Supporting e-learning with automatic glossary extraction:
Experiments with Portuguese

Rosa Del Gaudio
University of Lisboa
rosa@di.fc.ul.pt

António Branco
University of Lisboa
antonio.branco@di.fc.ul.pt

Abstract

This paper reports a preliminary work on automatic glossary extraction for e-learning purpose. Glossaries are an important resource for learners, in fact they not only facilitate access to learning documents but also represent an important learning resource by themselves. The work presented here was carried out within the project LT4eL which aim is to improve e-Learning experience by the means of natural language and semantic techniques. This work will focus on a system that automatically extract glossary from learning objects, in particular the system extract definitions from morpho-syntactic annotated documents using a rule-based grammar. In order to develop such a system a corpus composed by a collection of Learning Object covering three different domain was collected and annotated. A quantitative evaluation was carried out comparing the definition retrieved by the system against the definitions manually marked, On average, we obtain 14% for precision, 86% for recall and 0.33 for F_2 score.

Keywords

automatic definition extraction, glossary candidate detector, Portuguese, e-learning management systems

1 Introduction

The main focus of this work is supply Learning Management Systems (LMS) with a tool allowing an easy and quick glossary building. The definition extraction system presented here was developed with the practical objective of supporting the functioning of the module of Glossary Candidate Detector (GCD), in a Learning Management System (LMS).

The research underlying the extraction system presented here was carried out within the LT4eL project [1] funded by European Union (FP6). The main goal of this project is to improve LMSs by using language technology, in order to make more effective the retrieval and the management of learning materials and information.

In particular, the ILIAS[2] Learning Management Systems is being extended with new functionalities to support the different actors in e-learning environments. ILIAS is a fully fledged web-based learning management system that allows users to create, edit and publish learning and teaching material in an integrated system with their normal web browsers. At present, it is being extended with a automatic keyword extractor [8] and a GCD.

This GCD module permits the extraction of candidates to definitory contexts from learning objects. Such a module is thus meant to help content providers and teachers to speed up the process of building glossaries corresponding to the learning objects they are making available via the LMS to their users and students.

In particular the final user, the content provider or the teacher, will take advantage, in terms of time saving, in generating the glossary using the tool; instead of scrolling the entire document looking for terms and definition the system will do it for him presenting the result in a page (see Figure 1), and then the user can accept or discard each definition as well as modify it.

Beside the advantage in term of time saving this approach allows the construction of specific glossary for each learning object, and this positively influence the learning process in two different way. First, these glossaries can be used as a quick index to the information contained in the original document, second, the learned will have access not to a general definition of a concept but to the specific acception that the concept takes in that particular context.

In this paper we will focus on the module of the system that allow the extraction of definition from document written in Portuguese. Modules for Bulgarian, Czech, Dutch, English, German, Polish and Romanian are being developed by other partners. We present the methodology used to develop the system and its performance by comparing the results of the system against a test data made of texts belonging to the domains of computer science, information society and e-learning.

In this work, a *definition* (also called definitory context) is assumed to be a sentence containing an expression (the *definiendum*) and its definition (the *definiens*) and a connector between them. We identify three different connector: the verb "to be", all other verbs other than "to be" and punctuation mark such as ":". Here, we will be calling copula definition all those definitions where the verb "to be" acts as a connector, verb definition to all those definitions

---

that are introduced by a verb other than "to be", and punctuation definitions to the ones introduced by punctuation marks.

In the next Section we present a brief review of researches focused on automatic glossary building and definition extraction, with some references to question answering systems.

In Section 3 we present the corpus collected in order to develop and test our system. The corpus is composed by learning objects in three different domains. Part of the corpus, referred to in the remainder of this paper as the development corpus, was retained to identify the lexical and syntactic patterns to be taken into account by the system. A grammar for each definition type was developed, which are described in Section 4.

In Section 5, the results of the evaluation of the grammar, in terms of recall, precision and F2-score, are presented and discussed.

In Section 6, we provide an analysis of errors and discuss possible alternative methods to evaluate our system.

Finally in Section 7 conclusions are presented as well as possible ways to improve the system in future work.

