NLP Tools for CALL: the Simpler, the Better

Olivier KRAIF, Georges ANTONIADIS, Sandra ECHINARD, Mathieu LOISEAU, Thomas LEBARBÉ, Claude PONTON
LIDILEM, Université Stendhal Grenoble 3
Grenoble - France
{kraif; antoniadis; echinard; loiseau; lebarbe; ponton}@u-grenoble3.fr

Abstract

This paper addresses the problem of the implementation of Natural Language Processing (NLP) tools for Computer Assisted Language Learning (CALL). After analysing why some complex NLP applications appear to be inappropriate for CALL, we show, with precise examples, that the more promising progresses are expected to come from NLP basic processes, as morphological tagging or lemmatisation.

1 Introduction

One has to admit that Natural Language Processing (NLP) applications are not very widespread in the real world of CALL products. This is mainly due to three reasons: NLP techniques often lack reliability, NLP products and resources are quite expensive and difficult to implement, and the end-users (teachers, learners, conceptors, editors…) are not aware of NLP possibilities. The aim of this paper is to cope with the latter problem, by demonstrating that the two former obstacles can be overcome. Indeed, we think that many simple, modest and well mastered techniques can bring relevant improvements to the existing on/off-line CALL applications.

First, we review some of the most interesting achievements illustrating CALL/NLP enrichment. Through the examples of systems such as Alexia (Chanier & Selva, 2000) or Exills Platform, (Brun *et al*. 2002) we suggest how to use NLP resources in a simple and effective way.

To illustrate this approach, we describe a tool that we are developing for our own language curriculum, allowing to generate automatically gap-filling exercises. After a short description of the forthcoming developments of our NLP tools, we finally sketch the more promising applications of NLP for language learning.

2 NLP for CALL applications

In a communication of the TALN 2003 Conference, Jean Véronis implicated the natural "hubris" of the NLP community. Decades ago, Bar-Hillel (1964) already warned the community about its *penchant* for *hubris*, arguing that one should find "a judicious and modest use of mechanical aids". Indeed, from the beginning of NLP, in the early fifties (cf. the IBM-Georgetown translating machine, demonstrated in 1954), researchers have claimed that they could solve complete problems of human communication, like translation, using computing models of *encoding-decoding*. Half a century of huge progresses in computing has proven that human language and communication were not such simple matters.

Anyone has used a state-of-the-art machine translation system has faced the inherent limitations of NLP. The best systems, using a wide range of NLP techniques (morphological analysis, syntactic analysis, sense disambiguation, morphosyntactic generation, etc.) and resources (dictionaries, grammars, ontologies, etc.), often result in strange linguistic productions. Such tools may be very useful to make the communication easier, for instance in translating the global content of a website or an e-mail, but we're still a long way off getting enough quality to use them in the classroom context. Using one the most famous systems available on Internet, to translate the following English sentence:

The InSTIL SIG is delighted to announce our next symposium to be held in Venice on June 17-19, 2004

one gets the understandable but not completely correct Italian sentence:

Il SIG di InSTIL si diletta per annunciare il nostro simposio seguente da tenere a Venezia su giugno 17-19 2004

Would a teacher accept to use a dictionary giving wrong or erroneous information 1 time out of 10? Of course not. So, what about NLP tools?

We have also tested a didactic version of a well-known spell checker, designed to help learners in correcting their own productions. This product includes interesting features, like the automatic linking between analysed words and grammatical explanation. But it also gives access to the

complete syntactic analysis, which is often wrong, detect grammar errors where there are not: these functionalities are not well-mastered enough for such a didactic use. It is, according to us, a typical case of NLP *hubris*.

Anyway, we simply claim that Bar-Hillel's "modest use" is possible in the CALL field, and that some techniques are reliable enough to be implemented. For instance, instead of giving a *single* and *full* translation, a machine translation system could give a list of lexical and grammatical clues (rules, various equivalents, partial analyses, ambiguous terms) that were computed during the process, and that could give an interesting, and more reliable - though fragmentary - information to a learner.

