Machine Translation Viewed as Generation of Text with a Pre-defined Contents

A. V. ZUBOV

Minsk State Pedagogical Institute
of Foreign Languages,
ul Zakharova 21,
220662 Minsk, USSR

Proposes a way for organising a system of machine translation (MT) based on results obtained in linguistic investigations of text. It includes modelling of three stages in the process of human text translation: comprehension of the original, its interpretation and re-expression. The contents of the text (in a particular field of knowledge) under translation is set automatically in the form of a statistically compiled table of the main contents (TMC) and a sequence of paragraphs of definite logical contents. In the process of the MT system operation, each paragraph of the source text is presented with regard for probability and algorithmic rules, by a possible semantic-syntactic formula (SSF). A sequence of such formulae makes a semantic-syntactic translation formula. Positions of the latter are filled in with translation equivalents of TMC units chosen from the automatic dictionary, special concretising words with which each SSF is supplied, and units from four lists of two-word combinations of autosemantic words specific for the given field of knowledge.

The process of translation from one language to another, as certain researchers point out, consists of three stages: 1) to comprehend the original, 2) to interpret it, 3) to reformulate it [1, p. 59—88].

If one deals only with scientific and technical texts, the first stage can be reduced to word-for-word (i. e. philological) understanding of the original and to comprehension of stylistic factors of language expression.

The main task of interpreting the original is to determine the objective sense of the work, i. e. the reality described in the work should be conveyed without regard to the interpreter's likings and dislikings.

The original is reformulated according to the peculiarities of the corresponding language systems, to the linguistic peculiarities in the original, and to some other operations.

If one analyses existing machine translation (MT) strategies, one can see that they basically reduce comprehension of the original to an elementary philological understanding through a bilingual automatic dictionary (AD), with style factors being considered only in special cases.

MT so far completely lacks an interpretation stage of the original. In describing the situation conveyed by the translated text, existing MT systems do not take into account real relations between AD words.

And lastly, in operating MT systems, the translation stage proper differs principally from human reformulation of text. A human translator relies on the contents of the whole text and is aware of the specifics of the source and target languages, while the computer merely translates the text phrase by phrase.

Bearing in mind the aforesaid and also the fact that MT does not need to consider stylistic peculiarities of texts, one may discern two directions in research aimed at improving MT systems:

(1) Investigation of methods of relating AD words to information on relationships between objects, phenomena, and facts, conveyed by a certain text.

(2) Investigation of a general semantic formula of, the text and ways of developing it into a logical sequence of smaller semantic text units.

A possible approach to these problems is based on recent results in linguistic text research.

Any fragment of reality is some set of objects and facts interlinked by certain relations. Obviously a text describing that fragment of reality should contain two types of constituents: names and name combinations denoting objects, phenomena and facts, and some units reflecting interrelations of names in the given real situation being described.

Let us take a good look at what these constituents are from the viewpoint of text organisation as a unified whole.

Names and name combinations as substitutes for objects and events of objective reality have different values for expressing the contents of the text. Some of them represent the main contents of the text — they are semantic 'milestones', 'supports', 'piles', 'backbone' of the text. As a rule, these are the most frequent words and word combinations present in the prevailing number of text paragraphs. We will call them the main supporting units (MSU) of the text. A second group of supporting words goes along with the MSU, minutely describing the situation created by the MSU denotates. These are also names and name combinations of sufficiently high frequency in some text paragraphs [2, p. 176—178], Let us call them secondary supporting units (SSU).

A combination of MSU and SSU groups makes up the table of main text contents. The table can be built up automatically by means of, for example, altering values of the *K* coefficient calculated in the following way:

$$K = \frac{m \bullet F}{n \bullet N},$$

where *F* is the word (combination) frequency in the text, *m* is the number of paragraphs containing the word (combination), *n* is the total number of paragraphs in the text, and *N* is the total number of word usages in the text.

However MSU and SSU denotates can act within the limits of events described in a certain text at different times and places, with different reasons and aims. They are specified in the text paragraphs. Let us assume that each paragraph of a certain text contains a subset of words corresponding to a definite micro-situation. We can define the composition and structure of this word set if we assume that this micro-situation is not an indivisible unit, and that it consists of several elementary constituent situations (ECS). Analysis of different text types has delineated such ECS as 'state', 'action', 'feature', 'detail', 'place', 'time', 'means', 'cause' and so on. Thus, the specificity of each micro-situation can be defined by a certain ECS sequence and by filling each ECS individually with corresponding denotates. According to the aforesaid, while studying different texts each paragraph is assigned an ordered sequence of subgroups of specifier-words representing these denotates*. As a rule subgroups consist of verbs, participles and some nouns.

