Using Wikipedia to Collect a Corpus for Automatic Definition Extraction: Comparing English and Portuguese Languages

1 Introduction

Systems for the detection and extraction of definitions are being developed for different purposes, such as glossaries creation [5, 3], lexical databases [6], ontologies [2], question answering [1], etc. All these systems use annotated corpora to build a set of rules or patterns capable to identify a definition in a different text.

The basic structure of a definition should resemble an equation with the *definiendum* (what is to be defined) on the left hand side and the *definiens* (the part which is doing the defining) on the right hand side. Between the term defined, and its description there is a a connector, usually a verb or a punctuation symbol.

In general, works in this field are restricted in terms of number and types of definitions considered, they are based on specific limited corpora very domain specific, lacking of a general approach. This limitation is due to scarcity of corpora previously annotated with definition information, as these corpora are not usually available and the annotation process constitutes a very expensive task. In this work we propose to use wikipedia as a corpus to extract general domains definitions, that can represent a bootstrap in the construction of a automatic definition extractor. The corpus can be used to draw pattern or extract lexical information characterizing definitions.

The convenience of using Wikipedia as font for definition is based on the peculiar structure of its articles, following well-defined rules stated by Wikipedia itself that contributors should follow when write an article. In particular Wikipedia states that the first paragraph of each article should define the topic of the article.

In this paper, we focus on the issues arising when extracting a general balanced corpus composed by Wikipedia articles and the size of such a corpus. We presented a study using two different languages, that is Portuguese and English, two different algorithms, and corpora of 5 different sizes.

2 Wikipedia

Wikipedia represent probably one of he larger open source language repository: more than 7.5 million articles in more than 250 different languages. Besides the value giving by its size, another advantage is constituted by the structure and metadata enriching the plain text. Articles in Wikipedia are not isolated piece of information, indeed they are linked to each others through both great number of inter-reference link and a structured category system.

For this rich structured information, Wikipedia has been used in a variety of NLP related task, such as text classification [8], information retrieval[9], question answering, computing semantic relatedness[10], or named entity recognition. Regarding definition extraction, Wikipedia was used as the main font to address definitional questions in QA systems [4].

For our specific propose we exploit both the category structure and the article structure characterizing Wikipedia. In the next two subsections we will describe these characteristics.

2.1 Article Structure

The structure of each article follows well-defined rules. In particular, Wikipedia states that the first paragraph of each article should define the topic with a neutral point of view, but without being overly specific.

The article usually begins with a declarative sentence giving a concise definition, telling the nonspecialist reader what is the subject. The first occurrence of term defined is placed in boldface.

These guide-lines allow to extract automatically the first sentences as a definition, where the term defined is the title of the article, the first verb in main form is the connector verb and what follows is the *definiendum.*

2.2 Category Structure

Categories in Wikipedia are organized in a taxonomy-like structure. This means that categories do not form a strict hierarchy or tree, since each article can appear in more than one category, and each category can appear in more than one parent category. Furthermore, each category can have an arbitrary number of subcategories, where a subcategory is typically established because of a hyponymy or meronymy relation.

When browsing Wikipedia categories for articles there are two top categories, parents of all other categories denoting a top-level place to start browsing the "tree of all knowledge". They represent a top level entry in terms of encyclopedia article function and content. These two top categories are "*Fundamentals*" and "*Main Topics*".

"*Fundamentals*" is intended to contain all and only the few most *Fundamental* ontological categories which can reasonably be expected to contain every possible Wikipedia article under their category trees.This category has four subcategories.

"*Main Topics*" is an alternative root category, based on a somewhat more detailed initial classification. It has twenty-two sub-categories.

3 The English and Portuguese Wikipedia

We accessed and analyzed Wikipedia dumps through Java Wikipedia Library (JWPL), an open-source, Java-based application programming interface that allows to access all information contained in a Wikipedia [10].

	EN	PT
Pages	8,739,845	1,240,318
Categories	744,971	116,885

Table 1: Wikipedia Dump

The two wikipedia used in this work are based on the dump available in `http://dumps.wikimedia.org/backup-index.html`. The English dump is dated 3rd of August 2011, while Portuguese one is dated 30th of May 2011. In Table1 the size of the two wikipedia is show.

4 Extracting corpora of definitions

When using Wikipedia to build as a general corpus for improving automatic definition extraction there are several questions that must be addressed, such as representativeness, sample and balance of the corpus. This is due to the fact that the grown of wikipedia is not controlled, and a particular area could be more developed than another and there is no way to know where it happens.

