LEXICAL KNOWLEDGE FROM LARGE CORPORA AND ITS APPLICATION TO TEXT GENERATION

ISAHARA Hitoshi (Communications Research Laboratory)
INUI Hiroko (Institute of Behavioral Science)
UCHIDA Yuriko (Electrotechnical Laboratory)

Abstract

We provide an overview of the text generation module of the CONTRAST machine translation system, including its concept conversion modules, as well as its method of knowledge representation. We also point out problems for generating high-quality Japanese sentences from the intermediate representation, describe what we have gathered from corpora for compiling lexicon for text generation and provide an overview of the Real World Computing (RWC) Text Database Project, the results of which are being used to gather lexical information.

1. Introduction

To achieve high-quality translation between languages with extremely dissimilar sentence structures such as Japanese and English, we cannot rely solely on concepts in one language as the units of information to be processed. Rather, we must process the differences between conceptual systems of the languages. We have been engaged in researching text generation and concept conversion as well as the concept hierarchy used in those endeavors. The machine translation system, CONTRAST, utilizing techniques of concept conversion, was created by incorporating the results of that research. [1]

In this paper, we provide an overview of the text generation module of the CONTRAST machine translation system, including its concept conversion modules, as well as its method of knowledge representation. We also point out problems for generating high-quality Japanese sentences from the intermediate representation, and describe what we have gathered from corpora for compiling lexicon for text generation. In the appendix, we provide an overview of the Real World Computing (RWC) Text Database Project, the results of which are being used to gather lexical information.

2. Text Generation from Intermediate Representation

In this section, we briefly explain features of the generation module in CONTRAST and provide examples of rules obtained through an analysis of actual newspaper articles. We also describe our concept dictionary, which incorporates a concept hierarchy, and its application to concept conversion.

2.1. Text generation module and generation rules [2,3]

The English text generation system of CONTRAST begins with intermediate representation created by its

analysis module. Referring to the concept dictionary (which is shared with the analysis module), it creates English text according to generation rules which are created from knowledge extracted from the result of observation on real-life text data relating to English newspaper articles.

Production of a satisfactory text requires both knowledge of the language in which the text is written and knowledge of the organizational and writing style for the subject matter. Thus the text generation module has several characteristics:

(1) It makes no direct reference to the information of the original text and starts from a structured intermediate representation that is independent of source text language.

(2) Instead of referencing the original text, it generates text based on rules of generation derived from knowledge extracted from the result of observation on real-life text data.

(3) Not only information contained in the network structure of intermediate representation but also the network structure itself is instrumental in inter-paragraph construction.

(4) In addition to intermediate representation, structured background knowledge for the concept dictionary is used effectively for each stage of text generation serving as the basis of the writer's common-sense knowledge.

Because of these properties, the system can easily handle differences in inter-paragraph structure between Japanese and English articles. Moreover, it has an enhanced capability to generate information that does not appear in the original sentence but must be included in the translated text (such as articles). Fig. 1 shows examples of the generation rules.

2.2. Concept Dictionary and Concept Conversion [4, 5]

The concept dictionary contains all the general knowledge utilized by CONTRAST. Knowledge is represented as relationships between concepts. There are four methods for describing relationships between concepts: (1) hierarchical structure, (2) constraints on instances that can fill a slot, (3) subdivision by process models and part-whole relations, (4) explanations of concepts described as constraints. Figure 2 shows an example of an entry in the concept dictionary.

In (1), concepts are arranged in a hierarchical structure as upper and lower concepts. In (2), relationships between two instances, mainly case relations, are established through relators. The relators for slot names are also concepts, and can create instances. In (3), the process models subdivide event concepts into parts, and part-whole relations subdivide thing concepts into parts.

Regarding (4), the method explains one concept in terms of other concepts. Lines 5 to 7 in Fig. 2 show the constraint for *criminal, which means that an instance of *criminal (i.e. @criminal) is an instance of person (represented by "@@" in Fig. 2) who is the agent of some instance of *commit-a-crime (represented by "@1" in Fig. 2). In this paper, the name of the concept is preceded by a "*", e.g. *kidnap, and the instance of the concept by "@", e.g. @kidnap.

