Domain-Specific English-To-Spanish Translation of FrameNet

Mario Crespo Miguel[+], Paul Buitelaar*

[+] Facultad de Filosofía y Letras
Universidad de Cádiz, Spain
mario.crespo@uca.es

[*] DFKI GmbH - Language Technology Lab
Saarbrücken, Germany
paulb@dfki.de

Abstract

This paper is motivated by the demand for more linguistic resources for the study of languages and the improvement of those already existing. The first step in our work is the selection of the most significant frames in the English FrameNet according to a representative medical corpus. These frames were subsequently attached to different EuroWordNet synsets and translated into Spanish. Results show how the translation was made with high accuracy (95.9 % of correct words). In addition to that, the original English lexical units were augmented with new units by 120%

1. Introduction

Most lexical semantic resources have been created for English. This is mainly because most modern approaches to computational lexical semantics emerged in the United States. Some of these projects have been subsequently extended to other languages; however in all cases this requires a big human effort and time to create them. Transferring linguistic information automatically is therefore an attractive possibility to extend such resources to other languages.

FrameNet is a free on-line resource based on frame semantics and supported by corpus evidence (Ruppenhofer et al., 2006). It documents a range of different situations or frames and the list of lexical units that account for such frames in English. Frames are information packets about how to put across and understand information about a certain situation. Currently it covers 10195 different words in 795 different frames (approx. 14.14 per frame). The special conformation of FrameNet allows creating similar resources for other languages by maintaining the structural organization.

2. Approach

If the list of words of a certain language is limited, then the frames that are supported by them must be limited, too. However, the number of topics that human beings can talk about is unlimited; therefore frames combine with each other in our daily speech to convey information: *medicine*, *politics*, *family*, etc.

In addition to that, Frames can be described in terms of a variety of EuroWordNet concepts. *EuroWordNet* (Vossen, 1998) is a multilingual database like WordNet for several European languages *(Dutch, Italian, Spanish, German, French, Czech and Estonian). W*ords are divided into nouns, verbs, adjectives and adverbs and sorted into groups representing *concepts* (called synsets). Each WordNet represents a unique language-internal system of lexicalizations linked on the Princeton WordNet used as Inter-Lingual index. This allows for the use of the database in multilingual information retrieval. Table 1 shows the synsets attached to unit *cold.n* of the FrameNet frame *Medical_condictions*:

SYNSET ID	SYNSET IN ENGLISH	SYNSET IN SPANISH
10174608n	*cold.n* *common_cold.n* *respiratory_desease*	*resfriado* *resfriado_ común*
04422784n	*cold.n* *coldness.n*	*frialdad* *frío*
03916773n	*cold.n* *coldness.n* *low_temperature.n*	*frío* *temperatura_ baja*

Table 1: Lexical unit *cold.n* in EuroWordNet and synset equivalences in Spanish

From among them, *10174608n* represents a concept depicted by the situation *Medical_condictions*. Assuming that every FrameNet trigger could be attached to a certain synset, equivalences could be used to obtain the triggers that support such a frame in a target language.

According to Yarowsky (1995), the meaning of words in a specific text is consistent and uniform, that is, polysemous word usually reflect only one sense in a certain document. Moreover, words semantically close to the general subject of the document have a significant distribution. Our approach takes advantage of this by selecting the frames and synsets that co-occur in a particular document and interconnecting them.

3. Data Set

The exploitation of a representative medical corpus allowed us to study the real use of language in this

domain. A corpus of around 7 million tokens and 90.000 different lemmas was obtained from several on-line resources:

medlineplus.gov, a website on health information from the National Library of Medicine in the US, the world's largest medical library

familydoctor.org, a web site about health information that is operated by the American Academy of Family Physicians, a US-based medical organization representing family physicians and medical students

www.umm.edu, the website of the University of Maryland Medical Center, including general information about diseases and treatments

All the XML and HTML tags were removed and the resulting text was analyzed with TreeTagger[1]. Once the corpus was annotated with part-of-speech, lemma frequencies were counted, which provided us with information on the distribution of lexical triggers for FrameNet frames in the medical corpus.

4. Processing

4.1. Frame Selection

The selection of medical-oriented frames was conducted over t-test. It was tested for each set of frame triggers if the distribution that they have in our medical corpus was compatible with the distribution they have in the *British National Corpus* (Burnard, 2007). Only triggers occurring in the medical corpus were computed and frames with only one element like **Studying:** *study.v* or **Try_defendant:** *try.v* were taken out. Every frame group was checked at 99.5 percent significance level and the ones statistically significant were chosen. A set of 35 different frames was selected.

