Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation

Jun-ichi NAKAMURA, Makoto NAGAO
Department of Electrical Engineering, Kyoto University, Yoshida-honnachi, Sakyo, Kyoto, 606, JAPAN

Abstract

The automatic extraction of semantic information, especially semantic relationships between words, from an ordinary English dictionary is described. For the extraction, the magnetic tape version of LDOCE (Longman Dictionary of Contemporary English, 1978 edition) is loaded into a relational database system. Developed extraction programs analyze a definition sentence in LDOCE with a pattern matching based algorithm. Since this algorithm is not perfect, the result of the extraction has been compared with semantic information (semantic markers) which the magnetic tape version of LDOCE contains. The result of comparison is also discussed for evaluating the reliability of such an automatic extraction.

1 Introduction

A large dictionary database is an important component of a natural language processing system. We already know syntactic information which should be and can be stored in a large dictionary database for a practical application such as a machine translation system. However, we still need more research on semantic information which can be prepared for a large system. As a first step to construct a large scale semantic dictionary (lexical knowledge base) the authors of this paper have inspected a machine readable ordinary English dictionary LDOCE, Longman Dictionary of Contemporary English, 1978 edition [Proctor 1987]. Extracting semantic information from an ordinary dictionary is an interesting research topic. One of the aims of automatic extraction is to produce a thesaurus. Noël, for example, proposed the idea of thesaurus production from LDOCE in [Noël 1982]. Amsler also showed the result of automatic thesaurus production from a technical encyclopedia [Amsler 1987]. Boguraev and Alshawi have studied the utilization of LDOCE for natural language processing research in general [Alshawi 1987, Boguraev 1987].

In this paper, the automatic extraction of semantic relationships between words from LDOCE is described. For the extraction, the magnetic tape version of LDOCE is loaded into relational database system. Developed extraction programs analyze the definition sentence in LDOCE with a pattern matching based algorithm. Since this algorithm is not perfect, the result of the extraction has been compared with semantic information (semantic markers) which the magnetic tape version of LDOCE contains. The result of comparison is also discussed for evaluating the reliability of such an automatic extraction.

2 RDB Version of LDOCE

In general, a dictionary consists of a complex data structure: various relationships between words; grammatical information; usage notes, etc. Therefore, we need a special database management system to handle dictionary data. For instance, [Nagao 1980] shows such a system for retrieving a Japanese dictionary. In this paper, however, the authors are mainly interested in the definition and the sample sentence parts of LDOCE, instead of complex relations among information in the dictionary.

For the sake of efficiency (including the cost of system development) of LDOCE retrieval, we have decided to use a conventional relational database management system (RDBM). The RDBM which we use is running on the mainframe computer of Kyoto University Data Processing Center (Fujitsu M783, OS/IV F4 MSP, FACOM AIM/RDB).

For loading the magnetic version of LDOCE into this RDBM, we have extracted the following fields from LDOCE:

1. Head Word (HW); 2. Part-of-Speech (PS); 3. Definition Number (DN); 4. Grammar Code (GC); 5. Box Code (BC); 6. Definition (DF); and 7. Sample Sentence (SP).

The Box Code field contains various information such as semantic restrictions, etc, which are explained in section 4.1.

The fields 1 through 5 are almost the same as the original LDOCE data. (Several special characters are removed or changed into standard characters for simplicity of retrieval. The syllable division mark (.) is removed. Some of the font control characters are changed into ‘<’ and ‘>’.)

The definitions and the sample sentences are separated into a clause or a sentence. For example, definition 1 of the verb to abandon is:

1. to leave completely and for ever; desert

in the original data. This definition is transformed into two separate clauses in the RDB version:

1. to leave completely and for ever
2. desert.

Since every data in the RDB is represented in a tabular form, we have made three tables for the RDB version of LDOCE (LDOCE/RDB, see table 1 regarding the its record format):

459
3 Extraction of Semantic Information

One form of semantic information useful for natural language processing is a thesaurus (or semantic network), which basically describes semantic relations between words. To automatically produce the thesaurus from LDOCE, two programs have been developed:

1. Key Verb extraction program.
2. Key Noun and Function Noun extraction program.

These programs and the result of extraction are discussed in this section.