Words representing ECS build up to some extent the backbone of the text contents defined by the table of main text contents, but they do not complete its construction. The reason for this is that, besides supporting and situative words, there exists in any text one more word group called filler-words. They fill the constructed contents with qualitative features, i. e. they designate qualities of objects named by nouns or features of actions defined by verbs and participles in the text. This qualitative specification is done through preliminary fixing of free word combinations of a certain type available in the text: verbal-nominal, attributive, adverbial and nominal ones. As is stated elsewhere, the combinations are the main material for text (not phrase!) construction [4—6].

Now we pass on to the second type of text constituents expressing relations between names, and we ought to cite Professor V. G. Admoni's lines that '... some relations in objective reality which are most vital and important and systematically repeated (including those in human social practice and mental life), tend to fix themselves in the language not only lexically but in the form of syntactic structures, too. Thus these relations become potential practical logic universals implemented in syntactic structures of various languages of the world' [7, p. 3]. The same was expressed by A. M. Peshkovsky back in 1920: 'We all speak in certain cliches, use certain forms of combinations acquired in childhood with words and sounds of the given language and inherited by our generation from previous generations. These cliches ... always come to mind no matter what we might be talking about or listening to. That is our syntactic luggage which we have been carrying in our lifetime since childhood, as well as sound luggage, lexical luggage, semasiologic luggage ... which jointly make up what is called the Russian language. We fetch these cliches from our luggage and dress our idea up in them in a kind of lingual clothing everytime we have to say something. The more commonplace the cliche, the more we are accustomed to it, the greater the chance it will turn up right when we need it ...' [8, p. 427—428].

Many researchers think such cliches, speech patterns', and 'logical-syntactic structures' are psychic realities existing in our conscience together with units of other language levels participating in text construction [7, 9-11].

At the same time, modern psychology and psycholinguistics state that 'an idea is not necessarily verbally concretised into one sentence. A speech action can include several sentences which are somehow interdependent' [12, p. 146],

M. M. Bakhtin, a well-known specialist on the aesthetics of speech creation, has written: 'We speak only in definite speech styles, i. e. all our statements possess definite and relatively stable forms of constructing the whole. We possess a wide repertory of oral (and written) speech styles. In practice we use them safely and confidently but in theory we might be oblivious of their existence altogether ... Even in the freest, most casual conversation we cast our speech in definite stylistic forms which sometimes are conventional and trite, sometimes — more flexible, plastic, and creative ... We acquire language forms only through statement forms and only with these forms ... To learn to speak is to learn to make up statements (because we speak by statements and not by separate phrases...)' [13, p. 257—259].

All the quoted scientific opinions testify to the necessity of singling out in the text a unit larger than a phrase and possessing a certain semantic integrity and expressing a complex and complete idea. In text linguistics, this unit is called a superphrase unity of a complex syntactic entity. However, it has virtually no semantic nor formal borders. When dealing with written speech processing, one can put the text paragraph to correspond to the superphrase entity.

As certain linguistic research and psychological experimental work show, authors are not arbitrary in uniting groups of phrases in paragraphs. Dividing the text into paragraphs is dependent on semantic contents along with other factors. An author divides material into parts of a relatively complete sense. For example, subjects in one experiment who were told to divide a continuous text into paragraphs insisted that they divided it into topical parts. Moreover, these same experiments confirmed that, as a rule, parts which are singled out coincide with paragraphs [14—16].

Thus a paragraph can be viewed as a sort of fixed syntactic cliche reflecting relations between objects of reality, but then again, also as a part, a quantum of the general meaning of the text.

Scientists of different inclinations of thought have voiced the hypothesis that for each type of scientific and technical text (translated with computer aid) the number of paragraph types describing this or that fragment of reality will be small and finite.

* Actually this is but encyclopedic information which current ADs lack so much [3],

Consequently, every scientific and technical text can be represented by a certain sequence of paragraphs with definite subject-logical contents (for example: 'description of object action', 'description of errors in object actions', 'main stages of research description', and so on),

In turn, the semantic contents of each such paragraph can be represented in a special language by a semantic-syntactic formula based on semantic cases analogous to the cases distinguished by C. Fillmore [17, 18]. In addition, there are elements in this language that take into account conventional surface cases. Moreover, a paragraph of some subject-logical contents can be described with several semantic-syntactic formulae.

How is the relation between the basic constituents of the text — the main support words and semantic-syntactic paragraph formulae — brought about? For this purpose the formulae are marked with special signs indicating the support name incorporated in this or that argument of a semantic function.

Hence, any text contents referring to a certain scientific and technical sub-language can be represented by:

(1) The table of main text contents.

