A Design Framework for Foreign Language Learning Applications

Mihaela Colhon

University of Craiova
Department of Computer Science

mcolhon@inf.ucv.ro

ABSTRACT

In this article we present a method for generating and applying syntactic motivated patterns in order to develop a foreign language learning mechanism. The patterns have been extracted from a parallel corpus that has been automatically annotated for morpho-syntactic descriptions and syntactic constituents. The proposed language learning framework is not designed around the well-known list of words. Using this application, the user does not necessarily practice the foreign language lexicon, which is supposed to be known at a medium level. Instead, with this application, the user is guided to learn the so-called "translation knowledge".

Auhor Keywords

Machine Translation, Learning Schema, Human-Computer Interaction.

ACM Classification Keywords

I.2.1: Natural language interfaces; H.5.2: User-centered design.

INTRODUCTION

The task of natural language translation from one language to another is attributed to human intelligence. Besides the knowledge of the two languages, it requires an "understanding" of the source language text and to transform the mental picture created through understanding of the source text into its target language representation.

The design of the systems developed for foreign language instruction needs to be grounded on what we know about human learning, language processing and human-computer interaction. Nevertheless, machine translation provides a good starting point for foreign language learning giving a rough understanding of the natural language constructions, generating alternative translation choices and creating proper resources of example translations.

Human-Computer Interaction (HCI) deals with all the relevant aspects concerning the design, the implementation and the evaluation of interactive systems. From the HCI point of view, the interactive applications of these days have to face the following challenges:

- Speech/Text Recognition and Understanding
- Speech/Text Generation
- Interactive Machine Translation

Natural language processing (shortly, NLP) can be done on several levels. The first level implies phonology and phonetics data for speech recognition and understanding or morphology data in case of natural language texts recognition and understanding. The next levels imply syntactic, semantic and pragmatic based processing.

For the applications with natural language interfaces, a number of problems are encountered due to the ambiguity of natural language constructions, to the huge amount of involved lexical knowledge and of natural language utterances.

The main goal of this study is to develop a design framework for a foreign language learning software product. Because learners vary both in how they learn and what they want and need to learn, we can not say which is the best way of learning. When creating an application like this it is important to keep in mind that learning is a highly subjective process, and for this reason the user characteristics and ergonomy must to be carefully treated [9]. Trăușan-Matu [12] considers that a user friendly application must allow easy, effective, safe and without risks usage and must to be useful, easy to learn and easy to remember.

A possible solution consists in developing applications that can follow the user's knowledge and needs [8]. In this paper we give a possible design framework for foreign language learning applications.

The remainder of the paper is organized as follows: in Section 2 are presented the previously research works based on which the proposed foreign language learning mechanism is developed. The proposed framework for foreign language learning is detailed in Section 3 together with a particular case exemplification concerning nominal phrase learning mechanism. The manner in which the learning schema can be modeled upon the user knowledge level is presented in Section 4. Finally, the conclusions and future work guidelines can be read in Section 5.

AUTOMATIC TRANSLATION SYSTEM AS TARGET LANGUAGE LEARNING APPLICATION

Given a source-language (e.g., Romanian) sentence, the problem of machine translation is to automatically produce a target-language (e.g., English) translation. We found that translating lemmas and morpho-syntactic descriptors of the source language sentence words and then generating, the corresponding word-forms in the target language achieves better results than the baseline phrase-based translation model. In view of this, we have

developed a symbolic Machine Translation program [1] with English as Source Language and Romanian as Target Language, the eRoL System[1].

In any automatic translation system, getting syntactic data with the scope of producing linguistic information about the source sentence structure involves a pre-processing step of the source-language sentences also known as *parsing*. The resulted structure of a sentence has to indicate the relationships that exist between the words of that sentence or how the words are grouped into syntactic phrases like noun phrases (NPs), prepositional phrases (PPs), verb phrases (VPs), etc. Usually, all these information are stored in a tree representation.

LINGUISTIC RESOURCES

Parallel corpora can generate extremely valuable linguistic knowledge for machine translation studies such as: in *direct approaches*, parallel corpora are used to extract information about lexical units (how a particular word is translated in a certain construction), in *transfer-based approaches*, parallel corpora are used to extract transfer rules while in *statistical approaches*, these corpora are used to extract translation rules and to assign probabilities to possible translations.

Linguistic resources upon which the translation data are created are based on parallel natural language constructions extracted from multilingual corpora. Such a corpus, called JRC-Acquis, is the compiled part of the parallel texts from the Acquis Communautaire legislative documents. The Acquis Communautaire is a collection of parallel texts in 22 official European Union languages, including English and Romanian.