3.1 Key Verb Extraction Program

Most of the definitions of verbs in LDOCE are described as:

to VERB ... 

Usually VERB in this pattern expresses a 'key concept' of the defined verb. Therefore, we call this VERB a Key Verb.

For example, the verbs semantically related to the verb to hit have the following definitions:

- strike: to hit

| KW | PS | DN | DF |
|----|----|----|----|
| hit | v  | 1  | to leave completely and for ever |
| hit | v  | 2  | to leave (a relation or friend) in a thoughtless or cruel way |
| hit | v  | 3  | to give up, esp. without finishing |
| hit | v  | 4  | to give (oneself) up completely to a feeling, desire, etc. |
| hit | n  | 0  | the state when one's feelings and actions are uncontrolled |
| hit | adj | 0  | given up to a life that is thought to be immoral see also ABANDON (2,4) |

- beat: to hit many times, esp. with a stick
- kick: to hit with the foot
- knee: to hit with the knee

From this pattern of definitions, we can draw figure 1 which shows the semantic hierarchy around to hit: to beat, to kick and to knee are specialized verbs of to hit.

To expand this hierarchy, a program to extract the key verbs from a definition is developed. Table 3 (LDBV.D2) shows some examples of this extraction. In Table 4, the frequency of key verbs is listed. Most frequently used key verb is to make. Note that to make and to cause are used to define causative and transitive verbs respectively.

Table 1: Record Format and Size of LDOCE/RDB

| Column Name | HW | PS | DN | GC | BC |
|-------------|----|----|----|----|----|
| Head Word   | char(20) | char(10) | char(10) | char(36) | char(14) |
| Part of Speech | char(20) | char(10) | char(10) | char(36) | char(14) |
| Definition Number | char(20) | char(10) | char(10) | char(36) | char(14) |
| Grammar Code | I1HW | I1PS | I1DN | I1GC | I1BC |
| Box Code | BC |

Table 2: Definition Table (LDB.D2) of LDOCE/RDB

| HW | PS | DN | DF |
|----|----|----|----|
| abandon v | 1 | to leave completely and for ever |
| abandon v | 2 | to leave (a relation or friend) in a thoughtless or cruel way |
| abandon v | 3 | to give up, esp. without finishing |
| abandon v | 4 | to give (oneself) up completely to a feeling, desire, etc. |
| abandon n | 0 | the state when one's feelings and actions are uncontrolled |
| abandon adj | 0 | given up to a life that is thought to be immoral see also ABANDON (2,4) |

Table 3: Definition and Key Verb Table (LDBV.D2, part)

| HW | KW | PS | DN | DF |
|----|----|----|----|----|
| abase | make v | 0 | to make (someone, esp. oneself) have less self-respect |
| abase | make v | 0 | make humble |
| abash cause v | 0 | to cause to feel uncomfortable or ashamed in the presence of others |
| abate | become v | 1 | (of winds, storms, disease, pain, etc.) to become less strong |
| abate | decrease v | 1 | decrease |
| abate | make v | 2 | <hit> to make less |
| abate | bring v | 3 | <law> to bring to an end (esp. in the phr. <abate a nuisance>) |
Table 4: Frequency of Key Verbs

| KV     | COUNT(KV) |
|--------|-----------|
| make   | 1311      |
| be     | 875       |
| cause  | 641       |
| give   | 505       |
| put    | 446       |
| take   | 338       |
| move   | 383       |
| have   | 374       |
| become | 336       |
| go     | 263       |
| get    | 208       |

Traversing these relations between defined verb and key verb, a thesaurus (network) of verbs has been obtained approximately. Most of the verbs in this thesaurus make a tree-like structure shown in figure 1. However, several 'loops' are found. A 'loop' expresses a cyclic definition: to welcome is defined by to greet, and to greet is defined by to welcome. In the network, six typical cyclic definitions are:

- do: do (the verb to do does not have a key verb.)
- change: change, move, come, become
- go: go, leave
- get: get, receive
- stop: stop, cease
- let: let, allow, permit

Note that there are many other cyclic definitions in the network. However, most of them have a link to another verb; at least one of the verb in a cyclic definitions is defined by another verb.