2 Previous Work

Glossary building is often considered as an extension of term extraction, many systems start with the identification of relevant term in a certain domain [7] and then try to apply different techniques in order to build glossary for that specific domain. For instance [1] present a methodology in order to construct a glossary, supporting knowledge dissemination, in a collaborative research project, where the semantic unification represents a critic factor for the success of the project. Different resources are used in order to construct the glossary: a term extractor based on statistical techniques, corpora and other glossaries where definition are extracted using a context free grammar. Then non relevant definitions are filtered out using information about the domain covered by the glossary. The use of information about the domain is also used in [7] in order to improve results. In our work

Fig. 1: *An example of the outcome of the system*

we don't have such previous domain knowledge, the system is supposed to work with documents belonging to different domains.

DEFINDER [6] is based on a methodology that use just lexical and syntactical information. This is an automatic definition extraction system targeting the medical domain, where the data is composed of consumer-oriented medical articles. In terms of quantitative evaluation, this system presents 87% precision and 75% recall. This very high values are probably due to the nature of the corpus.

Turning more specifically to the Portuguese language, there is only one publication in this area. Pinto and Oliveira [11] present a study on the extraction of definitions with a corpus from a medical domain. They first extract the relevant terms and then extract definition for each term. An evaluation is carried out for each term; for each term recall and precision are very variable ranging between 0% and 100%.

By using the same methodology for Dutch as the one used here, Westbund and Monachesi [15] obtained 0.26 of precision and 0.72 of recall, for copula definitions, and 0.44 of precision and a 0.56 of recall, for other verbs definition.

Hearst [5] proposed a method to identify a set of lexico-syntactic patterns to extract hyponym relations from large corpora and extend WordNet with them. This method was extended in recent years to cover other types of relations[10].

In particular, Malaise and colleagues [9] developed a system for the extraction of definitory expressions containing hyperonym and synonym relations from French corpora. They used a training corpus with documents from the domain of anthropology and a test corpus from the domain of dietetics. The evaluation of the system using a corpus of a different domain, makes results more interesting as this put the system under more stressing performance. Nevertheless, it is not clear what is the nature and purpose of the documents making this corpora, namely if they are consumer-oriented, technical, scientific papers, etc. These authors used lexical-syntactic markers and patterns to detect at the same time definitions and relations. For the two different, hyponym and synonym, relations, they obtained, respectively, 4% and 36% of recall, and 61% and 66% of precision.

Answering questions asking for a definition is a particular dimension of the broad task of question answering that is very related to the main focus of this paper. The objective is that given an expression, a definition for it should be retrieved in a corpus or in the entire web. The main difference between this task and our work resides in the fact that we do not know beforehand the expressions that should receive definitions. This lack of information makes the task more difficult because it not possible to use the term as a clue for extracting its definitions.

Saggion [13] presents results of the TREC QA 2003 competition, where he tested his QA system against 50 definition questions. He just reports F5 score, where the recall is 5 times more important precision. His system obtained a F-score of 0.236, where the best score in the same competition was of 0.555 and the median was of 0.192.

```
−<s id="s183">
    <tok base="o" class="word" ctag="DA" id="t3698" msd="ms" sp="y">O</tok>
    <tok base="cabo" class="word" ctag="CN" id="t3699" msd="ms" sp="y">cabo</tok>
    <tok base="ser" class="word" ctag="V" id="t3700" msd="pi-3s" sp="y">é</tok>
    <tok base="a" class="word" ctag="DA" id="t3701" msd="fs" sp="y">a</tok>
    <tok base="parte" class="word" ctag="CN" id="t3702" msd="fs" sp="y">parte</tok>
    <tok base="mais" class="word" ctag="ADV" id="t3703" sp="y">mais</tok>
    <tok base="básico" class="word" ctag="ADJ" id="t3704" msd="fs" sp="y">básica</tok>
    <tok base="de" class="word" ctag="PREP" id="t3705" sp="y">de</tok>
    <tok base="uma" class="word" ctag="UM" id="t3706" msd="fs" sp="y">uma</tok>
    <tok base="rede" class="word" ctag="CN" id="t3707" msd="fs">rede</tok>
    <tok class="punctuation" ctag="PNT" id="t3708" sp="y">.</tok>
</s>
```

Fig. 2: *The sentence* `O cabo é a parte mais básica duma rede.` *(The cable is the most basic component of a network) in final XML format*

3 The Data Set

In order to develop and test our grammars, a corpus of around 270 000 tokens was collected. The corpus is composed of 33 different documents, which can be integrated in the an LMS as learning objects. These documents are mainly tutorials, PhD thesis, and research papers. They cover three different domains: Information technology for non experts, e-learning, and information society. This last part is composed by the Section 3 of Calimera guidelines. These guidelines have been compiled by the CALIMERA[3] Co-ordination Action, funded by the European Commission's, with the goal of explaining in an accessible way how technologies can be deployed to support digital services designed to meet real user needs.

