eRoL

Automatic Voice Translator for Romanian:

Building Resources for a Symbolic Machine Translation Program

MIHAELA COLHON

EDITURA UNIVERSITARIA
Craiova, 2013

Referenţi ştiinţifici:
Prof.univ.dr. DAN CRISTEA
Prof.univ.dr. COSTIN BĂDICĂ

Copyright © 2013 Universitaria
Toate drepturile sunt rezervate Editurii Universitaria

Descrierea CIP a Bibliotecii Naţionale a României

COLHON, MIHAELA VERONA
eRoL : automatic voice translator for Romanian: building resources for a symbolic machine translation program /
Mihaela Colhon. - Craiova : Universitaria, 2013
Bibliogr.
ISBN 978-606-14-0648-7

004

Coperta: Alexandra Cristea

For this research, the author M. Colhon has been funded by the strategic grant POSDRU/89/1.5/S/61968, Project ID 61986 (2009), co-financed by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007-2013.

Apărut: 2013
TIPOGRAFIA UNIVERSITĂŢII DIN CRAIOVA
Str. Brestei, nr. 156A, Craiova, Dolj, România
Tel.: +40 251 598054
Tipărit în România

To the memory of my father Şerban Ghindeanu

and of my grandmother Elena Ghindeanu

FOREWORD

Mihaela Colhon is a researcher that, recently, has acquired esteem in the Romanian NLP[1] circles, especially for her writings oriented towards the transfer of technologies from knowledge rich to knowledge poor languages. Such technologies have grown constantly during the last decade, because the researchers have great expectations that, through alignment of resources and transfer, they will shortcut to a certain extend the tedious and expensive path towards developing resources for their own languages.

It is notorious that Romanian is a language that lacks both resources and speech-oriented technologies. The technological level of Romanian, in comparison with other major European languages, has been described in the state-of-the-art compte-rendue *The Romanian Language in the Digital Age*, recently elaborated inside the consortium of META-NET projects[2]. With respect to speech technologies, Romanian is presented in this book as situated in the last category ("no support or very weak"), but has a somehow better position ("fragmentary support") in translation technologies, text analysis technologies and resources dedicated to text and speech.

Knowledgeable about this situation, Mihaela Colhon has proposed herself an extremely ambitious task: to develop and describe the architecture of a symbolic English-Romanian machine translation system activated by voice. It is well known the complexity of such systems. Building one from scratch is a tremendously difficult task, perhaps impossible to be accomplished by a unique researcher (even if during a period of 3 years, as has allowed her the financers of the project she has been involved in). To develop a symbolic Natural Language Processing machine resembles in many respects the handcrafts activity of a jeweller (in contrast, a statistical-based machine is like a fabric from whose platen rolls the outputs do not necessitate manual intervention). In this book, Mihaela Colhon proposes us to look at the linguistic details of the process of language translation, but with the eyes of a computational linguist.

[1] NLP stands for Natural Language Processing.

[2] www.meta.eu

But who will look for a thorough description of speech technologies will not find them here. Realistic in estimating the almost never ending complexity of developing a symbolic Machine Translation system, the research efforts described in this book concentrate towards finding techniques that would accelerate the efforts of creation of the resources involved in this process. Among the many paradigms, she chooses the transfer model. Given the complexity of the task, the author, inevitably, had to focus on only parts of it, and to offer a sketchy description to the rest of the activities.

The book is organised as follows. The first chapter shortly presents speech recognition technologies. It gives a general orientation to a reader which is unfamiliar with speech techniques and technologies and looks for a bird's eye vision over the domain.

In the second chapter, a comprehensive presentation of the domain of Machine Translation (MT) is offered. The main research trends in MT are briefly described, with a particular emphasis on the transfer-based paradigm. A parallel English-Romanian corpus, word aligned, with the English part annotated for syntactic structure, allows the author to make a contrastive study over syntactic peculiarities of English and Romanian. A number of differences that characterise the English and Romanian syntax are put in evidence in the 3rd Chapter. These observations are intended to signal tricky aspects of a process of automatic acquisition of transfer rules that would map syntactic phrases from one language to the other.

