PRICAI '90

The Automatic Extraction of Conceptual Items from Bilingual Dictionaries

TOKUNAGA Takenobu TANAKA Hozumi

Department of Computer Science Tokyo Institute of Technology
Ôokayama, Meguro, Tokyo 152, Japan
take@cs.titech.ac.jp

Abstract

To improve the quality of machine translation systems, we should step towards the deeper analysis at the conceptual level. This paper proposes a method to extract the items of the conceptual dictionary from a pair of machine readable bilingual dictionaries. The basic idea is to consider each word sense defined in the dictionaries as a conceptual item, and to make pairs of word senses corresponding to the word sense of each dictionary. The pair of word senses can be regarded as a concept item which is shared by both the languages. The paper also describes the outline of a preliminary experiment conducted to verify the method. The results of the experiment are promising and some remarks are also given. We conclude the paper with pointing out the possibility to construct the set of conceptual items which are shared by more than two languages, by extending our method.

1 Introduction

To improve the quality of machine translation systems, we should step towards the deeper analysis at the conceptual level. The surface analysis is not enough to select the appropriate word in the translation because of ambiguity in word senses [1]. Since there are two types of concepts, that is, the concepts which are unique to a language and the concepts which are universal over languages, it is necessary to distinguish them when we talk about the concepts.

Developing the machine translation systems with deeper analysis requires the dictionaries including at least the following information:

(1) the set of conceptual items,

(2) the mapping relation between the surface words and the conceptual items,

(3) the correspondence between the conceptual items of the source language and that of the target language.

There are several researches which aim to compile such dictionaries. Japan Electronic Dictionary Research Institute (EDR) is now compiling such dictionaries on a large scale [2]. Nirenburg et al. at Carnegie Mellon University has proposed a systematic method to construct a conceptual dictionary [3]. These attempts try to compile the dictionary by hand with the help of software tools. However this approach suffers from the problems such as huge amount of manual work, and the unstable result [4]. In particular, to obtain the correspondence between the conceptual items of each language ((3) above) requires enormous manual work. This is because the number of the candidates to be verified becomes the product of the number of conceptual items of each language.

Unlike this approach, this paper proposes a method to construct a conceptual dictionary which includes three types of informations mentioned above in an automatic way by using bilingual dictionaries. Here, a machine readable bilingual dictionary means a bilingual dictionary which is compiled for human and is in the computer readable format. It is very difficult to compile the complete conceptual dictionary in a fully automatic way. The results of the method may require some refinement and modifications by human. Our goal is rather to automate the compilation process as much as possible and to decrease manual work.

In this paper, we make an approximation in which each word sense defined in the bilingual dictionary is a conceptual item. Since each word sense represents a certain meaning and has the proper translation words in a bilingual dictionary, this approximation is reasonable in terms of word choice in the machine translation, and we can easily obtain both (1) the set of conceptual items and (2) the mapping relation between the surface words and the conceptual items. We use the term "word sense" and "conceptual item" interchangeably in the rest of the paper unless the distinction is explicitly mentioned.

The most difficult thing is to obtain the correspondence between the conceptual items of the source language and that of the target language. The rest of the paper focuses on this issue. To obtain the correspondence, we use a pair of bilingual dictionaries and introduce the three types of *translation circuits*. A translation circuit is a tuple which consists of four elements, a headword of each language and one of the word senses of these headwords, where the word sense of each language should have the headword of the other language as its translation word. We assume that the two word senses in a translation circuit represents the same meaning. According to the assumption, the pairs of word senses appearing in the translation circuits can be considered as conceptual items shared by two languages, while the word senses not appearing in any translation circuit can be considered as conceptual items which are unique to one of the languages.

The paper also describes the outline of a preliminary experiment conducted to verify this assumption. The results of the experiment are promising and some remarks are also given.

We conclude the paper with pointing out the possibility to construct the set of conceptual items which are shared by more than two languages, by extending our method.

