[MT Summit IV, July 20-22, 1993, Kobe, Japan]

An Example-Based Multilingual MT System in a Conceptual Language

Hiroshi Yasuhara
Japan Electronic Dictionary Research Institute, Ltd.
c/o Systems Laboratory OKI 11-22 Minato-ku 4-chome Tokyo,108 Japan
tel +81-3-3454-6228, fax +81-3-3454-5774
yasuhara@okilab.oki.co.jp

1. Introduction

A long history of the development of the machine translation systems presented several building methods. There are two judge points. The first decision is transfer vs interlingua. The second is rule base vs example base. These are independent selections. Therefore by the combinations of the above selections the machine translation systems are classified into four types. We call them RB_TR, EB_TR, RB_IL, and EB_IL respectively. RB means Rule Base, TR means TRansfer, EB means Example Base, and IL means InterLingua. RB_TR is the most orthodox system and has been developed by a lot of commercial systems. EB_TR was born in [Nagao,84] to gain more natural style of target sentences between structurally different languages' translation such as English and Japanese. [Sato,91], [Furuse,92], [Sumita,92] developed EB_TR based on a parallel corpus. RB_IL has also a long history. The main purpose of IL directs to develop a cost effective multilingual machine translation systems. There are two commercial RB_IL systems in Japan. EB_IL has the latest system of the four types. [Sadler,89] proposed a same philosophy based on Esperanto as an intermediate language. We took a more deep analysis because Japanese language has different syntactic structures against English.

MT development is very time consuming task and labour oriented task. In particular, building of lexical data is first ranked job in MT development. The works of collection and coding of more than several hundreds thousands vocabularies are over the single company. The language is common intellectual property of the human beings. The lexical data base evolves and should be maintained forever. EDR was founded by the government and eight Japanese computer companies. The goal is the development of theory-independent large scale dictionaries intended to the next generation natural language processing systems. The products will contribute to the developments of MT systems much focusing on language processing freed from the development of the own dictionary.

The second ranked job is grammar rules writing. It is intelligent works. It is not sure that better qualities of MT system will result from the more investment to the coders of the grammar rules. The rules are linked each other and the updating or incrementing of the rules needs the knowledge of total systems and part to part relations. When the scale of the grammar rules will be rising along with the system revisions, the maintenance of rules may reach the critical point. Once reaching such a point the system will cease to be updated. Or if you select another linguistic theory for a new system, the set of gathered rules are in vain.

In this paper we show an EB_IL system. In the following we use a term EB_MMT to specify our system. MMT means Multilingual Machine Translation. The key lexical resources in the system are Word Dictionary, Concept Classification Dictionary, and EDR Corpus which are products of EDR..

EB_MMT inputs syntactically parsed data of a source language sentence, analyzes the sentence into conceptual level interlingua by an example method and generates words list of target language sentence also based on an example method. Finally we present further research items to be remained.

2. Motivations of Multilingual MT in a Conceptual Language

Purpose: EDR introduced an interlingua into the lexicons and the natural language processing. There are no consensus with regard to universal language as an interlingua. Some people even negate the existence of an interlingua. However EDR took an approach based on interlingua. The main reason was cost reduction in multilingual MT development. We use a term "conceptual language" as EDR interlingua in the following. Conceptual language has more essential and positive roles in the natural language processing. The first point is the clearness in the meanings of the words. In generally speaking, conceptual language can be treated as a formal language. It is a good property for computers. The second point is universality. Conceptual language is language independent. EDR Concept Dictionary gives a frontier of new generation natural language processing. We believe conceptual language processing will provide us with a new relation between computers and human beings because computers share a common language with us. We have not yet enough knowledge about conceptual language. We should launch to the new paradigm. The most appropriate way is by development of a prototype system of mapping between natural language and conceptual language.

Goal: Every natural language is analysed to the conceptual language. Conversely the conceptual language can generate a natural language.

We have many kinds of difficulties in the conceptual language processing. In analysis we have problems of word sense selection and case role selection and in generation we have problems of word selection and function word selection. They are all tough problems.

3. A Japanese to English Example-Based Multilingual MT System

3.1 Characteristics of EB-MMT

There are three alternatives in the natural language processing as shown in the Fig.l. The first is dictionary base approach. Morphological analysis and generation belongs to dictionary base approach although heuristic rules may be used. Syntactic level adopts primarily rule base approach. However there is a possibility of employment of example base approach. To attain semantic level by the rule base approach was difficult. The example base approach was appeared as an attack on the difficulty. This approach is profitable in the case of EDR lexical resources.

