Semantic based generation of Japanese German translation system

- Result and Evaluation-

K. Hanakata
Institut f. Informatik
University of Stuttgart
Herdweg 51
D-7000 Stuttgart 1
F.R. Germany

A. Lesniewski
Standard Elektrik Lorenz AG,
Ostendstrasse 3
D-7530 Pforzheim
F.R. Germany

S. Yokoyama
Electrotechnical Laboratory
Umezono, Sakuramura, Niihari
Ibaraki 305
Japan

Abstract
Project SEMSYN*** achieved a state where a prototype system generates German texts on the basis of the semantic representation produced from Japanese texts by ATLAS/II of Fujitsu Laboratory. This paper describes some problems that are specific to our semantic based approach and some results of the evaluation study that has been made by the Germanist group.

I. Generation procedure in SEMSYN

This section summarizes the SEMSYN genration procedure. Those readers who are more interested in the SEMSYN system are recommended to read our previous COLING84[1] paper or the paper submitted to this conference[2]. The generation process begins with the conversion of the semantic networks, each represents one sentence, into a so-called IKBS (Instantiated Knowledge Base Schema.) The IKBS is an instantiation of case or concept schemata denoted by semantic symbols as nodes in the semantic network. A case schema contains three main description slots; **a)** roles of cases associated with the semantic symbol, **b)** transformation rules of schemata, **c)** choice of German syntactic realization schemata.

Being triggered by the semantic symbols of the given network, IKBS specifies the best basic syntactic structure associated with a German word by checking fillers of roles and converts them into functional roles within each German syntactic category. A German syntacto-morphological component called SUTRA-S[3] a extended version of SUTRA [4] generates German surface texts from the instantiated syntactic structure called IRS (Instantiaed Realization Schemata.)

Though English-like terms are used for semantic symbols, the choice of a German word associated with each semantic symbol and its syntactic structure very differ from the English corresponding one.

II. Some problems of semanitc based translation approach

There are some advantages as well as disadvantages of the semantic based approach, which we anticipated at the beginning of the project. Theoretically speaking, a reason why we adopted a semantic based approach againt the syntactic transfer approach is founded on the cultural difference and communication barriers between the two project groups that cooperate with each other to build up a translation system. Understanding the content of the origenal sentence from the given semantic representation the generation group could express it in a way that is common in its mother tangue, relatively free from the syntactic restriction and lexical corresponding terminology. It is a well known fact that one language of a culture can only be interpreted and not literally be translated into the other languages of different cultures, as it would be possible within the same cultural sphere. As the matter of fact we often took this advantage in our generation system.

On the other hand, exactly this freedom turned out frequently to be a disadvantage on the generation side. Dealing with real data (titles of sientific papers in the field of information technology from the Japanese data base JOIS) we encountered new problems we didn't expect before and recognized the limit of our approach. In the following we describe some of these problems:

(1) Lack of information in Japanese original text
We had also to come up with this well known problem such as lack of articles (definite or indefinite) and of distinction between numbers (singular or plural) for nouns as well as verbs. We embedded some heuristic rules in KBS and dictionary to add these syntactic features, if they must not be missed in the German text. There still exists deeper semantics which rules the decisions, but cannot be represented in general, except for very limited cases. Heuristic rules are based on our **ambiguity conservation principle**, i.e. we keep the ambiguity of input text as much as possible to avoid any active selection of one alternative, that might lead to a wrong expression from the view point of the author of the titles. Following examples show typical errors of numbers and articles generated by the present SEMSYN heuristics. They also illustrate how difficult it is to find a trade off between the ambiguity conservation and an active decision infered from the content:

E.g. 1:小型計算機を用いての大規模グラフィックプログラムの実行
SEMSYN generation: Die Verwendung von kleinen Computern zur Durchfuehrung von grossen graphischen Programmen
(The application of small computers for the execution of large graphic programms)
Comment: The author of the paper will discuss how to use a small computer to execute a very large graphic package, so readers may naturally assume one small computer instead of many small computers, though it is possible to assume the latter. On the other hand, it is generally assumed that a computer processes many programms. For this reason the latter plural case is more natural than the former case. However, it is a bad German to have neither a number feature nor an article as it is in the original text.

E.g. 2: 実時間のアプリケーションのための分散型オペレーティングシステムの核の開発
SEMSYN generation: Die Entwicklung des Kerns beim Betriebssystem vom verteilten Typ fuer real-time Anwendungen.
Correct German: Die Entwicklung des Kerns eines verteilten Betriebssystems fuer Echtzeitanwendungen
(The development of the kernel in the operating system of the distributed type for real-time applications)
Comment: It is assumed that the author developed the kernel of one distributed OS, instead of many distributed OS, for many applications.

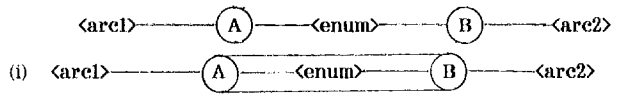
2) Ambiguity of conjunctions
One of the hard problems we expected in our semantic

representation was the ambiguity in the coordinating conjunctions in an attributive context such as:

<AP> A, B and C <PP>.

