HOW COULD RHETORICAL RELATIONS BE USED IN MACHINE TRANSLATION? (AND AT LEAST TWO OPEN QUESTIONS)

Ruslan Mitkov

Machine Translation Unit
University of Science Malaysia
11800 Penang, Malaysia
Fax (60-4) 873335,
Email ruslan@cs.usm.my

My position paper addresses more or less Workshop question No. 5: "How are rhetorical relations used in discourse understanding? How are linguistic clues and world knowledge brought to bear?"

The paper shows how rhetorical relations could be used in Machine Translation (MT). It introduces in brief, a discourse-oriented approach for MT which uses schemata of rhetorical predicates for describing the structure of a paragraph. At the same time, it poses at least two questions (in my opinion practically unsolved problems):

1) How can rhetorical predicates be computationally recognized?
2) Are the so far defined predicates sufficient and precise enough to describe the real world?

INTRODUCTION: DISCOURSE-ORIENTED MACHINE TRANSLATION

The discourse-oriented MT should be regarded as a very important research topic, since it is expected to make the translation more natural in MT systems. Unfortunately, not much attention has been given to this problem yet and the availability of a discourse component in a MT system has been reported very briefly in [7] only .

Most of the MT systems perform sentence-by-sentence translation. Only a few try to translate paragraph-by-paragraph and in these cases, the discourse structure of the output language is identical with that of the input language. However, I have shown that the discourse structures across the different sublanguages are not always the same for any pair of natural languages [5].

Paragraph-by-paragraph machine translation seems to be for now, an unjustifiably complicated task for practical needs. It involves the complete understanding of the paragraph, the determination of discourse topic(s), goals, intentions, so that the output can be produced in accordance with the respective discourse rules and purposes. However, recognizing topic, goal, intention by a computer program seems to be a very tough problem. Moreover, analyzing a paragraph is itself a complicated task which does not always yield satisfactory results.

On the other hand, translating sentence-by-sentence with the sequence of the original sentences preserved is a general approach, which guarantees in most of the cases an understandable output. However, in order for a translated message to sound as natural as possible, it should be conveyed in accordance with the discourse organization rules of the target language. If we examine more closely the work of a professional translator, we shall inevitably note that he/she does not always follow the order of sentences in the source text.

Taking into account the complexity of paragraph understanding and the necessity of observing the specific target sublanguage rules, I have been working on a practical discourse-oriented MT approach (within an English to Malay MT system) which analyzes a source paragraph as a schema of rhetorical predicates and generates the target text possibly as another schema of rhetorical predicates. Towards this end, I have developed a Text Organization Framework Grammar which maps source paragraph structures of rhetorical predicates into the specific target paragraph structures of rhetorical predicates [6].

SELECTION OF TEXT ORGANIZATION APPROACH

I have been studying different approaches which have been so far used to describe the organization of a given text (paragraph). From a practical point of view, I argue that the most appropriate approach would be the "schemata-based approach" introduced by K. McKeown [3] and used by other researchers .

Though some researchers point out the relatively missing flexibility of this approach, I found this approach more suitable for the needs of MT. The plan-based approach [4] seems to be too complicated and unrealistic to be implemented in an MT system because its rhetorical relations are dependent on an expected effect on the hearer achieved by their combination. In a MT system, as already mentioned, it is very hard, if practically not possible, to recognize automatically in a paragraph the goals and intentions of the speaker.

SUBLANGUAGES AND SCHEMATA

In the sublanguages I studied, however, I found out that the schemata of rhetorical predicates could not be always uniquely defined. There are sublanguages where more than one typical schema should be defined and consequently used. I examined numerous texts on which basis I defined "stable schemata". The schemata S_1, S_2, .., S_N can be considered "stable" if 1) $S_I/N \geq \delta$, $\forall I$ and 2) $\Sigma S_I/N \geq \gamma$ where N is the number of all examined texts, δ, γ are numbers in the interval (0,1) which we call "individual contribution minimum" and "global contribution minimum" respectively. The idea behind these mathematics is that schemata can be considered as "stable" if they as a whole represent a significant portion of all examined texts and yet every "stable" schema should be itself representative.

For translation from English into Malay, if more than one stable schema is available in the respective sublanguage, the stable schema, which is closest to the input of English text is chosen. For determining closeness, special metrics has been developed which takes into account not only the number of displaced predicates, but also the size of the displacement and the maximal length of matched substrings from the input and output schemata of rhetorical predicates.

We have studied the discourse structure of a few sublanguages (for both English and Malay), potential candidates for translation domains in our MT system: the

sublanguages of job vacancies, residential properties for sale, cars for sale and education advertisements from different newspapers in English and Malay.

