The machine translation system MATS - past, present & future

Per Weijnitz, Anna Sågvall Hein, Eva Forsbom, Ebba Gustavii, Eva Pettersson, Jörg Tiedemann

Department of Linguistics and Philology
Uppsala University, Uppsala/Sweden
{perweij,evafo,ebbag,evapet,joerg}@stp.ling.uu.se
anna@lingfil.uu.se

Abstract

This is a status report of the rule-based machine translation system MATS (Sågvall Hein et al., 2002). It describes the development into its current state, which has recently been focused on extending the linguistic resources to new domains and improving robustness. In cases where an input sentence is not fully processable by the system, partial results are taken care of and used in producing a translation by complementary methods. A road-map for future development is also sketched.

1. Introduction

MATS (Sågvall Hein et al., 2002) is a rule-based machine translation (MT) system in the traditional transfer paradigm, with separate modules for source language analysis, transfer, and target language generation. The core of the system is the MULTRA research prototype. The main goal in creating the MATS system was the scaling-up of MULTRA for industrial use. This was done in a cooperative project between with the Department of linguistics at Uppsala University, and Scania CV AB in Södertälje, as the main partners. In addition to the design and implementation of MATS, the project work included the redesign, porting and integration of MULTRA, the redesign and implementation of the dictionaries of the language modules as a lexical database, and the extension of the dictionaries and the grammars. The language domain was automotive service literature.

In this paper, we describe the development of the MATS system from the original version to its current state and the road-map for the future. We will start with a look back at the system preceding MATS, the translation engine MULTRA. The definitions *primary methods* and *primary modules* refer to the algorithms and implementations of the rule-based foundation of MATS. *Fall-back methods* and *fall-back modules* refer to the new extensions to MATS, that are used to complement their primary counterparts.

2. Past

2.1. MULTRA

MULTRA is a transfer-based system, primarily intended for high-quality translation within limited domains. Distinctive features of MULTRA are strict modularity, well-defined transfer based on unification, well-defined generation based on unification and concatenation, non-deterministic processing, a well-defined preference machinery in transfer and generation, and powerful tracing options. Swedish has been the only source language and English the primary target language.

Analysis in MULTRA is carried out by means of a chart parser, Uppsala Chart Processor, UCP (Sågvall Hein, 1983). Grammars and dictionaries are formulated in a procedural formalism (Ahrenberg, 1984; Sågvall Hein, 1984b; Dahllöf, 1989). Analysis structures are expressed as trees of attributes and values. Alternative sentence analyses are arranged in a preferred order (Sågvall Hein, 1994) by means of rules in a separate preference component. The preference rules are defined by the grammar writer.

Transfer and generation rules are expressed in PATR-like formalisms that were specifically developed for MULTRA (Beskow, 1993; Beskow, 1997a; Beskow, 1997b). Lexical and grammatical transfer rules are formulated in the same formalism, facilitating the transfer of lexical units in context. The transfer and generation rules are partially ordered; a more specific rule has precedence over a less specific one, i.e. a lexical transfer rule taking a context larger than the word itself into account will have precedence over a lexical rule applying to a word in isolation. For the transfer to be complete, all attributes in the analysis structure have to be mentioned by a rule. In other words, there has to be one transfer rule for each unique attribute in the feature structure.

2.2. MATS

MATS is an MT system into which MULTRA has been integrated. Processing in the MATS system proceeds in a number of distinct steps from an SGML version of the source document to an SGML version of the target document via MULTRA. The need for morphological analysis and generation has been replaced by a relational lexical database for both source and target language look-ups.

Following the design principle of the MULTRA machine translation system, MATS is strictly modular. Each step in the translation is carried out by a stand-alone module connected serially in a unidirectional data pipe, allowing a module to inspect the output of all preceding modules. The communication is text based which contributes to transparency and trace-ability, as it is possible to inspect the intermediate result coming from each processing step. These are the major processing steps:

1. Decomposition of the source SGML document.
2. Tokenisation of the sentences into words or phrases.
3. Look-up of morpho-syntactic and semantic information in the source database, and a default translation.
4. Parsing of the sentence into a complete analysis. UCP is replaced by UCP Light, which is a new implemen-

tation in C. A preference mechanism orders the analyses, in case the input is ambiguous.