In Chapter 4, the important issue of Word-Alignment (WA) is discussed. Although the literature is abundant in word alignment algorithms, and the pair English-Romanian is not bad represented as well, her own results are good enough to argue for dedicating a chapter on this issue. An interesting solution for the problem of multiple choice alignments is described: a set of filters eliminates the least probable alignments.

In Chapter 5, the author proposes an algorithm, called Treebank Generation Algorithm, for transferring constituent structures between the two parts of a word-aligned English-Romanian corpus, annotated at morphological level. The English part should also be syntactically annotated. The algorithm tries a bottom-up construction of the target tree by transferring non-terminal tags from the source language onto the

target language. A transfer is validated if the corresponding spans do match exactly through word-alignment.

The way the algorithm works makes it generates not only transfer rules but also a grammar of the target language, and this grammar will inevitably be influenced by the one of the source language – not an extremely desirable property. A suggestion would be to involve in the generation process also a grammar of the target language, not necessarily complete. Even if only partial, such a grammar would make sense in this process, because as long as it describes syntactic idiosyncrasies of the target language, the generation algorithm would cover the rest, mainly the common parts of the two grammars, and the overall result would be a much more expressive grammar, tailored to the necessities of the target language. Surprisingly, however, the evaluation shows rather high values of recall and precision. I say, surprisingly, because the corpus against which the algorithm was tested does not include all possible syntactic chunks in the annotation, up to the sentence level. As such, it is to be expected that symbols on the high layers of a sentence, even if correctly identified by the algorithm, will be reported as precision misfits. In my regard, this part needs at little bit more discussion. The author reports that an increase in the size of the corpus yielded slightly worse results (more chances to find weird examples). The generation methodology described in this chapter presents the syntactic components at an abstract level and, thus, all the particular information regarding the two languages are discarded. This property makes it language-pair independent.

It is no surprise in the author's notice that the WA information highly influences the accuracy of the Treebank Generation Algorithm. It would be, however, interesting to study the correlation between the precision of the WA and the one of the Treebank transfer. The algorithm is also unable to take decisions in cases of intersections of annotation, and we still don't know how many are these parallel structures in a bilingual corpus like the English-Romanian one.

Finally, the last chapter of the book deals with projecting transfer rules from a parallel corpus, syntactically aligned. We arrive therefore to the key issue of the whole enterprise, because syntactic transfer rules make the very core of a MT system embracing the transfer paradigm. Let's

revise shortly the logic of the author in those parts of the book describing the research pursued by herself (the Chapters 4, 5 and 6):

- There are good reasons to believe that, for most languages, the today technology has acquired good accuracy in segmentation at sentence and word level, as well as in POS/MSD-tagging and lemmatisation.

- There exist also many parallel collections of texts. Then, if you are interested to build a translator between languages *S* to *T*, you should look for such a corpus, tokenise it at word level and segment it at sentence boundaries on both sides. Then use sentence level alignment algorithms to align sentences of this parallel corpus (the literature is dense on them).

- Then go for a bilingual *S-T* dictionary and use the algorithm proposed by the author in Chapter 4 to align this corpus also at word level.

- The next step is to look for a good syntactic parser for the source language. Once you have it, use it to parse the parallel corpus on the source side. You can now use the Treebank generation algorithm, described in Chapter 5, to generate a syntactic annotation on the right side of the parallel corpus.

- Finally, use the algorithm described in Chapter 6 to generate transfer rules from the syntactic parallel corpus.

- Brilliant, isn't it? You are now very close of a transfer-based technology between *S* and *T*. Then, if you have a brand new text T_s in the source language *S* and want to translate it in the target language *T*, what you have to do is: segment, POS/MSD-tag , lemmatise and parse T_s, then apply the transfer rules and get the syntactic tree of the text in the language *T*, in which the terminal nodes carry MSD information. Still not the expected text T_t but very near to it. To get the text, exploit the alignment of terminal nodes and, through the bilingual lexicon, generate the lemmas of the words and, finally, input their MSD tags to a morpho-syntactic generator to produce the flexed forms of the words. Now, you are done!

