Supporting e-Learning with Language Technology for Portuguese

Mariana Avelãs, António Branco, Rosa Del Gaudio, and Pedro Martins

University of Lisbon
Faculdade de Ciências, Departamento de Informática
NLX - Natural Language and Speech Group
Campo Grande, 1749-016 Lisbon, Portugal
{mariana,antonio.branco,rosa,pmartins}@di.fc.ul.pt

Abstract. In this paper we report on the development and adaptation of language technology tools for Portuguese aimed at supporting e-Learning via the extension of a Learning Management System with new functionalities. We also describe how these tools were integrated into this Learning Management System and present results of both their intrinsic and extrinsic evaluation.

Keywords: Automatic keyword extraction, ontology, definition extraction, e-Learning, LMS, annotated corpus.

1 Introduction

The immense potential of Language Technology to enhance e-Learning has been repeatedly pointed out, and to a very large extent such potential remains to be explored. In this paper, we report on some first steps in that direction, discussing the application of some tools and resources for the computational processing of Portuguese with the aim of supporting e-Learning. More specifically, we report on the development and application of three tools aimed at enhancing learning activities in the scope of a Learning Management System (LMS).

One of the tools is a Keyword Extractor, which supports a new functionality with which the LMS was extended: once a text-based Learning Object is selected, a list of candidate keywords for that object can be automatically generated. This list can be subsequently filtered out by the users so that only the more relevant are retained and are persistently associated with that Learning Object. This functionality can be used by tutors in their task of meta-data annotation and thus helps to alleviate the burden of hand writing them, speeding up that process. It can also be used by students, who can obtain on the fly a list of some core concepts for a Learning Object they may have just imported into the LMS, and rapidly have a first glimpse of their content or relevance.

A second tool which was developed is a Definition Extractor, which supports another new functionality of the LMS, the Glossary Candidate Detector (GCD): from a given Learning Object selected, it is possible to generate a list of tentative

A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, pp. 192–201, 2008.

definitions that form a draft glossary; this glossary in turn can also be subsequently filtered out and extended by the users. Again, this functionality can be used by tutors in their task of meta-data annotation and helps to speed up that process. It can be used by students as well, who can obtain a draft overview of the concepts being defined in a Learning Object imported into the LMS.

Finally, the third tool developed was a Semantic and Multilingual Search Tool. A key component of this tool is an ontology and the annotation of the Learning Objects with their concepts: in the Learning Objects, each natural language expression conveying one of those concepts is associated to such concept via metadata annotation. Accordingly, the search tool developed permits to retrieve Learning Objects given the concept entered and its occurrence in the retrieved objects. Since the ontology is common for Learning Objects from different idioms, the set of retrieved objects can include also those not written in the language of the user, thus supporting cross-language search.

The results reported in the present paper were obtained in the scope of the LT4eL project activities. This is an FP6 European project whose goals were pursued with the cooperative contribution of 12 partners, including our team, under the coordination of the University of Utrecht. In the present paper, we focus in the tools and results contributed by our team.

In Section 2, we describe the corpus collected and annotated in order to support the development and the intrinsic evaluation of the tools. In Section 3, a detailed presentation of the keyword extractor is offered, while the glossary candidate extractor is discussed in Section 4. In Section 5, we present the ontology and the semantic search mechanisms it supports. In Section 6, we briefly describe how these tools were integrated in the LMS and what was the outcome of their extrinsic evaluation. Finally, Section 7 is devoted to conclusions.

2 The Corpus

In order to support the development and the intrinsic evaluation of the tools a corpus was developed. Given that the corpus was to be used also for the extrinsic evaluation, viz. as a repository of learning material in the LMS, we selected documents that can be taken as Learning Objects.

A Learning Object (LO) is any small, reusable chunk of instructional media, digital or non-digital, which can be used, re-used or referenced during technology supported learning and should be enriched with metadata (the actual standard is the "Learning Object Metadata") [1]. Keeping this in mind, we selected 31 documents, mostly of a tutorial nature, apt to be used as LOs, covering three domain areas, namely Information Technology (IT) for non experts, e-Learning (eL) and Information Society (IS). Table 1 shows the composition of the corpus.

