João Miguel Jones Ventura

MSc in Computer Science

Automatic Extraction of Concepts from Texts and Applications

Dissertação para obtenção do Grau de Doutor em Informática

Orientador : Joaquim Francisco Ferreira da Silva, Prof. Auxiliar, Universidade Nova de Lisboa

Júri:

Presidente:	Doutor Pedro Manuel Calvente de Barahona
Arguentes:	Doutor Paulo Miguel Torres Duarte Quaresma
	Doutor Pavel Bernard Brazdil
Vogais:	Doutor José Gabriel Pereira Lopes
	Doutor Nuno João Neves Mamede
	Doutor Joaquim Francisco Ferreira da Silva

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

May, 2014

Automatic Extraction of Concepts from Texts and Applications

To my parents, João and Rosabela Ventura.
To my sister, Amarílis Ventura and my niece, Ariana.
To Carmen Matos.

Acknowledgements

To my advisor, Prof. Doutor Joaquim Ferreira da Silva, Assistant Professor at FCT/UNL, I thank for all the support on this PhD thesis. We have been working together since my Master's course for about 8 years now, and starting from the first ideas for the extraction of relevant words from texts, to all the discussions that arose from there, I have learned so much with him.

I would also like to thank the members of the Thesis Advisory Committee, namely Prof. Doutor Gabriel Pereira Lopes and Prof. Doutor Gaël Harry Dias, for the advise given at the Thesis Plan and Thesis Proposal discussions. I thank also Prof. Doutora Maria Francisca Xavier of the Linguistics Department of FCSH/UNL for the feedback and manual evaluation of some of test cases of the Concept Extractor and applications.

I would also like to acknowledge the FCT/MCT, as the research for this thesis was supported by grant SFRH/BD/61543/2009.

Finally, I reaffirm the acknowledgments to my parents, João Ventura and Rosabela Ventura, my sister Amarílis Ventura, and Carmen Matos for all the support and understanding.

Abstract

The extraction of relevant terms from texts is an extensively researched task in Text-Mining. Relevant terms have been applied in areas such as Information Retrieval or document clustering and classification. However, *relevance* has a rather fuzzy nature since the classification of some terms as *relevant* or *not relevant* is not consensual. For instance, while words such as "president" and "republic" are generally considered relevant by human evaluators, and words like "the" and "or" are not, terms such as "read" and "finish" gather no consensus about their semantic and informativeness.

Concepts, on the other hand, have a less fuzzy nature. Therefore, instead of deciding on the relevance of a term during the extraction phase, as most extractors do, I propose to first extract, from texts, what I have called *generic concepts* (all concepts) and postpone the decision about relevance for downstream applications, accordingly to their needs. For instance, a keyword extractor may assume that the most relevant keywords are the most frequent concepts on the documents. Moreover, most statistical extractors are incapable of extracting single-word and multi-word expressions using the same methodology. These factors led to the development of the *ConceptExtractor*, a statistical and language-independent methodology which is explained in Part I of this thesis.

In Part II, I will show that the automatic extraction of concepts has great applicability. For instance, for the extraction of keywords from documents, using the *Tf-Idf* metric only on concepts yields better results than using *Tf-Idf* without concepts, specially for multi-words. In addition, since concepts can be semantically related to other concepts, this allows us to build implicit document descriptors. These applications led to published work. Finally, I will present some work that, although not published yet, is briefly discussed in this document.

Keywords: Concepts, extractor, application of concepts, keywords, semantic relations.

Resumo

A extracção de termos relevantes é uma área muito investigada em Text-Mining. Estes termos têm sido aplicados em áreas como *Information Retrieval*, entre outras. No entanto, a *relevância* tem uma natureza relativamente difusa, uma vez que a classificação de alguns termos como *relevante* ou *não relevante* não é consensual. Por exemplo, enquanto palavras como " presidente" e " república" são geralmente consideradas relevantes, e outras como " o" e " ou" não o são, palavras como " ler" e " terminar" não reúnem consenso.

Os conceitos, por outro lado, têm uma natureza menos difusa. Portanto, invés de decidir sobre a relevância de um termo durante a fase de extracção, como o fazem os extractores actuais, proponho extrair primeiro dos textos aquilo a que chamei *conceitos genéricos* (todos os conceitos) e adiar a decisão sobre a relevância para as aplicações a jusante, de acordo com as suas necessidades. Por exemplo, um extractor de palavras-chave poderá assumir que as palavras-chave relevantes são os conceitos mais frequentes nos documentos. Além disso, os extractores estatísticos actuais são incapazes de extrair palavras únicas e multipalavras usando a mesma metodologia. Estes factores levaram ao desenvolvimento do *ConceptExtractor*, uma abordagem estatística e independente da língua que é explicada na Parte I desta tese.

Na Parte II, irei mostrar que a extracção automática de conceitos tem grande aplicabilidade. Por exemplo, na extracção de palavras-chave de documentos, a utilização da métrica *Tf-Idf* apenas em conceitos produz melhores resultados do que o uso do *Tf-Idf* sem conceitos, especialmente para multipalavras. Além disso, visto que os conceitos podem estar relacionados semanticamente com outros conceitos, isto permite-nos construir descritores implícitos de documentos. Estas aplicações deram origem a trabalhos publicados. Por fim, apresentarei algum trabalho que, apesar de não estar publicado, será brevemente discutido neste documento.

Palavras-chave: Conceitos, extractor, palavras-chave, relações semânticas.

Contents

List of Figures

List of Tables

1 Introduction

The automatic extraction of relevant terms from texts has been an extensively researched topic in the Text Mining area. Relevant terms are informative words or sequences of words with a high semantic value, and they have been successfully used in diverse applications such as Information Retrieval, document clustering, and classification and indexing of documents.

However, a large majority of the work has been done on the extraction of relevant multi-word expressions. This means that the automatic extraction of relevant single-word units has been largely ignored. Nevertheless, it is easy to show that leaving out relevant single-words impoverishes, to a certain extent, a process of knowledge extraction. Take, for instance, the following excerpt from the English *Arthritis* Wikipedia document:

> Gout is caused by deposition of uric acid crystals in the joint, causing inflammation. (...) The joints in gout can often become swollen and lose function. Gouty arthritis can become particularly painful and potentially debilitating when gout cannot successfully be treated.

Although multi-word terms such as "uric acid", "uric acid crystals" and "gouty arthritis" would probably be captured by most modern multi-word extractors, informative single-word terms such as "gout", "joint" and "joints" would not. Similarly, the relevant single-words which compose some of the multi-word terms, such as "acid", "crystals" and "arthritis", would also be discarded by those extractors. Thus, much of the knowledge in this small excerpt would simply be ignored.

Furthermore, languages such as German and Dutch tend to have complex terms which are agglutinated into a single-word. For instance, the German word for "master's certificate" (Kapitänspatent) is the junction of "Kapitän" (meaning *sea captain*) and "patent" (*license* or *certificate*). This kind of relevant and complex single-word terms would also be left out by current multi-word extractors. Therefore, a unified approach for extracting relevant single-words and multi-word expressions using a similar methodology is a major motivation for this thesis.

However, the notion of *relevance* (as in *relevant single-words* and *relevant multi-word expressions*) has a rather fuzzy nature. Consider, for instance, Table 1.1 which presents an example of the manual classification about the relevance of some terms from the previous excerpt.

Table 1.1: Classification of some terms from the *Arthritis* document excerpt.

Relevant terms	**Non-relevant terms**	**Non-consensual terms**
uric	is/of/by/...	deposition
acid	in the/can often/...	inflammation
crystals	caused	swollen
uric acid	successfully	lose function
uric acid crystals	treated	painful
acid crystals	causing	debilitating
gout	particularly	potentially debilitating
arthritis	potentially	–
gouty arthritis	become	–
joint	–	–

For instance, while terms such as "uric acid", "gouty arthritis" or "joint" are usually considered relevant by human evaluators, less informative concepts such as "deposition", "inflammation", "swollen", "lose function", "painful" and "debilitating" gather no consensus. This happens mainly because Text-Mining is frequently used for Information Retrieval tasks, and concepts like these are usually considered as not informative enough for most tasks. For instance, concepts such as "painful" and "debilitating" are not considered relevant for tasks such as the extraction of keywords from documents since they usually do not describe the content of documents. But undeniably, these terms have a semantic value, and they may be useful for other kind of applications. Thus, a methodology for the extraction of *generic* concepts is also one of the main purposes of this thesis.

This thesis presents a unified and language-independent methodology for the extraction of single-word and multi-word concepts from texts. Given that different tasks may use concepts in different manners, this thesis also proposes that the relevance of a concept should depend on the specific needs of each task. Therefore, to support this view, this thesis also presents some applications which make extended use of the extracted concepts.

1.1 Motivations

1.1.1 A statistical approach for single-words and multi-words

Most methodologies for the extraction of relevant terms from texts are currently divided into linguistic, statistical or hybrid approaches. In a general way, linguistic and hybrid approaches tend to use syntactic filters and other language-dependent tools, and not all languages have high quality taggers and parsers available. This makes statistical methods more desirable when language independence is a requirement. Besides, relevancy is not completely determined by morphosyntactic patterns. For instance, "triangle angle" and "greenhouse effect" share the same *Noun-Noun* pattern, however, only the second one is usually considered relevant.

Regarding the statistical methods, the majority of work has been done on the extraction of multi-word expressions. This means that the automatic extraction of relevant single-word units has been largely ignored, and as I mentioned previously, leaving out the relevant single-words impoverishes, to a certain extent, the process of knowledge extraction.

Currently, as far as I know, there are no statistical extractors capable of extracting both relevant single-word and relevant multi-word expressions using the same base methodology. That poses an interesting challenge, and the development of such methodology is of great interest.

1.1.2 Extraction of *generic* concepts from texts

Given that the notion of *relevance* has a rather fuzzy nature, it is proposed in this thesis that what is routinely known as relevant single-words and relevant multi-word expressions are essentially the most relevant (or informative) concepts in texts. Yet, unlike relevant single-words and multi-word expressions, concepts have a less fuzzy nature. For instance, although concepts such as "inflammation" or "painful" would probably not be considered relevant enough for most Information Retrieval tasks, they are, without a doubt, informative concepts in the sense that they have some semantic value, i.e, they convey an idea, a thought. But regarding their relevance, their interest is, say, fuzzy, mainly because they are dependent on the task at hand. In this sense, while their interest may be low for a task of keyword extraction, because they may not describe enough a core subject, they may be of high interest for a generic knowledge extraction application.

Thus, instead of extracting relevant single-words and multi-word expressions from texts (and consequently, having to define beforehand what is and what is not relevant), this thesis is focused on the extraction of *generic concepts* from texts. By doing this, we postpone the decision about the relevance of concepts to the tasks that will use them, i.e, for the applications themselves. Therefore, the creation of an extractor capable of extracting both single-word and multi-word concepts is one of the main purposes of this thesis.

1.1.3 Applicability of extracted concepts

By extracting all concepts from a text (instead of the small subset of *relevant* terms), and *feeding* them downstream, we guarantee that most knowledge in the texts is made available to the downstream applications. Then, accordingly to the specific needs of the task, each application decides which concepts are relevant. For instance, while a task of keyword extraction may assume that the most relevant keywords are the most frequent and exclusive concepts in each document, a task of *generic* knowledge extraction or thesaurus construction may assume that all concepts are equally relevant for its analysis.

To support the view that the definition of *relevance* of a term is strongly up to the purpose of the tasks which will use the concepts, the creation of several applications which make extended use of concepts were also a major motivation for this thesis.

1.2 Main contributions of this thesis

The following subsections summarize some of the main contributions of this thesis.

1.2.1 Part I – ConceptExtractor

Part I of this thesis presents, as main contribution, the research which led to the implementation of the *ConceptExtractor*, an approach capable of extracting both single-word and multi-word concepts. *ConceptExtractor* was published in ICCS 2012, an *A*-type conference [VS12], and is capable of extracting concepts in text corpora with Precision and Recall values of about 90% for the tested corpora. Some of the innovations associated to this approach are:

- **The *RelVar* metric**
  The core of the extractor is the identification of semantic relations between pairs of words which are not necessarily contiguous. Concepts tend to co-occur at fixed positions relatively to each other and *RelVar* is a simple statistical metric to detect and quantify those situations.

- **Specificity of concepts**
  It is possible to quantify with the *ConceptExtractor* how specific a concept is in a certain text, in relation with other concepts. More specific concepts tend to carry, say, *more* semantic information.

- **Independence on cohesion metrics**
  Generally, multi-word extractors use cohesion metrics to identify the pairs of words which tend to co-occur statistically above average. This tends to fail with multi-word concepts for which one of the words is fairly common, as is the case of "typical antipsychotic" where the word "typical" is far from being used exclusively with "antipsychotic". *ConceptExtractor* does not depend on this type of metrics.

- **Identification of concepts**
  My research led to the consideration that there is an uniform, cross-language, *specificity* threshold value for concepts. Words and multi-words below a certain specificity value tend to be too generic/vague, carrying little semantic information. Therefore, they must not be considered concepts.

- **Language independence**
  The statistic character and the non-usage of morphosyntactic filters on this approach makes it independent of the language or application. This allows it to extract concepts in several languages and for several applications.

- **Applicability**
  There are several domains for which the automatic extraction of concepts may be useful, besides the ones presented in Part II of the thesis:

  - Enrichment of lexicons for Natural Language Processing.
  - Enrichment of terminological dictionaries.
  - Improvement of the automatic translation between languages, using concepts extracted from parallel texts, and identifying translation pairs by means of specificity values. The same concepts should have similar specificity values even in different languages.
  - Access to existing information in document collections, using extracted concepts as document descriptors. Users may search for specific documents using tools like search engines specifically tailored for this type of application.
  - Unsupervised document clustering for multi-language corpora.
  - Etc ...

1.2.2 Part II – Applicability of the extracted concepts

Part II of this thesis presents, as contributions, the research which led to the implementation of some applications using the concepts automatically extracted by the extractor described in Part I. Some of the work described in the second part was published in two Book Chapters [VS13a; VS13b]. Some of those innovations are:

- **Explicit document descriptors**
  Keywords of documents are essentially the most meaningful concepts occurring explicitly in the documents. By using the *ConceptExtractor* to automatically extract the concepts from documents, we are in fact reducing the search space from all possible sequences of single-words and multi-word expressions to a much smaller set of semantically meaningful concepts. Having a smaller set of terms to analyze, statistical metrics such as *Tf-Idf* can be applied successfully to both single-word and multi-word concepts, in order to find the best descriptors.

- **Implicit document descriptors**
  There are meaningful concepts that, although not occurring in the text of a document, are semantically related to its content. I call these the *implicit keywords* of a document. Concepts such as "car emissions", "Toxicology" and "acid rains" may be useful if automatically added as implicit keywords of a document about "air pollution", if those terms do not occur explicitly in that document. They may, for instance, provide a user of a search engine the access to documents that may not contain these keywords, but are semantically related to them.

- **Identification of semantic relations in collections of documents**
  By being semantically rich, concepts tend to relate with other concepts. For instance, "car" is related to "automobile" and to "means of transportation". When concepts tend to co-occur in the same documents of a document collection, it can be assumed that their meanings are somewhat related. Thus, by extracting concepts with extractor presented in Part I, and then using a statistical, language-independent metric, it can be measured how semantically related a pair of concepts is in a collection.

- **Identification of clusters of concepts**
  A cluster of a concept is a specific area on a text where a concept is relevant and tends to occur rather densely. When a concept occurs densely in an area, it usually implies that its meaning is being used in that area. The identification of clusters of concepts is essential for the next three contributions.

- **Measuring semantic relations in standalone documents**
  When two concepts tend to form clusters in the same areas of a document, it means that they may be semantically related at a low-level. For instance, in a paragraph describing *Gout* (an inflammation of the joints) it is said that gout is caused by deposition of uric acid crystals. If "gout" and "uric acid" are used densely in that paragraph, both terms will form clusters in that area.

- **Finding changes of topic in documents**
  Although structured documents such as papers, thesis and books, have clearly defined frontiers between passages (such as sections or chapters), some documents, especially web documents, are usually unstructured. However, most of these texts can be broken into fine-grained subtopics. *TextTilling* [Hea97] is a widely known algorithm in this area, and it will be shown that the usage of concepts can improve the performance of this algorithm.

- **Finding descriptive areas of documents**
  Many documents, such as encyclopedic articles, do not have a uniform distribution regarding the description of the underlying subject. In fact, some sections are more descriptive than others. When a lot of concepts occur densely in some specific areas in detriment of others, it may indicate that these areas can be more *interesting*

to readers. Clusters of concepts can be used to measure the density of concepts throughout a document.

- **Concept definition**
  As mentioned, a cluster of a concept occurs when a concept is being highly used in a specific area of a document. When a concept is being used in the same area as other concepts, in some cases it corresponds to its definition, especially when encyclopedic texts are being used as source.

1.3 Structure of this document

This thesis is divided in two parts: Part I deals with the automatic extraction of concepts from texts. It starts on chapter 2 by presenting some of the current state-of-the-art methods for the extraction of concepts. On chapter 3 an empirical definition of concepts will be presented. The purpose is to demonstrate that there are some relations between concepts which can be explored through a statistical approach. The rest of the chapter shows how the *ConceptExtractor* uses those relations to infer about the *specificity* of concepts. Finally, chapter 4 shows how the *specificity* of concepts allows us to separate concepts from non-concepts. The results for the *ConceptExtractor* will be presented in this chapter, including comparative results with some of the methods reviewed.

Part II presents some applications implemented during the research phase to make use of the extracted concepts. The purpose of this section is to support the view that the relevance of a term is mostly dependent on the goals of each task. Chapter 5 deals with the extraction of explicit and implicit keywords while chapter 6 presents a new methodology for the extraction of semantic relations using clusters of concepts. Since these applications are somewhat specific, I will present the state-of-the-art methods for each application in each respective chapter. Chapter 7 presents three other possible applications for concepts which, by lack of opportunity, were not extensively researched and did not led to effective publications. However, some preliminary results were obtained, and they represent, essentially, opportunities for future research. Finally, chapter 8 presents the conclusions.

Part I

Automatic extraction of concepts from texts

2

Current Work

In this chapter I present some current methodologies for the extraction of concepts that I am aware of, and which are representative of the possible types of approaches. Since relevant single-words and multi-word expressions can be considered as a subset of all concepts occurring on texts, as mentioned in Chapter 1, some extractors are also presented in this chapter. Because of the fact that there are no extractors capable of extracting both relevant single-words and multi-word expressions using the same methodology, I will present them separately.

2.1 Concept extractors

Unlike the following sections, which handle the extraction of relevant words and multi-words, the work discussed in this section is about methodologies which claim to extract complete concepts from the texts.

2.1.1 CICM – a linguistic approach

In their paper [ZW10], Zhou and Wang present a method for the extraction of concepts on texts and to discover inner semantic relations within concepts, namely the type of relation. To extract the concepts from the texts, they use lexical patterns. Since they are working with Chinese texts, they first designed a set of rules based on Chinese lexical patterns, for which the concepts tend to be on predefined lexical positions. To extract those concepts, they summarized the following criteria:

- **High accuracy:** Each lexical pattern on texts must be reflected by at least one linguistic rule.

- **High coverage:** Each concept belongs to only one of three groups (physical object, time object and generic concept).

Their idea is that a *chunk* of text extracted using a lexical pattern is a concept if it has been matched by several rules (high accuracy criteria) and has been matched sufficient times by each single rule (high coverage criteria). In their experiments, they identify concepts if the number of matching patterns is greater than 5 and each pattern is matched at least 14 times. Although they report a very high precision using this approach (about 98.5%), they also report very low recall results.

To fix the low recall problem, they propose the *CICM* (Concept Inner-Constructive Model) to recognize more concepts from text chunks. Their hypothesis is that concepts obey other rules inside the first rules. In practice the *CICM* is a list of *C-vectors*, where, for a concept $W = w_1w_2\cdots w_n$, the *C-vector* of a single word w_i is an ordered list with the words that occur before and after w_i in the analyzed texts. Since the word w_i can occur with other neighbor words on the texts, w_i may have more than one *C-vector*.

For constructing automatically the *CICM*, the authors use an external lexicon (the *HowNet* dictionary in their experiments), and keep only the *C-vectors* which form frequent patterns in the text. Because this generates a lot of *C-vectors*, the authors propose a method to cluster similar words – they compute the *distance* between two vectors using a *similarity* metric, and group the *C-vectors* which score higher than a given threshold based on a Gaussian function. Finally, to identify if a chunk of text is a concept, they compute the similarity to all their *C-vectors*, and use the same rules as before (number of matching patterns greater than 5 and each pattern matched at least 14 times). Later, they proceed to the identification of semantic relations between concepts, although that is outside the scope of this part of the Thesis.

Clearly, this method is not language independent, since the authors use lexical patterns which are specifically for the Chinese language, and an external lexicon is used to generate further rules to fix the low recall problem. A complete rewrite of the morphosyntactic filters would be necessary if this method was to be applied to other languages.

2.1.2 GARAGe – an approach using external lexicons

In their paper *Automated Concept Extraction from Plain Text* [GWP98], the authors describe a system for extracting concepts from unstructured text by identifying relationships between words based on a lexical database.

The main idea of this approach is that most concepts have semantic relations between them. The authors propose to represent those semantic relations by means of a structure which represents the text's thematic content. They call this structure a *Semantic Relationship Graph* (SRG), which is essentially a graph where the nodes are the concepts and the existence of semantic relations is given by the lines connecting the nodes.

For building the SRG, the authors start by breaking the text into its single-word components, originating a unordered list of unigrams. Then, for each single-word (called *base words*), they proceed to consult its occurrence and semantical relations on an external lexicon – *Wordnet* in this paper. If a semantic relation between two base words is found in the lexicon, even if by means of a third concept not occurring in the text, the relation is drawn between those two base words. There can be more than one "bridge" word not occurring on the texts between a pair of base words, up to a certain predefined number.

Finally, having the structure which relates the words to each other, the idea is that words which do not have any semantic relations, or have incomplete semantic relations, are considered outliers and non-concepts.

The use of external lexicons, such as Wordnet, may affect the quality of results, since these lexicons are usually not entirely complete. Therefore, some semantic relations may not be identified and, as such, some valid concepts may be discarded.

2.1.3 DIPRE – a domain-specific pattern-based approach

The work in [Bri99] presents a pattern-based approach for the extraction of concepts from texts. The idea behind this paper is that some domain specific concepts can be extracted from web documents by exploring recognizable patterns in the texts.

Specifically, in this paper, the author considers the problem of extracting books, namely author names and book titles (tuples *(name, title)*). He starts with a small seed of *(name, title)* pairs. Then, from these occurrences, he recognizes patterns from the citations of these books which will then be reused to find new books. Finally, with these new books, he is able to generate new patterns which will be used to find even more books. The process ends after some iterations.

The method proposed is called *DIPRE* (Dual Iterative Pattern Relation Expansion) which relies on the duality between patterns and relations. For the experiment about books, the author defines a pattern as a 5-tuple *(order, urlprefix, prefix, middle, suffix)* where *order* is a boolean value which is set to true if there is a pair *(author, title)* matching the pattern. If *order* is set to false, the pair *(author, title)* is switched as *(title, author)*.

An important component of this method is the generation of patterns, which takes a set of occurrences of books and converts them into a list of patterns. Since this is not a trivial task, the author uses a very simple set of heuristics. Also, since patterns can be too general or too specific, the author measures the specificity of a pattern as the length of the pattern string and rejects all patterns which length is greater than a given threshold. This allows him to get rid of generic patterns and empty ones, since highly specific patterns are still usable. Finally, for matching the patterns with the text on the documents, the author uses regular expressions.

By using predefined patterns, this approach is highly domain dependent. However, that seems to be the intention, as the author demonstrates by using it on a highly specific domain such as book titles and author names.

2.1.4 KOSMIX – an hybrid extractor

In [PRGM10] the authors propose a technique to extract concepts from large datasets, mainly web pages. The authors start by defining that their target are *k*-grams, representing entities, events or ideas, that are somewhat popular (for which most users may be interested in) and concise.

An important observation of this technique is that for a k-gram $a_1a_2..a_k$ with $k > 2$, it is not true that both $(k-1)$-grams ($a_1..a_{k-1}$ and $a_2..a_k$) are necessarily concepts. For instance, for the 3-gram "Manhattan Experimental Theater", "Manhattan Experimental" is not considered a concept but "Experimental Theater" is. For $k = 2$, they assume that both words must be concepts.

As for the procedure, the first step is the extraction of all k-grams in a dataset, up to a predefined size *n*, tagged with the frequency of occurrence of each k-gram. This *n* is set to 4 on their experiments, for which they claim to be the largest length of most concepts, given by the titles of the Wikipedia articles they had access to. Then, since they want concise concepts, their idea is that either a k-gram is a concept, or one of its $(k-1)$-grams are concepts. For that, they use the following indications:

- The frequency of occurrence of a concept should be higher than a given threshold (i.e., concepts must be "popular").
- Either the k-gram is *better* than all its $(k-m)$-grams, or it is not a concise concept.
- A concept must contain only portions of sentences that convey a single meaning or idea.

For the first indication, because concepts must occur more than a given threshold, this means that rare concepts may be ignored. For instance, the English wikipedia article *Otolaryngology* has only about 25 occurrences of this concept. Given its specificity, it is possible that a random corpus from Wikipedia documents may contain only one or two documents which refer to this medical specialty once or twice. However, the authors set a threshold of 100 to 1-grams, which means that *otolaryngology* may never be considered a concept as well as other highly specific and infrequent concepts. This is a factor which probably gives low recall value to this technique, but the authors chose to publish only Precision results.

For the second indication, the idea is that a k-gram must be "better" than all possible sub-$(k-m)$-grams. For that, they rely on the concept of *confidence*, which is basically the probability of occurrence of the k-gram given both its $(k-1)$-grams. A k-gram is a concept if its *confidence* is greater than a given threshold. For single-words, they use 100 for the frequency of occurrence threshold in their experiments.

For the last indication, this means that candidate concepts are rejected if they start or end with function words and verbs, or do not contain nouns. Although the criterion that states that concepts must neither start or end with function words is a sensible one, and also used on the extractor that I proposed in the context of this thesis, that is the *ConceptExtractor*, the criterion that states that concepts must not start or end with verbs means that concepts such as *Cry me a River*, a popular song by Justin Timberlake which starts with a verb should not be eligible as concept. As for the idea that concepts should have nouns, it may imply that concepts such as *White House* should not be eligible as well.

2.2 Relevant single-word extractors

Relevant single-word extractors are methods which are specifically tailored to extract single words from texts. These methods can be divided into four different categories: the linguistic approaches; the structure or knowledge-based approaches; the neural net approaches; and the statistical approaches. In this section I will present one or two prominent examples of each category.

2.2.1 Heid – a linguistic-based approach

In [Hei99], the author presents a method for the extraction of candidate single-word terms from German texts. His approach combines linguistic procedures based on pattern matching via regular expressions with a relative frequency comparison.

Before the extraction of candidate terms, the author specifies that the corpora used (German texts) must be preprocessed, specifically by tokenizing (word and sentence boundaries correctly identified), word class annotation (Part-Of-Speech tagging) and then lemmatization (grouping of different inflections of the same base word). The retrieval tool then operates on the previous parsed information, making use of lexical data such as lists of *grammatical words* and of sequence information, implemented as regular expressions over the sequences of characters. Those regular expressions are based on prefixes and suffixes which the author justifies as being more frequent in technical vocabulary than in general language.

Next, the author describes that the regular expressions based on suffixes and prefixes may extract words which are not relevant to his application. He describes that the use of some domain-specific morphemes, or regular expressions, specific to his "car manufacturing" corpus may improve the results, but may lead to over-specialization. The idea is that not all morphemes are usable for the task of relevant word extraction, so the author proposes to extract the best morphemes by comparing the frequency of occurrence of the morphemes in a technical corpora versus a general language corpora. The underlying assumption is that some words will be more frequent in a domain-specific text (as being more relevant for the topics of that text) than in a general, or domain-unspecific text. The most frequent morphemes, given by a predefined threshold, are used as regular expressions for the pattern matchings that follow.

This approach is quite dependent on linguistic tools such as POS tagging, lemmatization, and regular expressions matching. This means that it may be not easily portable to other languages. The author itself presents a section where he assumes the difficulties in adapting some of the tools from English to German. Also, this approach is mainly directed to domain-specific relevant words, as the author clearly states by using morphemes which are specific to the "car manufacturing" domain.

2.2.2 NN – an approach based on Neural Networks

Neural Networks have also been applied on the extraction of relevant single-words from texts. In [DMPPG02] it is presented a method to search for "featured words", which can describe topics of documents, and then find documents which matches user queries.

Their Neural Network model consists of several nodes. Each node is assigned with a word from a user defined search query with pre-assigned equal *energy*. The model then reads an article. The output of the article is a list of single-words obtained from the text. That list includes only the first 200 words of each document of a document-based corpus because, as the authors assume, a word in the title or in the summary of an article is more relevant than a word used in the body text. Next, a stop-word filter is applied, which has the particularity of removing unwanted words such as prepositions or articles ("and", "or", "the", etc.). Then, since words in the title or in the summary part of the document are considered more relevant, different weights are assigned to words accordingly to their place of occurrence.

For each article, if a match between a node (which is a word in the user query) and a word from the article is found, that node is fired and gets a higher energy. The strength of the energy change depends on the weight of the matching word, accordingly to its position of occurrence in the article. This process continues until the Neural Network reaches a state of equilibrium. This happens when no more nodes will change significantly their levels of energy. Finally, having a set of active nodes, the article with the higher energy will contain a larger number of searched words in its word list. This will associate user queries with documents.

Although this work is focused mainly in the search for candidate documents to satisfy user queries, it uses preprocessed lists of single-words. Those single-words may be what we consider as relevant single-words, since the authors use them as terms for identifying documents. However, Neural Networks, with their back-propagation computations, are known for being quite time consuming.

2.2.3 Luhn – frequency criterion

Luhn, in one of the first published papers concerning the extraction of relevant words [Luh58], suggests a method for the classification of words based on the frequency of occurrence of terms. According to the author,

> "... the justification for measuring the relevance of a word by the frequency of occurrence is based on the fact that a writer usually repeats some words when arguing and when elaborates certain aspects of a subject ..."

Luhn also suggests that words with a very high frequency of occurrence are usually considered common words, and words with low frequency of occurrence can be considered rare, both being irrelevant. Although this approach seems intuitive, it is not necessarily true. During my research I noticed that for some specific corpora, considering different languages, among the 100 more frequent words, in average, about 20%–30% could be considered relevant. Table 2.1 lists the relevant words among the 100 more frequent words in an English corpus made of Wikipedia medicine articles:

Table 2.1: Relevant words among the 100 more frequent ones in an English medicine corpus.

Word	Rank	Frequency
medical	34	9093
health	44	6950
patients	54	5715
research	57	5481
treatment	60	5127
disease	62	5048
medicine	65	4489
cells	67	4466
blood	70	4342
time	71	4222
people	78	3831
body	79	3794
study	80	3765
cancer	83	3711
care	85	3694
university	89	3476
patient	91	3350
human	93	3344
studies	95	3338
system	97	3298

Considering the fact that the mentioned corpus has about 10 million words in average, from which about 120.000 are distinct, it can be easily understood that with this criterion some or all of the words listed in Table 2.1 would be thrown away. Luhn's criterion becomes, in this case, quite restrictive. And if we consider the fact that the words in Table 2.1 came from Medicine texts, one can see the kind of the information that would be rejected: words like "medicine" and "health" are quite descriptive of the texts.

Other problem with this approach has to do with the thresholds. How can a threshold between very frequent words and relevant words be found? Or between relevant and rare words? This is a problem because not all words between those thresholds may be important. The author solves this problem partially using a list of common words that should be rejected on the final list. However, Luhn idealized his method for texts with an average of 700 distinct words (scientific papers), but nowadays it would be impracticable to maintain a list of irrelevant words on texts with 100.000 distinct words, for all possible languages and domains.

2.2.4 *TF-IDF* – a statistical approach

Tf-Idf (Term Frequency – Inverse Document Frequency) [SB88] is a statistical metric for calculating the relevance of words in documents. Essentially, this technique measures how important a certain word is on a document regarding other documents in the same collection. Basically, a word is more important in a certain document the more it occurs in that document, but if that word occurs in other documents, its importance decreases. Words that are very frequent on a single document tend to be more valued than common words that occur on more documents, like articles or prepositions.

Formally, being W a word, the importance of W for a document d_j in a corpus $\mathcal{D}$, it is defined by:

$$Tf\text{–}Idf(W, d_j) = Tf(W, d_j) \,.\, Idf(W, d_j) = \frac{f(W, d_j)}{size(d_j)} \,.\, \log \frac{\|\mathcal{D}\|}{\|\{d : W \in d\}\|} \,. \tag{2.1}$$

In equation 2.1, $\|\mathcal{D}\|$ means the number of documents on corpus $\mathcal{D}$; $\|\{d : W \in d\}\|$ is the number of documents containing term W and $size(d_j)$ the number of words on the document d_j. To prevent bias towards longer documents, probability $(f(W, d_j)/size(d_j))$ of term W in document d_j is commonly used instead of the absolute frequency $(f(W, d_j))$.

However, it must be considered that the main goal of *Tf-Idf* is to analyze the relevance of a word in a document regarding other documents, and not to analyze the relevance of a word in a corpus. A slight modification was made in an experiment in the context of this thesis, so that the relevance of a word could be obtained from a corpus: the score of each word was given by the maximum *Tf-Idf* value.

Unfortunately, *Tf-Idf* presents some problems for this task. It harms the relevant words that are relatively frequent because they tend to exist in a significant amount of documents. On the other hand, the *Idf* component also harms some words, specifically by not taking into account the distribution of the frequency of occurrence of a word in the documents. For instance, a word occurring 100 times on one document and just 1 time in another document gets the same *Idf* value that it would get if the distribution was 100 times in the first document and 100 times in the second one, or any other distribution as long as the number of documents having that word was the same. Finally, the *Idf* component may also have the problem of benefiting rare words, where, for instance, unique orthographic errors get the maximum *Idf* value.

2.2.5 Zhou2003 – another statistical approach

Zhou2003 is a metric proposed by Zhou and Slater [ZS03] for calculating the relevance of single-words in a text. It assumes that relevant words can be found in certain areas of the texts, either by being part of local topics or by being related to local contexts, therefore forming clusters in those areas. On the other hand, common and less relevant words should occur randomly in all the text, not forming significant clusters. This technique measures the relevance of a word according to the position of occurrence of each word in the texts.