2 Structure of Bilingual Dictionaries

2.1 Correspondence between word senses

This section gives the speculation on the structure of a bilingual dictionary. Consider two bilingual dictionaries, one from language L^a to language L^b ($D_{a \to b}$), the other of inverse direction ($D_{b \to a}$). With respect to $D_{a \to b}$, L^a is called the source language and L^b the target language. Suppose a headword a_i of $D_{a \to b}$ has m word senses ($a_i/1, \ldots a_i/m$), a headword b_j of $D_{b \to a}$ has n word senses ($b_j/1, \ldots b_j/n$), and $a_i/2$ has the translation word b_j, then we can have two types of edges, as illustrated in Figure 1. One goes from a headword to word sense(s) (type I), and the other goes from a word sense of the source language L^a to headword(s) of the target language L^b, representing the translation (type II). Note that there may be multiple edges that share the starting vertex. In such situations, type I edges denote that a word can have several meanings, and type II edges denote that a word sense can be translated to several words of the target language.

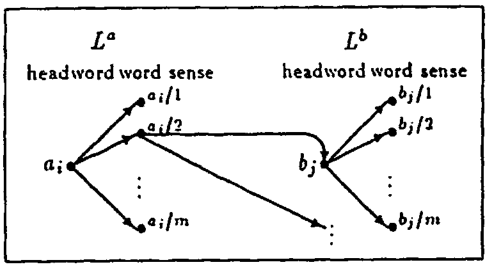

Figure 1 Structure of bilingual dictionaries

Figure 1 illustrates that the translation of $a_i/2$, which is a word sense of the headword a_i in $D_{a \to b}$, is b_j. The headword b_j has n word senses, however, it is not clear which of them has the same meaning as $a_i/2$. This is because bilingual dictionaries contain only the mapping relation from the word senses of the source language to the (head)word of the target language and not the correspondence between the word senses of each language. To extract this correspondence is the central issue of this paper.

2.2 Translation circuit

We regard a bilingual dictionary as a directed graph and call it a *translation graph (TG)* (See Figure 1). A translation graph from L^a to L^b is denoted by TG_{ab}, where the order of the subscription indicates the direction of the translation. TG_{ab} consists of a tuple $\langle H_a, S_a, T_b, E_{ab} \rangle$, where H_a, S_a and T_b are sets of vertices and E_{ab} is a set of edges. $H_a \subset L^a$ is a set of headwords of TG_{ab}, S_a is a set of word senses of H_a, and $T_b \subset L^b$ is a set of translation words of TG_{ab}. Some translation words of S_a may not appear in the headwords of TG_{ba}.

In case of a human translator to use TG_{ab}, first selection is done for one of the word senses of the headword to translate (e.g. a_i/k of the headword a_i), followed by selecting a translation (e.g. b_j). This process corresponds to the path $a_i \to a_i/k \to b_j$ and is called the *translation path*.

Next, we introduce a *bidirectional translation graph* TG_{ab+ba}, which is defined as a union of two translation graphs, TG_{ab} and TG_{ba}. In TG_{ab+ba}, a *translation circuit* is defined as a circuit which passes through $h_a(\in H_a) \to s_a(\in S_a) \to h_b(\in H_b) \to s_b(\in S_b) \to h_a$.

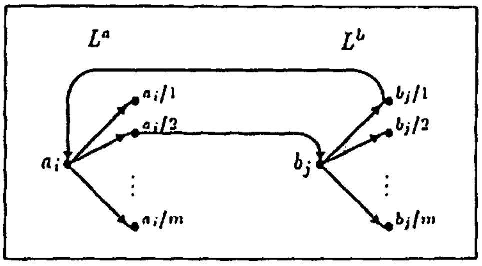

Figure 2 Example of type A translation circuit

Figure 2 is an example of a translation circuit that includes a_i, $a_i/2$, b_j, $b_j/1$. A translation circuit is denoted by a pair of word senses which are included in the circuit. For example, the translation circuit in Figure 2 is denoted by $\langle a_i/2, b_j/1 \rangle$.

For a pair of headwords, one from each language, we define three types of translation circuits which includes these headwords.

Definition: Type A translation circuit
Consider two headwords, $h_a \in H_a$ and $h_b \in H_b$. If there exists a unique circuit which includes h_a and h_b, then this circuit is *type A translation circuit*.

For example, in Figure 2 there exists a unique circuit which includes the headwords a_i and b_j, and the word senses $a_i/2$ and $b_j/1$. $\langle a_i/2, b_j/1 \rangle$ is a type A translation circuit by definition.

Definition: Type B translation circuit
Consider two headwords, $h_a \in H_a$ and $h_b \in H_b$. If there exist more than one circuits which include h_a and h_b, and all these circuits have a unique word sense of either language, then these circuits are *type B translation circuits*.

