An LCS-Based Approach for Analyzing Japanese Compound Nouns with Deverbal Heads

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

Abstract

This paper describes a principled approach for analyzing relations between constituent words of compound nouns whose heads are deverbal nouns. To develop an analyzer for the compound nouns with deverbal heads is an essential element of developing of a general compound analyzer, as they constitute a major part of the compound nouns. Our approach is based on the classification of deverbal nouns by their lexical conceptual structure (LCS) and the classification of nouns in general (to appear in the modifier position) vis-à-vis a few core LCS types (of head deverbal nouns). The experimental evaluation based on compound nouns with deverbal heads showed that over 99% of the compounds were accurately analyzed. The result of the experiment indicates that our approach is very promising.

1 Introduction

This paper proposes a principled approach to analyzing Japanese compound nouns whose heads are deverbal nouns (henceforth deverbal compounds for succinctness), using the framework of lexical conceptual structure (LCS).

Deverbal compounds constitute a major part of Japanese compound nouns, especially in technical terms where noun compounds are abundant,[1] and to develop a method to deal with deverbal compounds is an essential element of developing a full-scale, high-performance compound noun analyzer. We focus on compounds with only two constituent words, as more complex compounds are basically constructed by the iteration of binary rules (Kageyama, 1999).

[1] For instance, the ratio of technical terms that contains deverbal nouns is 78% in a dictionary of information processing (Aiso, 1993)

Deverbal nouns may behave as verbs, assigning thematic roles to the other noun constituent in a compound, or they may behave as nouns accepting a thematic role or modifying the other constituent. On this account, the relation between constituent words in a deverbal compound may be modification or *adjunct* relation (when the deverbal head functions as an ordinary noun), or it may be thematic role or *internal argument* relation (when the deverbal head takes a verbal role). Though there are various kinds of semantic relations that should be identified in compound nouns, the disambiguation of modification and internal argument is the first and essential element in compound noun analysis.

For the analysis of deverbal compounds, we propose a method based on LCS (Jackendoff, 1990; Kageyama, 1996). We will show that, by adopting LCS-based approach, it is possible to disambiguate the relation between constituent words in deverbal compounds with high accuracy. 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.

In the following, after briefly discussing the background and the overall framework of the approach, we sketch the basic idea of using LCS for the analysis of deverbal compounds. We will then elaborate our LCS system (TLCS) in section 4, noun classification in section 5, and procedure of analysis in section 6. In section 7, we will show the result of experimental evaluation, which is highly promising.

2 Background

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.

Lauer (1995) and Kobayasi et al. (1994) used corpus-based statistical techniques for bracketing compound nouns. Barker and Szpakowicz (1998) proposed a semi-automatic approach to analyzing and relations between constituents in nouns phrases. Statistical techniques are useful for broad-coverage shallow analysis when training methods are available.

On the other hand, semantic approaches explore types of relations between constituents in compounds (Levi, 1978; Isabelle, 1984). Iida et al. (1984) applied semantic framework of (Levi, 1978) and (Isabelle, 1984) to the analysis of English compounds. Semantic approaches are important for the deeper analysis, but the construction of the necessary lexicon tends to be too much dependent on a particular system and the lexicon is not easily extendable.

2.2 The basic design of our compound analyzer

In this paper we try to establish a semantic framework of the analysis of compound nouns without depending on particular applications or specific domain knowledge. It is because the formation of compound nouns mainly depends on lexical nature of their constituent words.

In semantic approach, it is important to clarify what kind of lexical nature is relevant to what kind of phenomena (i.e. relations) in what type of compound nouns, and to develop a framework and method of defining and describing the necessary semantic information consistently. The semantic framework should be clear enough so that the required lexicon can be extendable consistently to a very large size. The work reported in this paper, i.e. an LCS-based approach for deverbal compound analysis, constitutes a core of this overall framework.

3 Basic Framework: Compound Noun Analysis and TLCS

Recent studies in linguistics (Hale and Keyser, 1990; Rappaport and Levin, 1988; Jackendoff, 1990; Levin and Hovav, 1995; Kageyama, 1996; Sugioka, 1997; Sugioka, 2000) have shown that the semantic decomposition based on the LCS framework can systematically explain the word formation as well as the syntax structure. However existing LCS frameworks can not 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 or with the general idea of LCS.

In this section we will briefly sketch how the expanded categories for nouns and deverbal nouns based on the TLCS framework can be used in disambiguating the intra-compound relations in deverbal compounds. Our framework, as having a formal structure and procedural restrictions, also has an advantage that the consistency of describing lexical meanings can be maintained.

