Extraction of Definitional Contexts from Restricted Domains by Measuring Synthetic Judgements and Word Relevance

César Aguilar
Pontificia Universidad Católica de Chile
Santiago de Chile
caguuilara@uc.cl

Olga Acosta
Pontificia Universidad Católica de Chile
Santiago de Chile
olgalimx@gmail.com

Abstract

In this article we present an ongoing work for extracting conceptual information from specialized-domain texts. Concepts are forms of dividing the world in classes and they are the fundamental pieces for constructing ontologies. In this sense, ontology learning is the (semi-) automatic support for constructing an ontology. Input data are required for the ontology learning and this data are the basic source from which to learn the relevant concepts for a domain, their definitions as well the relations holding between them. With this necessity in mind, we propose here a methodology that takes into account the level of synthetic judgements and word relevance in a sentence in order to filter out and rank sentences. Sentences with high relevance and low level of synthetic judgements should have at least a predicative verb characteristic of analytical definitions for being good candidates.

1 Introduction

Concepts are one of the most fundamental pieces of the cognition: humans daily use concepts for interacting with others and the world. According to Smith (1988), concepts mirror the way that we divide the world into classes, and much of what we learn, communicate, and reason involves relations among these classes. Additionally, Rosch (1978) argues that concepts promote the cognitive economy because the human beings attempt to gain as much information as possible about its environment while minimizing cognitive effort and resources.

Currently, due to the accelerated growth of digital information on the Web and other media as well the urgent necessity of obtaining relevant information in a fast and efficient way from these huge text sources, automated methods or approaches have been developed. For instance, in Maedche and Staab (2004) define *ontology learning* as a number of complementary disciplines that feed on different types of unstructured and semi-structured data in order to support a semi-automatic ontology engineering process. In line with this, Cimiano (2006) describes various sub-processes for constructing an ontology from texts where the concept extraction is an important phase. So, the ontology learning needs input data from which to learn the relevant concepts for a given domain.

According to these ideas, in this paper we sketch a methodology for recognizing candidates to analytical definitional contexts, according to the work developed by Sierra *et al*. (2008). We organize our work as follows: in section 2 we present general information about analytical definitions and the automated extraction of conceptual information. In section 3 we describe the function of adjectives as modifiers of a noun as well the distinction among descriptive and relational adjectives and the relation of descriptive adjectives with synthetic judgements in an attributive form. In section 4 we summarize the methodology proposed. In section 5 we show some preliminary results. Finally, in section 6 we present the future work.

2 Conceptual information

We consider as *conceptual information* the information expressed by specialized definitions, particularly in analytical definitions constituted by *Genus Term* and *Differentia*, following the criteria formulated by Smith (2004). In fact, this author considers that information expressed by these kinds of definitions is relevant to create ontologies based in lexical relations, specifically hyponymy/hypernymy and meronymy/holonymy relations. Smith argues that these relations, from a philosophical point of view, are basic and universal.

2.1 Analytical Definitions

An analytical definition is a formula for describing a concept, denoted by a linguistic tag, in terms of a superordinate concept (*Genus Term*), and a differentia distinguishing the concept defined from others with the same Genus Term.

For example, the next definition provides a description of the concept *lightning conductor* using one of the most common verbs (i.e., to be) for introducing a definition. In this case, the genus is the concept *device* while the differentia describes the function of the *lightning conductor*:

> [Lightning conductor $^{\text{Term}}$] is a [device $^{\text{Genus Term}}$] [that allows to protect the electrical systems against surges of atmospheric origin $^{\text{Differentia}}$].

2.2 Definitional contexts

Sierra *et al*., (2008) proposed a based-pattern method for extracting terms and definitions in Spanish. This relevant information is expressed in textual fragments called definitional contexts (or DCs) and are constituted by: a term, a definition, and linguistic or metalinguistic forms, such as verb phrases, typographical markers and/or pragmatic patterns, for example:

> The **primary energy**, in general terms, is defined as an energetic resource that has not been affected for any transformation, with the exception of its extraction.

We can see here a DC sequence formed by the term *primary energy*, the definition *that resource that…* and the verb pattern *is defined as*, as well other characteristic units such as the pragmatic pattern *in general terms* and the typographical marker (bold font) that in this case emphasizes the presence of the term.

For achieving this objective, the authors employ verb patterns operating as connectors between terms and definitions. Such patterns syntactically are predicative phrases (or PrP), configured around a verb that operates as a head of this PrP (e.g., to be, to characterize, to conceive, to consider, to describe, to define, to understand, to know, to refer, to denominate, to call, to name).