Two segments of texts from a pair of parallel texts which represent reciprocal translations make a *translation unit* [13]. Phrase based machine translation techniques work with pairs of phrases, the so-called translation units, which are consistent with respect to the inner word transation-alignment: the words of a phrase are aligned to words of the other phrase and not to the outside words.

The current practice in phrase-based translation has shown that creating large syntactic phrase tables allow the learning of "translation knowledge". Indeed, most of the phrases syntactical motivated are expected to be translated without interleaving with other phrases/words. In general, noun phrases tend to obey the above rule to a much greater degree. Conversely, verb phrases usually suffer modifications in structure during translation, caused by adjunct movement [4].

A *parallel treebank* is a special type of parallel corpus that has been grammatically annotated in order to identify and label different syntactic information about the text. Such syntactic information usually implies incorporating into the text markers which indicates the syntactic dependencies relations or the phrase-based structures[2].

Techniques that were applied on the corpus sentences include tokenization, part-of-speech tagging and lemmatization. Part-of-Speech (POS) tagging, describes the annotated words in terms of grammatical tagging (Noun, Verb, Pronoun, etc.) and morphological information (sequences of codes about the inflectional features of the words such as gender, number, person, case, etc.). Often, POS tagging can include lemmatization, by indicating the lemmas of the words. A POS tagset generated during the MULTEXT-East project [5] includes *morphosyntactic descriptions* (*MSD*) for all the languages of the project (including Romanian).

The application described in this paper uses 4389 English-Romanian parallel patterns extracted from a English-Romanian Treebank [2] with syntactic constituents.

The Treebank was constructed upon 1420 sentences of the English-Romanian corpus developed at the Alexandru Ioan Cuza University of Iaşi by the Natural Language Processing Group of the Faculty of Computer Science. For this bilingual corpus, the English and Romanian parts of the JRC-Acquis corpus were used.

In the considered Treebank, each translation unit has representations at several levels:

- at lexical level (sequences of words);
- at POS level (part of speech of the annotated words);
- at phrase-based level (syntactic constituents)

Such a collection was designed to be used in a translation automatic mechanism with the scope of moving from words to phrases as the basic unit of translation [3].

PATTERN-BASED TRANSLATION MODEL

The knowledge enclosed in the eRoL system development is represented by syntactic patterns defined in terms of morpho-syntactic specifications and phrasal tag specifications in the form introduced by the Penn Treebank project[3]. In this manner, the system uses the so-called *informed language model* which is described by word-forms but also by MSD specifications or POS data.

The representation of the parallel English-Romanian parse trees of the Treebank are flattened into linear string form following the bracket representation for syntactic trees with constituents [11]. From the Treebank, the system considers the syntactic level representations of the parallel phrases by focusing mainly on POS tags instead of real words.

In Figure 1 it is given a screen shot of the eRoL Web interface.

[1] There is a web demonstrator of this system available at http://www.mcolhon.ro/erol/eRoLsystem.html.

[2] Traditionally, phrases markers are taken to be syntactic constituents of a sentence.

[3] The web address of the project is http://www.cis.upenn.edu/~treebank/.

Figure 1. The eRoL System Web Interface

In what follows we will exemplify the implemented translation mechanism, by considering the following English sentence:

A brown and beautiful cat plays with a big ball.

The input sentences are parsed with Stanford Parser[4] tool [6] in order to mark the syntactic phrases and the POS data of the sentences' words. For the considered example, the representation obtained with Stanford Parser has the following syntactic tree form:

```
[S [NP DT/a [JJ brown] CC/and [JJ beautiful] [NN cat]] [VP [VBZ plays] [PP IN/with [NP DT/a [JJ big] [NN ball]]]]]
```

The system finds in the English-Romanian Treebank the Romanian phrase corresponding to the last constituent of the given phrase, which is:

```
[PP IN/with [NP DT/a [JJ *] [NN *]]
```

and, after translations word-to-word are made, the generated Romanian phrase is:

```
[PP IN/(împreună)cu [NP o: Tifsr minge:Ncfsr voluminoasă:Afpfsm]
```

where `IN/(împreună) cu` represents the translation of `IN/with` and `[NP o: Tifsr minge:Ncfsr voluminoasă:Afpfsm]` is the translation of the English noun phrase `[NP DT/a [JJ big] [NN ball]]`.

FOREIGN LANGUAGE LEARNING FRAMEWORK

Any learning process is schema-based. The recent interests in language learning schema match up closely with an ongoing theme from cognitive psychology that bases all long-term learning on the construction of schemas. This view sees the learning of new material as involving integration into old material [7].

Using the linguistic resources constructed for the eRoL system, an application for English language learning can be developed for Romanian users. The presented learning mechanism can be applied to any other pair of languages by replacing Romanian with the name of the user native language and English with the foreign language name.