For a word w, the authors start with a list $L_w = \{-1, t_1, t_2, \ldots, t_m, n\}$, where t_i represents the position of the i-th occurrence of word w in the text and n represents the total number of words in the same text. Then, they obtain $\hat{u}$, which is basically the average separation between consecutive occurrences of word w for the case of uniform distribution of the occurrences.

$$\hat{u} = \frac{n+1}{m+1} \ . \tag{2.2}$$

The next step consists of the calculation of the average separation between real consecutive occurrences of the word w in the text; 3 consecutive occurrences are used for each calculation:

$$d(t_i) = \frac{t_{i+1} - t_{i-1}}{2} \qquad i = 1, 2, \ldots, m \ . \tag{2.3}$$

Then the approach identifies the points on L_w that form part of clusters. Basically a point forms part of a cluster if its average distance $d(t_i)$ is less than the average distance between occurrences for the case of the uniform distribution ($\hat{u}$). This way, $\delta(t_i)$ (equation 2.4) is obtained to identify which points t_i belong to clusters. In a parallel way, $v(t_i)$ (equation 2.5), which represents the local excess of words on position t_i, is also obtained. $v(t_i)$ basically measures the normalized separation to the average distance $\hat{u}$.

$$\delta(t_i) = \begin{cases} 1 & \text{if } \delta(t_i) < \hat{u} \\ 0 & \text{otherwise} \end{cases} \ . \tag{2.4}$$

$$v(t_i) = \frac{\hat{u} - d(t_i)}{\hat{u}} \ . \tag{2.5}$$

Finally, the score of the word w is measured by equation 2.6. Being the information about whether t_i belongs or not to a cluster in $\delta(t_i)$ and in $v(t_i)$ the normalized separation to the average distance, $\Gamma(w)$ has the value of $v(t_i)$ when t_i belongs to a cluster and zero otherwise.

$$\Gamma(w) = \frac{1}{m} \sum_{i=1}^{m} \delta(t_i).v(t_i) \ . \tag{2.6}$$

Although this is a very efficient and ingenious method to implement, it has also some problems regarding the very frequent relevant words. In fact, it harms the relevant words that are relatively frequent because they tend to occur throughout the texts and not only on local contexts. Also, by dealing exclusively with significant clusters, relevant words with low frequency of occurrence are also very harmed by this method.

2.2.6 Islands – yet another statistical extractor

The Islands extractor was developed in the context of my Master's Thesis [VS07]. It presents several statistical metrics and methods for calculating the relevance of single-words in corpora, as well as a method for the automatic extraction of the most relevant words.

The underlying idea of this work is that relevant words have a special preference to relate with a small group of other words. Having this in mind, I proposed two metrics to calculate the score of a word w based on the relations with its successor words (all words occurring right after w – equation 2.7) and with its predecessors (all words occurring just before w – equation 2.8).

$$Sc_{\text{suc}}(w) = \sqrt{\frac{1}{\|\mathcal{Y}\| - 1} \sum_{y_i \in \mathcal{Y}} \left(\frac{p(w, y_i) - p(w, .)}{p(w, .)} \right)^2} \ . \tag{2.7}$$

$$Sc_{\text{pre}}(w) = \sqrt{\frac{1}{\|\mathcal{Y}\| - 1} \sum_{y_i \in \mathcal{Y}} \left(\frac{p(y_i, w) - p(., w)}{p(., w)} \right)^2} \ . \tag{2.8}$$

$$p(w, .) = \frac{1}{\|\mathcal{Y}\|} \sum_{y_i \in \mathcal{Y}} p(w, y_i) \qquad p(., w) = \frac{1}{\|\mathcal{Y}\|} \sum_{y_i \in \mathcal{Y}} p(y_i, w) \qquad p(a, b) = \frac{f(a, b)}{N} \ . \tag{2.9}$$

$\mathcal{Y}$ is the set of words in the corpus, $\|\mathcal{Y}\|$ stands for its size and N is the number of words occurred in the corpus. $f(a, b)$ is the frequency of occurrence of the 2-gram (a, b) in the same corpus. The final score is given by $Sc(w)$:

$$Sc(w) = \frac{Sc_{\text{pre}}(w) + Sc_{\text{suc}}(w)}{2} \ . \tag{2.10}$$

After analyzing the words which were considered relevant, I proposed a metric based on the number of syllables of a word. The underlying idea is that it exists more words with 2, 3 and 4 syllables (depending on the language) than words with other number of syllables. As such, this class of words contains more semantical diversity. By applying the syllable analysis to the *score* of a word, Precision and Recall results of this metric were improved by an average of 20%.

However, words are only ranked in terms of how relevant they are relatively to each other. Some may be more relevant in some areas of the texts than their score can hint, so I've presented a method (the *Islands method*) which would extract the *local* relevant words. The idea of the *Islands* method is that a word w is relevant if it scores consistently higher than its immediate neighbors. If $r(w)$ is the score of a word given by $Sc(w)$ (or $Sc(w)$ with the syllable analysis), the relevance of w is given by equation 2.13.

$$Avg_{\text{pre}}(w) = \sum_{y_i \in \{\text{predecs of } w\}} p(y_i, w) \, . \, r(y_i) \; . \tag{2.11}$$

$$Avg_{\text{suc}}(w) = \sum_{y_i \in \{\text{succecs of } w\}} p(w, y_i) \, . \, r(y_i) \; , \tag{2.12}$$

$$\text{Relevance}(w) = \begin{cases} 1 & r(w) \geq 0.9 \times \max(Avg_{\text{pre}}(w), Avg_{\text{suc}}(w)) \\ 0 & \text{otherwise} \end{cases} \; . \tag{2.13}$$

The problem of this approach is that is only analyzes the immediate successors and predecessors of a word. Therefore, not all relevant words which are part of multi-words are correctly extracted. Furthermore, the syllable analysis tends to ignore small relevant words (like acronyms, such as "RAM", "ROM", "FBI", etc.) and larger relevant words (such as "electroencephalograph" or "otorhinolaryngology"). Large relevant words tend to be highly specific concepts, and may be of possible use for some applications.

2.3 Multi-word relevant expression extractors

Multi-word relevant expression extractors are methods which are specifically tailored to extract meaningful multi-words – sequences of 2 or more words – from texts. These sequences are also known as Multi-word Expressions (MWE) or Multi-word Units (MWU) and they include sequences having a idiosyncratic meaning, i.e., not assembled from the composition of the words in it (such as "raining cats and dogs"), and also sequences where their meaning may be taken from the semantics of each word in the MWE (such as "president of Pakistan"). Whatever the type, MWEs are expected to have a strong meaning.

The methods to extract MWEs can be divided into four different categories: the linguistic approaches; the structure or knowledge-based approaches; the neural net approaches; and the statistical approaches. In this section I will present one or two prominent examples of each category.

2.3.1 Hindi – a linguistic approach

The work of Sinha in [Sin11] is an approach which uses linguistic knowledge to extract MWE from the texts. In this specific work, the author is interested in extracting multi-words from Hindi texts, by applying a set of linguistic rules mostly specific for the Hindi language. The approach of Sinha starts by identifying sentence boundaries. Then, he makes a Part-Of-Speech tagging followed by a morphological analysis. Then Sinha applies a sequence of steps.

The first step is the identification of acronyms and abbreviations containing dots.

Acronyms and abbreviations in Hindi differ from Western languages (for instance, "Mohandas Karamchand Gandhi" may be abbreviated as "ma. ka. gaandhii", "mo. ka. gaandhii" or "ema. ke. gaandhii"). The identification of acronyms and abbreviations, with dots, is carried out using a rule base approach.

The next step is the Hindi chunker and verb-phrase form separation. Chunking is a process of performing shallow parsing of the sentence, where the words having affinity with each other at a syntactic level are grouped together. Since Hindi is a verb ending language, a finite state machine (FSM) is designed in a way such that it starts scanning the words from the rear end (right to left) for possible inclusion in the verb group, based on the POS tags and the morphemes.

The following step is the identification of replicating words and doublet class. Hindi, as other South-Asian languages, has replicating words which are used to emphasize an idea. For instance, "baRii baRii", which can be literally translated as "big big" in English, means in fact "quite big". As for doublets, they are pairs of words which are antonyms or synonyms/hyponyms of each other. An example for pairs of antonyms can be "din-raat" ("day night") which means "all the time" in English, while for synonyms can be "betaa-betii" ("son daughter"), meaning "family issues". Replicating words are identified using syntactic patterns and each word on a doublet is also identified as antonym or synonym using WordNet.

Next, it follows the identification of *vaalaa* morphemes. *Vaalaa* are multi-words which contain one word of the form "vaalaa", "vaalii", "vaale" or "vaalo.M", such as "*jaane vaalaa*" ("go *vaalaa*", that is, "about to go" in English) or "*doodh vaalii balti*" ("milk *vaalii* bucket", that is, "bucket filled with milk").

The next step is the identification of complex predicates and compound verbs. A complex predicate is a MWE where a noun, a verb or an adjective is followed by a light verb, and it behaves as a single verb unit. Some examples are "daan denaa" ("donation give" meaning "to donate" in English) and "mukka maaranaa" ("fist kill/beat" which means "to punch"), for which "denaa" and "maaranaa" are the *light verbs*. Sinha uses a list with 30 *light verbs*.

Then, it follows the identification of acronyms with no dots, such as "beejepii", which is the acronym of "Bharatiya Janata Party" without dots, from the first English characters. Finally, the last step is the identification of named-entities, for which it is used an in-house named-entity recognizer.

This work is a good example of an approach that uses linguistic knowledge in such an intensive way, that the rules are only applicable for the Hindi language itself.

2.3.2 Fips – another linguistic approach

The work in [WSN10] is a linguistic approach for the extraction of collocations. Collocations are sequences of words that co-occur more often than would be expected by chance. Examples of collocations are "crystal clear" or "cosmetic surgery". So, collocations may be considered a subset of the MWEs.

As the authors justify, previous linguistic extractors work by identifying collocations in a specific syntactic configuration, like (*Verb*, *Name*), and not defined in terms of linear proximity, as most statistical approaches usually do. This process is mostly made by a parser, and the identification of the collocations are made after the parsing process. The authors of this work propose that since collocations are made of frequently used and highly ambiguous terms, the identification of collocations should occur during the parsing process and not after, because this can help with the reduction of lexical ambiguities.

Fips is a grammar-based parser which uses left attachment and right attachment rules to build respectively left sub-constituents and right sub-constituents. The idea is that when a grammar rule is triggered in the text, the collocation procedure is invoked. This collocation procedure first verifies that both words of the collocation are associated in a lexical database to one or several collocations. Then, it searches the database for a collocation with both terms following a certain lexical pattern.

Similarly to other linguistic approaches already reviewed, this work is also quite dependent on the usage of POS taggers, parsers, grammars and lexical databases.

2.3.3 HELAS – a multi-word hybrid extractor

In his work *Multiword unit hybrid extractor* [Dia03], Dias presents a hybrid approach for the extraction of MWEs.

The author starts with a Part-Of-Speech tagged corpora. This POS tagged corpora is then divided into two sub-corpora: one containing words and the other containing POS tags. Each sub-corpus is then segmented into a set of positional n-grams. A positional n-gram is a vector of words in the form $[p_{11}, u_1, p_{12}, u_2, \ldots, p_{1n}, u_n]$ where u_i is any word in the positional n-gram and p_{1i} is the distance between word u_1 and word u_i. These positional n-grams allows the representation of non-contiguous multi-word expressions. The segmentation into positional n-grams of the sub-corpora allows to associate a positional n-gram of a word with the positional n-gram of its Part-Of-Speech counterpart. Having both sub-corpora referencing each other, the author then merges both into a custom-made positional n-gram notation (of the form $\left[p_{11}, u_1, \text{POS-tag}_1, \ldots, p_{1n}, u_n, \text{POS-tag}_n\right]$).

The following step is the evaluation of the cohesion between all the textual units contained in a positional n-gram, based on the concept of Normalized Expectation (NE) and relative frequency. The basic idea of the Normalized Expectation is to measure the cost of the loss of one element in the positional n-gram. For $f(.)$ being the frequency of a positional n-gram, NE is given by:

$$NE([p_{11}, u_1, \ldots, p_{1i}, u_i, \ldots, p_{1n}, u_n]) = \frac{f([p_{11}, u_1, \ldots, p_{1i}, u_i, \ldots, p_{1n}, u_n])}{\frac{1}{n}\left(f([p_{22}, u_2, \ldots, p_{2i}, u_i, \ldots, p_{2n}, u_n]) + \sum_{i=2}^{n} f([p_{11}, u_1, \ldots, p_{1i}, u_i, \ldots, p_{1n}, u_n])\right)}. \quad (2.14)$$

Since the author assumes that the average cost of the loss of an element, given by equation 2.14, is not sufficient, he uses a Mutual Expectation variant (equation 2.15) to refine the results. In practice, the author uses the Mutual Expectation to weight $NE(.)$ by the relative frequency of occurrence of the positional n-gram, mainly because there may be two positional n-grams with the same Normalized Expectation. It is given by:

$$\begin{array}{l} ME([p_{11}, u_1, \ldots, p_{1i}, u_i, \ldots, p_{1n}, u_n]) = \\ p([p_{11}, u_1, \ldots, p_{1i}, u_i, \ldots, p_{1n}, u_n]) \,.\, NE([p_{11}, u_1, \ldots, p_{1i}, u_i, \ldots, p_{1n}, u_n]) \end{array} . \tag{2.15}$$

where $p(v)$ measures the probability of occurrence of vector v. This allows Dias to get the most frequent and cohesive positional n-grams. Thus, by including POS data, Dias claims that the cohesiveness of words and the degree of cohesiveness with its associated POS tags may allow us to identify MWEs. The combination of both factors is expressed in equation 2.16.

$$\begin{array}{l} CAM([p_{11}, u_1, t_1, \ldots, p_{1i}, u_i, t_i \ldots, p_{1n}, u_n, t_n]) = \\ ME([p_{11}, u_1, \ldots, p_{1i}, u_i, \ldots, p_{1n}, u_n])^{\alpha} \,.\, ME([p_{11}, t_1, \ldots, p_{1i}, t_i, \ldots, p_{1n}, t_n])^{1-\alpha} \end{array} . \tag{2.16}$$

In equation 2.16, t_1, t_i and t_n are the corresponding POS tags. Variable α allows Dias to choose whether the process should be more oriented towards the cohesiveness of words or of POS tags. Finally, having *CAM* scores for all positional n-grams, the most relevant n-grams are selected using the GenLocalMaxs algorithm. The GenLocalMaxs algorithm is quite similar to LocalMaxs algorithm (section 2.3.6), where the underlying idea is that a n-gram is relevant if it scores higher than its neighbor n-grams.

This approach is quite similar to the work presented in section 2.3.6, although it is adapted for non-contiguous multi-word expressions. Results of this approach are variable, whether we are dealing with 2-grams up to 6-grams, but average Precision results seems to be around 60%.

2.3.4 TEG – another hybrid approach

TEG (Trainable Extraction Grammar) [FRF06] is a hybrid approach for the extraction of entities and relations at the sentence level, which combines a knowledge-based approach with a statistical machine-learning approach. The system is based on stochastic context-free grammars for which the rules of extraction are manually written.

The idea is that for each corpus for which information is to be extracted, entities and semantic relations can be described by means of a context-free grammar. For a specific experiment, the authors started by manually writing the extraction rules and tag the documents. A *TEG* rulebook consists of declarations and rules which basically follow the classical grammar rule syntax, with a special construction for assigning concept attributes. These concepts are entities, events and facts that the system is designed to extract, but two classes of symbols require further declaration: *termlists*, which are collections of terms from the same semantic categories, such as country names, cities, states, genes, proteins; and n-grams. The following shows an example of such rules:

```
termlist TLHonorific = Mr Mrs Miss Ms Dr;
(1) Person :- TLHonorific NGLastName;
(2) Person :- NGFirstName NGLastName;
(3) Person -> IsFriend Person;
(4) Text :- NGNone Text;
(5) Text :- Person Text;
(6) Text :- ;
```

In this example, the written rules are specific for a grammar to extract names of persons. To further improve the efficiency of this method, the authors train the grammar on a tagged corpus. The idea is that some rules are more *important* than others. That importance is given by the frequency for which each rule is "fired" in the training data. Each rule is then rewritten with the probability of occurrence on the training data and finally the grammar is set to extract the entities and relations for which it was trained.

The main problem of this approach is that it tends to be very domain-specific. For instance, to extract names of persons, a set of rules is written, but to extract names of companies, another set of rules has to be written. This makes the usage of this approach rater laborious each time the domain changes, because patterns must be changed whether names of persons or of companies are to be extracted.

2.3.5 Mutual Information, Chi-squared, Phi-squared – statistical metrics

MI, χ^2 and Φ^2 are metrics used in some statistical approaches for the extraction of MWEs, mostly collocations. These statistical metrics measure the tendency for a pair of words on a 2-gram to co-occur in sequence. When the 2-grams on a text are ranked by the score of one of these measures, the application of a threshold filter may eventually be used in order to find a possible separation between the MWEs and the non-MWEs.

Mutual Information

The original *Mutual Information* metric [Sha48] is mostly used to measure the uncertainty between two random variables. To measure the degree of "cohesion" between a pair of words, Church & Hanks proposed the *association ratio* metric in [CH90]. The *association ratio* is commonly known as *Mutual Information* or *Specific Mutual Information* in Computational Linguistics. Its expression is as follows:

$$MI(x\ y) = \log_2 \frac{p(x\ y)}{p(x).p(y)}\ . \tag{2.17}$$

$$p(x\ y) = \frac{f(x\ y)}{N-1} \qquad p(x) = \frac{f(x)}{N} \qquad p(y) = \frac{f(y)}{N}\ .$$

Functions $f(x)$ and $f(y)$ give the frequency of occurrence of the single-words in the texts and $f(x\ y)$ returns the frequency of occurrence of the 2-gram $x\ y$, i.e., of x occurring in a position i while y occurs in position $i+1$. N stands for the total number of words in the corpus. Although this metric returns good results for highly co-occurring pairs of words, it also benefits rare pairs. In fact, lets us suppose that a 2-gram occurs with the same frequency n as their unigrams x and y, namely, $f(x) = f(y) = f(x\ y) = n$. Then, assuming a big corpus, such that $N \gg 1$,

$$MI(x\ y) = \log_2 \frac{p(x\ y)}{p(x).p(y)} = \log_2 \frac{\frac{f(x\ y)}{N-1}}{\frac{f(x)}{N}.\frac{f(y)}{N}} \approx \log_2 \frac{\frac{n}{N}}{\frac{n}{N}.\frac{n}{N}} = \log_2 \frac{N^2.n}{N.n^2} = \log_2 \frac{N}{n} \ .$$

This shows that when n is low and N is high, $MI(.)$ values are also high. So, rare 2-grams are favored by this metric, especially when they occur once ($MI(x\ y) = \log_2 N$), for instance, orthographic errors.

Chi-squared

$\chi^2(.)$ is a statistical metric based on Pearson's coefficient [Pea00]. For the extraction of multi-words, this metric is used as follows:

$$\chi^2(x\ y) = \frac{N.\left(f(x\ y).f(\neg x\ \neg y) - f(x\ \neg y).f(\neg x\ y)\right)^2}{f(x).f(y).f(\neg x).f(\neg y)} \ . \tag{2.18}$$

As in the previous metric, $f(x\ y)$ measures the frequency of occurrence of the pair $x\ y$. $f(\neg x\ y)$ measures the frequency of occurrence for the cases when x does not occur before y and $f(x\ \neg y)$ measures the frequency of the cases when y does not occur after x. $f(\neg x\ \neg y)$ measures the frequency of 2-grams having neither x nor y. To find a threshold capable of separating relevant 2-grams from non-relevant ones, the χ^2 test is usually used. However, the χ^2 test is only applicable when the frequency of occurrence of the 2-gram is greater than 5, or else it cannot be considered valid. This makes the $\chi^2(.)$ measure and the χ^2 test unusable for a great number of 2-grams in the texts.

Phi-squared

$\Phi^2(.)$ is a statistical metric based on the $\chi^2(.)$. It was proposed in [CG91] to rank pairs of parallel texts.

$$\Phi^2(x\ y) = \frac{\left(f(x\ y).f(\neg x\ \neg y) - f(x\ \neg y).f(\neg x\ y)\right)^2}{f(x).f(y).f(\neg x).f(\neg y)} \ . \tag{2.19}$$

It is similar to $\chi^2(.)$, however divided by N. Unlike the $\chi^2(.)$, $\Phi^2(.)$ has the advantage of always returning values between 0 and 1 independently of the size of the corpus. But like $\chi^2(.)$, $\Phi^2(.)$ is strictly for 2-grams since both can not measure cohesions for more than 2 words.

2.3.6 LocalMaxs – a statistical approach

LocalMaxs is an algorithm presented in [SL99] for the extraction of MWEs from large corpora. Although Localmaxs may be used to extract other elements from texts, such as characters or morphosyntactic tag patterns, it is mostly used for the extraction of multi-words.

LocalMaxs, such as the metrics described in the previous section, is based on the idea that each n-gram has a kind of *glue* or cohesion between the words within the n-gram. Different n-grams usually have different cohesion values. For instance, there is a strong cohesion between the words "Alfred" and "Nobel" (forming the 2-gram "Alfred Nobel"), but not a strong cohesion within "or uninterrupted" or "of two". For the calculation of the internal cohesion of a generic 2-gram the authors propose $SCP(x\ y)$ which is given by:

$$SCP(x\ y) = p(x|y)\,.\,p(y|x) = \frac{p(x\ y)}{p(y)}.\frac{p(x\ y)}{p(x)} = \frac{p(x\ y)^2}{p(x).p(y)}\ . \tag{2.20}$$

$p(x)$ and $p(y)$ are the probabilities of occurrence of words x and y, while $p(x\ y)$ is the probability of occurrence of the 2-gram $x\ y$. However, to measure the cohesion of n-grams larger than 2-grams, the authors propose the $SCP\_f(w_1 \dots w_n)$ which is based on the idea of the Fair Dispersion Point Normalization and can be considered a generalization of equation 2.20.

$$SCP\_f(w_1 \dots w_n) = \frac{p(w_1 \dots w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} p(w_1 \dots w_i)\,.\,p(w_{i+1} \dots w_n)}\ . \tag{2.21}$$

Finally, for the extraction of MWEs, the authors present LocalMaxs. The idea behind LocalMaxs is that a multi-word should be considered relevant if its cohesion value is greater than the average of two maxima: the greatest cohesion value found in the contiguous (n-1)-grams contained in the n-gram, and the greatest cohesion value found in all contiguous (n+1)-grams which contain the n-gram. In a formal way, a sequence $W = (w_1 \dots w_n)$ is a MWE if and only if:

$$\text{for } \forall x \text{ in } \Omega_{n-1}(W),\ \forall y \text{ in } \Omega_{n+1}(W)$$
$$(\text{length}(W) = 2 \wedge g(W) > y) \vee (\text{length}(W) > 2 \wedge g(W) > \frac{x+y}{2})$$

Being $g(W)$ the value of $SCP\_f(W)$, $\Omega_{n-1}(W)$ and $\Omega_{n+1}(W)$ respectively the set of $g(.)$ values of all contiguous (n-1)-grams contained in the n-gram, and all contiguous (n+1)-grams which contain the n-gram, and length(W) the number of words in W. Thus, LocalMaxs extracts MWEs whose cohesion values form local maxima in the texts.

Although LocalMaxs is a statistical and language-independent method, it does not present high Precision and Recall values. Essentially, the recall is low for texts written in languages where the relevant units lie significantly on single-words, such as German and Dutch.

2.4 Summary of the related work

In a general way, extractors that are focused on the extraction of concepts tend to use language-specific or domain-specific tools. For instance, *CICM* (subsection 2.1.1) uses lexical patterns specific for Chinese as also an external lexicon (*HowNet*) to generate more lexical rules. *GARAGe* (subsection 2.1.2) uses another external lexicon (Wordnet), and *DIPRE* (subsection 2.1.3) uses predefined patterns to extract domain-specific concepts such as names of authors and titles of books. Other extractors, such as *KOSMIX* (subsection 2.1.4) mixes statistics with linguistics. By using POS taggers, these approaches are highly language-dependent, since not all languages have high quality linguistic tools.

As for single-words, the linguistic approaches tend to have the same language dependency problems as the concept extractors. POS tagging, lemmatization, and regular expressions matching, limit the usage of methods such as the one described in subsection 2.2.1, for other languages than German. On the other hand, approaches using Neural Networks, such as the one described in subsection 2.2.2, are known for being time-consuming mainly because of the calculation of the back-propagation.

As for statistical methods, Luhn's frequency criterion (subsection 2.2.3), although language-independent, is too simplistic. Not all frequent words are function words, and most rare words are indeed relevant. This poses some difficulties in setting thresholds. *Tf-Idf* (subsection 2.2.4), similar to Luhn's frequency criterion, also tends to harm the relevant words that are relatively frequent, while benefiting the rare ones (such as orthographic errors). Also, the *Idf* component is insensitive to the distribution of the frequency of a word in the documents. The method of Zhou et al. (subsection 2.2.5), by assuming that relevant words always make part of clusters, tends to harm the relevant words that are relatively frequent, as also the rare relevant words. Finally, the Islands method (subsection 2.2.6) tends to fail for words which are part of relevant multi-words, because it assumes that a relevant word has to score consistently higher than the immediate neighbors (predecessor and successor words). Also, the syllable analysis, which complements the $Sc(.)$ measure, ignores both small relevant words as well as the larger ones. Larger words tend to be highly specific concepts.

As for the multi-word extractors, linguistic and hybrid approaches also tend to be highly language or domain dependent. For instance, a highly complex set of linguistic rules for Hindi language is used in the work described in subsection 2.3.1. For *Fips* (subsection 2.3.2), POS taggers, parsers, grammars and lexical databases are used. For *HELAS* (subsection 2.3.3), although it uses statistics, the POS tagged corpora imposes a language dependency, while for *TEG* (subsection 2.3.4), the dependency is on the domain, since the manual creation of the grammar makes the changing of domains an extreme laborious process.

As for the multi-word extractors based on statistical metrics (subsection 2.3.5), since they use plain text corpora and only require the information appearing in texts, such systems are highly flexible and able to extract relevant units independently from the domain and the language of the input text. However, they have two major drawbacks: they rely on ad hoc establishment of global thresholds which are prone to error and only allow the acquisition of binary associations. *LocalMaxs* (subsection 2.3.6) circumvents those problems: the generalization of $SCP(.)$ to $SCP\_f(.)$ allows the extraction of multi-words greater than 2-grams, and it also provides a mechanism for inferring relevant n-grams from the analysis of the neighborhood, eliminating the necessity of thresholds. However, it does not present high Precision and Recall values.

In the next chapters I will try to answer to some of the challenges which are implicit on what I have just exposed. Mainly, I propose a method capable of extracting both single-word and multi-word concepts which is language and domain independent.

3

The *ConceptExtractor* approach

The *ConceptExtractor* is a statistical methodology for the extraction of single-word and multi-word concepts from texts. Since this thesis if focused mainly in the Text-Mining area, this chapter starts with an empirical definition of concepts in the context of this work. The main purpose is to demonstrate that there are specific relations between concepts which can be explored using a statistical approach. The latter sections will present the *ConceptExtractor* with greater detail.

3.1 An empirical approach to concepts

In a general way, a concept can be defined empirically as a word or a sequence of words which possess some semantic value. For instance, while words such as "president" and "republic" can be considered concepts, words such as "and", "of" and "or" do not have much of a semantic value. The former words possess some intrinsic semantic value, they have a meaning and convey an idea, while the latter belong to the class of function words and do not have any significant meaning. However, not all content words (nouns, most verbs, adjectives and adverbs) should be considered concepts because, as it will be shown, it may be essentially a matter of *degree*.

3.1.1 Compound concepts

Concepts, on its most basic form, are made of single-words. For instance, "president" is a concept, meaning essentially a leader, and "republic" is another concept, a specific form of governance. Both, isolated, have their own meanings.

But concepts may be formed by more than one word. For instance, the aggregation of "president" and "republic" forms a new compound concept "president of the republic". This compound concept is more specific than the single-word concepts which form it. In fact, we are not referring to any *president*, but specifically to the *president* of the *republic*. From the point of view of *republic*, we are not referring to any republican institution or representative, but specifically to its *president*.

So, apart from the non-compositional expression cases such as "hot dogs" and "raining cats and dogs", which have an idiomatic meaning, compound concepts are usually specializations of the single-word concepts that form it.

3.1.2 Edges of compound concepts

Another empirical property of concepts is that compound concepts tend to start and finish with single-word concepts, even when they are composed of only two words. The rationale is that the inclusion of function words in the edges of compound concepts causes an impression of incompleteness to the multi-word, as if some other concept should follow and complete it. This happens because function words provide the connection to other words. The following tables (tables 3.1, 3.2 and 3.3) present some multi-words from different languages.

Table 3.1: Some multi-words from an English corpus.

Multi-word
Autistic enterocolitis
Magnetic field imaging
Hopkins Center for Health Disparities Solutions
Medical Society of London
University of
using children in
by the
in case of

Table 3.2: Some multi-words from a Portuguese corpus.

Multi-word
Abastecimento público de água
Abdómen humano
Patologia clínica
Escola Portuguesa de Angiografia
e o aborto
Síndrome de
por causa de
da angústia respiratória do

Table 3.3: Some multi-words from a German corpus.

Multi-word
Abbott Laboratories
Charles Drelincourt der Jüngere
Cerebrale Bewegungsstörung
Homöoboxprotein DLX-3
Museum für Verhütung und
Psychotherapie in
im Fall von
Tuberkulose der

In each table, the first four examples represent compound concepts. The last four are not compound concepts, since they either start or end with a function word.

3.1.3 Tendency for fixed distances

Another empirical property of concepts is that the single-word concepts in compound concepts tend to be semantically related. In this thesis I explore that fact by measuring the tendency for a pair of single-words to co-occur in fixed positions relatively to each other. Consider the following tables (tables 3.4, 3.5 and 3.6) which present some pairs of words occurring in compound concepts and the frequency of occurrence of those pairs, for different relative positions between the words.

Table 3.4: Co-occurrence frequency of word pairs for different relative positions in an English corpus.

Pair	Multi-word	Frequency by relative position
(abortion, surgical)	surgical abortion	[0, 0, 0, **14**, abortion, 0, 1, 0, 0]
(abortion, induced)	induced abortion	[0, 0, 0, **43**, abortion, 0, 1, 3, 3]
(university, minnesota)	university of minnesota	[0, 0, 1, 0, university, 0, **29**, 0, 0]
(brain, implants)	brain implants	[1, 0, 1, 0, brain, **23**, 0, 0, 0]
(human, virus)	human immunodef. virus	[1, 1, 4, 0, human, 1, **25**, 1, 3]

Table 3.5: Co-occurrence frequency of word pairs for different relative positions in a Portuguese corpus.

Pair	Multi-word	Frequency by relative position
(abastecimento, água)	abastecimento de água	[0, 0, 0, 0, abastecimento, 1, **28**, 1, 0]
(aborto, legalização)	legalização do aborto	[0, 0, **26**, 0, aborto, 0, 0, 0, 0]
(etílico, álcool)	álcool etílico	[1, 0, 0, **16**, etílico, 0, 0, 1, 0]
(glândula, salivar)	glândula salivar	[0, 0, 0, 0, glândula, **22**, 0, 0, 1]
(síndrome, asperger)	síndrome de asperger	[0, 0, 0, 0, síndrome, 0, **27**, 0, 0]

Common to all tables is the fact that the pairs of words in compound concepts tend

Table 3.6: Co-occurrence frequency of word pairs for different relative positions in a German corpus.

Pair	Multi-word	Frequency by relative position
(therapie, antiretroviralen)	antiretroviralen therapie	[0, 0, 0, **11**, therapie, 0, 0, 0, 0]
(anatomie, pathologische)	pathologische anatomie	[0, 0, 0, **47**, anatomie, 0, 2, 0, 0]
(medizin, lizentiat)	lizentiat in medizin	[0, 0, **14**, 0, medizin, 0, 0, 0, 0]
(genetische, information)	genetische information	[0, 0, 0, 0, genetische, **12**, 0, 0, 0]
(chirurgie, plastische)	plastische chirurgie	[0, 5, 0, **27**, chirurgie, 0, 0, 1, 1]

to co-occur in fixed positions relatively to each other, forming specific multi-words, even when those multi-words have function words between the single-words. For instance, in Table 3.4, english word "surgical" occurs 14 times just before "abortion" and one time two words after. This comes from the fact that the concept "surgical abortion" occurs 14 times while "abortion by surgical [means]" occurs only once in this corpus. Similarly, "minnesota" occurs 29 times two words after "university" and just one time two words before. In fact, the concept "university of michigan" occurs 29 times while "minnesota state university" only occurs once. This analysis is also applicable to the compound concepts in the other languages (tables 3.5 and 3.6).

3.1.4 Specificity of concepts

Finally, concepts may have several degrees of specificity. If a term (be it a single-word or a multi-word expression) is not *promiscuous*, i.e., if it relates with only a few other terms (considering a limited neighborhood window and a considerable amount of text), there is a high probability that it represents a more specific concept. In fact, it can be easily recognized that terms such as "University" and "University of Minnesota" are both concepts. However, the later is more specific than the former, since it describes a specific *university*. On the other hand, function words such as "the" and "or" tend to relate with many words in English texts, so they are not specific at all. Appendix A shows some classification lists of concepts which illustrates the approach.

3.2 Exploring the tendency for fixed distances

The tendency for compound concepts to have fixed-distances between their single-word concepts is the starting point of the *ConceptExtractor* approach. This tendency is measured as follows:

For an individual word w from a corpus, $B_w = [b_1, b_2, .., b_m]$ is the list of all unique neighbor words of w. Each neighbor b_i occurs at different positions relatively to w, inside a window with size s. Positions of b_i can be positive or negative and are determined by considering that w is at the center of the window. For each pair (w, b_i), a list $X_{(w,b_i)}$ is obtained counting the co-occurrence frequencies by relative distance between w and b_i, such that:

$$X_{(w,b_i)} = [x_{-\frac{s}{2}}, \ldots, x_{-1}, x_1, \ldots, x_{\frac{s}{2}}] \,. \tag{3.1}$$

Thus, x_j is the co-occurrence frequency of word b_i at position j relative to w (examples, for $s = 4$, can be seen in section 3.1.3). Please consider the fact that, although in most examples throughout this thesis, the central word in $X_{(w,b_i)}$ is shown, it is only for illustrative purposes and does not make part of any calculations.

For a given $X_{(w,b_i)}$, the following metric computes the *relative variance* of the distribution of frequencies in $X_{(w,b_i)}$:

$$Rel\_var(X_{(w,b_i)}) = \frac{1}{s(s-1)} \sum_{j=1}^{s} \left(\frac{x_j - \bar{x}}{\bar{x}} \right)^2 \,, \tag{3.2}$$

where x_j is the value of the j-th element of the list $X_{(w,b_i)}$ and s is the length of the list (the size of the window); $\bar{x}$ stands for the average value of the frequencies in $X_{(w,b_i)}$:

$$\bar{x} = \frac{1}{s} \sum_{j=1}^{s} x_j \,. \tag{3.3}$$

It must be noted that, although $X_{(w,b_i)}$ represents a window ranging from $-s/2$ to $s/2$, $Rel\_var(.)$ computes the *relative variance* independently of the order of its elements. Therefore, in equation 3.2, $X_{(w,b_i)}$ is treated as a list ranging from 1 to s.