E.g.3
high speed bus, memory and switching in bit slice technology

The scope of context could be made unique, if the semantic network could allow such a node which denotes a subnetwork. The following conjunctive subnetwork is classified into three basic cases:



(i)

E.g.4 . . efficient algorithm and computation for parallel processors

(ii)

E.g.5 . . new algorithm and computation by a vector processor

(iii)

E.g.6 .. high-speed bus and memory management by stacks

In practice, however, we found that 90% of about 380 titles which contain conjunctions among 2000 titles we so far generated from the given semantic networks belong to the case (i); only about 8% are the case (ii), and the rest is the case (iii). This statistic results may be spesific for the titles, but this indicates that authors of titles are aware of the syntactic structural ambiguity and consequently try to avoid the above straightforward sequence of conjunctions except for the case (i).

Beside this statistic sample-based facts, the conjunctive ambiguity is further weakened by the fact that the generation system produces ambiguous titles according to our ambiguity conservation principle to let expert readers naturally infer which is meant by the author. At the moment we deal with the both cases (ii) and (iii) by exploiting this possibility to convey the ambiguity so as if it were the case (i).

Though this conjunctive ambiguity in semantic networks seemed to be a serious factor at our first glance on them, it fortunately turned out to be a very minor problem as the evaluation study indicates.

3) Stylistic problem

Generally speaking, a semantic based generation approach has a strong advantage as well as disadvantage in terms of sentence styles. The stylistic advantage is based on the large freedom of interpreting a given semantic representation. A serious disadvantage is the exactly the other side of this interpretation freedom. Following examples illustrate typical stylistic problems of our generation system:

E.g.7 ＰＤＩＬを用いた通信プロトコルの記述，シミュレーション，開発

SEMSYN generation: Spezifikation, Simulation und Entwicklung von Protokollen, fuer die PDIL verwendet werden, fuer die Kommunikation

(Specification, simulation and development of protokol, for which PDIL is applied, for the communication)

Comment: "fuer die PDIL verwendet werden, fuer die Kommunikation." should be expressed as ".. eines Kommunikationsprotokolls unter Verwendung von PDIL." (".. of a communication protocol by using PDIL", in stead of " for which PDIL.. .)

E.g.8 記憶の中の導出可能な情報の表現

SEMSYN generation: Die Repraesentation von Informationen, die ableitbar in einem Speicher gewesen werden.

(The representation of informations, which can be derived in a memory)

Comment: The clause " .., die ableitbar.." should be replaced by an adjective phrase "von der aus dem Speicher ableibaren Information".

E.g.9 ＣＯ２レーザ表面硬化法においてデータベースシステムを用いた条件設定法

SEMSYN generation: Die Verwendung von Datenbanksystemen zu einem Verfahren zur Aufstellung von Bedingungen beim Verfahren zur Verstaerkung von Oberflaechen des Kohlendioxidlasers

(The application of data base systems for the listing of conditions for the surface hardening procedure with CO2 laser.

Comment: Instead of repeating nominalized case frames for role purpose "Verfahren zur Aufstellung" and "Verfahren zur Verstaerkung" should the latter be expressed as "Oberflaechenverhaerterungsverfahren"

Though bad styled expressions may transmit the correct meaning, they substantially reduce the understandability of the generated texts. The stereotypical bad styles can be easily improved in some cases; however, the style conversion problems seem to have its inherent continuous depth from "easy to patch" to the infinite depth to be pursued in a long run.

4) Cultural difference problems

Before we started the project we discussed many problems that are specifically attributed to the well known cultural difference. In the following given are some of the real problems we encounted in dealing with title translations:

i) Focus shift

We have frequently to come up with the difference of focussing, that forces us in a conflict situation whether we should prefer fidelity of the translation to the common style of German titles.

E.g.10 通信プロセスのための仕様記述向け意味論

SEMSYN generation: Die Verwendung von Semantiken zur Spezifikation von Kommunikationsprozessen

(The application of semantics for specification of communication processes)

Comment: The original Japanese text does not contain an explicit word that coreponds to the semantic symbol "USE.ACT", that is infered by the analyzer. Generally speaking, however it sounds better in German if a expression explicates the meaning in a more resolved form, while ambiguous expressions or even fuzzy expressions are prefered in Japanese. In this example the purpose are expressed as "zur" implies the application of the semantics.

ii) Reversed causality

The most striking case that exemplifies the opposite relation between east and west is the reversed expression of causality, mostly would-be results are used instead of the cause in Japanese and vice versa in west. Following example demonstrates the fact:

E.g.11 計算機の専門知識のある教師を学校体系の内部で養成する問題

SEMSYN generation: Probleme bei der Ausbildung ueber Lehrer, die spezielles Computerwissen besitzen, innerhalb eines Schulsystems.