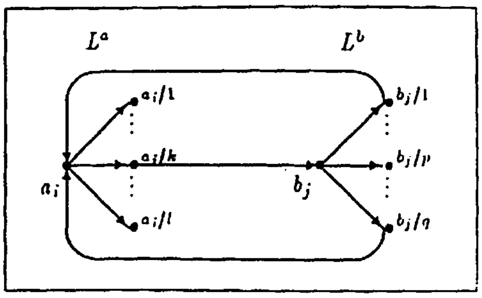

Figure 3 Example of type B translation circuit

Figure 3 shows an example of type B translation circuit. With respect to headwords a_i and b_j, there are two circuits, $\langle a_i/k, b_j/1\rangle$ and $\langle a_i/k, b_j/q\rangle$, and both include only one word sense of L^a, that is a_i/k. By definition, $\langle a_i/k, b_j/1\rangle$ and $\langle a_i/k, b_j/q\rangle$ are both type B translation circuits.

Definition: Type C translation circuit
Consider the two headwords, $h_a \in H_a$ and $h_b \in H_b$. If there exist more than one circuits which include h_a and h_b and they are not type B translation circuits, then they are all *type C translation circuits*.

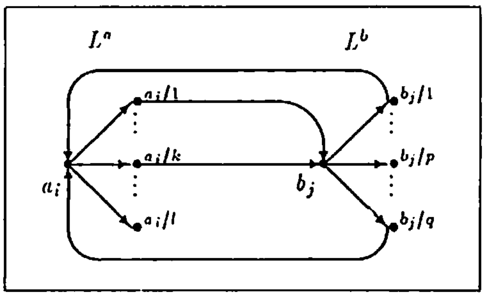

Figure 4 Example of type C translation circuit

Figure 4 illustrates an example of type C translation circuit. With respect to headwords a_i and b_j, there are four type C translation circuits, $\langle a_i/1, b_j/1\rangle$, $\langle a_i/1, b_j/q\rangle$, $\langle a_i/k, b_j/1\rangle$, and $\langle a_i/k, b_j/q\rangle$.

We call the pair of word senses, one from each language, which have the same meaning as *word sense pair*. A word sense pair can be considered as a conceptual item which is shared by two languages. As mentioned above, we assume that the two word senses included in a translation circuit have the same meaning. In other words, a translation circuit gives a word sense pair. However, it is obvious that some translation circuits do not give word sense pairs. The reasons why not all the translation circuits give the word sense pairs are given in the section 4. In order to check the validity of the above assumption we conducted an experiment, which is described in the next section.

3 Preliminary Experiment

This section describes a preliminary experiment conducted to verify our method. We used *Lighthouse English-Japanese dictionary* [5] and *Lighthouse Japanese-English dictionary* [6] for the following reasons:

- Both dictionaries are compiled by the same editors and on the same principle;
- Word definitions are clear, and the structure of the dictionary entry is convenient for computer processing.

3.1 Definition and representation of word sense

In the dictionaries, the word definition has the following hierarchical structure:

- In the English-Japanese dictionary, the words which has the same spelling but has the different meaning have different entries. For example, "*bank*$_1$" (a business ganization) and "*bank*$_2$ (land along the side of a riv have the different entries. While in the Japanese-Engli dictionary, some headwords (written in *Hiragana*) ha several Kanji notations even if each of these Kanji no tions represents the different meaning.
- In both the dictionaries, entries have numbered sube tries and each subentry is separated by ';' (semicolc and they are further separated by ',' (comma).

Thus commas are weaker separator than semicolons. T items separated by commas are interchangeable as the edit has declared. We therefore used semicolons as a word sei separator. We call the number of subentries *major num* and the displacement (base is 0) of each word sense in t subentry *minor number* of the word sense.

In both the dictionaries, optional words are enclosed by)' and alternative words are enclosed by '[ ]' in the trans tion words. As for the optional words, we use both the tra lation words with/without the optional words in both the d tionaries. As for the alternative words, we use all the tra lation words with each alternative words in the Japane English dictionary. In the English-Japanese dictionary, t alternative words are just deleted, because we can not id tify which portion of the translation words should be replac by the alternatives without segmentation of the translati word.

Some word senses have phrases, clauses or words wi comments as translation words. Obviously, such translati words do not appear as headwords in the dictionary of verse direction. Comments which is attached to translati words are deleted, because these comments can be regard as a explanation of the word sense.