3.1 Argument structure

The approach, using TLCS, is based on argument structure (Grimshaw, 1990). Argument structure is a simple semantic structure for a verb to express thematic role delivered to nouns which are arguments. There are two types arguments, i.e. external argument ('Agent') and internal argument ('Theme' and 'Goal')[3]. For example, we describe the argument structure of a Japanese deverbal 'sousa' (operate) as follows:

sousa (operate): (Agent ⟨ Theme ⟩).

Agent is the operator and Theme is the object that is operated. $\langle\cdot\rangle$ denotes an internal argument.

In deverbal compounds, it is only the internal arguments that can take a predicate-argument relation (Kageyama (1996)); elements of external arguments cannot establish a predicate-argument relation in deverbal compounds. This restriction is essential in disambiguating relations between constituent words in deverbal compounds.

[2] 'T' denotes the initial of terminology as well as the first author.

[3] In this paper, we limit the types of thematic role to three, i.e. Agent, Theme and Goal.

3.2 Relation between a noun and a deverbal noun

As mentioned, the relations between the words in Japanese deverbal compounds can be divided into two: (i) the modifier becomes an internal argument of the deverbal head, and (ii) the modifier functions as an adjunct. The disambiguation of these two relations is an essential element in compound noun analysis.[4]

Take, for example, the following two compounds.[5]

kikai	-sousa	kikai	-hon'yaku
machine	-operate	machine	-translate
(machine operation)		(machine translation)	

The modifier 'kikai' is the internal argument of the deverbal head in the former, while it is the adjunct in the latter.

3.3 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-á-vis a few canonical TLCS types (or semantic predicates) of the deverbal heads. Figure 1 shows examples of disambiguating relations using the TLCS types of deverbal heads.

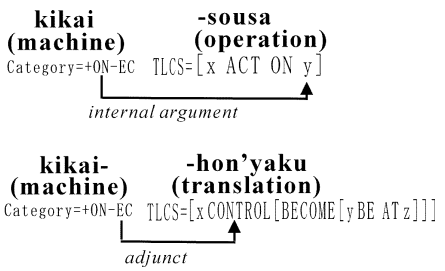

Figure 1: Disambiguating relations using TLCS types of deverbal heads.

The description in square brackets denotes the TLCS for the deverbal heads 'sousa' (operate) and 'hon'yaku' (translate). In TLCSes, the words written in capital letters are semantic predicates, 'x' denotes the external argument, and 'y' and 'z' denote the internal arguments. We will elaborate how the disambiguation is done in the following three sections.

4 TLCS

The first lexical information essential for the compound analyzer is the LCS-based classification scheme of deverbal nouns and the lexicon. Based on the existing work on LCS (Kageyama, 1996; Sugioka, 1997), we have established a TLCS, i.e. a set of original predicates and basic structure types that can describe the semantic structure of deverbal nouns for the Japanese compound noun analyzer.

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)

We established 12 TLCSes as listed in Table 1.[6] The number attached to each TLCS type in Table 1 will be used throughout the paper

[4] The level of the analysis is language independent, and the proposed approach can be applied to compounds in other languages when the lexicon is available.

[5] The first line denotes a Japanese compound, the second shows word-to-word translation and the third shows translation of the compound. A mark '-' designate that forming one word by connecting with preceding morpheme.

[6] In the actual establishment, we referred to the data of technical terms mentioned in section 7. Takeuchi et al. (2001) describes in detail how to establish LCS scheme and assignment of verbs to LCSes.

to refer to specific TLCS types. Examples of deverbal nouns are also given as well in Table 1.

In Table 1, as in Figure 1, words written in capital letters (such as 'CONTROL' and 'ACT') are semantic predicates. 'x' denotes an external argument, and 'y' and 'z' denote an internal argument (see (Kageyama, 1996; Levin and Hovav, 1995)).

TLCS **1** $\sim$ **4**, **8** and **9** represent different types of transitive verb. TLCS **11** and **12** are the types of intransitive verb. TLCS **5** represents the ergative verb and, **6**, **7** and **10** are unaccusative verbs.

The semantic predicates have combinatorial restrictions, which can be described in BNF as shown in Table 2 (⟨VERB⟩ is the root, i.e. the semantic structure of deverbal nouns).