Well, it remains to see how good is such a translation. Many questions regarding the evaluation are still unaswered. I am also convinced that the 3 main algorithms described in the Chapters 4, 5 and 6, could themselves be improved during a period of tests and enhancements. We would also

like to see, one day, the described technology and the built resources made accessible on a public site.

All these necessitate further attention. It is my sincere believe that the book you have in your hands will be used by many young researchers as an inspiring text for continuing the efforts to develop new resources for Romanian language. And this is, perhaps, the most significant contribution of Mihaela Colhon's book.

Dan Cristea

PREFACE

This book is a collection of articles and reports published by the author during the three years of post-doctoral research entitled „**eRoL: Automatic Voice Translator for Romanian**". All the translation studies presented in this book were used in order to develop a symbolic Machine Translation program with Romanian as Target Language.

Achieving fully automatic translators of high quality is a difficult task. A complete translation system usually implies some variety of intermediary linguistic representation which involves morphological, syntactic, and semantic analysis. Just like speech, the research community has been working on translation for the last 60 years, but until now performances have been achieved only for German ↔ English, German ↔ French, English ↔ Chinese, English ↔ French, English ↔ Italian, English ↔ Portuguese, English ↔ Spanish translators.

Although, not to the extent of the languages with greater electronic visibility, efforts have been invested by researchers in different places (Romanian, Republic of Moldova, Unites States, United Kingdom, Germany, Italy, etc.) to develop Romanian linguistic resources[3] such as corpora, dictionaries, wordnets and collections of linguistic data in both symbolic and statistical form (Cristea and Forăscu, 2006). These efforts made possible the developing of new processing tools for Romanian language, such as the system presented in this book.

The main characteristic of the **eRoL** system is represented by the fact that the system is dedicated to Romanian in a manner in which all the implemented translation representations and processing methods have Romanian as Target Language. For this first version of the **eRoL** system the author decided to choose English as Source Language because, by far, English is the most studied natural language in the NLP community. In this manner, the system could benefit from the existing English parsing tools or mechanisms that are already state-of-the in the NLP field.

The book is organized in six chapters as follows: Chapter 1, „*Speech Recognition Technologies*" is a survey of the way Speech Recognition

[3] Cristea, D., Forăscu, C.: Linguistic Resources and Technologies for Romanian Language", Computer Science Journal of Moldova, vol. 14, no. 1(40). (2006)

systems evolved in the last forty years. This chapter includes the basic terms and concepts of an Automatic Speech Recognition system and then focuses on a well-known English Speech Recognition system - Microsoft Speech Recognition as this product was chosen for the voice-user interface of the eRoL system.

Chapter 2 „*Machine Translation Systems. Current Approaches*", is like a small treatise of the current approaches concerning Automatic Machine Translation systems. Besides overviewing the current research in this field, the author points out from the beginning the importance carried by the syntactic information of the source texts during the process of generating accurate translations in the form of Target Language grammatical motivated texts.

Chapter 3 „*A Contrastive syntactic Study of English and Romanian Texts*" continues the study motivated in Chapter 2 by making a contrastive study over some common constructions in the considered languages from their syntactic information point of view.

Chapter 4 „*Automatic Lexical Alignment Between Syntactic Weak Related Languages. Application for English and Romanian*" addresses the problem of finding correspondences between parallel English and Romanian texts at their „anchor words" level (a category that includes all the content words but also some functional words of the two languages). This study was needed in order to support the word-alignments annotations of the English-Romanian corpus based on which the Treebank Generation Algorithm presented in Chapter 5 was developed.

Chapter 5 „*Generating Romanian Syntactic Trees from Parallel English Ones*" presents a mechanism named Treebank Generation Algorithm that constructs syntactic tree structures for Romanian phrases based on the structure of their translations in English. The algorithm is developed by taking into account the morpho-syntactic word annotations of an English-Romanian parallel corpus and is guided by the corpus word alignments. The resulted parallel syntactic sequences generated by this algorithm will be used to construct the main resource of the **eRoL** system upon which the translations are generated.

Chapter 6 „*Acquiring Syntactic Translation Rules from the English-Romanian Treebank*" presents the main translation resource of the **eRoL**