Automatic document enrichment for language learners

Hans Paulussen (*)
Francisco Bonachela Capdevila (*)
Pedro Debevere (**)
Maribel Montero Perez (*)
Martin Vanbrabant (*)
Wesley De Neve (**)
Stefan De Wannemacker (*)

(*) ITEC, KU Leuven, Kortrijk
(**) MMLab, Ugent, Gent

Thanks to the internet, foreign language (FL) learners nowadays have ubiquitous access to authentic text materials. On the other hand, direct access to extra information on unknown words often remains a clumsy undertaking. For example, if a learner wants to know the meaning of a new word, he has to open a separate window and decide on which lexicon or dictionary he wants to consult. Getting hold of extra information can thus become a time-consuming task. In order to facilitate access to such information in the context of CALL (computer assisted language learning) programs, we have developed, within the framework of the iRead+ project, an enrichment pipeline which enables to embed enrichment markup in reading texts. The resulting enriched documents can be used by developers of CALL applications. The aim of the iRead+ project is to enable automatic enrichment of texts with word-specific and contextual information in order to create an enhanced reading experience on tablet PC and to support automatic generation of grammatical exercises. In this talk, we present the architecture of the enrichment pipeline developed in the iRead+ project, and describe the proof of concept that is being developed related to FL learners, covering two applications of enriched documents: an enriched reading environment and an exercise generator.

The internet has changed our vision of the world considerably, and this has also had a positive impact on FL learning. Nowadays you can read any text in any language on the internet, listen to broadcasts sent from all over the world, and view video materials from your comfortable chair. In other words, authentic language samples can be easily accessed in a real-life cultural context. FL learners can benefit a lot from the overall availability of authentic input. Indeed, Google, Wikipedia, and other resources and databases of all sorts are directly accessible on the internet. However, notwithstanding the direct access to learning materials, we are in some way overwhelmed by the massive amount of information, so that it is not always that easy to get hold of the right information on a particular language problem or a cultural topic. At the moment, a learner eager to understand a new word in the text he is reading, will have to search for the word in a separate window, which can be cumbersome, especially when using a tablet: what starts off as a simple search can end up in a long diverting excursion on the internet. A better solution would be that a simple click on the word returns the required explanatory information.

Within the iRead+ project, we propose a possible solution which is based on a procedure that yields an enriched version of the text being read. The whole enrichment process is provided via a webservices workflow. The result is an enriched document (ENR), an XML document, containing both linguistic annotations (part-of-speech and lemmata) and semantic annotations based on the recognition and disambiguation of named entity references (NER). Named entities are words referring to persons, locations, and organisations. Application developers can use the enriched document to create, for example, language learning exercises or give extra information in documents presented to the reader via tablet computers or other devices.

The enrichment procedure requires two input files: the XML source file (SRC) and a control file (CTL) containing XPath expressions referring to the nodes in the source file that have to be annotated and enriched. Thanks to the combination of both source and control files, any type of XML document can be enriched, and sections not applicable for annotation (e.g. tables and figures) can simply be ignored. The enrichment pipeline consists of two steps: linguistic annotation and identification and disambiguation of named entities and multiword entities (MWE).

One of the challenges in the project consisted in improving the access to the databases used (e.g. DBpedia, Wiktionary, Cornetto), since each database has its own ontology structure. We opted for databases complying with the Resource Description Framework (RDF). In order to query the different databases in a similar way, a uniform query system has been developed. Although different queries are used to retrieve, for example, synonyms from Wiktionary or Cornetto, a uniform webservice overcomes the different ontology structures. In this way, the webservices implementation alleviates the need to know different ontologies.

At this moment, three proofs of concept are being developed and validated in the iRead+ project, each proof being related to a typical reading experience: the general reader, the language learner and the "struggling" reader. In the case of the language learner application, an enriched reading environment and an exercise generator is being developed. In both environments, a personal dictionary is used for the selection of appropriate text samples.

The enriched reading environment is an integrated resource which alleviates the burden of external resource consultations for the FL learner. The main advantage of this "enriched reading experience" resides in the fact that the annotation and the enrichments have been added automatically. The data enrichments make the text more understandable and therefore augment chances for language acquisition. The exercise generator automatically creates morphological and morphosyntactic exercises on the basis of linguistic annotation found in the enriched documents. This includes exercises on verb conjugation, the use of prepositions, gender and number of nouns and adjectives, etc.

The personal dictionary is considered a tool that helps selecting texts that match the language skills of the language learner. This data structure, which contains vocabulary that describes the language level of the user, is updated either proactively by the user himself every time that he finds an interesting or difficult word or by the iRead+ application when the learner makes a mistake in an exercise. This data structure evolves along with the skills of the learner, growing with new words and losing old ones once the learner is able to correctly use them in exercises. The algorithm that classifies and selects new texts for the user ranks the pool of texts based on the number and relevance of co-occurrences with the personal dictionary.