A word sense is represented by a pair consisting of hea words and their word sense numbers. A word sense num includes a part of speech, a major number and a minor nu ber. A major number is the number of the subentry, a a minor number the displacement of each word sense in t subentry.

Table 1 Symbols of parts of speech

	English-Japanese	Japanese-English
noun	n	n, c, u
verb	vi, vt	v, vi, vt
adjective	adj	adj
adverb	adv	adv
unknown	—	x

c: countable noun
u: uncountable noun
vi: intransitive verb
vt: transitive verb

In the experiment, we take into account only the wo whose part of speech is either noun, verb, adjective, or a verb. We assign the symbols shown in Table 1 to each part speech. The classification of part of speech is slightly diff ent between the dictionaries. Furthermore the part of spee of some word sense can not be identified automatically Japanese-English dictionary. Symbol 'x' in the Table 1 re resents this case.

3.2 Procedure of the experiment

Unfortunately, *Lighthouse dictionaries* are not yet in machine readable form. The procedure must start by making them machine readable, but due to financial constraints it is impossible to enter the entire dictionaries into the computer. We selected words of each languages from six essays in *The Last Badger* [7] by C. D. Lummis, which has Japanese translation, and manually entered dictionary entries of these words into the computer. These essays were chosen because each contains about 300 words and is written in simple English. Thus, 562 English and 779 Japanese entries were rendered machine readable.

Table 2 Statistics of the Englith-Japanese dictionary

The number of headwords	:	562
The number of translation paths	:	7824

Occurence of each parts of speech

n	vi	vt	adj	adv
3145	949	1993	1061	680

The number of word sense for each headword

Word sense	1	2	3
Occurence	60	72	54

4	5	6	7	8	9
60	39	31	32	32	24

10	11	12	...	avg
18	12	12	...	8.33

The number of translation words for each word sense

Translation	1	2	3	4	...	avg
Occurence	2243	1844	496	87	...	1.67

Table 3 Statistics of the Japanese-Englith dictionary

The number of headwords	:	779
The number of translation paths	:	4995

Occurence of each parts of speech

n	c	u	vi	vt	adj	adv	x
487	594	159	378	952	471	232	1082

The number of word sense for each headword

Word sense	1	2	3
Occurence	121	149	102

4	5	6	7	8	9
92	80	54	37	28	25

10	11	12	...	avg
12	9	13	...	5.16

The number of translation words for each word sense

Translations	1	2	3	4	...	avg
Occurence	3266	623	80	43	...	1.24

The number of words for each translation word

Words	1	2	3	4	5	6	7	avg
Occurence	3302	347	448	270	107	17	4	1.62

Next, the tuples consisting of a headword, a word sense and a translation word were automatically extracted using a simple program written in GNU Emacs Lisp. These tuples correspond to the translation paths described in section 2. 7,824 translation paths of English-Japanese and 4,996 of Japanese-English were extracted. Table 2 and Table 3 show some statistics of each dictionary.

Finally, the number of translation circuits were counted on a Prolog interpreter, in which a translation path is represented by a unit clause. All these translation circuits are checked if they give word sense pairs by hand. The results are shown in Table 4.

Table 4 The number of translation circuits (t/c) and word sense (w/s) pairs

Type	w/s pair	t/c	Ratio
A	237	244	97%
B	65	113	58%
C	18	82	22%

4 Results and Discussion

4.1 Dictionaries

Table 2 and Table 3 show that the English-Japanese dictionary has more word senses and translation words than the Japanese-English dictionary. This may be because both the dictionaries are compiled for Japanese readers. Using a Japanese-English dictionary compiled for English readers may give the different results.

In the experiment, when a translation word consists of multiple words, such as phrases or clauses, no translation circuit is extracted, because the dictionary does not have multiple words as its headword. However, as Table 3 shows, since 73% of the translation words in the Japanese-English dictionary consist of a single word, it is enough as a first order approximation to take into account only the translation words consisting of a single word.

Although the translation words in the English-Japanese dictionary are not segmented into words, we do not divide them into words. The reasons are, generally it is difficult to define a "word" in Japanese, and we do not take into account the translation circuits which include the headword consisting of multiple words in the experiment. Thus, the number of words for each translation word is missing in Table 2.

