6 The Machine Translation of Open-Ended Answers

This paper is to be presented at the ESOMAR conference in Vienna in August. ESOMAR is the European Society for Opionion and Marketing Research. The intended audience is not familiar with MT.

Ian D. K. Kelly
Quantime Ltd.

Summary

It is necessary in a multilingual survey to ensure the comparability of the data, no matter in which language it was collected. For purely closed-ended (multiple-choice) questions this is moderately straightforward, and depends largly on an accurate translation of the questionnaire. For open-ended questions, however, in which a free discursive or verbatim response is invited, answers given in different languages may be compared only after some form of translation. This translation is normally achieved by referring the verbatim responses back to a multi-lingual code-frame. The translation of this code-frame may well lie on the critical path of the data analysis: it is certainly a translation which cannot be completed until all the data is in.

Advancing technology, particularly in the field of Computational Linguistics, allows us to consider another approach. This is to translate automatically the verbatim responses themselves before coding. All the verbatim responses are then available in one language, and may be scanned, sorted, and otherwise usefully manipulated in that one language by the wide range of text-handling tools already existing. This both aids in the setting of code-frames, and makes the first step towards automatic coding in a multi-lingual environment.

Machine Translation (MT) has had some spectacular failures in the past, and is only just beginning to give useful results in particular restricted contexts. What is outlined here is an automatic translation system which we are developing, aimed at translating the great bulk of all open-ended responses. The techniques involve some highly sophisticated computing, and some flagrantly crude heuristics, and it is towards the judicious balancing of these conflicting elements that much of our research is directed, in order to achieve an automatic translation system that is sufficiently effective without excessive costs.

THE PROBLEM

Open and Closed Ends

When surveys of opinion are done for market research or other purposes, the raw data [1] in its crudest initial form consists of two kinds of answers - closed ended, or selections from pre-defined lists of possibilities, and open ended in which the respondent is invited to generate a free discursive ("verbatim") response, some of whose range of possibilities may be known in advance, but some of which will not. Examples of closed-ended questions are:

Q1: What is your age?
a) under 18
b) 18-26
c) 27-46
d) 47-52
e) 53 or over

Q3: Are you
a) Male?
b) Female?

Q2: Which of the following magazines do you read (if any)?
a) International Toenail Gazette
b) Epistemologist's Weekly
c) Purine & Pyridine Advertiser
d) The Jargonaut
e) Brass Planishing Today
f) Boilermakers' Bugle
g) The Lost-Wax Casting Chronicle
h) Flat Earth News

The respondent must be some one age, and hence exactly one of the possibilities offered may be selected: the respondent may read any, all or none of the stated periodicals, and may thus select any combination.

The answers to open-ended questions such as:

Q3: c) Other? (Please specify).

Q4: Why did you time-fly by Chrono-Stop Tours on this occasion?

may be partially guessed in advance ("convenience", "best price", "guaranteed absence of time-and-motion sickness"), but there is always the possibility of a completely unexpected answer ("I am on the wrong flight"; "I fell in love with the time-pilot"). When such open-ended questions have been asked their answers have to be inspected, and a classification scheme adopted and applied to facilitate their integration into the data analysis. These schema are called code frames, and, depending on the complexity of the questions and their answers, and their intended use, may be of almost any degree of complexity.

We have to cope more and more frequently with multi-lingual surveys. Such surveys may or may not cross national boundaries, but will inevitably create the problem of setting up compatible code-frames spanning linguistic boundaries, to allow the data generated in different languages to be compared. Once international code-frames have been set up, there is no need to translate the individual responses. Each response is matched against the code-frame by a native speaker of the language in which the response was given. Thereafter all open-ended responses may be examined by reference to the code-frame, irrespective of the originating language.

Frames and Translation

Code frames - especially multi-lingual code frames - are non-trivial to set up. Although some of the details of the code frames can be set up before the questions are asked, the frames cannot be completely set until all the responses have been received.

For the setting of frames it is necessary to examine the responses, possibly using concordance and KWIK techniques, and determine the common significantly recurring responses. But there are two problems: firstly, in an extensive multi-lingual survey it may not be possible to find one head that contains all the relevant languages. And secondly, the available automatic techniques do not bring together across the various langauges the different words and expressions that happen to mean the same thing, but are spelled differently. Hence the setting of multi-lingual code frames may well involve the slow interaction between people who do speak the relevant languages, and this slow interaction often lies on the critical path of the project of analyzing the data.

Depending on the way in which data has been collected, the verbatim texts of the open-ended responses of a survey may be available in a machine-readable form (as in [5]). In this case, if all the open-ended responses could be translated automatically into one language, then they could all be inspected by a single person, who would be able to apply the numerous text-handling tools now available. It might eventually be possible to make the target language of the translation a code frame automatically generated and selected, and speed up the whole coding process. Hence we are moved to consider the development of a Machine Translation system; that is, the development of computer software which will translate between human (spoken) languages.