The concept conversion module uses this information to alter the structure in the intermediate representation, i.e., paraphrase. This makes it possible to identify objects referred to using different forms of expression, and to use appropriate expressions at the time of language generation.

[Rules Determining Internal Structure of a Paragraph]

A number of scene nodes emanating from a single scene node are generated within a paragraph. The scenes are generated on the basis of the temporal sequence of events that took place. Event type nodes outside the scene are generated based upon causal relationships that are indicated by intermediate representation.

[Rules of Selection with Subject Candidates]

Nodes which are highly discriminated, that is, those which have name slots or those which have many other slots, are used as the subject. If the filler of an agent slot is the subject, the active voice is used, but if the filler of an object slot is the subject, the passive voice is used.

Fig. 1 Examples of the Generation Rules

```
(*criminal (generalizations *person)
           (specializations)
           (slots-of-instance (crime-is action))
           (constraints
                 ((@@ (concept = *person))
                  (@1 (concept = *commit-a-crime)
                      (agent = @@)))))
```

Fig. 2 An example of an entry in the concept dictionary

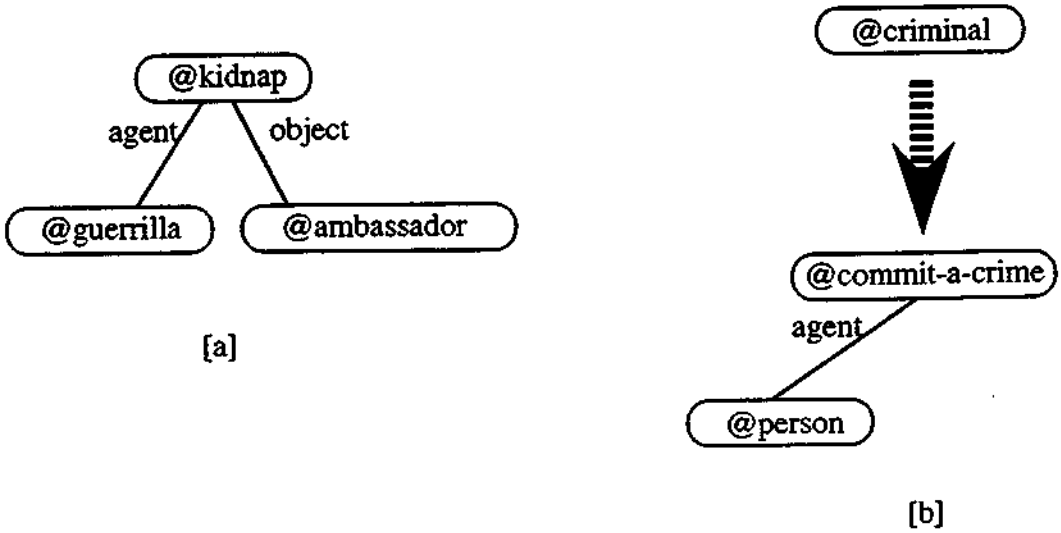

Fig. 3 Concept Conversion

When several representations are used in the text to signify one object, the analyzer can check the identity of these representations by the concept converter using the hierarchy and the constraints.

Suppose that the parts of two sentences in the input text are as follows:

".....the ambassador was kidnapped by the guerrilla....."
".....the criminal....."

When the analyzer finishes the first sentence above, it creates the semantic representation shown in Fig. 3-a. When the analyzer meets "criminal" in the second sentence, it creates @criminal and begins to check the identity of @criminal and other instances already created by the former sentences analysis. The analyzer, however, cannot check the identity of these instances, because there is no obvious information for the identification. (These concepts, i.e. *guerrilla, *ambassador and *criminal, are specializations of *person, but do not provide enough to confirm the identity without ambiguity. Of course, we can use the knowledge that a guerrilla is much more likely to be a criminal than an ambassador, however, a guerrilla can be kidnapped and this static inference makes an error.)

