Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach

Rosa Del Gaudio and António Branco

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

Abstract. In this paper we present a rule-based system for automatic extraction of definitions from Portuguese texts. As input, this system takes text that is previously annotated with morpho-syntactic information, namely on POS and inflection features. It handles three types of definitions, whose connector between *definiendum* and *definiens* is the so-called copula verb "to be", a verb other that one, or punctuation marks. The primary goal of this system is to act as a tool for supporting glossary construction in e-learning management systems. It was tested using a collection of texts that can be taken as learning objects, in three different domains: information society, computer science for non experts, and e-learning. For each one of these domains and for each type of definition typology, evaluation results are presented. On average, we obtain 14% for precision, 86% for recall and 0.33 for F_2 score.

1 Introduction

The aim of this paper is to present a rule-based system for the automatic extraction of definitions from Portuguese texts, and the result of its evaluation against test data made of texts belonging to the domains of computer science, information society and e-learning.

In this work, a *definition* is assumed to be a sentence containing an expression (the *definiendum*) and its definition (the *definiens*). In line with the Aristotelic characterization, there are two types of definitions that typically can be considered, the formal and the semi-formal ones [1]. Formal definitions follow the schema $X = Y + C$, where X is the *definiendum*, " = " is the equivalence relation expressed by some connector, Y is the *Genus*, the class of which X is a subclass, and C represents the characteristics that turn X distinguishable from other subclasses of Y. Semi-formal definitions present a list of characteristics without the *Genus*.

In both types, in case the equivalence relation is expressed by the verb "to be", such definition is classified as a copula definition, as exemplified below:

J. Neves, M. Santos, and J. Machado (Eds.): EPIA 2007, LNAI 4874, pp. 659–670, 2007.

FTP é um protocolo que possibilita a transfêrencia de arquivos de um local para outro pela Internet.
FTP is a protocol that allows the transfer of archives from a place to another through the Internet.

Definitions are not limited to this pattern [2, 3]. It is possible to find definitions expressed by:

- punctuation clues:
  TCP/IP: protocolos utilizados na troca de informações entre computadores.
  TCP/IP: protocols used in the transfer of information between computers.

- linguistic expressions other than the copular verb:
  Uma ontologia pode ser descrita como uma definição formal de objectos.
  An ontology can be described as a formal definition of objects.

- complex syntactic patterns such as apposition, *inter alia*:
  Os Browsers, Navegadores da Web, podem executar som.
  Browsers, tools for navigating the Web, can also reproduce sound.

The definitions taken into account in the present work are not limited to copula definitions. The system is aimed at identifying definitory contexts based on verbs other than "to be" and punctuation marks that act as connectors between the two terms of a definition. Here, we will be calling 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.

The research presented here was carried out within the LT4eL project[1] funded by European Union (FP6) whose main goal is to improve e-learning systems by using multilingual language technology tools and semantic web techniques. In particular, a Learning Management System (LMS) is being improved with new functionalities such us an automatic key-words extractor [4] and a glossary candidate extractor. In this paper, we will focus on the module to extract definition from Portuguese documents.

The reminder of the paper is organized as follows. In Sect. 2 we present the corpus collected in order to develop and test our system. In Sect. 3 the grammars developed to extract definition are described.

In Sect. 4, the results of the evaluation of the grammar, in terms of recall, precision and F2-score, are presented and discussed.

An errors analysis and a discussion on possible alternative methods to evaluate our system are provided in Sec. 5

In Sect. 6, we discuss some related work with special attention to their evaluation results, and finally in Sect. 7 conclusions are presented as well as possible ways to improve the system in future work.

[1] www.lt4el.eu

2 The Corpus

The corpus collected in order to develop and test our system is composed by 33 documents covering three different domains: Information Technology for non experts, e-Learning, and Information Society.

Table 1. Corpus domain composition

Domain	tokens
Information SocietyS	92825
Information Technology	90688
e-Learning	91225
Total	274000

Table 1 shows the composition of the corpus.

The documents were preprocessed in order to convert them into a common XML format, conforming to a DTD derived from the XCES DTD for linguistically annotated corpora [5].

The corpus was then automatically annotated with morpho-syntactic information using the LX-Suite [6]. This is a set of tools for the shallow processing of 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%).

The last step was the manual annotation of definitions. To each definitory context was assigned the information about the type of definition. 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 definitions, `punct_def`, for definitions 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. The reason could be that this domain is composed by documents conceived to serve as tutorials for non experts, and have thus a more didactic style. In Sect. 4, we will see how this difference can affect the performance of the system.

