Exploiting verb similarity for event extraction

Uso de la similitud verbal para la extracción de eventos

Lara Gil-Vallejo
Universitat Oberta de Catalunya
UOC Tibidabo
lgilva@uoc.edu

Resumen: Los objetivos de esta tesis son crear una clasificación verbal de forma automática y valorar la utilidad de las clases verbales obtenidas para la tarea de extracción de eventos en texto plano, específicamente en lo relativo a la identificación y clasificación de participantes en eventos expresados por verbos. Esta tarea será modelizada como etiquetación de papeles semánticos. Para ello, nos proponemos encontrar el tipo de información lingüística más productivo para esta tarea.
Palabras clave: clasificación verbal, clustering, extracción de eventos, roles semánticos.

Abstract: The main objectives of this thesis are to build a verb classification automatically and to assess its performance for the task of event extraction from plain text, specifically in terms of identifying and classifying participants in events expressed by verbs. This task will be formulated as Semantic Role Labelling. For this purpose, we aim to find the most productive linguistic features to define these verbal classes.
Keywords: verb classification, clustering, event extraction, semantic roles.

1 Motivation of the research

The development of the Internet and the ICTs (Information and Communication Technologies) has created a new scenario, changing how we interact among us and transforming the way we obtain information and knowledge. In the present situation everybody is able to produce information and make it publicly available. However, information does not constitute knowledge by itself: it has to be interpreted and connected with a previous background in order to be considered knowledge.

Currently, there is an exponentially growing volume of information of potential interest for the user. In addition, very often the information that users and professionals require can only be found in unstructured sources, such as press releases, news, application forms, etc.

In order to fully take advantage of the vast amount of information available, documents and texts have to be structured in a way that allows the extraction of the relevant information that, in turn, can be used to acquire knowledge. Extracting information from texts usually involves the use of specific resources and tools. In many cases, for languages other than English, these resources have to be created from scratch, with the subsequent costs and time investments. Our starting point is an annotated corpus with semantic and syntactic information associated to the most frequent verbs of Spanish. Our research will be focused on the application and extensibility of this corpus for an event extraction task.

2 Related work

Semantic Role Labelling (SRL) has been argued to be easily tailored to event extraction tasks because the boundaries of the event participants have a high degree of coincidence with those of the constituents labelled by SRL systems (Surdeanu et al., 2003). Christensen, Soderland and Etzioni (2010) showed that SRL

is a robust technique that can be used to perform information extraction with heterogeneous web data, obtaining an F-measure of 69.9. SRL has been also used to extract events from Wikipedia with an F-measure of 71.2 (Exner and Nugues, 2011) and to identify and tag events and related temporal information in Llorens et al., (2010) within the standard framework for event and temporal expressions annotation established in TimeML (Pustejovsky et al., 2003)

As for verb classification, starting with the work of Levin (1993) and the creation of VerbNet (Kipper, 2005), it has received increasing attention over the past years. Verb classifications are a useful way of organizing lexical information. Handcrafted classifications have been used for a range of tasks such as word sense disambiguation (Brown, Dligach and Palmer, 2014), semantic role labelling (Swier and Stevenson. 2004), information extraction (Mizuta et al., 2006), among others. There have been many efforts to automatically create an verb classification that resembles the work of Levin (Schulte im Walde, 2004; Li and Brew 2008, among others). However, to our knowledge, none of them has been tested in a task, with the exception of the work of Sun and Korhonen (2011), which was applied to argumentative zoning (Shutova, Sun and Korhonen, 2010) and metaphor identification (Guo, Korhonen and Poibeau, 2011).

VerbNet has been used for SRL with interesting results. Swier and Stevenson (2004) obtained a F1 of 65 on the task of argument identification and labelling using VerbNet 1.5. Pradet, Chalendar and Pujol (2013) also based their system in VerbNet and showed that handling specific constructions such as the passive voice could improve the results. They obtained a F1 of 70.48 on gold arguments.

3 Research description and hypothesis.

3.1 Description of the research

The first step will be to build a verbal classification automatically. In order to do so, we will apply clustering techniques to classify the most frequent verbs occurring in a Spanish corpus, taking into account different types of information available in the corpus.

We will evaluate the performance of our automatically built verbal classification for the event extraction task. Grishman (2003) defines events as "predications that involve multiple entities and modifiers". According to the same author, event extraction consists of:

a. Identifying instances of events of a particular type.
b. Identifying the arguments of each event.

We will concentrate our efforts in task (b), namely, identifying and classifying the participants of each event (entities and modifiers such as place, time, manner, etc.) for texts in Spanish. Given the similarities of this task with SRL, we will model it as a SRL problem.

The evaluation will be carried out using the information associated to the verb entries in the classification. This information will be feed to extraction patterns that will be applied to sentences containing events described by verbs in the classification. In addition, we will explore the extent to which participants in events described in sentences headed by out-of-classification verbs can be labeled.