As explained in Section 2.2, articles in Wikipedia are organized in order to follow a hierarchical structure, from more general to more specific topics. Following this tree is possible to extract articles on general topics, selecting the articles directly linked to these top level categories. It also true that Wikipedia does not guarantee that the domain are equally covered and with the same granularity. This means that going down along the category structures some domains begin to include very specific articles very soon.

Two algorithms to collect articles are here proposed. A first algorithm (Alg1) collect the same number of articles for each category below the top category separately. In this way we want ensure that each domain, represented by the children of top categories, has the same likelihood to be represented.

The second algorithm (Alg2), first gather together all the articles linked to the top category children and then collect randomly the articles till get the desired number. As for the first algorithm, if the number of articles is less than the corpus size, the operation is repeated with the categories in the next level of the tree.

Using these algorithms, we extract five corpora with different size, containing respectively 1000, 10000, 25000, 50000, 100000 articles. The question we want to address is which top category is better to start from, either "*Fundamentals*" or "*Main Topics*", in which way to harvest the tree and the influence of different corpus size.

We automatically extracted the first sentence of each article, as it represent a definition, marking the defined term, the connector verb and the *definiens*.

5 Analyzing Corpora

In order to analyze the corpora, we focus our attention on the first noun after the connector verb "to be". The verb "to be" when used as connector verb in a definition introduces a generic hyperonyms occurring in definitions. Several authors focus on words such as

Table 2: Alg1 *Fundamentals* EN

1,000	10,000	25,000	50,000	100,000
term	term	term	term	term
element	process	process	process	plant
study	form	type	organization	organization
name	element	name	plant	type
form	type	organization	name	name
concept	concept	form	type	process
process	name	plant	form	form
phenomenon	study	concept	concept	genus
group	organization	element	method	species
type	method	method	study	method
state	theory	study	compound	compound
model	system	system	element	concept
theory	phenomenon	theory	species	book
statement	group	book	genus	study
organization	set	group	theory	element
method	model	genus	system	system
field	field	species	book	group
system	plant	set	group	act
act	approach	compound	set	theory
ability	state	field	act	part
word	branch	branch	branch	set
principle	measure	phenomenon	research	technique
part	part	practice	practice	research
meson	book	model	technique	branch
genus	act	research	field	journal

Table 3: Alg2 *Fundamentals* EN

1,000	10,000	25,000	50,000	100,000
term	term	term	term	term
concept	organization	organization	organization	organization
form	process	process	process	name
study	form	form	type	type
process	type	type	form	plant
organization	concept	name	name	form
theory	name	concept	concept	process
state	element	study	method	book
element	study	method	study	method
type	method	system	theory	concept
phenomenon	theory	group	book	genus
name	system	book	group	group
group	group	theory	system	language
approach	book	element	plant	study
act	field	plant	set	species
ability	act	set	act	journal
system	state	field	field	system
science	set	act	practice	organisation
part	practice	research	branch	act
material	model	approach	research	part
book	research	practice	element	theory
practice	branch	branch	business	association
model	phenomenon	state	approach	research
emotion	plant	phenomenon	technique	body
body	approach	movement	movement	set

Table 4: Alg1 *Main Topics* EN

1,000	10,000	25,000	50,000	100,000
term	term	term	term	term
study	process	process	process	type
process	form	organization	organization	organization
system	organization	form	form	name
set	study	type	type	plant
research	type	name	name	process
branch	method	study	method	journal
form	name	method	list	form
concept	concept	concept	study	book
computer	computer	list	book	list
practice	system	computer	concept	method
organization	field	book	journal	computer
theory	branch	field	plant	language
period	research	system	computer	study
method	theory	practice	language	concept
application	language	language	system	research
word	art	theory	research	device
type	technique	research	device	system
name	practice	art	practice	group
language	science	group	group	species
field	set	journal	theory	act
event	group	branch	art	part
discipline	act	set	field	technique
act	book	technique	technique	company
state	device	area	set	school

Table 5: Alg2 *Main Topics* EN

1,000	10,000	25,000	50,000	100,000
term	term	term	term	term
study	process	process	organization	organization
system	type	organization	process	type
process	form	type	type	plant
method	method	form	form	name
practice	organization	method	method	process
organization	study	study	name	form
field	concept	concept	study	journal
concept	name	name	concept	book
application	business	practice	plant	method
research	practice	research	book	list
form	system	field	device	language
branch	field	system	language	device
act	branch	business	journal	study
technology	research	set	system	concept
technique	set	device	research	research
set	theory	theory	company	system
business	science	result	practice	company
theory	result	approach	act	act
ability	act	branch	list	computer
type	device	technique	business	group
time	approach	language	set	technique
science	technique	company	group	species
measure	language	group	technique	software
event	technology	act	software	school

“technique”, “method”, “process”, “function”, called class words, representing generic hyperonyms characterizing definitions[7].

