An Experiment on Synchronous TAGs for the Construction of a Transfer Module

Alexandre Agustini & Vera Lúcia Strube de Lima

PUCRS - Instituto de Informática
Av. Ipiranga, 6681 prédio 30 bloco 4
90619-900 Porto Alegre RS - BRASIL
{agustini,vera}@inf.pucrs.br

Abstract

This paper presents some considerations on the use of Synchronous TAGs for the design of a structural transfer module, which is the main component of transfer-based systems for Machine Translation. The transfer module establishes the correspondences between the structural representation of both the source and target languages. A study of a corpus from Economics was carried out in order to define structural divergences for the translation between the Portuguese and English languages.

1 Introduction

Machine translation (MT) has been a challenge for linguists and computer scientists over the last decades. During this period, plenty of progress was accomplished, though the results are not yet the ones expected.

Transfer based approaches to MT involve three main phases: analysis,. transfer and generation. During analysis, the syntactic and semantic structure of a sentence is made explicit through a source language (SL) grammar and semantic processing modules. The result of the analysis is one or more syntactic and semantic representations which are used to construct a syntactic and/or semantic representation in the target language (TL) through a series of transfer rules and according to a bilingual lexicon. From this representation a TL sentence is generated based on some form of mapping procedure [Hutchins & Sommers 92; Trujillo 95].

In this paper we describe a prototype implementation of a transfer MT module based on the Synchronous Tree-Adjoining Grammars (STAGs) formalism. STAGs [Shieber & Schabes 90] are a variant of Tree-Adjoining Grammars (TAG) to express the related representations of semantics and syntax in natural-language description.

2 Corpus based development

Our basic approach is corpus-driven. We started by collecting a source-language corpus (Portuguese sentences) in a limited domain. The corpus made up by 200 sentences was created randomly from an economics headlines database. About 50% of them were discarded because they were ill-formed senteces. The database had previously been generated from a news broadcasting system.

An English version of the corpus was produced by a native translator · with experience in the domain terminology. Finally, both corpuses were tagged and aligned in order to achieve:

- *virtual grammars*[1] for both the source

[1] In this context, *virtual grammar* refers to a syntactic structure subset necessary for parsing any input sentence and generating target structures occuring in the corpus.

and target corpuses: the subset of lexicalized trees necessary for syntactic/semantic analysis of source and target corpus was defined. These grammars are based on [Kipper 94] and [Becker et al. 94] technical reports.

- *lexicon coverage*: source and target lexical dictionaries were set. As we are working on a lexicalized model [Abeillé 90; Srinivas et al. 94], each lexical item anchors one or more syntactic structures.
- *translation discrepancies*: translation problems to be solved during transfer from source to target structures were addressed.

We found it helpful to divide translation problems into three different types: lexical, syntactic and lexical-semantic. These terms are used according to the following concept (according to [Dorr 94]): *Lexical problems* are concerned about finding correct choices for expressions that occur in the source and target languages. *Syntactic problems* feature syntactic properties associated with each language (i.e., properties that are independent of the actual lexical items that are used). Finally, *lexical-semantic problems* which feature properties that are lexically determined.

Some examples of divergences observed in the corpus are presented on Table 1. In case (1), problems originated from lexical gaps in the source and target languages are shown; the translation has to deal with structural problems and feature inheritance. Syntactic problems (2) usually have to do with word order, in the examples, adjectival phrase order. In the last one, a lexical-semantic problem, see (3), the Portuguese verb (*fazer*) and its complement (*leilão*) are translated into a verb (*to auction*) in English.

	Portuguese	English
(1)	*corretora* *parlamentares* *empreiteiras* *linhas aéreas*	firm of brokers members of parliament Contract construction companies airlines
(2)	*peso Mexicano* *bolsa de Nyork*	Mexican peso New York stock exchange
(3)	*fazer leilão*	to auction

Table 1: Example of some discrepancies

3 A Model Proposed and Implementation

Figure 1 illustrates the proposed model in which the structural divergence related information is modeled in two different dictionaries: a *structural dictionary* and a *bilingual dictionary.*

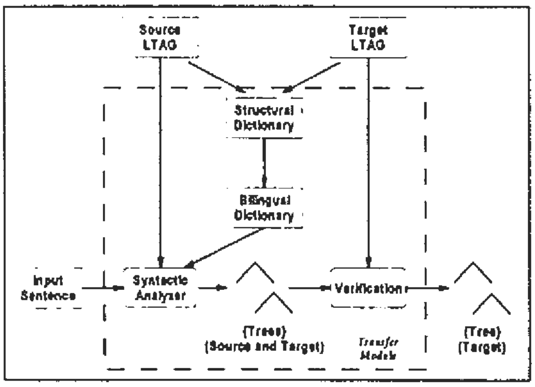

Figure 1: Overview of Proposed Model

The *structural dictionary* connects LTAGs structures that define both the source and the target languages. This structure maintains the node-to-node links between the source and the target grammars. All elementary trees in the source language are associated with trees in the target language. It has information about the inheritance of semantic attributes and also holds all the information for syntactic divergence resolution Table 2 illustrates some entries of the structural dictionary.

```
# format:
#  ( source_id : target_id ) → [ links ]

#  transitive verbs (NP object complement)
#  S( N V N )  -->  S( NP CV( V NP ) )
(  2:302) = [ $0:$0 , $1:$1 , $2:$3 , $3:$4 ] ;

#  adjectives
#  N( N Adj )  -->  NP( Adj NP )
(100:400) = [ $0:$0 , $1:$2 , $2:$1 ] ;
#  N( Adj  N )  -->  NP( Adj NP )
(101:400) = [ $0:$0 , $1:$1 , $2:$2 ] ;

#  adjectival phrase
#  +NPROP = Proper Noun
#  N( N Prep N )  -->  NP( Adj NP )
( 20:400) = [ $0:$0 , $1:$2 , $3[+NPROP]:$1 ] ;
```