The XML-based format version of the corpus went through a process of linguistic annotation. The corpus was automatically annotated with morpho-syntactic information using LX-Suite, a set of tools for the shallow processing of Portuguese with state of the art performance [2].

This annotation includes information about sentence and tokens boundaries, POS categories, and inflectional features and lemmas.

Finally, in the last step the output of the annotation tools was converted into a common, project internal, XML format, the LT4eLAna document format. The DTD of this format conforms to a DTD derived from the XCESAna DTD, a standard for linguistically annotated corpora [3]. This DTD structures the documents into paragraphs, sentences, chunks and tokens. The textual content of tokens is the actual text of the document while the attributes associated to the tokens encode linguistic and layout information. Markup for some other elements was yet added, namely for keywords, defined terms and defining text.

Over this version of the corpus in this final format, a phase of manual annotation of keywords and of definitions was carried out.

Concerning keywords, 29 documents were annotated (corresponding to 265 915 tokens) with 1 033 different types, which means a mean of 35.6 types per document.

Definitions were marked with the indication of the definiens and of the definiendum. Information regarding the type of definitions was also encoded, namely distinguishing four different kinds of definitions: definition introduced by the verb "to be", termed copula definitions; definitions introduced by other verbs; definitions introduced by a punctuation mark; and definitions of none of these previous three types. Table 2 displays the distribution of the different types of definitions in the corpus, and their breakdown by sub-corpora.

Table 1. Corpus domain composition

Domain	tokens
IS	92825
IT	90688
eL	91225
Total	274000

Table 2. The distribution of definitions

Type	IS	IT	eL	Total
Copula	80	62	24	166
OtherVerb	85	93	92	270
Punctuation	4	84	18	106
other	30	54	23	107
total	199	295	157	651

3 The Keyword Extractor

Keywords are (single or multi-word) terms that are presented to very briefly characterize a text and resume what it is about. In order to extract such terms automatically, a few algorithms, based on distributional statistics, were tested. In particular, project internal work provided an implementation of algorithms based on TF*IDF, RDIF and a term frequency adjusted version of IDF (ARDIF). Such tool, developed by Lemnitzer and Dergorski [4], took into account the linguistic information encoded in the corpora, in particular the base form of each word, the part of speech, and the morpho-syntactic features. These tools try to pay justice to the fact that good keywords have a typical, non random distribution in and across documents and that keywords tend to appear more often at certain places in texts (e.g. headings, etc.).[1]

[1] For full details, the reader is referred to [4].

These tools ran over the corpus described in the previous Section, and its outcome underwent a subsequent process of refinement. When looking at their results, it was apparent that some terms selected as candidate keywords were not apt to be considered keywords at all and could be very easily discarded. For instance, focusing on single-word keywords, this was the case of candidates made of punctuation marks or of a single preposition. Or, when taking multi-word expressions, for instance, that was the case of candidates starting with punctuation marks or prepositions.

In order to automatically refine such preliminary outcome, a system of pattern-based filters was developed. That filtering module is based on the use of four portmanteau tags that are in correspondence with the elements of the POS tagset used for the annotation of the corpus:

PLU - punctuation elements, that should be ignored completely.
FLU - lexical units that are not possible as single-word, though they can appear inside multi-word units but not at the initial or final position.
CMLU - lexical units which are admissible in multi-word lexical units, even at their beginning or end, but cannot form a single-word keyword.
MLU - admissible both as single-word keyword and as of a multi-word one.

Intrinsic evaluation was carried out at the output of this filtering of the first tentative results provided by the statistics-based tools. Scores for Precision, Recall and F-measure were obtained against the manually annotated documents reserved as test set. Table 3 displays the results for each base technique tried out, showing a slight advantage for the combination TFIDF-based algorithm followed by rule-based filtering.