In order to examine the corpora regarding their balance, the terms extracted were ordered from the more to the less frequent. The idea is that in the first places we expected to find generic word such those enumerate by Pearson [7]. If specific words appear, this means that the corpus over-represents a specific domain. We present, for space reason, only the first 25 terms for each algorithm and for each top category. Terms belonging to specific domains are underlined.

Tables 2, 3, 4, 5 show results for English. Regarding corpora with size 1000 and 10000, for both the algorithms and both top categories, the number of domain specific terms is very low (1 or 2). With bigger corpora the best results are obtained when *Fundamentas* is used instead of *Main Topics* and Alg2 instead of Alg1. Looking at the specific terms, we can see than when *Fundamentals* category is used the domains that are overrepresented are linked to editorial area (book and journal) and to the botanical

Table 6: Alg1 *Fundamentals* PT

1,000	10,000	25,000	50,000	100,000
termo	espiral	espiral	espiral	asteroide
nome	galáxia	galáxia	asteroide	espécie
conjunto	termo	termo	galáxia	género
conceito	número	nome	espécie	espiral
forma	nome	espécie	nome	nome
símbolo	espécie	tipo	termo	galáxia
processo	tipo	organização	tipo	termo
organização	doença	doença	organização	empresa
número	conjunto	conjunto	sistema	gênero
fenômeno	forma	forma	forma	tipo
sistema	organização	número	conjunto	sistema
tipo	processo	asteroide	empresa	organização
teoria	sistema	processo	processo	conjunto
expressão	conceito	sistema	doença	forma
estado	ramo	grupo	número	grupo
designação	grupo	conceito	grupo	unidade
revista	asteroide	ramo	ramo	família
ramo	movimento	empresa	unidade	instituição
movimento	estrutura	expressão	órgão	órgão
unidade	área	gênero	instituição	processo
palavra	designação	unidade	conceito	instrumento
espécie	gênero	estrutura	movimento	ramo
parte	método	área	expressão	programa
estudo	ciência	parte	instrumento	doença
ato	estudo	método	programa	símbolo

Table 7: Alg2 *Fundamentals* PT

1,000	10,000	25,000	50,000	100,000
termo	termo	espiral	espiral	asteroide
forma	organização	galáxia	galáxia	espiral
conceito	nome	termo	asteroide	espécie
nome	número	nome	nome	nome
conjunto	forma	organização	termo	galáxia
processo	tipo	tipo	espécie	empresa
organização	conjunto	forma	empresa	termo
movimento	espécie	conjunto	tipo	tipo
sistema	conceito	número	organização	organização
tipo	processo	sistema	sistema	sistema
estado	sistema	espécie	forma	grupo
estudo	movimento	processo	conjunto	conjunto
designação	ramo	doença	órgão	unidade
área	doença	empresa	processo	forma
parte	expressão	conceito	grupo	instituição
palavra	associação	ramo	instituição	órgão
fenômeno	grupo	grupo	doença	processo
símbolo	instituição	movimento	unidade	partido
revista	teoria	instituição	ramo	estação
ramo	designação	órgão	número	doença
prática	empresa	expressão	movimento	ramo
método	área	associação	conceito	instrumento
denominação	estudo	unidade	expressão	entidade
teoria	ato	asteroide	programa	movimento
órgão	ciência	área	associação	associação

Table 8: Alg1 *Main Topics* PT

1,000	10,000	25,000	50,000	100,000
termo	termo	nome	nome	género
nome	nome	termo	termo	nome
conjunto	tipo	tipo	espécie	espécie
sistema	conjunto	sistema	tipo	espiral
forma	sistema	conjunto	género	empresa
conceito	forma	forma	sistema	termo
tipo	processo	espécie	espiral	tipo
processo	ramo	processo	conjunto	jogo
computador	computador	jogo	forma	galáxia
área	língua	doença	jogo	sistema
ramo	organização	espiral	organização	programa
ciência	conceito	organização	instituição	instituição
técnica	espécie	ramo	processo	série
programa	método	língua	empresa	grupo
organização	expressão	conceito	língua	conjunto
palavra	dispositivo	empresa	programa	forma
expressão	designação	programa	galáxia	gênero
método	movimento	expressão	grupo	língua
estudo	estudo	método	número	organização
documento	ciência	dispositivo	ramo	unidade
dispositivo	área	movimento	doença	processo
designação	programa	instituição	conceito	banda
tecnologia	técnica	designação	escola	partido
revista	empresa	computador	método	ramo
instituição	instituição	grupo	instrumento	instrumento