4.2. Lexical Trigger Disambiguation

There were 35 different frames and 881 triggers to be disambiguated. 79 of these triggers (8.9%) were not present in EuroWordNet. The process of disambiguation was similar to the selection of frames. Firstly, each trigger was attached to all synsets in which it shows up and subsequently every synset was tested over a Statistical Hypothesis Testing. The most appropriate synset must be statiscally significant (medical frames must be related with medical synsets). If the synset was composed of one term we used chi-square and if more than one, we used T-test, both at 99.5 percent significance level and by using the *British National Corpus* as reference corpus. Synsets attached to a trigger that were not statiscally significant over the Statistical Hypothesis Testing were detached from the trigger. If none of the synsets were statiscally significant, we keep all of them matched to the trigger. On the other hand, if all of them were statiscally significant, we kept all of them attached, too. A certain word had been completely disambiguated if we got an only synset attached to the trigger.

EuroWordNet lexical information was used to extend the number of terms in the synset tested and to improve results. Figure 2 shows the procedure:

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

Trigger	Senses	hypernyms	Words used in T-test
alleviate.v	***00361385v*** *alleviate.v ease.v*	***01737017v*** *aid.v help.v*	*alleviate.v ease.v aid.v help.v* ***00361385v***
	00044854v *alleviate.v palliate.v*	***00140937v*** *ameliorate.v amend.v*	*alleviate.v palliate.v ameliorate.v amend.v,* ***00044854v***

Figure 2: Illustration of how to include lexical relations

We detected that results can be affected by the fact that in a certain medical corpus different topics co-occur, and therefore, the system would select more synsets than the medical ones as significant. The corpus was then split into twelve different sub corpora to check if synsets were consistent in different medical texts. Table 2 shows the different experiments conducted.

	EuroWordNet Senses for Lexical Triggers			
	1	**2**	**3**	**+3**
Initial state: no disambiguation	327 (37.1%)	152 (17.2%)	111 (12.5%)	212 (24%)
Experiment A	666 (75.5%)	52 (5.9%)	24 (2.7%)	17 (1.9%)
Experiment B	672 (76.2%)	51 (5.7%)	23 (2.6%)	13 (1.4%)
Experiment C	669 (75.9%)	52 (5.9%)	26 (2.9%)	12 (1.3%)
Experiment D	704 (79.9%)	37 (4.1%)	11 (1.2%)	7 (0.7%)
Experiment E	723 (82%)	28 (3.1%)	6 (0.6%)	2 (0.2%)

Table 2: Experiments on disambiguation

Experiment A. In this first case, disambiguation was carried out by computing t-test over the term frequencies of each synset attached to each trigger in our selection of frames.

Experiment B. In this case, experiment A was extended by adding the term frequency from immediate hyponym by using lexical-semantic relations provided by EuroWordNet.

Experiment C. As before, experiment A is extended with new terms. In this case, t-test is computed with the lexical units of the synsets and those from the immediate hypernym.

Experiment D. Disambiguation is conducted in two steps. Firstly we applied the procedure followed in experiment B. Secondly, we applied our procedure in experiment C on those triggers not disambiguated so far.

Experiment E. For experiment E the system was told to choose the synset with the best average in t-test among the triggers not disambiguated so far. T-test result had been summed up for all sub corpora and stored. 82% of the

triggers of our frame selection could be matched with a synset in WordNet.

5. Evaluation

Synsets attached to the frame were translated into Spanish and it was evaluated if these words represented the situation depicted by the frame. From the 1109 words obtained in Spanish, 95.9% of them were correct as shown in Table 3. Precision can be evaluated according to the number of lexical triggers correctly matched to WordNet synsets (95.9%) and recall according to the number of FrameNet triggers attached to WordNet (82%).

Table 3 describes the precision of the Spanish translation. The first column lists the 35 frames used. The second column shows the number of lexical units translated into Spanish for each frame. The third column displays how many of them were correct in Spanish according to the sense expressed by such a frame.

FRAMES	Spanish Lexical Triggers	
	Selected	Correct
Active_substance	16	15
Being_named	26	26
Being_obligated	7	7
Biological_urge	65	65
Body_mark	33	33
Cause_change_consist	15	15
Communicate_categor	19	19
Cure	37	34
Deny_permission	4	4
Documents	56	51
Duplication	10	10
Duration	15	15
Excreting	25	23
Food	90	89
Grinding	21	21
Health_response	10	10
Institutionalization	6	4
Intoxicants	79	63
Kidnapping	7	7
Likelihood	17	17
Locale_by_use	61	61
Medical_conditions	84	84
Medical_instruments	18	18
Medical_professionals	54	54
Medical_specialties	24	24
Observable_bodyparts	108	102
Ordinal_numbers	11	11
People_by_age	24	19
People_by_origin	3	3
Perception_body	27	25
Placing	108	106
Redirecting	3	3
Scarcity	12	12
Toxic_substance	13	13
Trust	11	11

Table 3. Spanish lexical triggers translated by use of EuroWordNet

We also translated the synsets selected in English to see how many of the original triggers were included, how many additional triggers can be found and the precision of such a translation into English. This is shown in Table 4 and Table 5. In our original selection there were 881 words, but the number was increased to 1942. 731 (87.2%) of the original triggers are represented in the current version (the upper bound was located at 91%). Most of the original triggers are included. FrameNet triggers were augmented by 120%.

