Transfer of regular expressions in example-based machine translation

Jacek Gintrowicz and Krzysztof Jassem (Adam Mickiewicz University)

The continuous growth of available text corpora in recent years has moved the focus of research in Machine Translation. Rule-Based Machine Translation (RBMT) [Arnold, 1994], the most successful approach in XX century, is gradually driven out by methods based on large linguistic corpora: Statistical Machine Translation (SMT) [Brown, 1990, Ney, 2005] and Example-Based Machine Translation (EBMT) [Nagao, 1984, Hutchins, 2005].

Experiments have shown that using corpora-based translation in its pure form rarely yields satisfying results. The limitations are of the dual nature: linguistic (people manage to find combinations of words and phrases that have never been used before) and computational (the larger the database, the longer the search for relevant information). Therefore, the actual trend in MT is to combine RBMT with EBMT or SMT [Somers, 1999]. Examples of such a combination (RBMT and EBMT) are described in [Carl, 1999, Thurmair, 2004].

Here, we describe the idea of combining the RBMT and EBMT approaches by applying transfer rules for regular expressions in EBMT.

The idea of using regular expressions for natural language processing is widely known. Regular expressions (regexps) are most frequently applied to searching for regularly structured fragments, like numbers or dates. Kartunnen (1996) suggests applying finite automata and transducers that represent regular expressions, for natural language texts. Oflazer (2004) shows the use of regexps for tokenization, shallow parsing or morphology. Hasan (2005) describes how regexps may be applied in SMT for sentence clustering.

Here, we present the formalism, developed for describing transfer rules in EBMT. Each transfer rule consists of:

A) Regular expressions defining the search patterns in: Instance, Source Example and Target Example

B) Transfer of the instance expression into an output expression.

Below we show an example of a transfer rule for translation from Polish into English. The rule converts a date format.

Suppose the instance sentence is:

Instance: Egzamin jest zaplanowany na 18.02.2008.

The database contains a following example and its translation.

Source Example: Egzamin jest zaplanowany na 15.01.2007.

Target Example: The exam is planned on 15/01/2007.

We expect the algorithm to generate the following Output Sentence based on Target Example:

Output Sentence: The exam is planned on18/02/2008.

The appropriate transfer rule looks like this (only selected tags are listed for clarity):

```
1.  <instance>([0-9]{1,2})[\.]([0-9]{1,2})[\.]([0-9]{2,4})</instance>
2.  <source>(?:([0-9]{1,2})[\.]([0-9]{1,2})[\.]([0-9]{2,4}))</source>
3.  <target>(?:([0-9]{1,2})[\/]([0-9]{1,2})[\/]([0-9]{2,4}))</target>
4.  <orders>
5.      <order sourceGroup="1" suffix="/" />
```

6. <order sourceGroup="2" suffix="/" />
7. <order sourceGroup="3" suffix=" “/>
8. </orders>

Line 1 defines the search pattern for an input sentence (here, the pattern matches a Polish date notation). The Instance must contain a matching string to become applicable to the rule.

Line 2 defines the search pattern for the source part of an example. (here, the pattern is identical to that for Instance). The source part of an example must contain a matching string to be considered as the source for translation

Line 3 defines the search pattern for the target part of an example. The target part of an example must contain a matching string to be considered as the target for translation

Lines 4. to 8. define the transfer of the regular expression:

Line 5 refers to the first group of the instance surrounded by braces, i.e. *[0-9]{1,2}*. It says that in the output sentence the matching string should be added a suffix “/”.

Lines 6 and 7 refer to the second and third group of the instance respectively.

The <orders> tag makes it possible to permutate the order of the groups. The groups are listed inside the tag in the order expected in the target sentence (here, the order of the target is identical to that of the source).

So far, transfer rules have been developed for:

– various formats of date
– various formats of time
– currency expressions
– metric expressions
– numbers
– Internet addresses.

The idea has been applied to a commercial system, Translatica Server. The system combines ideas of RBMT, EBMT and Translation Memory [Hodasz, 2004]. Translation Memory plays a double role in the system: Firstly, it is used as a corpus on which example-based translation is performed. Secondly, it serves as a set of reference translations to aid human translation.

Bibliography

Arnold D., Balkan L., Lee Humphreys R., Meijer S., Sadler L., *Machine translation: an introductory guide*., Manchester/Oxford: NCC Blackwell.viii, 1994; 240pp.

Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S., A statistical approach to machine translation. *Computational Linguistics* 16 (2), 1990; pp. 79-85.

Carl, M., Inducing Translation Templates for Example-Based Machine Translation., *Proceedings of the Seventh Machine Translation Summit* (MT-Summit VII). 1999; 250-258.

Hasan S., Ney H., Clustered language models based on regular expressions for SMT, *10th EAMT conference Practical applications of machine translation, 30-31 May 2005, Budapest*; pp. 119-125.

Hodász G., Gröbler T., Kis B., Translation memory as a robust example-based translation system, *9th EAMT Workshop, Broadening horizons of machine translation and its applications, 26-27 April 2004, Malta*; pp.82-89.

Hutchins J., Towards a definition of example-based machine translation, MT Summit X, Phuket, Thailand, September 16, 2005, *Proceedings of Second Workshop on Example-Based Machine Translation*; pp.63-70.

Karttunen L., Chanod J-P., Grefenstette G., Schiller A., *Regular Expressions for Language Engineering, Natural Language Engineering (1996)*, 2: 305-328 Cambridge University Press.

Nagao M., A framework of a mechanical translation between Japanese and English by analogy principle, *Artificial and human intelligence: edited review papers presented at the international NATO Symposium, October 1981, Lyons, France*; ed. A. Elithorn and R. Banerji. Amsterdam: North Holland, 1984; pp. 173-180.

Ney H., One decade of statistical machine translation: 1996-2005, MT Summit X, Phuket, Thailand, September 13-15, 2005, *Conference Proceedings: the tenth Machine Translation Summit*: invited paper; pp.i-12-17.

Oflazer K., Yilmaz Y, Vi-xfst: A Visual Regular Expression Development Environment for Xerox Finite State Tool, In *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, Pages 86--93, Association for Computational Linguistics, Barcelona, Spain, July 2004.

Somers H., Review article: Example based Machine Translation, *Machine Translation*, 14(2), 1999; 113-157.

Thurmair G., Comparing rule-based and statistical MT output, *LREC-2004 Workshop, 25th May 2004: The amazing utility of parallel and comparable corpora*; pp. 5-9.