MACHINE TRANSLATION

The History

Unfortunately, Machine Translation (or "MT" to its afficionados) is extremely difficult, and the general problem of translating free-running text on any subject is far from being solved as a possible computer application.

In the late 1950s it was confidently asserted that MT would be soon achieved - all that was needed was larger machines and a small amount of further study of language to determine those awkward little areas where a simple dictionary look-up does not suffice [6]. But by 1966 it was clear that the simple word-for-word approach would not work [7], and that human spoken languages (called "natural" languages to distinguish them from the "artificial" algorithmic languages being created for computers) are intensely complex. The interior systems which enable us to understand and generate language have so far resisted full analysis, or emulation by mechanical means [8] [9].

As a result of the ALPAC report [7] many research projects lost their funding, and MT was viewed as just another of those dead-ends in

science, like the search for the Philosopher's Stone. It is only now, as we begin to understand the degree of our ignorance about language, and the global need increases, that MT research has become respectable once more [10] [11] [12] [13]. The almost paralyzing effect of the policy of multi-lingualism adopted by the European Economic Cummunity in its official actions has resulted in the funding of at least two major MT research projects to the extent of five million dollars over three years [14].

The Generations of Translation

It is quite common to classify MT systems into generations [15].

First generation systems apply a simple word-by-word technique, with some heuristics to cater for idioms that otherwise defy this kind of analysis. These systems can be very effective, especially in a limited universe of discourse, where precise equivalent terms can be found easily and, more importantly, specified beforehand. Examples of basically first generation systems are [16], the present hand-held calculator-style translators, and Systran [17] - though this latter bears the same relationship to the original first-generation system from which it came [18] as the dinosaur did to the first primaeval newt [19].

Second generation MT systems are characterized by taking a syntactic approach [23]. An attempt is made to analyze completely the surface structure of the source text, and to map the text into another language by transfering that structure. These structural transformations may yearn towards the specification of a universal "interlingua" [24] [25], or eschew it with varing degrees of severity [10]. Second generation systems are capable of coping with greater difficulties and more tortuous utterances than lie within the purview of the first generation. They are effective but complex, and may well have yet to yield of their best.

Third generation MT systems are those that adopt an Artificial Intelligence (AI) approach, in an attempt to see a deep structure through the surface structure [20]. Although these may well ultimately offer the best solution, their practical use is still a long way in the future. They are still at the stage of translating very slowly and with great difficulty, particularly simple sentences drawn from very narrow pre-defined contexts [21] [22].

The Difficulty

Those of you who are not sensitive linguists, or have not studied the problem before, may be puzzled as to why machine translation should be so difficult. The difficulty is, briefly, that we do not say what we mean, but what we hope will prove sufficient for other people with similar cultural conditioning to infer what we mean from [9]. There is a story, for whose authenticity I cannot vouch, that a computer once translated the English proverb "Out of sight, out of mind" into the Russian equivalent of "Invisible idiot". One story I can vouch for is that the Russian sentence "na ulitsa bolshoi dvijenie" was once part of

the test input for a Russian-English translation. This was translated as "On a street large motion". This is fair enough, if a little startling to the uninitiated, because that is exactly what the Russian says: the trouble is, that is not what the Russian means. The sentence actually means "There is a lot of traffic in the street".

What we have here is an example of idiom: the use of a word or phrase to convey a meaning which cannot be deduced simply from the primary dictionary meanings, but can be found only by reference to the verbal or cultural context. Most of us are almost totally insensitive to idioms in our own language - we take them for granted and use them with the same unthinking ease we use the air - without a thought for its origin or composition. We become aware of idioms only when we learn a language that is foreign to us. This can be a painful awareness, and the factor that above all makes for the foreignness of other tongues. A British civil servant once commented that in Arabic every word has three meanings: its dictionary sense; the exact opposite; and something to do with a camel.

I remember as an eleven-year-old schoolboy thinking how absurd it was that in French I was forced to assert that I had (not "was") eleven years, I had hunger and I had cold. My French contemporary was no doubt thinking the English equally absurd for using the verb "to be" to assert the mere temporary possession of such attributes: it would be as logical to assert, he would say, that I am hiccoughs. A German schoolboy would agree with me that he was eleven, using the verb "to be"; he would agree with the French that he had hunger; but amidst the snows and ice of winter he would show greater linguistic stoicism, and distance himself from the discomfort with "it is to me cold".

And this is another source of difficulty in Machine Translation - the essential arbitrariness of many linguistic structures. Neighbouring languages may have similar words and similar structures: but we must no more expect identity of idiom any more than we should expect identity of vocabulary.

