PaTrans - A Patent Translation System

Bjarne Ørsnes, Bradley Music & Bente Maegaard
Center for Language Technology
Njalsgade 80
2300 Copenhagen S
Denmark
{bjarne,music,bente}@cst.ku.dk

Abstract

This paper describes PaTrans - a fully automatic production MT system designed for producing raw translations of patent texts from English into Danish. First we describe the backbone of the system: the EUROTRA research project and prototype. Then we give an overview of the translation process and the basic functionality of PaTrans, and finally we describe some recent extensions for improving processing efficiency and the translation quality of unexpected input encountered in real-life texts.

1 Introduction

PaTrans [1] is a fully-automatic machine translation system designed for English-Danish translation of patent texts. It is based on the linguistic specifications and to some extent on the software of the EUROTRA project of the European Community (Copeland et al., 1991a; Copeland et al., 1991b). PaTrans consists of a core grammar and translation module and a host of peripheral utilities: term databases, general databases, editors for pre- and postediting, document handling facilities, facilities for creating and updating term databases. In this short presentation we will concentrate on the grammar, lexicon and translation module and on some of the new features of PaTrans.

2 From EUROTRA to PaTrans

EUROTRA was the European Community MT research programme. The Community started the programme in 1982, with the goal of creating an advanced system for automatic translation capable of treating all the official working languages of the Community. When the programme finished in 1992, it had delivered a huge amount of research results and an implemented prototype of a multilingual translation system. The PaTrans development relies on the prototype resources (Maegaard and Hansen, 1995), the system architecture and linguistic specifications, as well as on the experienced staff created by EUROTRA.

2.1 The EUROTRA Prototype

EUROTRA was a transfer-based multilingual MT project. Because of the multilinguality, the prototype was quite "clean" in terms of separate modules for analysis, transfer and synthesis of the various languages and language pairs.

2.1.1 Software

The software component consisted of the translation kernel, used for analysis, transfer and generation. The translation kernel had mechanisms for treating grammar rules, dictionary information and mapping rules.

2.1.2 Lingware

For all languages, the project produced a large grammar and a general language dictionary. Though insufficient for the task at hand, the PaTrans development could build on the English and Danish grammars and dictionaries, as well as on the transfer module from English into Danish.

2.2 Customizing EUROTRA

Patent texts are characterised by the vocabulary they contain: terms belonging to the field treated, e.g. chemistry, and patent document terms of a more legal nature. But patent documents are also characterised by the frequency of some linguistic phenomena and the absence of others, e.g. we had to develop a treatment of lists and enumeration, and conversely we could simplify the treatment of modality considerably. The current maintenance and further development of the system continues this text type specific line. The success of the system is mainly based on this fundamental principle of tailoring it to a specific text type and subject field.

[1] PaTrans was developed for Lingtech A/S.

3 An overview of the Translation Process

3.1 Document handling

The document handling step has four main functions:

- **Format Preservation** Input to document handling is a text from a text processing system which has been marked up in SGML. The SGML codes denote e.g. titles, paragraphs, text segments that should not be translated, etc. All information about document layout is stored separately and taken away from the translation process.
- **Formula Recognition** The document handler automatically recognises certain text typical untranslatable units, such as chemical formulas and tables.
- **Term Recognition** Terms and multi-word units are also recognised at this stage. In this context, words are treated as terms if they are subject specific or if they have a unique translation in the given text type. They are recognised during text handling and have their translation equivalent attached to them along with morphosyntactic information for both source and target language.
- **Segmentation** Finally the text is separated into units for translation i.e. sentences for which various recognition patterns have been set up. In some patent texts of specfic subject fields, the sentences are incredibly long. In these cases, there is no point in trying to arrive at a complete parse of the whole sentence, since the parse is most likely to fail and processing will be too space and time consuming. Therefore the document handler attempts to arrive at a meaningful partition of the sentences by identifying sentence internal boundaries and submitting the individual subparts for translation.

3.1.1 Disambiguation

Before the text is passed on to the parser, it is subjected to a thorough process of disambiguation. This is one of the new features of PaTrans compared to the EUROTRA model and will be discussed in detail below.

3.1.2 Source language analysis

Since PaTrans is based on the transfer translation model the surface strings of the text are sequentially transformed into an intermediate representation defined by several mapping principles.

During source language analysis the sentences are assigned a surface syntactic structure. This surface syntactic structure is converted into a language-neutral transfer representation ordering the constituents of the sentence in a canonical order with heads preceeding arguments and arguments preceding modifiers (Copeland et al., 1991a). The transfer representation is a reflection of the argument structure of the predicates where information about surface syntactic realization appears as features on the individual nodes. Function words (conjunctions, determiners, prepositional case markers) are featurized and tense/aspect and negation represented in language-neutral features.