FRAMES	English Lexical Triggers		
	Original	Select	Overlap
Active_substance	14	23	11
Being_named	36	50	21
Being_obligated	18	16	9
Biological_urge	43	136	37
Body_mark	18	34	17
Cause_change_consist	11	21	11
Communicate_categor	14	40	14
Cure	26	57	24
Deny_permission	2	11	2
Documents	31	55	27
Duplication	12	27	12
Duration	13	44	13
Excreting	20	170	20
Food	66	170	63
Grinding	12	48	12
Health_response	9	16	9
Institutionalization	10	20	6
Intoxicants	36	111	32
Kidnapping	12	23	11
Likelihood	22	26	16
Locale_by_use	30	101	29
Medical_conditions	72	145	66
Medical_instruments	19	21	15
Medical_professionals	37	62	30
Medical_specialties	29	44	26
Observable_bodyparts	88	196	82
Ordinal_numbers	10	18	10
People_by_age	13	39	13
People_by_origin	13	4	1
Perception_body	13	28	11
Placing	68	135	61
Redirecting	3	5	3
Scarcity	5	12	5
Toxic_substance	6	18	6
Trust	7	16	6

Table 4. Triggers in the original FrameNet and after attaching new terms from EuroWordNet

Table 4 describes the overlap between original triggers and the ones from our translation. The table's first column lists our selection of frames. The second column shows the original number of triggers in FrameNet. The third column provides the number of triggers after translation. Finally, the overlap is displayed between the original FrameNet triggers and the ones translated by using EuroWordNet.

From the 1784 words obtained again in English, 91.7% of them were correct as shown in Table 5, which describes the precision of the English triggers that have been created by using EuroWordNet synset information. The table's first column lists our selection of frames. The second column shows the number of triggers after disambiguation and the third column how many of them were correct in English according to the sense of the frame.

FRAMES	English Lexical Triggers	
	Selected	Correct
Active_substance	23	18
Being_named	50	50
Being_obligated	16	16
Biological_urge	136	130
Body_mark	34	31
Cause_change_consist	21	20
Communicate_categor	40	37
Cure	57	51
Deny_permission	11	11
Documents	55	50
Duplication	27	27
Duration	44	41
Excreting	170	152
Food	170	143
Grinding	48	45
Health_response	16	16
Institutionalization	20	15
Intoxicants	111	99
Kidnapping	23	22
Likelihood	26	26
Locale_by_use	101	95
Medical_conditions	145	141
Medical_instruments	21	21
Medical_professionals	62	62
Medical_specialties	44	44
Observable_bodyparts	196	176
Ordinal_numbers	18	17
People_by_age	39	37
People_by_origin	4	4
Perception_body	28	26
Placing	135	116
Redirecting	5	5
Scarcity	12	12
Toxic_substance	18	15
Trust	16	16

Tabla 5. Number of correctly back-translated lexical triggers in English

6. Conclusions and Future Work

EuroWordNet has been the key element in all of the tasks in our approach:

- In the frame selection by extending each frame group of triggers with new terms.
- In the matching of triggers with synsets by augmenting the range of concepts with new words.
- In the translation of the English FrameNet triggers into Spanish. It provides translational equivalences among languages.

This approach not only provides us with a reliable way to transfer FrameNet triggers to other languages, but also with the match from FrameNet to (Euro)WordNet. WordNet provides a rich list of semantic relations (synonyms, hypernym, hyponym, roles, etc.) along with information about thematic roles that could be added to FrameNet. FrameNet frames coverage could be augmented by using the synonyms that WordNet provides.

We are planning to extend the translation to all FrameNet frames. Since our approach is corpus-based, the matching from triggers to WordNet synsets will be supported by texts with different topics. We hope this provide us with an accurate translation of all FrameNet English lexical triggers into other languages.

7. Acknowledgements

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors.

8. References

Reference Guide for the British National Corpus (XML Edition) edited by Lou Burnard, February 2007. http://www.natcorp.ox.ac.uk/XMLedition/URG/

Miller, George A., Christiane Fellbaum, Katherine J. Miller. 1993. *Five Papers on WordNet*.

Ruppenhofer J., Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Jan Scheffczyk.2006., *FrameNet II: Extended Theory and Practice*.

Vossen, P. (ed.) .1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Dordrecht: Kluwer Academic Publishers.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods, *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, p.189-196, June 26-30, Cambridge, Massachusetts.