3 Adjectives

Based on Demonte (1999), adjectives are syntactic units modifying the noun's meaning and associating it with one or various attributes. There are two kinds of adjectives which assign properties to nouns: descriptive and relational adjectives. On the one hand, descriptive adjectives refer to constitutive features of the modified noun. These features are exhibited or characterized by means of a single physical property: color, form, character, predisposition, sound, and so on: *la silla verde* (e.g., *the green chair*). On the other hand, relational adjectives assign a set of properties, i.e., all the characteristics jointly defining names as *sea*: *puerto marítimo* (e.g., *maritime port*). In terminology, relational adjectives represent an important element for building specialized terms, e.g.: *inguinal hernia*, *venereal disease*, *psychological disorder* and others are considered terms in medicine. In contrast, *rare hernia*, *serious disease* and *critical disorder* seem more descriptive judgments and closely related with a specific context.

3.1 Syntactical Identification of Non-Relevant Adjectives

In line with what was just mentioned, if we consider the internal structure of adjectives, two kinds of adjectives can be identified: permanent and episodic adjectives (Demonte, 1999). The first kinds of adjectives represent stable situations, permanent properties characterizing individuals. These adjectives are located outside of any spatial or temporal restriction (i.e., *psicópata- psychopath*). On the other hand, episodic adjectives refer to transient situations or properties implying change and with time-space limitations. Almost all descriptive adjectives derived of participles belong to this latter class as well all adjectival participles (i.e., *harto-jaded*, *limpio-clean*). Spanish is one of the few languages that in syntax represent this difference in the meaning of adjectives. In many languages this difference is only recognizable through interpretation. In Spanish, individual properties can be predicated with the verb *ser*, and episodic properties with the verb *estar*.

Another linguistic heuristics for identifying descriptive adjectives is that only these kinds of adjectives accept degree adverbs, and they can be part of comparative constructions, for example, *muy alto* (Eng.: *very high*). Finally, only descriptive adjectives can precede a noun because

—in Spanish— relational adjectives are always postposed, i.e.: *la antigua casa* (Eng.: the old house).

3.2 Synthetic Judgements and Descriptive Adjectives

According to Kant (2013), *analytic* sentences are those whose truth seems to be knowable by knowing the meanings of the constituent words alone (e.g., *gynecologists are doctors*), unlike the more usual *synthetic* ones (e.g., *gynecologists are rich*), whose truth is knowable by both knowing the meaning of the words and something about the world.

We believe that synthetic judgements in an attributive position (e.g., *rich gynecologists*) are common in non-relevant sentences in specialized domains. This kind of judgements can be recognized from the descriptive adjectives obtained by linguistic heuristics mentioned in section 3.1.

4 Methodology

We present here our methodology for extracting conceptual information from a medical domain corpus. The input data consist of a corpus with POS tagged with FreeLing (Carreras *et al*., 2004).

4.1 Sentence Segmentation

The heuristics assumed here in order to segment our corpus by sentences take into account that a sentence must be separated by a point, to have at least a main verb, and the number of words must be greater than 10 words because the most short DC would have a single word term, the most long predicative verb-is defined as, a possible article preceding genus, genus term and, in this case, some arbitrary limit of words for the *differentia*).

4.2 Filtering out Sentences by Predicative Verbs

The set of sentences obtained by the above step are filtered out by considering predicative verbs mentioned in section 2.2, that is, if there is at least a predicative verb; then it is a good candidate to DC. For the case of *to be*, if it is the first word of the sentence, then it is discarded.

4.3 Chunking

We have used the library of Natural Language NLTK (Bird, Klein and Loper, 2009) in the Python language, for implementing a chunker in order to extract descriptive adjectives with heuristics described in section 3.1.

In this work, we propose a phase of quantification of *synthetic judgments* in candidate sentences as a further filter of non-relevant sentences. We assumed here that synthetic judgments are descriptive adjectives in an attributive position (e.g., *rare syndrome*). So, the higher amount of synthetic judgments in a sentence, the more likely sentence is non-relevant. We considered the set of descriptive adjectives obtained by heuristics as a mechanism for this quantification of syntheticity.

Acosta, Aguilar and Sierra (2013) point out relational adjectives have a higher probability of being part of terms. The heuristics considered in this experiment are:

<RG><AQ>
<VAE><AQ>
<D.*|P.*|F.*|S.*><AQ><NC>

Where RG, AQ and VAE as tagged with FreeLing, correspond to adverbs, adjectives and the verb *estar*, respectively. The tags *<D.*|P.*|F.*|S.*>* correspond to determinants, pronouns, punctuation signs and prepositions. The expression *<D.*|P.*|F.*|S.*>* is a restriction to reduce noise, since elements wrongly tagged by FreeLing as adjectives are extracted without this restriction.