A NEW MT SYSTEM

Aims

With some knowledge of the difficulty, then, how should the construction of a system for automatically translating the open-ended (verbatim) responses of a survey be approached? Such a system would have to be effective, without absorbing enormous machine resources; sufficiently robust to be operated in a production environment, and sensibly priced.

A Machine Translation system is effective if it produces substantially correct translations - translations correct in substance, though not necessarily of a high literary merit - with a minimum requirment for post-editing. We need to be able to examine the information-content and begin to code the open data soon after its collection, which leaves very little time for the translation. This rules out any possibility of post-editing the translated text before using it for coding - there

simply isn't the time. During and after the coding process it may be possible to polish the translation and make it more literate before its inclusion in a final report, but the raw MT output, unedited and unchanged, must be precise enough and sufficiently readable to be coded without distortion of the balance and nuance of the data.

A robust system is one which can be operated and extended by user staff, who must not be asssumed to have the particular expertise of researchers in linguistics. As any system is used, further requirements are laid upon it: this will be particularly true of a linguistic system whose deficiencies of vocabulary must be repaired, and further contexts of use introduced.

It is beyond the scope of this paper to discuss pricing as such, but one consequence of the "sensible price" requirement is a limitation on the manpower that can be expended in creating such a system. We are at this time beginning the design and construction of an advanced second generation MT system directed specifically towards the information content of open-ended responses of opinion and market surveys. We estimate that with our new understanding of computational linguistics the investment of between one and a half, and two man-years will produce a usable system.

Outline design

In principle it does not matter how the machine translates, merely that it does. Whether the translation is achieved by the arcane alchemy of programmers lurking amongst nests of wires and tangles of incomprehensible printout, or by means of a multi-lingual gorilla somehow electronically entrapped in the black box is a matter of mere technical detail. Those of you, however, who are, like me, endlessly fascinated by such details may be interested in a very brief thumbnail sketch of the technicalities involved.

The translator we are in the process of constructing will have five main parts.

The first part is concerned with morphology and lexical analysis of the source language; the second performs a syntactic analysis of the surface structure. The third part performs the semantic matching, the fourth the syntactic transformation required by the target language, and the final fifth part generates the output.

Phase 1 is thus a dictionary look-up. But in order to keep dictionaries to a manageable size, words are stored only in their root form, with flags indicating the applicable morphology. Thus the English words:

love loved loving loves lover lovable
unloved unloving unlovable

would all be stored under the root "love", with flags indicating the possibility and use of each of these morphological variants. The dictionary is a complex interconnection of files which is able to pass

into phase 2 (syntactic analysis) information about the syntactic usage of the word in its various meanings, its collocations, and any longer idioms in which it might take part. It is, incidentally, in this phase that we must deal with mis-spellings and abbreviations introduced by the interviewers, or those entering the data.

Phase 2 is a recursive top-down fast-back table-driven syntax analyzer based on PROTRAN [24] [25]. This phase constructs a generalized tree with labelled nodes from the linear sequence of marked lexemes (dictionary entries) produced by phase 1. This analysis tree is what all the subsequent phases operate on, by re-labelling its nodes, and transforming its shape.

Phase 3 is primarily semantic, and it performs the kernel of the translation transformation - the vocabulary transfer. The semantic model used is that of vectorial sematics [26], which appears to have sufficient depth and complexity to deal with the proposed application, even though it is recognized to be structurally incomplete when compared against, say, Montague semantics [27] or AI models [22]. It is not intended to use this system to translate long, logically interconnected passages, nor is the system designed to cope with narrative describing the changing states of entities, and which can be understood only by tracking those changing states [28].

Phase 4 is effectively the reverse of Phase 2, producing a linear string out of a tree or net. This "soutax" phase is also based on PROTRAN.

The fifth phase performs the final morphological adjustments required by the target language - for example, making adjectives agree with their nouns in French, or adding an "n" to the English indefinite article where it occurs before a vowel.

Progress so far

We have working already the outlines of the syntax for the second (syntactic) phase. This is possibly the easiest part, as we are mainly concerned with translating short episodic phrases of a very simple structure: the abstract syntax required to describe longer fragments of free-running non-conversational prose is very complex [29]. We also have a substantial part of the first phase functioning in that a given word can be conjugated automatically through all the forms of which it might be a regular variant, which is done prior to inspecting the dictionary for the correct morphology.

VECTORIAL SEMANTICS

Roget's Thesaurus

Roget's Thesaurus [30] is a dictionary organized by meaning rather than by alphabetic sequence. It consists of (about) 1000 paragraphs, each of which contains the synonyms, near-sysnonyms and associated phrases of the basic idea embodied in and expressed by the paragraph's

headword. We will call this idea the paragraph's theme. P. M. Roget classified and subclassified the whole of human consciousness expressible in speech, singling out the key themes. The higher levels of this classification will not concern us here [31]. A paragraph may contain not only nouns, but also verbs, adjectives, adverbs and compound expressions relating to its theme. expressions relating to its theme. For example:

385. Refrigeration - N. refrigeration, reduction of temperature; cooling &lc.; ice &c. 383; solidification &c. 321; refrigerator &c."387.
V. cool, fan, refrigerate, refresh, ice; congeal, freeze; benumb, starve, pinch, chill, petrify, chill to the marrow, nip, cut, pierce, bite, make one's teeth chatter.
Adj. cooled &c.w v.; frozen; cooling &c.; frigorific‡

A given word may occur in several different paragraphs, being related to the concept of the headword of each of them. Some words are so placed because they are polysemes (homonyms) - words with two completely distinct meanings - for example, "bowl" in the senses of "wide-topped deep dish" and "throw or deliver a ball with a straight-arm action (as in cricket)", or "dear" in the senses of "cherished" and "expensive" [32]. But even a word with just one basic meaning can find itself listed in two or more Roget paragraphs, because that one meaning is associated with the several headwords.

There are different classes of this association: some words may have closely similar meanings (for example "little" in 32:smallness [33]); some associations are of degree (e.g. "warm", "torrid", "white hot" in 382:heat); some bear the relationship of component to compound (e.g. "composer" and "pianist" in 416:musician). There are associations of an encyclopaedic kind: for example "cannon ball" and "swallow flight" in 274:velocity, "wings", "ailerons" and "parachute" in 273a:aircraft; and there are associations which can only be described as "remote assonances in the mind" - for example, "trail of a red herring" in 615:motive, and "secret passage" in 627:method.

Relevance, Quantity and Type

So we can, for any given word, quote a string of 1000 numbers, each between 0 and 1, and each of which indicates the degree of relevance of one theme to the meaning of the word being thus described, here termed the defiendum [34]. Relevance is the loosest, most general kind of association, which subsumes all others. Examples of the values that might be assigned for the words "Adoration" and "Rubbish" are:

Adoration relevance	theme	Rubbish relevance	theme
0.75	990:worship	0.7	040:remainder
0.2	991:idolatry	0.5	517:unmeaningness
0.5	897:love	0.8	645:inutility
0.5	898:hatred	0.3	497:absurdity
		0.4	653:uncleanliness
		0.3	401:fetor

with all other values being zero. Note that because we are looking here only at relevance we do not distinguish between the kinds of relationship held. We might not ordinarily think of "hatred" as being relevant to "adoration", but a little reflection reveals that the qualities discernable in adoration may be seen in negative quantity in hatred.

Thus for each of the relevant themes there must also be specified the quantity in which it is possessed. We might at first suppose that these quantities would be numbers between -1.0 and +1.0, but this does not allow for the concept of excessive possesion. So we allow these quantity values to range between -2.0 and +2.0, with +1.0 representing the norm of complete inherence of the relevant quality, and -1.0 its precise antonym. Taking our previous examples of "adoration" and "rubbish" we might assign these quantities as:

Adoration quantity	theme	Rubbish quantity	theme
1.25	990:worship	0.5	040:remainder
-0.8	991:idolatry	0.8	517:unmeaningness
1.25	897:love	1.0	645:inutility
-1.25	898:hatred	0.3	497:absurdity
		0.5	653:uncleanliness
		0.3	401:fetor

Each theme relevant to the defiendum is related by some specific type of relation. There are possibly more of these types of relationship than there are ways of getting to heaven: for our purposes we need only distinguish between fourteen types of relationship in each of two classes: lexical and encyclopaedic. Lexical relationships are concerned purely with the base (dictionary) meaning of the defiendum, without which it simply would not be what it is. Thus aeroplanes (are intended to) fly, rubbish is of no utility, and the lemon is the fruit of the Citrus Limonum. But aeroplanes are also noisy, lemons are bitter, and rubbish often smells nasty. Each of these relationships is encyclopedaic, not lexical. One can easily conceive of silent aircraft and sweet lemons, and the development of these by the appropriate technology would not require ordinary dictionaries to be re-written. We could not, however, call the root of the solanum tuberosum "lemon" without making a dictionary change (and, perhaps, employing some pretty violent genetic engineering too); nor would the lexical meaning of

"aeroplane" permit us to develop an aeroplane which did not fly, but instead processed seaweed into sausages on the ocean floor: we may develop such a machine if we wish, but we cannot call it "aeroplane" without first warning the (somewhat bemused?) lexicographers.

