Specialized Corpora Processing with Automatic Extraction Tools

Yuliya Goncharova, Beatriz Sánchez Cárdenas

Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 293 – 297

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Specialized Corpora Processing with Automatic Extraction Tools

Yuliya Goncharova[a]*, Beatriz Sánchez Cárdenas[b]

[a] *LIT, LSHA,Université de Strasbourg, 22 rue René Descartes, Strasbourg 67000, France*
[b] *LexiCon research group, Universidad de Granada, 11 Calle Buensuceso, Granada 18002, Spain*

Abstract

This research describes a protocol for specialized corpus analysis using natural language processing (NLP) tools to define semantic hierarchies of verbs in the specialized domain of Volcanology. The experimental analysis was carried out with a domain-specific corpus in English of approximately 500,000 tokens composed of dissertations and scientific articles in the domain of Volcanology. The combination of semantic and syntactic analysis results in the verb macro structure that illustrates the evolution of the meaning from more general to more specific verbs.

Keywords: frame-based terminology; verb macro structures; semantic frames; automatic terminology extraction

1. Introduction

The role of the verb in terminology has been traditionally undervalued, However, Frame-based Terminology, a new cognitive approach to specialized language, shows that the elaboration of verb hierarchies, based on their meaning definitions, can effectively contribute to the building of ontologies in specialized domains (Faber, 2002) (Faber, 2012). However, experience has shown that defining corpora-based verb hierarchies manually without a purpose-built tool is a very time-consuming procedure. In this context, automatizing domain-specific corpora analysis with a view to extracting hierarchical lexical networks is an important step towards knowledge organization.

* Corresponding author. Tel.: +33 7 60 60 59 10
E-mail address: yuliya.goncharova@etu.unistra.fr

doi:10.1016/j.sbspro.2013.10.650

The design of a fully automatic tool for this type of application is not possible without a great deal of previous study. It was first necessary to perform an experiment with existing NLP tools on a domain-specific corpus. After a description of the analysis carried out, the results obtained are illustrated with corpus-based examples. This experiment thus usefully contributed to the construction of this tool for automatic information extraction and opened the door to further research.

Our analysis was applied to the domain of volcanology, and the domain-specific English corpus used had a total of approximately 500,000 tokens. The corpus was composed of dissertations and scientific articles in this field. The stages in the analysis were performed with the POS-tagger TreeTagger (Schmid, 1994), the concordance tool AntConc (Anthony, 2005), the Word Sketch module of SketchEngine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004), and original scripts in Perl.

2. Related works

Based on Frame Semantics (Fillmore 1979, 1982, 1985 in Petruck, 1996) and Constructional Grammar (Goldberg, 2003), Frame-based Terminology Theory (Faber, Linares & Exposito, 2005; Faber, 2002, 2009, 2012) describes the terminology in a specialized domain as a system of concepts interacting at the syntactic, semantic and pragmatic levels.

Frame-based terminology focuses on: (1) conceptual organization; (2) the multidimensional nature of terminological units; and (3) the extraction of semantic and syntactic information through the use of multilingual corpora (Faber, 2009).

From this perspective, the verbs that encode actions, processes, and events are the most important parts of the system.

However, what is invariably overlooked is the fact that predicates in specialized texts are also instrumental in the analysis of conceptual structure. Verbs can be related to the conceptual categories that characterize the arguments they normally appear with in medical texts. In such cases, we have even found that verbs can trigger entire hierarchies of concepts and provide useful information in the structure of conceptual categories (Faber, 2002).

At the same time, the preferences of verbs for certain concepts in these hierarchies can provide information concerning meaning structures in the domain (Faber & Mairal Usón, 1999).

3. Methodology

After a part-of-speech (POS) tagging by TreeTagger the corpus was processed with a specially designed Perl script in order to obtain the corpus keywords list and extract the most domain specific verbs. The script combined the term frequency - inverse document frequency value (TF-IDF) and a general language reference corpora. The TD-IDF value reveals words with high frequency that appear on few documents of the corpus.

