Construction of a Hierarchical Translation Memory

S. Vogel, H. Ney
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology
D-52056 Aachen, Germany
Email: vogel@informatik.rwth-aachen.de

Abstract

Translation memories are promising devices for automatic translation. Their main weakness, however, is poor coverage on unseen text. In this paper, the use of a hierarchical translation memory, consisting of a cascade of finite state transducers, is proposed. A number of transducers is applied to convert sentence pairs from a bilingual corpus into translation patterns, which are then used as a translation memory. Preliminary results on the German–English VERBMOBIL corpus are given.

1 Introduction

In recent years, example-based translation has been proposed as an efficient method for automatic translation (Sato and Nagao, 1990; Kitano, 1993; Brown, 1996). Translations are stored in a translation memory and used to construct translations for new sentences. In its simplest version, example-based translation boils down to using a database of source sentences with their translations. For many translation tasks, especially in computer assisted translation, this approach works with great success. For fully automatic translation the main problem is poor coverage on new data. To overcome this weakness, a hierarchical translation memory is proposed. Applying a cascade of finite state transducers, a source sentence is translated into the target language.

2 The Transducers

2.1 Overview

A translation memory is simply a collection of source–target string pairs. As a first step, these translation examples can be converted into translation patterns by introducing category labels, e.g. for proper names or numbers. To make the translation patterns even more useful, not only single words but complex phrases can be replaced by category labels. Which phrases to select for categorization depends on the application. For example, the corpus used for this study contains many time and date expressions. Therefore, a specialized transducer was constructed to recognize and translate such expressions.

Each transducer is a set of quadruples of the form:

label # source pattern # target pattern # score

Source patterns and target patterns may contain category labels. We call such patterns 'compound'. Transducers working only on the word level are called 'simple'. If a transducer contains recursive patterns, e.g. DATE # DATE und DATE # DATE and DATE # -3.0, it has to be applied recursively to the input.

The scores attached to the translation patterns can be viewed as translation scores. They are used to bias towards the selection of longer patterns and towards more likely translations in those cases where several target patterns are associated with one source pattern.

The transducers can be applied in both directions, i.e. for a given language pair, each language can be viewed as source language. Thereby, bilingual labeling is possible. This can be applied to convert a bilingual corpus into a selection of translation patterns which are formulated in terms of words and category labels.

2.2 Construction of the Transducers

The transducers should be selected in such a way as to minimize the need for recursive application in order to improve efficiency. Therefore, the patterns to search for are partitioned to form a cascade of transducers. Some transducers analyse parts of the sentence and replace it by a category label, which is then used at a later step by another transducer. The labeling of the days of the week or the names of the months is a prerequisite to apply more complex patterns for date expressions. The transducers currently used are listed in Table 1.

Table 1: List of transducers.

1. names (persons, towns, places, events, etc)
2. spelling (e.g. 'D A double L')
3. numbers (ordinal, cardinal, fractions, etc)
4. time and date expressions
5. parts of speech (for certain word classes)
6. grammar (noun phrases, verb phrases)

Some transducers are general in scope, e.g. the transducers for numbers, part of speech tags and grammar. Others are costumized towards the domain for which the translation system is developed. In the VERBMOBIL corpus, which is used for the experiments, time and date expressions are very prominent. To recognize these expressions, a small grammar has been developed and coded as finite state transducer. Actually, two transducers are used. On the first level, words are replaced by labels, like DAYOFWEEK = { Montag, Dienstag, ...}. On the second level, these labels are used to form complex time and date expressions. This second transducer works recursively, as simpler expressions are used to build more complex expressions.

Finally, a small grammar based on POS (part of speech) tags has been crafted manually. The purpose of this grammar is to recognize simple noun phrases. Extensions to handle the different word ordering in the verb phrases are under development.

2.3 Scoring

The scores attached to the translation patterns can be viewed as a kind of translation scores. In the current implementation a rather crude heuristic together with some manual tuning in the grammar transducer is applied. The idea is to give preference to longer translation patterns as they take more context into account and encode word reordering in an explicit manner. Thus, for simple and compound translation patterns the score is exponential to the length of the source pattern. The scores are negative by convention: not translating a word gives zero cost, translating it gives a benefit, i.e. negative costs. In future, scoring will be refined by using corpus statistics to assign probabilities to the translation patterns.

2.4 Bilingual Labeling

The sentence pairs in the bilingual training corpus can be segmented into shorter segments with the help of an alignment program (Och et al., 1999). This collection of segments could be used directly as a translation memory. However, to improve the coverage on unseen data, these segments are labeled. Applying the transducers as given in Table 1 transforms these segments into compound phrases.

The procedure is as follows:

1. For each transducer taken from the complete cascade – as given in Table 1 – apply the transducer to both, the source and the target sentences of the bilingual training corpus.

2. Find those sentence pairs which contain equal number and types of category labels for both sentences.

3. For sentence pairs which do not match in number and type of the category labels keep the original sentence pair.

Table 2 shows examples of some translation patterns which resulted from bilingual labeling.