Table 2: Selected Structural Dictionary Entries

The *bilingual dictionary* contains the rules for the resolution of lexical and lexical-semantic divergences. This dictionary manipulates the pairs of lexicalized items and points out one or more elementary structures of the structural dictionary to which the item is anchor. In this dictionary, derivation tree fragments can be defined, with the purpose of resolvihg lexical and lexical-semantic divergences. Furthermore, the dictionary can extend the rules contained in the structural dictionary to state the restrictions imposed by the accomplished lexical insertion. A fragment of the bilingual dictionary is presented in Table 3.

The transfer module receives a sequence of lexical items, generated by a lexicon-morphologic module. The output corresponds to one or more derivation trees in the target language with all structural modifications accomplished and *decorated* with the semantic features inherited from the source language.

Virtual grammars for source and target languages are described in an independent way and the notation introduced by [Kipper 94] for both grammars was used. The Portuguese grammar is a subset of the [Kipper 94] grammar and the English description was extracted from [Becker et al. 94].

```
#  format:
#  ( source_entry : translation )  =  [ anchor list ];
(hoje  : today)          = ;
(fazer : make)            = ;  [2];

%  (fazer : ##)   // redefinition of verb fazer
%  S(N V[#lex=fazer] N(N[#lex=leilao] Prep N))
%                  →
%    ( S ( NP  CV ( V[#lex=auction]  NP ) ) :
%     [ $0:$0, $1:$1, $2:$2, $6:$4];

# default
%(#lex : #lex)
```

Table 3: Selected Bilingual Dictionary Entries

Due to the incremental characteristic of the STAGs method, transfer functions were incorporated to syntactic analysis. The implementation involves two distinct steps: syntactic analysis (parser) and verification.

The *parser* uses a top-down algorithm for LTAG recognition. Each operation carried out by the parser in the SL enables one or more operations in the TL. The output is: for each SL syntactic structure a set of structures in the TL is generated.

During the process of analysis and translation, two types of attributes are manipulated: structural and semantic attributes. The structural attributes are inherent to each language and do not need to be transferred. On the other hand, semantic attributes are inherited by each one of the accomplished items of the pairs in lexicalized trees.

Finally, in the *verification step*, the unification of semantic features and the verification for structural consistency on the generated target trees is carried out. This process is based on the target LTAG grammar and inconsistent trees are discarded.

4 Conclusion and Future Work

This work investigated the use of the STAGs formalism for the treatment of lexical, syntactic and lexical-semantic divergences defined from a corpus in the field of Economics. Due to the extended domain of locality of LTAGs, it is possible to define regular correspondences among complex structures without the need of intermediary representations.

Although it was possible to set the translation rules for about 85% of the selected corpus (composed of 90 sentences), the model cannot yet be validated due to the short number of sample sentences.

Nowadays, we are starting to work on tagging and aligning tools for a bilingual corpus. These tools will allow us to set a more complex corpus of sentences to validate the work we have developed.

References

[Agustini 97] AGUSTINI, A. *Experiência de Utilização do Formalismo STAGs para a Construção de um Módulo de Transferência Estrutural.* Master's Degree in Informatics, PUCRS, RS, Brazil, Mater's Dissertation, 1997.

[Abeillé 90] ABEILLE, A. et. al. Using Lexicalized TAGs for Machine Translation. In *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, Finland, 1990.

[Beckeret et al. 94] BECKER, T. et al. A Lexicalized TAG for English. University of Pennsylvania, *Technical Report*, 1994.

[Dorr 94] DORR, B. J. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, v. 20, n. 4, 1994.

[Hutchins & Somers 92] HUTCHINS, W. J. & SOMERS, H. L. *An Introduction to Machine Translation.* Academic Press, Great Britain, 1992.

[Kipper 94] KIPPER, K. *Uma Experiência de Utilização do Formalismo de Gramáticas de Adjunção de Árvores para a Lingua Portuguesa.* Master's Degree in Informatics, CPGCC-UFRGS, Porto Alegre, RS, Brazil, Mater's Dissertation, 1994.

[Rambow & Satta 96] RAMBOW, O. & SATTA, G. Synchronous Models of Language. In *Proceedings of the 19th International Conference on Computational Linguistics*, California, USA, 1996.

[Shieber & Shabes 90] SHIEBER, S. M. & SHABES, Y. Synchronous Tree-Adjoining Grammars. In *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, Finland, 1990.

[Trujillo 95] TRUJILLO, A. Bi-Lexical Rules for Multi-Lexeme Translation in Lexicalist MT. *MI-95*, Leuven, Belgium, 1995.