The types of relationship we will particularize are:

0 No relationship
1 Defiendum results in, or is for the intention of theme; theme is caused by defiendum
2 Theme causes, results in, or is for the intention of the defiendum; defiendum is caused by theme
3 Defiendum is part or component of theme
4 Theme is part or component of defiendum
5 Defiendum is location or time of theme
6 Theme is location or time of defiendum
7 Defiendum is instrument of theme
8 Theme is instrument of defiendum
9 Defiendum is subclass or type of theme
10 Theme is subclass or type of defiendum
11 Defiendum is property of theme
12 Theme is property of defiendum
13 Defiendum is some degree of the theme (synonomy)
14 Some other or non-specific relationship

Examples of these might be:

Lexicographic			Encyclopaedic	
type	defiendum	theme	defiendum	theme
1	composer	415:music	aeroplane	402:sound
2	cure	662:remedy	expansion	382:heat
3	melody	415:music	mast	273:ship
4	river	337:water	wisdom	842:wit
5	concert	415:music	stag party	959:drunkenness
6	worship	1000:temple	beauty	127:youth
7	furnace	384:calefaction	information	625:business
8	transport	273:ship	brevity	842:wit
9	ketch	273:ship	bigamy	964:illegality
10	vehicle	273:ship	politician	504:madman
11	rhyme	597:poetry	poverty	559:artist
12	water	333:fluidity	lemon	436:yellowness
13	speed	274:velocity	genius	503:insanity
14	scholar	505:memory	oyster	804:poverty [35]

As examples let us take the values which might be assigned to quantify the semantics of the words "Adoration" and "Rubbish":

Adoration				Rubbish			
relv.	qty.	type	theme	relv.	qty.	type	theme
0.75	1.25	13	990:worship	0.7	0.5	9	040:remainder
0.2	-0.8	13	991:idolatry	0.5	0.8	12	517:unmeaningness
0.5	1.25	4	897:love	0.8	1.0	12	645:inutility
0.5	-1.25	4	898:hatred	0.3	0.3	13	497:absurdity
				0.4	0.5	12	653:uncleanliness
				0.3	0.3	1	401:fetor

Thesaurus Reduction

There are many instances in [30] where successive paragraphs exactly negate each other: since we may use negative quantities to represent opposite meanings we may drop one out of every such pair to produce a reduced thesaurus base. The choice of which within any pair of antonymic themes is to be dropped and which retained is completely arbitrary. We will assume in what follows that an efficiently reduced base is available.

Vector Spaces

Let R, Q and T be three vector spaces of dimensionality N, the number of themes in the reduced thesaurus base. We will call these the Relevance, Quantity and Type spaces.

For any defiendum we may arrange the sequence of numbers measuring relevance to the themes of the reduced thesaurus to form a vector in R. Each vector is an ordered sequence of N numbers, each between 0 and 1, with the nth number in the sequence representing the relevance of theme n of the reduced thesaurus to the defiendum. These relevance vectors will tend to be very sparse, most of their elements being zero. There is a convenient metric available for these vectors which will measure the standard Euclidean distance between any pair of relevance vectors, and thus indicate the mutual degree of relevance between two defienda.

If $\underline{a}$ and $\underline{b}$ are two vectors in R such that

$$\underline{a} = (a_1, a_2, a_3, \ldots, a_{N-1}, a_N)$$

$$\underline{b} = (b_1, b_2, b_3, \ldots, b_{N-1}, b_N)$$

then the standard Euclidean distance between $\underline{a}$ and $\underline{b}$, written "dist($\underline{a}$,$\underline{b}$)", is

$$\text{dist}(\underline{a},\underline{b}) = \sum_{i=1}^{N}(a_i - b_i)^2$$

The modulus or length of a vector $\underline{a}$, written $|\underline{a}|$, is its distance from the zero vector:

$$|\underline{a}| = \sum_{i=1}^{N} a_i^2$$

The vectors of space Q are similarly the relevant theme quantities

taken in order: for this space too the standard Euclidean metric is meaningful. The vectors of space T, the relevant theme relationship types, are purely conventional, however, and for them there is no meaningful metric.

Closeness of Meaning

The separation of two defienda may be measured by the quoted Euclidean metric in space R, or in space Q, or in any one of a series of semantic spaces S, S', S", etc. The simplest of these spaces is RQ, formed from the product of the elements of an R vector with the corresponding elements from a Q vector. Thus if defiendum D corresponds to vector $\underline{r}$ in R and vector $\underline{q}$ in Q, then it corresponds to vector $\underline{s}$ in RQ, where

$$\underline{r} = (r_1, r_2, r_3, \ldots, r_{N-1}, r_N) \quad \text{written as} \quad \{r_i\}(i)^N_1$$

$$\underline{q} = \{q_i\}(i)^N_1$$

and $$\underline{s} = \{r_i q_i\}(i)^N_1 .$$

A less simple but more sensitive semantic space would be the vectors

$$\{r_i q_i f(t_i)\}(i)^N_1$$

where $\underline{f}$ is a weighting function which allots each type of relationship a different weight. The intention of the weighting function is to make the dist metric a more sensitive measure of intuitively-perceived closeness of meaning. The general semantic space of this type would be composed of vectors

$$\{g(r_i, q_i, t_i)\}(i)$$

An even more general semantic space, in which we would have to forgo the Euclidean metric, is RxQxT, of dimensionality 3N. Each vector is then of the form

$$\{r_i, q_i, t_i\}(i)^N_1$$

We will call this space M, meaning space.

