MT at the Paragraph Level: Improving English Synthesis in SYSTRAN

Eduard Hovy

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
310-822-1511 ext 731
hovy@isi.edu

Laurie Gerber

SYSTRAN Systems Inc.
7855 Fay Avenue, Suite 300
La Jolla, CA 92037
619-459-6700 ext 119
lgerber@systransoft.com

Abstract. In Machine Translation (MT), output quality can be seriously affected by multi-sentence and multi-clause phenomena such as pronominalization, multisentence quotation, comma placement, etc. Yet almost all MT systems operate sentence-by-sentence. This paper describes the transfer of some research on paragraph structure from the research laboratory (USC/ISI) to commercial practice (SYSTRAN). It outlines the definition, implementation, incorporation, and evaluation of a paragraph tree and associated rules to handle the misplacement of quote marks over multiple sentences in Japanese-to-English translation.

1 Introduction

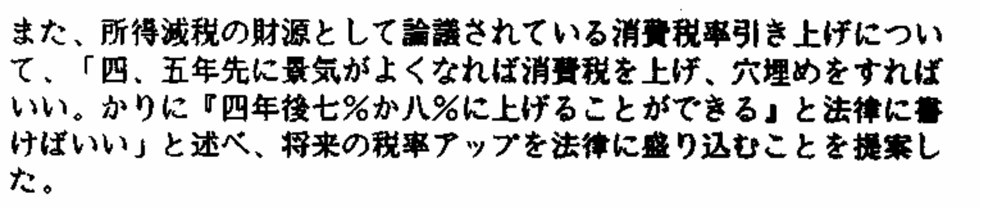
When the linguistic coverage and robustness of an MT system is satisfactory, the attention of the developer turns naturally toward improving output quality. And unfortunately there is always room for improvement—even for one of the oldest and most tested production systems, SYSTRAN.

With mature systems, quality problems often arise from the synthesis (generation) stage. This stems from the curious fact that MT systems overwhelmingly place most effort in the analysis stage, while human translators invariably prefer greater competence in synthesis. Once a system can achieve a halfway decent level of analysis, then the quality-determining problems begin. However, sophisticated synthesis is precisely where most systems fall short. It is too easy to produce a just-adequate sentence generator! This is a tragedy for generation research.

A component targeted for quality improvement in SYSTRAN is English synthesis in the Japanese-to-English system. This component tends to be underdeveloped compared to its source language analysis (whether measured in lines of code or in duration of developer effort). For language pairs with even moderately similar source language structures this imbalance has not proved to be an insurmountable problem. However, the structural differences between Japanese and English, and the underspecification in Japanese of items required for grammatical English, can result in hard-to-read output, no matter how successful the analysis and transfer stages have been. Indeed, the problems resulting from such disparity between analysis and synthesis can be observed in the output of most J-E MT systems. What is required is the ability to depart from source language sentence structure and 'rebuild' sentences according to English grammatical and pragmatic constraints, performing revision primarily on the English level to ensure that the output conforms to English norms.

One area that cries out for attention in MT is multisentence (discourse) phenomena. Almost all MT systems, production and commercial, work one sentence at a time. Typically, their only cross-sentence record is a list of referents for pronominalization. Yet many phenomena in language span sentence boundaries. And recent developments in Text Linguistics, Discourse Study, and computational text planning have led to theories and techniques that are potentially of great importance to MT systems.

As is apparent from Figure 1, an example of real Japanese input and raw SYSTRAN output, several multisentence phenomena have to be addressed.



また、所得減税の財源として論議されている消費税率引き上げについて、「四、五年先に景気がよくなれば消費税を上げ、穴埋めをすればいい。かりに『四年後七％か八％に上げることができる』と法律に書けばいい」と述べ、将来の税率アップを法律に盛り込むことを提案した。

In addition, if " 4 and 5 years first business becomes good, as revenue source of income tax reduction concerning the consumption tariff pulling up which is discussed, it increases consumption tax, it should have filled a gap. Temporarily " it is possible 4 years later to increase 7% or 8%, that " you should have written on law, " that you expressed, you proposed that future tariff rise is included in lav.

Figure 1. Japanese input and raw SYSTRAN output.