To better understand the mechanism of $Rel\_var(.)$, figures 3.1 and 3.2 show the distribution of frequencies for two pairs of words which occur in an English corpus – (*allergic*, *reaction*) and (*of*, *reaction*) [1].

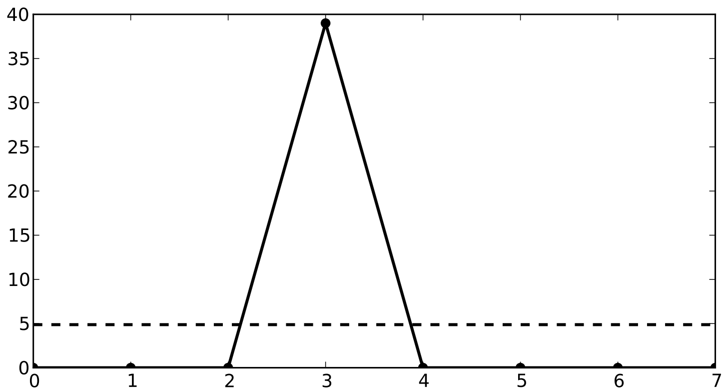


Figure 3.1: Representation of the frequencies of co-occurrence for the pair (*allergic*, *reaction*). $X_{(reaction,allergic)} = [0, 0, 0, 39, reaction, 0, 0, 0, 0]$, $\bar{x} = 4.875$ and $Rel\_var(.) = 1.000$.

[1]Please consider that when a pair is referred, the order of appearance of its elements is technically irrelevant, having no implication. For example, the pair (*allergic*, *reaction*) is the same as (*reaction*, *allergic*). However, in order to promote a quick understanding of some particular mechanism, it may be helpful to present the pair by using a particular order of appearance of its elements.

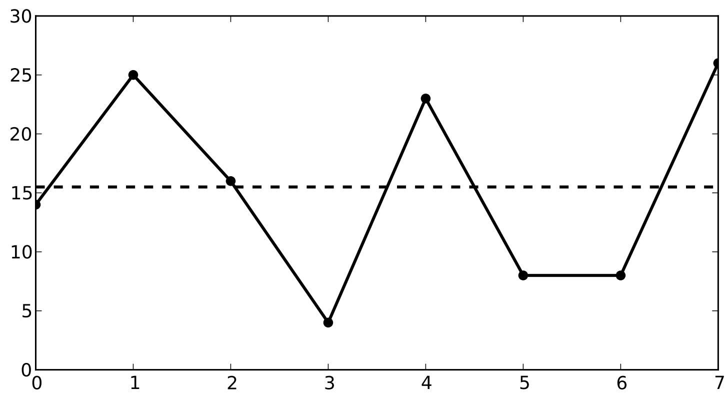


Figure 3.2: Representation of the frequencies of co-occurrence for the pair (*of*, *reaction*). $X_{(reaction,of)} = [14, 25, 16, 4, reaction, 23, 8, 8, 26]$, $\bar{x} = 15.5$ and $Rel\_var(.) = 0.0377$.

$Rel\_var(.)$ measures, essentially, the normalized distances from the points (in this case, frequencies by position) to an average value (the average frequency). These normalized distances are squared so the numbers don't cancel each others. The maximum value of 1.0 is given to lists where all frequencies except one are 0, as for the pair (*allergic*, *reaction*) in Figure 3.1. In this case, there is a clear peak in the position preceding the word "reaction" (from *allergic reaction*), and $Rel\_var(.) = 1.000$. For the pair (*of*, *reaction*) in Figure 3.2, since all frequencies are around the average value, there is no obvious preference for the pair to co-occur in a fixed position, having, thus, a lower $Rel\_var(.)$ value.

So, pairs (w, b_i) which show preference to occur at fixed positions are more valued than pairs which usually occur scattered.

The following tables (tables 3.7, 3.8 and 3.9) show some examples of $Rel\_var(.)$ values for pairs of words extracted from the corpora described in Table 4.1.

Table 3.7: Some $Rel\_var(.)$ values for pairs (b_i, reaction) from an English corpus.

Pair	**Frequency by relative position**	***Rel_var(.)***
(allergic, reaction)	[0, 0, 0, 39, reaction, 0, 0, 0, 0]	1.000
(autoimmune, reaction)	[0, 0, 0, 11, reaction, 0, 1, 0, 0]	0.825
(chemical, reaction)	[0, 1, 0, 10, reaction, 0, 0, 0, 1]	0.666
(adverse, reaction)	[0, 0, 3, 10, reaction, 0, 0, 0, 0]	0.594
(such, reaction)	[1, 1, 2, 1, reaction, 3, 4, 0, 1]	0.080
(of, reaction)	[14, 25, 16, 4, reaction, 23, 8, 8, 26]	0.037
(and, reaction)	[11, 8, 14, 5, reaction, 8, 10, 8, 23]	0.032
(in, reaction)	[5, 13, 8, 7, reaction, 11, 15, 12, 14]	0.014

Analyzing Table 3.7, the first line shows that the word "allergic" tends to occur in a fixed position, in that window, relatively to "reaction", forming the term "allergic reaction". Since it has a clear peak and all other frequencies are 0, this pair has a $Rel\_var(.)$ value of 1.0. For "autoimmune", although it shows a high preference for occurring one

Table 3.8: Some $Rel\_var(.)$ values for pairs (ácido, b_i) from a Portuguese corpus.

Pair	Frequency by relative position	*Rel_var(.)*
(ácido, láctico)	[0, 0, 0, 0, ácido, 19, 0, 0, 0]	1.000
(ácido, úrico)	[1, 1, 0, 0, ácido, 83, 0, 0, 1]	0.922
(ácido, desoxirribonucleico)	[1, 0, 0, 0, ácido, 11, 0, 0, 0]	0.825
(ácido, clorídrico)	[0, 0, 1, 0, ácido, 13, 0, 0, 1]	0.726
(ácido, pela)	[2, 9, 0, 0, ácido, 0, 3, 3, 8]	0.162
(ácido, ser)	[5, 4, 2, 2, ácido, 0, 0, 11, 9]	0.120
(ácido, nos)	[4, 2, 0, 0, ácido, 1, 3, 4, 3]	0.075
(ácido, para)	[8, 3, 5, 3, ácido, 2, 8, 7, 5]	0.026

Table 3.9: Some $Rel\_var(.)$ values for pairs (b_i, chirurgie) from a German corpus.

Pair	Frequency by relative position	*Rel_var(.)*
(orthopädische, chirurgie)	[0, 0, 0, 12, chirurgie, 0, 0, 0, 0]	1.000
(plastischen, chirurgie)	[0, 1, 0, 24, chirurgie, 0, 1, 0, 0]	0.834
(gesellschaft, chirurgie)	[0, 4, 62, 0, chirurgie, 0, 0, 1, 1]	0.811
(facharzt, chirurgie)	[2, 5, 23, 0, chirurgie, 0, 1, 0, 0]	0.522
(war, chirurgie)	[11, 1, 0, 0, chirurgie, 3, 3, 11, 9]	0.127
(er, chirurgie)	[22, 22, 1, 0, chirurgie, 0, 16, 8, 24]	0.104
(im, chirurgie)	[7, 7, 2, 0, chirurgie, 15, 8, 5, 3]	0.077
(die, chirurgie)	[33, 12, 21, 33, chirurgie, 4, 26, 8, 12]	0.045

position before "reaction" ("autoimmune reaction"), the fact that it occurs one time two words after "reaction", makes the $Rel\_var(.)$ value of the pair to be less than 1.0 (it is 0.825). On the bottom of the list, it can be seen that function words show no preference to occur in fixed positions relatively to the center word. Therefore, their $Rel\_var(.)$ values are lower.

Therefore, pairs such as (*allergic*, *reaction*) score higher than pairs such as (*in*, *reaction*), where co-occurrences are more scattered over the positions. Furthermore, since pairs such as (*allergic*, *reaction*) tend to have fixed distances between both words, it is likely that both are single-word concepts as both seem to form a compound concept ("allergic reaction"). On the contrary, the pairs on the bottom of the table score less on $Rel\_var(.)$ due to their more scattered distributions, being less likely to form compound concepts. This analysis is also applicable to the remaining tables (3.8 and 3.9).

However, although the evaluation concerning the fixed relative positions gives us a hint about whether or not two words are likely to be concepts, that still has to be assessed. In this methodology, that is done by measuring the *semantic specificity* (*specificity* for short) of words.

3.3 Specificity of single-word concepts

As mentioned in section 3.1.4, concepts may have several degrees of specificity. In other words, some concepts may have a more or less specific meaning than others. For instance, "arthritis", a disease which affects the joints, is less specific than "gout": there are many different types of arthritis (*osteoarthritis*, *rheumatoid arthritis*, *psoriatic arthritis*, *septic arthritis*, *reactive arthritis*, etc.), and *gout* is one of them. Therefore, by being a specific type of *arthritis*, "gout" is a more specific concept than "arthritis".

To measure the specificity of a word w in a corpus, let $B = [b_1, \ldots, b_m]$ be the list of all m unique words in the corpus. Equation 3.4 represents the distribution of all $Rel\_var(.)$ values that the word w has with all words b_i in B.

$$RDist_w = [Rel\_var(X_{(w,b_1)}), Rel\_var(X_{(w,b_2)}), \ldots, Rel\_var(X_{(w,b_m)})] \ . \tag{3.4}$$

$X_{(w,b_i)}$ is the list of the co-occurrence frequencies of the word b_i near word w (considering a fixed-size window), and $Rel\_var(X_{(w,b_i)})$ is the $Rel\_var(.)$ value for a pair (w, b_i), as in equation 3.2.

Finally, equation 3.5 is used to measure the specificity of w.

$$Spec(w) = Rel\_var(RDist_w) \ . \tag{3.5}$$

The underlying idea about $Spec(w)$ is that, if a single-word w is strongly associated (has higher $Rel\_var(.)$ values) with a few words in the corpus, and weakly associated with the rest of them, then w is a fairly specific concept. This mechanism can be understood by looking at the following figures, which shows the $RDist_w$ distribution for *medicine* (Figure 3.3) and *of* (Figure 3.4), on the English *Medicine* corpus.

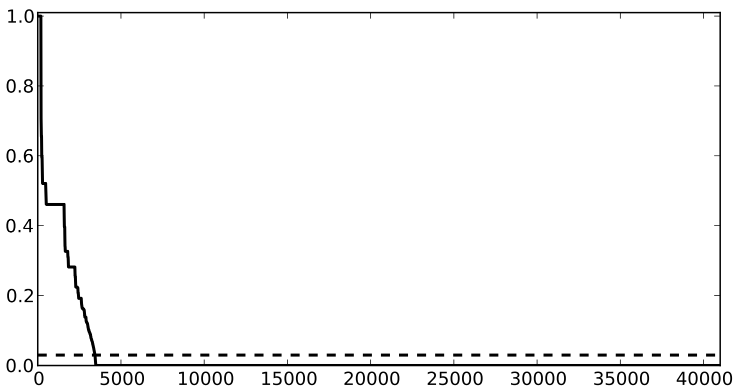


Figure 3.3: Ordered distribution of the $Rel\_var(.)$ values for pairs (*medicine*, b_i).

Figure 3.3 shows that the word *medicine* has high $Rel\_var(.)$ values with a few unique words of the corpus. Then, it has decreasing $Rel\_var(.)$ values until it reaches zero very quickly. In other words, it shows that the word *medicine* relates strongly (in terms of

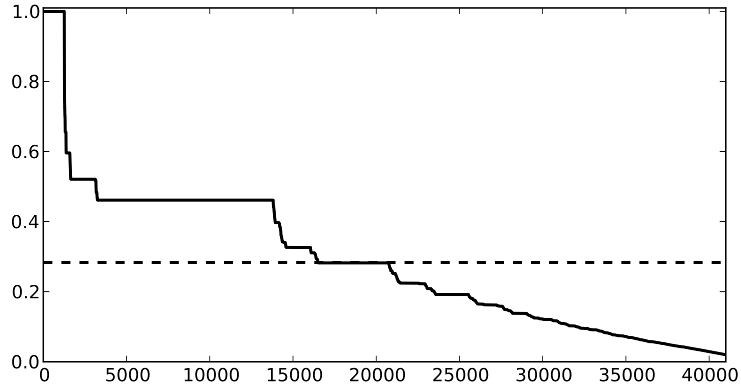


Figure 3.4: Ordered distribution of the $Rel\_var(.)$ values for pairs (*of*, b_i).

fixed-positions) with a few words of the corpus, and then it relates increasingly less and less with all other words of the corpus until it reaches zero – these are words with lower influence over the word *medicine*, and most do not occur near *medicine* at all.

On the other hand, in Figure 3.4, the $Rel\_var(.)$ values for the word *of* decreases very slowly. Basically, the word *of* maintains the tendency for having fixed-distance relations with much more words than *medicine*. Since $Rel\_var(.)$ (equation 3.2) measures the tendency for the occurrence of "peaks" in lists of numerical values, the $Rel\_var(.)$ value for the distribution in Figure 3.3 (*medicine*) is greater than the $Rel\_var(.)$ value for the distribution in Figure 3.4 (*of*).

The following tables (tables 3.10, 3.11 and 3.12) show some examples of $Spec(.)$ values for the same words translated into three different languages, corresponding to the three different test corpora used. As reference, the column *number of pairs (w,b_i)* on the tables measure the number of pairs (w,b_i) for which $Rel\_var(X_{(w,b_i)}) > 0$.

Table 3.10: Specificity of some words from the English corpus.

w	**# of Pairs (w, b_i)**	***Spec(w)***
gout	82	1.51×10^{-2}
arthritis	315	4.42×10^{-3}
inflammation	538	2.49×10^{-3}
in	34711	3.33×10^{-5}
of	41438	2.56×10^{-5}
the	55259	1.95×10^{-5}

Even though the three test corpora are not made of parallel translated texts (they are made of random Wikipedia documents from the *medicine* category), it can be seen that the relative specificity of the words are consistent for the three different languages. In fact, the word "gout" ("gota" in Portuguese and "gicht" in German), seems to be more specific than the rest – in each corpus it is the one which co-occurs with less words and scores higher than the rest. Furthermore, considering the translations, each word in the

Table 3.11: Specificity of some words from the Portuguese corpus.

w	**# of Pairs** (w, b_i)	$Spec(w)$
gota	121	1.06×10^{-2}
artrite	232	5.68×10^{-3}
inflamação	551	2.40×10^{-3}
em	30378	3.82×10^{-5}
o	37818	2.84×10^{-5}
de	58536	1.76×10^{-5}

Table 3.12: Specificity of some words from the German corpus.

w	**# of Pairs** (w, b_i)	$Spec(w)$
gicht	53	2.45×10^{-2}
arthritis	214	6.32×10^{-3}
entzündung	407	3.30×10^{-3}
von	32863	3.66×10^{-5}
in	44196	2.61×10^{-5}
die	63177	1.73×10^{-5}

tables keep essentially the same relative score positions. Finally, the words that represent concepts score consistently higher than the function words.

3.4 Specificity of multi-word concepts

Although $Rel\_var(.)$ gives some evidence about whether a pair of words (w, b_i) occurs at preferred relative positions, it is not reliable to assume that two strongly associated words are always part of a compound concept. In fact, the following tables (3.13, 3.14 and 3.15) show, for three different languages, some strongly associated pairs which do not form compound concepts.

Table 3.13: False compound concepts from the English corpus.

Pair	***Rel_var(.)***	**Frequency by relative position**
(the, safest)	1.000	[0, 0, 0, 0, the, **11**, 0, 0, 0]
(encoded, by)	0.965	[1, 5, 1, **579**, by, 0, 0, 2, 0]
(in, conjunction)	0.936	[1, 1, 1, 0, in, **171**, 0, 1, 1]
(physiology, or)	0.895	[1, 1, 0, **121**, or, 2, 1, 1, 0]
(or, indirectly)	0.828	[1, 1, 0, 0, or, **23**, 0, 0, 0]
(in, fact)	0.671	[14, 11, 8, 0, in, **307**, 8, 2, 15]

By looking at tables 3.13, 3.14 and 3.15, one can see that despite the fact the $Rel\_var(.)$ values of these pairs are high, they do not form compound concepts. For instance, "in fact" has a $Rel\_var(.)$ value of 0.671 essentially because its co-occurrence is relatively

Table 3.14: False compound concepts from the Portuguese corpus.

Pair	*Rel_var(.)*	Frequency by relative position
(equivale, a)	1.000	[0, 0, 0, **15**, a, 0, 0, 0, 0]
(por, detrás)	1.000	[0, 0, 0, 0, por, **10**, 0, 0, 0]
(por, exemplo)	0.929	[9, 9, 8, 10, por, **1759**, 1, 6, 14]
(o, acto)	0.886	[0, 0, 1, 0, o, **18**, 0, 0, 0]
(a, residir)	0.866	[0, 0, 0, 0, a, **15**, 0, 1, 0]
(acompanhado, por)	0.750	[1, 2, 0, **37**, por, 0, 0, 0, 2]

Table 3.15: False compound concepts from the German corpus.

Pair	*Rel_var(.)*	Frequency by relative position
(die, balsamtanne)	1.000	[0, 0, 0, 0, die, **17**, 0, 0, 0]
(von, kondomen)	0.910	[0, 0, 0, 0, von, **23**, 0, 1, 0]
(teilnahme, an)	0.893	[1, 1, 1, **59**, an 0, 0, 0, 0]
(metaanalyse, von)	0.858	[1, 0, 0, **14**, von, 0, 0, 0, 0]
(professoren, an)	0.811	[0, 0, 0, **10**, an, 0, 0, 0, 1]
(die, schultern)	0.794	[1, 0, 0, 0, die, **9**, 0, 0, 0]

high as a collocation, but it is not a concept nor necessarily part of a greater compound concept. This means that the specificity of a multi-word cannot be assessed entirely by the relation between pairs of words regarding their tendency to occur at relative fixed positions. However, it must be noted that the pairs listed in the previous tables are composed by at least one function word and that the *specificity* of function words ($Spec(.)$, equation 3.5) is usually a low value.

Table 3.16 illustrates some differences regarding the specificity values of the single-words between some strongly associated pairs.

Table 3.16: Comparison of the *Spec(.)* values for the single-words in some multi-words.

(A,B)	*Rel_var(.)*	*Spec(A)*	*Spec(B)*
(safest, procedures)	1.000	4.84×10^{-2}	1.51×10^{-3}
(rheumatoid, arthritis)	0.787	1.28×10^{-2}	4.42×10^{-3}
(autoimmune, reaction)	0.825	3.76×10^{-3}	1.73×10^{-3}
(the, safest)	1.000	1.95×10^{-5}	4.84×10^{-2}
(encoded, by)	0.965	6.17×10^{-3}	7.23×10^{-5}
(in, conjunction)	0.936	3.33×10^{-5}	6.54×10^{-3}

Although all pairs in the table have high $Rel\_var(.)$ values, the specificity values of the function words is lower than the specificity values of the true single-word concepts. This information is valuable to distinguish between concept and non-concept multi-words.

Being W a multi-word consisting in a sequence of words (w_1, w_2, ..., w_n), equation 3.6

(*unigram quality)* is used to measure the average specificity of a pair of single-words.

$$uq(w_i, w_j) = \sqrt{Spec(w_i) \,.\, Spec(w_j)} \,. \tag{3.6}$$

The geometric average is used because its results are closer to the lower values than the highest values, considering $Spec(w_i)$ and $Spec(w_j)$. So, by returning lower values when one of the words is not a single-word concept, $uq(.,.)$ *penalizes* these types of pairs. The following equation (equation 3.7 – *pair quality*) measures the tendency for two words on a multi-word to co-occur at a certain distance relatively to each other.

$$pq(w_i, w_j) = \frac{x_{j-i}}{\sum_{k \in Pos} x_k} \,. \tag{3.7}$$

The *pair quality*, $pq(w_i, w_j)$, measures the tendency for a word w_j to co-occur at position $j - i$ relative to w_i. This is done by dividing x_{j-i} (the number of co-occurrences of w_j at position $j - i$ relative to w_i), by the sum of all co-occurrences of w_j at any position relative to w_i. This sum is given by counting all x_k values of the list $X_{(w_i,w_j)}$ in equation 3.1. Also, $Pos = \{-\frac{s}{2}, \ldots, -1, 1, \ldots, \frac{s}{2}\}$ is the set of all relative positions in the window of size s. While $Rel\_var(.)$ checks for preferences at any position, $pq(.,.)$ checks for the preference at a certain position. As an example of $pq(w_i, w_j)$, consider the pair (*cardiopulmonary*, *resuscitation*) which has the following distribution of co-occurrence frequencies: [0, 0, 0, 0, *cardiopulmonary*, 23, 0, 0, 1]. The word *resuscitation* occurs 23 times one position after *cardiopulmonary*, meaning that $pq(\text{cardiopulmonary}, \text{resuscitation}) = \frac{23}{23+1+(0\times6)}$. In other words, *resuscitation* has a preference of 0.958 to co-occur one position right after *cardiopulmonary* (multi-word "cardiopulmonary resuscitation").

Finally, for a multi-word W = (w_1, w_2, ..., w_n) the following metric measures the specificity of W:

$$SpecM(W) = \left(\frac{1}{\binom{n}{2}} \sum_{\substack{i,j \,\in\, \{1 \ldots n\} \\ \wedge\, i<j}} uq(w_i, w_j) \,.\, pq(w_i, w_j) \right) .\, min(Spec(w_1), Spec(w_n)) \,. \tag{3.8}$$

The specificity of a multi-word W is measured by computing all single-word pair combinations of W in terms of the quality of their isolated single-words, which is given by $uq(w_i, w_j)$, and the quality of the pair, which is given by $pq(w_i, w_j)$. Basically, $pq(w_i, w_j)$ (*pair quality*) gives an hint whether a pair (w_i, w_j) forms a compound concept, by measuring the tendency for the pair to co-occur on certain positions, while $uq(w_i, w_j)$ (*unigram quality*) measures the average specificity of the words in the pair. Then, the multiplication by the minimum $Spec(.)$ value of the first and last words of the multi-word has the purpose of harming the multi-words that do not start or do not end with concepts, as described in section 3.1.2.

Tables 3.17, 3.18 and 3.19 present some multi-words from the three different test corpora in different languages, as described in Table 4.1.

Table 3.17: Specificity of some multi-words from the English corpus.

Multi-word (W)	***SpecM(W)***
extracorporeal membrane oxygenation	3.15×10^{-4}
cardiopulmonary resuscitation	6.83×10^{-5}
restrictive abortion laws	9.32×10^{-6}
sodium pertechnetate	5.55×10^{-6}
intrahepatic cholestasis of pregnancy	4.28×10^{-6}
ophthalmology training in	1.98×10^{-8}
by the fact that medicine	3.87×10^{-9}
of clinical chemistry and	3.54×10^{-9}
international association of	2.70×10^{-9}
in the	1.38×10^{-10}

Table 3.18: Specificity of some multi-words from the Portuguese corpus.

Multi-word (W)	***SpecM(W)***
fissura labiopalatal	7.94×10^{-4}
aborto cirúrgico	5.36×10^{-6}
acidente vascular cerebral	5.06×10^{-6}
complexo principal de histocompatibilidade	4.78×10^{-6}
infecção bacteriana	1.63×10^{-6}
do tronco cerebral	4.00×10^{-8}
de ventre	8.20×10^{-9}
complexo principal de	5.98×10^{-9}
de gestação	2.52×10^{-9}
para que	2.19×10^{-10}

Table 3.19: Specificity of some multi-words from the German corpus.

Multi-word (W)	***SpecM(W)***
nebennierenrindenstimulierenden hormons	2.16×10^{-3}
konus sehr weit fortgeschritten	8.63×10^{-4}
akute bronchitis	4.19×10^{-5}
chemische kastration	3.82×10^{-5}
anthroposophische medizin	2.05×10^{-6}
des menstruationszyklus	3.73×10^{-8}
der sehstärke mehr	3.73×10^{-8}
für uns	2.20×10^{-8}
in der schulmedizin	5.40×10^{-9}
der komplementärmedizin	5.27×10^{-9}

Despite the reduced number of multi-words in tables 3.17, 3.18 and 3.19, it allows us to conclude that the information from the relative $SpecM(.)$ values is consistent among the three languages. In fact, the multi-word concepts in the tables have higher specificity values than the non-concepts. For instance, "extracorporeal membrane oxygenation" has, undoubtedly, a more specific semantic value than "in the", for which no topic can be even vaguely suggested.

Furthermore, the separation between concepts and non-concepts seems to be highly obvious on these tables. Concepts seem to have specificity values above 1.0×10^{-6} while non-concepts seem to score below 10.0×10^{-8}. Although these are the specificity values for the examples on these tables, they suggest the existence of a specificity threshold which separates concepts from non-concepts.

The suggestion that such specificity threshold may exist was already described in section 3.1.4. Since we are now able to measure the specificity values for words and multi-words, the next chapter will detail the procedure that was used to find those specificity thresholds. The chapter will also include the details of the tested corpora and the results for the procedure.

4

The *ConceptExtractor* – corpora, methodology and results

This chapter presents the corpora, experimental methodology and the results of the extraction of concepts using the *ConceptExtractor*. I will start by explaining, in section 4.1, the tools for building the corpora used in the experiments. It is my belief that this method and tools for building corpora are simple enough to be useful for other researchers in Natural Language Processing. In section 4.2, I will illustrate how the specificity thresholds were found, by presenting the information about the test sets and the procedure to find the *best* threshold values which maximize the results. As it will be seen, those threshold values are quite similar for all the tested languages. Finally, I will also present the results of the *ConceptExtractor* including comparative results with some statistical methods.

4.1 The corpora

To build the Wikipedia-based corpora, I started by obtaining the titles of documents belonging to the *Medicine* category, down to a certain depth. *CatScan V2.0*β (`http://tools.wmflabs.org/catscan2/catscan2.php`) was the tool used. The Wikipedia article (`http://en.wikipedia.org/wiki/Wikipedia:CatScan`) describes *CatScan*:

> *CatScan* is an external tool that searches an article category (and its subcategories) according to specified criteria to find articles, stubs, images, and categories. It can also be used for finding all articles that belong to two specified categories (the intersection). *CatScan* is developed by the German wikipedian *Duesentrieb* and is run on the *toolserver*, a special machine used for such tools.

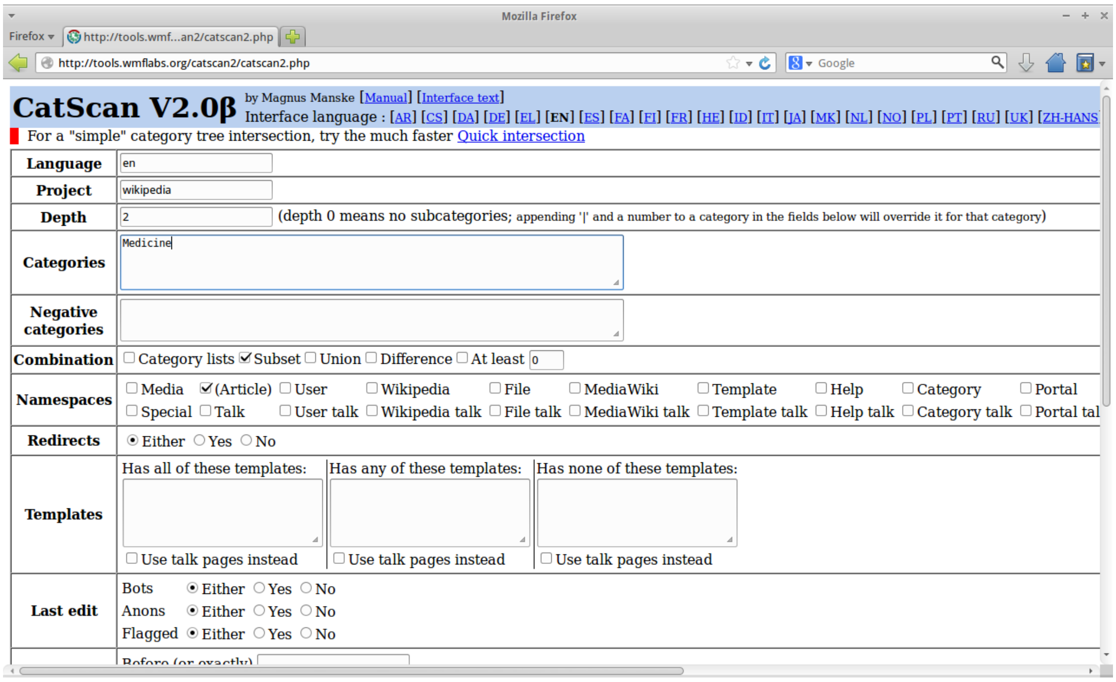


Figure 4.1: *CatScan V2.0β* web interface.

With the names of articles belonging to the *Medicine* category, the following step was the extraction of Wikipedia *XML dump files* with the content of the articles. *Export pages* are a Wikipedia web based service to export article pages in an XML format. Each language has its own export page:

- **English:** http://en.wikipedia.org/wiki/Special:Export
- **Portuguese:** http://pt.wikipedia.org/wiki/Especial:Exportar
- **German:** http://de.wikipedia.org/wiki/Spezial:Exportieren

Listing 4.1: XML Excerpt of English *Medicine* article.

```
<mediawiki xsi:schemaLocation="http://www.mediawiki.org/xml/export-0.8/ ...
  <page>
    <title>Wikipedia</title>
    <ns>0</ns>
    <id>18957</id>
    <revision>
    <text xml:space='preserve' bytes="68699">
      {{two other uses|the science and art of healing|pharmaceutical
      drugs|Medication}} '''Medicine'''
      ({{IPAc-en|'|m|e|d|s|i|n|audio=En-uk-medicine.ogg}},
      {{IPAc-en|'|m|e|d|i|s|i|n|audio=En-us-medicine.ogg}}) is the
      field of [[applied science]] related to the art of healing by
      [[diagnosis]], [[healing|treatment]], and prevention of [[disease]].
```

However, the *XML dump files* includes information, such as revision history, users, etc., which has no interest for building text corpora. Moreover, the text element is not raw text, but it includes many wikipedia tags, such as links to images, other articles, etc. I've created a *Python* library (publicly available at `https://github.com/joaoventura/WikiCorpusExtractor`) which creates corpora from a Wikipedia *XML dump file*, cleaning the text as a result. With this library, it is possible to create a corpus with one only document, or configure some parameters such as the minimum words by document or the maximum number of words in a corpus. Figure 4.2 shows an example output.

Listing 4.2: Excerpt of the output of the *Python* library to create corpora.

```
<doc id="xx" title="Autism">
  Some tokenized text, i.e., words and punctuation are separated by a space .
  Some special words like step-by-step or U.S.A. are correctly handled .
</doc>
<doc id="xxx" title="zzz">
  ...
</doc>
```

The final output, for each language, was then *gzipped* for smaller sizes.

As already mentioned, the corpora used in the experiments are composed of articles extracted from the Wikipedia *Medicine* category, for three different European languages, namely English, Portuguese and German. The articles belong to the *Medicine* main category or to a subcategory of *medicine* down to a certain depth, being "depth" the level of subcategories used (for instance, 1 means all direct subcategories of *Medicine*, while 2 includes also the subcategories of all direct subcategories of *Medicine*, and so on). Table 4.1 presents some basic statistics about the corpora.

Table 4.1: Basic statistics about the corpora based on Wikipedia *Medicine* articles.

Corpus	English	Portuguese	German
Number of documents	4 160	4 066	4 911
Total words	4 657 053	4 153 202	4 337 068
Average #words by document	1 120	1 022	884
Depth of subcategories	2	4	2

The target number of words for all corpora was around 4M – 4.5M words. To guarantee approximately the same number of words for all languages, it had to be added more documents to the German corpus, and documents of deeper categories had to be included on the Portuguese corpus. For the German case, this has to do with the fact that the German language tends to agglutinate many compound concepts into single-words, and so, by having a less number of words by document, the number of documents had to be increased. For Portuguese, because of the scarcity of documents belonging to the *medicine* category and direct subcategories (down to depth 2), I was forced to use documents down to depth 4.

4.2 Methodology and results

Although the definition of concept seems clear, there is sometimes a fuzzy area where some terms seem difficult to classify as concept or non-concept. Thus, it was asked to Prof. Dr. Maria Francisca Xavier of the Linguistics Department of FCSH/UNL to provide her expertise to the evaluation process. For that, 300 single-words and 300 multi-words were randomly extracted from each corpus. To guarantee enough statistical information for the experiment, each random term had to occur at least 3 times in the entire corpus. Finally, each term was manually classified as concept or non-concept. So, for each of the three languages, 2 test sets were used (single-word and multi-word), each with 300 elements. The multi-word sets contained from 2-grams to 5-grams. Excerpts of the classified lists can be found in Appendix A.

For each test set, the *Precision*, *Recall* and *F-measure* were calculated. These measures are give by equations 4.1, 4.2 and 4.3.

$$Precision = \frac{\#(\text{true\_concepts} \cap \text{considered\_concepts})}{\#\text{considered\_concepts}} \ . \tag{4.1}$$

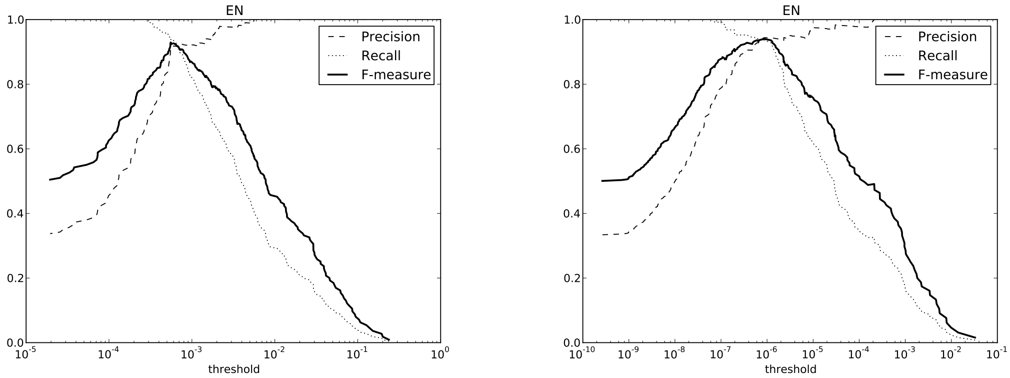
$$Recall = \frac{\#(\text{true\_concepts} \cap \text{considered\_concepts})}{\#\text{true\_concepts}} \ . \tag{4.2}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \ . \tag{4.3}$$

Precision, sometimes also called *positive predictive value*, measures the proportion of how many words and multi-words considered concepts by the method (*considered_concepts*) are indeed concepts (*true_concepts*). On the other hand, *Recall* measures how many true concepts, of the total number of true concepts (where *true concept* is a word or multi-word classified manually as concept), were correctly considered concepts by the method. *F-measure* is the harmonic average between *Precision* and *Recall*, tending essentially towards the lowest value.