Table 3. Keyword Extractor intrinsic results

	ADRIDF			RIDF			TFIDF		
	R	P	F	R	P	F	R	P	F
filtered	0.30	0.17	0.21	0.21	0.12	0.15	0.31	0.18	0.22

Given that the manual annotation of keywords was performed by a single annotator, in order to have a more reliable notion of the intrinsic performance of the tool an experiment was carried out to obtain a score for inter-annotator agreement on this specific task of keyword assignment.

Ten individual testers were given one LO from the corpus and were asked to extract the 10 keywords that should be assigned to that document. The agreement between testers was assessed by using the AC1 measure proposed in [5]. It scored 0.58 which indicates that the task is inherently quite difficult (even for humans).

Note that the scores displayed in the table above were obtained by comparison with the list proposed by a single annotator. Accordingly, a much more significant measure of the performance of the tool is to be collected with the AC1 score obtained for the comparison between the tool and the ten annotators.

The agreement between human testers and the tool scored 0.67. The list of keywords proposed by the "typical" tester (taken as the 10 most selected keywords by all the testers) is thus in agreement with the system more than the testers agree among each other. This is clearly an indicator of a very fgood performance of the system given the inherent difficulty of the task.

Finally, further pursuing the intrinsic evaluation of the keyword extractor, additional scores for the performance of this tool were obtained yet from another perspective. The first 20 keywords automatically extracted from a document were presented to 10 human testers. These testers were then asked to rate the keywords in a scale from 1 to 4 (very relevant, quite relevant, not relevant to the document, not a valid term). The average score was calculated over the entire set of 20 keywords, over the first 10 and the over the first 5. For the entire list of 20 keywords, a score of 2.34 was obtained; 2.08 was the score for the first 10 candidate keywords; and finally 1.94 was the score obtained for the first 5. These results are quite satisfactory: they indicate that the keywords automatically extracted are correctly ranked by the tool (with the more relevant being presented in the first positions) and that those higher ranked tend to be quite relevant.

4 The Glossary Candidate Detector

The Glossary Candidate Detector (GCD) was designed to automatically detect definitions, being able to tell apart the definiens from the definendum. A rule-based approach was adopted to develop this tool. The rules encode general patterns of candidate definitions whose basic components are some reserved words (e.g. verb "to be", etc.) and POS categories. The patterns were hand crafted on the basis of the analysis of the development data previously created, under the form of a corpus annotated with definitions.

To write down such rules, we resorted to lxtransduce. This is a tool that allows to build transducers specially suited to add or rewrite XML markup. It is a component of the LTXML2 tool set developed at the University of Edinburgh.[2]

In order to develop such transducer, three types of definition were identified and for each one a specific set of rules was written (for more details see [6]). Furthermore, the 274 000 token corpus was split in two parts, a development set, with 75% of the corpus, the remaining 25% for the test data.

Similarly to what was done for the keyword extractor, the GCD was evaluated both in a quantitative and in a qualitative manner. For the quantitative evaluation, the value of recall and precision was calculated at the sentence level. Recall here is the proportion of the sentences correctly classified by the system as containing a definition with respect to the sentences manually annotated as actually containing a definition. Precision is the proportion of the sentences correctly classified by the system with respect to the sentences automatically annotated. Furthermore, the F_2-measure[3] was also calculated. This score was

[2] http://www.ltg.ed.ac.uk/~richard/ltxml2/lxtransduce-manual.html

[3] $F_2 = \frac{(1+2)*Precision*Recall}{(2*Precisio)+Recall}$.

preferred to the simple F-measure in virtue of the type of task at stake. We are more interested in higher recall than in higher precision, given the application of the tool which is better to give more (possibly incorrect) definition candidates (with a higher recall, at the expense of a lower score in Precision) than to miss good definitions (in the opposite situation). We obtained a score of 0.14 for Precision, 0.86 for Recall and 0.33 for the F_2-measure.

On a par with this quantitative evaluation, a qualitative evaluation was carried out involving a group of users. We selected six MA students and presented them a LO with a list of definitions automatically generated using the tool—the LO was a 12 page introductory document on the use of Internet and the GCD had extracted 34 different definitions. Testers were instructed to read the document carefully and then score each definition using a rating scale from 1 to 4 (very good definition, good definition but not complete, acceptable definition, not a definition at all). The average score was 2.21, thus indicating that the candidate passages automatically extracted are on average considered good definitions according to human appreciation.

