MATS - A Glass Box Machine Translation System

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

Department of Linguistics, Uppsala University

Uppsala, Sweden

anna@ling.uu.se, {ebbag, evafo, joerg, perweij}@stp.ling.uu.se

1 Introduction

MATS is a fully automatic machine translation system with the unification-based translation engine MULTRA as its core (see e.g. Sågvall Hein, 1993, 1997). The system was developed in a co-operative project between the Department of Linguistics, Uppsala University, the bus and truck manufacturing company Scania CV AB, and the translation company Explicon AB[1]. The focus of the project was the scaling up of MULTRA for translation from Swedish to English in the automotive service domain (Sågvall Hein et al., 2002). The scaling-up effort also implied eliminating separate morphological processing by storing the lexical data in a bi-lingual lexical database with a built-in morphology (Tiedemann, 2002).

Here we focus on the transparency aspects of the MATS system. The translation proceeds in a number of distinct steps from an SGML version of the source document to an SGML version of the target document. The output of each step is, optionally, presented to the user for inspection. The transparency of the system is most useful for grammar developers and teaching purposes. Another interface will be developed for end users.

MATS runs via a web-based interface, and if a step fails for a certain input, the corresponding part is highlighted in a colour specific to that step. The outcome of the different steps is collected and summarised in an evaluation report. The interface provides a great variety of presentation, tracing, and evaluation options

[1] The project was supported in part by VINNOVA (Swedish Agency for Innovation Systems), contract no. 341-2001-04917.

2 MULTRA

MULTRA is a unification-based translation engine of four modules: analysis, preference, transfer, and generation. Analysis is carried out by means of a chart parser, Uppsala Chart Processor, UCP (Sågvall Hein, 1983; Weijnitz, 2002) with a procedural formalism. Transfer in MULTRA is based on unification, solely; generation, in addition, includes concatenation. Transfer and generation rules are expressed in PATR-like formalisms (Beskow, 1993, 1997). Default translations of words and phrases are stored in the lexical database. For the translation of lexical units in context, contextual lexical transfer rules are defined. Processing is non-deterministic. However, transfer rules, as well as generation rules are partially ordered and a more specific rule precedes a less specific one (Beskow, 1993, 1997).

3 MATS Architecture

Following the design principle of the MULTRA 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. A protocol specifies how a module communicates with downstream modules, using channels layered on top of the transportation stream. All transmissions are text based for transparency and trace-ability. The system is transparent, as it is possible to inspect the intermediate result coming from each processing step:

Sentence extraction

The text is extracted from the input SGML document and segmented into sentence units.

Word tokenisation

The sentences are tokenised into one-word units and multi-word units.

Source dictionary lookup
For each token, the lemma, the lexeme, the morpho-syntactic code, the semantic code, and the default translation is retrieved from the lexical database.

Code expansion
The morpho-syntactic and semantic codes are expanded to feature structures, one for each token. Sentence units are represented as lists of feature structures.

Source language analysis
The list structure is parsed and a grammatical feature structure of the source language is generated.

Transfer
The source language feature structure is transferred into a target feature structure.

Generation
The target feature structure is processed by the generation module. A string of target lemmas with feature structures is created.

Code composition
The feature structures are transformed into a compact code format.

Target dictionary lookup
The full word forms, corresponding to the lemmas and their codes, are retrieved from the target lexical database, generating a string of words.

Finish
The string is finalised with an initial capital letter and the proper assignment of signs of punctuation.

Evaluation
The output is evaluated, as reported below.

Re-creation of SGML document
The SGML document is re-created.

4 Evaluation Module

An evaluation module at the end of the pipe collects information from the previous modules, and displays three summary tables.

For an illustration, an evaluation report resulting from on-going work training the system is presented in the Appendix.

The first table gives an overview of the input, showing how many words and segments, e.g. sentences or list items, the system has to handle. The percentage of unique words and segments (type/token ratio) gives an indication of how repetitive the input is.