3 Applying the Transducers

The working of the transducers is best described as the construction of a translation graph. That is to say, the sentence to be translated is viewed as a graph which is traversed from left to right. For each matching source pattern, as encoded in the transducers, a new edge is added to the graph. The edge is labeled with the category label of the translation pattern. The translation and the translation score are attached to the edge. In this way a translation graph is constructed. In those cases, where a source pattern has several translations, one edge for each translation is added to the graph.

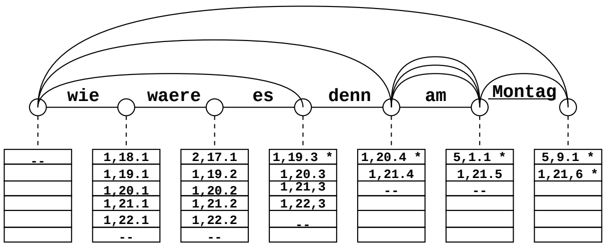
The left–right search on the graph is organized in such a way that all paths are traversed

Table 2: Compound translation patterns (CTP).

CTP # DATE_DAY ginge es wieder	# DATE_DAY it is possible again	# -4.6
CTP # SURNAME am Apparat	# this is SURNAME speaking	# -3.3
CTP # NP dauert DATE	# NP takes DATE	# -3.3
CTP # nehmen PPER NP DATE	# let PPER take NP DATE	# -4.6

in parallel and the patterns stored in the transducer are matched synchronously. For each node n and each edge e leading to n, all patterns in the transducer starting with the label of e are attached to n. This gives a number of hypotheses describing partially matching patterns. Already started hypotheses are expanded with the label of the edge running from the previous node to the current node. This procedure is shown in Figure 1. For a selection of translation patterns from the simple, word-based translation memory the hypotheses for partially matching patterns generated during the left–right traversal are shown as well as the resulting new edges.

The result of applying all transducers is a graph where each path is a (partial) translation of the source sentence. The path with the best overall score is used to construct the final translation. For good results, not only the scores from the transducers should be used in selecting the best path, but a language model of the target language should be included.



1 am # at, on, at the
...
9 Montag # Monday
...
17 waere es so moeglich # would that be possible
18 wie ist es bei Ihnen # how about you
19 wie waere es # how about
20 wie waere es denn # how about
21 wie waere es denn am Montag # how about Monday
22 wie waere es am Montag # how about Monday

Figure 1: Expansion of Pattern Hypotheses

3.1 Error Tolerant Match

To improve the coverage on unseen test data, it may be avantageous to allow for approximative matching. The idea is, to apply longer segments for syntactically better translations without loosing to much as far as the content of the sentences is concerned.

We use weighted edit distance, i.e. each error (insertion, deletion, substitution) is assiciated with an individual score. Thereby, the deletion or insertion of typical filler words can be allowed, whereas the deletion or insertion of content words is avoided.

3.2 Translation on Word Lattices

The approach described so far can be used for a tight integration of speech recognition and translation. Speech recognition systems typically produce word lattices which encode the most likely word sequences in an efficient manner. A direct translation on the lattice has, compared to transforming the lattice into an n-best list, translating each word sequence, and selecting the overall best translation, a number of advantages:

- all the paths can be covered, whereas in an n-best approach typically only a small fraction of the paths is considered;
- partial translation hypotheses are reused;
- acoustic scores can be taken into account when calculating an overall score for each translation hypothesis.

4 Experiments and Results

In this section, we will report on first experiments and results obtained with the cascaded transducer approach. Experiments were performed on the VERBMOBIL corpus. This corpus consists of spontaneously spoken dialogs in the appointment scheduling domain (Wahlster, 1993). The vocabulary comprises 7 335 German words and 4 382 English words. A test corpus of 147 sentences with a total of 1 968 words was used to test the coverage of the transducers and to run preliminary translation experiments.

In Table 3 the sizes of the transducers are given.

Table 3: Number of translation patterns of the transducers.

Transducer	Patterns
Name	442
Spell	60
Number	342
Date	334
POS Tags	6714
Grammar	124

4.1 Coverage

In a first series of experiments, the coverage of the cascaded transducers was tested. The sentences pairs from the training corpus were segmented into shorter segments. This resulted in 43609 bilingual phrases running from 1 word up to 82 words in length. The longest phrases were discarded as it is very unlikely that they will match other sentences. Thus, for the experiments only 40000 sentence pairs were used, the longest sentences containing sixteen source words.

Starting from those simple phrases, successively more transducers were applied up to the full cascade. In Table 4 the coverage for each level is shown. As expected, the coverage increases and nearly full coverage on the test sentences is reached. In the final step, the POS transducer and the grammer transducer are both applied.

The first column shows which transducers have been applied. In each step, one additional transducer is applied for bilingual labeling and for translation. Bilingual labeling reduces the number of distinct patterns in the translation memory, whereas the number of compound patterns increases. The last column shows the number of words in the test sentences not covered by the patterns in the translation memory. As can be seen, the coverage increases which each step. The large improvement in the final step results from applying the POS-tag transducer which coveres a large part of the vocabulary.