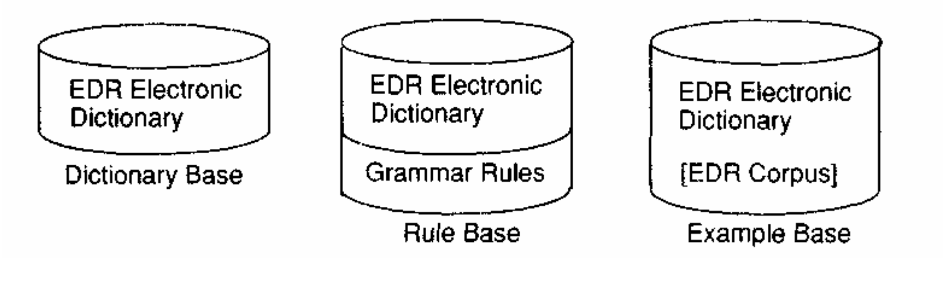
EDR Electronic Dictionary
Dictionary Base
EDR Electronic Dictionary
Grammar Rules
Rule Base
EDR Electronic Dictionary
[EDR Corpus]
Example Base

Figure 1. Strategies of Language Processing and their Linguistic Resources

In the natural language processing systems, disambiguations of the given sentences are main goal. The EDR conceptual language is less ambiguous forms than the natural language. The content words with homonymy and polysemy are selected to a single word sense concept. Functional words such as the prepositions are assigned to case role labels. The dependency relations in a sentence are represented as a semantic network. If there exists a high quality natural language processing system which has an ability to output the above mentioned conceptual language, computer can handle a variety of natural language texts. Then the information of human beings can be shared with computers. The interlingua method of MT contributes as first objectives to the communication of different languages' people, but the essential impacts are more broad communication. Computers can communicate each other and people can communicate with computers. These paradigms are dreamed at starting point of EDR's design of the conceptual language. It reminds us that Esperanto was devised as people's common language. The reason why the interlingua was not spread out into the information society, I believe, was the difficulty of analysing the natural language. The correct selections of just one exact word sense and exact case label from a lot of candidates. The rule base approach to the disambiguations of natural language has found difficulties of focusing the grammar rule writings. If the rules are decomposed to resolve the language phenomena, then the rules lost of the uniformity of the theory and the power of generality. They made threads of small classes.

The new approach called ecology of natural language [Walker,92] rebirthed from these thinkings. We can collect a lot of electronic text data. The computer is used to get linguistic knowledge from large scale text data. The situation trends toward a shift from an exact rule based system to a robust and plausible analogy based system. We look for the most plausible sentences in the corpus. Then the task of disambiguation is reduced to comparison with the corpus. We applied these ecological or corpus linguistic technologies to an MMT in a conceptual language.

3.2 Basic Framework

(a) Concept Dictionary

EDR defined concept primitives from word senses in each language. This strategy that concepts are extracted by the words gives a criterion about which concepts are primitives. We thought it is a difficult approach that a set of concept primitives is given a priori and then mapped to the words. The conceptual language consists of a combination of the concepts and concept relation labels, equivalent concepts in Japanese and English have the same concept identifiers. The formal specification of concepts are given in [EDR,93]

Concept Dictionary consists of two sets. One is Concept Classification Dictionary and the other is Concept Description Dictionary. EB_MMT uses Concept Classification Dictionary. It corresponds to thesauri in the natural language. Since concepts are disambiguated, a concept has a unique position in the Concept Classification Dictionary. There are allowed multiple passes that have multiple parents in order that the different views of one concept are permissible. Concepts are classified into things, events and attributes. Events class hierarchy has an important role in the determination of similar sentences. [Nirenburg,89] demonstrated such an ontology is effective in domain specific MT. On the other side EDR constructed the Concept Classification Dictionary on the general words.

(b) Similarity Function

As a fundamental tool we devised similarity functions which have central roles in the extraction of the most similar sentence from EDR corpus.The similarity function is related to a metric function in the EDR conceptual space. The naive sense that similar concepts exist near each other is applied to the similar degree in Concept Classification Dictionary. We have defined seven similarity functions in analysis and generation process: $\sigma a(W,W')$, $\sigma a(C,C')$, $\sigma a(S,S')$ and $\sigma g(W,W')$, $\sigma g\ (C,C')$, $\sigma g(CL,CL')$, $\sigma g(S,S')$ where σa, σg, W, C, CL and S represent similarity function of analysis, similarity function of generation, word, concept, concept language and sentence respectively. The detail specifications of these functions are defined in the [Cui,92], [Komatsu,93].The measurement of a degree of similarities is based on the Concept Classification Dictionary. That dictionary covers not only noun categories but also verb categories. The materials or things concepts are dispatched different branches from the events concepts. Sister or brother related concepts and father children related concepts have a high score of the similarity. But associated relations such as between food concepts and eat concepts have a very low score. They are stored in the Concept Description Dictionary. Since the similarity functions are based on the concepts, they are commonly used in any languages.