Table 2: Restriction of TLCS predicates

⟨VERB⟩	::=	⟨BE⟩|⟨MOVE⟩|⟨BECOME⟩
		|⟨CONTROL⟩|⟨ACT⟩
⟨BE⟩	::=	BE AT|NOT BE AT|BE WITH
⟨MOVE⟩	::=	MOVE TO
⟨BECOME⟩	::=	BECOME ⟨BE⟩
⟨CONTROL⟩	::=	CONTROL ⟨BE⟩|CONTROL
		⟨MOVE⟩|CONTROL ⟨BECOME⟩
⟨ACT⟩	::=	ACT|ACT ON

Structures which are not listed in Table 1, such as '[CONTROL [BE WITH]]' can be derived from the rules in Table 2. It is however unlikely for them to occur in the real data.

4.1 TLCS Components

Each TLCS in Table 1 consists of the combination of semantic components that form basic meaning unit, as follows:

- 'y BE AT z' means that 'y' exists at 'z' as a state or a place.
- 'BECOME' represents that a current state changes into the next state completely. For example, '[BECOME[y BE AT z]]' means changing into the situation that 'y' exists at 'z'.
- 'y MOVE TO z' means that 'y' changes its place to 'z'.
- 'x CONTROL' means 'x' controls a predicate represented by a component that follows.
- 'x=y' represents an internal argument 'y' can change the state by itself.
- 'x ACT ON y' means the continuous action of 'x' to 'y' without changing the state of 'y'.
- 'x ACT' means the continuous action of 'x'.
- 'x BE WITH y' means that 'x' owns 'y'.
- 'NOT' represents negation.
- '[FILLED]y' means that the argument 'y' is filled with something and the verb of this TLCS cannot take an internal argument.

Table 3 shows the typology of TLCSes in relation to the TLCS components. 'Both' in Table 3 indicates that TLCS **5** have both transitive and intransitive nature[7] which is known as the transitivity alternation. Table 3 also shows the relation between traditional grammatical categories of verbs and the TLCS-based categorization of verbs.

Table 3: Typology of TLCS

argument type	TLCS num.	key compo.
transitive	1,	ACT ON
	2,3,4,8,9	CONTROL
intransitive	6,7	BECOME
	10	BE AT
	11	ACT
	12	[FILLED]y
both	5	x=y

5 Categorization of Modifier Nouns

The second lexical information necessary for the compound noun analyzer is the categorization of nouns which are to be in the modifier position of the deverbal compounds. The essential underlying assumption is this: If, as claimed in (Jackendoff, 1990; Kageyama, 1996), the LCS (and TLCS) represents a linguistically proper lexical information of verbs (including deverbal nouns) and can contribute to explaining phenomena related to the argument structure in

[7] For example, transitive case is *kaigi-o shuuryou-suru* 'meeting-ACC end-do' [end a meeting], and intransitive case is *kaigi-ga shuuryou-suru* 'meeting-NOM end-do' [a meeting ends]. Here 'ga' and 'o' designate case markers of nominative (NOM) and accusative (ACC), respectively. This phenomenon also appears in English (ex. 'break'), which is called ergative (Kageyama, 1996).

a principled way, then, correspondingly, there should be some general and principled categorization of nouns that also consistently contribute to describing phenomena related to argument structure AND this categorization can be defined vis-á-vis some basic components of the LCS scheme in a generalized and consistent manner, together with some general grammatical characteristics. Below we explain the noun categorization thus established, together with the basic features of the categorization criteria.

- **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' (unaccusative) and '+ACC' (accusative).
  For example, 'kimitsu' (secrecy) and 'kioku' (memory) are '+ACC', and 'sougo' (mutual-ity) and 'kinou' (inductiv-e/ity) are '-ACC'.
- **Categorization by the basic components of TLCS**
  The basic components that contribute to the general categorization of nouns are 'ON', 'CONTROL', 'x=y' and 'BECOME [y BE AT z]'. They are used in constructing TLCS types **1**, **2**, **5** and **6**, respectively (see Table 1).
  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. In Table 4 and in the discussion below, the categories of 'ON', 'CONTROL', 'x=y' and 'BECOME [y BE AT z]' are denoted as 'ON', 'EC(external control)', 'IC(internal control)' and 'UA(unaccusative)'. Below are examples of modifier nouns categorized as negative or positive in terms of each of these TLCS components.

  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' (linearly) are '-IC'.

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

6 Procedure of 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 4. The number in the TLCS column corresponds to the number given in Table 1.