As potential practical applications for this project we can name the creation of a database for queries such as ReVerb (Fader et al. 2011)

3.2 Hypotheses

Our main hypothesis is that a verb classification automatically built from corpus data reveals patterns of homogeneous linguistic behaviour that are useful to identify and tag participants in an event. In addition, there are two sub-hypothesis:

- Both syntactic and semantic features are relevant in order to cluster. As for the organization of the features, we think that following the argument structure of verbs and associating features in subcategorization frames will lead to an optimal verbal classification, as opposed to using information from isolated constituents (arguments).
- The data obtained from the most frequent verbs is representative of a more general verbal behaviour. Therefore, at least some of the structures in which an out-of-classification verb participates will coincide with structures associated with verbs in the classification. A similarity criterion will enable the application of extraction patterns associated to verbs

in the classification to sentences headed by out-of-classification verbs.

4 Resources, methodology and experiments.

We will perform experiments with two corpora: the Sensem corpus (Alonso et al., 2007), that contains 25.000 sentences belonging to the news domain annotated manually at syntactic and semantic levels, and Ancora corpus (Taulé, Martí and Recasens, 2008), that contains 500.000 words annotated manually at the syntactic and semantic levels, also belonging to the news domain.

We are going to rely on the linguistic information available in these corpora for the most frequent verbs of Spanish. Specifically, we will use this information to characterize verbs. We will apply clustering algorithms to these verbs in order to create several automatic verbal classifications based on different linguistic features combinations (part of speech, syntactic function, constructions, selectional preferences, word embeddings, among others). We will evaluate the performance of these classifications in a SRL task. To do so, the most salient features which define each class will be used to create extraction patterns that will be applied to sentences containing verbs members of the classes. An evaluation will be carried out in order to assess the performance of different sets of patterns drawn from different feature sets. We will also evaluate the performance of extraction patterns built directly from corpus sentences, without clustering verbs. We will compare both approaches in order to test whether verbal classes are useful for the event extraction task.

Secondly, we will study the degree to which extraction patterns obtained from verbal classes can be used to identify and tag participants in events headed by out-of-classification verbs. To do so, we will measure the similarity between these verbs to each of the verbs present in the classification. We will use the information related to the most similar verbs present in the classification to label participants in the event.

5 Specific research questions for discussion

We would like to center the discussion of this project on the following research questions: first of all, we are concerned about the possible drawbacks of this approach regarding coverage and accuracy in the proposed task, and we would like to discuss some strategies to help overcome them. Secondly, we are currently performing experiments with hierarchical clustering algorithms, but given the amount of available clustering algorithms, we would like to address the issue of the suitability of this algorithm for the task. Finally, we would like to bring up the subject of possible applications of automatically created verb classifications to other tasks such as verb sense discrimination, implicit argument identification, event coreference, etc.

References

Alonso, L., J. A. Capilla, I. Castellón, A. Fernández-Montraveta and G. Vázquez. 2007. The sensem project: Syntactico-semantic annotation of sentences in Spanish. *Amsterdam studies in the theory and history of linguistic science series 4*, *292*: page 89.

Brown, S. W., D. Dligach and M. Palmer. 2014. VerbNet class assignment as a WSD task. In *Computing Meaning*, pages 203-216, Springer (Netherlands).

Christensen, J., S. Soderland and O. Etzioni. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 52-60, Association for Computational Linguistics, (Los Angeles).

Exner, P., and P. Nugues. 2011. Using semantic role labeling to extract events from Wikipedia. In *Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011). Workshop in conjunction with the 10th International Semantic Web Conference*, pages 23-24, (Aachen).

Fader, A., Soderland, S., & Etzioni, O. (2011, July). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1535-1545). Association for Computational Linguistics.

Grishman, R. 2003. Information extraction. In *The Handbook of Computational Linguistics and Natural Language Processing*, John Wiley & Sons, pages 515-530, Hoboken.

Guo, Y., A. Korhonen and T. Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 273-283, Association for Computational Linguistics, Stroudsburg.

Kipper Schuler, K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. Doctoral dissertation, PhD dissertation.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press, Chicago.

Li, J. and C. Brew. 2008. Which Are the Best Features for Automatic Verb Classification. In *Proceedings of the 46th Annual Meeting on Association for Computational Linguistics, pag*es 434-442, Ohio.

Llorens, H., Saquete, E., & Navarro, B. (2010, July). Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 284-291). Association for Computational Linguistics.

Mizuta, Y., A. Korhonen, T. Mullen and N. Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, *75*(6), pages 468-487.

Pradet, Q., G. De Chalendar and G. Pujol. 2013. Revisiting knowledge-based Semantic Role Labeling. In *Proceedings of the 6th Language & Technology Conference*, page 71, Poznań.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A. and Radev, D. R. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3, 28-34.

Surdeanu, M., S. Harabagiu, J. Williams and P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 8-15, Association for Computational Linguistics, Stroudsburg.

Shutova, E., L. Sun and A. Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002-1010, Association for Computational Linguistics, Stroudsburg.

Swier, R. S. and S. Stevenson. 2005. Exploiting a verb lexicon in automatic semantic role labelling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 883-890. Association for Computational Linguistics, Stroudsburg.

Schulte Im Walde, S. 2004. Automatic Induction of Semantic Classes for German Verbs. Doctoral dissertation, PhD dissertation.

Sun, L. and A. Korhonen. 2011. Hierarchical verb clustering using graph factorization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages. 1023-1033. Association for Computational Linguistics, Stroudsburg.

Taulé, M., M. A. Martí and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech.