G. Biswas et al. (Eds.) (2012). Proceedings of the 20th International Conference on Computers in Education. Singapore: Asia-Pacific Society for Computers in Education

Towards a system architecture for ICALL

Elena VOLODINA[a*], Hrafn LOFTSSON[b], Birna ARNBJÖRNSDÓTTIR[c], Lars BORIN[a] & Guðmundur ÖRN LEIFSSON[d]
[a]*Språkbanken (Swedish Language Bank), University of Gothenburg, Sweden*
[b] *School of Computer Science, Reykjavik University, Iceland*
[c] *School of Humanities, University of Iceland, Iceland*
[d] *School of Engineering and Natural Sciences, University of Iceland, Iceland*
*elena.volodina@svenska.gu.se

Abstract: In this paper, we present an on-going project whose overall aim is to develop open-source system architecture for supporting ICALL systems that will facilitate re-use of existing NLP tools and resources on a plug-and-play basis. We introduce the project, describe the approaches adopted by the two language teams, and present two applications being developed using the proposed architecture.

Keywords: Intelligent Computer-Assisted Language Learning, Natural Language Processing, System Architecture, Writing Feedback, Exercise Generation

1. Introduction

It is a remarkable fact that, despite the existence of various accurate Natural Language Processing (NLP) tools and resources that can potentially benefit language learning, very few projects are devoted to development of Intelligent Computer-Assisted Language Learning (ICALL)[1] applications.

This situation calls for a change. We have therefore joined our forces in order to design and develop open-source system architecture for supporting ICALL systems. We make the architecture open-source in order to encourage participation from other researchers and developers, and to facilitate re-usability of existing NLP tools and resources. To test the architecture, we are currently developing two specific applications for Icelandic and Swedish, described in the sections below.

2. An emerging ICALL platform for Icelandic

In the Icelandic part of the project, we are developing a platform which chains together various NLP tools. Internally, the platform uses the *Text Corpus Format* (TCF; an XML format), proposed in the *WebLicht* project [1], for communication of information between the various components. All annotations for a particular text are stored in a single XML file, where each annotation, e.g. at the level of tokens, part-of-speech (PoS) tags, or constituents, is stored in a separate layer. In addition to using the layers proposed in WebLicht, we have added a layer for information about grammatical errors. An important part of the design of the platform is language-independence, i.e. it should be simple to add

[1] Intelligence in CALL systems can be understood differently by different researchers. In this paper, we define ICALL as NLP-based CALL, i.e. intelligence in CALL is ensured through the use of NLP tools and resources like parsers, part-of-speech (PoS) taggers, corpora, lexicons, etc.

NLP tools for supporting various languages, the only requirement being that they must communicate with other tools in the platform using TCF.

Currently, no ICALL application exists for the Icelandic language. However, a free and open CALL application named *Icelandic Online (IOL;* http://icelandiconline.is) was launched in 2004. IOL is a pedagogically driven web course which has evolved over time [2]. It has, currently, almost 90,000 registered users, and has received universally positive feedback. In IOL, second-language learners of Icelandic receive feedback from a teacher regarding short written texts. Currently, teachers use special codes for hand-marking specific types of errors, i.e. spelling errors, feature agreement errors, case errors in objects of verbs, etc.

In order to automate part of the error-marking and to test our platform, we are currently in the process of developing a web service which allows students of IOL to send texts to the service for the purpose of detecting particular types of grammatical errors. This will allow students to correct potential errors, re-submit the texts for error detection again, and so forth, before finally submitting the text to the teacher. The web service merely identifies error candidates, but does not attempt to correct errors. Writing feedback is immensely labour intensive and this service will allow students to identify certain types of errors of form, allowing the teacher to focus on content feedback.

The web service submits Icelandic text (input by a student) to the platform, which, in turn, uses tools from the IceNLP toolkit [3], i.e. a tokeniser, a PoS tagger and a finite-state parser, to detect the following types of grammatical errors: (1) feature agreement errors in noun phrases, i.e. errors in gender, number and case; (2) feature agreement errors between subjects and verb complements; (3) feature agreement errors between subject and verbs, i.e. errors in person and number; and (4) incorrect case selection of verb objects. In addition, spelling errors are flagged.

IceNLP outputs TCF containing information from the analysis, i.e. about the individual tokens and their PoS tags, individual constituents and error candidates. The platform forwards the TCF to the web server, which converts it to a human readable HTML format and displays a resulting page to the student, containing the original text submitted and the error candidates highlighted.

3. Lärka – an emerging ICALL platform for Swedish

Language technology research has a long history in Sweden, going back to the 1960s. Most of the basic NLP components exist for Swedish in quite stable and mature forms, e.g. PoS taggers and parsers, annotated reference corpora, and large lexical databases with morphological analysers. Swedish ICALL, however, has a shorter history. Only two projects – ITG [4] and Grim [5] – have resulted in concrete ICALL applications that are used in real-life teaching settings. However, they both use a technology which was state of the art at the time, but which practical experience shows is not the optimal solution today.

The application developed by the Swedish partner is web-based, and has the working title *Lärka* (http://spraakbanken.gu.se/larka/). Its principle is to have all functionalities web-service-based to ensure flexibility and reuse. The main components of Lärka's architecture are the following.

(1) Lärka's *frontend* is the graphical user interface that handles user interaction, sends requests to the backend and assigns behaviour to buttons and fields.

(2) Lärka's *backend* consists of a number of web services for generating language training exercises, selecting distractors, interacting with databases and generating syntactic trees. The backend depends heavily on Korp and Karp described below.

(3) *Korp* (http://spraakbanken.gu.se/korp/) is Språkbanken's web-service based infrastructure for maintaining and searching a constantly growing corpus collection, at the moment amounting to about one billion words of Swedish text [6]. The corpora available through Korp contain multiple annotations, e.g. lemmatisation, compound analysis, PoS tagging, and syntactic dependency trees, which can form the basis for versatile exercises.

(4) *Karp* (http://spraakbanken.gu.se/karp/) is the corresponding web-service based infrastructure for maintaining and retrieving information from Språkbanken's collection of computational lexical resources [7].

Korp and Karp together provide the necessary information for Lärka's learning activities. Once the sentence or lexical information is retrieved, the relevant algorithm is applied to generate an output for the exercise. The output from Lärka's backend can be used by any program, not only Lärka's frontend, for example in apps for mobile phones.

Each exercise (or any other learner activity) is added as a separate module consisting of backend and frontend parts. Exercises can thus be developed separately and be "plugged in" with minimal efforts into the architecture.

At the moment, Lärka offers three (multiple-choice) exercise types, all based on authentic corpus data: (1) identifying parts of speech; (2) identifying syntactic relations; and (3) vocabulary exercises. Each of the exercise types can be deployed as a test or as a self-study exercise. Next on our "to-do" list is to extend Lärka's exercise scope, add syntactic trees to every sentence, and add an "encyclopaedia" of basic linguistic terms.

4. Concluding remarks

The main idea of our project is to propagate the re-use of existing accurate NLP tools and resources in language learning by designing and implementing a system architecture for ICALL, at the moment on a more abstract level – where our two subprojects share the general philosophy of making NLP components available via web services – and in the next phase of the project on the concrete level of having a common data exchange format (e.g. TCF). It is thus clear that ICALL researchers and developers can be affected by our project. In addition, language learners will also be affected because the system architecture and the two test applications will benefit language learners in the form of a more versatile and open-ended CALL experience, thanks to the NLP components.

References

[1] Hinrichs, M, Zastrow, T., & Hinrichs, E. WebLicht: Web-based LRT services in a distributed eScience infrastructure. *Proceedings of LREC 2010*. Valletta, Malta: ELRA.

[2] Arnbjörnsdóttir, B. 2004. Teaching morphologically complex languages online: Theoretical questions and practical answers. In P. J. Hendrichsen (Ed.*) CALL for the Nordic Languages*. (Copenhagen Studies in Language 30.) Copenhagen: Samfundslitteratur.

[3] Loftsson, H., & Rögnvaldsson, E. 2007. IceNLP: A Natural Language Processing Toolkit for Icelandic. *Proceedings of InterSpeech 2007, Special Session: 'Speech and language technology for less-resourced languages'*. Antwerp, Belgium.

[4] Borin, L., & Saxena, A. 2004. Grammar, incorporated. In P. J. Hendrichsen (Ed.), *CALL for the Nordic languages* (pp. 125–145). (Copenhagen Studies in Language 30.) Copenhagen: Samfundslitteratur.

[5] Knutsson, O. 2005. *Developing and evaluating language tools for writers and learners of Swedish.* Doctoral thesis in human-computer interaction. KTH, Stockholm.

[6] Borin, L., Forsberg, M., & Roxendal, J. 2012. Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul, *Turkey: ELRA.*

[7] Borin, L., Forsberg, M, Olsson, L.-J., & Uppström, J. 2012. The open lexical infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul, *Turkey: ELRA.*