5 Semantic Search Tool

The Semantic and Multilingual Search Tool aims at allowing semantic search within a collection of documents; in more concrete terms in view of the application at stake, within a set of LOs. This means that it is possible to search for a term and retrieve, for example, all documents containing not only that term but also its synonyms and related concepts (such as super and sub-concepts). Since the tool is based on aligned ontologies developed for different languages,[4] it is possible to search for a term from a language A and retrieve documents in languages other than A, allowing for a multilingual retrieval.

The Semantic Search tool builds on the Lucene retrieval engine [7] embedded in the LMS and is based on three resources: a domain ontology, a lexicon, and an annotation grammar.

The ontology resulted from the merge between the DOLCE top-ontology, intermediate concepts from OntoWordnet, and a domain-specific ontology developed from scratch. This latter part was built in a bottom-up manner using as starting point the collection of keywords automatically generated for the corpora of every language in the project. This collection was translated into English in order to end up with a common collection. This final collection offered a first list of concepts to be covered by the domain ontology. Additionally, when entering these concepts in the ontology, concepts important to establish intermediate levels with the upper ontologies were added.[5] The domain covered by the final ontology is the realm of Computer Science for Non-Computer Scientists and includes concepts related to operating systems, applications, document preparation, computer networks, markup languages, world wide web, etc. The ontology

[4] The languages concerned are: Bulgarian, Czech, Dutch, English, German, Polish, Portuguese, Romenian and Maltese.

[5] For a fully detailed account of the core ontology building process see [8].

includes about 950 domain concepts, including 50 concepts from DOLCE and about 250 intermediate concepts from OntoWordNet.

The lexicon, in turn, was built by collecting every possible lexicalization for each one of the concepts in the ontology. By the end of the lexicon development process, we ended up with a list with 917 entries and 1019 lexicalizations, where each concept is associated with its possible lexicalizations.

Finally, an annotation grammar was developed which has the lexicon as its central component. A first common template of this grammar was put forward by Simov and Osenova [9] and subsequently worked out to develop different grammars for every language, and in particular for Portuguese by us. When applied to an input LO, this grammar detects possible (single or multi-word) lexical units and suggests all concepts of the ontology possibly expressed by that lexical unit. In the process of annotating the corpus with concepts, the output of this grammar is validated by human annotators who can select the right concept, or reject all and/or suggest a new one.

The lexicon constitutes the main relay resource between the query entered to start a search for documents, the ontology and, consequently the semantic-based search. The words entered are looked up in the lexicon, and the concepts that are associated to them are the items actually used in the search process, which will retrieve those documents containing some occurrences of those concepts in the markup semantic layer underlying the raw text.

Given the nature of the semantic search functionality, it was submitted only to an extrinsic evaluation, as described in the next Section.

6 Integration in the LMS and Extrinsic Evaluation

Besides the intrinsic evaluation of the tools developed when applicable, an extrinsic evaluation was also carried out after the integration of the new functionalities supported by them in the LMS. The LMS used was ILIAS, an open-source, fully edged web-based LMS that allows users to create, edit and publish learning material in an integrated system with normal web browsers.[6]

Fig. 1 displays the Graphical User Interface (GUI) by means of which it is possible to invoke the keyword extractor over a certain LO, and automatically obtain a list of candidate keywords for that document. The pane shows the candidate keywords list proposed after pressing the "Generate KeyWords" button. The user can accept the proposed keywords by checking the boxes and can also add new ones in the text field below them.

Fig. 2, in turn, presents a sample of the outcome of calling the GCD over a given LO.