(c) EDR Corpus

Another important resource to EB_MMT is EDR corpus. Raw EDR Corpus is converted into an example database which consists of simple sentences. EDR Corpus behaves as a prescription of natural language analysis.The corpus may be considered to be our brain memory of language. The right or wrong is judged based on the data. The data can be considered as common knowledge of the language. Being stored in the computer memories, the data can continuously increase the volume of the knowledge.

EDR Corpus, which is a very large set of conceptually tagged sentences in a monolingual as shown in the Fig.2. In EDR there has been developed English and Japanese corpora. Our tagging system is complete in the sense of natural language processing systems. It includes source sentences, morphological structure, syntactic structure and semantic structure to every sentence. For example the syntactic representation of "They play tennis in the park." is "w#play w#they M(sbj); w#play w#tennis M(obj); w#play (w#park S(the) S(in)) M(pp)". The semantic representation is "c#play c#they agent > ; c#play c#tennis object > ; c#play c#park place > ".

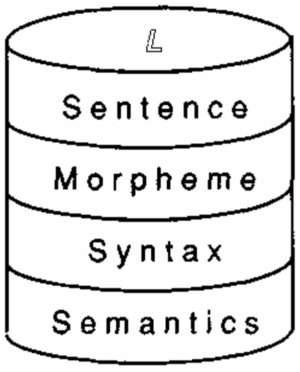

Figure 2. EDR Corpus for Language L

We had already collected 200,000 sentences of Japanese and English. EDR's event concepts have 29,676 concepts. If we assume every pair of head word and its concept identifies the usage uniquely, then there needs at least 80,794 sentences in the corpus. More theoretically speaking, the corpus will be ideal if the following are satisfied.

Registration criterion:
Given any corpus E, its any sentence a, some sentence P and some small numbers e, eel
$(0 < \varepsilon, \varepsilon cl < 1)$

$\sigma sen\ (\alpha,\beta) < \varepsilon \cup \sigma cl\ (\alpha cl,\ \beta cl) < \varepsilon cl$; σsen is a similarity function of sentence, σcl is a similarity function of an interlingua.

Then β and βcl can be registered to E.

Deletion criterion:
Given any corpus E, some sentence α in E and some number θ, θcl near to 1 $(0 < \theta, \theta cl < 1)$

$\sigma sen\ (E\text{-}\alpha,\ \alpha) > \theta \cap \sigma cl\ (E - \alpha cl,\ \alpha cl) > \theta cl$

Then α and αcl can be deleted from E.

We call such a corpus has uniform properties if ε, εcl, θ and θcl are constant in the corpus.

3.3 System Overview

We took an interlingua approach in multilingual MT. MT from a language L0 to any other language L1 proceeds as shown in Fig.3. In the prototype system L0 is equal to Japanese and L1 is equal to English. Step (1) analyses Japanese sentence to morphological elements. Step (2) analyses into syntactic representation. We assume these steps have already successfully finished and start from step (3). Semantics in this figure is identical to EDR conceptual language. Step (4) generates target language words from concepts and relations. At that time syntactic structure is determined by head constituents. Step (5), (6) are neglected in our EB_MMT. The step (3), (4) are the main parts. Concept Transfer Table is a kind of a mapping table between equivalent concepts. In principle EDR concepts are language independent but some words have no exact meanings in other languages.