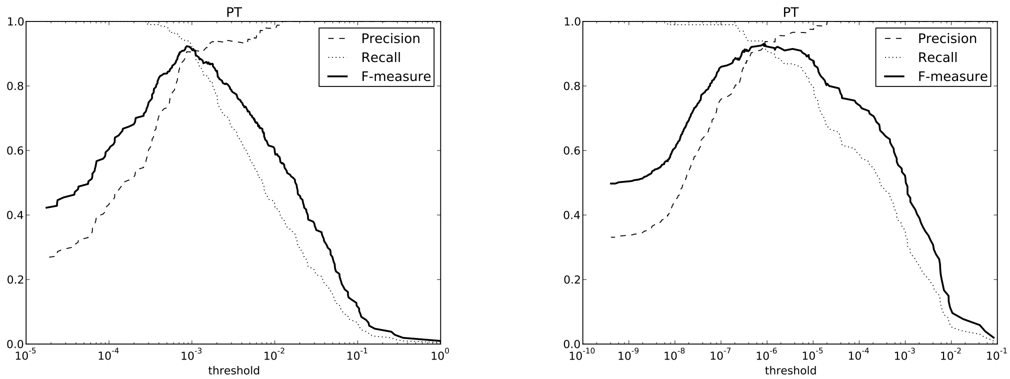
However, the specificity of words and multi-words only allows to have lists ranked by specificity. But given the empirical fact that concepts are more specific than non-concepts, as described in section 3.1.4, this means that there must be a certain specificity threshold for which above that threshold, a word or multi-word can be considered concept, and below that threshold, non-concept.

In order to find that specificity threshold, for each test-set I built a method to consider all possible thresholds and compute the Precision, Recall and F-measure for each case. The idea is that the specificity threshold which gives the best results should be the specificity threshold to separate concepts from non-concepts. Figures 4.2, 4.3 and 4.4 show the Precision, Recall and F-measure results by threshold, for each test set.



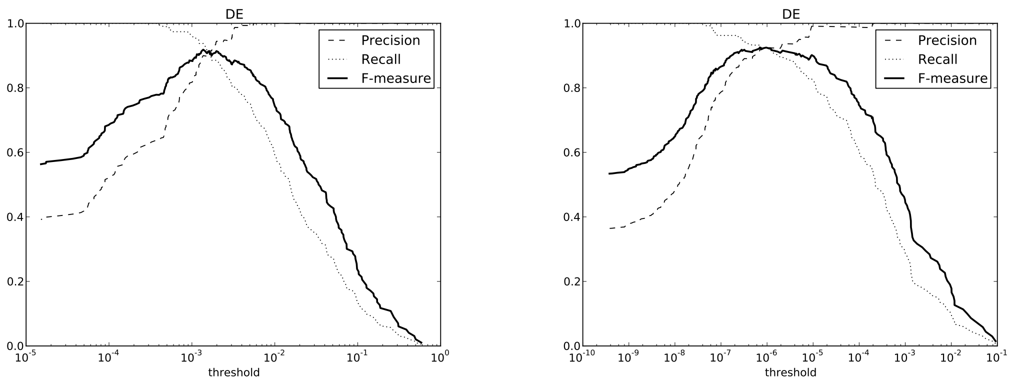
(a) EN unigrams (b) EN multiwords

Figure 4.2: Precision, Recall and F-measure for different thresholds in the English test sets.



(a) PT unigrams (b) PT multiwords

Figure 4.3: Precision, Recall and F-measure for different thresholds in the Portuguese test sets.



(a) DE unigrams (b) DE multiwords

Figure 4.4: Precision, Recall and F-measure for different thresholds in the German test sets.

As expected, for each test set, lower thresholds imply higher Recall and lower Precision values. This has to do with the fact that setting a low threshold means that every word and multi-word are considered concept – Precision is lower since many non-concepts are being considered concepts, but since all true concepts are being considered concept by the method, Recall is high. On the other hand, Precision is higher and Recall is lower for higher threshold values, since only the highly specific terms are being considered concepts, while the less specific ones are left behind. Hence the low Recall for higher thresholds.

However, as it is visible in the figures, there are certain threshold values for which the *F-measure* has a maximum value. Those values correspond to the best equilibrium between Precision and Recall. Table 4.2 shows the Precision, Recall and threshold for the maximum F-measure value of each test set.

Table 4.2: Precision, Recall and threshold values for the maximum *F-measure* value of each test set.

Test set	F-measure	Precision	Recall	threshold
Single-words – English	0.91	0.90	0.93	1.63×10^{-3}
Single-words – Portuguese	0.93	0.93	0.95	1.44×10^{-3}
Single-words – German	0.92	0.91	0.94	1.97×10^{-3}
Multi-words – English	0.94	0.93	0.95	6.10×10^{-7}
Multi-words – Portuguese	0.93	0.92	0.95	6.73×10^{-7}
Multi-words – German	0.93	0.92	0.94	6.99×10^{-7}

The thresholds corresponding to the maximum *F-measure* values were found for approximate threshold values, considering it being single-words or multi-words. This allowed me to choose, as language-independent thresholds, an average value for each group. These average threshold specificity values were set to 1.68×10^{-3} for all single-words and 6.60×10^{-7} for all multi-words, independently of its size. Therefore, for the *ConceptExtractor*, terms with specificity values above the average thresholds are to be considered concepts, below that, non-concepts. Table 4.3 shows the classification results for the *ConceptExtractor* method considering the mentioned average threshold values.

Table 4.3: Precision, Recall and F-measure values for the test sets considering the average threshold values.

Test set	Precision	Recall	F-measure
Single-words – English	0.90	0.93	0.91
Single-words – Portuguese	0.91	0.94	0.92
Single-words – German	0.89	0.94	0.92
Multi-words – English	0.93	0.93	0.93
Multi-words – Portuguese	0.92	0.95	0.93
Multi-words – German	0.91	0.94	0.92

The results are practically unchanged, since the thresholds for each single-word and multi-word test set are relatively similar. Precision and Recall values can be considered good given that no morphosyntactic information was used to focus the extraction to any particular language. Also, since the results between languages are relatively close in Table 4.3, I believe this can be considered a language independent approach.

Tables 4.4 and 4.5 show the comparison of the *ConceptExtractor* with some approaches mentioned in chapter 2. The basis for comparison were as follows: for single-words, since each method provides its own score metric, the comparison with the *ConceptExtractor* thresholds does not make sense. Therefore, the comparison for single-words was made using the results which maximized the F-measure value of each method. As for multi-words, since *LocalMaxs* (described in section 2.3.6) is capable of identifying relevant multi-words on a yes-no basis, the comparison was made using the classification results of *LocalMaxs*, and the classification results of *ConceptExtractor* using the given average thresholds in order to separate concepts from non-concepts.

Table 4.4: Precision and Recall values for different approaches – single-words.

Approach	Parameter	English	Portuguese	German
ConceptExtractor	Precision	**0.90**	**0.93**	**0.91**
	Recall	**0.93**	**0.95**	**0.94**
Tf-Idf	Precision	0.58	0.68	0.60
	Recall	0.85	0.73	0.86
Zhou	Precision	0.65	0.62	0.66
	Recall	0.73	0.66	0.67
Syllables	Precision	0.66	0.72	0.78
	Recall	0.78	0.84	0.80

Table 4.5: Precision and Recall values for different approaches – multi-words.

Approach	Parameter	English	Portuguese	German
ConceptExtractor	Precision	**0.93**	**0.92**	**0.91**
	Recall	**0.93**	**0.95**	**0.94**
LocalMaxs	Precision	0.75	0.77	0.76
	Recall	0.71	0.74	0.72

ConceptExtractor shows higher results than the other methods on the extraction of single-word concepts and multi-word concepts.

Regarding single-word extractors, although *Tf-Idf* is aimed to work only on documents, it was adapted such that the score of a word was given by its maximum *Tf-Idf* score obtained for some document, considering all documents of the corpus. Although the Recall is quite good on average, the low Precision comes from the fact that some concepts are relatively frequent in the corpus, attaining lower *Idf* values. As for the *Zhou* approach, it scores a word by measuring its capabilities to form local clusters in a corpus. However, in the tests it was noted that rare concepts are harmed by this metric since their tendency to form clusters is greatly diminished by their lack of occurrences. Finally, although the *Syllable* approach scores, in average, higher than the other methods, it tends to harm smaller concepts, such as "air", "CDC" (acronym for *Center for Diseases Control*) or "CBP" (acronym for *Calcium Binding Protein*), which do occur in the English corpus.

As for multi-words, regarding *LocalMaxs*, the lower results are due to the fact that the method classifies terms by comparing them with their immediate neighbors. For instance, irrelevant multi-words such as "which is", "from the", "rather than", "responsible for", among others, tend to be considered relevant by this extractor. This happens, essentially, because the inclusion of a new word before or after the multi-word does not increase its $SCP\_f(.)$ score. For instance, immediate neighbors of "responsible for" include terms such as "branch responsible for", "responsible for suppressing", "responsible for skin", etc. However, although they seem more relevant than "responsible for", these neighbors are infrequent resulting in lower scores. As for the recall, it may be due to the fact that the method tends to prefer the largest terms. For instance, "genetic information", which is undoubtedly a concept, is not considered as such by *LocalMaxs* because it has *better* immediate neighbors, such as "genetic information research" or "cell's genetic information".

4.3 Summary

In the first part of this thesis I presented a new methodology for the extraction of single-word and multi-word concepts from large texts. This methodology uses tools and ideas, such as the specificity of terms and the *Rel_var* metric, which may be potentially usable outside the scope of the extraction of concepts. For instance, the idea of specificity can be used in the identification of anchor points in parallel texts for the task of automatic translation: if the texts are truly parallel (one being the exact translation of the other), the specificity of a term in *language A* should be similar to the specificity of the translated term in *language B*.

Considering the limitations of most approaches regarding the dependence on tools which are language-specific, such as parsers, Part-of-Speech taggers, external lexicons, etc., the *ConceptExtractor* is a language-independent approach. However, the main criterion for its successful usage on untested languages is that the terms in an untested language must follow the same basic "rule" as on the tested languages – the single-word concepts in compound concepts must tend to co-occur in fixed positions relatively to each other. That is the basis of this approach.

Regarding other language-independent approaches, beside the fact that most are incapable of extracting single-words and multi-words using the same methodology, I've shown that the *ConceptExtractor* shows higher comparative results.

However, the *ConceptExtractor* is not without its drawbacks. Most of these drawbacks arise from the fact that some multi-word concepts, such as *President of the United*, score high in their specificity, although they are clearly incomplete. In this specific case, although one cannot say that *President of the United* does not contain any concept, clearly *President of the United States* or *President of the United Nations* are better and more complete concepts. These are frontier cases, although quite uncommon. A possible solution could be to include a new rule for concepts such as "*multi-word concepts must start and end with **complete concepts***". However, the problem would be to define programmatically or statistically, what a *complete concept* is. Algorithms such as *LocalMaxsLocalmaxs* could be of help for those highly specific situations, but not as complete replacements.

Another improvement could be done on the identification of synonyms and of singular-plural concepts. For instance, although *abortion* and *abortions* are the same basic concept, both the extractor and downstream applications are unaware of the similarity.

Finally, although the *ConceptExtractor* presents quite encouraging results, future work could be done in order to increase the performance of the extractor.

Part II

Applicability of automatically extracted concepts

5

Extraction of explicit and implicit keywords from documents

Part II of this thesis presents some applications for concepts automatically extracted by the *ConceptExtractor*, as described in Part I. In this specific chapter, I will present an approach based on concepts for the extraction of *explicit* and *implicit* keywords from documents. This approach is language-independent and comparative results for three different European languages will be presented. The work in this chapter was published in [VS13a].

5.1 About explicit and implicit keywords

Keywords are semantically relevant terms that are used to reflect the core content of documents. Some of the first works related to the automatic extraction of keywords were addressed in [Luh58], [Jon72] and [SY73]. However, in many applications, as in library collections, the extraction of keywords remains mainly a manual process.

In the context of this thesis, I argue that keywords are essentially concepts that are meaningful in the documents: they either describe the content of a document or of a part of a document. This approach starts by automatically extracting the concepts of the documents, using the *ConceptExtractor*. By doing this extraction, we are in fact reducing the search space from all possible sequences of single-words and multi-word expressions to a much smaller set of semantically meaningful concepts. Then, by applying *Tf-Idf* to the extracted concepts, the first ranked concepts are selected as explicit keywords of the document.

However, there are other meaningful concepts that, although they may not occur explicitly in a document, they are semantically related to the document content. These can be called the *implicit keywords*. They may, among other possibilities, provide a user of a search engine the access to documents that may not contain these keywords, but are semantically related to them. For instance, concepts such as "car emissions", "toxicology" and "acid rains" may be useful if automatically added as implicit keywords of a document about "air pollution", if those terms do not occur explicitly in that document.

To extract the implicit keywords of a document, the *Semantic Proximity* is calculated between concepts extracted from the corpus and each keyword of the document's explicit descriptor. The first ranked concepts, according to a defined metric, are selected as the document's implicit keywords and form the document implicit descriptor.

This chapter presents a statistical and language-independent approach to build document descriptors where each *global* document descriptor is made of two distinct descriptors: an *explicit descriptor*, containing explicit keywords, and an *implicit descriptor* with the implicit keywords.

Next section will describe the related work. In section 5.3, the explicit descriptor and its results will be presented, while the implicit descriptor and its results will be presented in section 5.4. A summary and conclusions for this chapter can be found in section 5.5.

5.2 Related work

Currently, there are two main methodologies for the extraction of keywords from documents: the supervised and the unsupervised learning approaches. Other division is usually made considering approaches that use linguistic tools, external lexicons, or statistical metrics. In the following subsections, I will review some work in order to frame the reader in the general shortcomings of current methods.

5.2.1 Noun-phrases as document descriptors – an unsupervised linguistic approach

In [CPGV05], the authors consider the usage of the Formal Concept Analysis (FCA) as an alternative to classic document clustering, regarding its applicability on search engines. More precisely, they defend that clustering techniques such as FCA allows for a quick focus on specific groups of documents and improves precision, as response to user queries. As attributes for clustering, they propose to use noun-phrases as document descriptors.

The authors start by extracting candidate phrases that may be relevant for the documents in which they appear. For that, they apply lemmatisation and Part-of-Speech tagging so that they may identify the grammatical category of words. Then, a specific linguistic pattern is applied, such that a phrase must start and end with a noun or adjective and might contain other nouns, adjectives, prepositions or articles in between. The ending result is a list of phrases and their frequency of occurrence.

The next step is the phrase selection, where different strategies are discussed. One of the strategies is to select the phrases with the highest frequency of occurrence covering the maximum number of documents retrieved. Other strategy is to use the frequency analysis, although restricting the set of candidate phrases to those containing one or more of the original query terms. The last strategy assigns higher values to those phrases that occur more frequently in the retrieved document set than in the whole collection, somewhat similar to *Tf-Idf*. The rest of the paper deals with the clustering process having the mentioned features as their basis.

However, by using lemmatisation and Part-of-Speech taggers, the authors make use of language-specific tools which may not be available for other languages.

5.2.2 UvT – an unsupervised hybrid approach

UvT [Zer10], is a linguistic and statistical approach for the extraction of keywords in scientific documents. In this approach, Zervanou starts with a linguistic preprocessing of the texts, namely its Part-of-Speech tagging and the identification of specific areas of the documents, such as the title, abstract, introduction, conclusions, acknowledgements and references. Then, the next step consists of the identification of candidate key-phrases, by means of predefined morphosyntactic rule patterns. These patterns are based on some well-defined grammatical sequences.

In order to reduce the variation of the results after the application of a statistical measure, the author proposes the *normalization* of the text. To reduce the morphological variation, he uses the Wordnet lexicon to obtain the lemmas of each candidate key-phrase, while for orthographic variations, such as hyphenated vs non-hyphenated compound phrases, they are treated by rule matching techniques.

Finally, the author applies the *C-value* measure to obtain a score for a multi-word. This *C-value* metric is essentially the multiplication of the frequency of occurrence of a phrase by its length. As with other linguistic approaches, the use of language-specific tools and, in this case, of linguistic rules to obtain lemmas and identify different orthographically written similar concepts, imposes a language dependency. For instance, Wordnet is not available for many languages and is not complete even for English (in the sense of including all possible combinations or relations). This may imply a lower than wished Recall. Furthermore, the usage of the length of a term in the calculation of the *C-value* may imply the removal of shorter keywords, such as *RAM* or *ROM* in an article about *Computer memory*.

5.2.3 Lexical chains – a supervised learning approach

In their paper [EC07], Ercan and Cicekli proposed a supervised learning method for the extraction of keywords by means of lexical chains. A lexical chain is a graph connecting semantically related words. To build the lexical chains of a text, the authors use Wordnet, specifically Wordnet's synonyms, hypernym/hyponym and meronyms. The end result is a graph (lexical chains), where the nodes are words and the relations are expressed in the connections between nodes.

The next step consists on the extraction of features for the supervised learning task. Ercan and Cicekli use, for each node (word), the frequency of occurrence of the word and the first and last positions of occurrence. Also, they use the type of semantic relation as a feature, giving it different weights accordingly to the type of relation (synonym relations weight more, while meronym relations less). Finally, the authors use a C4.5 decision tree induction algorithm on a manually classified test set.

Overall, the dependency on Wordnet makes this approach difficult to apply to other languages for which such external lexicons may not exist.

5.2.4 SVM – another supervised learning approach

The approach in [ZXTL06] is another supervised learning approach. It uses a Support Vector Machine to train a keyword extractor. An SVM is a supervised learning technique that tries to find a linear or non-linear hyperplane which best separates data of different classes. In this paper, the authors start with a linguistic preprocessing of the text, namely a word and sentence tokenization, and a Part-of-Speech tagging. Then, they use a tool to analyze the dependency relation between words on sentences. Finally, they obtain candidate terms up to 3-grams above a certain frequency threshold, and exclude words which are on a stop-word list. Wordnet is also used to conduct a stemming process.

To train the classifier, each candidate term includes Global Context features such as the *Tf-Idf* value and its positions of occurrence in the text, and Local Context features such as the Part-of-Speech descriptor and the dependency relations. The rest of the paper deals with the details of the classifier.

Similar to the approaches described above, the dependence on linguistic tools and external lexicons may lead to greater difficulty in applying this method to other languages for which such tools may not be available.

5.2.5 Wikipedia as data source – recent trends

The works of Xu *et al.* [XYL10] and Mihalcea and Csomai [MC07] are examples of the recent trend on some current approaches which use information from large structured data sources, such as Wikipedia.

For instance, in [MC07], an unsupervised learning approach, the authors use metrics such as *Tf-Idf* and χ^2 to rank terms by relevance to a document. The most relevant part, however, is that the authors use Wikipedia to do a word sense disambiguation, namely by checking if the highest ranked terms are related to Wikipedia articles. Accordingly to the authors, this allows the improvement of both Precision and Recall.

As for [XYL10], they devised an innovative supervised learning approach based on Support Vector Machines. The innovation on this work is related to the features fed to the classifier, which are based on the analysis of Wikipedia. For instance, for a given word x_i, they obtain a score proportional to the number of documents for which x_i is an out-link (a link to another Wikipedia document) and for which x_i is an in-link (a link in another Wikipedia document). Other feature is the category of the word, which is obtained through the category information of every article in which the word occurs, and through Wordnet. Finally, further information is obtained through the *infobox* table. The *infobox* is the fixed-format table that usually occurs on the top-right of Wikipedia articles and consists of structured information, as for example the area and population on articles about cities and countries.

As mentioned, the usage of Wikipedia as data source is a recent trend for which some of the current works are turning to. The main problem I can foresee lies in the fact that these approaches may tend to be overly dependent of Wikipedia, which although being a community project, it is not guaranteed it will always be available. Another possible problem lies in the fact that, given that Wikipedia is an encyclopedia of general knowledge, keywords of documents outside the scope of Wikipedia may not be represented inside the Wikipedia structure – documents of very specific areas of knowledge are good examples.

5.2.6 Keywords as relevant expressions – a statistical approach

The work in [SL10] is an example of an approach which uses only statistical tools to extract keywords from documents. As starting point, the authors use the LocalMaxs algorithm [SL99] to extract MWEs. This procedure essentially reduces the search space from all possible sequences of words in a document to only a selected few multi-words. As for metrics, the paper compares four different ones.

The first metric is *Tf-Idf*, which was already reviewed in section 2.2.4. Basically, *Tf-Idf* considers the frequency of occurrence of a term in a given document and in other documents of a collection. The idea behind *Tf-Idf* is that a term is more relevant as keyword of a document if it occurs frequently in that document but not in many more documents. *Tf-Idf* is usually considered the baseline method for which others are compared against.

The second metric is *LeastRvar*, which consists on the analysis of the words in the beginning and ending of a multi-word expression. For a multi-word $W = (w_1, w_2, \ldots, w_n)$, its expression is as follows:

$$LeastRvar(W) = least(Rvar(w_1), Rvar(w_n)) \ . \tag{5.1}$$

$$Rvar(w) = \frac{1}{\|D\|} \sum_{d_i \in \mathcal{D}} \left(\frac{p(w, d_i) - p(w, .)}{p(w, .)} \right)^2 \qquad p(w, .) = \frac{1}{\|D\|} \sum_{d_i \in \mathcal{D}} p(w, d_i) \ .$$

$\|D\|$ is the number of documents in the collection, while d_i is the *i*-th document in the collection. $Rvar(w)$ measures the variation of the probabilities of w along the documents in the collection. *LeastRvar* tends to privilege informative MWEs and penalize multi-words starting or ending with function words.

Another metric presented in the paper is the $LeastCv$ which is somewhat similar to $LeastRvar$, although based on the *coefficient of variation*. Its expression is:

$$LeastCv(W) = least(Cv(w_1), Cv(w_n)) \ . \tag{5.2}$$

$$Cv(w) = \frac{\sigma(w)}{\mu(w)} \qquad \sigma(w) = \sqrt{\frac{1}{\|D\|} \sum_{d_i \in \mathcal{D}} (p(w, d_i) - p(w, .))^2} \qquad \mu(w) = p(w, .) \ .$$

The last metric presented in this paper is $Mk(W)$ which considers the fact assumed by the authors that the keywords of a document tend to have a "optimum" number of characters.

$$Mk(W) = LeastRvar(W) \,.\, Median(W) \,.\, Ckl(W) \ . \tag{5.3}$$

$$Ckl(W) = \frac{1}{|Pnw(W) - T| + 1} \qquad Pnw(W) = \frac{Num\_chars(W)}{Median(W)} \ .$$

According to the authors, $Pnw(W)$ (*pseudo number of words* of W) returns a value close to the number of meaningful words of W. $Ckl(w)$, on the other hand, measures the deviation of the *pseudo number of words* of W to a fixed T (which is 2.5 or 3.5 in their experiments). Finally, $Mk(W)$ privileges MWEs that do not start or end in stop-words (given by $LeastRvar(W)$), are long (given by $Median(W)$) and have a specific *pseudo number of words* (given by $Ckl(W)$). The $Median(W)$ in $Mk(W)$ is the median number of characters of the individual words of W. In the same line of thought, the authors propose another metric, $Sk(w)$, to rank single-words as keywords of documents.

$$Sk(w) = Rvar(w) \,.\, Length(w) \ . \tag{5.4}$$

$Sk(w)$ privileges lengthier single-words which have high relative variations of probabilities along the documents in a collection.

The method and metrics presents in [SL10] is in fact language independent, since it does not use language-specific tools as the previously reviewed papers. However, to decide the relevance of a multi-word W as keyword, only by the analysis of the starting and ending words of W, is controversial, and seems to return good results only because it removes the errors imposed by *LocalMaxs* (specifically, by returning multi-word candidates with function words on the "edges"). With a better multi-word extractor (one that does not suggest keyword candidates starting or ending with function words), deciding the relevance of an entire expression only by the first/last word may not be valid.

Furthermore, benefiting larger expressions in the $Mk(.)$ metric may also be controversial. For instance, in the Wikipedia document about the musical band "The Doors", "The Doors" is a quite frequent expression. However, its $Mk(.)$ value would be low, essentially because of the lower median of the number of characters, and because of the lower $LeastRvar(.)$ due to the word "The".

Finally, as for the extraction of single-word keywords with $Sk(.)$, considering only lengthier keywords may not be a valid approach, as it can be exemplified with real keywords such as "RAM" or "ROM" in Wikipedia's "Computer Memory" article.

5.2.7 TLR11 – a comparison of statistical methodologies

The work in [TLR11] presents a comparison of statistical methodologies for the extraction of single-word and multi-word keywords from documents. Some of the metrics compared in the paper are already described in this section, such as *Tf-Idf* and *LeastRvar*. Other metrics such as φ^2 and MI (Mutual Information) were also used. The innovation on this comparison is the introduction of new measures, called *operators*.

The *Least Operator* is the same used in the *LeastRvar* measure (as in equation 5.1), adapted to work with single-words. Considering that MT stands for any of the mentioned metrics, the *Least_MT* operator is defined as:

$$Least\_MT(W) = \begin{cases} MT(W) & \text{if } W \text{ is a single-word} \\ Min(MT(w_1), MT(w_n)) & \text{if } W \text{ is a multi-word } (w_1, \ldots, w_n) \end{cases} . \tag{5.5}$$

As it can be seen, equation 5.5 shows that, for a multi-word, the *Least* operator follows the same idea used in the *LeastRvar* metric, whereas for single-words, it is just the application of the metric to the word. Another operator described is the *Bubbled Operator* which deals with the prefix P of a single-word.

$$Bubbled\_MT(W) = MT(P) . \tag{5.6}$$

Other operators described in the paper include the *Least Bubbled MT*, *Least Median* and *Least Bubbled Median*.

$$LM\_MT = Least\_MT(T) . Median(T) . \tag{5.7}$$

$$LBM\_MT = Least\_Bubbled\_MT(T) . Median(T) . \tag{5.8}$$

With these operators and metrics defined, the paper then presents the results. Results seem to demonstrate the equivalence of some metrics, especially *Tf-Idf* and φ^2 with the *Least* and *Least_Bubbled* operators. However, clean *Tf-Idf* (i.e., without any operators) clearly has better results. My analysis is that for multi-words, the use of the *Least* operator is not a valid methodology for accessing keywords, since it only analyses the edge single-words. As for single-words, the *Least* operator is innocuous. However, the usage of the *Median* operator (as in *LM Tf-Idf*, $LM\varphi^2$, and *LBM Tf-Idf*) may introduces errors by giving more weight to larger single-words. In a similar way, by removing all single-words with less than 6 characters, the authors are also removing valid smaller candidate keywords. As already mentioned, there are perfectly valid smaller keywords, such as *RAM* and *ROM* in documents about "Computer Memory".

5.2.8 Latent Semantic Indexing – another statistical approach

Latent Semantic Indexing [LD97] is a technique widely used in *Information Retrieval* to index documents of a collection and return them as response to user queries. Basically, the technique consists of the generation of a table which relates the occurrence of words with the documents where they occur. Then, a posterior "compression" (linear decomposition) of that table is made using a technique called *Singular Value Decomposition* (SVD). In this way, a table which maps thousands of words into documents is condensed into a table with 50–300 components.

However, the applicability of LSI to the extraction of document's keywords, outside the use case of *Information Retrieval*, is marginal. This occurs essentially because the components generated after the *singular value decomposition* process may only marginally resemble the original terms in the documents. So, for the purpose of keyword visualization, LSI bears little interest. Also, because of size constraints, the generation of the original table is usually done only with single-words, not including multi-words nor knowledge of the order of words. However, the order of words and multi-words imply fundamental semantic meaning. For instance, the multi-word "hot dog" in a particular document about food should be considered as an integral concept, instead of the isolated terms "hot" and "dog". The word "dog", isolated, has little to none resemblance to food.

5.3 The explicit descriptor

The explicit descriptor is a set of keywords that occur explicitly in documents. For the purposes of this thesis, the explicit descriptor of a document is formed by 20 keywords: the 10 best scored single-words and the 10 best scored multi-words. To extract the keywords from a document, *Tf-Idf* (see the description in section 2.2.4) is applied to the concepts extracted from the documents.

5.3.1 Methodology and results

To evaluate the results of *Tf-Idf* applied to the extracted concepts, I've built corpora for English, Portuguese and German languages from Wikipedia XML dump files, with a procedure quite similar as described in section 4.1. However, articles of all categories were used for this experiment, instead of articles just from the *medicine* category. Table 5.1 presents some statistics about the corpora used.

Table 5.1: Basic statistics about the corpora based on Wikipedia *generic* articles.

Corpus	English	Portuguese	German
Number of documents	2 714	1 811	4 682
Total words	12 176 000	11 974 000	11 305 000
Average #words by document	4 486	6 611	2 414

For each corpus, the keywords of ten random document were randomly extracted and evaluated by three independent reviewers who had full access to the documents. The reviewers were instructed to consider as keywords the concepts that described the document or sections of it.

Table 5.2 shows the titles of the randomly selected documents and the reviewers' classification rates. The classification rates were obtained by measuring the number of "correct" keywords in which the majority of the reviewers agreed on. For instance, given a document, if $2/3$ of the reviewers agreed on the same keywords obtaining a rating of 0.80 and the third reviewer obtained a rating of 0.90, the overall rate for that document would be set to 0.80.

Table 5.2: Titles of documents and the reviewers' classification.

EN		**PT**		**DE**	
Doc. Title	**Cl.**	**Doc. Title**	**Cl.**	**Doc. Title**	**Cl.**
Abortion	0.95	Era dos Descobrimentos	0.85	Adolph Hitler	0.85
Brain	0.90	Al-Andalus	0.95	Genetik	0.70
Nostradamus	0.75	Direitos animais	0.85	Demokratie	0.75
Dog	0.95	Ácido desoxirribonucleico	0.90	G. Rossini	0.75
Saint Peter	0.75	História de Espanha	0.90	Immunsystem	0.85
Imagism	0.90	Gato	0.90	Kairo	0.75
Monopoly	0.90	W. A. Mozart	0.90	Microsoft	0.80
Desert	0.90	Teosofia	0.70	Papageien	0.50
Plate Tectonics	1.00	Vasco da Gama	0.85	Pflicht	0.90
History	0.75	Nazismo	0.85	Wolga	0.75
Average	**0.875**	**Average**	**0.865**	**Average**	**0.76**

The average classification rate for the three languages is 0.83, although it is lower for German mainly because two of the three reviewers had to rely on automatic translators for this language.

Tables 5.3, 5.4 and 5.5 show the explicit descriptors of the English *Brain* document, *Ácido desoxirribonucleico* Portuguese document and *Immunsystem* German document. Considering, for instance, the English *Brain* document, although some terms may not be accepted as correct keywords ("phenomena are identical", "central nervous"), most terms describe the core content of the documents.

Table 5.3: Explicit descriptor – *Brain* English document.

Single-word	***Tf-Idf(.)***	**Multi-word**	***Tf-Idf(.)***
brain	0.2046	spinal cord	0.0113
neurons	0.0346	cerebral cortex	0.0073
disease	0.0185	artificial intelligence	0.0057
animals	0.0179	optical lobes	0.0046
nervous	0.0167	olfactory bulb	0.0046
cells	0.0157	central nervous	0.0044
brains	0.0153	brain stem	0.0040
intelligence	0.0147	Parkinson's disease	0.0039
body	0.0145	simple reflexes	0.0030
vertebrates	0.0142	phenomena are identical	0.0030

Table 5.4: Explicit descriptor – *Ácido desoxirribonucleico* Portuguese document.

Single-word	***Tf-Idf(.)***	**Multi-word**	***Tf-Idf(.)***
DNA	0.0962	dupla hélice	0.0136
ADN	0.0805	informação genética	0.0112
cadeia	0.0418	pontes de hidrogênio	0.0080
bases	0.0364	dupla cadeia	0.0056
proteínas	0.0315	cadeias de ADN	0.0056
células	0.0257	sequência de DNA	0.0056
dupla	0.0227	cadeia simples	0.0055
sequências	0.0214	DNA de cadeia	0.0055
transcrição	0.0205	cadeia de ADN	0.0048
sequência	0.0199	material genético	0.0047

Table 5.5: Explicit descriptor – *Immunsystem* German document.

Single-word	***Tf-Idf(.)***	**Multi-word**	***Tf-Idf(.)***
Zellen	0.0424	angeborene Immunabwehr	0.0082
Immunsystem	0.0423	zytotoxischen T-Zellen	0.0065
Immunsystems	0.0351	dendritische Zellen	0.0065
Erreger	0.0338	angeborenen Immunabwehr	0.0065
Immunabwehr	0.0334	körpereigene Zellen	0.0062
T-Zellen	0.0265	Zellen des Immunsystems	0.0061
Makrophagen	0.0210	adaptiven Immunabwehr	0.0049
Infektion	0.0207	adaptive Immunabwehr	0.0049
Granulozyten	0.0206	Antigene erkennen	0.0049
Krankheitserreger	0.0204	schweren Ketten	0.0048

Table 5.6 shows the evaluation of the approach. Precision gives the average rate of correct keywords in each descriptor; Recall measures the rate of concepts that did not need to be exchanged by "better" keywords outside the descriptor. As reviewers' agreement was not 100%, these values were measured assuming the majority of their choices.

Table 5.6: Results for the explicit keyword extraction using *Tf-Idf* with concepts.

Corpus	**Single-words** Precision	Recall	**Multi-words** Precision	Recall
English	0.89	0.80	0.87	0.79
Portuguese	0.88	0.86	0.91	0.83
German	0.89	0.89	0.85	0.80

The results are quite similar for the three languages, despite some slight differences, and show that the combination of *Tf-Idf* with the *ConceptExtractor* is able to extract document keywords to build explicit descriptors. The reader could be tempted to assume that the application of *Tf-Idf* to all sequences of words in a document could provide similar results, but as I'll show in the next subsection, *Tf-Idf* does not handle multi-words well.

5.3.2 Comparative results

I have also compared the use of *Tf-Idf* only on concepts (referred to as ***Explicit*** in the comparison tables) with other extraction methods. Table 5.7 compares the *Explicit* method with *Tf-Idf (without concepts)*, while Table 5.8 compares it with *LeastCv*, *LeastRvar* and *Mk[2.5]*, as described in [SL10]. *Tf-Idf (without concepts)* is the use of *Tf-Idf* applied to all words and multi-words in a document whether or not they are concepts.