These three domains are evenly represented in the corpus. One third is composed of documents that are mainly tutorials focusing on basic notions and tools in the domain of computer technology (tutorials on using text editors, HTML, Internet, etc.). The second third of the corpus is composed of documents (mainly articles and PhD thesis) on e-learning concepts, experiments, and governmental policies. The last third is the Section 3 of the Calimera guidelines.

Table 1 shows the composition of the corpus. All the documents were originally in several different file formats (.pdf, .html, etc.). They were processed in order to be converted into a common XML format, conforming to a DTD derived from the XCES DTD for linguistically annotated corpora.

The corpus was then automatically annotated with morpho-syntactic information using the LX-Suite [14]. This is a set of tools for the shallow processing of

---

[3] http://www.calimera.org

Domain	tokens
IS	92825
IT	90688
e-Learning	91225
Total	274000

Table 1: *Corpus domain composition (IS: Information Society; IT: Information Technology)*

Type	IS	IT	e-Learning	Total
is_def	80	62	24	166
verb_def	85	93	92	270
punct_def	4	84	18	106
other_def	30	54	23	107
total	199	295	157	651

Table 2: *The distribution of types of definitions in the corpus*

Portuguese with state of the art performance. This pipeline of modules comprises several tools, namely a sentence chunker (99.94% F-score), a tokenizer (99.72%), a POS tagger (98.52%), and nominal and verbal featurizers (99.18%) and lemmatizers (98.73%).

In Figure 2, we present a sample of the final result. Each sentence is delimited by tags `s`, and each token by the tag `tok`. Of particular interest for the development of our grammars are the attribute `base`, containing the lemma of each word, the attribute `ctag`, containing the POS information, and the `msd` with the morpho-syntactic information on inflection.

Subsequently, the definitory contexts in the corpus were manually annotated. This involves the explicit mark up of the sentence containing the definition (`definingText` tag), of the *definiens* (`markedTerm` tag), and also of the type of definition (`defType1` attribute). Figure 3 shows an example of the mark-up of a definitory context.

Besides the three definition types referred to in the beginning of this paper, a fourth category was also introduced for those that are not captured under any of the other three types. Accordingly, the definition typology is made of four different classes whose members were tagged with `is_def`, for copula definitions, `verb_def`, for verbal non copula definition, `punct_def`, for definition whose connector is a punctuation mark, and finally `other_def`, for all the remaining definitions. Table 2 displays the distribution of the different types of definitions in the corpus.

The domains of Information Society and Information Technology present a higher number of definitions, in particular of copula definitions. This is due to the fact that most documents belonging to these domains were conceived to serve as tutorials for

non experts, and have thus a more didactic style. The part of the corpus concerning the e-learning domain is mostly composed by research papers and PhD thesis, where the goal is less didactic. In Section 5, we will see how the difference in the objectives of the documents, irrespective of the domain they belong to, may affect the performance of the system.

4   The Rule-Based System

In order to take advantage of the XML format of the corpus, a regular expression based tool for pattern matching that was XML aware was convenient. The tool opted for was the LXtransduce[4]. It is a transducer which adds or rewrites XML markup on the basis of the rules provided. Lxtransduce is an updated version of fsgmatch, the core program of LT TTT[4].

LxTransduce allows the development of grammars containing a set of rules, each of which may match part of the input. In case of successful match of the rule with some part of the input, it is possible to replace the matched text or wrap it with a xml tag. Rules may contain simple regular-expression, or they may contain references to other rules in sequences or in disjunctions, hence making it possible to write complex procedures on the basis of simple rules. The outcome of a rule may be used to instantiate variables that can be used later on by other rules. A grammar is thus composed by several rules, where a main rule calls the remaining ones.

All the grammars we developed present a similar structures, they start with simple rules for matching basic expressions, such as conjunctions, articles or nouns. Then further rules may proceed on the basis of the outcome of those rules aiming at matching for complex noun phrases, for instance. As expected, special focus is given to verbs and syntactic patterns surrounding verbs and possibly other connectors in definitory contexts.

A development corpus, consisting of 75% of the whole 270 000 token corpus, was inspected in order to obtain generalizations helping to concisely delimit lexical and syntactic patterns entering in definitory contexts. This sub-corpus was used also for testing the successive development versions of each grammar.

