Ju.D.Apresjan, I.M.Boguslavsky, L.L.Iomdin, A.V.Lazursky, L.G.Mitjushin, V.Z.Sannikov, L.L.Tsinman

Institute for Information Transmission Problems
Russian Academy of Sciences, Moscow

Automatic Dictionaries in the ETAP-3 System

From: AAAI Technical Report SS-93-02. Compilation copyright © 1993, AAAI (www.aaai.org). All rights reserved.

1. ETAP-3: General Information

ETAP-3 is an outgrowth of the ETAP-2 MT system (see [1-2]) and, in its present state, has evolved into a multipurpose linguistic processor [3]. It is implemented on a VAX-6340 computer. The programming languages used are PL/1 and C. The processor's two basic options are communication with relational databases in unconstrained natural language and bidirectional (English-to-Russian and Russian-to-English) machine translation (MT). The latter system can be joined to the former, so that the database can be accessed in either of the two working languages. The formats of linguistic knowledge representation are highly standardized and completely language independent, so that the number of working languages can be easily enlarged. Linguistic knowledge itself (morphological, syntactic and transfer rules, morphological, combinatory and semantic dictionaries) is not domain-oriented and does not require any re-adaptation (apart from the domain-specific part of the dictionaries) when transferred to a new object domain. The algorithms are geared exclusively to the formats of linguistic knowledge representation and do not require any re-adaptation either.

In fact, both basic options may be conceived of as machine translation systems, the main difference being that the MT system proper translates scientific and other technical texts from one natural language into another, while the system of natural language interface with databases translates texts (queries) from a natural language into the Structured Query Language (SQL).

In processing queries and hard science texts VAX-6340 takes from 4-5 to 10-15 seconds respectively for a sentence of average complexity and length (up to 15 words for queries and up to 30 words for translated sentences). The quality of machine translation is comparable to that produced by a human. Here is a sample of the MT system performance in direct and back translation:

In recent years, government regulations, technical demands, microprocessor development and a host of other influences have forced a revolution in chemical, biochemical and industrial instrumentation →

V nedavnie gody pravitel'stvennye postanovlenija, texnicheskie trebovanija, razvitie mikroprocessorov i mnozhestvo drugix faktorov vyzvali perevorot v ximicheskom, bioximicheskom i promyshlennom oborudovanii →

In recent years government decrees, technical demands, development of microprocessors and a set of other factors caused a revolution in chemical, biochemical and industrial equipment.

2. Linguistic Background

Linguistically, ETAP-3 takes its inspiration from Mel'chuk's "Meaning ↔ Text" theory [4]. There are three levels of text representation in the MT system and four levels of representation in the system of natural language interface with databases. This difference is accounted for by the fact that natural languages display sufficient linguistic similarity at the level of deep syntactic (=normalized) structure to cope with the mismatch and effect a transition from the source to the target language. However, there is a more fundamental difference between queries in a natural language and their SQL counterparts, so that to overcome the mismatch in such a pair it becomes necessary to reach the level of semantics, which acounts for the additional level of representation.

The three levels of text representation shared by both systems are morphological, syntactic, and deep syntactic. As is clear from the above, there is a semantic level in the system of natural language interface with databases. At all the levels, apart from the morphological, sentence representations are written in the same formalism of dependency trees.

In what follows we concentrate on the MT systems only.

Morphological representations are strings of lexeme names, with a set of morphological characteristics (specifications of number, tense, aspect etc.) assigned to each such name. In case of homonymous wordforms alternative sets of characteristics are generated. Syntactic and deep syntactic representations, or structures, are

linearly ordered dependency trees, with wordform representations for nodes and subordination relations for labelled arcs (in all, 40 to 60 relations are used to account for various universal or language specific syntactic constructions.) All the levels are kept strictly apart although in the process of transfer hybrid structures combining the elements of adjacent levels cannot be avoided.

3. Basic Components of MT System

Two basic components of each MT option (English-to-Russian and Russian-to-English) are **grammar** (i.e. sets of rules) and **dictionaries**.

The core of the system are parsing rules. They map a pair of wordform representations occurring in the processed sentence onto a (hypothetical) binary dependency subtree. A sophisticated system of filters and other devices is applied to extract the correct dependency tree (or a number of such trees in case of syntactic homonymy) out of the set of admissible hypotheses. The resulting tree (syntactic structure of the sentence) is transformed into the deep syntactic structure which is free of whatever is language- specific in the source language. It is on the level of the deep syntactic structure that the transfer phase (translation proper) takes place, during which all structural units (lexical units and syntactic dependency relations) are replaced by units of the target language. At the next phase called expansion the obtained structure is equipped with a number of language specific elements, including the semantically motivated morphological characteristics, and the result of this operation is in its turn submitted to syntactic synthesis to produce the fully accomplished syntactic structure of the target language. At the last stage morphological rules are applied to produce the target sentence in conventional orthography.