The output of source language analysis is thus a tree with multilayered information including syntactic and morphosyntactic features, as well as the syntactic/semantic relationships between the predicators and the arguments.

At all levels, sets of preference rules based on heuristic principles select among competing analyses, e.g. for PP-attachment (Bennett and Paggio, 1993).

3.1.3 Transfer

PaTrans adheres to simple transfer, i.e. the substitution of source language lexical units with target language lexical units by means of lexical transfer rules, [2] while the source language structural representation is mapped directly onto the target language transfer representation which is input to the generation module. There are two main reasons why complex transfer (i.e. transfer where the structure of the input representation is altered) is kept at a minimum:

- Complex transfer is costly inasmuch as the general applicability of the rules is usually very restricted.
- A transfer rule applies to any object matching its left-hand side and performs the mapping defined on the right-hand side. Due to the 'fail-soft'-mechanism (discussed below), the structure of the objects which the transfer rules must apply to cannot be fully predicted. In order for complex transfer to work in all cases, rules must be set up not only for correctly parsed input structures, but also for the special fail-soft structures. For this reason, complex transfer is costly and is only used for frequent phenomena considered crucial for good translation, e.g. converting certain English *ing*-forms into Danish relative clauses.

3.1.4 Target syntactic generation

During generation, the transfer representation is mapped onto a target syntactic structure through intermediate representational levels. At the first level, the target language lexical units are looked up in the lexical database and monolingually relevant features are calculated on the

[2] Recall that this only applies to words of the general vocabulary which require disambiguation during analysis and not to terms

basis of the language-neutral representation, e.g. tense and aspect.

At the second level (the relational level) surface syntactic functions are calculated and certain function words, such as prepositional markers are inserted. Finally, the relational structure is mapped onto the level defining the constituent structure of the target language sentence. At this level all information with independent lexical expressions is present.

3.1.5 Target morphological generation

PaTrans has a highly developed morphological module which provides an almost complete coverage of Danish inflectional morphology. The module is based on structure building rules which allow for downwards expansion. Regular inflection, syncope and gemination is accounted for while only completely irregular word forms will have to be coded in their entirety. PaTrans also has a limited strategy for translating compounds compositionally. Generally, compounds are coded in the (terminological) dictionaries, but the parser tries to translate compounds which are not coded in the dictionaries by translating their individual subparts.

3.1.6 Document generation

Finally, the document generation module inserts all SGML-markers and all items which have been marked as untranslatable (tables, formulas, numbers etc.), and a separate conversion programme converts the output into WordPerfect format. [3]

4 The lexica

PaTrans distinguishes two kinds of vocabularies: the general vocabulary and the terminological vocabularies.

- The general vocabulary is stored in a monolingual English dictionary, a monolingual Danish dictionary separated into a into syntactic and a morphological level, and a bilingual transfer dictionary.
- The terminology is divided into subject specific databases. As PaTrans is used for a number of different subject fields, the priority of the databases is user-defined and flexible. The user specifies which term bases are to be used for a translation job, and in which order of priority. When a term is found in one term base, it is not looked up further in the subsequent databases.

4.1 PaTerm Coding Tool

For ease of maintenance and updating, PaTrans has a special coding tool. As mentioned above, the PaTrans term bases contain terms as well as words and expressions which behave like terms, i.e. which have unique translations. New terms occur in each and every patent document which is submitted for translation. Consequently, it is important that the user, who is not necessarily a computational linguist, can encode terms in an efficient and precise way. The PaTerm coding tool provides a screen with fields to fill in, and in most cases an answer is proposed by the system, so that the user has to make just one acceptance keystroke. Care has been taken to present the most frequent, and therefore most probable, answer on the top of the list. PaTerm asks the minimum number of questions and computes the remaining linguistic information from the answers received. This also saves time for the user.

5 Special Features

5.1 Error Recovery

Since the system runs in a practical environment, it must never fail to produce an output, even if it encounters an unanalysable sentence. Consequently, a fail-soft mechanism was introduced. The fail-soft mechanism works at all levels of representation. If the parser fails to assign a well-formed structure to the input, a path is selected from the chart which spans the greatest amount of the input and already created constituents are collected. The quality of fail-soft output varies considerably and recent work has attempted to improve the results of fail-soft. Disambiguation of individual words, the selection of appropriate readings and the determination of individual constituents at a very early stage are crucial in arriving at a 'best-fit' parse.