A good example of such an approach is the Alexia System (Chanier & Selva, 2000). The authors start from explicit didactic principles: they note that lexical activities yield better results when they correspond to the organisation of the *mental lexicon* of the learner. To generate automatically activities that fit with this observation, they use an electronic dictionary, organised as a word net (i.e. entries of the dictionary are connected with each other by semantic links like hyponymy, hypernymy or antonymy). Given a textual corpus, such a dictionary allows to extract automatically word occurrences that are related to a specific semantic field (e.g. *salaires*, *patron*, *usines*, *licencierait*, *dirigerai*, *condition*, *emploi*). From the extracted *concordance*, i.e. the given occurrences within their local context, Alexia automatically generates a fill-gap exercise. To be efficient, the extraction of such a concordance involves simple NLP techniques as morphosyntactic tagging and lemmatisation:

- With *tagging*, words are given tags which describe morphological properties like: (IT) *imparando -> verb*, *gerund*
- Lemmatisation gives the canonical form of a word: *imparando -> imparare*.

These tools are very useful to process the highly inflected languages, because they make possible the search of every occurrence of a given verb (e.g. *imparare*), in whatever form (e.g. *imparo*, *impari*, *impara, etc.*).

The Exills platform, developed by Xerox (Brun et al., 2002), is another interesting example of "modest and judicious use" of NLP. In the Exills environment, the learner is immersed in a virtual world where his avatar can interact with other learners, and with robots. To compute its mission, the learner has to fulfil various communication tasks. In order to force the learners to use the target language between each other, the exchanges are controlled by a language identifier, based on statistic NLP models. At every moment, comprehension aids using NLP are proposed:

- the learner can access to a dictionary, with a contextual disambiguation of the entry: for instance, for (EN) *mouse*, only the "computer" sense will be given if the previous exchanges concern the computer field;
- a *phonetiser* allows to give the probable spelling corresponding to a wrongly spelled word: if a French learner searches an article on **Venezzia*, the phonetiser can give the correct Italian form;
- a *conjugator* allows to display the inflected forms of a given verb;
- the *Xelda tagger* can indicate the morphological description of an unknown word.

Morphological tagging and lemmatisation are reliable tools, because the various possible analyses may be given to the learner, without any risky conjecture on a precise *interpretation* (which is often outside the scope of a computer).

3 An NLP-based activity generator

Following this example, we have developed a simple NLP-based tool for our own language curriculum.

Given a tagged and lemmatised corpus[1], this tool aims at generating gap-filling exercises. A set of generation rules allows to select the form and the grammatical features of the words that are to be removed, and the information to be shown in the gaps.

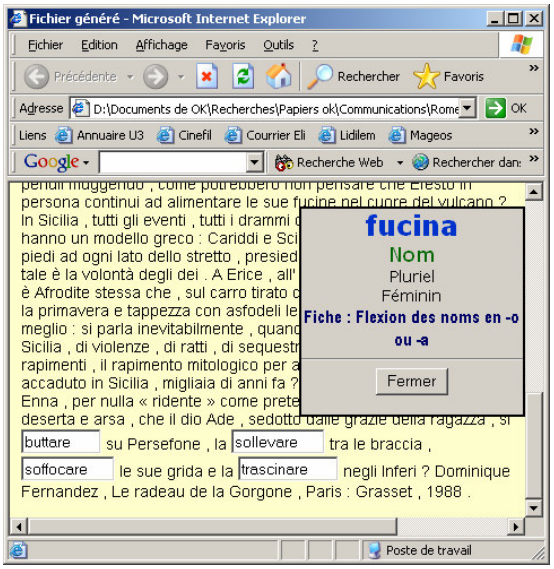

Fig.1 - Generated gap-filling exercise with contextual aid

[1] Using Xelda, a Xerox XRCE product (see http://www.xrce.xerox.com)

Selection criterion	Type	Example of activity	Expected answer	Involved NLP functions
Semantic	lexical spotting	Spot every word related to the (IT) "*aula*" topic	Spotting of "*tabella*", "*insegnante*", "*lezione*", "*alluno*", etc.	morphosyntactic tagging, lemmatization, semantic net interrogation
Semantic	lexical question	Give the word corresponding to the following definition: "*sala per pubbliche riunioni*"	Entering of "*aula*"	morphosyntactic tagging, lemmatization, dictionary interrogation
Semantic	lexical question	Give an Italian translation for "*to learn*"	Entering of "*imparare*"	morphosyntactic tagging, lemmatization, bilingual dictionary interrogation
Morpho-syntactic	gap-filling	Replace every infinitive verb in the gaps, using the appropriate tense	Replacement of "*imparare*" by "*imparassi*"...	morphosyntactic tagging, lemmatization
Morpho-syntactic	lexical question	What would be the contrary of the adverb "*difficilmente*"?	Entering of "*facilmente*"	morphosyntactic tagging, lemmatization, semantic net interrogation
Morpho-logical	lexical spotting	Spot every word derived from the verb (IT) "*tradurre*"	Spotting of "*traduttore*", "*traduzione*", "*ritradotto*", etc.	morphosyntactic tagging, lemmatization, stemming
Morpho-logical	gap-filling	Fill every gap by a word of the "*tradurre*" verb family	Entering of "*traduttore*", "*traduzione*", "*ritradotto*", etc	morphosyntactic tagging, lemmatization, stemming

Tab.1 - Examples of activity generators

Moreover, clicking on the text words gives access to additional information, in order to provide the learner with a comprehension aid: grammatical information, manually added notes, or links to external resources (text documents, websites, etc.).

In the example displayed on figure 1, gaps have been created for every verb at the "*passato remoto*" tense. Moreover, links are automatically added depending on lexical and grammatical criteria specified by the rules.

Such a tool is valuable for the teacher, because rules can be reused on any text, and for the learner, for whom it yields more autonomy.

These parameters are accessible to the end-user through a control panel. This control panel shall be relevant from the didactic point of view, that is why the controls may be transcribed into a parameter set. By the mean of a simple form, the user may define:

a) Which are the units to be removed from the text. Any linguistic feature should be used for this definition: lemma (e.g. *imparare*), part-of-speech (ex. *verb*), morphosyntactic description (ex. *past tense*), or even sense (e.g. "classroom" semantic field - this functionality has not been implemented yet).

b) What information has to be given in the gap: none, the lemma, the morphosyntactic features, a synonym, a definition, (not implemented yet) etc.

c) If the removed words should appear or not as an ordered list in the text header.

d) If the learner's answer should initiate a feed-back process immediately after it was entered.

It is clear that the definition of linguistic features in a) involves a simple transcription process in order to determine the parameters of the NLP script: the tagged and lemmatized texts handled by the generator use specific codes for morphosyntactic description. Declarative features as "Verbo, Prima coniugazione, Indicativo, Presente, Prima persona, Singolare" will be transcribed into a parameter set, related to XML attributes of our annotation format: base="er$", ctag="verb", msd="IndP SG P1".

Using additional lexical resources, as dictionaries, word nets, other activities may be generated following the same model (i.e. starting from a tagged text). For instance, a semantic net allows to find related senses of a given word (synonyms, antonyms, etc.): as for Alexia, gaps can be selected on a semantic basis. Table 1 displays some example of lexical activities based on this framework: gap-filling, lexical spotting and

lexical questions. The two latter functionalities have not been implemented yet in our prototype.

To avoid errors in the word selection, every ambiguous form (i.e. forms that bear multiple analyses like *studi*-congiuntivo or *studi*-indicativo) may be discarded. Anyway, such errors may be minor, because there is no strong assertion behind the fact of removing a word. The instructions going with the exercise may take into account possible mismatches, e.g.: "Replace, *when it is the case*, the infinitive form by the conjugated *passato remoto* form".

4 Prospects

During this prototype designing, three application fields have appeared to be very promising for NLP-based tools:

- *Activity generation*: examples of lexical activities have been given, but the range of possibility is large: exercises about flexional[2] or derivational[3] morphology, corpus mining using monolingual or bilingual concordancer as suggested by Nerbonne (2000), etc.

- *Interactive aids*: NLP can make easier the access to relevant linguistic resource, as specific grammar points or dictionary entries, allowing a more adapted and context sensitive search.

- *Evaluation*: this point is indeed the more difficult for CALL. Usually the learner can be only evaluated in the case of yes / no questions or multiple-choice tests. To go further, we think that the more realistic and promising application concerns the evaluation of simple lexical productions. We are currently studying a three levels protocol for the evaluation of a given answer with respect to the expected correct answer. If the given answer is different, three cases are considered:

1- Spelling error: if the entered chain does not exist in an inflected form dictionary, one can suppose that it bears a spelling error. If the chain is very close to the correct answer, a message can be displayed, warning about the spelling error. Else, a list of resembling existing words can be proposed to the learner, asking him to make a choice.

2- Morphosyntactic level: at this stage, the answer is integrated in the linguistic context of the activity (for instance, the sentence where the gap was done, in a gap-filling exercise), in order to compute a morphosyntactic analysis with tagging and lemmatization. If the lemma is the same than the lemma of the correct answer, a warning can be displayed about the difference in the morphosyntactic features (e.g. "wrong tense", "wrong number", etc.).

3- Semantic level: in the case of a different lemma, a semantic word net is searched in order to check whether a close semantic link (synonymy, hypernymy, hyponymy, meronymy, antonymy) can be found between the given answer and the expected one. Then, a warning can be displayed such as "be more precise", "not exactly", etc.

5 Conclusion

With the example of machine translation, we have shown that traditional NLP applications, as spell and grammar checkers, translators etc. cannot be used *as they are* for CALL. But these applications include many simple components, as morphological *taggers*, lemmatisers, conjugators, phonetisers, etc. that are sufficiently efficient and well-mastered to deserve from now on a place in the CALL field. In the future, more complex tools, specially designed for didactic purposes, may be developed from these NLP components, and new pedagogical practices may appear, taking advantage of these new possibilities. We are now implementing a platform, called MIRTO (Antoniadis & Ponton 2004), dedicated to host the forthcoming NLP tools in a coherent framework, and to make them accessible to teachers and designers without programming skills: with MIRTO, we would like to show the great potential of a complete NLP-based authorware.

[2]examples at: http://www.pomme.ualberta.ca/devoir/

[3] http://www.robobunny.com/cgi-bin/dislexicon/dlc

References

G. Antoniadis & C. Ponton. 2004. *Mirto : un système au service de l'enseignement des langues*. UNTELE'2004, 17-20 mars 2004, Compiègne.

Y. Bar-Hillel. 1964. *The future of Machine Translation*, In "Language and Information : Selected Essays on their Theory and Application", Addison-Wesley Publishing Company, Inc., pp. 180-184.

C. Brun, T. Parmentier, A. Sandor, F. Segond. 2002. *Les outils de TAL au service de la e-formation en langues*. In "Multilinguisme et traitement de l'information", sous la dir. de F.Segond, Hermès, Paris, 2002, pp. 223-250

T. Chanier, & T. Selva. 2000. *Génération automatique d'activités lexicales dans le système ALEXIA*. In "Sciences et Techniques Educatives", 7, 2 : 385-412. Paris : Hermes

J. Nerbonne. 2000. *Parallel texts in computer-assisted language learning*. In "Parallel Text Processing", Jean Véronis Editor, Kluwer Academic Publishers, Dordrecht, p. 299-311.