Finally, Figure 3 shows the results of a semantic search triggered by the query made of the word "editor". It is worth noting that when a semantic search is performed, besides the relevant documents, the fragment of the ontology surrounding the concept used in the search is also displayed in the panel right

[6] http://www.ilias.de/

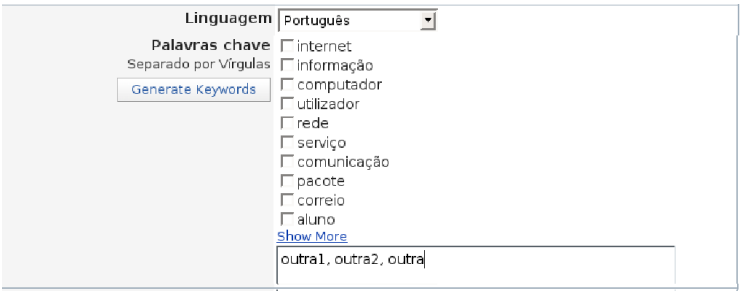

Fig. 1. User Interface ILIAS - Keyword Generator

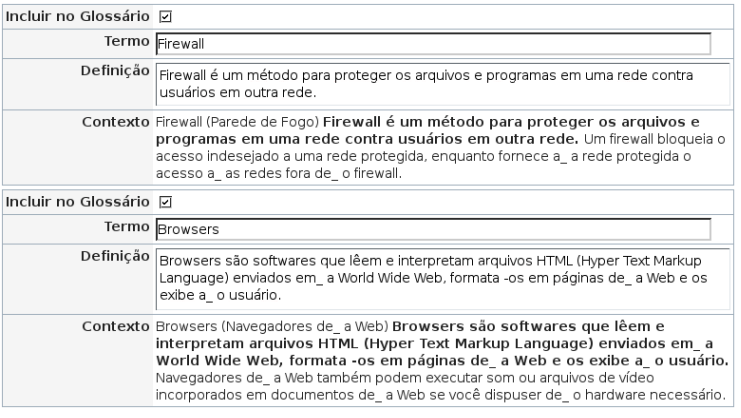

Fig. 2. User Interface ILIAS - Glossary Candidate Detector

inferior corner. The nodes of this fragment are clickable and allow the launching of a new search of LOs with occurrences of the concept clicked on.

The extrinsic evaluation was designed seeking to get some insight on the satisfaction of the potential end-users with respect to the new functionalities. This evaluation was based on the user scenario methodology [10]. Scenario here is meant to be "a story focused on a user, which provides information on the nature of the user, the goals he wishes to achieve and the context in which the activities will take place".

There were scenarios developed for two roles, i.e. for two kinds of users, students and tutors. For each role, two scenarios were created, one aimed at assessing the Keyword Extractor and GCD, and another aiming at assessing the semantic search. A group of at least 6 students participated in the student scenarios and the tutor scenarios were performed by 3 university teachers.

Regarding the extraction of keywords and the use of GCD by tutors, the participants were requested to generate a list of keywords and a glossary using the tools in order to make a certain LO available for a particular course. All testers (100% of score) agreed that both tools are useful, in particular for people

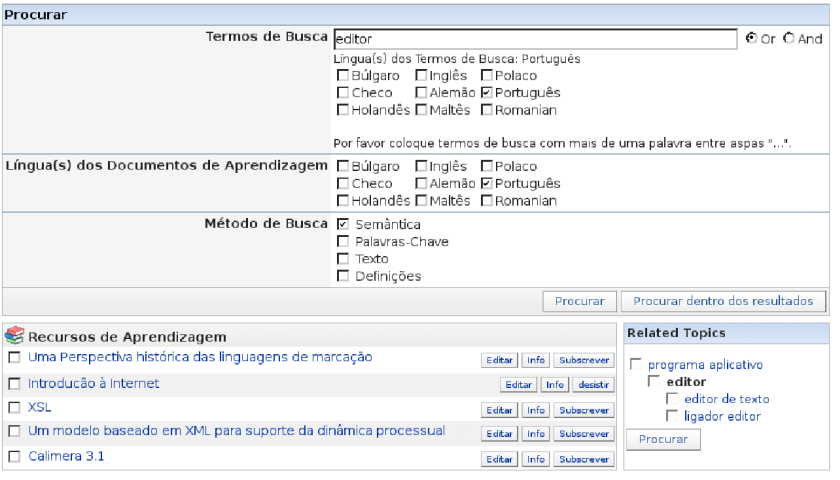

Fig. 3. User Interface ILIAS - Search

responsible for adding metadata to content. Although 30% of the testers think that the tools could be improved, they would use them if available.