Table 5.7: Comparison of methods for explicit document descriptors – single-words

Approach	**Parameter**	**English**	**Portuguese**	**German**
Explicit	Precision	0.89	0.88	0.89
	Recall	0.80	0.86	0.89
Tf-Idf (without concepts)	Precision	0.87	0.86	0.87
	Recall	0.79	0.86	0.88

Table 5.8: Comparison of methods for explicit document descriptors – multi-words

Approach	**Parameter**	**English**	**Portuguese**	**German**
Explicit	Precision	0.87	0.91	0.85
	Recall	0.79	0.83	0.80
Tf-Idf (without concepts)	Precision	0.50	0.52	0.49
	Recall	0.35	0.38	0.37
LeastCv	Precision	0.63	0.62	0.59
	Recall	0.57	0.61	0.62
LeastRvar	Precision	0.65	0.64	0.61
	Recall	0.66	0.64	0.63
Mk[2.5]	Precision	0.73	0.71	0.75
	Recall	0.69	0.72	0.71

The *Explicit* method, which is *Tf-Idf* applied to the concepts, scores higher than the others methods. Although *Tf-Idf (without concepts)* shows similar results to the *Explicit* method for single-words, it scores poorly for multi-words. This happens because *Tf-Idf* tends to assign high values to rare sequences, such as "do ADN" and "ADN é" ("of DNA" and "DNA is", respectively) which in this case occurs only in the Portuguese *Ácido desoxirribonucleico* document. These kind of sequences are filtered when multi-word concepts are extracted, hence the good results of *Tf-Idf* with concepts.

As for *LeastRvar* and *Mk[2.5]* (the *Mk* metric uses *LeastRvar* under the hood), their lower results are due to the fact that the *LeastRvar* metric tends to benefit multi-words which are rare in the documents, including the documents for which the keywords are being retrieved. For instance, "STOCK EXCHANGE" (all characters uppercased – from the card *Advance to Stock Exchange*) is considered by *LeastRvar* and *Mk[2.5]* as the first ranked keyword of the *Monopoly* document, and considered much better than "Stock Exchange". This happens because the *all-uppercase* term is quite rare (it occurs only 2 times and only in the *Monopoly* document) while *Stock Exchange* is quite frequent in the *Monopoly* document, and it occurs also in other documents. Both methods benefit too much the rare terms (and often, odd terms) rather than less rare terms with a slightly wider meaning. For instance, for the *Explicit* method, "Stock Exchange" is the third ranked concept in the *Monopoly* document, while the first ranked one is "Parker Brothers", the publisher's name. Both terms appear in other documents beside the *Monopoly* document, although without much relevance.

5.4 The implicit descriptor

The implicit descriptor of a document is a set of keywords that do not occur explicitly in a document but whose meanings are semantically related with the content of the document. For instance, a document may focus on topics such as "air pollution", "carbon monoxide" and "ground level ozone", but concepts such as "lung cancer" or "water cycle", although not occurring explicitly in the document, could enrich the global document descriptor, since they are semantically related with its content. A richer descriptor provides an extended semantic scope which may be useful in Information Retrieval or Web Search applications, just to name a few examples.

Basically, and in a practical sense, the implicit keywords of a document are concepts from other documents of a corpus which have strong *Semantic Proximity* values with most of the keywords of the document's explicit descriptor. The *Semantic Proximity* is composed of two factors – the *inter-document proximity* and the *intra-document proximity*, which will be explained in the next subsections.

5.4.1 Inter-document proximity – correlation

If we consider a collection of documents from different subjects, there is a high probability that terms that are specific to a certain subject appear only in documents that deal with this subject. Therefore, we can consider that these terms may be related at a subject level.

In a practical way, the idea behind the *Inter-document Proximity* is that, two terms A and B with the tendency to occur in the same set of documents of a collection (considering, say, the natural diversity of subjects in a collection) are probably related at a specific subject level. In this sense, they can be considered *semantically close*. To measure the tendency for a pair of terms A and B to co-occur in the same documents of a collection, I use $Corr(A, B)$ which is given by equation 5.9.

$$Corr(A, B) = \frac{Cov(A, B)}{\sqrt{Cov(A, A)} \, . \, \sqrt{Cov(B, B)}} \; . \tag{5.9}$$

$$Cov(A, B) = \frac{1}{\|\mathcal{D}\| - 1} \sum_{d_i \in \mathcal{D}} d(A, d_i) \, . \, d(B, d_i) \; . \tag{5.10}$$

$$d(A, d_i) = p(A, d_i) - p(A, .) \qquad d(B, d_i) = p(B, d_i) - p(B, .) \; . \tag{5.11}$$

$$p(A, d_i) = \frac{f(A, d_i)}{size(d_i)} \qquad p(A, .) = \frac{1}{\|\mathcal{D}\|} \sum_{d_i \in \mathcal{D}} p(A, d_i) \; . \tag{5.12}$$

Equation 5.9 is based on Pearson's correlation coefficient. $Corr(A, B)$ measures the covariance of terms (A, B) along the collection of documents $\mathcal{D}$. In the previous equations, $\|\mathcal{D}\|$ is the number of documents of the collection, d_i is the *i-th* document in $\mathcal{D}$, $size(d_i)$ is the number of words in document d_i and $f(A, d_i)$ is the frequency of term A in document d_i. $Corr(A, B)$ values ranges from -1 to $+1$: it gets negative results when A tends to occur in documents where B does not, values near zero occur when the correlation is weak, and values close to $+1$ when the correlation tends to be strong. Tables 5.9, 5.10 and 5.11 show $Corr(A, B)$ values for some pairs of terms from the tested corpora.

Table 5.9: Correlation values for some pairs in the English corpus.

Term A	**Term B**	***Corr(A,B)***
suanpan	Chinese abacus	1.000
Social anarchism	Collectivist anarchism	1.000
Anarchism	Anarchists	0.908
supply	demand	0.809
opera	La Clemenza di Tito	0.614
Rossini	opera	0.444
Kigali	Rwanda	0.435
airplane	automobile	0.023
Microsoft Windows	fail	0.008
car	computer	0.006

Table 5.10: Correlation values for some pairs in the Portuguese corpus.

Term A	**Term B**	***Corr(A,B)***
U2	Bono	0.987
Python	Guido van Rossum	0.962
Aristóteles	Platão	0.856
Anarquismo	Anarquista	0.813
John Lennon	Beatles	0.727
Nazi	Hitler	0.525
John von Neumann	computador	0.303
electricidade	aeroporto	0.010
carro	computador	0.005
Tio Patinhas	generosidade	0.002

Table 5.11: Correlation values for some pairs in the German corpus.

Term A	**Term B**	***Corr(A,B)***
organische Säure	Ascorbinsäure	0.965
Dennis Hopper	Born to Be Wild	0.950
Verdauung	Digestion	0.775
Anime	Hentai	0.697
Microsoft	Windows	0.587
Turbo Pascal	Compiler	0.531
Betriebssystem	Windows	0.136
Strom	Flughafen	0.009
Flugzeug	Lebensmittel	0.005
Sonne	Auto	0.001

Tables 5.9, 5.10 and 5.11 show that the higher $Corr(A, B)$ values are for pairs that are related, while lower values are for pairs which bear no relation. For instance, "Social anarchism" and "Collectivist anarchism" in the English corpus are 100% correlated mainly because they occur only in one document (the same document, English wikipedia article *Anarchism*). On the other hand, pairs such as "computer" and "car" are not related since they do not occur consistently in the same set of documents.

However, there are pairs that are highly related but whose $Corr(A, B)$ value is just moderately high. For instance, considering the pair "Nazi" and "Hitler" in the Portuguese corpus, although the pair is known for being highly related, both words tend to occur isolated of each other on the documents of the collection. By themselves, "Nazi" and "Hitler" are subjects relatively used in other contexts beside the one they have in common (for instance, there is a document briefly comparing *Hitler* to other dictators in Europe, without mentioning the *Nazi party*). Similarly, in the English corpus, although the pair "airplane" and "automobile" is related by the fact of both being means of transportation, they do not tend to occur in the same set of documents. This means that there is no collection-wide subject about means of transportation in that corpus.

5.4.2 Intra-document proximity – word distance

The correlation between pairs of terms, as mentioned in the previous subsection, has the problem of not being sensitive to the specific and local information inside documents. For instance, in the English corpus, the correlation between "suanpan" and "Chinese abacus" is 1.0, which is the same value as the correlation between "suanpan" and "Babylonian abacus", since all these terms occur only in the English *Abacus* document. However, since "suanpan" is a Chinese abacus, it should be desirable to have "suanpan" *more strongly related* with "Chinese abacus" than with "Babylonian abacus". In fact, inside the *Abacus* document, "suanpan" occurs in the same section as "Chinese abacus", while "Babylonian abacus" occurs five sections before. This led to the creation of the *Intra-document Proximity*.

The idea behind this metric is that two terms are more strongly related if they tend to occur near each other inside a document. Thus, the *Intra-document Proximity* between two terms A and B is defined as:

$$IP(A,B) = 1 - \frac{1}{\|\mathcal{D}^*\|} \sum_{d \in \mathcal{D}^*} \frac{dist(A,B,d)}{farthest(A,B,d)} \; . \tag{5.13}$$

$$dist(A,B,d) = \sum_{o_i \in Occ(A,d)} nearest(o_i,B,d) + \sum_{o_k \in Occ(B,d)} nearest(o_k,A,d) \; . \tag{5.14}$$

In equation 5.13, $\mathcal{D}^*$ is the set of documents containing terms A and B, while $\|\mathcal{D}^*\|$ is the number of documents in that set. In equation 5.14, $Occ(A,d)$ stands for the set of all occurrences of A in document d, while $nearest(o_i,B,d)$ gives the distance, in words, from occurrence o_i to the nearest occurrence of B in d, distances being positive numbers.

Considering equation 5.13, $dist(A,B,d)$ represents a global distance between A and B, considering all occurrences of both terms in d. This distance is normalized by the maximum global distance between A and B considering all possible distributions of occurrences in d, which is given by $farthest(A,B,d)$. This *extreme case* happens when all occurrences of one term are located at the beginning of d and the occurrences of the other term, at the end of document d. $farthest(A,B,d)$ is given by:

$$farthest(A,B,d) = C_1 - C_2 + C_3 - C_4 \; . \tag{5.15}$$

$$\begin{aligned} C_1 &= f(A,d)\,.\,(size(d) - f(B,d)) & \quad C_2 &= \frac{(f(A,d)-1)^2 + f(A,d) - 1}{2} \\ C_3 &= f(B,d)\,.\,(size(d) - f(A,d)) & \quad C_4 &= \frac{(f(B,d)-1)^2 + f(B,d) - 1}{2} \end{aligned} \; . \tag{5.16}$$

On equation 5.16, $f(A,d)$ and $f(B,d)$ are the number of occurrences of terms A and B in document d, while $size(d)$ represents the total number of words on document d.

$farthest(A, B, d)$ is the sum of the closest distances between the occurrences of terms A and B where all the occurrences of A are located contiguously at the beginning of document d and all the occurrences of B are located contiguously at the end of d.

To calculate $farthest(A, B, d)$, be $f(A, d)$ and $f(B, d)$ the number of occurrences of A and B in d. Then, the closest occurrence of B from the 1$^{\text{st}}$ occurrence of A (the one at the very beginning of d) is at distance $size(d) - f(B, d)$, where $size(d)$ is the number of words of d. Similarly, the closest occurrence of B from the 2$^{\text{nd}}$ occurrence of A is at distance $size(d) - f(B, d) - 1$. The last occurrence of A (the one in the $f(A, d)^{\text{th}}$ position), is finally at distance $size(d) - f(B, d) - f(A, d) + 1$, and the sum of all these distances is given by $f(A, d) \times (size(d) - f(B, d)) - \sum_{i=1}^{i=f(A,d)-1} i$. Similarly, the same reasoning is valid regarding the distances from all occurrences of B to occurrences of A in d. Because $\sum_{i=1}^{i=f(A,d)-1} i$ is equal to $((f(A, d) - 1)^2 + f(A, d) - 1)/2$, so $farthest(A, B, d)$ is equal to $C1$-$C2$+$C3$-$C4$.

Tables 5.12, 5.13 and 5.14 show $IP(A, B)$ values for some pairs of terms from the tested corpora.

Table 5.12: $IP(A, B)$ values for some pairs in the English corpus.

Term A	Term B	*IP(A,B)*
suanpan	Chinese abacus	0.966
suanpan	Babylonian abacus	0.665
airplane	automobile	0.807
airplane	crash	0.760
airplane	disease	0.704
airplane	electricity	0.621
airplane	Mozart	0.000
health	disease	0.798
health	computer	0.699
health	opera	0.657

Table 5.13: $IP(A, B)$ values for some pairs in the Portuguese corpus.

Term A	Term B	*IP(A,B)*
Mozart	Wolfgang	0.815
Mozart	ópera	0.807
Mozart	piano	0.804
Mozart	clarinete	0.661
Mozart	Barack Obama	0.000
Nova Iorque	Manhattan	0.901
Nova Iorque	Wall Street	0.870
Nova Iorque	Brooklyn	0.838
Nova Iorque	Estados Unidos da América	0.767
Nova Iorque	Angola	0.634

Table 5.14: $IP(A, B)$ values for some pairs in the German corpus.

Term A	**Term B**	***IP(A,B)***
Diode	Strom	0.837
Diode	Silizium	0.728
Diode	p-n	0.657
Diode	Hochfrequenz	0.467
Diode	reiten	0.000
Turbo Pascal	Compiler	0.886
Turbo Pascal	Programmiersprache	0.804
Turbo Pascal	Entwicklungsumgebung	0.790
Turbo Pascal	Prolog	0.751
Turbo Pascal	Software	0.362
Turbo Pascal	Bildschirm	0.000

Unlike previous tables where the information is ranked by the score of the metric being presented, the pairs in tables 5.12, 5.13 and 5.14 are combined by decreased semantical relevance with the term in the left column. For instance, in Table 5.12, for the English corpus, "suanpan" is more *semantically close* to "Chinese abacus" than "Babylonian abacus" as it was intended. Still in the same table, "airplane" is *semantically closer* to "automobile" (both are means of transportation) and "crash" than with "disease", "electricity" or "Mozart". For the Portuguese corpus, Table 5.13, good examples of the $IP(A, B)$ metric are also given. For instance, "Nova Iorque" (New York) is computed by $IP(A, B)$ as being *semantically closer* to New York streets ("Wall Street"), boroughs ("Brooklyn" and "Manhattan"), and "Estados Unidos da América" (United States of America) than "Angola". As most people knows, New York is a city in the United States of America, and not in Angola, the African country (although there is a village called Angola in New York State, and maybe that is why the pair still gets a score of 0.634 instead of a much lower one). For the German examples in Table 5.14, the fact that "Turbo Pascal" is more a "Programmiersprache" (Programming Language) than a "Software" or "Bildschirm" (screen) provides also a good evidence of the results of this metric.

5.4.3 Semantic Proximity

Finally, the Semantic Proximity between two terms A and B is defined as the multiplication of $Corr(A, B)$ by $IP(A, B)$. However, since it was intended to use *intra-document proximity* ($IP(.)$) only as a tuning factor to discriminate cases such as the one of "suanpan" and "Chinese abacus", it was preferred to add more weight to the $Corr(A, B)$ factor in the calculation of the Semantic Proximity, hence the square root on $IP(A, B)$:

$$SemProx(A, B) = Corr(A, B) \,.\, \sqrt{IP(A, B)} \,. \tag{5.17}$$

Table 5.15 shows some examples of pairs of terms and their $SemProx(.)$ values from the English corpus.

Table 5.15: $SemProx(A, B)$ values for some pairs in the English corpus.

Term A	Term B	*SemProx(A,B)*
diesel	engines	0.68
Ocean earthquake	tsunami	0.65
natural hazard	earthquakes	0.54
earthquake	tsunami waves	0.49
Google	engine	0.11

The *Semantic Proximity*, as it can be seen from the examples in this table, allows to quantify the semantic relatedness of a pair of terms. Higher values are for pairs for which their meanings are more related than for pairs for which we can recognize a greater semantic distance. For instance "diesel" and "engines", "Ocean earthquake" and "tsunami", vs “Google” and “engine”.

5.4.4 Ranking implicit concepts

As mentioned in the introduction of this chapter (section 5.1), to extract the implicit keywords of a document, the *Semantic Proximity* is calculated between concepts extracted from the corpus and each keyword of the document's explicit descriptor. The first ranked concepts are selected as the document's implicit keywords and form the document implicit descriptor.

So, for a document d, let k_i be the i-th ranked keyword of the explicit descriptor of d. If C is a concept not occurring in d but strongly related to most of the explicit keywords in d, then C is a strong candidate as an implicit keyword of d. Therefore, the following metric measures how a concept C is ranked as being an implicit keyword of d:

$$score(C, d) = \sum_{i=1}^{n} \frac{SemProx(C, k_i)}{i} \ . \tag{5.18}$$

In $score(C, d)$, n is the size of the explicit descriptor of d, which was set to 20 as referred. So, equation 5.18 considers the Semantic Proximity between the concept C and each explicit keyword k_i in d. It also considers the ranking of keyword k_i in the explicit descriptor, which is i. In this way, concepts which are strongly related with the top explicit keywords in the explicit descriptor (lower values of i) are considered more descriptive of document d, than concepts that are related with the bottom explicit keywords (higher values of i). Finally, since we are applying a *sum*, the greater the number of explicit keywords of d an implicit concept strongly relates with, the higher the probability that it gets a good score in the implicit descriptor.

Table 5.16 shows the first ranked implicit keywords for document *Economics* of the English corpora, as well as the $score(.)$ and $SemProx(.)$ values for the pairs. For comparison, Table 5.17 shows the ranked content of the explicit descriptor of the same document.

Table 5.16: First implicit keywords of the English Wikipedia *Economics* article.

Concept	***score(.)***	***SemProx(.)***	**Explicit Keyword**
supply curve	1.83	0.95	quantity supplied
		0.93	quantity demanded
		0.85	price
		0.82	supply
		0.79	quantity
demand curve	1.75	0.92	demand
		0.91	quantity supplied
		0.88	quantity demanded
		0.82	price
		0.75	quantity
Austrian school	0.45	0.24	economic
		0.23	economics
		0.15	Keynesian economics
		0.11	theory
		0.11	classical economics
		0.10	price
mercantilism	0.40	0.56	classical economics
		0.20	economics
		0.20	economic
Thomas Malthus	0.39	0.38	classical economics
		0.23	economics
		0.11	economic
		0.10	theory

Table 5.17: Explicit descriptor of the English Wikipedia *Economics* article.

Rank	**Single-word**	**Multi-word**
1	economics	quantity demanded
2	economic	quantity supplied
3	supply	mainstream economics
4	demand	classical economics
5	price	Keynesian economics
6	quantity	neoclassical economics
7	analysis	price stickiness
8	theory	Labor economics
9	market	John Stuart Mill
10	economy	Stuart Mill

From Table 5.16, "supply curve" and "demand curve" are the top implicit keywords. As it can be seen, they relate very strongly with the first ranked explicit multi-word concepts ("quantity supplied" and "quantity demanded") and with the 3rd to 6th ranked explicit single-word keywords. Keyword "Austrian school" (which is a school of economic thought, by the way), although it does not have strong semantic relations with the explicit keywords, as the previous examples, it is ranked third because it relates with many explicit keywords in the explicit descriptor. Finally, both "mercantilism" and "Thomas Malthus" (a British economist from the 18th century) are moderately related with "classical economics", "economics" and "economic", which are ranked in the first positions of the explicit descriptor.

5.4.5 Experimental conditions and results

For evaluating the results concerning the implicit descriptors, I used the same corpora as mentioned in Table 5.1 (English, Portuguese and German Wikipedia documents of several different and random subjects) and the same documents for which the explicit keywords were extracted (titles of the documents are in Table 5.2). Thus, for each document d of each language test-set, the following process was used:

- Take the 20 explicit keywords (10 single-words and 10 multi-words) of document d.
- Compute the $SemProx(C, k_i)$ between each concept C extracted from the corpus, but not occurring in d, and each explicit keyword k_i of d.
- Then compute $score(C, d)$.
- Finally, take the first 20 concepts ranked by $score(., d)$ and consider them as the implicit descriptor of d.

Tables 5.18, 5.19 and 5.20 show the first ten implicit keywords of documents from the different corpora.

Table 5.18: First ten implicit keywords of the English Wikipedia *Brain* document.

score(.)	**Implicit keyword**
1.465	peripheral nervous system
1.277	transverse nerves
1.276	CNS
0.666	Purkinje
0.664	Purkinje cells
0.663	cerebellar cortex
0.663	granule cells
0.661	cerebellar nuclei
0.659	Purkinje cell
0.650	cerebellum

For each implicit descriptor an evaluation of the Precision results was made. The criterion followed by the reviewers was that an implicit keyword should be accepted as

Table 5.19: First ten implicit keywords of the Portuguese Wikipedia *Teosofia* document.

score(.)	**Implicit keyword**
2.798	Ísis sem Véu
1.727	Olcott
1.351	fenómenos psíquicos
1.320	pesquisas psíquicas
0.876	tradições religiosas
0.676	Grécia antiga
0.459	relações sexuais
0.295	Sociedade Torre de Vigia
0.214	Testemunhas de Jeová
0.198	Budismo

Table 5.20: First ten implicit keywords of the German Wikipedia *Immunsystem* document.

score(.)	**Implicit keyword**
0.530	Komponenten des Immunsystems
0.509	Reaktion des Immunsystems
0.431	Eukaryoten
0.410	Lymphozyten
0.395	eukaryotischen Zellen
0.358	Zellteilung
0.348	Adolf von Behring
0.348	Emil Adolf von Behring
0.327	Hormone
0.321	Antikörpern erkannt

correct only if they recognized that, although not occurring in the document, the keyword was semantically related to its contents. Recall was not evaluated since it would be impractical to find concepts in about 2000 other documents of the corpora (or 4000 in the case of the German corpus) which could be considered better than some of the implicit keywords. Table 5.21 shows the measured Precision results.

Table 5.21: Precision values for the implicit descriptors.

Corpus	**Precision**
English	0.84
Portuguese	0.87
German	0.83

Although the results are slightly lower than those obtained for the explicit descriptors, I believe that they are still good enough for applications benefiting from the extension of the semantic scope of each document. The global computation time for building all explicit and implicit descriptors took about 2 hours for each language, in a relatively modern computer (Intel Core 2 Duo, 4 GB RAM, Linux Ubuntu OS).

5.5 Summary

In this chapter I have presented a language-independent method for the automatic building of document descriptors formed by explicit and implicit keywords. The method starts by identifying concepts on the documents that are then used as explicit keywords. It was shown that, for this task, *Tf-Idf* returns the best results when using concepts, especially for multi-words.

I have also proposed metrics to identify semantic relations between terms in order to measure the relevance of a concept as implicit keyword of a document. Implicit keywords offers an extended semantic scope to the global descriptors of documents, with great applicability.

This methodology is independent of any language-specific tools, as I've tried to show by obtaining similar results for the different languages.

6

Extracting semantic relations from standalone documents using clusters of concepts

The extraction of semantic relations from texts is currently gaining increasing interest. However, a large number of current methods are language and domain dependent, and statistical and language-independent methods tend to work only with large amounts of text. This leaves out the extraction of semantic relations from standalone documents, such as single documents of unique subjects, reports from very specific domains, or small books.

A method to extract semantic relations inside documents was presented in the previous chapter. However, to measure semantic relatedness inside standalone documents only by means of distances between words is not without its flaws. Inconsistencies arise, for instance, when words of two different paragraphs are considered semantically related only because the paragraphs are near each other, even when the paragraphs are semantically unrelated at a lower level.

In this chapter, I will present a statistical method to extract semantic relations from standalone documents using clusters of concepts. Clusters are areas in the documents where concepts occur more frequently. When clusters of different concepts occur in the same areas, they may represent highly related concepts.

This method is language independent and comparative results for three different European languages will be shown. The work in this chapter was published in [VS13b].

6.1 Introduction

The extraction of semantic relations between concepts is a hot topic. Semantic relations between concepts have been used with several degrees of success in various *Natural Language Processing* applications, such as word sense disambiguation [PP06], query expansion [HTC06], document categorization [TYB03], question answering [SJFHTsK05] and semantic web applications [SAK03].

However, most methodologies for the extraction of semantic relations from texts have scalability issues. For instance, while some methods extract semantic relations by exploring syntactic patterns in texts, others use external semantic lexicons such as thesauri, ontologies or synonym dictionaries. These kind of approaches are deeply language and domain dependent. On the other hand, most statistical methods are language-independent but tend to have the need for large amounts of text in order to be effective.

This poses a problem for the extraction of semantic relations from standalone documents. Standalone documents are, essentially, isolated or single documents, such as documents of unique subjects or domains, reports from very specific fields of expertise or even small books. The specificity of some fields of expertise in some of these documents may imply that no external ontologies exist for those domains, and given the small amount of text in those documents, statistical methods, with their correlation-like metrics, are not efficient. As these isolated and autonomous documents are also a source of knowledge, a local, more document-centric analysis is required.

In chapter 5, I've presented a method to extract semantic relations inside documents using the distance between words. However, relying on the distance between words to infer semantic relatedness has some flaws. Consider the following quotation which shows two successive paragraphs from the English Wikipedia *Arthritis* article:

> **Lupus**
> Lupus is a common collagen vascular disorder that can be present with severe arthritis. Other features of lupus include a skin rash, extreme photosensitivity, hair loss, kidney problems, lung fibrosis and constant joint pain.
>
> **Gout**
> Gout is caused by deposition of uric acid crystals in the joint, causing inflammation. There is also an uncommon form of gouty arthritis caused by the formation of rhomboid crystals of calcium pyrophosphate known as pseudogout. (...)

Figure 6.1: Two successive paragraphs from the *Arthritis* article – English Wikipedia.

Although these paragraphs occur near each other, there is no clear evidence that **hair loss** or **extreme photosensitivity**, from the *Lupus* paragraph, is related with *rhomboid crystals of calcium pyrophosphate* from the *Gout* paragraph.

This chapter presents a statistical and language-independent method for the extraction of semantic relations between concepts in standalone documents. We start by extracting the concepts from a document, and for each concept, we identify its clusters. Since relevant concepts on a document tend to form clusters in certain specific areas, clusters occurring in the same areas may represent highly related concepts. Although we are able to measure the degree of semantic relatedness between concepts, the type of relation (still) cannot be inferred.

This chapter is structured as follows: the next section reviews the related work. Section 6.3 presents the method for the identification of clusters and for the extraction of semantic relations from them. Section 6.4 shows the results of this approach. In section 6.5 it will be briefly shown how this methodology may work on collections of documents and, finally, section 6.6 presents the conclusions for this chapter.

6.2 Related work

Current surveys on the matter of the discovery of semantic relations between concepts on unstructured texts ([WLB12], [Bie05], [GMM03]) have identified at least three classes of approaches: linguistic approaches, approaches which use external lexicons, and statistical approaches. In the following subsections, I will review some of the related work in order to frame the reader in the general shortcomings of current methods.

6.2.1 Gre93 – a comparison of a linguistic and a window-based approach

The paper of Grefenstette [Gre93] presents an evaluation of techniques for the automatic extraction of semantic relations in large corpora, namely a syntactic and a window-based approach. The first technique, the linguistic one, extracts the context of each word, throughout a corpus which was previously divided into lexical units via a regular grammar. A list of context-free syntactic categories in a normalized form is assigned to each lexical unit. Another grammar selects a most probable category for each word, and finally a syntactic analyzer chunks nouns and verb phrases, and creates syntactic relations within and between chunks. The context of a noun are all the adjectives, nouns and verbs for which the noun has syntactic relations with. The second technique consists of the analysis of the neighborhood of a noun within a fixed-sized window. The neighbors are looked up in a lexicon for their probable Parts-of-Speech and, finally, the context of a noun are all nouns, adjectives and verbs inside the window up to a distance of ten, all within the same sentence.

Once the contexts of each noun are derived, their similarities are compared using a weighted *Jackard* measure. For each noun, another noun whose context is the most similar is elected. Results are evaluated (Grefenstette uses *Roget's Thesaurus* and an online dictionary as gold-standards), and the syntactic approach is considered superior for the general cases, while the window-based approach is considered to favor rare words.

The syntactic approach is clearly language-dependent. On the other hand, deriving the context of a noun, in the window-based approach, by its immediate 20 neighbors, may not be sufficient to identify all possible semantic relations in texts.

6.2.2 Wanderlust – a linguistic approach using Dependency Grammar Patterns

Wanderlust [AB09] is a procedure which uses deep linguistic patterns to extract semantic relations from natural language texts. The main hypothesis behind the algorithm is that certain grammatical structures exist which are universally valid and therefore allow for the extraction of arbitrary semantics.

The method works as follows: the authors start with a deep linguistic analysis of sentences using a *link grammar*. This *link grammar* connects terms by means of their grammatical relations. For instance "D" is used to connect a determinant to a noun, while "S" is used to connect a subject to a verb. If a direct relation does not exist between a pair of terms, an *indirect* connection (via intermediate terms) is used. These paths are called *linkpaths*. However, not all *linkpaths* are considered valid, specially when they belong to terms which are not explicitly related in a sentence. In this case, the authors have classified a set of valid *linkpaths* and computed a coefficient based on the frequency of the positive cases.

The authors then proceed with an use case on Wikipedia articles and discuss their results, including possible errors. However, this approach is clearly language-dependent.

6.2.3 NHN08 – a linguist approach to extract semantic relations from Wikipedia *hyperlinks*

A more recent trend in the extraction of semantic relations is the usage of semi-structured textual resources, such as Wikipedia. In [NHN08], the authors present a method which explores the *hyperlink tags* in Wikipedia texts.

The method starts with the preprocessing of a Wikipedia document. Specifically, they trim the document into sentences, chunk sentences into semantic phrases, and tag the individual words with their Part-of-Speech. After this preprocessing, sentences are parsed into a structure tree and *hyperlink* tags, where they occur, are also added to the tree. Later, the type of semantic relation between the entities is extracted, using previously defined syntactic patterns, and dividing the object (i.e, whatever occurs **before** the *type-of-relation* pattern) from the subject (i.e., whatever occurs **after** the *type-of-relation* pattern). Finally, objects and subjects are semantically identified with the help of their *hyperlinks* or by using other syntactic patterns when there is no information on the *hyperlinks* (or there is no *hyperlinks*).

This method is clearly language-dependent as its usage for other languages may imply the rewriting of most patterns.

6.2.4 MN03 – an approach using external lexicons

A paper by Mohit and Narayanan [MN03] represents another class of approaches, those which use external lexicons. In this paper, the authors have compiled a set of 100 news stories from the Yahoo News Service, with topics related to Criminal Investigation. Then, they used FrameNet [BFL98] to compile a lexicon from crime related frames, such as "Arrest", "Detain" and "Verdict".

Next, with a system named *GATE*, they have compiled a precise set of patterns and evaluated manually the performance of the system. Since they had low recall values, they used Wordnet [Mil95], another lexicon, to extend the crime related lexicon. To extend the lexicon, the authors used a metric that considered the frequency of occurrence of Wordnet nodes in the first extraction with the frequency of occurrence of the nodes in the general text. However, this methodology is clearly language-dependent.

6.2.5 RAC05 – identification of lexical patterns using Wordnet

The work in [RCAC05] presents another approach that uses external lexicons. In this case, it is oriented towards the extraction of lexical patterns that may represent semantic relations between concepts on Wikipedia articles.

Their procedure starts with the collection of entries from Wikipedia documents and their disambiguation using Wordnet. The output is a list of disambiguated entries. The next step consists of the extraction of patterns representing semantic relations between the entries. For that, the authors use the *hyperlinks* from unknown concepts to concepts already in the disambiguated list. If a relation is found in Wordnet, the sentence where the *hyperlink* occurs is collected in its Part-of-Speech form. The third step consists of the derivation of lexical patterns from the collected sentences. To do that, an edit distance calculation is used to group somewhat similar patterns.

Finally, patterns are generalized by joining the similar tokens of each group. With these patterns, the authors proceed to their experimentation on Wikipedia texts to identify new semantic relations on Wikipedia articles.

6.2.6 Bra06 – a statistical approach using Latent Semantic Analysis

[Bra06] is a paper that presents a method for the identification of semantic relations between entities using Latent Semantic Analysis. The method starts with the extraction of all named entities from a database of 158,492 English texts. All the extracted named entities are then given to an LSI algorithm (Latent Semantic Indexing) to be treated as indexing units in the creation of the LSI representation space. Since LSI vectors correspond grossly to the frequency of occurrence of the entities in the documents, a cosine metric is employed to measure the relatedness of any two vectors, and consequently, of any two entities.

The extraction of named entities is a subtle way of correcting the problems which LSI has when dealing with multi-words. Most uses of LSA/LSI are based on single-words.

The LSI table is usually built using the frequency of occurrence of a term or entity in the documents of a collection. This approach is somewhat similar to what I have done with the Correlation, in section 5.4.1. The major downside of this approach regarding standalone documents, is that it requires a large collection of documents in order to be effective.

6.2.7 PARR12 – a statistical approach using a KNN classifier

The work in [PARR12] presents a statistical approach for the extraction of semantic relations using a *K-Nearest Neighbor* algorithm. A KNN algorithm is a non-parametric method for classification and regression that predicts objects "values" or class memberships based on the k closest training samples in the feature space. For their experiment, the authors used a set of 327,167 Wikipedia documents and prepared two data-sets: one containing 775 words and another containing concept definitions (327,167 words). For each word of the smaller set, the training set, they used the data available in *DBPedia.org* (a community effort to extract structured information from Wikipedia) as definition of the word (in practice, a vector of defining words). Finally, they trained the KNN classifier with two different statistical measures: the gloss overlap of the definitions d_1 and d_2 of concepts c_1 and c_2 (as in equation 6.1) and the cosine between vectors f_1 and f_2 of definitions d_1 and d_2 (as in equation 6.2); further details in [PARR12].

$$similarity(c_1, c_2) = \frac{2\,.\,|d_1 \cap d_2|}{|d_1| + |d_2|}\ . \tag{6.1}$$

$$similarity(c_1, c_2) = \frac{f_1 \cdot f_2}{\|f_1\| \cdot \|f_2\|} = \frac{\sum_{k=1}^{n} f_{1k} \cdot f_{2k}}{\sqrt{\sum_{k=1}^{n} f_{1k}^2} \cdot \sqrt{\sum_{k=1}^{n} f_{2k}^2}}\ . \tag{6.2}$$

The paper indicates that both metrics return similar results. Although the authors use a statistical approach, the definitions of each word are obtained using an external lexicon, with all the shortcomings already mentioned regarding the language-dependence. Finally, this approach may also need a sufficient number of entities to derive relationships, a number which may not exist on standalone documents or small corpora.