(2) Some semantic-syntactic text formula including a sequence of semantic-syntactic paragraph formulae with designated support words.

(3) Lists of specifier words which are singled out by preliminary analysis of a representative corpus of the given sub-language texts, and which accompany every text paragraph.

(4) Lists of pre-selected word combinations typical for the sub-language.

Processing along this pattern a sufficiently representative corpus of different language texts allows for the singling out of the most wide-spread paragraph types according to their subject-logical contents and their subsequent characterising with certain semantic-syntactic formulae and specifier words.

Now let us view machine translation of texts from this viewpoint.

First, a person looks through the foreign scientific or technical text on a certain field of knowledge that is to be translated, and marks each paragraph of the text in the corresponding paragraph type code (according to the prepared table), i. e. he assigns certain subject-logical contents to a paragraph.

Then, using the above formula, the computer statistically selects supporting words in the whole text, and using them, looks for a possible target-language paragraph type for each source-language paragraph type. Using a random number generator, the table of main text contents, and types of previous paragraphs, it selects a possible semantic-syntactic formula for the given paragraph. Finally, a semantic-syntactic formula of the text is arrived at.

The formula is filled with specific vocabulary by successively filling its slots with words and word combinations included in the table of main text contents, in the lists of specifier words assigned to each paragraph, and with words from combinations of the above type. This filling simulates to some extent the peculiarity of human text generation when actualisation of some text elements is necessarily accompanied by automatic choice of other elements referring to the given contents.

What does the procedure yield? Firstly, the established relations between elements of automatic dictionary correspond just to the relations between reality objects described in the source text,

Secondly, translation, i. e. text generation, proceeds along a pre-defined semantic-syntactic formula of the entire text.

And thirdly, grammatic elements in the semantic-syntactic text formula exclude altogether cases of disagreement, omissions and other grammatic drawbacks so frequent in existing MT results.

REFERENCES

1. Levy, J. *Art of translation.* Moscow: Progress Publishers, 1974.
2. Boldak, I. A. On selecting key words in a scientific text. In: *Experimental analysis of oral and written texts.* Minsk: Minsk-GPIIYa, 1981, 176—178.
3. Leontyeva, N. N.; Volkovyskaya, Ye. V, et al. Encyclopedic functions dictionary and its role in automatic indexing. *Nauchno-tekhnicheskaya informatsiya, Ser.* 2, 1978, No. 7, 23—29.
4. Dolgova, O. V. *Syntax as a science on speech construction.* Moscow: Vysshaya Shkola Publishers, 1980.
5. Zilberman, L. I. *Structural-semantic text analysis.* Moscow: Nauka Publishers, 1982.
6. Ter-Minasova, S. G. *Scientific-linguistic and didactic aspects of word combination.* Moscow: Vysshaya Shkola Publishers, 1981.
7. Admoni, V. G. Syntactic semantics is a semantics of syntactic structures. In: *Problems of syntactic semantics. Proceedings of the scientific conference.* Moscow, 1976.
8. Peshkovsky, A. M. *The Russian syntax in scientific presentation.* Moscow: Gosizdat, 1920.
9. Arutyunova, N. D. *The phrase and its sense:* Moscow: Nauka Publishers, 1976.
10. Lekomtsev, Yu. K. The psychic situation, the phrase and the semantic sign. In: *Transactions on semiotic systems. VI. Uchenye Zapiski Tartuskogo Gosudarstvennogo Universiteta.* Tartu, 1973, Vyp. 308, 444—463.
11. Solntsev, V. M. *The language as structural systems formation.* Moscow, 1977.
12. Zarubina, N. D. On psycholinguistic grounds for acceptability of the phrase and the superphrase entity as units in teaching a foreign language. In: *Psychology of grammar.* Moscow: MGU Publishers, 1968.
13. Bakhtin M. M, Speech genre problem. In: Bakhtin, M. M. *Oral creation aesthetics.* Moscow: Iskusstvo Publishers, 1979.
14. Galperin, I. R. *The text as an object of linguistic study.* Moscow: Nauka Publishers, 1981.
15. Serkova, N. I. On internal composition of syntactic unities. *Vestnik MGU. Ser. filologii,* 1967, No. 4.
16. Silman, T. I. *Problems of syntactic stylistics.* Leningrad, 1967.
17. Bogdanov, V. V. *Semantic-syntactic organisation of the phrase.* Leningrad: LGU Publishers, 1977.
18. Fillmore C. A case for case. In: *News from foreign linguistics. Vyp. X. Linguistic semantics.* Moscow: Progress Publishers, 1981, 369—495. (Translated from English).

[*All references are in Russian*]