The concept converter invoked by the analyzer reconstructs @criminal into the structure shown in Fig. 3-b using the information written in the constraints entry of *criminal in the concept dictionary shown in Fig. 2.

In the concept hierarchy of the concept dictionary, *kidnap is a specialization of *commit-a-crime and *guerrilla is a specialization of *person. Finally, the analyzer decides that these two structures show the same object. The analyzer begins to check the identity of the structures shown in Fig. 3-a and Fig. 3-b. There are many non-exclusive concepts which are specializations of *person, such as *ambassador, *criminal, *guerrilla, and they are often involved in one text. There are fewer non-exclusive concepts which are specializations of *event and involved in one text. For example, *kidnap and *hijack and *commit-a-crime and *move are mutually exclusive concepts. The analyzer begins to check the structure from an instance of an event (in this case, @kidnap), making it possible to limit the search.

The analyzer in CONTRAST can extract semantic information as precisely as necessary. When we obtain information from input text, the analyzer should simply create the semantic structure using the concepts of the language. Applications of this type of operation are low level database access and machine translation between languages which have almost the same concept hierarchy in a restricted area. Afterward, if one needs more precise information or information represented by other concepts (for example, if one wants to translate texts including some objects which are represented by different concepts in the source language and the target language), the concept converter can reconstruct the semantic information using concepts peculiar to that language.

For example, "younger sister" is represented with the single word "妹(imouto)" in Japanese and there is no single English word that corresponds to a "younger sister." Thus there is a concept representing "younger sister" in Japanese but not one in English. In English, there is a *sister but not a *younger-sister. Japanese does have the word, "姉妹(shimai)," which can represent "sisterhood" or "sisters," but it does not correspond to the English word "sister" which represents a single person. "姉妹(shimai)" should be related to a concept different from *sister or to a set of concepts including sisterhood (i.e., relation) and sisters (i.e.,

a set of people) in the word-to-concept transformation dictionary. It seems that the difference between the sets of concepts included in languages may be the origin of the difference in the style of the utterance in each language. Natives seem to use concepts of their language to construct their thought. Both *younger-sister and *sister are concepts in our concept dictionary. They are connected by the generalizations and specializations entries and the constraint entries.

There is information in the constraint entry that the instance of *younger-sister is one of the instances of *sister whose age is less than some of the instances of the *sister. In case of Japanese-to-English translation:

(1) when the analyzer meets the Japanese word "妹(imouto)," (younger sister) it creates the instance of *imouto, i.e. @imouto;
(2) the generator consults the concept-to-word transformation dictionary for English but can not find the word corresponding to *imouto in English;
(3) the concept converter reconstructs @imouto into @sister who is younger than someone and consults the dictionary again; and
(4) the generator generates the surface text "younger sister" using *sister and its additional information.

3. Lexical Information for Generating High-quality Japanese Sentences

We developed an English text generator based on the method described above which determines the structure of a paragraph using knowledge extracted from the result of observation on real-life text data, as well as the proper event concept via concept conversion. During our effort, we realized that it is necessary to gather lexical information, e.g. selectional constraints, for high-quality Japanese sentence generation.

Our verb lexicon comprises a hierarchy of verbs explained below and selectional constraints for each verb. This information is used for selecting verbs, structures and co-occurring noun phrases during text generation.

3.1. Development of Verb Hierarchy in Japanese

We are developing a verb hierarchy in Japanese based on the verb's syntactic roles in natural language processing. [6] We have classified verbs not by their objective meaning, as was done by the Word List by Semantic Principles (i.e., Bunrui-Goi-Hyou [7]), but by the co-occurrence between the verb and the noun phrases. For example, a transitive verb and an intransitive verb, both of which can refer to the same event are classified as members of the same category in Bunrui-Goi-Hyou. In our hierarchy, however, each word is classified in a different category, because our classification is based on the meanings of the nouns that share the same surface case (e.g., nominative case). This kind of hierarchy is very useful for both analysis and generation in natural language processing.