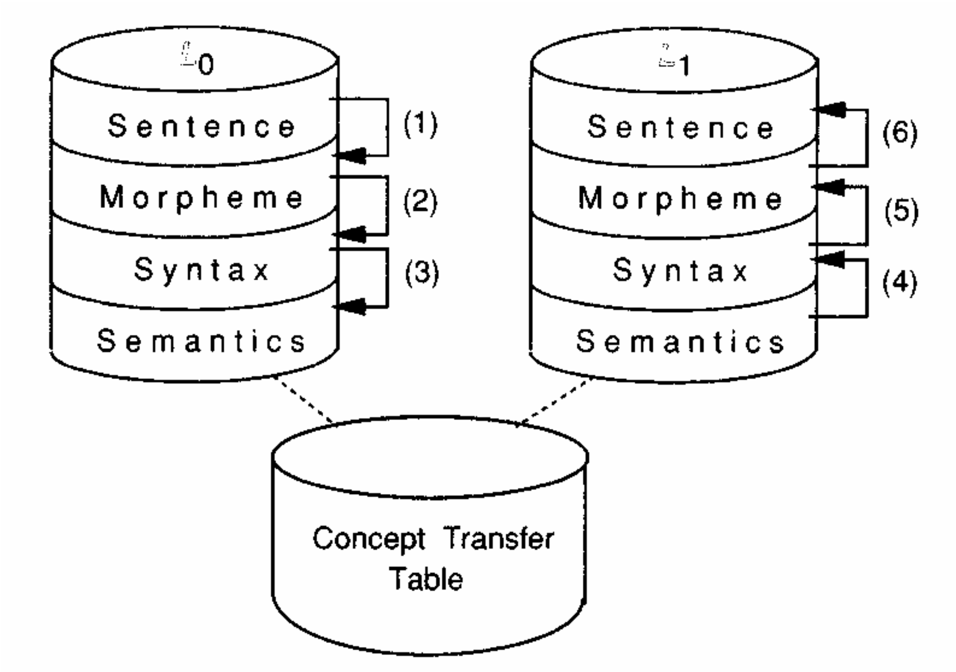

Figure 3. Analysis and Generation Steps of the System

3.4 Semantic Analysis and Generation

Simple Japanese sentence has a syntactic structure that a head verb dominates noun phrases with a particle. Semantic analysis outputs a conceptual language from a syntactic structure. English generation outputs an English sentence from the conceptual language. The Fig.4 shows the data structure of input, conceptual language, and output.

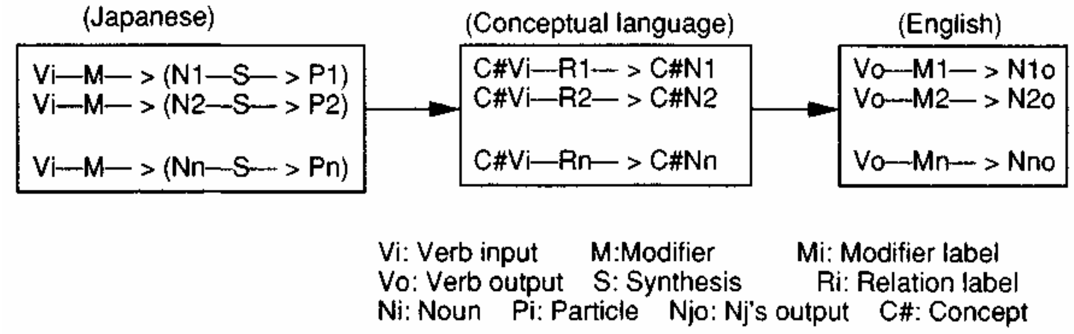

Figure 4. Data Structures of Japanese Analysis and English Generation

Japanese semantic analysis at first selects the most similar sentence in the Japanese corpus. We assume there are sentences of which heads match exactly an input head verb. A head verb dominates sentence structure. Therefore it is difficult to select a good example without exact matches of the head verb. Sentence similarity function o-a(S,S') uses case constituent words with particles. It calculates the similarity of the words A and B based on the shared numbers of their word senses and their super concepts. The function sums up all word similarity functions a-a (W,W') of the content words weighted with particles. The highest score sentence in the corpus is selected as an example sentence. And the corresponding conceptual language will be used for the concept selection of each word. The concept similarity function σ-a(C,C') calculates the degree of the similarity between word senses of the input word and the concept corresponding to the same modifier label.

This process is equivalent to case frame matching of the head verb. The corpus based similarity matching has more flexible search than rigid case frame matching because Japanese sentence has omissions of any obligatory cases. Similarity functions need no semantic markers. Concept Classification Dictionary supplies more dense set of semantic marker. The order of semantic categories is ten thousands and two digits bigger than traditional semantic markers. The next step is an assignments of concept identifier to every content word and case relation to every modifier. A conceptual language corresponding to the most similar sentence will guide a production of a conceptual language corresponding to the input sentence. The case relation label Ri is deduced from the example sentence. And the node C#Ni is derived from by the concept similarity function between the concepts of input word Ni and the corresponding concept of the example sentence. It should be noticed that the procedure of the semantic analysis is one to one corresponding of the surface structure. It is a week point of this method. In particular Japanese and English are very different languages. A word for word translation is lacks of natural sentence style. We think some mechanisms of transfer or paraphrasing might be needed in the conceptual level. But in the prototype system we delayed the introduction. However we set a simple transfer mechanism using bilingual word dictionary from a source language to a target language. For example a Japanese phrase ' 水泳(swimming) を (a objecive particle) する (practice)' corresponds to 'practice swimming' but in English a single verb 'swim' is better. Similarily a Japanese verb '散歩する(stroll)' does not correspond to English phrase 'take a walk'. These phrase for word or word for phrase transfer will be supported by the bilingual dictionary.