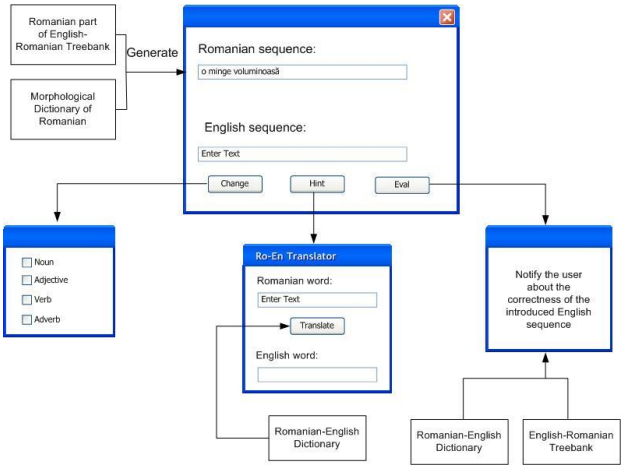

Figure 2. The Design Framework of the Application

In Figure 2 is shown the design of the proposed learning application. Basically, it consists of a simple main window where the Romanian sequences are displayed with the scope of receiving the user translations. The main window has three controls designed for the following tasks:

- the "Change" button let the user to decide which kind of patterns wants to practice: Noun Phrases, Adjective Phrases, Verb Phrases or Adverb Phrases;
- the "Hint" button uses a Romanian-English dictionary resource in order to help the user with the Romanian to English word based translations;
- the "Eval" button starts the evaluation of the English sequence introduced by the user and highlight the user's errors (if it is the case). If the evaluation ends successfully, another Romanian sequence is generated in order to be translated.

In what follows we give the algorithm that guides the learning process by means of the patterns extracted from the English-Romanian Treebank.

Foreign Language Learning Algorithm

```
1. listPOS_CW <-{Noun, Adjective, Verb, Adverb}
2. CW <- listPOS_CW.pop()
3. while (!user.wantsToStop())
3.1. extract all patterns including CW
3.2. do
3.2.1. generate sequences of words based on extracted patterns
3.3. until (user.wantsToStop() ||
       user.wantsToChange() ||
       user's translations are correct)
3.4 if (user's translations are correct)
3.4.1 CW <- listPOS_CW.pop()
3.5 endif
```

[4] Because the English sentences are processed using Stanford Parser (web page: http://nlp.stanford.edu/software/lex-parser.shtml), PENN Treebank parse trees are generated. As a direct consequence, the English words are annotated with PENN POS tags, as this is the tagging standard used by Stanford Parser.

```
3.6. if (user.wantsToChange())
3.6.1. CW <- user's CW choice
3.7 endif
4. endwhile
```

This learning schema, we consider, will enable users to practice the foreign language learning in a gradual manner, starting with easy sequences formed by few words and continuing with larger sequences. Also, the order in which the patterns are generated ensures a proper language learning method as the application starts by generating noun phrases and ends with the verb phrases (more complex than the nominal ones). This order is ensured by the `listPOS_CW` stack. Obviously, the user can change the category of natural language constructions he wants to practice (by selecting the "Change" button from the main window).

Exemplification. Noun Phrases Learning Mechanism

In order to exemplify the manner in which the learning process is ensured in the proposed framework, in this section we will present the NP learning design.

The simplest patterns for Noun Phrases consist of a single noun. In a first phase the application will generate sequences for `N*` patterns like `Ncmsoy`. Here are several entries of the Morphological Dictionary of Romanian [10] corresponding to this MSD sequence: "fratelui" (in En. "of the brother"/ "to the brother"), împăratului (in En. "of the king"/ "to the king"), "miliardarului" (in En. "of the billionaire"/ "to the billionaire").

The user is passed to the next level of learning process if its translations for the current level are correct. A next level for NP learning is generated by using patterns which include noun tokens but also extra function words tokens like in the following examples:

- "această chinezoaică" (in En. "this Chinese woman") or "această antilopă" (in En. "this antelope") corresponding to the pattern `Dd3fsr/această Ncfsrn,`
- "unui cortegiu" (in En. "of a procession"/ "to the procession") or "unui idol" (in En. "of an idol"/ "to the idol") corresponding to the pattern `Timso/unui Ncms,`
- "în conformitate cu materia" (in En. "in accordance with the matter") or "în conformitate cu decizia" (in En. "in accordance with the decision") corresponding to the pattern `Spca/în_conformitate_cu Ncfsry,`
- "astronava acestuia" (in En. "its spaceship") or "căruța acestuia" (in En. "its carriage") corresponding to the pattern `Ncfsry Pd3mso/acestuia.`

If the user matches the translations for the received sequences, more complex patterns will be generated. The larger the sequences are, the fewer the possible translations will be. This is determined by the fact that a sequence complete from the meaning point of view, usually has a unique translation.