Our hierarchy of verbs is made manually. The composition process [8] is as follows:

(1) Most "Sahen" verbs, i.e. verbs formed by verbal nouns and the auxiliary verb "する(suru)" (do), were extracted from a database of words occurring in the Asahi Shimbun (newspaper) over a one-month period. Our KWIC (key word in context) program [9] extracted 1436 verbs.
(2) Verbs associated with upper concept words of each word extracted in step 1 were gathered manually.
(3) The most suitable verbs of the candidate verbs gathered in step 2 were selected as the upper concep

verbs of the extracted verbs. Fourteen of the 1436 extracted verbs have multiple meanings, so two or three upper concept verbs were assigned to these verbs. The upper concept verbs were added to the set of verbs in this hierarchy.

(4) The network structure (hierarchy) of the verbs gathered from steps 1 and 3 was created.

(5) Errors in the selection of the upper concept verbs were corrected on close inspection of the network.

Steps 4 and 5 were repeated until our network consisted of 1717 verbs. In this hierarchy, the first layer has 273 verbs and the most precisely explained verb "言う(iu)" (say) has 5 layers. Part of the hierarchy is shown in Fig. 4, where IDs are concept numbers in this system and BNs are classification numbers of Bunrui-Goi-Hyou.

(6) Constraints (such as case relations and a process model) were added to each verb in the hierarchy.

(7) Errors of the constraints were corrected on close inspection of the network.

Japanese "Sahen" verbs are those composed of morphemes, Chinese characters followed by "する(suru)". We also categorized Japanese "Sahen" verbs based on their morphological structure. Through morphological analysis of "Sahen" verbs, we can explore not only their superordinates but also the way each "Sahen" verb restricts the meaning of its superordinate. [10]

作る TSUKURU(ID:10157, BN:) make
  造成する ZOUSEISURU(ID:876, BN:) prepare the ground for housing
  製造する SEIZOUSURU(ID:790, BN:1.386) manufacture
  制作する SEISAKUSURU(ID:789, BN:1.386) manufacture
  生産する SEISANSURU(ID:783, BN:1.3802) produce
    量産する RYOUSANSURU(ID:1422, BN:1.3802) mass-produce
    増産する ZOUSANSURU(ID:866, BN:1.1580+1.3802) increase production
  制作する SEISAKUSURU(ID:772, BN:1.320) make
  作成する SAKUSEISURU(ID:502, BN:1.386) draw up
    醸成する ZYOUSEISURU(ID:727, BN:1.123) brew
    演出する ENSYUTSUSURU(ID:106, BN:1.3832) stage
  作曲する SAKKYOKUSURU(ID:501, BN:) write music
  加工する KAKOUSURU(ID:124, BN:1.386) process
策動する SAKUDOUSURU(ID:10158, BN:) maneuver
  暗躍する AN'YAKUSURU(ID:63, BN:1.345) be active behind the scenes
刷る SURU(ID:10159, BN:) print
  印刷する INSATSUSURU(ID:89, BN:1.3821) print
    印字する INJISURU(ID:90, BN:) type

Fig. 4 Verb Hierarchy

We have completed steps 1 through 5 and are in the process of adding the constraints. The selectional constraints are described by the pairs consisting of a postpositional and a noun (or nouns). So far we have no restriction on selection of nouns, but when we finish adding constraints we will classify nouns in the constraint field of the description of verbs. This classification will be based on the co-occurrence of nouns and verbs. By repeating this process, we can develop a classification of nouns and verbs which is suitable for natural language processing.

3.2. Extraction of Co-occurrence Information from Corpora

To gather information for describing the selectional constraints, we extracted co-occurrence information from RWC Text Databases (See Appendix) using support tools developed by the Information-technology Promotion Agency (IPA). We extracted triplets, i.e. "(noun, postpositional, verb)", from all verbs and adjectives in RWC-DB-TEXT-95-2. We extracted about 100,000 triplets from the 3000 articles in RWC-DB-TEXT-95-2, which contain about 64,000 verbs and 6,000 adjectives. Our lexicon, compiled using the co-occurring data, includes descriptions of the selectional constraints with frequency. Therefore, it enables high-quality sentence generation. For instance, we can distinguish preferences between voices and preferences between nominal and verbal usage of "Sahen" verbs. In line with RWC's principle explained in the appendix, we will make the co-occurrence data available to the public.