4.2 Type A translation circuit

As Table 4 shows, 244 type A translation circuits were extracted. 7 translation circuits out of them did not give a word sense pair, the rest did. Thus we have obtained word sense pairs from type A translation circuit with 93% correctness.

The main reason for the errors are:

- Some word senses lack of proper translation words,
- Some headwords in Japanese-English dictionary are not assigned a parts of speech (symbol 'x' in Table 1).

For example, English word "hostess" has the following three word senses:

(1) a woman who receives guests and provides food, drink, etc. for them.

(2) airhostess

(3) a young woman who acts as a companion, dancing partner, etc. in a social club.

And Japanese word "HOSUTESU"[1] has the following two word senses:

(i) a woman who receives guests and provides food, drink, etc. for them.

(ii) a young woman who acts as a companion, dancing partner, etc. in a social club.

Although we obtained a word sense pair $\langle(3),(i)\rangle$, the correct correspondences are $\langle(1),(i)\rangle$ and $\langle(3),(ii)\rangle$. This error occurred because word sense (1) does not have "HOSUTESU" as a translation word but word sense (3) does, and word sense (i) has "hostess" as a translation word but (ii) does not. Word sense (ii) has word "barmaid" as a translation word. This is the former case of the error.

As for the latter case, a translation circuit which includes "think (vi)" and "KANGAERU (vt)" does not give a word sense pair. This is because the part of speech of "KANGAERU" could not be identified by the analysis (denoted by symbol "x" as described in section 3), and results the paring of inconsistent parts of speech.

4.3 Type B translation circuit

113 type B translation circuits are extracted, and 65 of them (58%) gave word sense pairs and 48 of them (42%) did not. Table 5 shows how many word senses each single word sense has correspondence with. As for type B translation circuits, since only the half of them give word sense pairs, human assistance is required to extract word sense pairs. However, as shown in Table 5, almost all the relation between word senses of each headword of each language are 1 to 2 or 1 to 3. it is possible to select the word sense pairs out of them without much labor, if an appropriate man-machine interface is provided.

Table 5 Type B word sense pairs

Combination	1 to 2	1 to 3	1 to 4	1 to 5	1 to 8
Occurence	72	15	8	10	8

There are the cases in which a word sense of the dictionary includes the multiple word senses of the other dictionary. This is the main reason to give the type B translation circuits. In particular, a headword which has many word senses tends to have a general meaning at the beginning of the word definition, and which subsumes the rest of the word senses. Figure 5 shows the type B translation circuits extracted from the headwords "write" and "KAKU". In this example, the word sense (1) of "write" represents the general meaning of "write" and it includes the rest of the word senses (2) through (5). On the other hand, the word sense (i) of "KAKU" represents the general meaning of "KAKU" and has correspondence with the word sense (1) of "write" and with (2) through (5) as well.

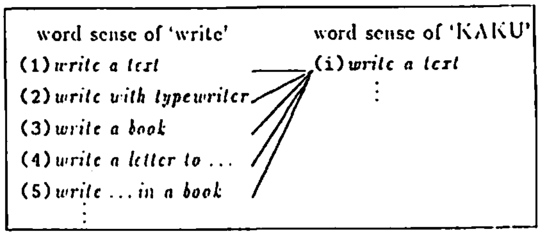

Figure 5 Example of extractec word sense pairs

4.4 Type C translation circuit

82 type C translation circuits were extracted and 18 of them (22%) gave word sense pairs and 64 of them (78%) did not. In case of type C, the correspondence between word senses are many to many with respect to a certain pair of headwords. Table 6 shows the combinations of the number of word senses and its occurrence. Since only about the one fifth of translation circuits give word sense pairs, human assistance is required to extract the word sense pairs just same as the case of type B translation circuit. But this case requires more manual labor than the case of type B, due to the more number of the combinations.

Table 6 Type C word sense pairs

Combination	2 to 2	2 to 3	2 to 6	4 to 8
Occurence	8	30	12	32

4.5 Symmetry between Dictionaries

There are many translation paths that do not constitute a circuit. As mentioned earlier, the primary reason is that the experiment used only a part of the dictionaries. Besides, there are also other reasons in such kind of experiment as Byrd pointed out [8], and the considerable fact that in the English-Japanese dictionary, a word not translated to a word but rather to phrases or clauses. This tendency is prominent in verbs, where some English verbs are translated to Japanese verbs with their objects of action. In order to handle such cases, Japanese translation should be segmented into words and some semantic structure should be constructed first. With that preparation, there is a possibility to make correspondence between a word sense and the semantic structure. Such cases are left for future works.