The held out 25% of the corpus was thus reserved for testing the system and to obtain the evaluation results reported below, and was not used in the development phase.

At the present stage of development, three grammars were developed, one for each one of the three major types of definitions, namely copula, other verbs, and punctuation definitions.

Copula definitions

In order to develop a grammar for definitions based on the copula verb "to be", the information contained in the developing corpus was exploited. The copula definitions manually marked in the developing corpus were gathered. Then all information was removed except for the information on part-of-speech, in order

to abstract possible useful syntactic patterns for definitory contexts organized around this type of connector. Every such pattern was listed, one per line, and sorted. This permitted to construct a list of different syntactic patterns and associate them with corresponding frequency. Patterns occurring more than three times in the development corpus were implemented in this sub-grammar. The following rules are a sample of the rules in copula sub-grammar.

```
<rule name="Serdef"> <!-- To Be 3rd person pl and s -->
<query
match="tok[@ctag = 'V' and @base='ser' and
(@msd[starts-with(.,'fi-3' )]or @msd[starts-with(.,'pi-3' )])]
</rule>

....

<rule name="copula1">
<seq>
<ref name="SERdef"/>
<best>
<seq>
<ref name="Art"/>
<ref name="adj|adv|prep|" mult="*"/>
<ref name="Noun" mult="+"/>
</seq>
<ref name="tok" mult="*"/>
<end/>
</seq>
</rule>
```

The second rule is a complex rule composed by other that use other rules, defined in the grammar. This rule matches a sequence composed by the verb "to be" followed by an article and one or more nouns.Between the article and the noun an adjective or an adverb or a preposition can occur. The rule named SERdef matches the verb "to be" only if it occur in the third person singular and plural of the present and future past.

Verbs definitions

To extract definitions whose connector is a verb other than "to be", first all such verbs appearing in the developing corpus were collected, to which some synonyms were also added. We exploited the possibility provided by Lxtransduce to have a separate lexicon where the verbs, were listed with some information about the use of the verb in definitory contexts. For example, it is possible to specify whether a particular verb occurs in a definition only in passive form, or only in reflexive form, etc.

We decide to exclude some verbs initially collected from the final list because their occurrence in the corpus were very high, but their occurrence in definitions were very low. Their introduction in the final list would not improve recall and would have a detrimental effect on the precision score. The following rule is a sample of how verbs are listed in the lexicon.

```
<lex word="significar">
<cat>act</cat>
</lex>
```

In this example the verb *significar* ("to mean") is listed, in his infinitive form that correspond to

```xml
<s id="s259">
  <definingText continue="y" def="m146" def_type1="is_def" id="d32">
    <tok base="um" class="word" ctag="UM" id="t4804" msd="ms" sp="y">Um</tok>
    <markedTerm id="m147" kw="y">
      <tok base="firewall" class="word" ctag="CN" id="t4805" msd="ms" sp="y">firewall</tok>
    </markedTerm>
    <tok base="ser" class="word" ctag="V" id="t4806" msd="pi-3s" sp="y">é</tok>
    <tok base="um" class="word" ctag="UM" id="t4807" msd="ms" sp="y">um</tok>
    <tok base="conjunto" class="word" ctag="CN" id="t4808" msd="ms" sp="y">conjunto</tok>
    <tok base="constituir,constituído" class="word" ctag="PPA" id="t4809" msd="ms" sp="y">constituído</tok>
    <tok base="por" class="word" ctag="PREP" id="t4810" sp="y">por</tok>

    <tok base="hardware" class="word" ctag="CN" id="t4811" msd="ms" sp="y">hardware</tok>

    <tok base="e" class="word" ctag="CJ" id="t4812" sp="y">e</tok>
    <tok base="por" class="word" ctag="PREP" id="t4813" sp="y">por</tok>

    <tok base="software" class="word" ctag="CN" id="t4814" msd="ms" sp="y">software</tok>

    <tok base="que" class="word" ctag="CJ" id="t4815" sp="y">que</tok>
    <tok base="ter" class="word" ctag="V" id="t4816" msd="pi-3s" sp="y">tem</tok>
    <tok base="como" class="word" ctag="CJ" id="t4817" sp="y">como</tok>
    <tok base="função" class="word" ctag="CN" id="t4818" msd="fs" sp="y">função</tok>
    <tok base="controlar" class="word" ctag="V" id="t4819" msd="inf-nInf" sp="y">controlar</tok>
    <tok base="o" class="word" ctag="DA" id="t4820" msd="ms" sp="y">o</tok>
    <tok base="acesso" class="word" ctag="CN" id="t4821" msd="ms" sp="y">acesso</tok>
    <tok base="a" class="word" ctag="DA" id="t4822" msd="fs" sp="y">a</tok>
    <tok base="uma" class="word" ctag="UM" id="t4823" msd="fs" sp="y">uma</tok>
    <tok base="rede" class="word" ctag="CN" id="t4824" msd="fs">rede</tok>
    <tok class="punctuation" ctag="PNT" id="t4825" sp="y">.</tok>
  </definingText>
</s>
```

