A Distributed Database for Mobile NLP Applications[*]

Petr Homola
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
CZ-118 00, Prague, Czech Republic
homola@ufal.mff.cuni.cz

Abstract

The paper presents an experimental machine translation system for mobile devices and its main component — a distributed database which is used in the module of lexical transfer. The database contains data shared among multiple devices and provides their automatic synchronization.

1 Introduction

In Europe, machine translation (MT) is very important due to the amount of languages spoken there. In the European Union, for example, there are more then 20 official languages. Some of them have very few native speakers and it is quite problematic for institutions and companies to find enough translators for comparatively rare language pairs, such as Danish-Maltese. We have developed an experimental MT system for Central and East European languages which is in detail presented in (Homola and Kuboň, 2004); at the moment, we have resources for German, Polish, Czech, Slovak and Russian. As the languages are syntactically and, except of German, lexically related, the system is rule-based. All components of the system are implemented in Objective-C (ObjC) and have been ported to the *iPhone*.

2 Architecture of the MT System

The basic version of the system consists of the following modules:

Morphological analyzer. Since the languages have rich inflection, a word has usually many different endings that express case, number, person etc. It is necessary to assign a lemma and a set of morphological tags to each word form.

Shallow parser. The parser analyzes constituents of the source sentence, but not necessarily whole sentences.

Lexical and structural transfer. The lexical transfer provides a lemma-to-lemma or a term-to-term translation. The structural transfer adapts the syntax of the phrases so that they are grammatical in the target language.

Morphological synthesis of the target language. This final phase generates proper word forms in the target language.

The shallow parser uses the dynamic algorithm described in (Colmerauer, 1969) with feature structures being the main data structure. The hand-written rules are fully declarative and defined in the LFG format (Bresnan, 2001), i.e., they consist of a context-free rule and a set of unificational conditions. The transfer (lexical and structural) is followed by the syntactic and morphological synthesis, i.e., the syntactic structures which represent the source sentences are linearized and proper morphological forms of all words are generated, according to the tag associated with them.

3 Lexical Transfer

The dictionaries are sub-components of the transfer module. Their task is to provide lexical translation of constituents analyzed by the shallow parser. The dictionary contains translation pairs for words and

[*]The research presented in this paper has been supported by the grant No. 1ET100300517 of the GAAV ČR.

Proceedings of the ACL-08: HLT Workshop on Mobile Language Processing, pages 27–28,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

phrases. Most items contain an additional morphological or syntactic information such as gender, valence frames etc.

The creation of the dictionaries is a very time-consuming task and they can never cover the complete lexicon of a language. In a production environment, it is inevitable to add new items to the database as new texts are processed. The typical workflow is as follows:

1. During the translation of a document (possibly on a mobile device), unknown words or phrases are found. In the translation, they appear in the source form since the system does not know how to process them. After the processing of the whole document, all found unknown words are added to the database with a remark that the words are new to the system.
2. The new items are transmitted to the computer of a translator whose task is to translate them. Moreover, most items will be assigned a morphological or syntactico-semantical annotation for the structural transfer.
3. The manually updated items are distributed to all instances of application, i.e., to all devices the MT system is installed on, so that they are available for future use by all users of the system.

The capacity of the used mobile device is sufficient to store the lexicon persistently but one could run into problems trying to keep the whole lexicon in memory. For this reason, we use a ternary tree as an index which is kept in memory while full items of the lexicon are loaded from a persistent repository at the moment they are needed.

4 Distributed Database

The database can be used on multiple devices and it is synchronized automatically, i.e., an update of an object is transmitted to all other instances of the database. The synchronization can be deferred if the modifier or the receiver of the update are offline. In such a case, the database is synchronized as soon as the device with the database has access to the internet. Due to the offline synchronization, synchronization conflicts can arise if two or more users update an object simultaneously. If the users have changed different properties of the same object, the changes are merged automatically. Otherwise, the administrator of the database has to resolve the conflict manually.

The distributed database consists of the following components:

Object repository. A local repository of ObjC objects so that the database is accessible even if there is no internet connection.

Transceiver. A communication module that sends/receives updates to/from the relay server. It includes a local persistent cache for updates which is used if there is no internet connection.

Relay server. A server that accepts updates and distributes them to other instances of the database. This component ensures that the database is synchronized even if two or more users are never online at the same time.

It is noteworthy that there is no replica of the database on the server, it only serves as a temporary repository for updated records that cannot be synchronized immediately because a receiving device may be offline at the moment another device has committed an update (this is the expected situation for mobile devices such as PDAs and smartphones).

Currently, the distributed database is being used as a collaboration platform in the Czech Broadcasting Company (Český rozhlas).

5 Conclusions

We have presented an experimental MT system that works on the *iPhone* and described how it uses a distributed object database with automatic synchronization to keep the lexicon of the system up-to-date on all devices it is installed on. We believe that the presented database is an effective way to keep frequently updated data up-to-date on multiple computers and/or mobile devices. The system is developed in Objective-C thus the code base can be used on the *iPhone* and on Macs, and it can be easily ported to systems for which the GNU C Compiler is available.

References

Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell Publishers, Oxford.

Alain Colmerauer. 1969. Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur. Technical report, Mimeo, Montréal.

Petr Homola and Vladislav Kuboň. 2004. A translation model for languages of acceding countries. In *Proceedings of the EAMT Workshop*, Malta.