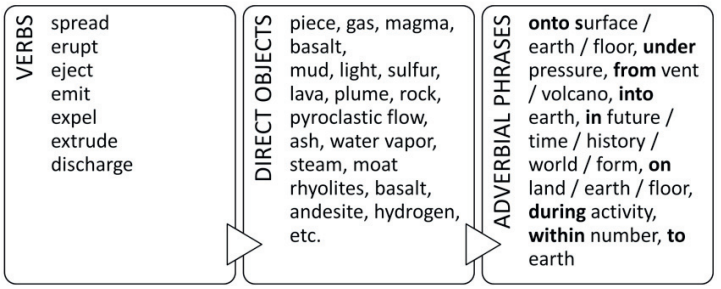

Fig. 1. Volcano activity verbs' arguments

In the second stage the corpus was sent to the Word Sketch module of SketchEngine, and sketches were obtained for each verb. The sketches contained the information regarding direct objects and adverbial phrases (Fig. 1). For this reason the corpus processed by TreeTagger was transferred to the AntConc concordance tool for manual analysis. These steps made it possible to study the selected verbs in context by using regular expressions with POS-tags and by extracting word clusters (2-6 words, frequency > 3). The typical arguments of the verbs were identified on the basis of their most characteristic semantic features.

This step allowed us to define the semantic frame of the domain-specific event (volcano eruption), based on causation verbs in the domain of Volcanology (*spread, erupt, eject, emit, expel, extrude, discharge*) (Fig. 2). The semantic frame seems to be an optimal context for semantic analysis. Since it is conditioned by the event and not by the verb, it gives a coherent representation of the whole set of concepts. The frame is the departure point for the study of verb arguments because these arguments can be subdivided according to the semantic roles. This analysis seems to be more complete than the classification based on the syntactic distribution of contexts (Bourigault & Jacquemin, 1999).

During an eruption [Action], a volcano [Agent] erupts substances [Theme] from the planet's depth [Location] under pressure and at very hot temperatures [Condition] through a natural vent in a planet's surface [Location] onto the planet's surface in some direction [Location], eventually forming some structure of solidified substances [Result].

Fig. 2. Semantic frame Volcano_eruption

The core roles of this frame are Agent and Theme. As the Agent is restricted to *volcano*, the verb semantic hierarchy is based on the Theme role. Although the Theme role is generally taken by a noun or a noun phrase, it is also possible for this slot to be empty. Of all the arguments available on this position, some are common to several verbs. For example:

(1) Volcano extrudes lava/basalt
Volcano erupts lava/basalt

At the same time, even though some arguments differ (2) they can still be switched without any important change in meaning (3):

(2) Volcano erupts magma
Volcano extrudes lava
(3) Volcano erupts lava
Volcano extrudes magma

Nevertheless, not all arguments can be switched.

(4) *Volcano erupts light/infrared rays
*Volcano extrudes gas/steam

Consequently, the Theme elements can be classified based on syntactic, pragmatic and semantic features (Faber & Mairal Usón, 1999). All the entities designated by noun phrases in the syntactic frame are concrete, inanimate and natural, but their physical state is different. They can therefore be divided into semantic subclasses (Fig. 3). This distribution is illustrated in Figure 4.

Semantic sub-class	Tag	Examples
-	-	A volcano erupts
Solid substance	SS	basalt, rock
Gas substance	GS	steam, gas
Liquid substance	LS	lava, magma
Physical phenomenon	PP	light, infrared rays

Fig. 3. Semantic groups of noun [Theme]

Verb	Noun Theme				
	-	SS	GS	LS	PP
Spread	-	+	+	+	-
Erupt	+	+	+	+	-
Eject	-	+	+	+	-
Emit	-	-	+	+ ?	+
Expel	-	-	+	+	-
Extrude	-	+	-	+ ?	-

Fig. 4. Verbs' arguments groups' distribution [Theme]

4. Results

The verb argument patterns shows that some verbs accept more types of semantic subclasses than others. This implies that the meaning of these verbs is more general:

> The full set of syntactic and semantic information is distributed throughout the domain in terms of inheritance mechanisms. Predicates on more specific levels of the hierarchy (and thus more constrained in their semantic scope) have meanings which pinpoint very specific areas of meaning, and this reduction in semantic scope brings with it a corresponding reduction in syntactic potential (Faber & Mairal Usón, 1999).

All this information makes it possible to build a verb hierarchy typical of the semantic frame of VOLCANO_ERUPTION. This frame combines general language definitions with the semantic frame and typical semantic configurations of verb arguments (Fig. 5).