4.4 Weighting Words

We evaluated relevance of simple words by means of a corpus comparison approach by applying the relative frequency ratio (Manning and Schütze, 1999) between two different corpora as in (1). Given that the syntactical pattern of most common terms in Spanish is <NC><AQ> (Vilvaldi, 2004), we take into account only nouns and adjectives in both corpora:

$$weigth(w_i) = log_2\left(\frac{f_{w_{i,D}}}{N_{w_{i,D}}} \Big/ \frac{f_{w_{i,R}}}{N_{w_{i,R}}}\right) \quad (1)$$

Where$f_{w_{i,D}}$, $N_{w_{i,D}}$correspond to the absolute occurrence frequency of w_i and the size of the domain corpus, respectively. Similarly, $f_{w_{i,R}}$, $N_{w_{i,R}}$ correspond to absolute occurrence frequency of w_i and the size of the reference corpus. The measure in (1) is only calculated for w_i's, where $\frac{f_{w_{i,D}}}{N_{w_{i,D}}} > \frac{f_{w_{i,R}}}{N_{w_{i,R}}}$. Otherwise, w_i can be used as part of a list of non-relevant words for purposes of quantifying non-relevance in sentences. On the other hand, words only occurring in domain are weighted as in (2). We assume that the reference corpus is large enough for filter out non-relevant words, hence words only occurring in the domain corpus will have a higher probability of being relevant so that the word's frequency can reflect its importance:

$$weight(w_{i,D}) = log_2(1 + f_{w_{i,D}}) \qquad (2)$$

4.5 Relevance of Sentences

The ranking of sentences is done by adding up the individual ranks of words present in the sentence. Formally, if s (that is, a sentence) has a length of n words, $w_1 w_2 \ldots w_n$, where n>10, then the ranking of the candidate s is the sum of the weights of all the individual words $w_i \in W$, where W are all of the relevant words weighted as mentioned in section 4.4. In contrast, if $w_i \notin W$, then its weight is zero.

5 Preliminary Results

Considering descriptive adjectives automatically extracted by heuristics for quantifying syntheticity, the first results show to be a good filter in order to remove non-relevant fragments by setting thresholds related with the number of descriptive adjectives in sentences. At the same time, the ranking of words achieves to sort sentences according to its relevance for the domain. Additionally, given that only sentences with predicative verbs are considered, a subset of the better ranked sentences are analytical DCs.

If we take into account words where relative frequency in reference is greater or equal than in domain (given its higher occurrence in reference than in domain, we assume they are non-relevant words) as part of this list for removing non-relevant sentences by setting thresholds (here, nouns and adjectives are included) improve significantly the results.

6 Future results

In a future phase of this experiment, we will implement a syntactic phase in order to remove more non-relevant sentences. For instance, sentences with *to be* verb are the most common sentences and which produce so much *noise* in results. Given this, we consider that a syntactic phase capable to assure the occurrence of specific syntactic structures will be an important advance in order to perform a better filtering.

On the other hand, we will continue with the recollection of more information for increasing the sections of science and technology in our reference corpus, in order to improve the word weighing and the calculation of relevance sentences.

Acknowledgments

This paper has been supported by the National Commission for Scientific and Technological Research (CONICYT) of Chile, Project Numbers: 3140332 and 11130565.

References

Olga Acosta, Gerardo Sierra and César Aguilar. 2011. Extraction of Definitional Contexts using Lexical Relations. International Journal of Computer Applications, 34(6): 46-53.

Steven Bird, Ewan Klein and Edward Loper. 2009. Natural Language Processing whit Python. O'Reilly, Sebastropol, Cal.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In Proceedings of the 4th International Conference on Language Resources and Evaluation LREC 2004, ed. by Maria Teresa Lino *et al.*, pp. 239-242. ELRA Publications, Lisbon, Portugal.

Philipp Cimiano. 2006. Ontology Learning and Population from Text. Springer, Berlin.

Violeta Demonte. El adjetivo. Clases y usos. La posición del adjetivo en el sintagma nominal. In *Gramática descriptiva de la lengua española*, ed. by Ignacio Bosque and Violeta Demonte. Vol. 1, Ch. 3, pp. 129-215. Espasa-Calpe, Madrid.

Immanuel Kant. 2013. Crítica de la razón pura, edited and traslated to Spanish by Pedro Ribas. Taurus, Madrid.

Alexander Maedche and Steffen Staab. 2004. Ontology Learning. In Handbook on Ontologies, ed. by Steffen Staab and Rudi Studer, pp. 173-190. Springer, Berlin.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Mass.

Rosch. 1978. Principles of categorization. In Cognition and Categorization, ed. by Elinor Rosh and Barbara Lloyd, pp. 27-48. Lawrence Erlbaum Associates, Hillsdale, NJ.

Gerardo Sierra, Rodrigo Alarcón, César Aguilar and Carme Bach. 2008. Definitional verbal patterns for semantic relation extraction. Terminology, 14(1): 74-98.

Barry Smith. 2004. Beyond concepts: ontology as reality representation. In Formal Ontologies in Information Systems, ed. by Achille Varzi and Laure Vieu, pp. 73-84., IOS Press, Amsterdam.

Edward Smith. 1988. Concepts and Thought. In Psychology of human thought, ed. by Robert J. Sternberg, pp. 19-49. Cambridge University Press, Cambridge, UK.

Jorge Vivaldi. 2004. Extracción de candidatos a términos mediante la combinación de estrategias heterogéneas. Ph. D. Dissertation. IULA-UPF, Barcelona.