During our effort to describe constraints using co-occurrence data, we realized that just gathering triplets as mentioned above is not sufficient for processing certain linguistic phenomena, such as idiomatic expressions and phenomena which need to be analyzed using longer context, as with phrases modifying nouns which modify verbs. Figure 5 shows some examples of these problems.

Therefore, we are not only gathering simple triplets of nouns, postpositionals and verbs, but also co-occurring data in longer contexts from the same newspaper articles. Currently, the data is described using surface words (or phrases), however, we will develop a format to represent the syntactic and semantic structure of the sentences for our lexicon.

4. Conclusion

To achieve high-quality sentence generation, we have developed a text generation module and a concept converter for English. At present, we are applying knowledge we have gained in the process of developing these modules to the creation of a lexicon for generating Japanese sentences. We are also engaged in basic research on a co-occurrency in a longer context.

References

[1] Isahara, H. and Y. Uchida: "Analysis, Generation and Semantic Representation in CONTRAST— A Context-Based Machine Translation System —," Systems and Computers in Japan, Vol. 26, No. 14, 1995.

[2] Uchida, Y., H. Isahara, S. Yokoyama and P. Juola: "English Text Generation from Contextual Representation Structure Based on Real Text Data Dependent Knowledge [No. 1]," Bulletin of the Electrotechnical Laboratory, Vol. 55, No. 11, 1991.

(1) 誘拐の疑いで逮捕する。YUUKAI NO UTAGAI DE TAIHO SURU.

(Arresting someone on suspicion of kidnapping)

*疑いで逮捕する。UTAGAI DE TAIHO SURU. (Arresting someone on suspicion)

Cf. 疑いを持つ。UTAGAI WO MOTSU. (having a suspicion)

The noun "疑い(utagai)" needs modifier phrase.

(2) 私としては賛成できない。WATASHI TOSHITE WA SANSEI DEKINAI.

=私は賛成できない。WATASHI WA SANSEI DEKINAI. (I can not agree.)

宝石としては紅玉を意味する。HOUSEKI TOSHITE WA KOUGYOKU WO IMI SURU.

(As for a jewel, it means a diamond.)

*=宝石は紅玉を意味する。HOUSEKI WA KOUGYOKU WO IMI SURU.

(A jewel means a diamond.)

The idiomatic expression has to be analyzed properly.

(3) 一年間に首相が三人交代する。ICHINENKAN NI SYUSYO GA SANNIN KOUTAI SURU.

(Three persons take prime minister in one year.)

首相が三人交代する。SYUSYO GA SANNIN KOUTAI SURU.

(Three persons take prime minister)

*一年間に首相が交代する。ICHINENKAN NI SYUSYO GA KOUTAI SURU.

(Taking prime minister in one year.)

首相が交代する。SYUSYO GA KOUTAI SURU. (Taking prime minister.)

When a term is assigned, the number of people will be obligatory.

Fig. 5. Examples of problems on the generation of Japanese sentences

[3] Uchida, Y., H. Isahara, S. Yokoyama and P. Juola: "English Text Generation from Contextual Representation Structure Based on Real Text Data Dependent Knowledge [No. 2]," Bulletin of the Electrotechnical Laboratory, Vol. 56, No. 5, 1992.

[4] Isahara, H. and S. Ishizaki: "Natural Language Understanding System with Concept Hierarchy," Proc. of Pacific Rim International Conference on Artificial Intelligence 90, 1990.

[5] Isahara, H. and S. Ishizaki: "Concept Representation of Machine Translation System CONTRAST," (in Japanese) Transactions of IPaJ, Vol. 35, No. 6, 1994.