ADJUST THE LEARNING SCHEMA UPON THE USER KNOWLEDGE LEVEL

The manner in which the application is designed permits automatic adjustments upon the user's level of expertise. Indeed, the level of difficulty for the generated sequences that must be translated can be automatically adjusted by restricting the data involved in the generation phase. More precisely, the application could use:

- only a part of the Romanian Morphological Dictionary which, in this case, means fewer words that will be used in the generated Romanian sequences
- patterns limited to medium size for beginners or maximum size for advanced users

Using these simple restrictions users with less expertise could be trained using simple sequences made from several words (that will be repeated if the user fails with his translations) while the advanced users will be trained using the whole lexicon – meaning, a large vocabulary that has to be passed and using complex sequences made upon the largest patterns from the Treebank.

CONCLUSIONS AND FUTURE WORK

In this paper, we present a design framework for foreign language learning applications. The proposed learning mechanism ensures the development of a proper learning schema using the existing linguistic resources for the involved languages, in our case Romanian and English. Our proposal addresses the usability issues concerning this kind of applications by proposing a progressive learning schema. Our permanent concern is to improve the set of parallel patterns and also the entries of the used English-Romanian dictionary in order to cover more utterances in the considered languages.

ACKNOWLEDGMENTS

The author wants to thank the Natural Language Processing Group of the Faculty of Computer Science, Alexandru Ioan Cuza University of Iaşi, for providing the English-Romanian corpus upon which the presented study was made.

REFERENCES

1. Colhon, M. 2013. *eRoL: Automatic Voice Translator for Romanian. Building Resources for a Symbolic Machine Translation Program*, Universitaria Publishing House, Craiova 2013, 151 pages, ISBN: 978-606-14-0648-7.
2. Colhon, M. 2012. Language Engineering for Syntactic Knowledge Transfer, In *Computer Science and Information Systems Journal*, 9(3), ISSN 1820-0214, 1231-1247.
3. Colhon, M. 2012. Acquiring Syntactic Translation rules from a Parallel Treebank, *Journal of Information and Library Science INFOtheca*, 2, XIII (Dec. 2012), 19-32, Serbian Academic Library Association, ISSN: 1450-9687 (e-magazine)
4. Colhon, M. 2011. A Contrastive Study of Syntactic Constituents in English and Romanian Texts, In

Proceedings of the Workshop "Language Resources and Tools with Industrial Applications", Eds. Iftene A., Trandabăţ D.-M., 11-20
5. Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M. and Vitas, D. 2003. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages, In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, 25-32, Budapest, Hungary.
6. Klein, D., Manning, C. D. 2003. Accurate Unlexicalized Parsing", In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423-430
7. MacWhinney, B. 1995. *Evaluating foreign language tutoring systems*. In: Holland, Kaplan and Sams, 317-326.
8. Pribeanu, C. 2001. *Human Computer Interfaces Design*, MATRIX ROM Bucureşti Publishing House (in Romanian).
9. Pribeanu, C. 1999. *Human Computer Interaction*, Didactic and Pedagogical Publishing House, Bucureşti (in Romanian).
10. Simionescu, R., Cristea, D., and Haja, G. 2012. Inferring diachronic morphology using the Romanian Thesaurus Dictionary. In *A. Moruz, et al. (eds.). Proceedings of the 8th International Conference "Linguistic Resources And Tools For Processing Of The Romanian Language"*, "Al.I.Cuza" University Publishing House, Iaşi, 85-92, ISSN 1843-911X.
11. Taylor, A. 1996. Bracketing Switchboard: An Addendum to the Treebank II Guidelines, http://www.seas.upenn.edu/~jmott/prsguid2.pdf
12. Trăuşan-Matu, Ş. 2000. *Evolved Human-Computer Interfaces*, MATRIX ROM Bucureşti Publishing House (in Romanian)
13. Tufiş, D. and Ion, R. 2007. Parallel Corpora, Alignment Technologies and Further Prospects in Multilingual Resources and Technology Infrastructure. In *Proceedings of SPED 2007*.

APPENDIX A. GLOSSARY OF NOTATION

The following table gives the notation used in this paper:

MSD tag	The meaning of the notation (according to MULTEXT-East lexical specifications)
Afpfp(s)	Adjective qualifier positive feminine plural(singular)
Dd3fpr-	Determiner demonstrative third feminine plural direct
Tif(m)sr(o)	Article indefinite feminine(masculine) singular direct(oblique)
Ncf(m)sr(o)y(n)	Noun common feminine(masculine) singular direct(oblique) +definiteness(-definiteness)
Pp(d)3mso	Pronoun personal(demonstrative) third masculine singular oblique
Spsa	Adposition preposition simple accusative