5. Transfer of source language structures into target language structures. Transfer rules may add, delete or replace data to adapt the sentence representation to the target language. They may also identify contexts where the default translations retrieved from the database are not appropriate, and provide better alternatives.

6. Generation of a target language string. This is based on a grammar describing the target language syntax.

7. Look-up of target language full-form words and phrases. A primary database key consists of a lemma and morpho-syntactic information, represented as a code.

8. Phonotactic adjustments. The phonotactic module makes it possible to modify generated target strings on the basis of the immediate preceding and succeeding context. The rule format admits constraints both on the surrounding surface strings and their corresponding feature structures. The only rule implemented so-far, changes the English indefinite article from *a* to *an* in the appropriate contexts.

9. Text finish, such as capitalising the initial characters of sentences.

10. SGML document reconstruction. The SGML document is reconstructed with the translated segments in the right place.

The coverage of fully translated sentences in the document is reported by an evaluation module, as well as a module-by-module error overview for not translated sentences (Sågvall Hein et al., 2003).

MATS is controlled by a web based interface targeted at developers of the linguistic resources and evaluators of the system.

3. Present

3.1. The MATS system

A general problem with rule-based systems is robustness; translation usually fails if the input is not covered by the grammars or dictionaries.

MATS is being redesigned with the focus on robustness. When a module fails, a secondary fall-back method is used to do its processing. The goal of the redesign is to have a system that is capable of dynamically lowering the ambition of good translation quality when the need arises. By using the available partial results from the preceding modules, parts of or all of the sentence may be well translated without full support from the grammars.

The following extensions are currently implemented:

- A mechanism that makes use of incomplete parses. When the parser is not able to produce a complete analysis, it selects a set of analyses for substrings that as a set cover the complete input string. Each analysis is then translated separately. Finally, the translated substrings are concatenated. The concatenation operation introduces a syntactic dependency between the syntax of the translated string and the input string. The success of this procedure depends to a large extent on the language pair, and how the string is segmented into partial analyses. The latter can be refered to as the edge selection problem, as substrings are represented by passive edges in the chart. A successful edge selection method will select edges that are likely to be well translated by the subsequent processing steps, and try to avoid edges with agreement dependencies on other edges. The reason for this is that the subsequent processing steps currently translate each part independently of the other parts, in the same way they translate complete sentences independently of other sentences.

- A robust transfer mechanism. The parsed analysis is represented as a tree of attributes and values. The transfer mechanism transforms that tree into a tree representing the equivalent analysis in the target language. This involves deleting, adding and replacing features and structures in the tree. An important task is to detect contexts in which a token's default translation is inappropriate and a context-sensitive translation should be generated. The robustness consists of a new default rule that applies when there is no applicable rule for the current tree node in the transfer rule grammar. It copies the tree node from the source language side to the target language side. This means that the transfer grammar does not have to describe all possible analyses, as the default copy rule will apply for the complement of the structures covered in the transfer rule grammar. In other words, only those transfer rules that represent structural shifts and context-sensitive lexical translations need to be defined in the grammar. The completeness criterion may still be maintained due to the default principle.

- A fall-back generation module. The task of MULTRA's generation module changed when incomplete parses were put to use. The input structure used to represent full sentences only. Now an input structure may also represent a part of a sentence, when a sentence has been split by a partial analysis. Each such partial structure is then generated on its own. In subsequent processing steps, the generated output of each chunk is concatenated in the same order. If the generation grammar does not cover the structure to generate, representing either a full sentence or a part of a sentence, control is handed over to a fall-back module. As each input structure is always processed by the primary generation module first, the fall-back module will only be used for the parts of a sentence that are not covered by the grammar. The module is still being developed, but is currently working in the following way. It retrieves all token translations from the transferred input structure, and simply orders them in the original source language order. The success of this

crude method, basically the simplistic direct translation method, depends on the language pair, and what segment is being processed. In Swedish-English, it works best for short phrases like NPs.