Since no reader of LDOCE can understand the meaning of these verbs only from the dictionary, these may be a kind of bug of the dictionary. However, these cyclically defined verbs seem to correspond to semantic primitives, which are first introduced to AI works by [Schank 1975]. Semantic primitives may be defined outside of linguistic words. Details of the result of extraction are discussed in [Nakamura 1986].

3.2 Key Noun and Function Noun Extraction Program

We can apply a similar algorithm to definitions of nouns, although the pattern of definitions of nouns is more complex than that of verbs. Inspecting definitions with LDOCE/RDB, most of them are classified into two forms:

1. {determiner} {adjective}* Key Noun {adjective phrase}*
2. {determiner} {adjective}* Function Noun of Key Noun {adjective phrase}*

The first one is a simple form and many of them express is-a relations between a defined noun and a key noun. For example, abandon: the state when one's feelings and actions are uncontrolled shows that

abandon in-a state.

The second form expresses more complex semantic relations between nouns.

abey: the group of people living in such a building shows that

abey is-a-group-of people.

A function noun, therefore, explicitly expresses the semantic relation between a head word and a key noun.

With terms of a semantic network, defined nouns and key nouns are nodes in a semantic network, and function nouns (when function noun is empty, its function noun is regarded as kind) express the name of a link between nodes. The following nouns (41 nouns, in total) are considered to be function nouns, which are manually extracted.

- is-a: kind, type, ...
- part-of: part, side, top, ...
- member-ship: part, member, group, class, family, ...
- action: act, way, action, ...
- state: state, condition, ...
- amount: amount, sum, measure, ...
- degree: degree, quality, ...
- form: form, shape, ...

A program to extract key nouns and function nouns from the definitions of nouns is developed. Table 5 shows a part of the key noun and function noun table in the LDOCE/RDB (LDBN.D2) generated by this program.

As shown in table 6, the key noun of highest frequency is person (2174 times) and for function noun is type (1064 times) except null function noun (pattern 1).

Traversing is-a relation, for example, a thesaurus has been obtained [Nakamura 1987]. Table 7 shows a part of the automatically obtained thesaurus, whose 'root' word is person: actor is a-kind-of person; comedian, extra, ham, and mime are a-kind-of actor; comedienne is a-kind-of comedian.

4 Comparison between Result of Extraction and BOX Code

The thesaurus produced from LDOCE by the key noun and key verb extraction programs is an approximate one, and, obviously, contains several errors. The key noun of abbreviation 1, for example, is shorter in table 5, because the current program ignores formed words. However, it should be making. (Even if we changed the extraction algorithm, still we have a problem that making is not a simple noun, but a gerund. We need to define noun-verb semantic relations.) To evaluate the quality of the produced thesaurus, the noun part of the thesaurus has been compared with the semantic markers in LDOCE.
Table 5: Definition, Key Noun and Function Noun Table (LDBN.D2, part)

| HW | DN | KN | FN | DF |
|----|----|----|----|----|
| abandon | 0 | state | | the state when one's feelings and actions are uncontrolled |
| abandon | 0 | freedom | | freedom from control |
| abbey | 1 | building | | (esp. formerly) a building in which Christian men (monk <s>) or women (nun <s>) live shut away from other people and work as a group for God |
| abbey | 1 | convent | | monastery > or convent |
| abbey | 2 | people group | | the group of people living in such a building |
| abbey | 3 | house | | a large church or house that was once such a building |
| abbreviation | 1 | shorter act | | the act of making shorter |
| abbreviation | 2 | word form | | a shortened form of a word, often one used in writing |

Table 6: Frequency of Key Nouns and Function Nouns

| KN | COUNT(KN) | FN | COUNT(FN) |
|----|-----------|----|-----------|
| person | 2174 | 36583 |
| to | 1660 | 1064 |
| something | 668 | 838 |
| act | 655 | 603 |
| piece | 479 | 557 |
| state | 294 | 498 |
| part | 261 | 327 |
| group | 255 | 306 |
| any | 253 | 247 |
| quality | 232 | 246 |
| types | 226 | 208 |
| set | 206 | 200 |
| action | 205 | 182 |

