Evaluation Methods of a Linguistically Enriched Translation Memory System

Gábor Hodász

Pázmány Péter Catholic University
Faculty of Information Technology
Práter utca 50/a.
Budapest 1083, Hungary
hodasz@itk.ppke.hu

Abstract

The paper gives an overview of the evaluation methods of memory-based translation systems: Translation Memories (TM) and Example Based Machine Translation (EBMT) systems. After a short comparison with the well-discussed methods of evaluation of Machine Translation (MT) Systems we give a brief overview of current methodology on memory-based applications. We propose a new aspect, which takes the content of memory into account: a measure to describe the correspondence between the memory and the current segment to translate. We also offer a brief survey of a linguistically enriched translation memory on which these new methods will be tested.

1. Introduction

The literature of MT Systems discusses the theme of evaluation exhaustively and plenty of methods and measures arose in the past decades. Whilst in the case of translation memories and (in many respects very similar) EBMT systems the methods are less comprehensive. The evaluation of a machine translation system aims to measure a distance (or similarity) between the output of the system and a human translation used as a gold standard. On the other hand the evaluation of the memory-based systems examines the effectiveness of the reuse of the segments in the memory. Beyond that the aim of such systems is to help the work of a (potentially professional) translator, therefore the "usefulness" of the output can be defined as well. The method of evaluation also can rely on the co-operation of the user, i.e. the human scoring of the suggested translation sentence-by-sentence, because this is the way of usage of a TM [Somers2003].

2. The MetaMorpho TM system

The MetaMorpho TM system is a linguistically enriched translation memory, based on sub-sentential segments and a similarity measure rested on morpho-syntactic similarity [Hodasz-Pohl2005]. Our aim was to develop an improved TM system that uses linguistic analysis in both source and destination language sides to yield more exact matches to the source sentence. The MetaMorpho TM stores and retrieves sub-sentential segments and uses a linguistically based measure to determine similarity between two source-language segments, and attempts to assemble a sensible translation using translations of source-language chunks if the entire source segment was not found.

To describe the basic operation of the proposed TM engine is not the topic of this paper. To get a detailed description, see [Hodász-Grőbler-Kis2004]. The atomic actions are:

(1) the attempt to translate a single source segment, and

(2) adding a new translation unit (a pair of a source and target segment) to the translation memory once the human translator confirmed it. (See Figure 1.)

Note that some gaps may remain in the composite translation: the operation can still finish with success. Experience with fully automatic translation shows that a human translation even with gaps could be more useful than a target segment translated in a fully automatic manner.

2.1. NP alignment and sentence skeletons

The MetaMorpho TM uses a shallow noun phrase parser (NP-parser) on both source and destination sides to cut the sentences into NPs and sentence skeletons (sub-sentential segments). An NP-alignment module synchronizes the NPs of the source and target language sentences. These pairs are stored in the memory, and the rest of the sentences remain the skeleton. Therefore it is possible that the skeleton contains NPs in the case if the alignment module couldn't couple them. This can happen if a source language NP was translated to a non-NP (e.g. VP) in target language, or the alignment score is too low for a decision. This is the case in (Example 1.): the phrase *"in a variety of ways"* is an NP, in the target language sentence the corresponding phrase "sokféle módon" is an NP as well, but the score of the link between them is not enough to align them.

The translation is assembled from stored translations of noun phrases and the morpho-syntactic skeleton of the source unit. A morphological transformation is applied according to the input. The morpho-syntactic skeleton is a sequence of lemmas and morpho-syntactic parses of the words in the source unit, with a symbolic NP slot at the place of each noun phrase. In order to store and retrieve skeletons and NPs, the system uses an automatic NP alignment method [Hodasz-Pohl2005].

See the example of tiled-translation from English to Hungarian (Example 1.).

According to the above we have 2 sub-systems to evaluate: the NP-alignment module (including the shallow NP-parser) and the similarity search module. The 3rd is the evaluation of the whole system.