6.2.8 TC03 – a comparison of statistical measures and methods

The paper by Terra and Clarke [TC03] presents a comparison of statistical metrics to measure similarity between words, and three approaches for extracting semantic relations from texts. The metrics are the *Pointwise Mutual Information*, χ^2*-test*, *Likelihood ratio*, *Average Mutual Information* for when contexts are not available. When contexts are available, the metrics are the *Cosine of Pointwise Mutual Information*, L_1 *norm*, *Contextual Average Mutual Information*, *Contextual Jensen-Shannon Divergence* and *Pointwise Mutual Information of Multiple words*.

To identify semantic relations in texts, the authors present a comparison between a window-oriented approach, a document-oriented approach and a syntax-based approach. The window-oriented approach, similarly to what was done in [Gre93], consists in the measurement of the frequency for which a pair of terms co-occur in the same window. On the other hand, the document-oriented approach consists of the measurement of the frequency for which a pair of terms co-occur in the same documents, quite similar to the *Correlation* presented in section 5.4.1. Finally, the syntax-based approach uses language-specific tools, such as parsers and Part-of-Speech taggers, to identify words of the "correct" grammatical categories to be used in conjunction with a document-based or window-based approach.

The best results are for a window-based approach using the *Pointwise Mutual Information* metric. Similarly to the work in [Bra06], correlation-like approaches tend to require large collections of documents in order to be effective, which is not the case of standalone documents or small corpora. On the other hand, the method to compute correlations using fixed-sized windows could work for standalone documents. However, from my observations, the distance between occurrences of some related concepts can be more than the 16 words which the authors propose.

6.3 Clusters of concepts – extracting semantic relations

Clusters of concepts occur when the distances between successive occurrences of a concept are less than what would be expected by chance. In other words, a cluster is a specific area in a text where a concept is relevant and tends to occur rather densely. For instance, consider the following paragraph from the English Wikipedia article *Arthritis*:

> **Gout.**
> **Gout** is caused by deposition of **uric acid** crystals in the joint, causing inflammation. (...) The joints in **gout** can often become swollen and lose function. (...) When **uric acid** levels and **gout** symptoms cannot be controlled with standard **gout** medicines that decrease the production of **uric acid** (e.g., allopurinol, febuxostat) or increase **uric acid** elimination from the body through the kidneys (e.g., probenecid), this can be referred to as refractory chronic **gout** or RCG.

Figure 6.2: A paragraph from the *Arthritis* article – English Wikipedia.

This paragraph is the only place, in the *Arthritis* article, where *gout* and *uric acid* occur. Since both concepts occur rather densely only in this paragraph, each one forms a cluster here. And since both concepts form a cluster in the same area, we consider the concepts to be highly related. Undoubtedly, *gout* and *uric acid* are related concepts ("**gout** is caused by deposition of **uric acid** crystals in the joint") and highly relevant in this paragraph.

6.3.1 Identifying clusters of concepts

In a formal way, a cluster of a concept exists where the distances between some of its successive occurrences are less than what would be expected by chance. So, the question is how to define the expected behavior of a concept on a document. Be $L_C = \{t_1, t_2, \cdots, t_m\}$ the list of the t_i positions where a concept C occurs in a document of size n. From L_C, we can obtain $\hat{u}_a$ (as in equation 6.3) which measures the average separation that would exist if C occurred uniformly (or randomly) on the document:

$$\hat{u}_a = \frac{n+1}{m} \,. \tag{6.3}$$

The underlying idea is that, for two successive occurrences (t_i, t_{i+1}) of C, if their separation is less than $\hat{u}_a$, both are part of a cluster, else, they are not.

Unfortunately, $\hat{u}_a$, as it is, tends to favor rare words. For instance, a concept which occurs 4 times in a document of size 2000 will have $\hat{u}_a \approx 500$. If the occurrences are spread over 4 successive paragraphs of size 200, the distances between successive occurrences of the concept will be always less than 500 – the maximum distance would be 400, for one occurrence in the beginning of one paragraph and the next occurrence in the end of the following paragraph. Thus, this rare concept will always form a cluster, but, instead of being highly concentrated on a single paragraph or two, the concept is weakly scattered over four paragraphs. To allow clusters over such distances may be too much, so we must impose an upper limit for such rare cases.

Figures 6.3, 6.4 and 6.5 show, on the left side, the number of paragraphs (y-axis) by paragraph size (x-axis, in words), and on the right side, the average number of words in a paragraph (y-axis) by document size (x-axis/10).

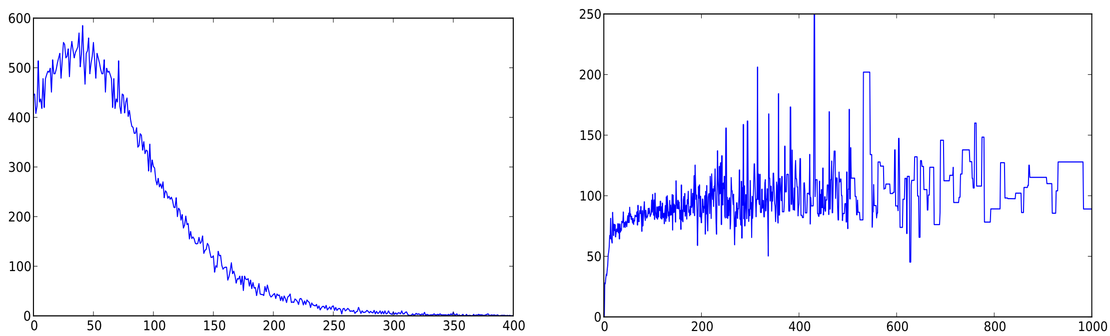


Figure 6.3: Paragraph analysis on a corpus of English documents.

As it is evident in the figures, the behavior of the paragraphs tends to be quite similar for all tested languages. On the left side, it can be seen that there is a peak of paragraphs, with about 50 words, and that 95% have less than 150 words. On the right side, except for small documents with less than 100–200 words, the average paragraph length is independent of the size of documents. For the purpose of this thesis, since we assume that clusters are somewhat associated with paragraphs (or parts of paragraphs), and since 95%

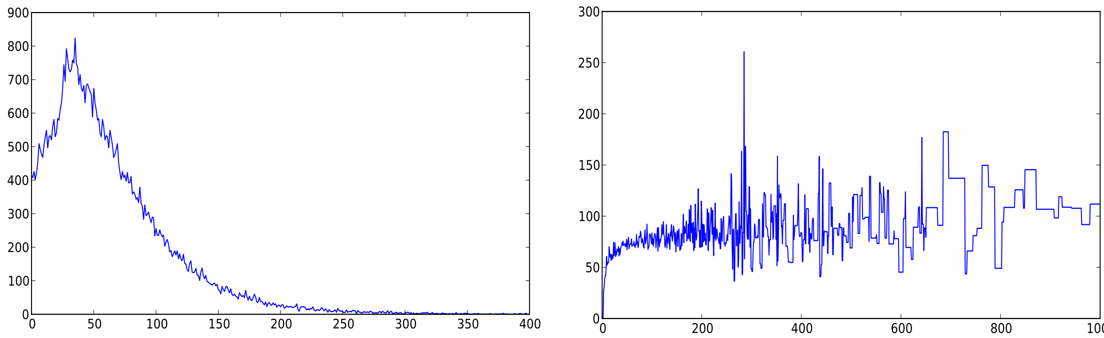


Figure 6.4: Paragraph analysis on a corpus of Portuguese documents.

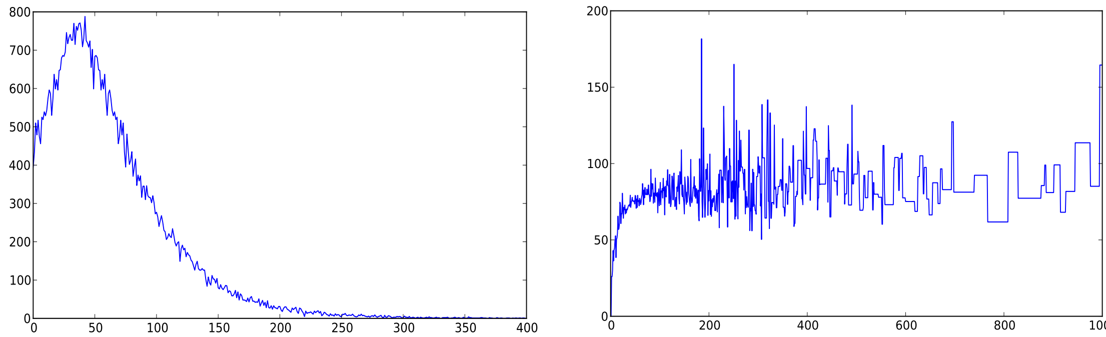


Figure 6.5: Paragraph analysis on a corpus of German documents.

of paragraphs have less than 150 words, I suggested an **upper limit of 150 words**. This means that no cluster may be formed where the distance between successive occurrences of a concept is greater than 150 words, independently of its frequency of occurrence.

On the other hand, $\hat{u}_a$ also tends to harm frequent concepts. For instance, in a typical document of size 2000, a relatively frequent concept, which may be one of the most relevant keywords, occurs in average 60 times ($\hat{u}_a \approx 33$). Since there is a great number of paragraphs that are about 50 words long, a frequent concept may not form clusters in those paragraphs, for instance, if it occurs only 2 times in the paragraph but in distinct edges, since the distance would be greater than *33*. Considering this, I suggested a **lower limit of 50 words**. This means that a cluster will always be formed where the distance between successive occurrences of a concept is less than 50 words, independently of its frequency of occurrence.

Formally, being $L_C = \{t_1, t_2, \cdots, t_m\}$ the list of the positions where concept C occurs, (6.4) measures the new proposed average separation to consider whether C occurs randomly in a document:

$$\hat{u} = \begin{cases} 150 & \text{, if } \hat{u}_a > 150 \\ 50 & \text{, if } \hat{u}_a < 50 \\ \hat{u}_a & \text{, otherwise} \end{cases} . \tag{6.4}$$

The next step consists of the calculation of the cohesions between successive occurrences of C, given by equation 6.5:

$$coh(t_i, t_{i+1}) = \frac{\hat{u} - d(t_i, t_{i+1})}{\hat{u}} \ . \tag{6.5}$$

$$d(t_i, t_{i+1}) = t_{i+1} - t_i \ . \tag{6.6}$$

Basically, the cohesion measures the distance between successive occurrences (t_i, t_{i+1}), proportional to $\hat{u}$. If the distance is small, the cohesion will tend to 1.0, else, it will tend to values less than zero. Zero stands as the frontier case, where the distance will be equal to $\hat{u}$.

The final step consists of traversing the L_C list and join together occurrences belonging to the same clusters, since a concept may form more than one cluster (or none). Figure 6.6 shows a pseudo-code sample for finding clusters in L_C.

```
def findClusters(Lc):
    clusterList = ClusterList()
    currCluster = Cluster()
    for (ti, ti+1) in Lc:
        if (coh(ti, ti+1) > 0):
            // Add the pair to the cluster
            currCluster.addPair(ti, ti+1)
            currCluster.addCohesion(coh(ti, ti+1))
        else:
            // If cluster is not empty
            if (currCluster.numberPairs() > 2):
                currCluster.computeAverageCohesion()
                clusterList.add(currCluster)
            // start a new empty cluster
            currCluster = Cluster()
    return clusterList
```

Figure 6.6: Pseudo-code for finding clusters in L_C.

The final cohesion value for each cluster is the arithmetic average of the positive cohesion values for the successive occurrences of the concept in the cluster. As it can be understood by the pseudo-code in Figure 6.6, no cluster can contain any pair of successive occurrences for which its cohesion is negative. Also, although not required, in my tests I enforce that a cluster, to be valid, must have at least 3 occurrences of the concept (or 2 pairs as in the pseudo-code).

6.3.2 Intersection and semantic closeness of clusters

As previously mentioned, the underlying idea is that a pair of concepts is highly related if they tend to make clusters in the same areas of a document. Thus, the purpose behind the intersection is to find whether two clusters occupy the same area of a text. So, for two clusters $C_A = \{p_{A1}, p_{A2}, \cdots, p_{An}\}$ and $C_B = \{p_{B1}, p_{B2}, \cdots, p_{Bm}\}$, where p_{Xi} is a position where concept X occurs in the text, the intersection is measured using equation 6.7:

$$intersection(C_A, C_B) = \frac{span(C_A, C_B)}{spanMin(C_A, C_B)} \ . \tag{6.7}$$

$$span(C_A, C_B) = min(p_{An}, p_{Bm}) - max(p_{A1}, p_{B1}) \ . \tag{6.8}$$

$$spanMin(C_A, C_B) = min(p_{An} - p_{A1}, p_{Bm} - p_{B1}) \ . \tag{6.9}$$

The size of a cluster is given by the difference between the rightmost and the leftmost positions of the concept in the cluster. Therefore, $spanMin(C_A, C_B)$ gives the size of the smallest cluster. On the other hand, $span(C_A, C_B)$ measures the size of the real intersection between clusters C_A and C_B. As for equation 6.7, $intersection(C_A, C_B)$, it measures essentially how much of the smaller cluster is intersected. Equation 6.7 returns values between $-\infty$ and 1.0, where 1.0 occurs when one cluster is completely inside the other, and values less than 0.0 occur when the clusters do not intersect.

Since we are now able to measure intersections between clusters, the Semantic Closeness for a pair of concepts (A,B) is measured using equation 6.10.

$$SC(A, B) = AvgIntersection(A, B) \, . \, AvgCoh(A) \, . \, AvgCoh(B) \ . \tag{6.10}$$

$AvgIntersection(A, B)$ is the average of all positive intersections between clusters of concepts A and B (i.e., only when $intersection(C_A, C_B) > 0$), and $AvgCoh(A)$ and $AvgCoh(B)$ stand for the average of all cohesions for all clusters of A and B respectively. Pairs of concepts for which their clusters are strongly intersected and the individual clusters are cohesive, are highly related. Tables 6.1, 6.2 and 6.3 show *Semantic Closeness* values between some pairs of concepts from documents of the tested corpora.

Table 6.1: Semantic Closeness for terms in the *Arthritis* article - English Wikipedia.

Term A	**Term B**	***SC(A,B)***
gout	gouty arthritis	0.671
gout	uric acid	0.604
rheumatoid arthritis	osteoarthritis	0.472
medications	exercise	0.067
rheumatoid arthritis	psoriatic arthritis	0.000
systemic	history	0.000

The tables clearly show that the results are quite balanced among all languages. Top results are for pairs of concepts whose relations are pretty obvious in the respective documents. For instance, in the English *Arthritis* article, *gout* is synonym of *gouty arthritis* and *uric acid* causes *gout*. In the Portuguese article, Aminoacyl-tRNA (*aminoacil-trna*) is

Table 6.2: Semantic Closeness for terms in the *Metabolismo* article - Portuguese Wikipedia.

Term A	Term B	*SC(A,B)*
gaminoacil-trna	aminoácidos	0.768
insulina	glicogénio	0.627
glicose	gluconeogénese	0.443
ácidos gordos	ácidos tricarboxílicos	0.282
via	energia	0.049
álcool	ferro	0.000

Table 6.3: Semantic Closeness for terms in the *Autismus* article - German Wikipedia.

Term A	Term B	*SC(A,B)*
intelligenz	sprachentwicklung	0.657
frühkindlichen autismus	atypischer autismus	0.512
autismus	sprachentwicklung	0.264
intelligenz	autismus	0.208
autismus	begriff	0.048
wissenschaftler	diagnosekriterien	0.000

an enzyme to which an amino acid (*aminoácido*) is cognated, and insuline (*insulina*) is a hormone to process glucose, where glycogen (*glicogénio*) is glucose stored in cells.

Bottom results are essentially for pairs which are not usually related, such as *systemic* and *history*. However, there are also cases for which, although the pair seems related, the relation is not explicit in the document. For instance, *rheumatoid arthritis* and *psoriatic arthritis* are two types of arthritis, but they are different types of arthritis, with different causes and different symptoms, therefore, they are not related at a low-level (rheumatoid arthritis affects tissues and organs while psoriatic arthritis affects people who have the chronic skin condition, psoriasis).

6.4 Experimental conditions and results

For evaluating the results of this approach, it was used the same *Medicine* corpora as described in section 4.1. Table 6.4 represents some basic statistics about the corpora, namely the number of documents and words, the average number of words by document and the depth of the subcategories.

Table 6.4: Basic statistics about the corpora based on Wikipedia *Medicine* articles.

Corpus	English	Portuguese	German
Number of documents	4 160	4 066	4 911
Total words	4 657 053	4 153 202	4 337 068
Average #words by document	1 120	1 022	884
Depth of subcategories	2	4	2

10 random documents with a minimum of 2000 words were extracted from each corpus. Then, for each document, the concepts were extracted using the *ConceptExtractor*. Table 6.5 shows the titles of the selected documents.

Table 6.5: Random documents extracted from the English, Portuguese and German Wikipedia.

English	**Portuguese**	**German**
Arthritis	Esclerose tuberosa	Schuppenflechte
Orthotics	Ácido desoxirribonucleico	Homöopathie
Pediatric ependymoma	Transtorno mental	Keratokonus
Effects of benzodiazepines	Cinética enzimática	Nosokomiale Infektion
Mutagen	Sistema imunitário	Tuberkulose
Canine reproduction	Bactéria	Phagentherapie
Schizophrenia	Antidepressivo	Krim-Pfingstrose
Menopause	Terapia genética	Verhaltenstherapie
Glucose meter	Micronutriente	Oberkieferfraktur
Rabbit haemorrhagic disease	Sistema circulatório	Sexualwissenschaft

Finally, for each document, 30 pairs of concepts were extracted and their *Semantic Closeness* computed (as in equation 6.10). It resulted in a list with 300 pairs of concepts, for each language, indexed by document title, which was manually classified as being related or not. The criterion for the classification was that a pair of concepts should only be classified as related if those concepts were explicitly related in their document of origin. This implies that the documents had to be available for reading. As an example of the criterion, Table 6.6 shows the classified results for the English Wikipedia article *Pediatric ependymoma*.

Table 6.6: Classification results for the English article *Pediatric ependymoma*.

Pair		**Pair**	
0.697 gene expression – telomerase	X	0.000 occur – tend	
0.657 mutations – ependymoma	X	0.000 arise – kit	
0.554 tumor suppressor – nf2	X	0.000 favorable – frequently	
0.492 classification – ependymoma	X	0.000 intracranial – correlated	
0.333 tumors – ependymomas	X	0.000 inversely – supratentorial	
0.327 genes – notch		0.000 significantly – remains	
0.312 expression – pediatric ependymomas	X	0.000 loss – down-regulation	
0.226 suppressor genes – mutations	X	0.000 loss – tyrosine	
0.204 pathway – pediatric ependymomas	X	0.000 men1 – inversely	
0.189 tumor suppressor – ependymomas	X	0.000 remains – candidate genes	
0.132 genes – p53	X	0.000 mmp14 – ependymomas	X
0.065 progression – p53		0.000 mmp2 – lethargy	
0.000 location – neurofibromatosis		0.000 mutations – mmp14	
0.000 chromosome – genomic hybridization	X	0.000 outcome – myxopapillary	

Since the extracted lists were sorted by rank, in order to obtain Precision and Recall values, a threshold had to be enforced, such that above the threshold a pair was to be *automatically* considered relevant, and below, non-relevant. That threshold was set empirically on 0.1. Table 6.7 shows the results of this approach.

Table 6.7: Precision and Recall results for the concept cluster's approach.

Language	Precision	Recall
English	0.91	0.83
Portuguese	0.92	0.85
German	0.89	0.79

As it can be seen, the cluster's approach is quite balanced for all tested languages. *Precision* measures how many of the pairs above the threshold are indeed related while *Recall* measures how many of the really related pairs (the ones classified with an 'X') are correctly above the threshold. Both metrics return results as percentages. As expected, recall results are lower than Precision results: given the lack of statistical information in a single document, this approach is not able to correctly identify all possible relations. For instance, in Table 6.7, the pair (*mmp14– ependymomas*) is a good example: *MMP14* is an enzyme related with *ependymomas*; however, since *mmp14* only occurs 2 times in the document, and both occurrences are relatively distant, it never forms a cluster. Rare, scattered concepts, are problematic for this approach. However, for most practical applications, higher precision values are more relevant than higher recall values.

6.5 On collections of documents

As already mentioned, because of the ability to do a local analysis on a document, I believe that this method can aid other methods when dealing with collection of documents. As a brief example, Table 6.8 shows the correlation values (using $Corr(.)$ as in equation 5.9) for some concepts co-occurring with *gout* in the documents of the English corpus.

Table 6.8: Pearson correlation values for concepts co-occurring with *gout* – English corpus.

Concept	*Corr(., gout)*
lawrence c. mchenry	0.544
dr johnson	0.544
hester thrale	0.544
samuel swynfen	0.544
christopher smart	0.544
gouty arthritis	0.352
arthritis	0.257
uric acid	0.198

In this example, the higher correlated concepts are person's names. They come from a document that relates the health of these persons with gout. By being rare in the corpus, these names are extremely valued by correlation metrics. However, especially for applications such as the creation of thesauri, this type of knowledge may have little interest. As an exercise, in Table 6.9 it is shown the same concepts, but including the results of the *Semantic Closeness*, as well as the arithmetic average value between the correlation and the Semantic Closeness.

Table 6.9: Concepts co-occurring with *gout* in the English corpus.

Concept	***Corr(., gout)***	***SC(., gout)***	**Average**
gouty arthritis	0.352	0.67	0.511
uric acid	0.198	0.60	0.399
arthritis	0.257	0.36	0.301
lawrence c. mchenry	0.544	0.00	0.272
dr johnson	0.544	0.00	0.272
hester thrale	0.544	0.00	0.272
samuel swynfen	0.544	0.00	0.272
christopher smart	0.544	0.00	0.272

Gouty arthritis, *uric acid* and *arthritis* are concepts explicitly related with *gout* in some documents of the English corpus. Sorting by the average value allows them to appear in the first positions of the ranking. This type of knowledge may be of interest for specific applications.

This procedure is somewhat similar with the approach presented in section 5.4 for the creation of the *implicit descriptor* of a document, specifically with the combination of the *inter-document proximity* with the *intra-document proximity*. However, $SC(.,.)$ is more severe than the procedure described in section 5.4.2.

6.6 Summary

In this chapter I have presented a method for the extraction of semantic relations from standalone documents. These are documents that, given their specific domains and text size, external ontologies may not exist and standard statistical methods such as the correlation may not work.

This methodology works by identifying clusters in order to measure the Semantic Closeness between pairs of concepts. By measuring the intersection between clusters of different concepts, we are able to measure their semantic relatedness. The results of the method were presented for three different European languages.

I have also shown with a small example, that the local analysis done by this approach may aid statistical methods, such as those based on correlations, when extracting semantic relations from collections of documents.

In general, although precision results are quite encouraging, this procedure is only able to extract semantic relations which are explicit in the texts. This is shown by the lower recall results. Future work should be done to address this issue.

7

Other applications of concepts – opportunities for future research

This chapter presents some other applications for concepts which, essentially by lack of opportunity during the research phase of this thesis, were not extensively researched and therefore did not led to effective publications. However, these applications do suggest possible routes for future research, which is why they are present in this thesis.

This chapter is structured as follows: section 7.1 deals with the segmentation of topics of documents. That section suggests that the usage of concepts may improve results for a baseline topic segmentation algorithm. On section 7.2, it is suggested, through a simple experimentation, that certain areas or topics of documents are more descriptive than others, and those may be identified by means of clusters of concepts. Finally, in section 7.3 it is shown another experiment, with clusters of concepts, to search for the definition of concepts.

7.1 Document topic segmentation – an opportunity for concepts

Topic segmentation of documents is the task of dividing the text of a document into shorter, topically coherent sets of sentences and paragraphs. Dividing a text into different topics is not a simple task, and it is largely dependent on the domain or application. For instance, if we want to segment a book, probably it makes sense to segment it into its different chapters. For a court transcript, probably we might be concerned with the segments in which different arguments or pieces of evidence are being discussed. For an article, probably it makes sense to segment it into its different subsections. However, problems arise essentially when books, transcripts, articles or other texts do not have indications about their section and subsection structure elements. This task has been approached in many ways and I'll briefly review the two major methodologies below.

7.1.1 Current work

Some methodologies are based on the insight that people talk about different topics in different ways, i.e, by using different words to refer to different things. For instance, if we are discussing a particular subject, we use a particular set of words relevant to that subject. The shift to a different subject implies the use of a different set of words. Therefore, a change in topic is associated with a change in the vocabulary. *TextTilling* [Hea97] is considered a baseline method for this type of methodologies and it is reviewed in the following subsection.

The second insight is that the boundaries between topics tend to have their own characteristic features, independent of the subject matter. When switching from one topic to another, signals tend to be made to the audience in various ways. For instance, there are various cue words and phrases (*discourse makers*) that provides cues about the discourse structure, and words like *okay*, *anyway*, *so* or *now* can signal the end of one topic and the beginning of another. In certain domains, there can be specific cues, such as the mention of "the next item on the agenda is" in formal meeting transcripts. Outside the domain of written texts, small pauses on speeches may be indicative of topic shifts as well as non-linguistic features such as changes in the physical posture of the speaker or of the audience. However, this particular line is outside the scope of this chapter.

7.1.2 TextTilling approach

The TextTilling algorithm [Hea97] is considered a baseline method for the topic segmentation of documents. It has three main parts: (1) Tokenization; (2) Lexical Score Determination and (3) Boundary Identification.

The tokenization refers to the division of the text into individual lexical units. All markup elements are ignored and all words in the text are converted to lowercase characters. Individual words are compared against a stop-word list and only valid words are used on the lexical score phase. Valid words are then reduced to their root by a morphological analysis function, converting regularly and irregularly inflected nouns and verbs to their roots. Finally, the text is subdivided into pseudo-sentences of size w to allow for comparison of equal-sized units.

The lexical score phase consists in the determination of the measure of similarity between adjacent blocks of text, in this case, of pseudo-sentences. For each interval (or gap) i between two consecutive pairs (b_1, b_2) of pseudo-sentences, its score is measured using equation 7.1.

$$score(i) = \frac{\sum_t w_{t,b_1} \cdot w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \cdot \sum_t w_{t,b_2}^2}} \,. \tag{7.1}$$

Variable t ranges over all terms registered during the tokenization phase (unigrams excluding stop-words) and $w_{t,b}$ is the weight assigned to term t in block b, which is essentially the frequency of occurrence. Equation 7.1, measures the similarity between consecutive pseudo-sentences for each gap i.

A different score formula is given in the paper which considers the number of new terms that appear in the pseudo-sentences. However, the author does not consider the results using that metric.

The last step consists in the identification of topic boundaries. Essentially, boundaries are identified when major gaps occur between pairs of adjacent pseudo-sentences. Steep gaps indicate large dissimilar topics. Hearst suggests in her paper to use a low-pass filter to smooth the plot in order to remove small irrelevant changes in the vocabulary, and suggests also the use of a cutoff as function of the average and standard deviation of the scores. Figure 7.1 shows the similarity plot and the topic boundaries suggested by the TextTilling algorithm for the English Wikipedia *Arthritis* document.

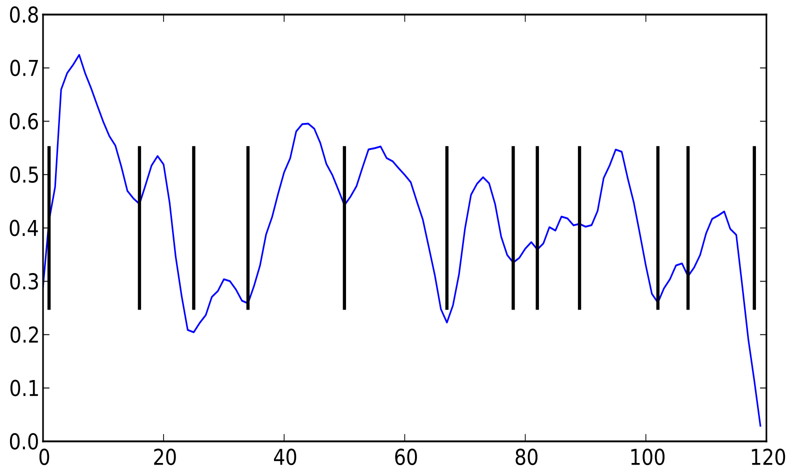


Figure 7.1: TextTilling similarity plot and suggested topic boundaries for the English *Arthritis* document.

The blue line corresponds to the similarity score for the gap at the x-axis pseudo-sentence, while the black vertical lines correspond to the suggested topic boundaries.

7.1.3 TextTilling with concepts

The tokenization phase of the TextTilling algorithm uses stop-word lists to validate words and also includes a procedure to reduce nouns and verbs to their morphological roots. This implies the usage of predefined lists and tools which may not be available for many languages. Also, this phase does not include multi-words.

I've made some experiments comparing the use of concepts with the original TextTilling approach. Although further research should be done to confirm if the use of concepts with TextTilling yields better results than not using concepts, preliminary results seem to indicate that there are some benefits. First, the truly language-independent source of the *ConceptExtractor* allows to implement the TextTilling algorithm independently of the text language. Also, using concepts allows the TextTilling approach to focus on the truly relevant terms instead of all single-words. This allows for a better separation of topics by their concepts. For instance, Figure 7.2 shows the comparison of TextTilling similarities with concepts versus without concepts for the English Wikipedia *Hormone* document.

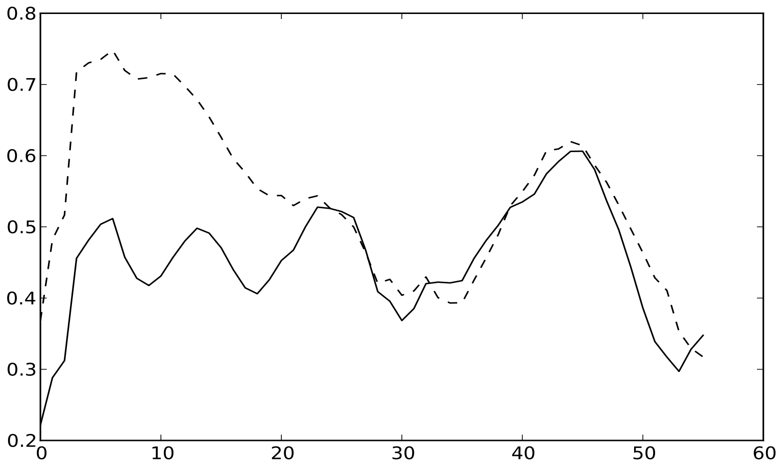


Figure 7.2: Comparison of TextTilling similarities with concepts (solid line) versus the original (dashed line) for the English Wikipedia *Hormone* document.

A manual analysis on the English Wikipedia *Hormone* article indicates that the document is divided on the following sections starting at the sentences in parenthesis: *Introduction* (1), *Hormones as signals* (9), *Interactions with receptors* (17), *Physiology of hormones* (31), *Effects of hormones* (45) and *Chemical classes* (48). The comparison of both approaches in the figure indicates that the concept-based approach (solid line), clearly indicates the shift in topics, specially for the topics starting at sentences 9 and 17. The original approach (dashed line – without concepts) does not indicate clearly the shift on those initial topics. For the original approach, the topic at sentence 17 is non-existent and there is a false positive at sentences 34/35. Finally, both methods fail to identify the last topics, although the concept-based approach (solid line) has slight indications on sentence 43 and sentence 53. Another example of comparison between both approaches can be found in Figure 7.3, for the English Wikipedia *Amygdalin* document.

The *Amygdalin* Wikipedia article has the following sections: *Introduction* (1), *Chemistry* (7), *Laetrile* (17), *Toxicity* (28), *Cancer Treatment* (41), *Initial studies at Sloan-Kettering* (47), *Subsequent clinical studies* (58) and *Advocacy and legality* (67). The comparison of both approaches in Figure 7.3 indicates that the concept-based approach (the solid line) is capable of identifying some topics for which the original approach is not capable, such as the topic at sentence 43 and the topic starting at sentence 67.

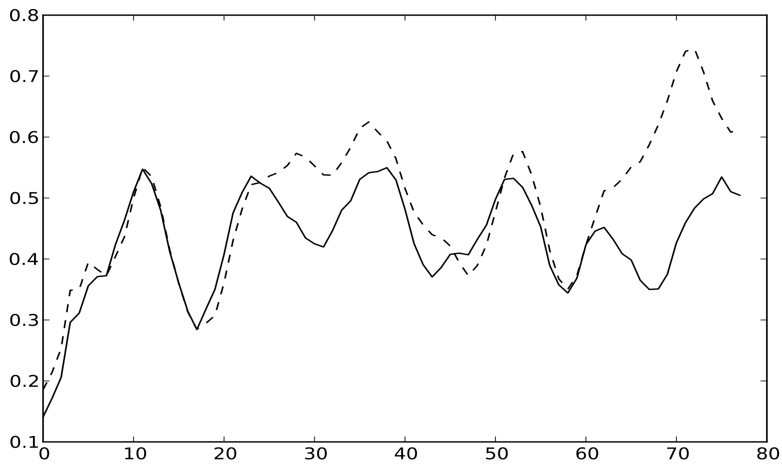


Figure 7.3: Comparison of TextTilling similarities with concepts (solid line) and the original (dashed line) for the English Wikipedia *Amygdalin* document.

These preliminary results are encouraging and indicate that baseline algorithms, such as TextTilling, could benefit from the use of concepts in order to improve their performance.

7.2 Finding the most descriptive areas of documents – applicability for clusters of concepts

This section of the thesis presents an experiment made with clusters of concepts with the purpose of identifying the areas or topics of documents which are more descriptive.

The underlying idea is that some documents, specially encyclopedic documents such as Wikipedia articles, usually start with a brief introduction of the subject being discussed, and are then divided in sections and subsections which present additional information about the main subject. Algorithms such as TextTilling are very efficient in identifying the boundaries of topics. However, not all sections and subsections are equally important in the context of the document. For instance, in the Wikipedia *Abortion* article, one could consider that the topic about the different abortion methods is more relevant to the topic than the history of abortion.

The definition of "importance" regarding the different topics and subtopics in a document is certainly related with the expectations of possible readers. In the *Abortion* article example, some readers could actually be looking for the history of abortion, but, since it is an encyclopedic document, it is safe to say that the majority of users are looking for those specific topics which allows them to increase their knowledge about the main subject in a more generic way.