Fig. 3: *A definitory context containing the definition of "firewall": ...* `uma firewall é um conjunto de hardware e software...` *(...a firewall is a composition of hardware and software...)*

the attribute `base` in the corpus. The tag `cat` allows to indicate a category for the lexical item. In our grammar, `act` indicates that the verb occurs in definitions in the active form. A rule was written to match this kind of verbs:

```
<rule name="ActExpr">
<query match="tok[mylex(@base)
and (@msd[starts-with(.,'fi-3')]
or @msd[starts-with(.,'pi-3')])]"
constraint="mylex(@base)/cat='act'"/>
<ref name="Adv" mult="?"/>
</rule>
```

This rule matches a verb in present and future past (third person singular and plural), but only if the base form is listed in the lexicon and the category in equal to `act`. Similar rules were developed for verbs that occurs in passive and reflexive form.

Punctuation definitions

In this sub-grammar we take into consideration only those definitions introduced by colon mark. This happens to be the more frequent pattern in our data. The following rule characterize this grammar. It marks up sentences that starts with a noun phrase followed by a colon.

```
<rule name="punct_def">
<seq>
<start/>
<ref name="CompmylexSN" mult="+"/>
<query match="tok[.˜'ˆ:\$']"/>
<ref name="tok" mult="+"/>
<end/>
</seq>
</rule>
```

5 Outcomes

In the present section we report on the results obtained for the sub-grammars, and for the larger grammar made of the composition of them. Scores for Recall, Precision and F2-measure, for developing corpus (dv) and for test (ts) corpus are indicated. These scores were calculated at the sentence level, that is a sentence (manually or automatic annotated) is considered a true positive of a definition if it contains a part of a definition. Recall is the proportion of the sentences correctly classified by the system with respect to the sentences (manually annotated) containing a definition. Precision is the proportion of the sentences correctly classified by the system with respect to the sentences automatically annotated. The option for an F2 measure instead of F1 one is justified by the context in which these grammars are expected to operate.

Since the goal is to help the user in the construction of a glossary, it is important that the system retrieve as many definition candidates as possible. These candidates will be presented to the user in a graphical interface that will allow him to very quickly delete the bad candidates and keep a list with the accepted definition (instead of manually scanning the document to find each one of those definitions, in a much more time consuming way). Hence, to obtain a good recall is more important that obtain a good precision, in case the latter can be traded for the first.

We presented results for the different domains in the corpus, because we expected some difference in the performance due to the different nature of the material belonging to each domain.

Copula definitions

Table 3 displays the results of the copula grammar. These results can be put in contrast with those obtained with a grammar that provides values that can be seen as baseline scores. This simple grammar was developed to extract all sentences as definitory context provided they contain the verbal form of the verb "to be", in the third person singular and plural of the present and future past and in gerundive and infinitive form (no further syntactic patterns were taken in account).[5] In future experiments more patterns will be introduced in order to improve recall.

	Precision		Recall		F2	
	dv	ts	dv	ts	dv	ts
IS	0.40	0.33	0.80	0.60	0.60	0.47
IT	0.26	0.51	0.56	0.67	0.40	0.61
e-Learning	0.13	0.16	0.54	0.75	0.26	0.34
Total	0.30	0.32	0.69	0.66	0.48	0.49

Table 3: *Results for copula grammar*

Verbs definitions

The results obtained with the grammar for other verbs are not so satisfactory as the ones obtained with the copula grammar. This is probably due to the larger diversity of patterns and meaning for each such verb. In order to improve these results a deeper analysis of each verb pattern is required.