(Problems of training teachers, who own special computer knowledge, within the school systems)

Comment: Here the Japanese original text means that the special computer knowledge is a result of the training. If the teachers have already this special knowledge, they don't need the training. Therefore, it must be expressed as "so as to have ."

At the moment neither our analyzer nor generator can afford such a deep understanding of input texts. Our approach is still open to enrich the TRAIN scheme to represent causal relation of the TRAIN concept which forces to reverse the causality of given meaning.

III. Evaluation

About 20% of the translation results produced from the available semantic networks are evaluated. In order to avoid the misunderstanding it is worth to make it clear that this evaluation was not done by the so-called blind test, instead, all semantic networks are already used as our training samples. This is because at the time when the evaluation study started we had only 2000 semantic networks available. The evaluation results are summarizes as follows:

Grade	Fidelity	Grammaticality
1	68.0%	16.7%
2	29.7%	48.0%
3	2.3%	19.0%
4	0.0%	16.3%
5	0.0%	0.0%

Explanation:

Grade Grammaticality

1: Exactly the same meaning as the original text
2: Almost same content
3: Still acceptable and informative
4: Only partially acceptable
5: Nothing to do with the content of the original text

Grade Fidelity

1: Correct style, syntax and morphology
2: Correct syntax and morphology, but stylistic defect and vice versa
3: Still readable, but substantial mistakes in syntax, morphology and style
4: Almost unreadable as German text
5: Not German

Based on this evaluation results we sorted our error sources. Following results show the error classification from which the readers can figure out the development state of our system.

Error classification	Occurrence among 300 titles
Fidelity	
(a) Lexicon(not standard terms, inappropriate terms);	105
(b) Selection of prepositions	89
(c) Word construction (noun compounds)	88
(d) Articles (def., indef.,)	50
(e) Number (sing., plur.)	43
(f) English terms (technical terms instead of German)	27
(g) Relative clause instead of Np, PP	19
(f) Focus	15
(h) "Unter Verwendung von"	10
(i) Possessive attribute "von" instead of genetive	8
Grammaticality	
(i) Selection of preposit	34
(ii) Word compounds	23
(iii) Articles	14
(iv) Relative clauses	12
(v) Focus	9
(vi) Conjunction alignment	5
(vii) Numbers	4
(viii) Attributes	3

The above classification indicates that dictionary problem cannot be solved in a short term. Especially in our approach, a semantic symbol generally corresponds to an upper concept, under which an appropriate German term is registered as a specialization. Therefore the terminology selection within a lexical entry is indirectly done through its context. Again, this very advantage of expression freedom causes a bad selection of a target word. We need time to polish our semantic German terminology data base so that system can select right German words in general.

The noun compound is a specific problem in German. By constructing a noun compound a stylistic problem may elegantly be solved (cf. e.g.9, 10), because otherwise using a modifier (possessive attributes, qualifiers and quantifiers, etc) results in an awful expressions that can not be compared with an alignment of English terms.

We also found a conflict situation in connection with the selection of technical terms. While we prefered common English technical terms in the field of information processing as CS experts for the reason of easy understanding, evaluators emphasize the authority of national standard technical terms (DIN), e.g. CRT (Datensichtgeraet), real-time (echtzeit), etc.

The reason why the German ideom "unter Verwendung von" was frequently used can be attributed to the semantic symbol "USE.ACT", often infered (about 10%) by the analysis system. (Note: USE.ACT covers "verwenden (use)", "anwenden (apply)", "Gebrauch machen (make use of)", but also <instrument> are for "mit (with)", "mit Hilfe von (with the help of)", etc). This means that the explication of USE.ACT of an implied meaning in the original Japanese text may either elucidate the situation in German (this is often the case) or make expression harder. By the same token a postpositional phrase or adjective phrase of an original text may awkwardly be expressed in a German relative clause. As the modifier and USE.ACT cases above mentioned, exemplify the situation, the over analysis and over-expression are specific to our semantic based approach and could be avoided in other transfer approaches.

IV. Conclusion

We discussed some problems of our semantic based approach. Many of them are also common to other aproaches. However, our approach seems to be open for continuous improvement in dealing with these problems.

We express our sincere thanks to the ATLAS/II group of Fujitsu Laboratory, Kawasaki for making semantic representations available for our generation

Reference

[1]Laubsch,J.,D.Roesner,A.Lesniewski,Hanakata,K.: "Language generation from conceptual structure: Synthesis of German in a Japanese/German MT project", in COLING-84, Stanford, 1984

[2]Roesner,D., Hanakata,K.: "When Mariko talks to Siegfried"; submitted to COLING-85 Bonn, 1985

[3]Emele,M.,Momma,St.:"SUTRA-S; Erweiterungen eines Generator-Front-Ends fuer das SEMSYN Projekt, Studienarbeit, Inst.f. Informatik, Univ. Stuttgart, 1985

[4]Buseman,S.: "Oberflaechentransformationen bei der automatischen Generierung geschriebener deutscher Sprache", Diplomarbeit Univ. Hamburg, Fachb. Informatik, 1983