Many functions will give us variously useful measures of distance between two such vectors, and part of the research that has yet to be done covers the development of suitable metrics. The simplest is $D(\underline{a},\underline{b})$, defined by:

Let $\underline{a}$, $\underline{b}$ be vectors in M

Let $$\underline{a} = \{r_{ai}, q_{ai}, t_{ai}\}(i)^N_1$$

$$\underline{b} = \{r_{bi}, q_{bi}, t_{bi}\}(i)^N_1$$

$$D(\underline{a},\underline{b}) = \sum_{i=1}^{N} [(r_{ai} q_{ai} - r_{bi} q_{bi})^2 f(t_{ai}, t_{bi}) + g(t_{ai}, t_{bi})]$$

where $\underline{f}$ and $\underline{g}$ are given by suitable weighting functions.

Ideal Bases

It would be a very remarkable coincidence if Peter Mark Roget in 1852 had chanced upon the most economic orthogonal base for the vector spaces R, Q, T and M proposed here: it would be even more remarkable were this space to have dimensionality of exactly 1000. We have already contemplated a reduction by the removal of antonyms without a basic change in arrangement: other changes are possible and desirable. To produce the best set of themes to constitute a base for M there would have to be additions, deletions and rearrangments.

Additions are needed to cope with new ideas not catered for in the base themes. The 1953 edition of Roget needed a paragraph for "Aircraft", which was not anticipated in 1852: the 1999 edition may perhaps need a paragraph detailing the various kinds of space-ship. As well as these natural extensions arising from our slowly changing perception of the universe and our place in it, there are deficiences - holes - in the set of themes for expressing even simple current ideas: "asprin" and "Neaderthal" are words it is instructive to try and place in M.

Deletions are needed, as even a reduced thesaurus is not a linearly independant set of orthogonal vectors. It must be possible, for example, to reduce the set {175:Influence; 462:Answer; 496:Maxim; 500:Sage; 512:Omen; 527:Information; 537:Teaching; 617:Plea; 668:Warning; 695:Advice; 696:Council; 697:Precept}.

More efficient selection of the thesaurus themes without necessarily reducing their number is also called for, so that vectors defined in the space have as many as possible zero elements. This is roughly equivalent to requiring that the set of themes be an orthogonal base for the vector space.

Some careful research is needed into the structure and content of an ideal base, and some of our development effort must go towards this.

Deficiencies, Structures, and Operators

There are many points yet to be cleared up. The model for meaning given here does not cope with "deficient" words, "structural" words or "formal" words.

A defiens is classed as deficient if its length in R is less than 1. A deficient words does not "have enough meaning" in M - it cannot be fully described in M. This may be because M lacks the axes - the themes - to describe M. In this case the dimensionality of M can be increased and another theme added to the base: words like "Neaderthal" and "asprin" would be dealt with this way, for example. There has to be a limit to the number of axes added to the base, however, and this approach must be stopped before it begins adding one more axis to the base for every proper name encountered.

A defiens may be deficient because the defiendum is structural; that is, its meaning may be derived from the existing base, but not by

simple mixture, rather by a structural composition. Again such words can be dealt with by augmenting the base, but this seems inelegant because uneconomic.

And a word may be deficient because it is intrinsically deficient. These will tend to be the formal words of the language that do not in themselves mean anything, but constitute the "connective tissue" of the body of an utterance. Almost all prepositions and conjunctions fall into this class, as do definite and indefinite articles and those odd little particles which defy placement by the schoolmen's attempts to apply Latin grammar to all languages - for example the interrogative "cu" of Esperanto, and the emphatic " " of Russian. Formal words of the source language will tend to be fully absorbed by the syntactic phase: the formal words of the target language will, in general, be generated by the soutax phase.

Some words act not as full vectors, but as operators upon other vectors: "too" and "very", for example, will stretch the q component of the vectors to which they apply; "somewhat" and "slight" will shrink q values, and "not" negate them. These operational words must be identified separately, and their corresponding operators applied before the language transfer transformation takes place.