From our investigations on these sublanguages we have drawn three main conclusions:

1) The stable schemata for English and Malay are *not always identical and do not occur equally frequent*
2) For some sublanguages there are more than *one stable schema*
3) For some sublanguages there exists *no stable schema*

These conclusions are important for MT because in the third case there is no need for discourse transition rules and the translation should be undertaken sentence-by-sentence.

THE BIG PROBLEM: IDENTIFICATION OF RHETORICAL PREDICATES

During the analysis, rhetorical predicates should be recognized. In certain sublanguages this can be done by means of key words and other clues [5]. However, in general this seems to be a very complicated problem and extensive world knowledge and inferencing mechanisms are needed. How could a program recognize a sentence (proposition) as amplification, attributive, etc. rhetorical predicate? For our sublanguage-based MT needs, I am considering two approaches for the identification of rhetorical predicates.

One approach would be to define "verb frameworks" characteristic of a verb within the sublanguage. Each verb should be associated with possible rhetorical predicates and the predicate should be identified on the basis of the logical structure of the analysis. However, this approach may not be powerful enough in certain cases. Consider the sample text from [2] describing Kyushu Daigaku (Kyushu University):

> "A national, coeducational university in the city of Fukuoka. Founded in 1910 as Kyushu Imperial University. It maintains faculties of letters, education, law, economics, science, medicine, dentistry, pharmacology, engineering, and agriculture. Research institutes include the following: the Research Institute of Balneotherapeutics, the Research Institute of Applied Mechanics, the Research Institute of Industry and labor, and the Research Institute of Industrial Science. Enrollment was 9,425 in 1980".

It will be quite difficult, however, using only verb framework, to recognize the first, the third, fourth and the last sentence as rhetorical predicates. An useful approach in this case would be to use a domain knowledge which would enable the recognition of the rhetorical predicate after a semantic analysis. For instance a proposition describing entities which are in 'sub-part' relation should be classified as a constituency predicate. This 'sub-part' relation could be easily recognized, provided it has been already described in the domain knowledge base. Consider again the sample text under the assumption that such a knowledge base exists. In this case, from the 'is-a' relation ("Kyushu Daigaku"- "University"), from the respective 'sub-part relations' ("university"-"faculty", "research centre") and the 'has' relation ("university" -"enrollment of students"), the program could assign to the above sentences identification (1. sentence), constituency (3., 4. sentences) and attributive (last sentence) predicates, respectively.

Consider, however, the second sentence. Is it "amplification"? If yes, how is the program supposed to conclude that this sentence is an elaboration of the first one? How feasible is in general the computational recognition of the rhetorical

predicates? And here comes an important question: how much domain and world knowledge, as well as AI inferencing techniques, are needed?

And if yes, does not it seem that "amplification" is not fine and precise enough (I can give many examples of propositions to which the rhetorical predicate "amplification" is to be assigned, because they simply do not fit the definition of the rest of the predicates)? Should not one introduce an additional predicate called e.g. "initiation" which would be associated with the act of founding, setting up, opening, organizing etc. something? This gives a rise to a second important question. Is the set of rhetorical predicates given in [1] , [3] , [8] or [9] sufficient and precise enough to describe the real world? But if we propose additional predicates, how far should we go?

REFERENCES

[1] Grimes J. - The thread of discourse. Mouton, The Hague, 1975

[2] Kodansha Encyclopedia of Japan, Vol.4, Kodansha Ltd., Tokyo, 1983

[3] McKeown K. - Text generation: using discourse strategies and focus constraints to generate natural language text. Cambridge University Press, 1985

[4] Mann W., Thompson S. - Rhetorical structure theory: description and construction of text structures. In Kempen G. (Ed.): "Natural language generation: new results in artificial intelligence, psychology and linguistics", Dodrecht, Boston, 1987

[5] Mitkov R. - Multilingual generation of public weather forecasts, Proceedings of the SPICIS'92 (Singapore International Conference on Intelligent Systems) Conference, 28 September-1 October 1992, Singapore

[6] Mitkov R. - Discourse-based approach in machine translation From Proceedings of the International Symposium on Natural Language Understanding and Artificial Intelligence, Fukuoka, Japan, 13-15 July, 1992

[7] Nirenburg S. - A distributed generation system for Machine Translation: Background, Design, Architecture and Knowledge Structures, CMU-CMT-87-102, Pittsburg, 1987

[8] Shepherd H. - The fine art of writing. Macmillan Co., New York, 1926

[9] Williams W. - Composition and rhetoric. D.C. Heath and Co., Boston, 1983