As for the students, they received the task of summarizing a scientific paper. The participants were split into two subgroups. A target group with access to the new functionality of automatic generation of keywords and definitions, and a control group with no access to these extensions of the LMS. With respect to satisfaction, 67% of the testers were very satisfied with the list of keyword and 80% would use this tool for selecting a document in a collection. Nevertheless, 50% think that some important terms are missing. Regarding the glossary, all testers agreed that definitions were of a good quality, even if some definitions were missing. All testers agreed on the usefulness of this tool for this particular task and they would use the tool for extracting definitions from other papers. Besides checking satisfaction of users, the abstracts developed by the two groups were also evaluated using as metric the number of relevant concepts covered by abstracts. It turned out that the abstracts produced by the target group had a best coverage than the abstracts of the control group. On average, abstracts produced by the target group mentioned 5.5 relevant concepts while abstracts produced by the control group mentioned 4.2.

Regarding the semantic search functionality, tutors were given the task of refining a list of prerequisites for a given course, and to identify those LOs in the LMS repository which would help a student to learn about those prerequisites. Although for all testers it was easy to locate the relevant topics and identify relevant documents, 50% of them were not able to find some topics that they thought should be present. All testers agreed on the advantages of using such a tool in a virtual learning environment.

Students, in turn, were provided with a quiz with multiple choice questions, and were asked to try to find the documents containing the relevant answers. 83% of the testers found that their search terms returned mostly relevant content

and 67% reported that the use of an ontology helped them in completing the task; 83% pointed out that ontology browsing and semantic search permit linking concepts in a way they were not aware of before.

7 Conclusions

In this paper we presented language technology resources and tools developed with the purpose of enhancing e-Learning by supporting new language processing-based functionalities embedded in an LMS. These tools were assessed under intrinsic and extrinsic evaluation.

Overall, the results coming out of the evaluation and reported above are positive and very encouraging. They provide an objective ground to the repeated claim that there is an important potential to be explored in what concerns the application of language technology to enhancing e-Learning.

References

1. LTSC: Learning technology standards committee website, `http://ltsc.ieee.org/`
2. Silva, J.R.: Shallow processing of Portuguese: From sentence chunking to nominal lemmatization. Master's thesis, Universidade de Lisboa, Faculdade de Ciências (2007)
3. N., I., K., S.: Xml, corpus encoding standard, document Xces 0.2. Technical report, Department of Computer Science, Vassar College and Equipe Langue et Dialogue, New York, USA and LORIA/CNRS, Vandouvre-les-Nancy, France (2002)
4. Lemnitzer, L., Lukasz, D.: Language technology for elearning: Implementing a keyword extractor. In: Fourth EDEN Research Workshop Research into online distance education and eLearning. Making the Difference, 2006 in Castelldefels, Spain (2006)
5. Gwet, K.: How to estimate the level of agreement between two or multiple raters. In: Handbook of Inter-Rater Reliability, Gaithersburg, Maryland (2001)
6. Gaudio, R.D., Branco, A.: Learning to identify definitions using syntactic feature. In: Progress in Artificial Intelligence, 13th Portuguese Conference on Aritficial Intelligence, Guimarẽs, Portugal. Springer, Berlin (2007)
7. Gospodnetic, O., Hatcher, E.: Lucene in Action. Manning Publications (2004)
8. Osenova, P., Simov, K., Mossel, E.: Language resources for semantic document annotation and crosslingual retrieval. In: LREC 2008 (2008)
9. Simov, K., Osenova, P.: A system for a semi-automatic ontology annotation. In: Proceedings of Workshop on Computer-aided language processing CALP
10. Carrol, J.M.: Scenario-based design. John Wiley and Sons, Inc., Chichester (1995)