Throughout this paper, we have only considered bidirectional word sense correspondence. There may be an approach claiming that if a word w_a can be translated to a word w_b, then w_b can be necessarily translated to w_a. In some cases, however, the inverse translation requires strong contextual support, and since our aim is to make correspondence between word sense of each language, we would do well to avoid such contextual dependency.

5 Concluding Remarks

This paper has proposed a method of extracting the information of conceptual dictionary from a pair of machine readable dictionaries. In particular, we have focused on extracting the correspondence between the word senses of each language. The outline of the procedure is as follows:

[1] Capitalized words denote Japanese

- modeling a pair of bilingual dictionaries as *translation graph*,

- extracting three types of *translation circuit* (type A, B, C),

- decide the *word sense pair* based on the translation circuit extracted in the previous step.

We conducted a preliminary experiment with an English-Japanese and a Japanese-English dictionary to verify the method. With about 600 headwords of each dictionary, we could extract 244 type A, 113 type B, and 82 type C translation circuits. Type A translation circuits give word sense pairs with 97% correctness, but type B and type C translation circuits give word sense pairs with only 58% and 22% correctness respectively. Thus some human assistance is required to obtain word sense pairs from type B and type C translation circuits.

Since we have used only a part of the dictionary in the experiment, it is very difficult to estimate how many word sense pairs can be extracted from the entire dictionaries. Some translation circuits may not be constructed because the data we entered into the computer lack the entries which match the translation words. Actually, only 1022 translation words out of 3302 in the Japanese-English data, which consist of a single English word, are contained in the English-Japanese data as headwords.

We obtained about 240 word sense pairs out of about 600 headwords of each dictionary automatically, that is 40% of the headwords. With a simple extrapolation of the rate, we will be able to obtain about 24,000 word senses out of medium size dictionaries, which have 60,000 headwords. The word sense pairs which can be extracted from type B and type C translation circuits with human assistance will increase the figure. This is quite a promising result as a first step, although we are forced to reexamine the translation paths that did not form circuits. We plan to repeat the experiment using a greater number of words.

In this paper, we have considered word sense pairs only between two languages. We conclude the paper with pointing out the possibility of applying this method to more than two languages. The outline of the procedure is as follows, where D_{x-y} denotes a bilingual dictionary from a language L^x to a language L^y. Consider $n+1$ languages L^c, $L^1, \dots L^n$.

(1) Extract word sense pairs between L^c and L^1 using D_{c-1} and D_{1-c}.

(2) Assign words of each language $L^2, \dots L^n$ to each word sense of L^c extracted in step (1).

(3) Extract $n-1$ sets of word sense pairs from L^c and $L^2, \dots L^n$ using words assigned in step (2) and D_{i-c}.

(4) Divide or integrate word senses of L^c.

Step (1) and (3) can be automated, but (2) and (4) require human assistance. The set of word senses of L^c may be a basic data to construct the interlingual concept items.

References

[1] S. Nirenburg, V. Raskin, and A. B. Tucker. The structure of interlingua in TRANSLATOR. In S. Nirenburg, editor, *Machine translation – Theoretical and methodological issues –*, chapter 6, pages 90–113. Cambridge, 1987.

[2] EDR. Concept dictionary. Technical Report TR-009, Japan Electronic Dictionary Research Institute, 1988.

[3] S. Nirenburg and V. Raskin. The subworld concept lexicon and the lexicon management system. *Computational Linguistics*, 13(3-4):276–289, 1989.

[4] M. Kiyono. A method for stabilizing word meaning in the concept dictionary. In *the 3rd Annual Conference of JSAI*, pages 383–386, 1989. (in Japanese).

[5] S. Takebayashi and Y. Kojima, editors. *Lighthouse English-Japanese dictionary.* Kenkyusya, 1984.

[6] Y. Kojima and S. Takebayashi, editors. *Lighthouse Japanese-English dictionary.* Kenkyusya, 1984.

[7] C. D. Lummis. *The Last Badger.* Syobunsya, 1988.

[8] R. J. Byrd. N. Calzolari, M. S. Chodorow, and M. S. Klavans, J. L. Neff. Tools and methods for computational lexicology. *Computational Linguistics*, 13(3-4):219–240, 1987.