Session 2: CURRENT RESEARCH

FUNCTIONS REQUIRED OF A TRANSLATION SYSTEM[1]

Gilbert W. King
IBM Research Center

Abstract

The fundamental problem of automatic analysis of grammatical structure and resolution of meaning by machine is as far away from comprehensive solution as programming chess playing. However, the resolution of linguistic difficulties when the clue lies in local context has been solved to a degree by various techniques. For practical purposes there are several other, but essential, processing problems to be solved in a translation system.

It is shown here that these problems, and the local context, have been handled by a single table lookup in a large memory with conditional addressing.

The Operational Point of View

The program at IBM Research has been to examine the question of automatic translation of languages from an operational point of view, rather than as an interesting academic exercise. This has resulted in a description of a system in terms of the functions required, and in estimates of the parameters involved.

The objective is to provide output which will convey sufficient information to attract readers. To this end it is not necessary to make perfect translations.

Our presumption is that high-quality translations will only be achieved in successive stages. One has only to consider sentences such as "cherchez la femme"," 'cat' has a prime number of letters", " 'To whit' said the owl", to suspect that some sentences may never be translated[2] by machines or humans. Nevertheless, it is an empirical fact that information can be conveyed by some kind of rendering of one language into another. The amount of information

---

1 This work has been largely supported by the Intelligence Laboratory, Rome Air Development Center.

2 Or never should be translated. Stephen Potter points out that "Hamlet, I am thy father's ghost" would become in Afrikaans "Omlet, ek is de papap spook".

conveyed is a function of the quality, and it is an observed fact that useful amounts can be conveyed by rather primitive procedures. There are several factors involved more important than a complete syntactic and semantic analysis.

Table I is a priority list of the necessary functions and the status of their solution. Generally speaking, all these features may be solved by a large, fast memory with adequate addressing facilities. One is the facility of being able to match on longest sequences, by the simple device of running backwards through the dictionary. (Rapid access is made to a point just beyond the entry of interest.) Another facility is that of being able, on the basis of one lookup, to stuff a prefix in front of the succeeding word.

TABLE I

	Function	Status
1.	Very large fast memory	In operation
2.	Word list	Adequate and growing
3.	Local context	Solved for adjacent words, except for stuffing backwards
4.	Punctuation	Processed without difficulty
5.	Format	Control resolved
6.	Correct spelling	Solved
7.	Automatic input	Temporary measures
8.	High quality, large font, fast output printer	Control solved; equipment not available
9.	Proper nouns	Solved except in some cases
10.	Reliability	Satisfactory
11.	Word order	Temporary solution for some cases
12.	Automatic parsing	Not yet completely formulated
13.	Semantic analysis	Problem not yet formulated
14.	Pronouns	Unsolvable (?)

1. Memory

It is extremely unlikely that the meaning of a word can be "computed" from its spelling, so that one is forced to a table lookup procedure in the processing of languages. Elsewhere we have estimated that between 200, 000 and 500, 000 entries will be required in a dictionary to translate arbitrary texts of most languages; and that

by the time automatic translation is in production the demand will be for a rate of 100 words/sec.

These demands have been met, and described elsewhere, in a currently fully operating memory, the AN/GSQ-16(XW-1) equipment (Photostore). The present capacity and access time are lower than ultimate requirements, but by obvious improvements in technique, or by multiplexing, the above requirement of 10^8 to 10^9 bits with 10-millisecond random access will be met by the time the lexical material is at hand.

A very important feature of this memory is that its capacity is unlimited, for all practical purposes, so that there is no concern over adding entries to get out of difficulties in the analysis.

2. Word List

One of the least excusable failures, even on an experimental basis, in attempting to translate arbitrary text is to regret that "the word is not yet in the dictionary". The easiest problem to solve in a program for automatic language translation is to prepare a list of all the appropriate stems of the input language, and arrange to have at least some English equivalent come out. There is no need to wait for a completed method, for, in practice, the cases when a misleading translation is made are quite rare.

3. Local Context

It has long been known that word-for-word "translation" is incomprehensible, But it was estimated several years ago [3] that local context, i. e. , adjacent words, would resolve structural and meaning difficulties 85% of the time. Our results seem to verify this.

There are several ways by which this observation may be implemented. A crude method, available when memory capacity is no problem, is merely to have as entries various word pairs, triplets, etc. , and to give an equivalent of them as a whole. A more sophisticated technique is to pass clues from one word to another by address modification.

Consider the structure "I shall be coping with . . . ". When certain auxiliary verbs, such as "shall be", are looked up, the appropriate information concerning them is read out. In addition a prefix, located in the entry, can be stuffed in front of the following

---

Abraham Kaplan, "An Experimental Study of Ambiguity of Context", 1950.

word. In this case it could be the old English "a-", so the next search is not for "coping" but for "a-coping". The latter is an entry, giving the information for the participial form of the verb "to cope". In this way the completely distinct meaning of the noun "coping" (part of a. wall) is not involved.

Another example is "the curfew tolls the knell of parting day", where a class of words of which "day" is a member stuffs "de-" in front of "parting" to form "departing", thus resolving the many interpretations of "parting".[4]

Such prefixes need not have a grammatical history, such as "a-" or "de-". In order to take advantage of the resolution afforded by adjacent words, many empirical or actual grammatical word classes are established, each stuffing some artificial prefix.

If in our first example the word following "shall be" were not a participle, its prefixed form would not be an entry. The longest match would then be on only the prefix, which would give a vacuous output. Then the search on the word of the text following the auxiliary verb would proceed in the normal way.

These devices, exploiting local context, allow the first table-lookup procedure to give an output which is considerably better than a word-for-word rendering. Because authors unconsciously use local rather than remote context to resolve ambiguities, this exploitation results in a surprisingly big step toward intelligibility.

4. Punctuation, etc.

There are a number of features in printed material, not specifically part of a language, which help in conveying meaning and contribute to the structure of sentences. They are punctuation, capitalization, and italicization, etc. [5] Without them even the most elegant translation procedures would be ineffectual, so that provision must be made for their automatic recognition and processing.

Related to these are the use of Arabic numerals, Roman characters (when they are not the input alphabet), Greek symbols, and scientific notations. Abbreviations can be troublesome when

---

4 Backwards stuffing was not used in the examples given in the Appendix.

5 The European habit of spreading the letters of a word for emphasis is an unsolved problem.

punctuation or capitalization is not involved, as there is confusion with suffixes. Most of these problems have been resolved by the use of "space" before and after the abbreviation as a regular character. [6]

5. Format

There is no point in spending time and effort on an operational translation scheme if no one is going to read the output. Since most people have already too much to read in their native language, the translator has a very serious problem in finding a market for his product. A factor which has much higher priority than polish, or even completeness of translation, is readability.

Readability requires formal control: margins left and right, top and bottom; paragraphing; pagination; space for equations, drawings, and pictures; and correct hyphenation. These are tedious details, requiring little sophistication (although hyphenation is fairly subtle); but experience has shown that without them, even good translation has so little appeal as not to be read at all. Furthermore, in order for psychologists to test the value of the translation, or rendering, as a means of conveying information, it is essential that clutter due to poor format be removed to reduce the noise in the test procedures.

6. Correct Spelling

Most readers of the machine output will have the duty of preparing reports, and will be accustomed to looking for typographical errors in reading. The scanning of output from a translation machine is hindered by the occurrence of incorrect spelling of the English words. Primarily this is easily solved by careful editing of the words in the dictionary. However, in the final synthesis of the English sentence, suffixes and prefixes are added, and the many irregularities must be handled. For example, if the input language has a word, Re where R is a stem and gets converted, say, to "drag", and e is an ending which indicates the past participle normally formed by the English ending "ed", the normal synthesis would give "draged". Those verb stems ending in "g" which must be doubled before "ed" are made to stuff a prefix, say g, in front of the f(e) . Thus the

6 This is another example where flexible addressing is of value. "Space" does not always have to be treated as a character, part of a word.

address is gf(e), whose translation is "ged" (not "ed"). Similar procedures change "y" to "ies", etc.

These procedures are all carried out with exactly the same address-modification algorithms used to execute the other functions previously described.

7. Input

Naturally, automatic input is a necessity for the ultimate translating system. But even in the development phase, fairly soon one has to face up to problems "in the large", those which only exposure to large bodies of text will reveal. Thus even for experimentation, automatic input becomes a necessity.

We have demonstrated that it is quite easy to train typists with no knowledge of the input language to type accurately on a Flexowriter at the normal typing speed.[7] Thus a team of typists can prepare a large body of input texts on punched paper tape.

The ultimate answer is of course an automatic print reader operating at 1000 characters/sec, such as is now under development for Rome Air Development Center at Baird-Atomic Corporation. Such a device must ultimately be able to recognize all kinds of fonts and symbols. The encoding of the recognized characters does not offer any serious problems, as the table-lookup procedure is independent of code length.

8. Output

As remarked under previous items, readability is a prime necessity even in early stages of development of machine translation. It is essential to have a large number of symbols in each font, upper and lower case, punctuation, and fonts for headlines. Ultimately, all the fonts now used in the publishing business will be needed, to encourage the reading of our output. Studies made by Case Institute of Technology have shown that the amount of material read by research chemists is a direct function of the readability, defined in obvious ways.

Although the solution of the question is an engineering problem quite distinct from automatic translation, it is worth noting that the complex coding requirements for setting up high-equality output printing can be handled in the basic table-lookup procedure. Apart from

---

7 This is true even for Cyrillic characters.

the linguistic information gained, table lookup also serves to make any code change that is required, in the same operation.

9. Recognition of Proper Nouns

Experience has shown that proper nouns are quite frequent in typical text, and therefore they must be recognized and treated in an appropriate way.

Frequently used proper nouns can of course be listed. Unusual ones will not be found in the lists, and then can be transliterated.[8] A difficulty arises with proper nouns which except for the capitalization are common nouns (e. g. , "Blanc", "White"). These have to be discovered and listed in capitalized form. This still leaves a problem when such nouns appear as the first word of a sentence. Partial translation is frequent, e.g. [9],

самсонова (Samsonova) → The most sonova

In European languages, proper names are usually preceded by initials, and this may be used to give the clue to transliterate.

A special difficulty arises in Russian, even in the case of human transliteration. If a proper name of a Western European is transliterated into Cyrillic, and then back into Roman characters, the original spelling is considerably distorted; e. g. ,

Herbert → герберт → Gerbert

Whipple → уипел → Uipel

Not infrequently, the Russian form is a (capitalized) common noun, e. g. ,[9]

Cohen, Cohn → кон → Horse

10. Reliability

The error rate of the processing need only be an order-of-magnitude better than the typographical error rate of the input text. This means that the cost of the equipment can be considerably reduced, or its speed greatly enhanced.

Certain operations, however, can cause a series of defects if a mismatch is made. A feature of the table-lookup algorithm is that it is undisturbed by repetition of (critical) entries, either locally

---

8 This proceeds automatically, by table lookup of each letter, soon after the location has been passed where the word would have been.

9 Note that only the stem form is found.

or at remote locations. In this way, very high reliability can be achieved.

11. Word Order

The reading of output in which the order of words is not typical of English becomes very tiresome. Some alleviation has been achieved by translating word pairs, whereby multiple meanings are also resolved. In general, however, word order cannot be corrected in a single table lookup.

12. Automatic Parsing

In many cases correct rendering can be achieved only by a fairly complete parsing of the sentence, and this can be executed only on the basis of the information gained in the first lookup. In particular, the operation of pairing parts of speech to form a phrase is often initiated by the second member of the pair. It is therefore fundamentally impossible to make these pairings in a single unidirectional pass through the sentence. A facility to proceed in either direction would enable parsing of many sentences in the first lookup, but more difficult constructions will always require several passes.

13. Semantic Analysis

Finally we come to perhaps the most interesting feature of language translation, namely, the selection of the proper words of the output language directly on the basis of meaning in circumstances where syntax alone fails. Obviously a method for doing this must be devised, but experience has shown that the frequency of this difficulty in obfuscating the transmission of information is quite low, and for practical purposes the problem can be given the lowest priority.

Several observations may be made to substantiate this conclusion. We have all read without difficulty books in which many words occur whose meaning we do not know. And even "Twas brillig, and the slithy toves. . . " has been translated into French, German, and Latin. One can read and understand text in which a considerable fraction of the words have been deleted. This is because natural language is about 50% redundant in meaning content.[10]

Any language already has many words with distinct meanings, resolvable, as a rule, by the reader from context. There is no serious further degradation in a pidgin which has an additional set of

10 This is in addition to the well-known 50% redundancy in symbolism.

words with multiple meanings: one can learn this pidgin, too. These additional words are in fact precisely those with multiple meanings in the input language, which the reader of the input language has to put up with.

14. The Pronoun Problem

A difficult question which transcends sentence-by-sentence analysis is the correct translation of pronouns by identifying antecedents that lie in preceding sentences. The English "the" was originally a pronoun, and its correct usage likewise depends on an antecedent, which however may never be explicitly stated. There is then a generalized pronoun problem. Explicit or implicit definitions frequently are made early in the text. Later they are referred to by a short form, and must be identified if the short form is to be correctly translated.

Summary

If one looks at the automatic translation of languages from an operational point of view and accepts the proposition that perfection is a long way off, one finds that the impediments to the transfer of information are largely of a secondary nature. These must be solved before one really encounters the fundamental problem of automatic parsing and semantic analysis.

Most of the secondary problems have been resolved by a technique of addressing a large memory, and can be accommodated in the first table lookup. Examples are given in the Appendix.

APPENDIX
SOME EXAMPLES OF RUSSIAN SENTENCES PROCESSED WITH A SINGLE TABLE LOOKUP

¢зта крупная победа колхозов и совхозов республики достигнута благодаря высокому мастерству прославленных хлопкоробов ¢узбекистана.

This big victory kolkhozes and sovkhozes republic reached due to high skill celebrated cotton growers Uzbekistan.

¢та хе тенденция обнаруживается и в военном бюджете ¢франции, пишет газета.

The same tendence detect and in military budget France, write newspaper.

¢кто не знает магнитофона? ¢эта магнитная пленка может буть размножена и разослана по студиям телевидения.

Who not know magnetic sound recorder? This magnetic film can be multiplied and distributed by studios television.

¢мировой рекорд скорости 2,260 километров в час был установлен в мае 1958 года на американском самолете "¢ф-104", таким образом показанная 31 октября советским летчиком ¢г. ¢к. ¢мосоловым на самолете ¢е-66" максимальная скорость превшает мировой рекорд на 244 километра в час.

¢рекордный полет проходил в нтжних слоях стратосферы. ¢по правилам ¢международной авиационной федерации сфсаси летчик должен был дважды на определенном отрезке пути показать максимальнюу скорость. ¢в одном из заходов он достиг скорости 2,504 километра в час, что в 2.3 раза превышает скорость звука.

World record speed 2, 260 kilometres in hour was fixed in May 1958 year on American aircraft "F-104", thus showed 31 October Soviet pilot G. K. Mosolovym on aircraft "E-66" maximum speed exceed word record on 244 kilometre in hour.

Record flight passed in lower layers stratosphere. By rules International aviation federation (FAI) pilot should have twice on definite section way show maximum speed. In one of approaches it reach speed 2, 504 kilometre in hour that in 2. 3 times exceed speed sound.

Note: The underscored characters are actually printed in red on the outprint printer when they are not found on the dictionary.