7.2.1 Clusters of concepts as indicators of topic importance

Since all tested corpora are made of encyclopedic articles from Wikipedia, the experiments in this section were made considering that the importance of a topic or section in an article is related to the semantic richness and descriptive ability of the topic. In practice, the importance of a topic is related with the amount of concepts being used in its text. The rationale for this condition is that since readers are looking for knowledge, a high use of concepts in a topic denotes that a complex discussion on a particular subject is taking place on that same topic. Figure 7.4 shows the number of active concept clusters for each sentence on the English Wikipedia *Encephalitis* article (Y-Y axis is normalized to the maximum number of concept clusters found in a sentence).

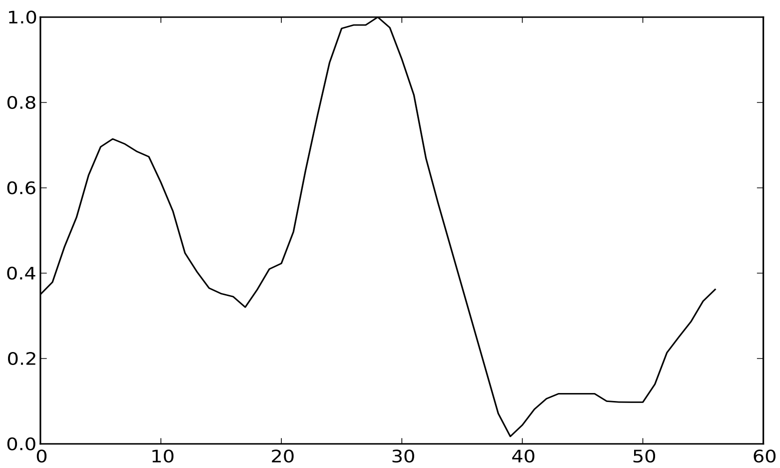


Figure 7.4: Histogram of concept clusters in sentences for the English Wikipedia *Encephalitis* document.

The *Encephalitis* article has the following topics (starting sentence index between paragraphs): *Introduction* (0), *Viral cause* (4), *Bacterial cause* (9), *Diagnosis* (16), *Treatment* (30), *Prevention* (41), *Encephalitis lethargica* (45), *Limbic system encephalitis* (51) and *Epidemology* (53).

Figure 7.4 shows two areas where concepts are being densely used: the first starting at sentence 17, reaching its peak at sentence 28, and ending at sentence 39; and the second one starting at sentence 1, reaching its peak at sentence 6, and ending at sentence 17. The first and largest graph curve includes the *Diagnosis* and the *Treatment* topics, reaching the peak in the end of the *Diagnosis* topic. The second graph curve starts with the *Introduction* and includes also the *Viral cause* and *Bacterial cause* topics, reaching the peak at the *Viral cause* topics.

A visual confirmation clearly indicates that the *Diagnosis* and the *Treatment* topics are the most complex and descriptive subtopics in the *Encephalitis* document, including concepts such as *herpes simplex virus*, *varicella zoster virus*, among others. The topics *Introduction*, *Viral cause* and *Bacterial cause*, also includes a complex discussion of certain concepts. Finally, the remaining topics are more direct and less descriptive.

7.2.2 Defining the boundaries and ranking by complexity

Be $hist = [h_1, h_2, \ldots, h_n]$ a vector for which h_i is the number of active clusters of concepts in sentence i. Be $cuts = [c_1, c_2, \ldots, c_{m+1}]$ and $peaks = [p_1, p_2, \ldots, p_m]$ two vectors where (c_k,c_{k+1}) are the positions of the lowest $hist$ values before and after the peak value in $hist[p_k]$. Vector $cuts$ contains the positions of the boundaries between peaks and is obtained using the *identification of topic boundaries* procedure as described in [Hea97].

Be $height$ a vector which has the height of a peak p_k relatively to its left and right immediate valleys. For each peak p_k, $height[p_k]$ is calculated as follows:

$$height[p_k] = 0.75\,.\,(hist[p_k] - max(hist[c_k], hist[c_{k+1}])) \tag{7.2}$$

Basically, $height[p_k]$ measures 75% of the shortest height from the peak in p_k to one of its immediate left and right valleys.

Now, be (a_k,b_k) a pair of indexes in $hist$ such that $c_k < a_k < p_k < b_k < c_{k+1}$ and $hist[a_k] = hist[b_k] = hist[p_k] - height[p_k]$. a_k and b_k are indexes near p_k for which their $hist$ value is 75% distant from the peak value (in relation to $height[p_k]$). Finally, a triplet ($height[p_k]$, a_k, b_k) contains the height of a peak p_k and the starting and ending positions a_k and b_k.

In a practical way, we want to find the a_k and b_k sentences for which the number of active clusters is some percentage of the total number of active clusters in the peak sentence. That percentage depends on the height of the peak and the depth of the valleys. Finally, a_k is a sentence before the peak while b_k is a sentence after the occurrence of the peak. Sorting peaks by their heights allows us to rank the sentences between a_k and b_k by decreasing order of importance.

Figure 7.5 shows the concept cluster histogram for English *Abortion* article (Y-Y axis is normalized to the maximum number of concept clusters found in a sentence).

The *Abortion* article is a quite long article and it includes the following sections: *Introduction* (1), *Induced abortion* (13), *Spontaneous abortion* (24), *Medical abortion* (36), *Surgical abortion* (47), *Other methods* (64), *Unsafe abortion* (92), *Incidence and motivation* (112), *Gestational age and methods* (120), *Motivation* (129), *History* (139), *Abortion debate* (154), *Modern abortion law* (164), *Sex-selective abortion* (184), *Anti-abortion violence* (194), *Art, literature and film* (204) and *Abortion in other animals* (226). Table 7.1 presents the $height$ results and boundaries for the cluster histogram approach for the same document.

Medical abortion and *surgical abortion* are considered by this approach as the most complex and descriptive discussed topics. In fact, by looking at the document's text, it can

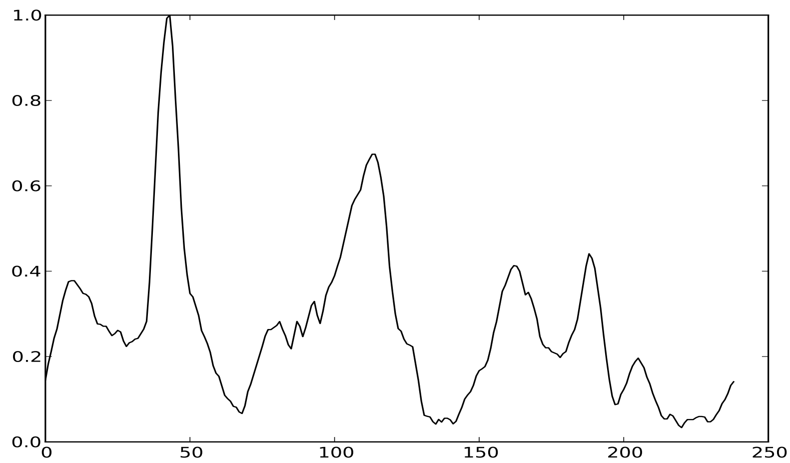


Figure 7.5: Histogram of concept clusters in sentences for the English Wikipedia *Abortion* document.

Table 7.1: Ranking results for the *Abortion* article - English Wikipedia.

height[p_k]	a_k	b_k	**Topics**
0.582	37	48	*Medical abortion*, *Surgical abortion*
0.297	100	119	*Unsafe abortion*, *Incidence and motivation*
0.182	183	192	*Sex-selective abortion*
0.155	156	170	*Abortion debate*
0.088	5	17	*Introduction*
0.081	200	209	*Anti-abortion violence*, *Art, literature and films*
0.040	76	83	*Safety*
0.009	87	87	*Safety*
0.004	216	217	*Art, literature and films*
0.002	136	139	*Motivation*

be seen that both topics are the most descriptive and rich in the document, by using concepts such as names of *abortifacient pharmaceuticals* and techniques such as *vacuum aspiration*. The following topics are *Unsafe abortion* and *Incidence and motivation*. The later is not particularly interesting, but the former includes a large discussion about the safety of this procedure. The third topic, *Sex-selective abortion*, is also a highly complex and descriptive topic. The rest of the topics are increasingly less descriptive.

7.2.3 Possible applications

As already mentioned, this approach reflects an experiment, and only preliminary results were obtained. However, this kind of approach may be applicable, for instance, to guide a reader on selecting a different order of topics to read in a document, instead of the traditional top-down linear method. Another possibility is to guide an automatic approach for the extraction of document's summaries.

7.3 Extraction of concept definitions – a quest for clusters

The definition of a concept is a textual description of a term that states its meaning, or describes the concept. The extraction of definitions from texts can be useful in several scenarios, such as the automatic creation of glossaries for building dictionaries or in question answering systems.

This section presents an experiment that uses clusters of concepts to find the best definitions of concepts. The rationale is that on encyclopedic texts, such as the ones used throughout this thesis, a concept is defined by using other concepts. Take, for instance, the following excerpt of a paragraph from the English Wikipedia *Arthritis* document:

> **Gout** is caused by deposition of **uric acid crystals** in the **joint**, causing **inflammation**. The **joints** in **gout** can often become **swollen** and lose **function**. **Gouty arthritis** can become particularly **painful** and potentially **debilitating** when **gout** cannot successfully be **treated**. When **uric acid levels** and **gout symptoms** cannot be controlled with **standard gout medicines** that decrease the production of **uric acid** (e.g., **allopurinol**, **febuxostat**) or increase **uric acid** elimination from the **body** through the **kidneys** (e.g., **probenecid**), this can be referred to as **refractory chronic gout** or **RCG**.

The previous quotation describes the causes of the medical condition *gout*, and the description is made with the use of other related concepts such as *uric acid crystals*, *joint*, *inflammation*, *swollen* and *gouty arthritis*, just to name a few.

The experiment in this chapter uses clusters of concepts. The idea is quite similar to the one presented in section 7.2.1: complex and highly descriptive areas of text, where a concept occurs with other concepts, can be quite descriptive of that concept. In other words, the description of a concept tends to be highly related with the occurrence of its clusters together with the occurrence of clusters of other concepts.

Other authors have done work in this area, therefore the next subsection reviews some of their work.

7.3.1 Current work

In the paper [GB07], the authors present a rule-based approach for the extraction of definitions in Portuguese. The input for their system is a Part-of-Speech annotated text with inflection features. Their idea is that the definitions of concepts in Portuguese texts follow specific patterns. Some of these patterns are the use of instances of the verb *to be* (Ex: "FTP é um protocolo de rede."), the use of punctuation clues (Ex: "TCP/IP: protocolos utilizados na troca de informações entre computadores.") or other linguistic expressions and patterns. They have compiled a grammar to parse their texts, and results were obtained. The low value for the precision indicates that their procedure has much room for improvement. Despite that fact, the use of POS taggers to explore syntactic patterns makes this approach largely language-dependent. [TBR10] and [PDSSOLKW07] are similar approaches, respectively for Arab and Slavic languages.

The paper in [BRP09] presents a different method for the extraction of definitions. The authors propose a machine learning approach, in particular, an evolutionary algorithm (genetic algorithm), to learn the best linguistic rules to extract definitions using a pattern-based approach. Their results seem to confirm that the genetic algorithm is capable of recreating most of the manually crafted rules. Similar to the previous approaches, this one also uses Part-of-Speech taggers and other linguistic tools, which makes it largely language-dependent.

7.3.2 Using clusters of concepts to find concept descriptions

The idea of the approach presented in this section is that the areas of text where a concept co-occurs densely with other concepts may be quite descriptive. A cluster of a concept indicates a dense area of occurrence. Therefore, clusters of concepts are the starting point of the proposed approach.

Be $C_{t_j,i,d}$ the i-th cluster of term t_j in document d. The score of cluster $C_{t_j,i,d}$, as the area of document d where the definition of the term t_j can be found, is measured using equation 7.3:

$$score(C_{t_j,i,d}) = size(C_{t_j,i,d}) \,.\, cohesion(C_{t_j,i,d}) \,.\, \sum_{l=0}^{n} intersection(C_{t_j,i,d}, C_{t_k,l,d}) \qquad \forall t_k \in concepts(d) \tag{7.3}$$

In equation 7.3, $size(C_{t_j,i,d})$ is the size of cluster $C_{t_j,i,d}$, $cohesion(C_{t_j,i,d})$ is the cohesion of cluster $C_{t_j,i,d}$, and $intersection(C_{t_j,i,d}, C_{t_k,l,d})$ measures the intersection between clusters $C_{t_j,i,d}$ and $C_{t_k,l,d}$. Finally, $concepts(d)$ is the list of all concepts occurring in document d. The definitions of these elements can be found in subsections 6.3.1 and 6.3.2.

Basically, the score of cluster $C_{t_j,i,d}$ as being the definition of term t_j, depends on the size of the cluster, the internal cohesion of the occurrences of t_j in the cluster, and on the occurrence of other clusters of concepts t_k which intersect cluster $C_{t_j,i,d}$. In other words, the text where the cluster $C_{t_j,i,d}$ occurs is more relevant as being a definition of concept t_j if it is large, highly cohesive, and other concepts co-occur densely in that same area. This is consistent with the fact that areas of text where many concepts occur can be considered highly descriptive.

Tables 7.2, 7.3 and 7.4, show the results of this approach for some concepts, namely *uric acid* and *arthritis* on the English corpus, and *amnésia* on the Portuguese corpus. The corpora used for the experiments are the same Medicine corpora as described in Table 6.4.

These tables show that the metric proposed in equation 7.3 is quite capable of ranking

Table 7.2: Definition results for *uric acid* – English Wikipedia corpus.

Score(.)	**Doc. Title**	**Text Excerpt**
1088.67	Antioxidant	Uric acid. Uric acid is by-far the highest concentration antioxidant in human blood. Uric acid (UA) is an antioxidant oxypurine produced from xanthine by the enzyme xanthine oxidase, and is an intermediate product of purine metabolism.
999.04	Glycogen storage disease type I	and uric acid compete for the same renal tubular transport mechanism. Increased purine catabolism is an additional contributing factor.
739.11	Arthritis	of uric acid crystals in the joint, causing inflammation. There is also an uncommon form of gouty arthritis caused by the formation of rhomboid crystals of calcium pyrophosphate known as pseudogout.
...	...	...

Table 7.3: Definition results for *arthritis* – English Wikipedia corpus.

Score(.)	**Doc. Title**	**Text Excerpt**
611.88	Childhood arthritis	Childhood arthritis (JA) also known as juvenile arthritis is any form of arthritis or arthritis related conditions which affects individuals under the age of 16. Juvenile arthritis is a chronic, autoimmune disease.
509.32	Arthritis	Arthritis (from greek "arthro-", joint + "-itis") is a form of joint disorder that involves inflammation of one or more joints. There are over 100 different forms of arthritis. The most common form, osteoarthritis (degenerative joint disease), is a result of trauma to the joint.
467.64	Arthritis	Rheumatoid arthritis is a disorder in which the body's own immune system starts to attack body tissues. The attack is not only directed at the joint but to many other parts of the body. In rheumatoid arthritis, most damage occurs to the joint lining and cartilage.
...	...	...

Table 7.4: Definition results for *amnésia* – Portuguese Wikipedia corpus.

Score(.)	**Doc. Title**	**Text Excerpt**
523.85	Memória	Amnésia. Amnésia é a perda parcial ou total da capacidade de reter e evocar informações. Qualquer processo que prejudique a formação de uma memória a curto prazo ou a sua fixação em memória a longo prazo pode resultar em amnésia.
414.99	Síndrome de Wernicke-Korsakoff	pela amnésia anterógrada , amnésia retrógrada e muito comumente a confabulação e uma desorientação temporoespacial. Acompanham esses sintomas uma severa apatia e desinteresse por parte do doente, que muitas vezes não é capaz de ter consciência de sua condição.
323.35	Amnésia	Amnésia anterógrada, é a perda de memória para eventos que ocorrem posteriormente ao acometimento da doença, ou seja, é a deficiência em formar novas memórias, como ocorre na doença de alzheimer. Amnésia Retrógrada, nesta outra forma de amnésia ocorre o inverso da amnésia anterógrada.
...	...	...

the text excerpts by complexity of description. For instance, the first result of *uric acid* is in fact the description of uric acid. Since there is no *Uric acid* document, the best description is found on the *Antioxidant* article, and, therefore, *uric acid* is defined as being an antioxidant. The following results are related with other contexts, and are less descriptive of the concept.

As for *arthritis* (Table 7.3), the first result is the definition of *juvenile arthritis* and the second results is the generic definition of *arthritis*. This occurs mainly because the description on the *Arthritis* document uses specific concepts such as *osteoarthritis*, which reduces the cohesion of the *arthritis* cluster.

As for the Portuguese concept *amnésia* (Table 7.4), the first result is the generic definition of the concept while the following results describe specific cases of amnesia.

On a final note, one could be tempted to consider more relevant, as definition of a concept t_j, those text areas (or clusters) for which concept t_j would occur more frequently. This idea is somewhat similar to the one behind *Tf-Idf*, which relates relevancy to the frequency of occurrence. However, see Table 7.5, which shows the definition results for concept *gout* including the frequency of occurrence of *gout* in the cluster originating the definition.

Although the cluster in the second result of Table 7.5 has 10 occurrences of *gout*, the *best* definition (which is in the first row of the table) has only 7 occurrences. However, the first definition includes a myriad of other concepts, such as *uric acid*, *joints*, etc., while the second definition (which is not a definition at all), has quite less concepts, since it is

Table 7.5: Definition results for *gout* – English Wikipedia corpus.

Score(.)	**#"gout"**	**Doc. Title**	**Text Excerpt**
1088.03	7	Arthritis	Gout. Gout is caused by deposition of uric acid crystals in the joint, causing inflammation. There is also an uncommon form of gouty arthritis caused by the formation of rhomboid crystals of calcium pyrophosphate known as pseudogout.
458.44	10	Samuel Johnson's health	Gout. Johnson suffered from what he and his doctors labeled as gout starting in 1775 when he was 65, and again in 1776, 1779, 1781, and 1783. He told William Boswles, in 1783, that "the gout has treated me with more severity than any former time".
135.66	3	Health effects of coffee	Gout. Coffee consumption contributes to a decreased risk of gout in men over age 40.
...	...	...	...

only a description of someone affected by the disease. Therefore, using the factor which considers the intersection to other clusters allows to compensate for these cases.

The results in this section are quite encouraging, but further research should be done in order to obtain concrete Precision and Recall values.

7.4 Summary

In this chapter I have presented three possible applications for automatic extracted concepts. For TextTilling, an algorithm to automatically detect topic boundaries, I have shown that the use of concepts may improve its performance, mainly because concepts allow the algorithm to enhance more clearly the boundaries of topics.

A methodology for finding the most descriptive areas of documents was also presented. It uses the number of concept clusters by sentence to indicate the level of complexity of a text area. Text areas where many concepts occur may be considered highly descriptive. This approach may be of interest to knowledge discovery applications.

Finally, it was presented an approach for the automatic extraction of concept definitions. The idea is that a concept is usually defined by using other related concepts, and clusters of concepts are used. A score metric was proposed, and the results are interesting. This approach may be of interest for applications such as question answering systems.

The experiments have shown that the results of these applications are encouraging, and future work could be done in any of them.

8

Conclusions

The extraction of relevant terms from texts is an extensively researched task in Text-Mining. However, it is not easy to classify many terms as *relevant* or *not relevant* because usually there is no consensus about the semantic value or the informativeness of some less clear terms. Concepts, on the other hand, have a less fuzzy nature. Instead of deciding on the relevance of a term during the extraction phase, which most extractors do, I proposed to extract what I have called *generic concepts* from texts and postpone the decision about relevance for downstream applications, accordingly to their needs.

Furthermore, current methodologies for the extraction of concepts from documents have shortcomings. In a general way, non-statistical methods tend to explore lexical patterns, use external lexicons (such as *WordNet*), or use Part-of-Speech taggers and other tools, which makes them highly language-dependent. On the other hand, most statistical approaches are language independent, but they can not cope with single-words and multi-words using the same approach. Moreover, statistical methods for the extraction of relevant single-words tend to harm frequent or large single-word concepts. As for the statistical methods for the extraction of relevant multi-words, they either extract only 2-grams, or, as LocalMaxs [SL99], are capable of extracting n-grams larger than 2 but present modest Precision and Recall values.

In Part I of this thesis, I've proposed *ConceptExtractor*, a statistical and language-independent approach for the extraction of single-word and multi-word concepts from texts. *ConceptExtractor* is able to identify both single-word and multi-word concepts, independently of the frequency of occurrence, in different languages, without privileging any language in specific. It presents Precision and Recall values around 90%.

In Part II of this thesis, I've presented some applications for the automatic extracted concepts. In chapter 5, I proposed a language-independent method for the automatic building of document descriptors formed by explicit and implicit keywords. Explicit keywords correspond to the most *Tf-Idf*-scored concepts and I've shown that *Tf-Idf* returns significantly better results when applied to concepts, specially for multi-words. I have also proposed metrics to identify semantic relations between terms in order to measure the relevance of a concept as implicit keyword of a document. Implicit keywords may offer an extended semantic scope to the global descriptors of documents, with great applicability, for example in Information Retrieval or in search engine contexts. In other words, the access to documents is no longer limited by the information they contain explicitly, but also by the information given through the implicit concepts. Implicit concepts, although not explicit in the documents, are related to its content. Results lead us to conclude that these automatically extracted keywords show the core content of the documents and form efficient document descriptors.

In chapter 6, I have presented a method for the extraction of semantic relations from standalone documents. Standalone documents are, essentially, isolated or single documents, such as those containing unique subjects or domains, reports from very specific fields of expertise or even small books. This methodology works by identifying clusters of concepts as being specific areas in a text where a concept is relevant and tends to occur rather densely. Clusters allow to measure the Semantic Closeness between pairs of concepts, considering the *intersection* of the corresponding clusters and their internal cohesion. Results of this method were presented for three different European languages and showed consistency and credible values for Semantic Closeness between pairs of concepts. Precision and Recall values are quite encouraging.

Chapter 7 presented some applications for concepts which were not extensively researched and did not led to publications. I have shown in this chapter that the use of concepts may improve the performance of topic segmentation algorithms, such as TextTilling. Also, an application for finding the most descriptive areas of documents was also presented. Descriptive text areas are areas of documents where concepts occur rather densely and concept clusters are used to identify the denser areas. This approach may be of interest to knowledge discovery applications. An approach for the extraction of concept definitions was also presented in this chapter. From the results, it is possible to conclude that the definition of concepts tend to be in areas of the text of higher density of concept clusters. This approach may be of interest for applications such as question answering systems.

The results of the applications presented in chapter 7 are quite encouraging, and deserve further research. Future work could be done in those applications.

Finally, *ConceptExtractor* is not without its drawbacks. Most of these drawbacks arise from the fact that some multi-word concepts score high in their specificity, although we can not say that they are complete. The inclusion of a new rule such as "*multi-word concepts must start and end with **complete concepts***" could help to define the solution, although the programmatic or statistical solution is not easy to define. Algorithms such as Localmaxs could be of help for those highly specific situations, but not as complete replacements. The identification of singular-plural concepts (such as *abortion* and *abortions*) and of synonyms, by the extractor, would also be desirable.

Regarding Precision and Recall values, although the results of the *ConceptExtractor* are quite encouraging, future work should be done to increase the performance of the extractor.

9

Bibliography

[AB09] A. Akbik and J. Broß. "Wanderlust: Extracting Semantic Relations from Natural Language Text Using Dependency Grammar Patterns". In: *Proceedings of the 18th International World Wide Web Conference*. Madrid, Spain, 2009. URL: http://citeseerx.ist.psu.edu/viewdoc/summary;?doi=10.1.1.204.5708.

[BFL98] C. F. Baker, C. J. Fillmore, and J. B. Lowe. "The Berkeley FrameNet Project". In: *Proceedings of the 17th international conference on Computational linguistics*. Vol. 1. COLING '98. Montreal, Quebec, Canada: Association for Computational Linguistics, 1998, pp. 86–90. DOI: 10.3115/980451.980860. URL: http://dl.acm.org/citation.cfm?id=980860.

[Bie05] C. Biemann. "Ontology Learning from Text: A Survey of Methods". In: 20.2 (2005), pp. 75–93. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.90.4046.

[BRP09] C. Borg, M. Rosner, and G. Pace. "Evolutionary algorithms for definition extraction". In: *Proceedings of the 1st Workshop on Definition Extraction*. WDE '09. Borovets, Bulgaria: Association for Computational Linguistics, 2009, pp. 26–32. ISBN: 978-954-452-013-7. URL: http://dl.acm.org/citation.cfm?id=1859765.1859770.

[Bra06] R. Bradford. "Relationship Discovery in Large Text Collections Using Latent Semantic Indexing". In: *Proceedings of the Fourth Workshop on Link Analysis, Counterterrorism, and Security, SIAM Data Mining Conference*. Bethesda, MD, U.S.A., 2006.

[Bri99] S. Brin. "Extracting Patterns and Relations from the World Wide Web". In: *Selected papers from the International Workshop on The World Wide Web and Databases*. WebDB '98. London, UK, UK: Springer-Verlag, 1999, pp. 172–183. ISBN: 3-540-65890-4. URL: http://dl.acm.org/citation.cfm?id=646543.696220.

[CG91] K. W. Church and W. A. Gale. "Concordance for parallel texts". In: *Proceedings of the 7th Annual Conference of the UW Centre of the new OED and Text Research, Using Corpora*. Oxford, UK, 1991.

[CH90] K. W. Church and P. Hanks. "Word association norms, mutual information, and lexicography". In: *Computational Linguistics* 16.1 (1990), pp. 22–29. ISSN: 0891-2017. URL: http://dl.acm.org/citation.cfm?id=89086.89095.

[CPGV05] J. M. Cigarrán, A. Peñas, J. Gonzalo, and F. Verdejo. "Automatic Selection of Noun Phrases as Document Descriptors in an FCA-Based Information Retrieval System". In: *Formal Concept Analysis - Third International Conference, ICFCA 2005*. Vol. 3404. Lens, France: Springer-Verlag, 2005, pp. 49–63. ISBN: 978-3-540-32262-7. URL: http://link.springer.com/chapter/10.1007%2F978-3-540-32262-7_4.

[DMPPG02] A. Das, M. Marko, A. Probst, M. A. Porter, and C. Gershenson. "Neural Net Model for Featured Word Extraction". In: *CoRR* cs.NE/0206001 (2002). URL: http://arxiv.org/abs/cs.NE/0206001.

[Dia03] G. Dias. "Multiword unit hybrid extraction". In: *Workshop on Multiword Expressions of the 41st ACL meeting*. Sapporo, Japan, 2003, pp. 41–48. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.7510.

[EC07] G. Ercan and I. Cicekli. "Using Lexical Chains for Keyword Extraction". In: *Information Processing and Management: an International Journal archive*. Vol. 6. 2007, pp. 1705–1714.

[FRF06] R. Feldman, B. Rosenfeld, and M. Fresko. "TEG – A hybrid approach to information extraction". In: *Knowledge and Information Systems* 9.1 (Jan. 2006), pp. 1–18. ISSN: 0219-1377. DOI: 10.1007/s10115-005-0204-y. URL: http://dx.doi.org/10.1007/s10115-005-0204-y.

[GB07] R. D. Gaudio and A. Branco. "Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach". In: *Proceedings of the 13th Portuguese Conference on Artificial Intelligence, EPIA 2007*. Vol. 4874. Lecture Notes in Computer Science. Guimarães, Portugal: Springer Berlin Heidelberg, 2007, pp. 659–670. ISBN: 978-3-540-77000-8. DOI: 10.1007/978-3-540-77002-2_55. URL: http://dx.doi.org/10.1007/978-3-540-77002-2_55.

[GWP98] B. Gelfand, M. Wulfekuhler, and W. F. Punch. “Automated Concept Extraction From Plain Text”. In: *Papers from the AAAI 1998 Workshop on Text Categorization*. 1998, pp. 13–17.

[GMM03] A. Gómez-Pérez and D. Manzano-Macho. *A survey of ontology learning methods and techniques*. Deliverable 1.5. OntoWeb Consortium, 2003. URL: http://www.deri.at/fileadmin/documents/deliverables/Ontoweb/D1.5.pdf.

[Gre93] G. Grefenstette. “Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches”. In: *Corpus processing for lexical acquisition*. Cambridge, MA, USA: MIT Press, 1993, pp. 205–216. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.1354.

[Hea97] M. A. Hearst. “TextTiling: segmenting text into multi-paragraph subtopic passages”. In: *Computational Linguistics* 23.1 (Mar. 1997), pp. 33–64. ISSN: 0891-2017. URL: http://dl.acm.org/citation.cfm?id=972684.972687.

[Hei99] U. Heid. *A linguistic bootstrapping approach to the extraction of term candidates from German text*. 1999. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.5364.

[HTC06] M.-H. Hsu, M.-F. Tsai, and H.-H. Chen. “Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison”. In: *Information Retrieval Technology*. Springer, 2006, pp. 1–13. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.5585.

[Jon72] K. S. Jones. “A statistical interpretation of term specificity and its application in retrieval”. In: *Journal of Documentation* 28 (1972), pp. 11–21.

[LD97] T. K. Landauer and S. T. Dutnais. “A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge”. In: *Psychological review* (1997), pp. 211–240.

[Luh58] H. P. Luhn. “The automatic creation of literature abstracts”. In: *IBM J. Res. Dev.* 2.2 (Apr. 1958), pp. 159–165. ISSN: 0018-8646. DOI: 10.1147/rd.22.0159. URL: http://dx.doi.org/10.1147/rd.22.0159.

[MC07] R. Mihalcea and A. Csomai. “Wikify!: linking documents to encyclopedic knowledge”. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. CIKM ’07. Lisbon, Portugal: ACM, 2007, pp. 233–242. ISBN: 978-1-59593-803-9. DOI: 10.1145/1321440.1321475. URL: http://doi.acm.org/10.1145/1321440.1321475.

[Mil95] G. A. Miller. “WordNet: A Lexical Database for English”. In: *Communications of the ACM* 38.11 (1995), pp. 39–41. URL: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.83.1823.

[MN03] B. Mohit and S. Narayanan. “Semantic extraction with wide-coverage lexical resources”. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003*. Vol. 2. NAACL-Short ’03. Edmonton, Canada: Association for Computational Linguistics, 2003, pp. 64–66. URL: http://dl.acm.org/citation.cfm?id=1073505.

[NHN08] K. Nakayama, T. Hara, and S. Nishio. “Wikipedia Link Structure and Text Mining for Semantic Relation Extraction”. In: *SemSearch 2008: CEUR Workshop Proceedings*. 2008. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.143.1537.

[PARR12] A. Panchenko, S. Adeykin, P. Romanov, and A. Romanov. “Extraction of Semantic Relations between Concepts with KNN Algorithms on Wikipedia”. In: (2012), pp. 78–88. URL: http://cental.fltr.ucl.ac.be/team/~panchenko/cdud-camera-ready.pdf.

[PRGM10] A. Parameswaran, A. Rajaraman, and H. Garcia-Molina. “Towards the web of concepts: extracting concepts from large datasets”. In: *Proc. VLDB Endow.* 3.1-2 (Sept. 2010), pp. 566–577. ISSN: 2150-8097. URL: http://dl.acm.org/citation.cfm?id=1920841.1920914.

[PP06] S. Patwardhan and T. Pedersen. “Using WordNet-based context vectors to estimate the semantic relatedness of concepts”. In: *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*. Trento, Italy, 2006, pp. 1–8. URL: http://www.patwardhans.net/papers/PatwardhanP06.pdf.

[Pea00] K. Pearson. “On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that can be reasonably supposed to have arisen from Random Sampling”. In: *Philosophical Magazine* 50 (1900), pp. 157–175. URL: http://www.economics.soton.ac.uk/staff/aldrich/1900.pdf.

[PDSSOLKW07] A. Przepiórkowski, L. Degórski, M. Spousta, K. Simov, P. Osenova, L. Lemnitzer, V. Kuboň, and B. Wójtowicz. “Towards the automatic extraction of definitions in Slavic”. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*. ACL ’07. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 43–50. URL: http://dl.acm.org/citation.cfm?id=1567545.1567554.

[RCAC05] M. Ruiz-Casado, E. Alfonseca, and P. Castells. "Automatic extraction of semantic relationships for Wordnet by means of pattern learning from Wikipedia". In: *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB 2005).* Alicante, Spain: Springer Verlag, 2005, pp. 67–79.

[SY73] G. Salton and C. S. Yang. "On the specification of term values in automatic indexing". In: *Journal of Documentation* 29.4 (1973), pp. 351–372.

[SB88] G. Salton and C. Buckley. "Term-weighting approaches in automatic text retrieval". In: *Information Processing and Management*. Vol. 24. Pergamon Press, 1988, pp. 513–523.

[Sha48] C. E. Shannon. "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27 (1948), pp. 379–423, 623–656. URL: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

[SAK03] A. Sheth, I. B. Arpinar, and V. Kashyap. "Relationships at the heart of semantic web: Modeling, discovering, and exploiting complex semantic relationships". In: *Enhancing the Power of the Internet Studies in Fuzziness and Soft Computing*. Springer-Verlag, 2003, pp. 63–94. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.1516.

[SL99] J. Silva and G. Lopes. "A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multi-word Units". In: *Proceedings of the 6th Meeting on the Mathematics of Language*. Orlando, USA, 1999, pp. 369–381. URL: http://hlt.di.fct.unl.pt/jfs/MOL99.pdf.

[SL10] J. Silva and G. Lopes. "Towards automatic building of document keywords". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. COLING '10. Beijing, China: Association for Computational Linguistics, 2010, pp. 1149–1157. URL: http://dl.acm.org/citation.cfm?id=1944566.1944698.

[Sin11] R. M. K. Sinha. "Stepwise mining of multi-word expressions in Hindi". In: *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. MWE 2011. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 110–115. ISBN: 978-1-932432-97-8. URL: http://dl.acm.org/citation.cfm?id=2021121.2021143.

[SJFHTsK05] R. Sun, J. Jiang, Y. Fan, T. Hang, C. Tat-seng, and C. M. yen Kan. "Using syntactic and semantic relation analysis in question answering". In: *Proceedings of the Fourteenth Text REtrieval Conference*. 2005. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.7043.

[TLR11] L. Teixeira, G. P. Lopes, and R. A. Ribeiro. "Automatic Extraction of Document Topics". In: *Technological Innovation for Sustainability - DoCEIS 2011*. IFIP AICT series 349. Springer Berlin Heidelberg, 2011, pp. 101–108. URL: http://link.springer.com/chapter/10.1007%2F978-3-642-19170-1_11.

[TC03] E. Terra and C. L. A. Clarke. "Frequency estimates for statistical word similarity measures". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Vol. 1. NAACL '03. Edmonton, Canada: Association for Computational Linguistics, 2003, pp. 165–172. DOI: 10.3115/1073445.1073477. URL: http://dl.acm.org/citation.cfm?id=1073477.