Summary

None of this discussion of Vectorial Semantics has made any assumption about the identity of the language from which the defienda arise. Indeed, it is the whole point of the argument that words from (at least) two different languages should be referred to the same themes, and expressed as vectors against the same axes. The discovery of a term in the target language corresponding to a term in the source language then becomes a process of identifying which of the listed vectors in M representing terms in the target langauge are closest, under the chosen metric, to the vectors corresponding to terms encountered in the source language text.

As dictionaries grow the collocation information will grow, and hence the avoidance of infelicities such as "on a street large motion". The heart of any translation system is its dictionary: the quality of the translation is limited by the quality of the dictionary. As the practical success of Systran has shown, the effects of even quite severe deficiencies in syntactic areas can largely be avoided by excellence in the dictionary.

SYNTAX

Introduction

Syntax has been described as "what you can say", and semantics as "what it means when you have said it". The full syntax of any natural language is long, complex and irregular. Except for languages like Esperanto (and perhaps not even then) there is no simple way to classify the forms of all possible utterances.

Luckily, we do not have to. The verbatim responses we are considering tend to be short, partial utterances: fragments, not complete sentences. Their syntactic form will not, therefore, be as involved as that of sentences taken from (in order of complexity) technical literature, poetry, or (worst of all) colloquial speech. The responses, moreover, have been filtered by interviewers who are not impartial transcribers like tape-recordsers, and most of the gross grammatical lapses will have been corrected at this first data-collection stage.

Sentential Models

We are concerned for the moment only with Indo-European languages, for which a simple partial phrase-structure grammar will be adequate. A grammar which assumes much more watertight classes of word category than are really observed in free colloquial speech will be sufficient, and moderately easy to specify. The full power of general recursion will not in practice be exercised by the utterances under consideration, although for simplicity of expression of the syntax general recursion will not be forbidden. In fact, the syntax will be couched in a permissive rather than a prescriptive manner: the syntax will permit more forms than will actually occur - this does not matter, provided it describes correctly the forms that do occur.

The top-down form of analysis normally presumed by PROTRAN is not the most efficient for a grammar where there is a doubt as to the identity of the top level. In the context of the texts we will be translating there is no unique formal target for the syntax of all the utterances. A basically bottom-up or "island-collision" parser would be more efficient, and these will be investigated. Until such a parser is available we will be expressing the PROTRAN input correlator to reflect this top-level variance. Because the practical instances we expect to encounter will not excercise the full complexity of the syntax analyzer, this will not be a severe loading on the time taken to perform the syntax analysis.

CONCLUSION

As yet there is very little to show - the parts of the system that are working are not easily demonstrable in isolation. But though the shape of the building is obscured under a mass of software scaffolding, we are confident that it will eventually be a renowned work of architecture - spacious, exciting, and indicative of the future.

References

[1] Etymologically the word "data" is the regular plural of a Latin second declension neuter noun "datum". The language of this paper is English, not Latin, and notwithstanding the strictures of Fowler [2] and Gowers [3], we here adopt the standard British English treatment of

the word as a non-count or mass noun (see Jespersen [4]), by analogy with words like "bread" and "wine", or - more precisely, as there is no really usable singular count-noun which corresponds - "furniture".

[2] Fowler, H. W., Modern English Usage, Oxford, Oxford University Press at the Clarendon Press, 1958 (first published 1926), p.103.

[3] Gowers, Sir Ernest, revised Fraser, Sir Bruce; The Complete Plain Words, London, Penguin Books 1980, ISBN 0 14 02.0554 3, pp.183, 277.

[4] Jespersen, Otto, Essentials of English Grammar, London, George Allen & Unwin, 1979 (first published 1933), ISBN 0 04 425005 2, Chapter 21 (note particularly the last paragraph).

[5] Miah, Barbara, Quancept Users' Manual, London, Quantime Ltd., 1981

[6] For an eminently readable potted history of MT see Professor Yorick Wilks, "Time flies like an arrow" and "Frames for machine translation", New Scientist, 15/22/29 December 1977.

[7] Automatic Language Processing Advisory Committee, National Academy of Sciences, National Research Council, Language and Machines: Computers in Translation and Linguistics, Rept: Pub. No. 1416, Washington D.C., 1966 ("The ALPAC Report").

[8] Simon, J. C. (Ed.), Spoken Language Generation and Understanding, Proceedings of the NATO Advanced Study Institute held at Bonas, France, June 26 - July 7, 1979, D. Reidel Publishing Company, Dordrecht, Holland, 1980.

[9] "The silent adjustments to understand colloquial language are enormously complicted"; Tractatus Logico-Philosophicus, 4.0002, Ludwig Wittgenstein, Routledge & Kegan Paul, London, 1922 (1962 edition).

[10] EUROTRA, the machine translation system being designed by order of the Commission of the European Communities; for full references apply to DG XIII, Luxembourg.