Table 4: Effect of selected transducers on coverage on test corpus.

Transducers	Patterns	Compound	not covered
None	40000		273
Name	39624	1259	254
+ Spell	39508	1468	249
+ Number	38669	11181	238
+ Date	36118	14684	215
+ Grammar	35519	15682	9

4.2 Translation

First experiments have been performed to test the approach for translation. So far, no language model for the target language is applied to score the different translations.

For the sentence 'Samstag und Februar sind gut, aber der siebzehnte wäre besser' the best path through the resulting translation graph gives a structure as shown in Figure 2. In Table 5, some translation examples for test sentences not seen in the training corpus are given.

Table 5: Three translations generated from the hierarchical translation memory.

Ich werde mit dem Flugzeug kommen.
I will come with the plane.
Ja, wunderbar. Machen wir das so, und dann treffen wir uns dann in Hamburg. Vielen Dank und auf Wiederhören.
Well, excellent. Shall we fix this, and then we will meet then in Hamburg. Thank you very much goodbye.
Das kann ich nicht einrichten. Ich habe eine Chance ab dreiundzwanzigsten Oktober. Ist es da bei Ihnen möglich?
It can I not arrange. I have a chance from twenty-third of October. Is it as for you possible?

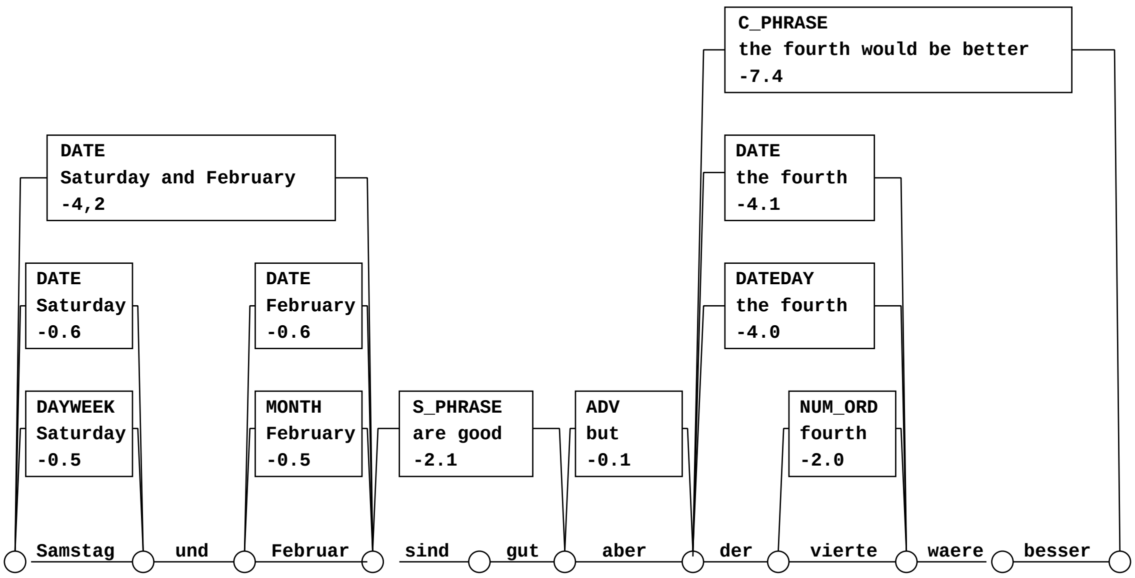


Figure 2: Translation example

5 Summary and conclusions

In this paper a translation approach based on cascaded finite state transducers has been presented. A small number of simple transducers is handcrafted and then used to convert a bilingual corpus into a translation memory consisting of source pattern – target pattern pairs, which include category labels. Translation is then performed by applying the complete cascade of transducers.

First experiments have shown the potential of this approach for machine translation. Good coverage on unseen test data could be obtained.

The main difficulty in this approach is to define a consistent scoring scheme for the different transducers. Especially, a good balance between the grammar and the word-based translation memory is necessary. This will be the main focus for future work.

As already mentioned, a language model for the target language has to be integrated into the scoring of the translation hypotheses. Finally, the transducer based approach to translation will be tested on word lattices as produced by speech recognition systems.

Acknowledgement. This work was partly supported by the German Federal Ministry of Education, Science, Research and Technology under the Contract Number 01 IV 701 T4 (Verbmobil).

References

R. D. Brown. 1996. Example-based machine translation in the pangloss system. *Proceedings of the 16th International Conference on Computational Linguistics*, 169–174, Copenhagen, Denmark, August.

H. Kitano. 1993. A comprehensive and practical model of memory-based machine translation. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, volume 2, 1276–1282. Morgan Kaufmann.

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 20–28, University of Maryland, College Park, MD, USA, June.

S. Sato and M. Nagao. 1990. Toward memory-based translation. *Proceedings of the 13th International Conference on Computational Linguistics*, vol. 3, 247–252, Helsinki, Finland.

W. Wahlster. 1993. Verbmobil: Translation of face-to-face dialogs. *Proceedings of the MT Summit IV*, 127–135, Kobe, Japan.