Interestingly, there are some fundamental difficulties in combining advanced MT with fail-soft strategies. The most striking example of this is the fact that PaTrans aims at a very deep analysis of the source text and at the same time the formalism allows for non-monotonic mappings between levels of representation. Due to the unexpected and to some extent unpredictable structure of fail-soft analyses, subsequent grammar rules may fail to apply, resulting in output representations where information e.g. about the degree of adjectives and other information stemming from function words has been lost. Current efforts consequently aim at preserving information at all levels.

5.2 Tagging

Before the text is submitted to the parser, the text is tagged, i.e. the tagger tries to determine the part-of-speech of the individual words based

[3] Until now, all texts have been delivered in WordPerfect, but the conversion programme may of course be adapted to other text processing systems.

on local cooccurrence restrictions. There are two reasons why the tagger has been integrated into the system:

- Since the overall translation system is unification-based, words are disambiguated by the application of all possible rules, which is highly inefficient.
- If the sentence is fail-softed, one intermediate analysis is picked from the chart, which means that all words may not have been disambiguated properly by the grammar rules. If, however, the words have been disambiguated and impossible readings have been discarded prior to parsing the 'best-fit'-parse is considerably better than it would otherwise have been.

The tagger is a public-domain, rule based tagger. It has been trained on a corpus of the Wall Street Journal and on patent texts within the subject field. In addition, it has been augmented with several 'local' contextual rules developed by the linguists working with PaTrans. The integration of the tagger has not only provided for more effecient processing but, more importantly, also for a higher quality of the translations of fail-softed sentences. Current efforts aim at improving the performance of the tagger.

5.3 Preparsing

The original EUROTRA-parser has been augmented with special rules which apply before the actual grammar rules (Music, 1993). The goal is to enable more efficient handling of long sentences that are otherwise unprocessable given moderate resources. With pre-rules, sentences are segmented via pattern-matching, before they are sent to the parser. In this way, the number of parse paths that the system has to consider is reduced considerably.

To give greater power to the preparser, pre-rule application has been made cyclic. This means that the output from one rule application (or one application cycle) is used as input to a new cycle which starts at the beginning of the rule set. In principle then, any rule can feed (i.e. create the preconditions needed for application of) any other rule, while at the same time allowing prioritization of rules. The pre-rules not only add structure to the input, they are also used for lexical disambiguation based on collocatives and immediate context. Where the rule based tagger described above is able to determine the part-of-speech of individual words based on prior training and contextual rules, pre-rules can select individual readings of words within the same part-of-speech. Pre-rules have been developed for lexical disambiguation and for parsing of adverbial phrases, complex verb groups, coordinated *that*-clauses, indexed lists, valency-bound prepositional phrases and explicitly marked intervals (e.g. *from ... to*, *between ... and*). The effects of pre-rules are twofold: On the one hand they assign structure to the input at a shallow level, which nevertheless increases processing efficiency considerably, on the other hand they also improve fail-soft results since inappropriate readings of words in a given context are discarded at an early stage.

6 Performance

PaTrans is in everyday use at the translation agency Lingtech where it is being used for all texts which are suited for it in its current version, i.e. chemical, biochemical, medical etc. patents, and gradually also a considerable amount of mechanical patents. PaTrans is making the translation process faster and more efficient, and it has proven to be a good business for Lingtech, saving around 50% of the raw translator cost.

7 Conclusion

PaTrans is a running production translation system producing cost-effective raw translations of patent texts. But PaTrans is also a project which combines academic research and practical applications and which has shown that MT is viable in limited domains. Current work concentrates on improving the coordination of the rule-based part of the system and the fail-soft component.

References

Bennett, P. and Paggio, P., editors (1993). *Preference in Eurotra*, volume 3 of *Studies in Machine Translation and Natural Language Processing*. Commission of the European Communities, Luxembourg.

Copeland, C., Durand, J., Krauwer, S., and Maegaard, B., editors (1991a). *The Eurotra Linguistic Specifications*, volume 1 of *Studies in Machine Translation and Natural Language Processing*. Commission of the European Communities, Luxembourg.

Copeland, C., Durand, J., Krauwer, S., and Maegaard, B., editors (1991b). *The Eurotra Formal Specifications*, volume 2 of *Studies in Machine Translation and Natural Language Processing*. Commission of the European Communities, Luxembourg.

Maegaard, B. and Hansen, V. (1995). PaTrans - Machine Translation of Patent Texts. From Research to Practical Application. In *Convention Digest: Second Language Engineering Convention, London*, pages 1–8.

Music, B. (1993). Preparsing in the PaTrans MT System. In *Bits & Bytes: Datalingvistisk Forenings Årsmøde nr. 3*, pages 82–90. Institut for Sprog og Kommunikation, Odense Universitet.