Table 2. The distribution of types of definitions in the corpus

Type	Information Society	Information Technology	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

```
- <s id="s204">
  - <definingText def="m106" def_type1="is_def" id="d01">
      <tok base="o" class="word" ctag="DA" id="t4097" msd="ms" sp="y">O</tok>
    - <markedTerm dt="y" id="m106" kw="y">
        <tok base="tcp" class="word" ctag="PNM" id="t4098" sp="y">TCP</tok>
      </markedTerm>
      <tok base="ser" class="word" ctag="V" id="t4099" msd="pi-3s" sp="y">é</tok>
      <tok base="o" class="word" ctag="DA" id="t4100" msd="ms" sp="y">o</tok>
      <tok base="protocolo" class="word" ctag="CN" id="t4101" msd="ms" sp="y">protocolo</tok>
      <tok base="que" class="word" ctag="CJ" id="t4102" sp="y">que</tok>
      <tok base="dividir" class="word" ctag="V" id="t4103" msd="pi-3s" sp="y">divide</tok>
      <tok base="a" class="word" ctag="DA" id="t4104" msd="fs" sp="y">a</tok>
      <tok base="informação" class="word" ctag="CN" id="t4105" msd="fs" sp="y">informação</tok>
      <tok base="em" class="word" ctag="PREP" id="t4106" sp="y">em</tok>
      <tok base="pacote" class="word" ctag="CN" id="t4107" msd="mp" sp="y">pacotes</tok>
    </definingText>
  </s>
```

Fig. 1. The sentence `O TCP é um protocolo que divide a informação em pacotes (The TCP is a protocol that splits information into packets)` in final XML format

In Fig. 1, we present a sample of the final result. 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.

3 The Grammars

The grammars we developed are regular grammars based on the tools lxtransduce, a component of the LTXML2 tool set developed at the University of Edinburgh[2]. It is a transducer which adds or rewrites XML markup on the basis of the rules provided.

Lxtansduce allows the development of grammars containing a set of rules, each of which may match part of the input. Grammars are XML documents conforming to a DTD (lxtransduce.dtd). The XPath-based rules are matched against the input document. These 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.

All the grammars we developed present a similar structure and can be divided in 4 parts. The first part is composed by simple rules for capturing nouns, adjectives, prepositions, etc. The second part by rules that match verbs. The third part is composed by rules for matching nouns and prepositional phrases. The last part consist of complex rules that combines the previous ones in order to match the *definiens* and the *definiendum*.

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

A development corpus, consisting of 75% of the whole 274 000 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.

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

A sub-grammar for copula definition. Initially we developed a baseline grammar for this type of definition. This grammar marked as definition all that sentences containing the verb "to be" as the main verb of the sentence. In order to improve this grammar with syntactic information, copula definitions manually marked in the developing corpus were gathered. All information was removed except for the information on part-of-speech in order to discover the relevant patterns. Patterns occurring more than three times in the development corpus were implemented in this sub-grammar. Finally the syntactic patterns of all the sentence erroneously marked as definition by the baseline grammar were extracted and analyzed, in order to discover patterns that were common to good and bad definition. We decide not to implement in our grammar patterns whose occurrence was higher in the erroneously marked definitions that in the manually marked ones. We ended up with a grammar composed by 56 rules, 37 simple rules (capturing nouns, adjectives, prepositions, etc), 5 rules to capture the verb and 9 to capture noun and prepositional phrases and 2 rule for capturing the definitory context.

The following rule is a sample of the rules in the copula sub-grammar.

```
<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>
```

This is a complex rule that make use of other rules, defined previously 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 occurs in the third person singular or plural of the present or future past or in gerundive or infinitive form.

A sub-grammar for other verbs definition. In order to develop a grammar for this kind of definitions we start immediately to extract lexico-syntactic patterns. In fact it is hard to figure out how a baseline grammar could be implemented. We decided to follow the same methodology used for copula definition.

In a first phase we extracted all the definitions whose connector was a verb other than "to be", and collected all such verbs appearing in the developing corpus, obtaining a list of definitory verbs. This list was improved by adding synonyms. We decided to exclude some verbs initially collected from the final list because their occurrence in the corpus is very high, but their occurrence in definitions is very low. Their introduction in the final list would not improve recall and would have a detrimental effect on the precision score.

In a second phase we divided all the verbs obtained in three different classes: verbs that appear in active form, verbs that appear in passive form and verb that appear in reflexive form. For each class a syntectic rule was wrote. This information was listed in a separate file called lexicon.

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 corresponds to 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 is equal to `act`. Similar rules were developed for verbs that occur in passive and reflexive form.

A sub-grammar for punctuation definition. In this sub-grammar, we take into consideration only those definitions introduced by a colon mark since it is the more frequent pattern in our data. The following rule characterizes this grammar. It marks up sentences that start with a noun phrase followed by a colon.

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

4 Results

In this section we report on the results of the three grammar. Further more the results of a fourth grammar are presented. This grammar was obtained by