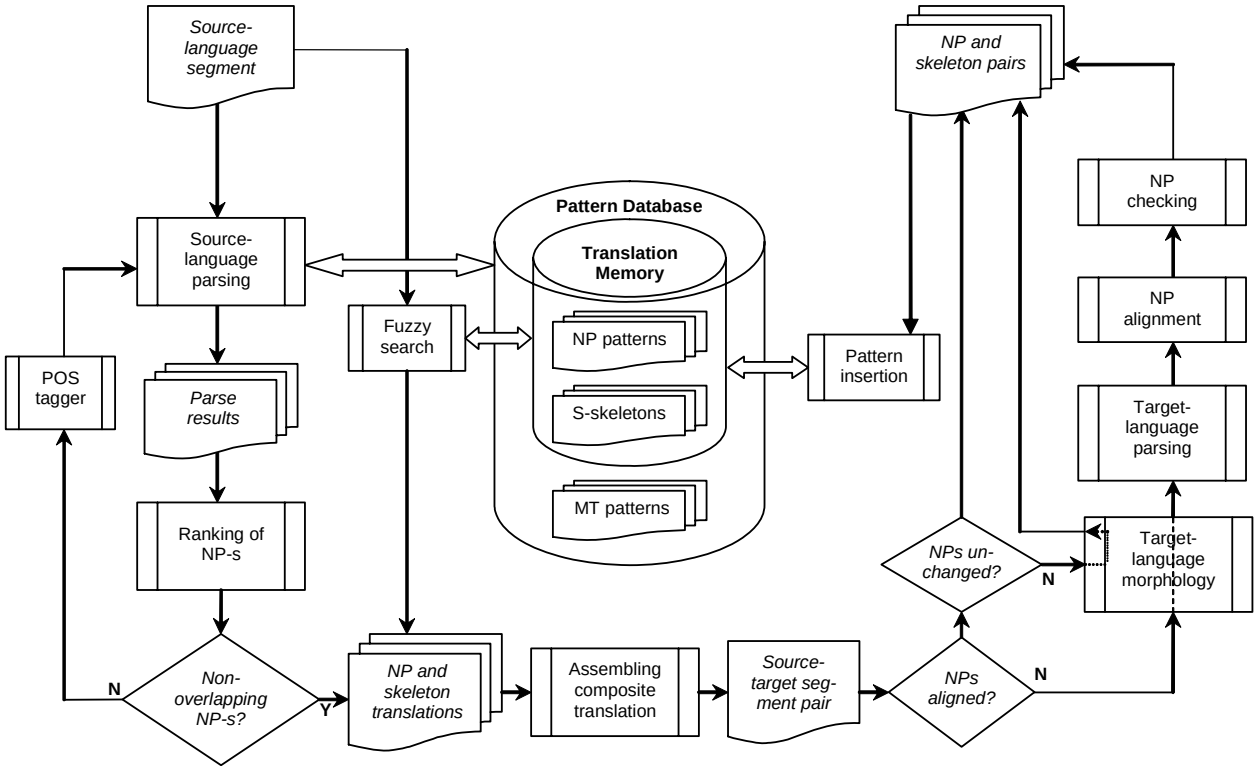

Figure 1. The basic processes of MetaMorpho TM

Sentence to translate:

Microsoft Windows 2000 makes it possible to configure hard disk drives in a variety of ways.

Sentence skeleton in memory:

[01] make$_{PERS3}$ possible to configure [02] in a variety of ways.
[01]$_{NOM}$ sokféle módon lehetővé teszi a beállítását [02]$_{DAT}$.

NP pairs found in memory:

Num	*Memory – ENG*	*Memory – HUN*
[01]	Microsoft Windows 2000	Microsoft Windows 2000
[02]	hard disk drive	merevlemez

Tiled translation:

[Microsoft Windows 2000] sokféle módon lehetővé teszi a beállítását [merevlemez][ek][nek].

Example 1. Tiled translation of sentence skeleton and NPs

3. Evaluation of the modules

Evaluation of memory-based systems ought to examine the effectiveness of reusing segments in the memory and require a measure to the "usefulness" of the output. According to these aspects we claim that it is necessary to distinguish the evaluation methods of MT Systems and Memory-Based Systems.

We give an overview of the most important methods (advantages and drawbacks), and present our measure to describe the correspondence between the memory and the current sentence to translate. We discuss the evaluation of the subsystems separately, presenting both automatic and manual methods.

3.1. Classification of evaluation methods

Evaluation strategies are divided into two major categories:

1. "black box": the system's operation is examined purely as it's input-output behavior without the consideration of the internal operation. This type of evaluation is suited to users (translators).
2. "glass box": the sub-modules are examined as well, their effect on the overall system is assessed. This method is relevant to developers and researchers.

Other classification divides the methods according to the automation of evaluation:

1. automatic (objective): large corpus can be evaluated in short time, however the "objectiveness" is depending on the defined measure and the selected corpus.

2. manual (subjective): more time consuming method and the results depends more on the persons who evaluate, but this method is closer to the real-life application of the system.

The evaluation method probably contains other important or useful features

3.2. NP-aligner module

The evaluation of the NP-parser and the NP-aligner module is less interesting inquiry, because a classical precision/recall method is convenient. One way is to use a reference corpus and evaluate automatically, the other way is to evaluate the output "by hand".

For the evaluation of the other modules, we use a corpus annotated and aligned by hand, not to accumulate the lapses of the NP-aligner.

3.3. Similarity search module

The similarity search module, which is obviously the most important one to increase the reusability of the segments in the database, can be evaluated in several ways.

Our method has two important features, which distinguish it from others.

1. we do not care about the number of the found similar segments, just the first one counts, herewith evaluating the sorting algorithm as well and preferring the fewer but more relevant matches (this is more helpful for the user).
2. we exclude the whole sentence matches (full correspondences), because they are trivial solutions and do not give picture about the effectiveness of the reuse of the patterns.

