A Polish-to-English Text-to-text Translation System Based on an Electronic Dictionary

Krzysztof Jassem
Dept. of Computational Linguistics and Artificial Intelligence
Faculty of Mathematics and Computer Science
Adam Mickiewicz University
Matejki 48-49, 60-769 Poznan, Poland
jassem@math.amu.edu.pl

POLENG is a Polish-to-English text-to-text translation system based on an electronic dictionary. The dictionary software enables the storing of lexical data in a finite automaton. The translation software uses the Arity Prolog interpreter in order to obtain a phrasal structure of an output expression.

1 The dictionary

The process of creating a Polish-to-English electronic dictionary destined to be used in computerised text translation was performed out in the following steps:
1. *The preparatory phase.* In this phase classification files of the inflection of Polish words as well as the coding of Polish and English inflection paradigms were prepared.
2. *The phase of creating the dictionary of canonical forms.* This phase was carried out by lexicographers aimed by an interactive computer application. Each entry in the dictionary is supplied with inflection codes of its Polish and English parts as well as other syntactic-semantic information. The inflection code of the Polish part of an entry is a reference to a set of inflection endings stored in one of classification files prepared in phase 1. The format of the inflection code of the English part is "self-constructive", i.e. it enables the generation of appropriate inflected forms from a canonical form without the necessity of a time-consuming look up of any classification file. (Designing a "self-constructive" code for a highly flexional language like Polish would have been a complex task).
3. *The phase of generating the SGML-type dictionary of inflected forms.* This phase is executed automatically. The inflected forms are generated on the basis of Polish inflection codes attached to all entries in the dictionary of canonical forms. Only inflected forms of Polish words (phrases) are created in this phase. Each form inherits the syntactic-semantic information from its canonical form. English equivalents of Polish inflected forms are **not** derived. The derivation of an appropriate English form is left to the morphological synthesis in the translation process. Storing the dictionary as an SGML-type document aims at comfortable browsing of its contents as well as facilitating its use in an application other than the POLENG translation algorithm.
4. *Converting the dictionary into two modified finite-state automata.* This phase is executed automatically in order to optimise the access time. The first automaton stores single words. Its alphabet coincides with the Polish orthographic alphabet. Reaching a terminal state of the automaton is equivalent to finding a Polish inflected form in the dictionary. Whenever a finite state is reached, references to the table of morphological features and the table of canonical forms are obtained. Due to the references, all morphological and syntactic-semantic information as well as the canonical forms of the English equivalents of the found word are achieved. The second automaton stores the lexical phrases. The alphabet of the automaton is the set of identifiers of the words stored in the dictionary of single words. This means that the process of searching a phrase is executed "word by word" (in contrast to the "letter by letter" search in the automaton of single words).

2 The translation algorithm

The translation algorithm is non-modular: its only results are the phrasal structure and the surface form of the English expression corresponding to the Polish input. The processes of syntactic parsing, semantic analysis, transfer and morphological generation are not separated. The grammar assumed in parsing Polish expressions consists mostly of DGC rules. A specific algorithm is used for parsing verbal phrases. The algorithm deals with a characteristic feature of Polish syntax: an almost arbitrary order of verb modifiers. Types and admissible orders of modifiers of a given verb are listed in the dictionary. A few English verbs may correspond to one Polish verb depending on the type and the order of its modifiers. For each type of a modifier of a Polish verb, the type of the corresponding modifier of the English equivalent is given in the dictionary. The translation algorithm searches for the constituent verb of a clause, consults the information extracted from the dictionary and first checks for the constructions admissible for the verb. This approach enables the analysis of the Polish input expression, the choice of the appropriate English verb equivalent and the synthesis of the correct English output. If verb modifiers in the clause fulfil none of the types given in the dictionary for the predicate, then default values for the English verb equivalent and the English modifier types are taken.

The algorithm makes it possible to transfer special verbal constructions called (-T) constructions and (T-shifted)

constructions, where T denotes a type of a modifier. (-T) constructions are used in analysing object questions and relative object clauses in which a modifier of the type T does not explicitly occur, although the type T-modifier is required for the verb according to the dictionary information (e.g. in the sentence "He is a man **I was talking to**" the object of the clause "I was talking to" appears "outside" the clause). The (T-shifted) constructions are used in analysing sentences in which an object of the Polish expression should be transferred into the subject of the English expression (e.g. the Polish sentence: "Nie powiedziano **mi** (object) o tym"; the English translation: "**I** (subject) have not been told about that").

In a Polish sentence verbs are characterised both by pre- and post-modifiers. However, the sequence of words between the subject and the predicate in a sentence is subject to a number of constraints and is therefore amenable to deterministic parsing. This deterministic part of the algorithm has a notable impact on the effectiveness of the translation process.

A few heuristic methods have been developed in order to limit the search space and thus achieve better efficiency. The "method of filters" consists in checking "negative rules" first. The success of a negative rule is equivalent to a failure of the hypothesis. The method "replace by alternative" consists in replacing two rules with the same left-hand symbol and the same beginning of their right sides by one rule - e.g. two rules **A→BC**, **A→BD** are replaced by one rule **A→B(C or D)**. This improves effectiveness because of a single (rather than double) attempt to expand the same non-terminal symbol to the given string of terminals. The method "from longest to shortest" says that if a symbol **A** occurs as a non-last right-hand symbol of a rule - e.g. in the rule **D→AE** - and the grammar includes more rules than one to replace the symbol **A** - e.g. rules: **A→BC**, **A→B**, then it is more effective to make the algorithm check the "longer" rule **A→BC** before checking the "shorter" rule **A→B**. This makes it possible to block backtracking in the rule **D→A!E**. The method "replace symbol by parameter" may be used when right-hand sides of productions for different symbols start with the same (sequence of) symbol(s). For example, two rules **A→BC, D→BE** may be replaced by one rule: **F(P)→ B((C and P is P1) or (D and P is P2))**.

3 Status of the system

Currently the POLENG dictionary consists of about 2000 lexemes which corresponds to about 30000 inflected forms, mainly from the domain of computer science. The research plans for near future include consulting a large corpus of computer texts in order to create a bilingual dictionary which will enable the translation of a wide range of Polish computational texts into English.

The starting point for the translation algorithm was a parser of Polish sentences described in (Szpakowicz, 1986). Most of the expressions parsed by that system are transferable in the system POLENG. There are still a lot of grammatical, syntactic and semantic problems to be solved. The problem of assigning a correct tense of the English output (there are only 3 tenses in Polish) is currently solved on the basis of the surface structure of the Polish input expression (the solution is far from perfect). The problem of determiners is not solved at all (all noun groups are assumed definite). The solutions of these two problems will be sought in the near future.

References

Szpakowicz S.1986. Formalny opis skladniowy zdan polskich. Wydawnictwa Uniwersytetu Warszawskiego. Warszawa.

Courtois B. 1990 : Un système de dictionnaires électroniques pour les mots simples du français. In *Langue française*. Larousse. Paris.

Roche E. 1992. Experiments in dictionary compression. Paris.

Jassem K. 1997. Classification of Polish nouns for the electronic flexional dictionary. In *Sprache und Datenverarbeitung. International Journal for Language Data Processing*. To appear.