POLENG - Adjusting a Rule-Based Polish-English Machine Translation System by Means of Corpus Analysis

Krzysztof Jassem, jassem@amu.edu.pl

Filip Graliński, filliposg@venus.wmid.amu.edu.pl

Grzegorz Krynicki, krynicki@ifa.amu.eu.pl

Adam Mickiewicz University, Poznań

1. Introduction

POLENG is a transfer text-to-text Polish-English MT system. It operates on an electronic dictionary annotated for morphological, syntactic, and partly semantic information. The dictionary is based on the corpus of Polish texts from the domain of computer science. The local version of the system has the form of a two-window editor. The client-server version is available at http://poleng.wmid.amu.edu.pl/. The system is intended to deal with HTML Polish texts as well as co-operate with Microsoft Word. The detailed structure of the system is described in (Wypych 2000). In our short presentation we will concentrate on corpus-based methods we adopt in order to improve the efficiency of the translation.

2. General properties of the POLENG system

1.1 Translation algorithm

The POLENG system is based on a uni-directional translation algorithm. The tree for representing English output expressions is built in the course of the analysis of Polish input. The translation algorithm employs transfer rules. The rules for parsing Polish input conform to the Phrase Structure Grammar. Definite Clause Grammar rules are applied to the analysis of nominal, adjectival and adverbial phrases. Verb phrases are parsed with an extended grammatical formalism because of the free order of verb arguments in Polish. The main reason behind our adopting PSG is that the grammars currently available for Polish (Szpakowicz 1983, Świdziński 1992) have been written in the formalism close to PSG. A few alternative approaches, like Dependency Grammar or Head-driven Phrase Structure Grammar, were also considered as potential formalisms for the POLENG parser. None of these grammatical formalisms, however, has been adopted for Polish to the extent that would allow its practical implementation. The research on the HPSG for Polish conducted in Warsaw by the team of Leonard Bolc is still in its development phase. Our own attempts at incorporating some DG techniques into the translation algorithm have been only tentative.

Moreover, in the context of translating the WWW pages, PSG has an additional advantage of being more convenient for handling HTML tags (Jassem 2000).

1.2 Dictionary

The POLENG translation algorithm consults a dictionary stored in the SGML format. Each entry in the dictionary is supplied with inflectional information about the Polish lexeme and all of its English equivalents. An entry may also contain a description of the syntactic behaviour of the Polish headword and its equivalents (e.g. specifiers and complements for verbs, nouns, adjectives and prepositions, predicative vs. attributive character for adjectives, syntactic role of adverbs and conjunctions, etc.), some basic semantic information about nouns (’human’, ’animate’, ‘abstract’, ‘concrete’, etc.), as well as contextual qualifiers (‘internet’, ‘hardware’, etc). Within each entry, idiomatic expressions can be included for its headword.

For the sake of shortening the access time, the dictionary is converted into two finite-state automata. One of them stores single words and allows “letter by letter« search, the other one stores the lexical phrases and allows “word by word« search.

2. Corpus Analysis in the design and optimisation of the POLENG system

2.1. Collection Procedure

The first step in the creation of the dictionary for the POLENG system was the semi-automatic collection of the corpus of Polish computer texts. The main sources used in the compilation of the corpus were selected web pages published on the Internet. The resulting corpus of Polish computer texts consisted of over 1,100,000 word tokens (Graliński 2000).

2.2. Lemmatisation

The whole corpus had to be subjected to morphological analysis that would produce the complete list of lemmas contained in the corpus. After the lemmatisation of the corpus by means of the SAM lemmatiser (Szafran 1997), the list of 19,842 lexemes was obtained. In the process of lemmatisation, however, it turned out that 3044 Polish lexemes were not present in the dictionary of the lemmatiser. One of the reasons was that Polish texts related to computer science contain a substantial number of neologies, unexpected derivations and jargon coinages, whilst SAM does not perform any derivational analysis. The list of such words formed the basis for defining derivational rules that would allow for recognition and translation of those words .

2.3. Automatic derivational analysis

It is a well-founded intuition that storing all the possible derivatives as separate dictionary entries is uneconomical if not unfeasible. Let's consider compound adjectives such as *japońsko-angielski* (Eng. Japanese-English), *domowo-biurowy* (Eng. *home-office*). Assuming that we have a set of N adjective entries and all two-element combinations of these adjectives are possible, we would have to create N^2 lexical entries to include each one of them. Thus, formulating a single rule for compounding seems to be a better solution than laborious typing-in of all the derivatives.

A set of derivational rules has been formulated that allows for the recognition and translation of word formations absent from the dictionary and derived from words present in the dictionary. For example, consider the rule for processing the adjectives that end with ***-cyjny*** (such as *nota**cyjny***, Eng. *nota**tional***) and that are coined from the nouns with the suffix *–cja*, (i.e. *nota**cja,*** Eng. *notation*):

```
$/cja+cyjny$:$.pos$=noun:$pinfl$=A10:$.e$al:$einfl$=A1
```

The rule is devided into five fields separated by colons. The first field is the pattern of derivation: "replace the suffix *-cja* with the suffix *-cyjny*«. The second field specifies that the base word has to be a noun. The third and the fifth field define the inflectional codes for Polish derivatives and for their English equivalents respectively. Finally, the fourth field is used to generate English equivalents: "add the suffix *–al* to the equivalent of the base form«.

In all, the analysis of the corpus resulted in the formulation of 34 rules for different types of prefixation, suffixation and compounding.