Punctuation definitions

As can be seen in Table 2, only 4 definitions of this type occur in the documents of the IS domain and 18 in e-learning domain. Consequently, this grammar for punctuation definitions ended up by scoring very badly in these documents. Nevertheless, the global evaluation result for this sub-grammar is better than the results obtained with the grammar for other verb definitions.

All-in-one

Finally, Table 7 presents the results obtained by a grammar that combines all the other three sub-grammars below. This table gives the overall performance of the system based on the grammars developed so far, that is this result represents the performance the end user will face when he will be using the glossary candidate detector.

To obtain the precision and recall score for this grammar, it is not anymore necessary to take into account the type of definition. Any sentence that is correctly tagged as a definitory context (no matter which definition type it receives) will be brought on board.

As can be seen, the recall value remains quite high, 86%, while it is clear that for the precision value (14%), there is much room for improvement yet.

---

[5] Note, however, that it is not clear how baseline results could be obtained for the other types of definitions.

	Precision		Recall		F2	
	dv	ts	dv	ts	dv	ts
IS	0.11	0.12	1	0.96	0.27	0.29
IT	0.09	0.26	1	0.97	0.22	0.51
e-Learning	0.04	0.5	0.82	0.83	0.12	0.14
Total	0.09	0.13	0.98	0.95	0.22	0.31

Table 4: *Baseline results for copula grammar*

	Precision		Recall		F2	
	dv	ts	dv	ts	dv	ts
IS	0.13	0.08	0.61	0.78	0.27	0.19
IT	0.13	0.22	.63	0.66	0.28	0.39
e-Learning	0.12	0.13	1	0.59	0.28	0.27
Total	0.12	0.14	0.73	0.65	0.27	0.29

Table 5: *Results for verb grammar*

6 Discussion

As expected, the results obtained with documents from the Information Society and Information Technology domains are better than the results obtained with documents from the e-Learning domain. This confirms our expectation drawn from the style and purpose of the material involved. Documents with a clear educational purpose, like those from IS and IT sub-corpora, are more formal in the structure and are more directed towards explaining concepts, many times via the presentation of the associated definitions. On the other hand, documents with a less educative purpose present less explicit definitions and for this reason it is more difficult to extract definitory contexts from them using basic patterns. More complex pattern and a deep grammar are likely to be useful in dealing with such documents.

Also worth noting is the fact that though the linguistic annotation tools used score at the state of the art level, the above results can be improved with the improvement of the annotation of the corpus. A few errors in the morpho-syntactic annotation were discovered during the development of the grammars that may affect the performance of the grammars.

Another issue about evaluation of our result is that determining the performance of a definition extraction system is not a trivial task. Many authors have pointed out that a quantitative evaluation as the one we carried out in this work may not be completely appropriate [12]. A qualitative approach to evaluation has to be taken in account (see for example [6]). For this reason we are planning to evaluate the effectiveness and the usefulness of the system with real users, by the means of user scenario methodology.

7 Conclusions and Future Work

In this work, we presented preliminary results of a rule-based system for the extraction of definitions from corpora. The practical objective of this system is to support the creation of glossaries in e-learning environments, and it is part of the LT4eL project aiming at improving e-learning management systems with the human language technology.

	Precision		Recall		F2	
	dv	ts	dv	ts	dv	ts
Calimera	0.00	00	0.00	00	0.00	00
IS	0.48	0.43	.68	0.60	0.60	0.53
e-Learning	0.05	00	0.58	00	0.13	00
Total	0.19	0.28	0.64	0.47	0.35	0.38

Table 6: *Results for punctuation grammar*

	Precision		Recall		F2	
	dv	ts	dv	ts	dv	ts
IS	0.14	0.14	0.79	0.86	0.31	0.32
IT	0.21	0.33	.66	0.69	0.38	0.51
e-Learning	0.11	0.11	0.79	0.59	0.25	0.24
Total	0.15	0.14	0.73	0.86	0.32	0.33

Table 7: *The combined result*

The better results was obtained with the system running over documents that are tutorials on information technology, where it scored a recall of 69% and a precision of 33%. For less educational oriented documents, 59% and 11%, respectively, was obtained.

We also studied its performance on different types of definitions. The better results were obtained with copula definitions, with 67% of recall and 51% of precision, in the Information Technology domain.

Compared to work and results reported in other publications concerning related research, our results seem thus very promising. Nevertheless, further strategies should be explored to improve the performance of the grammar, in particular its precision.

