4L−13

Japanese-to-German spoken-language translation utilizing empirical linguistic knowledge

Michael Paul Osamu Furuse Hitoshi Iida

ATR Interpreting Telecommunications Research Laboratories
e-mail: {paul,furuse,iida}@itl.atr.co.jp

1 Introduction

Within the context of Transfer-Driven Machine-Translation (TDMT) we are pursuing a multilingual approach for spoken-language translation using empirical linguistic knowledge. One of the target languages under investigation is German. In this paper we will describe how the empirical linguistic information provided by the transfer component is used to handle the free word order and complex inflectional phenomena of German.

2 Transfer

In order to handle the incremental translation, TDMT uses a *constituent boundary parsing* method in an exampled-based framework, simultaneously parsing the input and applying transfer knowledge by means of pattern matching [2].

A *pattern* expresses a meaningful unit of linguistic structure and is defined as a sequence consisting of variables ("placeholders" for constituents) separated by symbols representing the constituent boundaries. Boundaries can either be *functional words*, e.g. the Japanese particles は, が, を, or *part-of-speech bigrams*, which are markers inserted to the input when no surface word divides an expression, e.g. <*noun-noun*> in the case of compound nouns.

The incremental pattern matching algorithm is based on the idea of *chart parsing* and carried out in a bottom-up and left-to-right fashion. It derives all possible segmentations of the input sentence, whereby the structural ambiguity is restricted using the best-only substructures according to a semantic distance calculation.
Simultaneous with the structural parsing TDMT applies its transfer knowledge to the instantiations of the segmented input parts. The transfer knowledge associates each pattern of the source language with a corresponding linguistic expression of the target language, based on the empirical knowledge obtained during the processing of the training data. The target expressions consist of the word translations of the instantiated pattern variables enriched with the linguistic constituent markers of the target language, e.g. applying the pattern "Xは $\Rightarrow [X]_{subject}$" to the substructure "食事は" and using the dictionary entry "食事 $\Rightarrow [Essen]_{noun}$" yields in the transfer result "$[[Essen]_{noun}]_{subject}$". Thus, by choosing the most appropriate substructures, the transfer component delivers the target expressions most consistent with TDMT's empirical knowledge to the generation module.

3 Topological Fields

One of the main characteristics of German is its rather free word order within a sentence. In TDMT we use a *topological field* approach, whereby a sentence model consists of several "fields", whose linear composition specifies the chosen word order within the sentence.

The structure of a German clause depends on the position of its finite verb, which can appear in the *first*, *second* or *final* position. The finite verb position plus the non-finite parts of the predicate define the topological structure as follows:

first:	—	*finite verb*	middle field	*non-finite verb*	post field
second:	pre field	*finite verb*	middle field	*non-finite verb*	post field
final:	pre field	—	middle field	*verb-complex*	post field

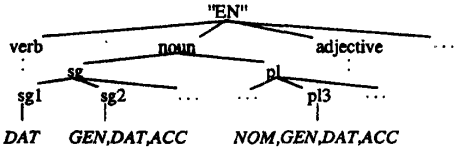
In contrast to English, the *pre field* is not exclusively reserved for the subject, and the topological fields of the German clause can be filled by almost any linguistic constituent. However, the *middle field* is in some sense the default field.

In TDMT this coarse sentence model is refined by introducing new topological fields for the linguistic constituents specified in the transfer result (subject, object, time and locative phrases, etc.). For each sentence type we define a *default template*, which represents the most general order of its fields. Thereby we have to obey some heuristics concerning both syntactic ordering principles, e.g. the *unmarked order* (subject, indirect object, object) or complement and adjunct order, and pragmatically based regularities, e.g. focus.

4 Generation

Given the transfer result each linguistic constituent will be analyzed both on the sentence-level (selection of a topological field) and on the word-level (inflection phenomena). On the word-level we use a classification-based approach, which considers morphological regularities of German as the basis for the definition of a fine-grained, word-class specific classification, i.e. words with the same morphological behavior are grouped together in classes. Additionally it uses morphosyntactic features (phonological properties) in refining the class hierarchy [1].

In addition to the *fullform-lexicon* for storing the non-inflecting word-classes (e.g. adverbs), the *stem-lexicon* contains information about the classification of each word-stem, and the *inflectional allomorph lexicon* (IAL) relates each inflectional morph to all its possible combinations of morphosyntactic information. Each entry in the IAL is a n–ary tree, of which the nodes describe the classes and the leaves contain the appropriate inflectional information. The figure below shows the IAL-tree of the suffix "EN". Given a noun (conjugation in *number* and *case*) ending in "en", e.g. "Essen" (*meal*), further subclassification due to the word-class of its stem, e.g. singular and plural classes, leads to a leaf containing the *case* attributes for the given input.



Thus the word-level analysis can be performed by means of simple operations on n–ary trees. An input word that is not included in the *fullform-lexicon* has to be decomposed into a candidate stem and the corresponding prefixes and suffixes. Taking into account substrings of the longest possible prefix/suffix, all valid *prefix–stem–suffix* combinations and the corresponding inflectional information of the given input are analyzed.

These segmentation processes can be reversed, in order to generate well-formed surface words in the generation output, i.e. given a word-stem and some inflectional attributes the corresponding prefix/suffix can be found in the IAL.

In TDMT the empirical linguistic knowledge of the transfer will be used to extract the appropriate constituents out of the (possibly) ambiguous word analysis, e.g. the German word "essen" is both a noun (*meal*) and a verb (*to eat*), but because of the linguistic marker $[]_{noun}$ of the input "$[[Essen]_{noun}]_{subject}$" the noun will be analyzed.

On the sentence-level the analyzed linguistic information of the target expressions will be propagated to the word analysis of their subparts and the result is added to the respective topological fields. In our example the marker $[]_{subject}$ forces the propagation of the nominative case to the noun "Essen", whose result will be updated to the *subject* field of our sentence model.

After the transfer result analysis the contents of each topological field are inflected and the concatenation of these surface strings results in the output of the generation module.

5 Application

An illustration of the generation is chosen from our *travel conversation* domain. Given the input "食事はついていますか" the transfer result is $[[[Essen]_{noun}]_{subject}[inbegriffen\ sein]_{vp}]_{yn-q}$. The finite part of the verb phrase is the auxiliary verb "sein" (*to be*) and the adjective "inbegriffen" (*included*) forms the non-finite part. Due to agreement in *number* and *person* the subject attributes, analyzed on the word-level, are propagated to the finite verb. Since yes/no-questions ($[]_{yn-q}$) are *verb-first* clauses, the *verb-first* template will be chosen, whereby the subject field is part of the *middle field* in the coarse model, shown above. Thus the subject is located between the finite and non-finite predicate parts. The inflection of each topological field results in the generation output:

"食事はついていますか"

⇓

	Ist	das Essen	inbegriffen	?
[pre]	[finite]	[middle]	[non-finite]	[post]

References

[1] W. Finkler and G. Neumann, MORPHIX: A Fast Realization of a Classification-Based Approach to Morphology, *4. Österreichische Artificial-Intelligence-Tagung*, pp. 11–19, Springer, Berlin, 1988.

[2] O. Furuse and H. Iida, Incremental Translation Utilizing Constituent Boundary Patterns, *Proc. of the 16th COLING*, Copenhagen, Denmark, 1996.