Table 7: Example of Thesaurus (person)

| HW | DN | DF |
|----|----|----|
| person | 0 | a person whose job is to keep and examine the money accounts of businesses |
| accountant | 0 | certified public accountant |
| CPA | 0 | infnl a person of the highest class or skill in something |
| ace | 2 | a person who takes part in something that happens |
| actor | 2 | an actor who a tells jokes or does amusing things to make people laugh |
| comedian | 1 | an actor who a tells jokes or does amusing things to make people laugh |
| comedienne | 0 | a female comedian (1) |
| extra | 2 | an actor in a cinema film who has a very small part in a crowd scene and is |
| sundry | 0 | extra (4) |
| ham | 3 | an actor whose acting is unnatural, esp. with improbable movements and expressions |
| mine | 3 | an actor who performs without using words |

Figure 2: Hierarchy of Semantic Markers in LDOCE

4.1 Semantic Markers in LDOCE: BOX Code

The magnetic version of LDOCE has a special field related to semantic markers, which is called as BOX code fields, although it does not appear in the printed version of LDOCE. Some of the BOX code fields (called BOX1, for instance) express semantic restrictions for a noun governed by a verb or an adjective, and a semantic classification of a noun. For example, the semantic restriction for a subject of the verb to travel is marked as 'Human'; the noun person is classified as 'H.' This shows that the verb to travel may govern the noun person in its subject position. The LDOCE uses 34 markers for expressing this restriction (Table 8).

These semantic markers have a hierarchy as shown in figure 2. For example, 'Human', 'Plant', and 'Animal' are subclasses of 'animate (Q).'

In the following part of this section, the comparison between semantic markers of LDOCE and the thesaurus constructed from the definitions of nouns in LDOCE is discussed from the view

Table 8: Semantic Markers in Box Code of Nouns and their Frequency (Part)

| type of code | box1 | DN=0,1 |
|--------------|------|--------|
| A Animal     | 957  | 386    |
| B Female Animal | 26 | 15     |
| C Concrete   | 350  | 181    |
| D Male Animal | 27  | 21     |
| E 'B' + 'L'  | 257  | 187    |
| F Female Human | 413 | 314    |
| G Gas        | 111  | 79     |
| H Human      | 3457 | 2426   |
| I Inanimate  | 42   | 26     |
| J Movable    | 5794 | 3927   |
| K Male ('D' + 'M') | 2 | 2 |
4.2 Nouns Marked as 'Animate'

Nouns related to the concept animate have a relatively simple structure in the thesaurus, as animate is often used as an example of a thesaurus-like system. Examples of the words marked as 'animate (Q)' and related nouns, especially marked as 'plant + animal (V)', are shown in table 9.

The produced thesaurus contains more than 60% of the words marked as simple concepts, such as 'plant' (table 10), 'animal', and 'human (person in definitions)', in correct positions. As shown in table 10, for example, 645 words are traversed from plant in the produced thesaurus; 370 words (62.4%) of these words are marked as 'Plant.'

However, the produced thesaurus does not capture disjunctive concepts such as 'animal or plant (V)' correctly. In the definition of crossbred (table 9), the produced thesaurus only uses plant as a key noun, and ignores animal. This is a typical problem in the current produced thesaurus.

Note that the distinction between 'animate (Q)' and 'animal or plant (V)' (animate without human) seems to be difficult for the lexicographers: brized is marked as Q; crossbred, however, is marked as V, for example.

4.3 Nouns Marked as 'Abstract'

In LDOCE, many nouns (about 46%, table 8) are marked as 'abstract', and they are not classified into more detailed subclasses. On the other hand, function nouns work as a key for sub-classification in the produced thesaurus. In section 3.2, some of the function nouns are listed as action, state, amount and degree. These function nouns classify abstract nouns.

For example, there are 597 nouns whose function noun is act, and 584 nouns (97%) of them are marked as 'abstract'; there are 398 nouns whose function noun is state, and 391 nouns (98%) of them are 'abstract.' The distinction between 'state' and 'act', for instance, is useful for natural language processing in general.

4.4 Nouns Marked as 'Inanimate'

Some 'inanimate' nouns are correctly identified in the produced thesaurus (table 11). Especially, 39% of nouns under the noun liquid have 'Liquid' markers, and 56% of nouns under the noun gas have 'Gas' markers.

However, many 'inanimate' nouns are defined by substance in LDOCE. Sub-classification of these noun is expressed with a compound word (or an adjective) as shown in table 11: coke is a solid substance; fluorine is a non-metallic substance. Since the current extraction program does not handle a compound word, the thesaurus cannot express these classification.

