is very high. Table 2 shows the details of how the rules are applied to disambiguating the relations between constituents in the deverbal compounds. These results indicate that our set of LCS and categorization of modifiers has the enough to disambiguate the relationships we assumed.

Table 2: Combination of modifiers and TLCS of deverbal heads,and statistics of the correct analysis

role	mod. cat.	TLCS	Jap.(%)	Eng. (%)
adjunct	-ACC	any	263 (36.7)	84 (75.0)
	any	10,11,12	88 (12.3)	4 (3.6)
	-ON	1	95 (13.3)	10 (8.9)
	-EC	2,3,4	186 (25.9)	14 (12.5)
	-AL	5	26 (3.6)	0 (0.0)
	-UA	6,7	59 (8.2)	0 (0.0)
	total		717	112
role	mod. cat.	TLCS	Jap.(%)	Eng.(%)
int. argu.	+ACC	8, 9	74 (14.9)	15 (18.3)
	+ON	1	89 (17.9)	19 (23.2)
	+EC	2,3,4	249 (50.0)	43 (52.4)
	+AL	5	57 (11.4)	3 (3.7)
	+UA	6,7	29 (5.8)	2 (3.4)
	total		498	82

7 Discussion

Roughly speaking, our LCS-based 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. The reason is that the most of modifiers indicate adjective function by adding suffixes in English. While in Japanese, adjectival nouns of modifiers have no inflecitons, then the semantic-based approach is needed for Japanese compound noun analysis.

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 of an adjective and a noun. We consider dealing with them as each different words like 'right_adj', 'right_noun' in future work.

8 Conclusion

This paper proposes a principled approach for analysis of semantic relations between constituents in compound nouns based on lexical conceptual structure we call it TLCS. The results of experiment for Japanese compounds and English compounds show our approach is highly promising, also the contribution of the lexical factor to disambiguation rule.

References

Hideo Aiso. 1993. *Dictionary of Technical Terms of Information Processing (Compact edition)*. Ohmusha. (in Japanese).

Cecile Fabre. 1996. Interpretation of Nominal Compounds: Combining Domain-Independent and Domain-Specific Information. In *Proceedings of COLING-96*, pages 364–369.

Jane Grimshaw. 1990. *Argument Structure*. MIT Press.

Ken Hale and Samuel J. Keyser. 1990. *A View from the Middle Lexicon (Lexicon Project Working Papers 10)*. MIT.

Jin Iida, Kentaro Ogura, and Hirosato 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.

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

Ray Jackendoff. 1990. *Semantic Structures*. MIT Press.

Michael Johnston and Federica Busa. 1998. The Compositional Interpretation of Nominal Compounds. In E. Viegas, editor, *Breadth and Depth of Semantics Lexicons*. Kluwer.

Taro Kageyama. 1996. *Verb Semantics*. Kurosio Publishers. (In Japanese).

Maria Lapata. 2002. The Disambiguation of Nominalization. *Association for Computational Liguistics*, 28(3):357–388.

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

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

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

Malka Rappaport and Beth Levin. 1988. What to do with θ-roles. In W. Wilkins, editor, *Thematic Relations (Syntax and Semantics 21)*, pages 7–36. Academic Press.