4. Dictionaries of the MT System

In all, the MT system makes use of four dictionaries: two monolingual morphological dictionaries (for each of the two working languages) and two combinatory dictionaries for Russian and English, which are nearly monolingual (except for the translation zone and certain translation rules in the entries of both dictionaries). All four dictionaries are addressed in both directions of translation.

While morphological dictionaries contain information for morphological analysis/synthesis only, the combinatory dictionaries store extensive information on the syntactic, semantic, co-occurrence and other properties of lexemes to provide all the material necessary for parsing and subsequent transfer. In particular, combinatory dictionary entries specify for every keyword

(1) its part of speech;

(2) the simplest and most common translation into the target language;

(3) syntactic features of the keyword allowing it to occur or precluding its occurrence in certain syntactic constructions; the total number of such features operational in the system is 150 for English and around 200 for Russian; for example, such verbs as *say, declare, report, make, help, feel* etc. are assigned the feature "passto" which marks their ability to govern the *to*-infinitive in the passive voice, as in *He is reported to have failed* or *He was made to stay*;

(4) its semantic features (desrciptors), such as 'action' (for *work, production, walk)*, 'information' (for *text, picture, sign, data*), 'parameter' (for *velocity, length, pressure, colour*), and the like; the total number of such semantic features is over 80;

(5) its government pattern, with further indications of the morphological, syntactic, lexical or semantic requirements which its potential actant dependents are expected to meet; for example, the verb *to rent* will be supplied with the indication that it may govern a direct NP object, a counteragent phrase in the form *from somebody* with further semantic specification that it must be a 'person' or an 'institution', a "price phrase" in the form *for how much* with further semantic specification that it must be 'money', and a duration phrase in the form *for how long* with further semantic specification that it must be 'time'; cf. *He rented a house from his neighbour for a year for the funny price of 50 dollars.*

Apart from that, a dictionary entry may contain references to specific (narrow lexical scope) rules which should be activated in case the given keyword occurs in the processed sentence, and the so-called dictionary rules which are word-specific. For example, the dictionary entry for *attention* will contain reference to the speci-

fic transfer rule providing for the idiomatic translation of the phrases like *to pay attention → udeljat' vnimanie, to draw attention → privlekat' vnimanie, to focus attention → sosredotochivat' vnimanie* and so on. The same rule will be mentioned in the dictionary entries of all the nouns in whose context support verbs are translated in a way different from the one indicated in their common translation zone. Such rules (whose total number in the system goes well into a thousand) may pertain to all the non-morphological phases of sentence processing.

The extent of sophistication necessary to compile such dictionary entries makes it virtually impossible to devise any automatic procedure of compiling combinatory dictionary entries directly from either a corpus of examples or from man-made monolingual or bilingual dictionaries, although the corpus and the "human" dictionaries are extensively used to produce the automatic combinatory dictionary. Besides that, prototypical dictionary entries can be selected for large lexicographic classes to facilitate the actual work of compilation.

Even morphological dictionaries for highly inflected languages such as Russian would hardly allow fully automatic compilation, although of course means can be devised to substantially facilitate this work. ETAP-3 uses a special algorithm based on a simple question-answering system, which helps to produce morphological dictionary entries semi-automatically. In the system, a short dialogue helps create entries generating whole paradigms of Russian words which, in the case of verbs, may contain as many as 250 wordforms.

References

1. Ju.D. Apresjan et al. Lingvisticheskoe obespechenie sistemy ETAP-2. (The Linguistics of the ETAP-2 system.) Moskva: Naula, 1989.
2. Ju.D. Apresjan et al. ETAP-2: The Linguistics of a Machine Translation System, META, 1992, vol. 37, No 1, pp. 97-112.
3. Ju.D. Apresjan et al. Lingvisticheskij processor dlja slozhnyx informacionnyx sistem. (A linguistic processor for advanced information systems.) Moskva: Nauka, 1992.
4. I.A. Mel'chuk. Opyt teorii lingvisticheskix modelej tipa "Smysl ↔ Tekst. (The theory of Meaning ↔ Text type linguistic models.) Moskva: Nauka, 1974.