Analysis of Definitions of Verbs in an Explanatory Dictionary for Automatic Extraction of Actants based on Detection of Patterns

Noé Alejandro Castro-Sánchez and Grigori Sidorov

Natural Language and Text Processing Laboratory,
Center for Research in Computer Science (CIC),
National Polytechnic (Technical) Institute (IPN),
Av. Juan Dios Batiz, s/n, Zacatenco, 07738, Mexico City,
Mexico
noe.acastro@gmail.com, sidorov@cic.ipn.mx

Abstract. Due to the importance that verbs have in language an identification of their actants (obligatory complements) is important for understanding of the meaning of sentences. Usually, the solution of this problem in natural language processing is based on machine learning approaches, which are trained on large sets of tagged texts. We show that it is possible to work with other kind of sources, i.e., explanatory dictionaries. Dictionary definitions have patterns that provide enough information for identifying actants. We develop a heuristic approach in order to obtain this information and developed an algorithm for detection of actants in texts.

Keywords: actants extraction, definition pattern, explanatory dictionary.

1 Introduction

The verb is considered as a linguistic center of a sentence. It expresses states, actions or processes that involve some other entities. These entities are known as *actants*, a term proposed by L. Tesnière in 1959. This term was motivated by an analogy with the theatre play, containing the action (verb), the actors (actants) and the decoration (circumstances). The actants are defined as obligatory expressions demanded by verbs: their absence would produce ungrammatical sentences. The circumstances can be omitted without harming a sentence being their principal function to broaden its meaning [6].

The correct identification of these entities is crucial for various linguistic theories and is widely used in various tasks of natural language processing. Usually they are extracted from a corpus using the concept of subcategorization frames on the basis of their grammar category, number of actants, and their syntactic position. Majority of these methods are developed for English ([1, 4, 5, 10, 17]), also there are works related to Spanish ([14]), Hungarian ([16]), Czech ([15]), Bulgarian ([13]), Italian ([9]), etc.

Other data sources also have been used for this task, such as the Web ([10, 18]), multilingual resources and texts ([1]), bilingual dictionaries, as in [5], where the dictionaries are used to extend the number of entries in a dictionary of valences, assuming that verbs with similar translations should have the same valence structure.

In this work, we use a large explanatory dictionary of Spanish to identify the actants required by verbs. We assume that definitions of verbs in dictionaries provide enough information to identify the actants, also providing some selectional restrictions that would help to identify them in sentences.

2 Patterns in Definitions

We used for our experiments the dictionary of Spanish Royal Academy which contains 162,362 definitions (senses) grouped in 89,799 lexical entries. From these, 12,008 lexical entries correspond to verbs, which contain 27,668 definitions (senses). The distribution of number of definitions (senses) among verbs corresponds to the well-known Zipf law, where there are few verbs with many senses, while the majority of verbs have only one sense.

A typical dictionary definition presents two terms: *genus* or hyperonym, and *differentia*, which shows the characteristics that distinguish the lexical entry from other items grouped within the same genus. This last term contains some elements named *pattern in definition*, which are not strictly related with the semantic content of the lexical unit, but are included to help the correct use of the defined term, providing some contextual restrictions ([5, 19]). We use this information to identify the verb actants. This can be appreciated carrying out the exchange task, a process that allows extraction of a pattern from the definition. For example:

Sostener: Sustentar, mantener firme algo
(Hold: To take something firmly).

Suppose that we have the following sentence:

Juan sostiene los libros
(John holds the books).

The exchange is carried out by replacing the defined term by its definition, as it is shown in the following sentence:

Juan sustenta, mantiene firme los libros
(John takes the books firmly).

We can see in the last sentence, that it was necessary to remove the element "*algo*" ("*something*") contained in the definition, because it is replaced by the noun phrase "*los libros* " ("*the books*"). This situation is called *pattern in definition*, and it exposes the actant that is required by the verb *sostener* (*to hold*).

3 Processing of Definitions

Verb meanings are extracted from the dictionary and grammatically tagged with the FreeLing parser, an open source text analysis tool for various languages including Spanish [11].

We carry out processing of definitions based on the following algorithm, which we will explain in detail in the following sections:

- Identify the patterns in definitions.
- Extract the verbs that have any pattern in their first sense.
- Identify the senses of other verbs where the verbs obtained at the previous step are genus of the definition.

3.1 Extraction of Patterns from Definitions

First, the patterns in definition are identified. In order to do that we consider that nouns and indefinite pronouns are the only parts of speech that can represent patterns. We measure their frequency in definitions, see Table 1.

Table 1. Frequencies of the usage of nouns and indefinite pronouns in definitions.