[6] Kawada, R., H. Inui and H. Isahara: "Verb Classification of Case Frames in Japanese," (in Japanese), 94-NL-101-16, IPSJ, 1994.
[7] National Language Research Institute (eds.): "NLRI Publications Source VI Word List by Semantic Principles," (in Japanese) Shuei Shuppan, 1964.
[8] The Institute of Behavioral Sciences: "Report on the Development of the Verb Concept Network," (in Japanese), 1994.
[9] Juola, P. and H. Isahara: "An Efficient KWIC System for Japanese Text," Bulletin of the Electrotechnical Laboratory, Vol. 55, No. 10, 1991.
[10] Inui, H., F. Motoyoshi and H. Isahara: "Categorization of Sahen verbs based on their morphological structure," (in Japanese), 95-NL-110-15, IPSJ, 1995.

Appendix: RWC Text Database -- A Publically Available Very Large Tagged Japanese Corpus --

Very large and richly annotated corpora are indispensable for corpus-based natural language processing research. The Real World Computing (RWC) Program is a research program funded by the Japanese Ministry of International Trade and Industry (MITI) and aims at developing flexible information processing systems for diversified information in the real world by introducing intuition-like processing functions. Research under the RWC Program is conducted mainly by the Real World Computing Partnership (RWCP), a consortium of 16 Japanese companies and 4 foreign research institutions. RWCP established the RWC Database Working Group in 1994 to gather and utilize real world knowledge. The Working Group is building four databases: text, speech, image and multi-modal databases. Since 1994, the text group of the RWC Database Working Group has been building an annotated text database of modern Japanese for the research and evaluation of NLP technology.

The RWC text database should:

(1) be very large scale,
(2) include accurately annotated corpora,
(3) be balanced, and gathered from actual texts, and
(4) include bilingual and speech corpora.

In principles, the corpora should be:

(1) AVAILABLE WITHOUT CHARGE to the public for research and evaluation of NLP technology,
(2) built under COOPERATION AND DISPERSION, and
(3) GENERAL AND INDEPENDENT of any one specific linguistic theory.

The following are the RWC text databases currently available.

(1) RWC-DB-TEXT-94-1

Morphologically analyzed data of MITI's report (Manually post-edited. Including MITI white papers

for 1993-1995.)

(2) RWC-DB-TEXT-94-2

Morphologically analyzed data of Japan Electronics Industry Development Agency's annual report. (Manually post-edited. Survey report on the trend of natural language processing.)

(3) RWC-DB-TEXT-95-1

Differential data of the results of morphological analysis of the CD-Mainichi Shimbun (newspaper) (covers articles from the Mainichi Shimbun from 1991 to 1994). This very large tagged corpus is the result of automatic morphological analysis of all sentences in CD-Mainichi Shimbun from 1991 to 1994. This 4-year database comprises about 100 million words (or morphemes).

(4) RWC-DB-TEXT-95-2

Differential data of the results of morphological analysis of the CD-Mainichi Shimbun (Manually post-edited, 3000 articles from 1994). This is the result of post-editing all sentences in the 3000 articles extracted from RWC-DB-TEXT-95-1. The 3000 articles, the length of which ranged from 400 to 600 characters, were randomly selected.

(5) RWC-DB-TEXT-95-3

Articles tagged with UDC. (30,000 articles from 1994.)

Since RWC-DB-TEXT-95-1 is very large, it was only processed mechanically. Therefore it may include errors, although they are insignificant for practical use. We manually post-edited parts of RWC-DB-TEXT-95-1 to make RWC-DB-TEXT-95-2. This database will give basic data for estimating the accuracy of automatic tagging. Furthermore, RWC-DB-TEXT-95-2 can be used as training data for a morphological analyzer to make the remaining data in RWC-DB-TEXT-95-1 more accurate.

Reference

Isahara, H., F. Motoyoshi, T. Tokunaga, M. Hashimoto, S. Ogino, J. Toyoura and R. Oka.: "Building text database with POS tags by RWCP," (in Japanese), 1st Annual Meeting of Japanese Society of Natural Language Processing, 1995.