This kind of structure not only portrays the semantic hierarchy but also the semantic subclasses in the verb definition that could be used for a domain dictionary or a terminological database. It can be also applied to domain-specific ontology building or unsupervised document clustering.

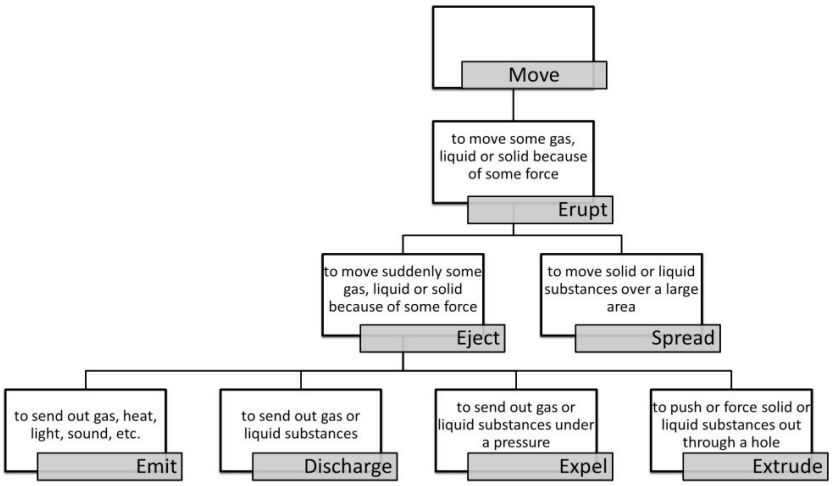

Fig. 5. Verb macro structure and definitions

5. Conclusion and perspective

The analysis described in this article illustrates the steps that should be followed to automatically obtain the semantic hierarchy and argument patterns within a specialized subdomain. The protocol has two advantages. On the one hand, it is a first step towards building a corpus-based ontology of concepts coupled with their definitions. On the other hand, it reveals the potential of verbs to combine in argument patterns of a certain semantic type.

Such a tool has a practical application both in Terminology and in Natural Language Processing. It can be used as an independent module, as well as an automatic terminology extraction tool. In the first case, it can be applied to domain-specific ontology building and to the specification of verb entries. In the second case, the module helps with the selection of verbs in specialized texts and the automatic recognition and organization of terms. Future research will focus on the protocol-based tool development.

Acknowledgements

This research has been carried out within the framework of the project RECORD: Representación del Conocimiento en Redes Dinámicas [Knowledge Representation in Dynamic Networks, FFI2011-22397], funded by the Spanish Ministry for Science and Innovation.

We are grateful to the Faculty of Languages and Applied Human Sciences of the Strasbourg University for the possibility to present our work at the V International Conference on Corpus Linguistics (CILC2013).

References

Anthony, L. (2005). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Professional Communication Conference, 2005. IPCC 2005. Proceedings. International* (pp. 729–737).

Bourigault, D., & Jacquemin, C. (1999). Term extraction+ term clustering: An integrated platform for computer-aided terminology. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics* (pp. 15–22). Retrieved from http://dl.acm.org/citation.cfm?id=977039

Faber, P. (2002). Terminographic definition and concept representation. *Training the Language Services Provider for the New Millennium. Oporto (Portugal): University of Oporto*, 343–354.

Faber, P. (2009). The cognitive shift in terminology and specialized translation. *MonTI. Monografías de Traducción e Interpretación*, (1), 107–134.

Faber, P. (ed.) (2012). *A Cognitive Linguistics Vuew of Terminology and Specialized Languag*e. Berlin: Mouton de Gruyter.

Faber, P., Linares, C. M., & Exposito, M. V. (2005). Framing Terminology: A Process Oriented Approach. Retrieved from http://www.erudit.org/livre/meta/2005/000255co.pdf

Faber, P., & Mairal Usón, R. (1999). *Constructing a lexicon of English verbs*. Berlin; New York: Mouton de Gruyter.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in cognitive sciences*, *7*(5), 219–224.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). ITRI-04-08 The Sketch Engine. *Information Technology*, *105*, 116.

Petruck, M. R. L. (1996). Frame semantics. *Handbook of pragmatics*, 1–13.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Presented at the International Conference on New Methods in Language Proceeding, Manchester, UK.