Table 4: Categorization of combination of modifier nouns and TLCS of deverbal heads.

role	modifier category	TLCS
adjunct	-ACC	any
	-ON	1
	-EC	2,3,4
	-IC	5
	-UA	6,7
	any	10,11,12
role	modifier category	TLCS
int. argument	+ACC	8, 9
	+ON	1
	+EC	2,3,4
	+IC	5
	+UA	6,7

TLCS types **10**, **11**, and **12** do not take the internal argument relation in compounds.[8] TLCS types **8** and **9** take the internal argument relation when the modifier is '+ACC'.

Some TLCS types are formed into the groups that correspond to modifier categories in Table 4. 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.

[8] Even though TLCS **10** has the argument 'y', we have found that the verbs in **10** always behave as ordinary nouns in compounds.

The actual analysis proceeds as follows:

Step 1 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 2 If the modifier has the category '-ACC', 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.

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.

7 Experimental Evaluation

7.1 Experiments and Results

We applied the method to 1223 two-constituent compound nouns with deverbal heads. 816 of them are taken from a dictionary of technical terms (Aiso, 1993), and 415 from news articles in a newspaper (Nikkei newspaper). In the experiment, we assumed that the compounds were segmented.

According to the manual evaluation of the experiment, 99.3% (1215 words) of the results were correct. The performance is very high. Table 5 shows the details of how the rules are applied to disambiguating the relations between constituent words in the deverbal compounds. All in all, it shows that 366 or about 30% of the disambiguation is done in Step 3 above, by referring to the relation between TLCS categories and the noun categories (while the single most frequently used category of disambiguation is '-ACC' used in step 2). This means that the method we proposed is highly important in disambiguating the intra-term relation of deverbal compounds.

Table 5: Statistics of effective rules applied to the correct analysis

process	applied rules	frequency
Step 1	in case of **10**,**11**,**12**	88
Step 2	-ACC	263
Step 3	total in step 3	366
	-ON	95
	-EC	186
	-IC	26
	-UA	59
Step 4	internal argument	498
	total	1215

7.2 Diagnosis and Discussions

We found that a small number of modifier nouns deviate from our assumptions (among the 1223 terms, we found eight). Typical cases are:

jiki	-shahei	jiki	-kioku
magnet	shield	magnet	memorize
(magnet shield)		(magnet memory)	

In the former case, the modifier is the internal argument, while in the latter it is the adjunct. The TLCS of the former deverbal head is the type **3** and that of the latter head is **2** in Table 1. The TLCS of each deverbal head is different, however, the categorization of the both TLCS types is the same according to the definition in Table 4. We categorize these TLCS types into the same group of combination rule because the TLCS **2**, **3** and **4** are a kind of causation type in Table 4. At the moment, we do not take account of the difference between TLCS **2** and **3** with the negative predicate 'NOT' because we do not have enough evidence about this. Further examination is needed to deal with cases of this type.

8 Conclusions

In this paper, we proposed a linguistic method for the analysis of compound nouns with deverbal heads. The main element of the method is the introduction of the original lexical conceptual structure, TLCS, for the deverbal nouns

and the consistent categorization of the modifier nouns. Though the relations that can be disambiguated by the current framework is limited, the performance by precision is very high.

As the method is based on the linguistically clear and well-motivated perspective, the method is very promising for further extension of token coverage of the same type of phenomena as it will be easy to extend the lexicon while keeping the consistency. In addition, though not reported here, it will be possible, according to our current examination, to extend the coverage of the types of relations to be disambiguated with extensions of this framework.

Acknowledgments

We thank Prof. T. Kageyama, for giving us important suggestions on the nature of LCS. We also thank Nihonkeizashinbunsha for allowing us to use their newspaper articles (1992–1998). This work was supported in part by Japan Society for the Promotion of Science (grant no. 14780313).

References

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

K. Barker and S. Szpakowicz. 1998. Semi-Automatic Recognition of Noun Modifier Relationships. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics*, pages 96–102.

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

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

J. 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.

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

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

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

T. Kageyama. 1999. Word Formation. In *The Handbook of Japanese Linguistics*, pages 297–325. Oxford Blackwell.

Y. Kobayasi, T. Tokunaga, and H. Tanaka. 1994. Analysis of Japanese Compound Nouns using Collocational Information. In *Proceedings of COLING-94*, pages 865–869.

M. 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.

B. Levin and M. R. Hovav. 1995. *Unaccusativity.* MIT Press.

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

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

Y. Sugioka. 2000. *Transitivity Alternations in Deadjectival Verbs.* COE International Symposium at Kanda University of International Studies. http://coe-sun.kuis.ac.jp/coe/ public/paper/outside/sugioka3.pdf.

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.