[TYB03] D. Tikk, J. D. Yang, and S. L. Bang. "Hierarchical text categorization using fuzzy relational thesaurus". In: *KYBERNETIKA-PRAHA*. Vol. 39. 5. 2003, pp. 583–600. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.109.649.

[TBR10] O. Trigui, L. H. Belguith, and P. Rosso. "An automatic definition extraction in Arabic language". In: *Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems*. NLDB'10. Cardiff, UK: Springer-Verlag, 2010, pp. 240–247. ISBN: 3-642-13880-2, 978-3-642-13880-5. URL: http://dl.acm.org/citation.cfm?id=1894525.1894559.

[VS07] J. Ventura and J. F. Silva. "New Techniques for Relevant Word Ranking and Extraction". In: *Proceedings of the 13th Portuguese Conference on Artificial Intelligence, EPIA 2007*. Vol. LNAI 4874. Guimarães, Portugal: Springer-Verlag, 2007, pp. 691–702. ISBN: 978-3-540-77000-8. URL: http://link.springer.com/chapter/10.1007%2F978-3-540-77002-2_58.

[VS12] J. Ventura and J. F. Silva. "Mining Concepts from Texts". In: *Proceedings of the International Conference on Computational Science – ICCS 2012*. Vol. 9. Omaha, Nebraska, U.S.A.: Elsevier, 2012, pp. 27–36. URL: http://www.sciencedirect.com/science/article/pii/S1877050912001251.

[VS13a] J. Ventura and J. F. Silva. "Automatic extraction of explicit and implicit keywords to build document descriptors". In: *Proceedings of the 16th Portuguese Conference on Artificial Intelligence, EPIA 2013*. Vol. LNAI 8154. Angra do Heroísmo, Azores, Portugal: Springer-Verlag, 2013, pp. 492–503. ISBN: 978-3-642-40668-3. URL: http://link.springer.com/chapter/10.1007/978-3-642-40669-0_42.

[VS13b] J. Ventura and J. F. Silva. "Using clusters of concepts to extract semantic relations from standalone documents". In: *Proceedings of the 16th Portuguese Conference on Artificial Intelligence, EPIA 2013*. Vol. LNAI 8154. Angra do Heroísmo, Azores, Portugal: Springer-Verlag, 2013, pp. 516–527. ISBN: 978-3-642-40668-3. URL: http://link.springer.com/chapter/10.1007/978-3-642-40669-0_44.

[WSN10] E. Wehrli, V. Seretan, and L. Nerima. "Sentence analysis and collocation identification". In: *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications*. MWE 2010. Beijing, China: Association for Computational Linguistics, 2010, pp. 28–36. URL: http://www.aclweb.org/anthology/W10-3705.

[WLB12] W. Won, W. Liu, and M. Bennamoun. "Ontology learning from text: A look back and into the future". In: *ACM Comput. Surv.* 44.4 (Sept. 2012), 20:1–20:36. ISSN: 0360-0300. DOI: 10.1145/2333112.2333115. URL: http://doi.acm.org/10.1145/2333112.2333115.

[XYL10] *Keyword Extraction and Headline Generation Using Novel Word Features*. Atlanta, Georgia, 2010, pp. 1461–1466.

[Zer10] K. Zervanou. "UvT: The UvT term extraction system in the keyphrase extraction task". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval '10. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 194–197. URL: http://dl.acm.org/citation.cfm?id=1859664.1859706.

[ZXTL06] K. Zhang, H. Xu, J. Tang, and J. Li. "Keyword extraction using support vector machine". In: *Proceedings of the 7th international conference on Advances in Web-Age Information Management*. WAIM '06. Hong Kong, China: Springer-Verlag, 2006, pp. 85–96. ISBN: 3-540-35225-2, 978-3-540-35225-9. DOI: 10.1007/11775300_8. URL: http://dx.doi.org/10.1007/11775300_8.

[ZS03] H. Zhou and G. W. Slater. "A metric to search for relevant words". In: *Physica A: Statistical Mechanics and its Applications* 329.1-2 (2003), pp. 309–327. ISSN: 0378-4371. DOI: http://dx.doi.org/10.1016/S0378-4371(03)00625-3. URL: http://www.sciencedirect.com/science/article/pii/S0378437103006253.

[ZW10] J. Zhou and S. Wang. "Concept Mining and Inner Relationship Discovery from Text". In: *New Advances in Machine Learning*. Ed. by Y. Zhang. InTech, 2010. ISBN: 978-953-307-034-6. DOI: 10.5772/9383. URL: http://www.intechopen.com/books/new-advances-in-machine-learning/concept-mining-and-inner-relationship-discovery-from-text.

A

Classification tables for *ConceptExtractor*

A.1 Classification for single-word concepts – English corpus

Table A.1: Classification for single-word concepts – English corpus.

Word	**C.**	**Word**	**C.**	**Word**	**C.**	**Word**	**C.**
often		nervousness	X	pieces	X	by	
have		result		had		rare	X
truly	X	almost		they		be	
thus		developed		receptors	X	silent	X
other		more		1998		koch	X
palliative	X	1		daily	X	vaginismus	X
risk		means		all		request	X
keller	X	raised		required		entrepreneurs	X
include		amino	X	concerned		foreign	X
during		microleakage	X	haplotypes	X	relying	X
working	X	eradicate	X	spring	X	integration	X
rhesus	X	opened		said		when	
come		no		help		these	

Continues on next page

Table A.1 – continued from previous page

Word	C.	Word	C.	Word	C.	Word	C.
punitive	X	that		use		due	
b		pulse	X	did		vessel	X
saccharin	X	still		as		referred	
him		make		large		were	
used		between		staff	X	regional	X
foul	X	recovery	X	however		guo	X
manifestations	X	mds	X	shown		without	
own		achieved		martin	X	others	
cytokine	X	part		similarities	X	copies	X
using		only		dmt	X	average	X
cbp	X	is		this		his	
an		obama	X	neglected	X	effort	X
chiropractic	X	extraction	X	diagnosed		contributions	X
october	X	3		mild		expressive	X
describes		necessary		overall	X	psychiatry	X
director	X	there		origin	X	made	
regarding		such		adults	X	proponent	X
because		s		according		at	
to		making		resulting		even	
disinhibition	X	specificity	X	who		anxious	X
from		for		would		metabolic	X
caused		camp	X	currently		departments	X
should		week	X	some		on	
scarce	X	cat	X	conical	X	date	X
project	X	18		sotai	X	dementia	X
immortality	X	them		argues	X	cdc	X
also		allow		in		books	X
sample	X	lymphomas	X	elimination	X	much	
fluency	X	added		implants	X	has	
cp	X	organ	X	position	X	especially	
do		sublingual	X	hyperekplexia	X	ad.	X
each		stromal	X	theories	X	survivors	X
rockefeller	X	same		runs	X	widely	
president	X	failure	X	ethan	X	viruses	X
warsaw	X	close		mcbride	X	injection	X
its		quote	X	agent	X	active	X
every		dioxide	X	infant	X	i	

Continues on next page

Table A.1 – continued from previous page

Word	**C.**	**Word**	**C.**	**Word**	**C.**	**Word**	**C.**
italy	X	r	X	analysis	X	further	
eventually		included		up		different	
known		called		into		arrogance	X
invention	X	can		with		embryo	X
fritz	X	fiber	X	providers	X	suspension	X
cytosine	X	tellurium	X	goal	X	out	
prevent	X	encoded	X	unidentified	X	are	
been		controversy	X	selected		nose	X
whose		later		many		will	
early		usually		first		crowns	X
computers	X	24		or		resistance	X
records	X	above		named		air	X
valuable	X	root	X	regenerated	X	she	
both		so		wilkins	X	disciplines	X
e		applied		any		vector	X
30		pathological	X	detailed	X	never	
new		until		subjects	X	a	
and		wilson	X	towards		half	X
researched	X	difference	X	references	X	either	
may		sailors	X	her		against	
not		generally		orders	X	sham	X
believe	X	grandparents	X	he		therapists	X
experimental	X	gac	X	full	X	leader	X
majority	X	since		after		before	
affect	X	within		those		replaced	

A.2 Classification for multi-word concepts – English corpus

Table A.2: Classification for multi-word concepts – English corpus.

Word	C.	Word	C.
duplicates of the		islamic world	X
distinguish from		forbidden in	
men who		exacerbation of	
steptoe and		hypersonic sound	X
surge of		cell-mediated immunity	X
doubled in		solute carrier	X
resorted to		exceptions to this	
restrictive abortion laws	X	run-in phase	X
kolli hills	X	botanic gardens	X
australasian society of		fact very	
robotics and		accusing the	
illustrated with		halves of	
environmental factors	X	fermentable fiber	X
gd nct	X	germ theory of disease	X
ligamentous laxity	X	bloodless surgery	X
la graufesenque		perforated by	
i've been		smokeless tobacco	X
ego-dystonic sexual	X	airway or	
stick to		incurable disease	X
cheat day	X	one hundred	
disengagement theory	X	overvalued idea	X
theorizes that the		mung beans	X
iso 14001	X	bioethicist jacob	X
activation is		removing the	
pgrs are		il28b gene	X
sport psychology	X	loops of	
lysyl oxidase	X	biventricular pacing	X
ranch in		wouldn't have	
kinotannic acid	X	pioneers of	
strategic alliance with		conclusion of	
steamed rice	X	m.d. degree	X
mad2 and		sharpe 2006	X
vitality and		bach flower	X

Continues on next page

Table A.2 – continued from previous page

Word	C.	Word	C.
left of		tachycardia is	
arca and		rahe stress	X
trustees of		gastrointestinal tract	X
intravitreal injection	X	rheumatoid arthritis	X
circulation and		redirects here	
investigational new		singularity is	
enzyme in		fruit or	
etiology of		senior-loken syndrome	X
books on the		decussation of	
paroxysmal nocturnal hemoglobinuria	X	maziar ashrafian bonab	X
working against		diffuse through	
sania nishtar	X	ratios in	
paramedian clefts	X	authorizes the	
refuge in		il-5 and	
appearing in the		melt and	
begins the		low-fat diets	X
junfeng was		mount sinai school of	
ragna rok	X	identical to	
cone cells	X	vinca alkaloids	X
susan dimock	X	recovering from	
gus and wes	X	asperger syndrome	X
neutralized by		searching for	
succumbed to		openness to	
often associated		blood-brain barrier	X
contributors to		contemplation and	
reflections on		constant and	
faux pas	X	wheaton franciscan	X
is a protein that in		erythroid progenitors	X
absorption and		odium attaching	X
unlock the		conscientious objectors	X
symbol for		confessed to	
dna-binding protein	X	faced with	
uttar pradesh	X	phil brewer	X
transgenic mouse	X	tertiary structure	X
retinitis pigmentosa	X	reluctance to	
diminishes the		organization based in	
commotio cordis	X	brugsch papyrus	X

Continues on next page

Table A.2 – continued from previous page

Word	C.	Word	C.
assurance of		unit in	
baylor college	X	brook university school of medicine	X
hope is a		shell shock	X
tomb of		anaesthetics and	
networks to		want to	
listening to		stack of	
absorb the		levonorgestrel-only users	X
zone and		gentamicin is also	
you are		electroconvulsive therapy	X
seen by		organizational effects	X
sublingual immunotherapy	X	gus and	
broca's area	X	post-traumatic stress	X
comment on the		whole-genome shotgun	X
abandonment of		asthma and	
retrieved from		impulsive and	
safely and		military personnel	X
asteroid m	X	penicillin and	
complaints commission	X	bombings of hiroshima and	
publishers of the		simulating a	
francs to		faked his	
thirty years	X	abundant protein	X
westminster hospital	X	mein kampf	X
josé farmer	X	biodefense and	
fossils from		physiology of	
heal the		analogy with	
frequented by		kshara sutra	X
evicted from		sabrina fullerton	X
zygomatic arch	X	marketed as an	
inductive logic	X	ctla4 is	
sums of		galveston national laboratory	X
number in		winter months	X
pangamic acid	X	thinness as	
monte albán		sagara sanosuke	X
ibs-like symptoms	X	pontifical academy of	
aero-digestive tract	X	vary based	
scientific journals	X	absence of	
antagonists such as		haitian health	X

Continues on next page

Table A.2 – continued from previous page

Word	**C.**	**Word**	**C.**
daneeka is		students and	
shih ch'un	X	depolarizing current	X
usable cannabis and		preclinical studies	X
autumn of		segment on	
criminalization of		matriculants to the	
appendages of		glutinous rice	X
defence of		natriuretic peptide	X
the boys		probiotics can	
platelet-derived growth	X	ukrain is	
must also		decaffeinated coffee	X
carcinomas of the		connects the	
gall bladder	X	michelson-morley experiment	X
creams and		modification is	
predeceased him		critics of	
culminating in		rundown to	
hope for		auditioned for	
necessary to carry		lighter than	
ignored by		planck institute of biochemistry	X
avoided the		multivitamins in	
lichen planus	X	ongoing medical	X
ola mau	X	belief in	
saliva and		securing of	
ceroid lipofuscinosis	X	aspires to	
polst form	X	point out	
bystander effect	X	planets and	
dipped in		exclusively on	
dolly the		work for	
rohs 2		steady and	
depletion of		homosexuality as a mental disorder	X
reconciles with		recessive disorder	X
she's a		stuyvesant high school	X
benefit from		hundreds of	
deaminases acting on rna	X	ticks of	
the demands of		responsible for	
are available		rather than	
the effectiveness of		it has been found that	
the role of		lack of	

Continues on next page

Table A.2 – continued from previous page

Word	**C.**	**Word**	**C.**
an additional		to create	
with that of		have been	
an roi		during the	
which is		which are	
have been		responsible for	
are available		from the	

A.3 Classification for single-word concepts – Portuguese corpus

Table A.3: Classification for single-word concepts – Portuguese corpus.

Word	C.	Word	C.	Word	C.	Word	C.
úlcera	X	heston	X	brigam	X	primeiro	
senhores	X	quando		leal	X	day	X
consome	X	sagan	X	paz	X	córtex	X
tais		mas		procedimentos	X	columbia	X
lá	X	contém		comuns	X	sempre	
manifesto	X	próstata	X	após		na	
dos		lino	X	solvente	X	nome	
orgânicos	X	escuro	X	ao		às	
através		tinha		suas		cor	X
bem		utilizados	X	apenas		seis	X
cardíaca	X	outros		pais	X	pelos	
estados		emocional	X	quebra	X	vários	
expõe	X	anime	X	filmes	X	no	
entanto		embolia	X	qualidade	X	tipicamente	X
hipotálamo	X	todas		tanto		foi	
princípios	X	importante		das		uso	
assistência	X	ela		mesmo		trabalhou	X
sua		circuncisão	X	pode		sacral	X
fez		notável	X	contra		resultado	X
bifocais	X	pois		esse		qual	
esta		popularidade	X	emenda	X	parte	
fim		meio		do		maior	
compatriotas	X	extensão	X	versão	X	todo	
uma		integração	X	é		radical	X
todos		um		blanca	X	possuem	
sono	X	constituição	X	mais		xaropes	X
afogamento	X	br	X	diferentes		relacionados	X
feita		e		síntese	X	taxonómicos	X
apesar		que		disco	X	atualmente	
sertaneja	X	está		constantino	X	dosagem	X
skinner	X	considerado		seja		global	X
recursos	X	ele		até		ser	
a		dois		são		podem	

Continues on next page

Table A.3 – continued from previous page

Word	C.	Word	C.	Word	C.	Word	C.
teve		purulento	X	especiarias	X	sob	
4		igreja	X	clínica	X	o	
já		kg	X	abuso	X	funcional	X
criada	X	entre		bacteriana	X	podendo	
oral	X	proteína	X	sem		estudantes	X
com		esofágico	X	não		sendo	
nos		à		assim		lisossomal	X
embora		também		oxidados	X	essa	
as		lancet	X	luz	X	treze	X
roberts	X	direcional	X	para		pyne	X
esses		alexandria	X	porém		três	
seus		geralmente		açúcar	X	há	
profundamente	X	antes		solar	X	exemplifica	X
relatada	X	segundo		isso		holismo	X
lema	X	1918	X	este		devido	
durante		falar	X	vez		instrumento	X
the		seria		investigador	X	químico	X
servidos	X	tai	X	onde		aos	
capacidade	X	sangue		agent	X	atriz	X
europeu	X	de		comício	X	alguns	
depois		acumulado	X	desde		sobre	
republicano	X	duas		cerca		algumas	
obra	X	fisiológica	X	em		quanto	
8		6		baniszewski	X	por	
20		suspeita	X	p53	X	quantidades	X
visão	X	estão		passou		ching	X
primeira		co	X	vasos	X	ou	
ponto		pernambuco	X	eram		forma	
tipo		centro	X	experiência	X	caso	
muitos		revisar	X	inguinais	X	transferência	X
hélices	X	seu		próprio		análise	X
menos		2010		conhecido		demonstração	X
trabalhar	X	single	X	era		nova	
fitzgerald	X	cálcio	X	sejam		evolução	X
monoclonais	X	cefaléia	X	classificação	X	ainda	
integrada	X	desse		sepultado	X	se	
nasceu	X	jovem	X	roubo	X	francês	X

Continues on next page

Table A.3 – continued from previous page

Word	**C.**	**Word**	**C.**	**Word**	**C.**	**Word**	**C.**
causa		história		colchões	X	ocorre	
como		uv.	X	pela		dose	X
drogas	X	portuguesa	X	qualquer		modo	
tadalafila	X	sofrem	X	só		discriminar	X
distribuída	X	da		manter		foram	

A.4 Classification for multi-word concepts – Portuguese corpus

Table A.4: Classification for multi-word concepts – Portuguese corpus.

Word	**C.**	**Word**	**C.**
diretrizes da		dignitária da ordem	X
vazamento de		proporcionou uma	
we're only in	X	elevar a	
billy the kid	X	repetições de	
multas e		capitão gregório	X
motoo kimura	X	tourette é	
desenvolveu em		excessiva de	
tocante à		prelazia do opus dei	X
impulsionador da		purificação de	
tubas uterinas	X	toracotomia de emergência	X
detectam a		inclusive a de	
dada a		portanto um dos componentes	
escândalo do		cerveja e	
coletânea de		aconselha que	
consideravelmente mais		bosio e col	X
anel benzênico	X	menciona que	
sentido de		razões pelas quais	
amadurece e		afectar a	
excluído do		portaria 518	X
chamando-a de		começo da década	X
chegado à		flambada e	
permanência do		ministra-chefe da casa civil	X
introduzindo o		almofadas hemorroidárias	X
mirtazapina é		óxido nitroso	X
conjuntamente com		grade de orientação	X
comutação de		transformando-os em	
feito no		sobreviveu a	
perspectivas de		nutricionistas do	
apreensão e		suplementos de	
obsessiva com		assistentes sociais	X
futura esposa	X	dubin-johnson é uma	
abertura da		porque estas	
autorizou o		prevenida através do	

Continues on next page

Table A.4 – continued from previous page

Word	C.	Word	C.
pense que		humanist association	X
vagos e		sarna sarcóptica	X
láctico e		mudam-se para	
fúria narcisista	X	prematura de	
intracerebral hemisférica	X	manchas vermelhas na	
preparou para		reduzidos em	
piloto de		declarar a	
variantes de		enrolados em	
organizam em		serenoa repens	X
aprender a		sutras de	
associação com		autossômica recessiva	X
taxa mais		pertencente à classe	X
perversão sexual	X	pediatria e	
obtém-se a		bordetella pertussis	X
tenderia a		cena de	
angariação de fundos para		isentos de	
batizada com o nome de		pólo de	
rainha vitória	X	encontrar um	
concorreu a		fecundar a	
floresta e		especializar em	
material da		segmento de dna	X
carbonato de cálcio	X	protecção contra	
obstrutiva e		terapia ocupacional	X
reciclagem de		impediu de	
publicamente sua		inaugural em	
contratura do		olivier e	
roteiristas de		importantes no	
empurramento com		aforismos de	
anemias hemolíticas	X	ipseo new	X
linguísticas e		hélice é	
condições de higiene	X	proporcionar a	
aktion t4	X	convertase da via	X
juscelino kubitschek	X	agrupadas de	
age através		processou a	
afetadas pela		observado em	
tem sido		neutralização da	
psicopatologia geral	X	consulta com	

Continues on next page

Table A.4 – continued from previous page

Word	C.	Word	C.
victorino de sousa	X	grande-colar da ordem	X
é grande		habituados a	
lâmpadas de		agraciada com	
bete balanço	X	auxilia no	
trailer de		inteiramente à	
estuda os		xbox 360	X
muita água		jugular interna	X
quadril é		gentílicos e topónimos	X
aprendem a		sergei rachmaninoff	X
autoestima pode		semelhança dos	
imersão em		desfaz a	
reconheça a		relataram que	
básicas de		transito intestinal	X
corticais e		localização e	
servia de		obrigatoriedade de	
living daylights	X	naufrágio do navio	X
distribuído para		caule e	
subgrupos de		défice de	
hospitalização de		dúzias de	
cercam a		claviceps purpurea	X
ácidos graxos	X	paramahansa yogananda	X
westwood village	X	cursa com	
josef stangl	X	registados em	
musculatura do		descendem de um	
especial de rastreamento de		deduziu que	
ocasionadas por		instâncias psíquicas	X
opióides e		kofi kingston	X
metformina e		ciente da	
ernest becker	X	rua da	
plasticidade fenotípica	X	estearato de	
rizomas lenhosos	X	consoante a	
rex allen	X	manuais e	
incumbência de		waldeck e	
reproduz-se por		comparar os	
princesa de		preenchida com	
wilmar de oliveira	X	lombar é	
veio ao		câncer no	

Continues on next page

Table A.4 – continued from previous page

Word	C.	Word	C.
comprimento de		sofriam com	
obrigando a		visto que	
neuroma de amputação	X	pentecostais e	
diferenciação das		retrato de	
incapacidade de se		offender index	X
casando com		indicação da	
erupções cutâneas	X	óvulos não	
seis em		estágios iniciais da	
figurados do sangue	X	frankfurt am maine	X
cava superior	X	restos tumorais	X
oxigenação cerebral	X	solas dos pés	X
elenco de		gás carbônico	X
incentivado por		insucesso de	
energia vital	X	lester young	X
lâmpada de		metabolizada no	
elaborar um		cox é	
lidam com		promessas de	
discretos e		provocadas pela	
glicerol e		realçar o	
esposa maria	X	automáticas e	
roda no		expressar um	
secretada na		baseia-se nos	
palidez e		pomadas e	
declarações polêmicas	X	norman granz	X
filia-se ao		negociado com	
eletrônica de		jyh cherng	X
substituto de		alberto eduardo	X
esclarecer a		reivindicação dos	
unidos com a sua		toracotomia de	
saxofone tenor	X	equação de	
biossíntese de		entrada de água	X
abdullah ibn al-jarrah	X	transmitida através	
eliminando assim		assombrará o mundo	
journal of		crê-se que	
seu nome		xix e início do século	
que é		forma de	
membro da comissão	X	escola médica	X

Continues on next page

Table A.4 – continued from previous page

Word	C.	Word	C.
exibida no brasil	X	toda a	
mais tarde		que pode	
entre outros		jornais locais	X
ser vivo	X	inibidores enzimáticos	X
contra a guerra	X	governo dos estados unidos	X
acordo com		apesar de	

A.5 Classification for single-word concepts – German corpus

Table A.5: Classification for single-word concepts – German corpus.

Word	C.	Word	C.	Word	C.	Word	C.
zunächst		haben		dem		im	
betonung	X	belangen	X	lassen		muss	
methoden	X	zeigen		belege	X	pcr	X
tiefer	X	diesen		darauf		karl	X
kosmetischer	X	einzelnen	X	selben		hat	
zusammenwirken	X	weiteren		mehr		am	
einzurichten	X	entwickelt	X	erkennen	X	dass	
synthetisierten	X	dna-analyse	X	scientific	X	number	X
regulär	X	warburg	X	david	X	gibt	
zahnhalteapparat	X	sogenannten		letzte	X	die	
intravenöse	X	fähigkeiten	X	besitzen	X	ein	
bakterielle	X	aufgrund		eine		als	
postdoktorand	X	durch		wie		den	
hyperplasie	X	erlassene	X	seltene	X	zum	
seiner		keine		für		auf	
fehlbildung	X	indikation	X	dänemark	X	b.	
regel		sehr		mary	X	4	
ausprägung	X	napoléon	X	ohne		also	
neurologie	X	eingesetzt		drittel	X	seit	
obdachlose	X	mehrheit	X	hoppe	X	ob	
bevölkerung	X	promovierte	X	mitte	X	beim	
internationalen	X	tätigkeiten	X	medicine	X	1964	
unterschiedlichen		während		schnell	X	so	
tuberkulose-erkrankung	X	sollte		monos	X	esche	X
kaiser-wilhelms-akademie	X	vorfeld	X	engeren	X	30	
herkömmliche	X	ersten		dies		akh	X
allerdings		außerdem		dossier	X	bedarf	X
anstellung	X	ebenfalls		cohnheim	X	über	
auswärtsdrehung	X	anerkanntes	X	äußert	X	baby	X
analoga	X	john	X	je		a	
beschleunigung	X	entfernt		nephron	X	aus	
längs	X	außer		unter		von	
weiterer		besteht		jakob	X	bis	

Continues on next page

Table A.5 – continued from previous page

Word	C.	Word	C.	Word	C.	Word	C.
schützt	X	daraus		ihrem		heute	
erdbeben	X	stammes	X	saturn	X	lehnt	X
philosophie	X	allgemeine	X	und		bei	
verschiedene		zahlreichen	X	sowohl		etwa	
kapazitäten	X	worden		wollte		μm	X
beiträge	X	gesetz		einen		sind	
mandibulae	X	erstmals	X	anzahl	X	beide	
universitäten	X	solche		anfang	X	meist	
übersetzt	X	ihrer		zeit		pro	
begegnungen	X	mutterleib	X	seine		1963	
interessen	X	ethnologie	X	wort	X	ist	
arzneimittel	X	psychologie	X	finsternis	X	aber	
nicht		oder		kurz		z	
ärztlichen	X	leitung	X	diesem		was	
synaptischen	X	befasste	X	dort		ab	
ct-koronarangiographie	X	gegenüber	X	dekaden	X	selbst	
interleukine	X	regulatoren	X	tenor	X	ester	X
penisverkrümmung	X	später		de		15	
frankfurt	X	manuelle	X	kann		dann	
tätigkeit	X	immer		1944		war	
behandlungsdauer	X	mikrobiologie	X	tagen	X	gegen	
könnten		noch		der		zu	
beschneidung	X	meldung	X	berater	X	wird	
pigmentosa	X	hatte		des		8	
beispielsweise		ungeladene	X	um		er	
kurort	X	hier		zur		an	
fachrichtungen	X	verminderung	X	einsatz		bzw	
präoperativen	X	eitrigen	X	diese		da	
portugal	X	stumm	X	nach		dazu	
schädlich	X	medizinern	X	werden		seinem	
roberto	X	bänden	X	jens	X	das	
kommt		dabei		allem		1	
gemeindepfarrer	X	ehrenbürger	X	science	X	vielen	
begann		ihren		ziel	X	vor	
sklerose	X	konnte		gruppe		folgen	
entsprechende	X	ermöglichte	X	phase	X	vom	
vorgestellt	X	1949		sie		ihm	

Continues on next page

Table A.5 – continued from previous page

Word	**C.**	**Word**	**C.**	**Word**	**C.**	**Word**	**C.**
begleiterkrankungen	X	französischen	X	zunge	X	folge	
ernennung	X	wurden		wieder		gingen	X
seinen		zwei		sir	X	es	
diagnose	X	können		krebsen	X	man	
vorstellung	X	monaten	X	bereits		sgb	X

A.6 Classification for multi-word concepts – German corpus

Table A.6: Classification for multi-word concepts – German corpus.

Word	C.	Word	C.
therapierbarkeit und		beschleunigung der	
antelope valley california poppy reserve	X	magens bei	
literarisches werk	X	empfindlich ist	
gesichertes wissen	X	absetzen des	
spielt eine wichtige	X	bindungen zu	
bewegungsfähigkeit des		fruchtblätter sind zu	
assoziation mit anderen		edwin smith	X
anthropologin und		nèi jing	X
neubildung von blutgefäßen	X	schädigungen des	
haemophilus influenzae	X	bekanntesten ist	
bahnbrechend und		libri duo	X
ventromedialen präfrontalen	X	funktionsminderung der	
nuklearmedizinischen verfahren	X	eizelle nicht	X
horst-eberhard richter	X	jahren begann er	
vergrößert werden		anna arfelli	X
leprahilfswerk iran	X	amtes als	
ernährt sich von den		züchtung und	
bewirkte die positive entscheidung	X	entlang einer	
eitriges sekret	X	siegeszug des	
komposita der stammsilbe	X	häufiger betroffen als	
kommentaren und		baroness murphy	X
studienaufenthalten in		maßgeblich an der	
verbiegung der		ehesten mit	
ritterorden vom heiligen grab zu		billigend in kauf	X
engagierten sich		gesundem gewebe	X
verschluss des		gelesen und	
trochanter major	X	richtungen der	
aufrichtung der		händen und	
grundgedanke der		nun in	
ruhr-universität bochum	X	laboratoriums der	
fliegende augenklinik	X	offenen brief an die	
beidäugigen sehens und		macfarlane burnet	X
rechtsanwalts und		im betrieb	

Continues on next page

Table A.6 – continued from previous page

Word	C.	Word	C.
prozent zu		floh er	
mitleidenschaft gezogen	X	farben der	
rückübertragung des		funktionsweise der	
bundeswehrzentralkrankenhaus koblenz	X	intensivierung der	
morphologie und		penny brookes	X
dura mater	X	sah er	
renato dulbecco	X	linderung von	
erwähnte er		ehe stammt	
künste und wissenschaften	X	blutes und	
nachfolgerin wird die		überschneiden sich	
aufrechterhaltung der		straßburger zeit	
feldberg foundation	X	alternativen zur	
renal-tubuläre azidose typ	X	allergy and	
beugeseiten der		gedenktafel an	
rückhalt in		nannte es	
problemorientiertes lernen	X	verlegte er	
fälle kommt es		westküste der	
damalige präsident	X	gefallen ist	
kolorektales karzinom	X	angina pectoris	X
vergebener wissenschaftspreis	X	kreuzbein und	
ordentlichen professor für		organell der	
befehlshaber der sicherheitspolizei	X	gibt es	
vorsicht geboten	X	vorkämpfer der	
großer bedeutung	X	lilly and	
kaiser-wilhelm-akademie für das militärärztliche	X	columbia universität	X
effektivität dieser		eingehängt und	
fossa pterygopalatina	X	wohl der	
ernährt sich von		saures milieu	X
nomina anatomica	X	bestrahlung die	
sicherstellung der		abgezogen werden	
leon orris jacobson	X	uwe henrik peters	X
alveolaris inferior	X	knochenmark und	
stiftungsvorstands des		herausgelöst und	
vermindert werden		barkas smh	X
automotive medicine	X	nach einiger	
fachwissenschaftler der medizin	X	bad herrenalber	X
unterstützen die		eigenaktionen des	

Continues on next page

Table A.6 – continued from previous page

Word	**C.**	**Word**	**C.**
mao zedong	X	boucher de	
expertin für		year against	X
endothelial growth factor	X	tor seidel	X
levin jacobson	X	heilkraft der	
eigenschaft als		gustav adolf	X
verfärben sich die		hiatus oesophageus	X
zubereitung von		regelwerk der	
zdrawko georgiew	X	silberdistel als	
jungen als auch mädchen	X	wachheit und	
röhrenförmige herzen	X	sanatorium schloss	X
grundstück in		vakant wurde	X
seenot und		gmds und	
verweisen auf		schnell zu	
aufführung des		ersteller der	
destilliertes wasser	X	chemischen und	
betrieblichen gesundheitsmanagements	X	vorgesetzter war	
beyond words	X	besten mit	
maximiliansorden für wissenschaft	X	orgastischen potenz	X
zehntes kind	X	eignung des	
neutrophiler granulozyt wandert	X	folgeschäden wie	
übertragbarkeit von		epitope der	
seine sporenlager		physik an der	
protozoen oder		ausbau der	
primärem hyperparathyreoidismus	X	serengeti darf nicht sterben	X
tumorforschung und		berufenes mitglied	X
elektromagnetisches feld	X	jahres 2011	X
trizyklische antidepressiva	X	gelenkkapsel und	
restriktive regelungen	X	inspektionen vor ort	X
sabina spielrein	X	auffüllen des	
zeitschrift für		firmensitz in	
ausschneiden der		prävalenz in	
zeylmans van emmichoven	X	ester und	
vergiftungserscheinungen führen	X	kranker menschen	
umgebenden haut	X	kovalenko medal	X
tropischer pflanzen	X	umschreibung für	
tropenmedizinischen gesellschaft	X	akupunktur und akupressur	X
pathologisch-anatomischen institut	X	shanghán lùn	X

Continues on next page

Table A.6 – continued from previous page

Word	C.	Word	C.
geändert werden		individuum und	
einzelligen organismen	X	chalara fraxinea	X
präfrontalen cortex	X	durchtritt durch	
portugiesisch und		arylsulfatase b	X
klonierung von		situation des	
mittwoch im		sahen sich	
thyroidea inferior	X	womit er	
diplomate of		pol der	
erbkranken nachwuchses	X	zonierung der	
kurmethode auf		frauenarzt in	
fortpflanzungsfähigkeit beeinträchtigen	X	beschäftigte er	
deshalb besonders	X	verordnung zum	
bent brigham	X	gekauft und	
psychoaktiven substanzen	X	bundesbeauftragte für	X
ambulant durchgeführt	X	cold spring harbor	X
absolvent der		wenden ein	
gefahren und		sekrete der	
registrieren zu		studie an	
psychologische diagnostik	X	qualifizierte er sich	
abhilfe zu schaffen	X	muster für	
medicinisch-chirurgische zeitung	X	konzentriert sich	X
ulf von euler	X	zeige sich	
nicht-invasive methoden	X	ledige mütter	X
meg patterson	X	eindruck des	
theosophical society	X	francis galton	X
apparativer sprechhilfen	X	wartezeit von	
reichsleitung der		pius ix.	X
stadtverordneter der		entwickeln sich	
leutnant der		bereits in	
aufzugeben und		findet diese	
angehöriger des wissenschaftlichen beirates	X	rostflecken und pusteln	X
geschlossene reposition	X	holding gmbh	X
auf dem gebiet		auch bei	
kann eine		von den	
mit den		ist der	
es gibt		er auch	
ablauf der frist		durch das	

Continues on next page

Table A.6 – continued from previous page

Word	**C.**	**Word**	**C.**
in dieser funktion		vorhanden sein	
machtübergabe an		unter dem titel	
nationalpreis der ddr		vor allem	
für den		ist in	
erhalten hatte		mit einem	
rolle zu spielen		aber auch	