The second table gives an overview of the system recall, a form of blackbox evaluation, showing how many words and segments the system did handle. On the word level, it gives a measure of the degree of source dictionary coverage (recall) for the current input. On the segment level, it gives an indication of the degree of grammar (and dictionary) coverage for the current input, i.e. the total number of (fully) translated segments, and the number and percentage of (fully) translated unique segments. Fully translated segments are those that passed the system with no error reports from any module, while translated segments include fully translated segments and segments that passed the system with reported word level errors.

The third table gives an overview of some error reports from the system modules, a form of glassbox evaluation, showing how many words were missing from the dictionaries, how many codes were missing from the code files, and how many segments the modules did not handle, as reported by the modules themselves. It shows, in one respect, each module's contribution to the total recall. The number of words missing from the translation dictionary is currently measured before word sense disambiguation takes place during the source language analysis. When more segments pass the analysis module, it would probably be better to do it after analysis.

Apart from the report, the evaluation module also produces log files of missing words and segments, which the system could not handle. Word level errors (except for code errors) are logged in word list files, which can be used for updating the lexicon. Logs on code errors are given in the same format as for segment level errors, as the context of a word with a code error is crucial for tracing the source of an error. Segment level errors (and code errors) are logged as SGML fragments, which can be fed back to the MATS system. In this way, it is possible to concentrate on "one kind of error" at a time, e.g. for diagnostic or progression evaluation. It is also possible to log segments that were translated, with or without reported errors, in the same format as above. The output from translating such logs could be used, for example, for evaluation of translation quality.

5 Conclusions

In the MATS system the translation process is made fully transparent in all aspects relevant to the user. This is due to the modularity of the system, and the logging of the individual steps. The glass-box character of the system and the fine-grained error report makes it a highly useful tool for development and teaching purposes.

References

Beskow, Björn. 1993. Unification-based transfer in machine translation. *Reports from the Department of Linguistics, RUUL #24*, Uppsala University.

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

Sågvall Hein, Anna. 1994. Preferences and linguistic choices in the Multra machine translation system. In Robert Eklund, editor, *Proc. of '9:e Nordiska Datalingvistikdagarna' NODALIDA'93*, pp. 267-276, Department of Linguistics, Stockholm University. Stockholm, June 3-5.

Sågvall Hein, Anna. 1997. Language Control and Machine Translation. In*: Proc. of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*. July 23-25. St. John's College, Santa Fe, New Mexico.

Sågvall Hein, Anna, Eva Forsbom, Jörg Tiedemann, Per Weijnitz, Ingrid Almqvist, Leif-Jöran Olsson & Sten Thaning. 2002. Scaling Up an MT Prototype for Industrial Use - Databases and Data Flow. In *Proc. of the 3rd International Conference on Linguistic Resources and Evaluation* (LREC'02) Vol V, pp. 1759-1766.

Tiedemann, Jörg. 2002. MatsLex - A Multilingual Lexical Database. *In Proc. of the 3rd International Conference on Linguistic Resources and Evaluation* (LREC'02) Vol VI, pp. 1909-1912.

Weijnitz, Per. 2002. *Uppsala Chart Parser Light - Improving Efficiency in a Chart Parser*. Master's Thesis. Department of Linguistics, Uppsala University.

Appendix: Evaluation Report

Input Overview

Words	Total	Unique	
Words	44107	15.58%	6874

Segments	Total	Unique	
Segments	7414	63.57%	4713

System Recall

Words	Total		Unique	
Source Language Words	99.15%	43730	97.70%	6716

Segments	Total		Unique	
Fully Translated	39.26%	2911	24.89%	1173
Translated	42.97%	3186	29.13%	1373

Error Reports

Words	Total	Unique
Source Language Words	377	158
Translation Links	6622	750
Target Language Words	180	13
Target Language Code	70	8

Segments	Total	Unique
Not Parsed	161	141
Partially Parsed	3658	3055
Not Transferred	250	14
Not Generated	159	130