Word	Frequency
Algo (something)	3,462
Alguien (somebody)	2,154
Cosa (thing)	1,594
Persona (person)	1,175
Lugar (place)	408
Cuerpo (body, section)	349
Animal (animal)	300
Agua, acción, tierra, fuerza, … (water, action, land, force, …)	< 300

It can be seen that the most frequent elements represent very general classes of words. For example, *algo* (*something*) can be matched with any word that refers to any non-animated thing, *alguien* (*someone*) matches with any word that refers to a person, *cosa* (*thing*) behaves in the same way as *algo* (*something*), etc. For our experiment, we select from the list above eight most frequently used elements; let us name them *general patterns*. It can be seen that some of these elements are synonyms, i.e., in the same context it is possible to exchange them, say, to change *alguien* (*somebody*) with *persona* (*person*), or *algo* (*something*) with *cosa (thing)*, etc.

In the majority of cases the patterns observed in the definition do not correspond to the complete list of valences (subcategorization frame). For example,

Conducir: Llevar, transportar de una parte a otra.
(Drive: move, transport from one part to another).

This definition does not contain information about the direct object that is necessary. Our preliminary evaluation is that more than 80% of definitions are incomplete in this sense. Nevertheless, each definition contains information about some valences and the proposed algorithm allows their extraction by making substitutions. Again, our preliminary evaluation is that in more than 60% of cases the absent actants are restored using this methodology.

Considering that usually the first sense of a lexical entry is the most frequently used, we extracted verbs that have general patterns in their first sense. The number of these verbs in the dictionary is 2,748.

3.2 Processing of Genus

Almost all definitions included in the dictionary follow the typical formula represented by *genus + differentia* (see Section 2). The predictable position of these elements allowed us to identify them in an automatic way.

From 12,008 verbs that we considered, 3,751 are used as genus. Usually, verbs that have general patterns in their first sense are used as genus. In the following table we show the frequency of the five most frequent verbs.

Table 2. Verbs with highest frequency in definitions as genus.

Verb	Frequency
Hacer (to do)	1,462
Dar (to give)	969
Poner (to put)	836
Quitar (to remove)	503
Echar (to throw)	271

We got 5,987 definitions that have a genus that is part of the verbs that have general pattern in their first sense (see above).

After this processing we got the information of actants of dictionary verbs that have patterns.

4 Mapping between Patterns and Actants in Sentences

After this processing we detected patterns in definitions of verbs. The next step is to apply them for identifying the actants in the input sentences. One manner of identifying actants is usage of lexical relations (hyperonim/hiponym) that are detected

using ontology like WordNet, for example. However, we considered that the same dictionary used at the first stage can also help us to carry out this task.

The mapping is based on the following algorithm. We maintain the list of the processed words for avoiding possible circles.

1. Identify nouns and indefinite pronouns in sentences. They are the candidate actants.
2. Start with one of the candidates getting its meaning from the dictionary. Clear the list of the processed words.
3. Identify the genus of the definition.
4. Compare this genus with the verb patterns possibly present in the sentence.
   4.1. If genus is different from all patterns, verify that it is not present in the list of the already processed genus and go to step 2, taking the new genus as the candidate for replacing and add the new genus to the list of the already processed words. If it is already processed then go to step 5.
   4.2. Otherwise mark the found actant and proceed to step 5.
5. Determine if there are more candidates to process.
   5.1. If that is the case, go to step 1, with the next candidate.
   5.2. Otherwise finish.

Let us apply the algorithm to the following sentence:

María regaló una trenca a su esposo
(Mary gave a duffle-coat to her husband).

In this example, we find as candidates the words "*trenca*" ("*duffle-coat*") and "*esposo*" ("*husband*"). The next step is to get the meaning of the verb:

Regalar: Dar a alguien, sin recibir nada a cambio, algo…
(Give a present: Give something to someone, without expecting compensation …).

The patterns in this definition are "*alguien*" ("*someone*") and "*algo*" ("*something*"). We execute the previously presented algorithm for the first candidate:

1. ***Trenca (duffle-coat)***. *Abrigo corto, con capucha y piezas alargadas*
   (a coat which has toggle buttons and usually has a hood).
2. Genus: *Abrigo* (*coat*).
3. Compare genus with patterns: *abrigo = alguien* or *abrigo = algo*
   (coat = someone or *coat = something)*
   3.1. Words are different. Go to step 1 of the algorithm with genus "*abrigo*" ("*coat*").

Now we repeat the algorithm with the new word.

1. ***Abrigo (coat)***. *Cosa que abriga (something to wrap up).*
2. Genus: *Cosa* (*something*).
3. Compare genus with pattern *cosa = alguien* or *cosa = algo*
   (*something = someone* or *something = something*)?
   3.1. *Cosa* (*something*) is equal to *algo* (*something*) (see sub-section 3.1). We found the actant here.
4. Go to step 1 and continue with other candidates.