Table 9: Alg2 *Main Topics* PT

1,000	10,000	25,000	50,000	100,000
termo	termo	nome	género	género
conjunto	nome	termo	nome	espiral
forma	conjunto	tipo	termo	nome
processo	sistema	sistema	tipo	galáxia
nome	tipo	conjunto	sistema	espécie
sistema	forma	forma	empresa	empresa
tipo	processo	processo	espécie	termo
designação	ramo	organização	organização	tipo
computador	conceito	ramo	conjunto	gênero
técnica	organização	empresa	espiral	asteroide
revista	computador	conceito	forma	sistema
organização	língua	programa	processo	organização
expressão	dispositivo	espécie	grupo	grupo
conceito	movimento	doença	gênero	conjunto
área	método	dispositivo	ramo	jogo
ramo	empresa	número	número	forma
grupo	expressão	movimento	jogo	instituição
ato	estudo	método	galáxia	unidade
dispositivo	designação	dia	dia	partido
ciência	área	designação	instituição	processo
tecnologia	técnica	instrumento	doença	freguesia
programa	doença	língua	programa	comuno
estudo	ciência	grupo	movimento	programa
estrutura	parte	expressão	conceito	órgão
palavra	programa	computador	método	instrumento

area (plant). When using *Main Topics*, at least other two over-represented domains are added, that is computer science (computer, software, language) and business (business and company).

Tables 6, 7, 8, 9 present the word lists for Portuguese corpora. As for English, the best results are obtained when Alg2 is used in conjunction with *Fundamentals* category. Regarding over-represented domain the situation is worst. As for English, we have the editorial area (revista =”magazine”), but then we have the health field (doença=“illness”), the astronomic domain (asteroide=“asteroid”, galáxia=“galaxy”), the math domain (espiral= “spiral”, número=“number”). When analyzing the word lists for *Main Topics* we find again the computer science domain (computador=“computer”, língua=“language”) but then we have also a number of other terms indicating very different domains such as jogo=“game”, dia=“day”, frequesia=“municipality”, etc.

6 Discussion and Conclusions

The word lists presented in the previous Section allows us to draw some final observations. In general corpora extracted starting from *Main Topics* are most affected by over-represented domains, especially when considering the three biggest corpora. This can be explained by the fact that this category has 22 children, representing specific domains. It turns more likely to encounter a over-specified area, composed for example by a list of all galaxy or of all plants. When comparing English and Portuguese experiments, Portuguese corpora present a greater number of over-represented domains. A possible explanation takes in consideration the size of Wikipedia, as Portuguese Wikipedia is by far smaller than the English ones, the number of article on general topics run out sooner.

To conclude, in this paper we show a method for building a corpus of definition using Wikipedia, applicable to different languages. We discuss two different algorithms and two different starting point categories. For both languages, the *Fundamentals* category in combination with Alg2, seem to devolve a more balanced corpus. Furthermore, a list of class word were extracted, that by itself represent a valuable resource in the definition extraction field.

References

[1] X. Chang and Q. Zheng. Offline definition extraction using machine learning for knowledge-oriented question answering. In D.-S. Huang, L. Heutte, and M. Loog, editors, *ICIC (3)*, volume 2 of *Communications in Computer and Information Science*, pages 1286–1294. Springer, 2007.

[2] M. C. de Freitas. *Elaboração automática de ontologias de domínio: discussão e resultados.* PhD thesis, Pontifícia Universidade Católica de Rio de Janeiro, 2007.

[3] R. Del Gaudio and A. Branco. Learning to identify definitions using syntactic feature. In *Progress in Artificial Intelligence, 13th Portuguese Conference on Aritficial Intelligence, EPIA 2007*, pages 659–670, Guimarẽs, Portugal, December 2007. Springer Berlin.

[4] I. Fahmi and G. Bouma. Learning to identify definitions using syntactic feature. In R. Basili and A. Moschitti, editors, *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*, Trento, Italy, 2006.

[5] S. Muresan and J. Klavans. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2002.

[6] J. Nakamura and M. Nagao. Extraction of semantic information from an ordinary english dictionary and its evaluation. In *In Proceedings of the Twelfth International Conference on Computational Linguistics*, 1988.

[7] J. Pearson. *Terms in Context.* John Benjamins Publishing Company, 1998.

[8] N. Tomuro and A. Shepitsen. Construction of disambiguated folksonomy ontologies using wikipedia. In *People's Web '09: Proceedings of the 2009 Workshop on The People's Web Meets NLP*, pages 42–50, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[9] T. Zesch and I. Gurevych. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 2007.

[10] T. Zesch, C. Müller, and I. Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In N. C. C. Chair), K. Choukri, B. Maegaard, J. O. Joseph Mariani, S. Piperidis, and D. Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.