The module for derivational analysis was tested on the 3044 Polish lexemes that had previously been not recognised by the SAM lemmatiser. The result was that 50% of them could be identified by the word-formation analysis presented above. Out of the identified word formations ~78% were translated correctly or comprehensibly.

2.4. The use of the corpus of Polish texts in the lexicographic work

In their description of the entries in the electronic dictionary, lexicographers were supposed to describe the linguistic properties of Polish lexemes and find their suitable translation according to the predefined formalism. Lexicographers were equipped with a specially designed concordancing program and could refer to the corpus at any time. The amount of detail required on the English side of the dictionary entry was conditioned by the context the entry headword appeared in, as testified by the corpus:

- appropriate translations of the corpus occurrences of the given headword were described in the entry of that headword as its English equivalents;

- English equivalents that could not be distinguished by means of the syntactic-semantic formalism were attached priorities depending on how often they were considered appropriate equivalents of the headword lexeme in the corpus;
- within the boundaries of the adopted syntactic-semantic formalism, the syntactic and semantic relations formed by the corpus occurrences of the headword were to be described in the dictionary and provided with their English equivalents.

3. Optimisation of the dictionary and the POLENG translation system

Machine Translation Systems that apply parallel corpora in the process of translation have been developed for several years by now (Somers 1998). Example-based Machine Translation involves considerable difficulties: it requires the collection of large parallel corpora and assumes the application of advanced methods of source and target language text alignment. Additionally, sophisticated methods need to be developed in order to match and recombine the examples in both languages.

A technically easier approach, that may at the same time prove effective in solving certain linguistic problems, relies on the application of the target language corpus only. This approach can efficiently support the translation process based on rules and dictionary.

3.1. Translation of Polish Genitive Noun Phrase into English

Polish noun phrases that consist of a noun followed by a noun in genitive case are usually rendered into English as *compound nouns* (a noun phrase composed of two nouns separated by a space or hyphen) or as *periphrastic possessive* (a noun phrase with *of*), e.g. the English equivalent of the Polish Genitive Noun Phrase *transfer danych* is *data transfer*, while the phrase (e.g. *zawartość pliku* is translated as *contents **of** a file*). The assumptions and the task can be formulated as follows:

Assumptions:

1) the input phrase consists of two Polish nouns: N_1 and $N_{2\ GENITIVE}$,
2) the English equivalents of both Polish nouns are given: $E(N_1)$ and $E(N_2)$,
3) the correct translation of the input phrase is only the one of the two phrases: $E(N_1)\ E(N_2)$ and $E(N_1)\ of\ E(N_2)$.

The task:

Find the correct translation of the input phrase.

This problem can be solved by means of English corpus in the following way:

1) count the occurrences of the phrases of the form $E(N_1)\ E(N_2)$ and $E(N_1)\ of\ E(N_2)$ in the corpus of English texts,
2) choose the more frequent one as the equivalent of the input phrase.

An analogous problem and a similar solution for Japanese-English translation is discussed in (Sumita, Iida 1991).

3.2. The Choice of the Best Equivalent for a Polish Word

It is often the case that lexical information given for a Polish lexeme with more than one English equivalents does not suffice to determine the most appropriate equivalent in a given context. In some cases the corpus of English texts may help to chose the most appropriate equivalent for the Polish lexeme. This may be achieved by means of the following algorithm:

1) search the dictionary for the equivalents of words that surround an ambiguous word in the source language text,
2) find the occurrences of similar portions of text in the target language corpus (by ‘similar portions’ we mean ones that contain equivalents of the neighbouring words and one of the several equivalents of the word we want to disambiguate),
3) choose the equivalent with the greatest number of occurrences.

At present, we are planning to use Internet search engines (such as AltaVista) to evaluate this technique: the ‘corpus’ will be all the WWW pages written in English.

References

[*]F. Graliński. 1999. Applying a Corpus of English Texts to Machine Translation from Polish to English. *Speech and Language Technology, Vol. 3.* Poznań.

[*]F. Graliński.2000. The Report on Lemmatisation and Disambiguation of a Corpus of Polish Texts Related to Computer Science. *Speech and Language Technology, Vol. 3.* Poznań.

[*]F. Graliński. G. Krynicki.2000. Word-formation Analysis in Polish-to-English Machine Translation. *Speech and Language Technology, Vol. 3.* Poznań.

[*]K. Jassem. 2000. Dealing with Free Order and Non-Language Markers in a Top-Down-Left-First Algorithm, *Speech and Language Technology, Vol. 4.* Poznań.

H. L. Somers. 1998. "New Paradigms" in MT: the State of Play now that the dust has settled. *10th European Summer School in Logic, Language and Information*. Saarbruecken.

E. Sumita, H. Iida.1991. Experiments and Prospects of Example-Based Machine Translation. *19th Annual Meeting of the Association for Computational Linguistics*. Berkley, California.

K. Szafran. 1997. SAM-96 — The Morphological analyser for Polish. DIALOG'97. Yasnaya Polyana.

S. Szpakowicz 1983. Formalny opis składniowy zdań polskich. Wydawnictwo Uniwersytetu Warszawskiego. Warszawa.

J. Świdziński 1992. Gramatyka formalna języka polskiego. Warszawa.

[*]M. Wypych. 2000. Designing a Client-Server Architecture for the MT system POLENG. *Speech and Language Technology, Vol. 4.* Poznań.

[*]These papers are available at http://poleng.wmid.amu.edu.pl.