3.4. Manual (subjective) methods

The manual methods usually build upon a few human translators who evaluate the result by hand. There are two main ways: one is that the translator scores the result usually on a 1-4 scale from "absolutely useless" to "no changes needed". The other is that the translator modifies the result to get an acceptable translation and the system counts the post-edit steps needed. All these methods have the advantage that they model the real-life usage of the system: either measures the "contentment" of the user, or the "usefulness" of the suggested translation.

3.4.1. Automatic (objective) methods

The automatic methods are based on a bilingual parallel corpus in which all sentences have been translated by humans and are used as a gold standard. The result of the TM System will be compared to this standard. This evaluation is not depends on the skills and opinions of a single (or a few) human translator. The speed of evaluation is higher; a bigger corpus can be evaluated in unit time. The drawback is the need of a reference corpus: it can be subjective and several possible translations would be acceptable for the human translator.

One automatic way of evaluation is to count an edit-distance that is similar to the manual method above; the other is to use some kind of a "similarity score", such as BLEU/NIST score to evaluate the result sentence.

Our automatic method itself is based on the widely used BLEU score with the 10-fold cross-validation of the corpus [Papineni2002].

Our manual method is based on the evaluation of each output by a user, simply counting the post-editing steps needed. Counting keystrokes is a useful measure because it relates to the kind of task that is relevant for usefulness of the system to a translator. Of course this evaluation could be subject to criticism regarding subjectivity and small numbers of judges.

The comparison of the results of automatic and manual methods can be regarded as an "evaluation of evaluations".

4. Importance of Memory Content

The main difference between the evaluation of machine translation and memory-based systems is that in the latter case the content of memory is an important condition. Therefore our method takes into account not only the amount, but also the "cohesion" of the corpus in the memory and a small test corpus to translate. The more is this "cohesion" the more is the penalty on the result of the system (naturally it is easier to translate a sentence of which words are in the memory many times). We suggest a measure to characterize the "cohesion" of the corpus and calculated on the basis of the repetition of n-grams in the text.

4.1. n-gram based coherence measure

As discussed above, the aim of this measure is to characterize the similarity between the content of the memory and the current segments to translate. We take the 1-grams, 2-grams, …, n-grams of the test corpus, examine their number of occurrences in the memory and divide this number with the total number of n-gram of the corpus. In this manner we get a relative frequency of each n-grams. We summarize these frequencies with weights and a result is a weighted average, which presents the similarity between the corpus in the memory and the test corpus:

$$coherence(T_{TC}, T_{DB}) = \sum_n w_n \cdot \frac{\sum_{g_n \in TC} count_{DB}(g_n)}{|g_n|_{DB}}$$

where $count_{DB}(g_n)$ is the frequency of a given word n-gram in the memory, $g_n \in T_{TC}$ are the word n-grams of the test corpus, $|g_n|_{DB}$ the total number of n-grams in the memory, w_n is the weigh of a given n, that $\Sigma w_n = 1$.

Our proposal to the values of the parameters:

$$N = \max(n) = 6$$

$$w_n = \frac{n}{\sum_{i=1}^{N} i}$$

Thus we consider similarity up to a maximum of word 6-grams, and the longer the similarity, the bigger the weight.

Some consideration to this measure:

1. the longer n-grams contain shorter ones, therefore we count them more than once. A solution is if we start counting the longer ones, and later do not count the included or overlapped shorter ones.
2. both the memory and the test corpus contains several meaningless words (stop-words), which ones are of no importance from the point of corpus coherence. But these stop-words have more or less equal distribution and have a characteristic frequency of the given language. Therefore it is reasonable not to deal with them.
3. in agglutinative languages, like Finnish or Hungarian, the character-based similarity is not a proper way to compare words, because an inflected form of the same word is count in coherence, but will be found as different words. One solution is to operate with bigger corpora, therefore the probability of identity will be bigger. Another possibility is to allow some character differences at the end of each word. This is rough approximation, but can give a suitable solution. Because in MetaMorpho TM the source language is English we don't have to deal with this problem.

We need to test and possibly improve our measure on further investigation.

5. Conclusion and future work

In this paper we presented the evaluation process of a linguistically enriched translation memory. We claim that evaluation of a TM or an EBMT system is different from the well-elaborated evaluation of machine translation systems: the evaluation of the memory-based systems examines the effectiveness of the reuse of the segments in the memory. We presented the most important methods from the literature and added a completely new approach: the measurement of the similarity between the corpus in memory and the test corpus to translate.

In the near future we evaluate our system and test the above-defined coherence measure.

6. References

Hodász G., Grőbler T., Kis B. (2004), Translation memory as a robust example-based translation system. In: *Proceedings of the Ninth EAMT workshop*, University of Malta, Valletta, pp. 82-89.

Hodász G., Pohl G. (2005), MetaMorpho TM: a linguistically enriched translation memory. In: *International Workshop, Modern Approaches in Translation Technologies (ed. Walter Hahn, John Hutchins, Cristina Vertan)* ISBN 954-90906-9-8, Borovets, Bulgaria

Papineni, Kishore & Roukos, Salim et al. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics*

Somers, H. (2003). An Overview of EBMT In *M. Carl. and A. Way. (eds.) Recent Advances in Example-based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.3—57.