While humans produce each new sentence properly embedded in the context of the ongoing discourse and situation, sentence-by-sentence MT causes such problems as:

1. *Erroneous quotation scoping:* In a direct quote in Japanese, the reporting verb of the sentence (the main clause) follows the quote itself (the dependent clause), while in English it normally precedes the quote. Inverting the main and dependent clauses is manageable when the quote is a single sentence, but when it spans multiple sentences, the system currently has no way to determine at which sentence the quote began, and is hence incapable of placing the main clause correctly. As a result, and as seen in Figure 1, quoted multisentence text is translated very oddly by the SYSTRAN J-E system.
2. *Inadequate pronominalization:* The system cannot know what personal pronoun ("he", "she", or "it" to use when its referent lies in an earlier sentence. This problem occurs especially often in J-E translation since Japanese frequently omits sentence subjects; when the system attempts to create and insert a pronoun in the English it has no knowledge of previously introduced referents and hence has no alternative but to guess a pronoun.
3. *Inappropriate comma insertion:* SYSTRAN's synthesis module currently contains a set of rules that govern the insertion of commas into the final English text. Even a cursory analysis of typical output shows that these rules do not operate adequately. One reason is that comma placement in English is partially prosodic, based on the rhythm and balance of clauses in the text; without knowing the length and internal structure of the paragraph, comma insertion rules have no way of determining appropriate placement points.
4. *Incorrect relative pronoun selection:* The choice of relative pronoun ("that", "in which", "which", "to whom", etc.) is not always trivial, and the behavior of the current synthesis rules in the system reflect that fact. Since relative pronouns refer to entities outside of the relative clause, rules for proper pronoun usage must be able to locate and inspect the appropriate referent.

These problems can only be solved if SYSTRAN'S transfer rules are given the ability to work across sentence boundaries. Thus, our project involves the definition and implementation of a paragraph structure and associated code to reduce or solve these problems. Intentionally, the paragraph structure is kept very simple, and is allowed to evolve only as needed to solve new problems. Emphasis of our work is on practical, measurable success, instead of on purely theoretical research considerations.

2 The Collaboration

SYSTRAN Software, Inc. (SSI):
SSI's system development process is highly refined and efficient, making it possible to quickly and efficiently build and upgrade the capabilities of MT systems. SSI has introduced many evolutionary improvements over the years, but historically most activities and processes have necessarily been oriented toward production.

Recently, SSI has placed a high priority on focused research that will lead to the introduction of new techniques and tools. A partnership with a research group is an ideal avenue for this. If successful, a collaboration would provide improved output quality for the J-E system, with the further possibility of implementing the new technology also in other language pairs. The collaboration is thus a vehicle for SYSTRAN to introduce cutting-edge technology into its system. In addition, it is a mechanism by which SYSTRAN personnel can learn more about the latest technical developments in the research sector.

USC/ISI:
Over the past two decades, USC/ISI has performed research on sentence generation, text and sentence planning, and (lately) Machine Translation. The Penman sentence generation system [Penman Group 89] has served as a benchmark generator since the mid-1980's; multisentence text planning techniques have been developed; and the JAPANGLOSS MT system, which performs Japanese-to-English translation of open-domain texts using an advanced hybridization of statistical and linguistic/symbolic techniques [Knight et al. 94, Yamada 96], is currently serving as the model for new Arabic-to-English and Spanish-to-English systems.

At the time of writing, several components have achieved significant levels of capability (JAPANGLOSS has interested two commercial endeavors). However, research and development continues, and the acid test—full operational deployment—still awaits. Therefore, the NLP team at USC/ISI is eager to participate in a collaboration with an MT company: not only does this afford the chance to test some modules and ideas in real-world settings, but the success of such an endeavour will help demonstrate the validity of new techniques beyond the research level. Such success will hopefully work beneficially for both MT theory and MT practice.

Finally, a successful collaboration will help justify the financial 'rat-hole' of MT research.