In general, we will seek to take advantage of a module that allows deep syntactic analysis, able to deal with anaphora and apposition, for instance. At present, a grammar for deep linguistic processing of Portuguese is being developed in our group [3]. We plan to integrate this grammar working in our system.

It is also worth noting that some authors[2, 9] noticed that some verbs are better clues for definitions than others, and that it is possible to measure this property in order to decide which verbs include in the searching patterns and which to exclude. For example, verbs that occur in many sentences, but only in few cases introduce a definition may perhaps be ignored. A more permissive approach is to parametrize the system in order to let the user choose whether he wants to use all the patterns in the grammar, obtaining a good recall but a worst precision, or to restrict the extraction only to the more promising patterns, thus improving precision at the expense of a worst recall. The same strategy can be applied to all type of definitions, not only to verb definitions.

Regarding punctuation definition, the pattern in the actual grammar can also be extended. At present, the pattern can recognize sentences composed by a simple noun followed by a colon plus the definition. Other rules with patterns involving brackets, quotation marks, dashes will be integrated.

Finally, in this work we ignored an entire class of definitions that we called "other definition", which represents 16% of all definitions in our corpus. These definitions are introduced by lexical clues such as *that is, in other words*, etc. This class also contains definitions spanning over several sentences, where the terms to be defined appears in the first sentence, which is then characterized by a list of features, each one of them conveyed by expressions occurring in different sentences. These patterns need thus also to be taken into account in future efforts to improve the grammar and its results reported here.

References

[1] *Interoperability of Enterprise Software and Applications*, chapter Methodology for the Definition of a Glossary in a Collaborative Research Project and its Application to a European Network of Excellence. Springer London, 2006.

[2] R. Alarcn and G. Sierra. El rol de las predicaziones verbales en la extraccin automtica de conceptos. *Estudios de Linguistica Aplicada*, 22(38):129–144, December 2003.

[3] A. Branco and F. Costa. LXGRAM – deep linguistic processing of Portuguese with HSPG. Technical report, Departement of Informatics, University of Lisbon, 2005.

[4] C. Grover, C. Matheson, A. Mikheev, and M. Marc. Lt ttt - a flexible tokenisation tool. In *In Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000.

[5] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[6] J. Klavans and S. Muresan. Evaluation of the DEFINDER system for fully automatic glossary construction. In *Proceedings of the American Medical Informatics Association Symposium (AMIA 2001)*, 2001.

[7] L. Kozakov, Y. Park, T. Fin, Y. Drissi, Y. Doganala, and T. Cofino. Glossary extraction and utilization in the information search and delivery system for ibm techical support. *IBM System Journal*, 43(3):546–563, 2004.

[8] L. Lemnitzer and L. Degórski. Language technology for elearning – implementing a keyword extractor. In *EDEN Research Workshop "Research into online distance education and eLearning. Making the Difference*, Castelldefels, Spain, October 2006.

[9] V. Malais, P. Zweigenbaum, and B. Bachimont. Detecting semantic relations between terms in definitions. In *the 3rd edition of CompuTerm Workshop (CompuTerm'2004) at Coling'2004*, pages 55–62, 2004.

[10] J. Person. The expression of definitions in specialised text: a corpus-based analysis. In M. Gellerstam, J. Jaborg, S. G. Malgren, K. Noren, L. Rogstrom, and C. Papmehl, editors, *7th Internation Congress on Lexicography (EURALEX 96)*, pages 817–824, Goteborg, Sweden, 1996.

[11] A. S. Pinto and D. Oliveira. Extracção de definições no Corpógrafo. Technical report, Faculdade de Letras da Universidade do Porto, 2004.

[12] A. Przepiórkowski, L. D. adn Miroslav Spousta, K. Simov, P. Osenova, L. Lemnitzer, V. Kubon, and B. Wójtowicz. Towards the automatic extraction of definitions in Slavic. In J. Piskorski, B. Pouliquen, R. Steinberger, and H. Tanev, editors, *Proceedings ofo the BSNLP workshop at ACL 2007*, Prague, 2007.

[13] H. Saggion. Identifying definitions in text collections for question answering. In *LREC 2004*, 2004.

[14] J. R. Silva. Shallow processing of Portuguese: From sentence chunking to nominal lemmatization. Master's thesis, Universidade de Lisboa, Faculdade de Ciências, 2007.

[15] E. Westerhout and P. Monachesi. Extraction of Dutch definitory contexts for elearning purposes. In *CLIN proceedings 2007*, 2007.