In this example we found the actant at the second iteration.

5 Conclusions

The aim of this work was to explore new resources and methodologies for automatic identification of the actants of verbs. Traditional approaches are based on the usage of texts corpora as training sources for automatic learning systems. Our proposal is based on the recursive analysis of dictionary definitions.

More or less regular structure that lexicographers use while developing definitions allows applying of simple heuristics for identifying the definitions terms and analyzing their structure. A typical definition of a verb contains the actants that can be iteratively reduced to words with very abstract meaning providing semantic restrictions that can be used for identifying the actants in sentences (like *something, someone, place*, etc.). We found that definitions of many verbs in Spanish can be transformed into definitions with abstract words that help in identifying their actants. Our preliminary evaluation is that at least in 70% of cases the absent actants are restored using this methodology.

We showed that the proposed method can be applied for detection of actants in texts.

As a future work we can mention more extended experiments and evaluation, analysis if verbs with the same genus keep the same number and type of actants.

Acknowledgements. Work done under partial support of Mexican Government (CONACYT projects 50206-H and 83270, SNI) and National Polytechnic Institute, Mexico (projects SIP 20080787, 20091587, 20090772, 20100773, 20100668; COFAA, PIFI).

References

1. Alcina, A., Valero, E.: Análisis de las definiciones del diccionario cerámico científico-práctico, sugerencias para la elaboración de patrones de definición. Debate Terminológico, 4 (2008)
2. Aone, Ch., MacKee, D.: Acquiring Predicate-Argument Mapping Information from Multilingual Texts. In: Corpus processing for lexical acquisition, pp. 191--202. MIT Press, Cambridge , MA (1996)
3. Brent, M.: Automatic acquisition of subcategorization frames from untagged text. In: 29th Annual Meeting of the Association for Computational Linguistics, pp. 209--214. Berkeley, CA (1991)
4. Brent, M.: From grammar to lexicon: unsupervised learning of lexical syntax. Computational Linguistics 19.3. 243--262 (1993)
5. Cordero, S.: Diccionario De la Lengua Española. Secundaria (Diles): Planta para su elaboración con algunos apuntes básicos de metalexicografía. Káñina, Revista Artes y Letras, XXXI. 167--195 (2007)

6. Frías, X.: Introducción a la semántica de la oración del español. Revista Philologica Romanica (2001)
7. Fujita, S., Bond, F.: An Automatic Method of Creating Valency Entries using Plain Bilingual Dictionaries.
8. Gahl, S.: Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, pp. 428--432. Association for Computational Linguistics, Morristown, New York (1998)
9. Ienco, D., Villata, S., Bosco, C.: Automatic Extraction of Subcategorization Frames for Italian. International Conference on Language Resources and Evaluation IREC (2008)
10. Kawahara, D. Kurohashi, S.: Case frame compilation from the web using high-performance computing. In: International Conference on Language Resources and Evaluation. 1344--1347 (2006)
11. Atserias, J., Casas B., Comelles, E., Gonzáles, M., Padró, L., and Padró, M. FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library. In: Fifth international conference on Language Resources and Evaluation, Genoa, Italy (2006) http://www.lsi.upc.edu/ nlp/freeling.
12. Manning, C.: Automatic acquisition of a large subcategorization dictionary from corpora. In: 31st Annual Meeting of the Association for Computational Linguistics, pp. 235--242. Association for Computational Linguistics, Columbus, Ohio (1993)
13. Marinov, S., Hamming, C.: Automatic Extraction of Subcategorization Frames from the Bulgarian Tree Bank (2004)
14. Monedero, J., et al.: Obtención automática de marcos de subcategorización verbal a partir de texto etiquetado: el sistema SOAMAS. Procesamiento del lenguaje natural. 17. 241--254 (1995)
15. Sarkar, A., Zeman, D.: Automatic Extraction of Subcategorization Frames for Czech.
16. Séreny, A., Simon, E., Babarczy, A.: Automatic Acquisition of Hungarian Subcategorization Frames. In: 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, Budapest, Hungary (2008)
17. Ushioda, A., Evans, D., Gibson, T., and Waibel, A.: The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In: Workshop on the Acquisition of Lexical Knowledge from Text, pp. 95--106. Association for Computational Linguistics Morristown, New York (1993)
18. Usun, E., et al.: Web-based Acquisition of Subcategorization Frames for Turkish. In: 9th International Conference on Artificial Intelligence and Soft Computing. IEEE Computational Intelligence Society (2008)
19. Wortjak, G.: Reflexiones acerca de construcciones verbo-nominales funcionales. Revista de Estudos Linguísticos da Universidade do Porto - Vol. 1, 257--280 (1998)