3 Related Work

Given the impact of multisentence phenomena on text quality, it is astounding that almost no literature exists on the subject in MT. The closest work we know of is the research on text planning within Natural Language Generation. Much of this work is devoted to developing discourse (paragraph-length and longer) structures. Rhetorical Structure Theory [Mann & Thompson 88] defined a set of approximately 25 relations that govern how adjacent clauses (and blocks of clauses) are related to each other in coherent text. These relations include ELABORATION, CAUSE, JUSTIFY, and MOTIVATE; many of them are signaled in English with a cue word or phrase (e.g., “for example”, “in order to”, “finally”).

In subsequent work, [Hovy & Maier 93] collected over 300 relations from numerous sources into a taxonomy of approx. 120 relations. Such relations are used in automated text planning systems such as [Hovy 88, Moore 89] to build the tree structure that represents the coherence organization of the clauses of text. Once such a tree structure has been built for a text, it is possible to perform a series of sentence planning tasks such as aggregation (to remove redundancy), sentence length determination, internal sentence organization of relative clauses, pronominalization, and so on, as described in [Nirenburg et al. 89, Rambow & Korelsky 92, Panaget 94, Dalianis 96, Hovy & Wanner 96].

With regard to SYSTRAN's J-E synthesis, it is not necessary to implement such complex text structures in order to realize improvements in the problem areas mentioned above, A much simpler paragraph structure is adequate (and provides a basis out of which a more complex structure can be developed). We call this structure the *paragraph tree.*

4 Theory and Implementation

4.1 The Paragraph Tree for Multisentence Quotes

We define an extremely simple paragraph structure in order to represent the primary groupings of clauses and sentences within paragraphs as follows:

> **Definition:** A *paragraph tree* is a tree of nodes, in which the top node (level 0) represents the whole paragraph, each node at level 1 represents a single multisentence grouping of interest, and each leaf node represents a single sentence.

A simple paragraph tree is shown at left in Figure 2.

```
             PAR                                PAR
     _____/ | | | \_____                ____/ | | \ ____
    /    __ / | \___     \            /   ____| |___   \
   /    /     |     \     \          /   /         \    \
SENT1 SENT2 SENT3 SENT4 SENT5      SENT1 QUOTE   SENT5 SENT6
                                         / | \
                                        /  |  \
                                    SENT2 SENT3 SENT4
```

Figure 2. Simple paragraph tree (left) and tree containing multisentence quote (right).

Our general approach to multisentence problems is as follows. For each grouping of sentences of interest to us, we create a new type of node for the paragraph tree. Thus, for example, we group all clauses contained within a multisentence quote under a QUOTE node, as shown on the right in Figure 2.

We first implemented a module to construct simple paragraph trees from the data structures built up in the SYSTRAN J-E system. We then extended the module to be able to add internal layers to the tree, and tested this by addressing the erroneous quotation scoping problem. The new internal node binds together any sequence of quoted sentences and hence enables transfer rules to identify the correct English (pre-quote) position of the reporting verb (the "said" of a "he said" portion) of a Japanese quote, and to move it there. This problem arises because sentence-at-a-time processing has no way of determining where the reporting verb should be moved to (let alone of moving it there!), as shown schematically by:

Example input (Japanese word order):

...XXXXXXXXX. XXXX. He "XXXXXX. XXXXXXXX. XXXX" announced. XXXXX....

Desired output:

...XXXXXXXXX. XXXX. He announced "XXXXXX. XXXXXXXX. XXXX". XXXXX....

Sentence-at-a-time output:

...XXXXXXXXX. XXXX. He "XXXXXX. XXXXXXXX. ?? announced XXXX". XXXXX....

4.2 Implementation

The construction of the paragraph tree and the manipulation of quotes required operations and rules of the following kind. Note that these rules apply also to sentence-internal quotes:

0. Operations:

```
1. create-node (sentence)            2. subordinate-node (newnode parent)
3. return-active-node ()             4. set-active-node (node)
5. find-parent (node)                6. find-first-child (node)
7. find-next-child (parent node)     8. find-last-child (node)
9. add-information (info node position)
```

1. Paragraph tree building rules:

```
- Build rule 0:
  a. at each paragraph break, create new top node PAR
  b. set it to be the active node
- Build rule 1 (default):
  a. at each new sentence, create a new sentence node
  b. subordinate it to the active node
```