4.5 Other Typical Nouns

Several typical nouns in the produced thesaurus are also compared with markers of LDOCE. Because the current system cannot distinguish senses of nouns, nouns which have several different senses causes a problem. A typical example is found in the definitions whose key noun is case. As shown in table 12, attach case and test case are both defined by case; these express completely different concept. In 30 nouns whose key noun is case,
Table 12: Nouns whose key noun is case

| HW       | B1 | KN  | DF                                      |
|----------|----|-----|-----------------------------------------|
| attaché case | J  | case | a thin hard case with a handle, for carrying papers |
| test case  | T  | case | a case in a court of law which establishes a particular principle and is then as a standard against which other cases can be judged |

Table 13: Nouns related the noun cloth

| HW      | B1 | FF | DF                                      |
|---------|----|----|-----------------------------------------|
| canvas  | J  |     | strong rough cloth used for tent, sails, bags, etc. |
| denim   | S  |     | a strong cotton cloth used esp. for jeans |
| serge   | J  | type | a type of strong cloth, usu. woven from wool, and used esp. for suits, coats, and dresses |
| tweed   | S  | type | a type of coarse woolen cloth woven from threads of several different colours |

16 nouns are 'movable (J)', and 14 nouns are 'abstract.'

Difficulty of semantic marking is also found. For example, lexicographers could not mark 'movable (J)' and 'Solid' systematically. For example, some nouns whose key noun is cloth are marked as 'Solid', and others are marked as 'movable (J)' (table 13). This is a problem in gathering of semantic information itself.

5 Conclusion

The extraction of semantic relations between verbs and nouns from LDOCE is discussed. Data from the magnetic version of LDOCE is first loaded into a relational database system for simplicity of retrieving. For the extraction of semantic relations, programs to find key verb, key noun, and function noun have been developed. Using these programs, the thesaurus is automatically produced.

To evaluate the quality of the noun part of the produced thesaurus, it is compared with the semantic markers in LDOCE. Although the produced thesaurus has several problems such as the difficulty of expressing disjunctive concepts, the comparison between the produced thesaurus and semantic markers in LDOCE shows the possibility of sub-classification of 'abstract' nouns.

Acknowledgements

The authors grateful to Prof. Jun-ichi Tsujii for his fruitful comments on this work. We also wish to thank Mr. Motohiro Fujigaki, Mr. Nobuhiro Kato, and Mr. Keiichi Sakai who inspected LDOCE data carefully.

References

[Alshwai 1987] Alshwai, H., Processing Dictionary Definitions with Phrasal Pattern Hierarchies, *Computational Linguistics*, Vol. 13 (1987).

[Amsler 1987] Amsler, R. A., How Do I Turn This Book On?, *Proc. of Third Annual Conf. of the UW Centre for the NOED*, pp. 75-88 (1987).

[Boguraev 1987] Boguraev, B., Experiences with a Machine-Readable Dictionary, *Proc. of Third Annual Conf. of the UW Centre for the NOED*, pp. 37-50 (1987).

[Nagao 1980] Nagao, M., Tsujii, J., Ueda, Y., Takiyama, M., An Attempt to Computerized Dictionary Data Bases, *Proc. of COLING80*, pp. 534-542 (1980).

[Nakamura 1986] Nakamura, J., Fujigaki, M., Nagao, M., Longman Dictionary Database and Extraction of its Information, Report on Cognitive Approaches for Discourse Modeling, Kyoto University (1986) (in Japanese).

[Nakamura 1987] Nakamura, J., Sakai, K., Nagao, M., Automatic Analysis of Semantical Relation between English Nouns by an Ordinary English Dictionary, Institute of Electronics, Information and Communication Engineers of Japan, WGNLC, 86-23 (1987) (in Japanese).

[Noël 1982] Michels, A., Noël, J., Approaches to Thesaurus Production, *Proc. of COLING82*, pp. 227-232 (1982).

[Procter 1987] Procter, P., Longman Dictionary of Contemporary English Longman Group Limited, Harlow and London, England (1978).

[Schank 1975] Schank, R. C., Conceptual Information Processing, New York, North Holland (1975).