AUTOMATIC CORPUS-BASED TRANSLATION OF A SPANISH FRAMENET MEDICAL GLOSSARY

Editorial Universidad de Sevilla

Calidad en Edición Académica
Academic Publishing Quality

Avalado por
FECYT
ANECA
Promovido por
une

Colección Lingüística

Mario Crespo Miguel

AUTOMATIC CORPUS-BASED TRANSLATION OF A SPANISH FRAMENET MEDICAL GLOSSARY

Editorial Universidad de Sevilla

Sevilla 2021

Colección Lingüística
Núm.: 65

Esta publicación, que ha contado con el apoyo y financiación del Contrato Programa 2020 del Departamento de Filología de la Universidad de Cádiz, se enmarca en los proyectos «Comunicación especializada y terminografía: usos terminológicos relacionados con los contenidos y perspectivas actuales de la semántica léxica» (Ref. FFI2014-54609-P) del Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia. Subprograma Estatal de Generación del Conocimiento (convocatoria 2014 del Ministerio de Economía y Competitividad) y «Bases metodológicas y recursos digitales para la creación de un léxico relacional de usos terminológicos de la semántica léxica (TerLexNet)», solicitado en la Convocatoria 2020 de Proyectos de I+D+i del Ministerio de Ciencia e Innovación. Igualmente se inscribe en los proyectos «Lingüística y nuevas tecnologías de la información: la creación de un repositorio electrónico de documentación lingüística» (Ref. FEDER-UCA18-107788), perteneciente a los Proyectos de I+D+i del Programa Operativo FEDER Andalucía 2014-2020, y «Lingüística y Humanidades Digitales: base de datos relacional de documentación lingüística» (Ref. PY18-FR-2511) de la Convocatoria 2018 de Ayudas a proyectos I+D+i (Modalidad «Frontera Consolidado») en el ámbito del Plan Andaluz de Investigación, Desarrollo e Innovación (Junta de Andalucía, PAIDI 2020).

© EDITORIAL UNIVERSIDAD DE SEVILLA 2021
Porvenir, 27 - 41013 Sevilla.
Tlfs.: 954 487 447; 954 487 451; Fax: 954 487 443
Correo electrónico: secpub4@us.es
Web: <https://www.publius.us.es>

Impreso en papel ecológico
Impreso en España – Printed in Spain

ISBN: 978-84-472-3027-3
Depósito Legal: SE 1510-2021
Maquetación: sputnix diseño editorial – www.sputnix.es
Impresión: Podiprint

Índice

Tables

Figures

Chapter 1
Introduction

In this chapter, we will briefly give an overview of this book. Our task, the Automatic Corpus-based Induction of a Spanish FrameNet Medical Glossary, is motivated by the demand for more linguistic resources for the study of languages and the improvement of those already existing. The main contributions include 1) the development of a reliable way of selecting the FrameNet frames that conceptually reflect the domain of medicine, 2) the accurate translation of such frame selections into Spanish, 3) the improvement of the coverage of these frames with new triggers and semantic relations provided by EuroWordNet.

1.1. Motivation

The purpose of language is communication. Human beings use languages as a means of interacting with one another and to talk about the world. Languages are therefore meaningful and expressive instruments. Linguistic analysis and study must always bear in mind that languages would not exist beyond the goal of communication. However, languages differ from one another. They can be considered as different versions of the same "instrument of communication" (Dummet, 1991). The organization and structure is completely different from one language to another, but they all have the same goal.

Broadly speaking, the act of communicating is divided into a structural part (words, syntactic structures, pronunciation, etc.) and a semantic part (the meaning conveyed). Likewise, linguistic study consists of different areas: whereas phonological, morphological and syntactic analyses focus on formal parts of languages, semantics deals with the analysis of the meaning expressed by that formal part. A complete analysis of a language must include all these areas and the idea that everything in languages exists for the sake of communication. In natural language processing (henceforth, NLP), the number of projects dealing with semantics is smaller than projects related to other fields like morphology or syntax. The reason for such a shortage in computational linguistics can be attributable to the fact that meaning cannot be observed directly.

Semantics, in relation to morphology or syntax, is a more difficult linguistic level to formalize and, hence, be automatically processed.

In this work, we will review some of the most important projects devoted to describing the semantic properties of languages and we will determine if the information they provide can be projected and translated cross-linguistically. These projects include FrameNet (Baker *et al.,* 1998), PropBank (Palmer *et al.,* 2005), WordNet (Miller *et al.,* 1993) or VerbNet (Kipper *et al.,* 2000). Finally, we will transfer part of FrameNet into Spanish. In particular, we will focus on the FrameNet frames most closely related to the domain of medicine.

1.2. Goals

Most of the semantically-oriented projects have been created for English, mainly because most modern approaches to computational lexical semantics emerged in the United States. This situation is changing over time and some of these projects have been subsequently extended to other languages; however, in all cases, much time and effort need to be invested in creating such resources. Because of this, one of the main purposes of this work is to investigate the possibility of extending these resources to other languages like Spanish. As we will see, the special structure of FrameNet offers an opportunity to create similar resources for other languages. FrameNet aims to explain how languages account for daily situations linguistically.

Focusing on the frames that best represent the domain of medicine, we present a statistical method which, assisted by the word associations proposed by WordNet, can create a medical FrameNet selection and translation for Spanish and can also improve the trigger coverage of the English FrameNet. Results will be checked manually to evaluate the reliability of the system.

1.3. Contributions

We have developed a method of matching frame predicates with WordNet synsets by using the contextual information provided by a representative corpus (*COCA corpus*). This approach has been used to disambiguate the FrameNet triggers according to WordNet. Once the matching has been done, all the information that WordNet provides can be used both to translate the unit into other languages and, in this case, to extend the coverage of FrameNet with new units.

More than 90% of the triggers in our frame selection match one or two synsets in WordNet and 95.6% of the translated words were correct in Spanish. This approach

provides us with a reliable way to transfer FrameNet triggers to other languages. In addition to this, our medical frame selection was widened with new units by 204%. This approach could be also used to improve the range of the current English FrameNet.

1.4. Outline of this work

This book is structured as follows.

In Chapter 2, we will present some of the most important resources in Lexical Semantics. We will discuss in detail their linguistic representations applied and features selected. We will pay special attention to those resources developed for Spanish. After that, we will discuss some of the techniques for the cross-linguistic transfer of information.

In Chapter 3, we will discuss our corpus-based approach. Firstly, we will describe how different text domains select different frames. Accordingly, texts about 'communication' will select frames such as *conversation, questioning, statement*, etc. and frames such as *judgment* or *categorization* will show up in texts about 'cognition'. Our approach aims to choose the range of frames that account for the medical domain. Roughly speaking, our approach uses the information regarding word distribution from a medical corpus to select a representative set of frames and translate them into other languages. Translation is supported by EuroWordNet, the extension of the Princeton WordNet for some European languages. Finally, this chapter briefly introduces the concept of a Statistical Hypothesis Testing, a key element in this work.

In Chapter 4, we will show the different experiments conducted for medically-oriented frame selection. Frame gathering is based on the results after carrying out Statistical Hypothesis Testing on the FrameNet frame triggers. Output was compared to a manual benchmark to check the suitability of the selection.

In Chapter 5, we will describe several experiments that were carried out to match the triggers of the frames selected with synsets of EuroWordNet. Both semantic properties of words and frequency distribution are applied to attach a particular trigger to the WordNet synsets. Various experiments were conducted and, at the end, the results are evaluated.

In Chapter 6, we summarize and discuss the basic results of the present work and outline some important future directions.