[11] Bruderer, H. E., Handbook of Machine Translation and Machine-Aided Translation - Automatic Translations of Natural Languages and Multilingual Terminology Data Banks, Amsterdam, North-Holland Publishing Company 1977.

[12] Lawson, Veronica; "Tigers and Polar Bears", Newsletter No. 11, Natural Language Translation Specialist Group, British Computer Society, 1981 (revised from "The Incorporated Linguist" 18.3, 1979).

[13] Loh Shiu-Chang, A Bibliography of Machine Translation, The Chinese University of Hong Kong, 1978 (Available through the British Computer Society - apply to The Treasurer, NLTSG (BCS), 72 Brattle Wood, Sevenoaks, Kent).

[14] Overcoming the Language Barrier, Verlag Dokumentation, Munich 1977. Proceedings of the third European Congress on Information Systems and Networks, EEC. ISBN 3-7940-5184-X.

[15] Melby, Dr. Alan K., "Design & Implementation of a Computer Assisted Translation System" Brigham Young University, in Newsletter No. 10, NLTSG (BCS), 1980.

[16] Loh Shiu-Chang and Kong L., "Computer Translation of Chinese scientific journals" in [14] pp 631-645.

[17] Toma, Dr. Peter, "An operational Machine Translation System", in Translation, Ed. Richard W Brislin, New York; Gardner Press, 1976, pp 247-259. (SYSTRAN)

[18] Toma, Dr. Peter, "The SERENA System, Part I, Morphology, Machine Translation Programming Paper I", Georgetown University, Washington D.C., 1959.

[19] This is not intended to be insulting. The dinosaurs were a very succesful life-form that dominated the earth for 160 Million years: it is only by modern standards that they seem cumbersome.

[20] Chomsky's use of these terms has been frequently misrepresented: for his original intention see Aspects of the Theory of Syntax, Noam Chomsky, M.I.T., Cambridge Mass., 1965.

[21] Winograd, Terry, Understanding Natural Language, New York, Academic Press 1972.

[22] Schank, R. C., and Colby, K. M., eds, Computer Models of Thought and Language, San Francisco, Freeman, 1973.

[23] Chandioux, John; "Creation of a Second-Generation System for

Machine Translation of Technical Manuals" in [14]

[24] Kelly, Ian D. K., "PROTRAN - A generalized translation tool for natural and algorithmic languages", in Overcoming the Language Barrier [14]

[25] Kelly, Ian D. K., "PROTRAN - An Introductory Description of a General Translator", in Practice in Software Adaption and Maintenance, Proceedings of the SAM workshop, Berlin 1979, North Holland Publishing, 1980, ISBN 0-444-85449-5

[26] Kelly, Ian D. K., "Thesaurus Vectors", Newsletter No. 9, NLTSG (BCS), 1980.

[27] Montague, Richard; Formal Philosophy; ed. Thomason, Richard H., New Haven, Yale Univerity Press 1974.

[28] Schank, R. C., and the Yale AI Project, SAM - A Story Understander, Research Rpt. 43, New Haven:Yale University Department of Computing Science, August 1975.

[29] Sager, Naomi; Natural Language Information Processing - A Computer Grammar of English and its Applications; Reading Mass. USA, Addison-Wesley 1981, ISBN 0-201-06769-2

[30] Roget, Dr. Peter Mark, Thesaurus of English Words and Phrases; Classified and Arranged so as to facilitate the Expression of Ideas and to assist in Literary Composition ("Roget's Thesaurus") 1852; abbridged with additions by John Lewis Roget and Samuel Romilly Roget, Penguin Books, 1953.
The 1953 edition has 1000 numbered paragraphs, plus a few extra (100a, 356a, etc.); the 1966 edition has been re-arranged, revised, and re-numbered by Robert A. Dutch into exactly 990. Since writing, the 1982 edition of Logmans, reworked by Susan M. Lloyd, has been published. All references here are to the 1953 edition.

[31] Some of the aspects of this classification are quirky, to say the least. "Food", for example, is classified not under "Organic Matter", but under "Space, Motion, with reference to direction". The universal classification attempted, even as an heirarchical structure, is probably impossible.

[32] The perception of the costliness of loved ones is fairly widespread: this particular homonymy occurs in Italian and French as well.

[33] The mode of reference to a particular theme of [30] adopted here is to give the paragraph's reference number followed by the headword. Thus 32:Smallness; 897:Love etc.

[34] "Defiens" and "Defiendum" are perhaps inappropriate terms, as it is not up to individual volition to choose the meaning of words. What is being spoken of here is not true definition, but description. A definition cannot be wrong: a description, being an attempt to reach something in the real world outside of itself, can be.

[35] Dickens, Charles; The Pickwick Papers