2. Quotation tree rules:

```
- Build rule 2 (quotation start):
  a. on encountering an open quote, create a new intermediate node QUOTE
  b. subordinate it to the active node
  c. set it to be the active node
  d. perform build rule 1
- Build rule 3 (quotation end):
  a. on encountering a close quote, reset the active node to the parent
     of the active node
  b. perform build rule 1
```

3. J-E transfer rules:

```
- Transfer rule 1: locate the reporting verb (and adjuncts) after the
  close quote in the final child of a QUOTE node
- Transfer rule 2: move it to just before the open quote in the first
  child of the QUOTE node
```

The paragraph tree builder was implemented as a standalone module, separate from the existing SYSTRAN J-E system. The major change required to the existing system was one function, activated after completed processing of each sentence, to save the results (intermediate data structures and final output string) in a growing record. When the system encounters a paragraph break, the new module traverses the saved data structures, creates the paragraph tree, performs the final transfer, and re-synthesizes.

The raw and treated outputs for a short example are shown in Figures 3 and 4.

5 Evaluation

Overall, the initial experiments proved very encouraging. We collected a set of 57 paragraphs containing one or more quotes (of various types: multi-sentence, single-sentence, or simple noun phrase) each. With not much effort, using approx. 10 rules, it was possible to treat the various types of quotes correctly in about 66% of cases. (Correct treatment means different things in different contexts: no change for an NP quote, but inversion for a single- and multi-sentence quote.) Correctly inverted sentences showed clear improvement, as illustrated in Figure 4. In only about 5% of cases were quotes treated incorrectly; the remainder were quotes simply missed by the rules, requiring additional rule writing.

Unfortunately, due to the fact that the paragraphs suffer from many problems, the improvement of quote handling alone did not result in significantly improved evaluation scores. To evaluate the results, we asked 8 native English speakers unaffiliated with the project but familiar with MT in general to score the outputs. From 36 texts we extracted the quoted passages, in both raw and treated form. Each scorer scored half raw and half treated passages, in random order; no person saw the same passage twice. Each passage was scored by 4 people.

Scorers were asked to rate the fluency of each text on a five-point scale:

```
How fluent is the following text?
Assign one score:
  1. Perfect—absolutely standard English
  2. Not perfect, but not so bad
  3. Problematic, but partly understandable
  4. Understandable in places, but really not acceptable
  5. Garbage—totally incomprehensible
```

The results are as follows:

Raw scores: Total: 491 Average: 13.64 Ave/text: 3.41
Treated scores: Total: 468 Average: 13.00 Ave/text: 3.25
Difference: 23 Average diff/text: 0.160
Std Dev: Raw scores: 0.927 Treated scores: 0.942 Overall: 0.935
Variance: 0.874
Confidence at 0.1: Raw scores: 0.254 Treated scores: 0.258 All scores: 0.256

6 Conclusion

We are very encouraged by the ease with which the paragraph tree module was implemented and added to the system, and by the success of the multisentence quote experiment. This is however only the beginning—the more common problems remain, and we plan to address pronominalization, comma insertion, and relative pronouns next. Should this series of experiments prove

```
At the time of opening as for President Kojima "the indefinite element...
stops being enough wants doing," that it said, it made that ordinary become
expectation and interest are moved aside. In addition, also President Gordon
Ladry...runs to the celebration and "the place where the media laboratory
makes idea and dream expand, raises imaginative power. It can become the
treasure house of reform and progress. Also by all means succeeding, we
want turning to center of the Japanese image media," that you cheered of
encouragement.

X Root
   S At the time of opening as for President Kojima
   Q "<
      S the indefinite element... stops being enough wants doing,
   T >"
   P that
   B it
   S said,
   M
   B it made that ordinary...are moved aside.
   S In addition, also President Gordon Ladry...runs to the celebration and
   Q "<
      S the place where—raises imaginative power.
      B It can become the treasure house of reform and progress.
      S Also by all means succeeding...Japanese image media,
   T >"
   P that
   B you cheered of encouragement
   Z .
```

Figure 3. Raw output and associated paragraph tree. Portions elided for space.

```
At the time of opening as for President Kojima said, "the indefinite element
...stops being enough wants doing," it made that ordinary become expectation
and interest are moved aside. In addition, also President Gordon Ladry...
runs to the celebration and you cheered of encouragement "the place where
the media laboratory makes idea and dream expand, raises imaginative power.
It can become the treasure house of reform and progress. Also by all means
succeeding, we want turning to center of the Japanese image media".

X Root
   S At the time of opening as for President Kojima
   S said,
   B
   Q "<
      S the indefinite element ... stops being enough wants doing,
   T >"
   B it made that ordinary...are moved aside.
   S In addition, also President Gordon Ladry...runs to the celebration and
   B you cheered of encouragement
   Q "<
      S the place where...raises imaginative power.
      B It can become the treasure house of reform and progress.
      S Also by all means succeeding, we want... Japanese image media,
   T >"
   Z .
```

Figure 4. Treated output and associated paragraph tree. Portions elided for space.

successful, it is natural to consider placing the paragraph module inline. But even with a closer coupling, in which the paragraph module is situated immediately prior to SYSTRAN'S synthesis stage, the only major effect would be to add appropriate instructions to the intermediate data structures, using existing SYSTRAN data fields and codes, that cause the English synthesis module to generate appropriate language.

We hope that the multisentence planning module will evolve to assume increasing importance in the SYSTRAN J-E system, and eventually also in its other language pairs as well. The stepwise development of the multisentence module will be mirrored by a staged incremental development of the paragraph tree itself: its internal node types, data structures, and average depth. Information will only be introduced into the tree if it is directly relevant to rules that affect the output; in this sense, the multisentence module will act as a proving ground for research ideas in the general text planning community.

This interesting and exciting project simultaneously provides an opportunity to validate theoretical research, improve commercial performance, and establish effective technology transfer from research to industry.

Acknowledgments

Thanks to Alan Nash of SSI for implementation and discussions.

References

[Dalianis 96] Dalianis, H. 1996. *Aggregation in Natural Language Generation.* Ph.D. dissertation, Stockholm University.

[Hovy 88] Hovy, E.H. 1988. Planning Coherent Multisentential Text. *Proceedings of the 26th Annual Conference of the Association for Computational Linguistics,* Buffalo, NY (163-169).

[Hovy &: Maier 93] Hovy, E.H. and E. Maier. 1993. Parsimonious or Profligate: How Many and Which Discourse Structure Relations? Unpublished ms.

[Hovy &: Wanner 96] Hovy, E.H. and L. Wanner. 1996. Managing Sentence Planning Requirements. In *Proceedings of Workshop on New Directions in Planning and Natural Language Generation* at the 12th European Conference on Artificial Intelligence, Budapest, Hungary (12-18).

[Knight et al. 94] Knight, K., I. Chander, M. Haines, V. Hatzivassiloglou, E.H. Hovy, M. Iida, S. Luk, A. Okumura, R. Whitney, K. Yamada. 1994. Integrating Knowledge Bases and Statistics in MT. In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas,* Columbia, MD (134-141).

[Mann &: Thompson 88] Mann, W.C. and S.A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3) (243-281). Also available as USC/Information Sciences Institute Research Report RR-87-190.

[Moore 89] Moore, J.D. 1989. *A Reactive Approach to Explanation in Expert and Advice-Giving Systems.* Ph.D. dissertation, University of California in Los Angeles.

[Nirenburg et al. 89] Nirenburg, S., V. Lesser, and E. Nyberg. 1989. Controlling a Language Generation Planner. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence,* Detroit, MI (1524-1530).

[Panaget 94] Panaget, F. 1994. Using a Textual Representation Level Component in the Context of Discourse or Dialogue Generation. In *Proceedings of the 7th International Workshop on Natural Language Generation,* Kennebunkport, ME (35-42).

[Penman Group 89] Penman Natural Language Group. 1989. The Penman Primer, User Guide, and Reference Manual. USC/Information Sciences Institute, Marina del Rey, CA.

[Rambow & Korelsky 92] Rambow, O. and T. Korelsky. 1992. Applied Text Generation. In *Proceedings of the 3rd Applied Natural Language Processing Conference,* Trento, Italy (40-47).

[Yamada 96] Yamada., K. 1996. A Controlled Skip Parser. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas,* Montreal, Quebec (14-23).