3.2. Linguistic resources

Lexical data used by the system are stored in a central relational database, MatsLex (Tiedemann, 2002). Tools and interfaces have been implemented to maintain the database and its content. The database comprises a set of tables with morphological, syntactic, and semantic information with appropriate relations between them. The same structure is used for each language in the lexicon. New languages can easily be integrated by creating a new copy of the relational structure. Lexemes from different languages are linked using bilingual link tables. Source language lexemes are connected with their default translations in the target language. Link tables are directional, i.e. source language lexemes are linked to target language lexemes. However, a directional link table can also be used in the inverse direction if necessary.

Command-line tools and web-interfaces are used for maintaining and updating the database. The translation engine does not use the lexicon directly. Instead, run-time lexicons are extracted from the database containing necessary information for different steps in the translation process. This ensures a time-efficient and consistent behaviour of the translation engine and makes it possible to compare different versions of the lexicon.

The original lexical database, developed for the truck-manufacturing domain, is compiled from approximately 8,200 Swedish-English entries. The development of this dictionary was a major step towards an industrial use of the system for translating documents within this domain (Sågvall Hein et al., 2002).

At a later stage, a new lexical database was developed to fit the domain of agricultural EC documents (Weijnitz et al., 2004). Dictionary entries were extracted from a parallel corpus comprising agricultural EC reports provided by the European Translation Service (SDT) within the project *Extension of EC Systran to Danish and Swedish into English, Commission contract SDT/MT2003-1*. The corpus contained approximately 71,000 Swedish tokens and about 86,000 English tokens, and the resulting lexical database includes almost 6,000 Swedish-English entries. This new domain is quite different from that of truck manuals, in that the truck literature has shorter and less complex sentences and exposes a more extensive use of imperative forms than the agricultural reports. Thus, a thorough adjustment of the grammars is needed to adapt them to the new domain. This work has been initiated, which among other things has resulted in an extended coverage of subordinate clauses and complex nominal phrases.

Currently, there is work in progress to create a new lexical database and adjust the grammars to cover the domain of reports from the Swedish Security Service (SÄPO). For this purpose, SÄPO has provided extractions from their term database, containing approximately 13,200 Swedish base forms and their English translations, and further a parallel corpus comprising four reports of all in all around 2,200 tokens per language.

4. Future

MATS has evolved from being a pure rule-based transfer system, into a system with fall-back methods that are related to direct translation. Future efforts will have to go into further development of both linguistic resources, algorithms and implementations. The coverage of the lexical database will be extended, as new domains are added. The analysis grammar will be refined with the aim of covering more phrase types and sentence types. The transfer rules will be adapted to match the advancements of the analysis grammar and new contexts in which the default translations need replacement.

The system will be developed in parallel with the linguistic resources. A number of projects are currently in focus:

- Morpho-syntactic disambiguation and word sense disambiguation. Currently, there is no module explicitly handling morpho-syntactic disambiguation of the input strings. When the input is successfully parsed, most morpho-syntactic ambiguity is however resolved, or may otherwise be handled by means of preference rules. Adding the fall-back mechanism makes the problem of morpho-syntactic ambiguity more emergent. Often the partial analyses do not provide enough context for the ambiguity to be resolved. The obvious solution will be to extend the pipe with a morpho-syntactic tagger. A related issue concerns word sense disambiguation, which could further reduce ambiguity and help the fall-back modules.

- Improved edge selection method. When the parser is not able to produce a complete analysis, it selects a set of analyses for substrings that as a set cover the complete input string. The selection is in itself a parsing task. It is not desirable to implement the edge selector as a parser using a grammar, as such rules would be better to implement in the main parser grammar directly. Instead, it may use heuristic rules or possibly a statistical model. The current implementation uses a greedy left-to-right, longest-edge-first strategy.

- Translation of productively formed compounds. Since compounds may be formed productively in Swedish, the translation of these can not be handled by simple dictionary look-up. Hybrid methods, combining rule-based analysis with statistically based selection of target constructions, have been tested for translation of German compounds to English (Rackow et al., 1992). We intend to experiment with similar strategies for Swedish compounds.

- A replacement of MULTRA's generation module is under development, which will support a wider range of operators, such as disjunction.

- A more powerful fall-back generation method is needed. There are a number of drawbacks with the current method that need to be resolved. It cannot

insert function words, and lacks the ability of knowing whether its output is likely to be syntactically correct. The new generation method should preferably not rely on a hand crafted generation grammar for complete sentences, as there is already such a grammar for MULTRA's generation module.

5. Summary and conclusions

The development of the machine translation system MATS has been described, followed by a road-map for the future. MATS has evolved from being a pure rule-based transfer system, into a system with fall-back methods that are related to direct translation. Future development will focus on both linguistic resources, algorithms and implementations.

The current version of MATS, with the full set of implemented fall-back strategies as out-lined above, has been applied to three different domains, i.e. automotive service literature, EU agricultural texts, and reports from SÄPO. The results indicate that the system is close to a commercial application, where publishing quality may be achieved after due post-editing. However, an intermediary step, before going to the market, will be setting up co-operative projects with the potential clients for customisation and additional training to a performance level that is mutually agreed upon. One such project has been set up, and one is underways. Such a commercialisation strategy presupposes additional governmental funding.

Acknowledgements

This project was supported by VINNOVA (Swedish Agency for Innovation Systems), contracts no. 341-2001-04917 (Ett regelbaserat maskinöversättningssystem för svenska [A Rule-base Machine Translation System for Swedish], MATS), 2002-02407 (Korpusbaserad maskinöversättning [Corpus-based Machine Translation], KOMA), and 2003-01580 (En svensk Systranmodul [A Swedish Systran Module]).

6. References

Ahrenberg, Lars, 1984. De grammatiska beskrivningarna i SVE.UCP [the grammatical descriptions in SVE.UCP]. In (Sågvall Hein, 1984a), pages 1–13.

Beskow, Björn, 1993. Unification-based transfer in machine translation. RUUL #24, Uppsala University.

Beskow, Björn, 1997a. *Generation in MULTRA*. Uppsala University. Department of Linguistics.

Beskow, Björn, 1997b. *Morphology in MULTRA*. Uppsala University. Department of Linguistics.

Dahllöf, Mats, 1989. *En lexikonorienterad parser för svenska [A lexicon-oriented parser for Swedish]*. Master's thesis, Gothenburg University.

Rackow, Ulrike, Ido Dagan, and Ulrike Schwall, 1992. Automatic translation of noun compounds. In *Proceedings of COLING*.

Sågvall Hein, A., E. Forsbom, J. Tiedemann, P. Weijnitz, I. Almqvist, L.-J. Olsson, and S. Thaning, 2002. Scaling up an MT prototype for industrial use - databases and data flow. In *Proceedings of the 2nd LREC*, volume V. Las Palmas de Gran Canaria, Spain.

Sågvall Hein, Anna, 1983. A parser for Swedish. status report for SVE.UCP. UCDL-R-83-1, Center for Computational Linguisics, Uppsala University.

Sågvall Hein, Anna (ed.), 1984a. *Föredrag vid De nordiska datalingvistikdagarna 1983 [Talks at the Nordic computational linguistics' days 1983]*, UCDL-R-84-1. Center for Computational Linguisics, Uppsala University.

Sågvall Hein, Anna, 1984b. Regelaktivering i en parser för svenska (SVE.UCP) [Rule-activation in a parser for Swedish (SVE.UCP)]. In (Sågvall Hein, 1984a), pages 187–199.

Sågvall Hein, Anna, 1994. Preferences and linguistic choices in the Multra machine translation system. In Robert Eklund (ed.), *Proceedings of '9:e Nordiska Datalingvistikdagarna' (NODALIDA'93)*. Department of Linguistics. Stockholm University. Stockholm.

Sågvall Hein, Anna, Eva Forsbom, Per Weijnitz, Ebba Gustavii, and Jörg Tiedemann, 2003. MATS - a glass box machine translation system. In *Proceedings of the Ninth Machine Translation Summit (MT SUMMIT IX)*. New Orleans, Louisiana, USA.

Tiedemann, Jörg, 2002. MatsLex - a multilingual lexical database for machine translation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, volume VI. Las Palmas de Gran Canaria, Spain.

Weijnitz, P., E. Forsbom, E. Gustavii, E. Pettersson, and J. Tiedemann, 2004. Mt goes farming: Comparing two machine translation approaches on a new domain. In *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*. Lisbon, PT. Forthcoming.