English generation proceeds as follows. It looks for the most similar conceptual language in the English Corpus. It uses a concept language similarity function σg(CL,CL') which is built up concept similarity function σg(C,C') between a concept of a given conceptual language and its corresponding concept of the corpus. The σg(C,C') has high score when words sets corresponding to C and C' have same set of parts of speech and C and C' have sister position in the Concept Classification Dictionary. The Concept Classification Dictionary also plays the essential role. The score is a summation of the concept similarity of each children nodes weighted by the case relation. The highest score concept language in the corpus is selected as an example. Its corresponding syntactic representation of an English sentence selects head verb and its obligatory case constituents of the given conceptual language by a sentence similar function σg(S,S') which is built up by word similarity function σg(W,W'). Optional case constituents are selected as another similar search of the corpus.

4. Open Problems

Our approach is empirical. Henceforth it is necessary to evaluate similarity functions and corpus itself by a lot of data, and compare with other approaches. EB_MMT is a prototype. We have not a confirmative data though the author thinks this example-based multilingual MT approach is prospecting. The following will be research items on this approach.

(1) Corpus

How to build a corpus with uniform properties. Is it possible to be uniform?
How many sentences will satisfy the uniform properties? Is it finite order?
How much computing power does it need to test registration and deletion criteria.
Relation between sentence similarity and its interlingua similarity

(2) Conceptual language

What kind of metric or similarity is suitable for the conceptual language.
Mathematical modelling of CL

(3) Analysis and generation

If similarity degree is higher than certain value, is it confirmed to analyze or generate successfully?
Is it necessary of conceptual transfer between two languages?

5. Conclusion

We have discussed a basic framework of example-based multilingual MT system. The system quality heavily depends on the quality of the Concept Classification Dictionary and EDR Corpus. We have not yet fully evaluated these lexical resources. However the introduction of similarity in the natural language processing provides us with a quantitative discussion.

Although the system deals with multilingual MT, the methodology may be applied to a monolingual semantic analysis and generation. We expect the example-based semantic processing will rapidly grow in accordance with the technology of EDR Corpus and Concept Dictionary. Then it will advance a new field of semantic level applications.

Bibliography

[Cui, 93] Cui, J.; Komatsu, E. and Yasuhara, H. : "A Calculation of Similarity between Words Using EDR Electronic Dictionary", Reprint of IPSJ, Vol. 93, No.l (January 1993) (in Japanese)

[EDR, 93] EDR: EDR Electronic Dictionary Specification Guide, TR-041 (1993)

[Furuse, 92] Furuse, O. and Iida, H. : "An Example-Based Method for Transfer-Driven Machine Translation", Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, pp 139-150 (June 1992)

[Komatsu, 93] Komatsu, E.; Cui, J. and Yasuhara, H. : "English generation from EDR concept relation representation" Proc. of IPSJ 45th, pp323-324 (August 1992)

[Nagao, 84] Nagao, M. : "A Framework of A Mechanical Translation between Japanese and English by Analogy Principle, Artificial and Human Intelligence (A. Elithorn and R. Banerji, editors) Elsevier Science Publishers, B.V. (1984)

[Nirenburg, 89] Nirenburg, S. and Levin, L.: Knowledge Representation Support, Machine Translation Vol. 4 pp25-52 (1989)

[Sadler, 89] Sadler, V.: Working with Analogical Semantics: Disambiguation Techniques in DLT, Foris Publications, Dordrecht Holland, (1989)

[Sato, 91] Sato, S. : "Example-Based Translation Approach" Proc. of International Workshop on Fundamental Research for the Future Generation of Natural Language Processing, ATR Interpreting Telephony Research Laboratories, pp. 1-16 (1991)

[Sumita, 92] Sumita, E. and Iida, H. : "Example-Based Transfer of Japanese Adnominal Particles into English", IEICE TRANS. INF. &SYST., VOL. E75-D, NO.4 (July 1992)

[Walker, 90] Walker.D.: "The Ecology of Language", Proc. of International Workshop on Electronic Dictionaries, Oiso, Japan, pp. 10-22 (November 1990)