Combining Rules and CRF Learning for Opinion Source Identification in Spanish Texts

Aiala Rosá[1,2], Dina Wonsever[1], and Jean-Luc Minel[2]

[1] Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay
{aialar,wonsever}@fing.edu.uy
[2] Université Paris Ouest Nanterre la Défense, Nanterre, France
jminel@u-paris10.fr

Abstract. In this work we present a system for the automatic annotation of opinions in Spanish texts. We focus mainly in the definition of a TFS-style model for the predicates of opinion and their arguments, in the creation of a lexicon of opinion predicates and in two additional variants for identifying the source of opinions. The original system extracts opinions and all its elements (predicate, source, topic and message) based on hand-coded rules, the first variant uses a CRF model for learning the source, assuming that the predicate is already tagged, and the second variant is a combined version, with the result of source recognition via the rule-based system being added as an additional attribute for training the CRF model. We found that this hybrid system performs better than each of the systems evaluated separately. This work involved the construction of several resources for Spanish: a lexicon of opinion predicates, a 13,000 word corpus with whole opinion annotations and a 40,000 word corpus with annotations of opinion predicates and sources.

Keywords: opinion extraction, hybrid approach, rule-based system, conditional random fields.

1 Introduction

An interesting task for various Natural Language Processing applications is the identification of the points of view or positions of different sources, generally people of public importance, about different topics. To answer questions such as *What is X´s opinion on the topic Y?*, *Who said something on the subject Y?*, *Who approves or disapproves of some issue Y?*, it is essential to be able to extract occurrences of opinions of different persons, in journalistic texts such as editorials or news articles.

There are systems (Appinions[1], EMM News Explorer[2]) that offer such services for English. These systems rely on different types of resources such as specialized lexicons and annotated corpora, besides general-purpose resources in natural language processing (i.e., taggers, lexical databases like WordNet, parsers, etc.).

[1] appinions.com

[2] emm.newsexplorer.eu

J. Pavón et al. (Eds.): IBERAMIA 2012, LNAI 7637, pp. 452–461, 2012.

In the case of Appinions, an important resource is the MPQA corpus [20], which contains annotations for different elements of an opinion (source, topic, polarity, etc.). This system also uses dictionaries of subjective words [7].

Other work on the identification of opinions is also based on the use of a specialized vocabulary: positive and negative verbs and adjectives [8], or verbs that introduce reported speech [9, 11].

For recognition of the sources, [18] defines a repertoire of source introducing predicates (SIP), where each entry has an associated semantic class and some syntactic information.

There are no specific resources for opinion extraction in Spanish so we had to rely on general purpose lexical resources. These resources proved useful as an initial basis, but they had to be adapted, and to be contrasted with examples from a corpus. This is the case of ADESSE[3], a lexical database for Spanish that provides lists of verbs belonging to different semantic classes and numerous corpus examples showing different syntactic configurations of the arguments inside each semantic class. Its adaptation to the type of text that we are interested in (journal articles) is not trivial, since many of the examples come from literary texts, where we find highly ambiguous cases, such as for instance occurrences of the verbs *greet*, *pray* or *sign* within the class communication (*"Ojalá que venga", reza. / "I hope he comes", he prays; Nunca dejé de rezar por ti. / I kept praying for you.*).

In a previous paper [15] we presented a rule-based system for the recognition of opinions and their elements (the predicate, the source, the topic and the message) in Spanish journalistic texts. In the present paper we focus primarily on opinion source identification. We report on the use of a CRF classifier for source recognition; on a combined system that includes the rule-based system output as an input attribute for training, significantly improving the results of the rule-based system and the CRF system; and on a specialized co-reference resolution module for the recovery of omitted sources. The combined system for source recognition achieves 83% of exact F-measure, this result being similar to those reported for English and Chinese [6, 11, 21].

The following resources were created and will be publicly available: a) a lexicon of 155 opinion predicates in which for every element we provide syntactic and semantic information, necessary for the recognition of different patterns for the predicates and their arguments, b) two annotated corpora: a 13,000 token corpus annotated with opinions and their elements and a 40,000 token corpus annotated with predicates and sources.

The organization of this paper is as follows. In Sect. 2 we present some works related to the source recognition task. In Section 3 we present our definition for opinion. In Sect. 4 we describe an opinion predicate lexicon. In Sect. 5 we present the automatic systems for source recognition: we first describe briefly the rule-based system and then we present the CRF classifier, the combined system and the co-reference module for opinion sources. Finally, we conclude in Sect. 6.

2 Related Work

Regarding the identification of opinions, one of the most important references is the annotation schema for opinions and emotions presented in [20]. This model specifies

[3] adesse.uvigo.es

the different kinds of expressions to be considered for the study of opinions: explicit mentions of private states (*The U.S.* ***fears*** *a spill-over*), speech events expressing private states (*"The U.S. fears a spill-over,"* ***said*** *Xirao-Nima*), expressive subjective elements (*The report is full of* ***absurdities***), and objective speech event (*Sargeant O'Leary* ***said*** *the incident took place at 2:00pm*). In our work we have considered most of these expressions, except the expressive subjective elements, so we included in our study all cases of reported speech (objective and subjective).

To our knowledge, there are not systems for opinion source identification in Spanish texts. Different works focus on source identification for English [2, 3, 5, 6, 8, 21] and Chinese [11]. Almost all these authors apply machine learning methods, only [11] has developed a rule-based system obtaining better results than most of the listed systems. Some authors [8, 21] use a semantic role tagger, this tool is considered very important for source identification by some authors who have studied this problem [17, 18]. We did not have access to this type of resource for Spanish.

In addition, there are some works on reported speech identification, the typical mechanism for citation. Both studies analyzed [9, 14] propose rule systems. In the first case, the speech verb, the source and the reported clause are identified for each reported speech instance. In the second case, only direct speech is recognized.

3 Opinion Definition

In our work, the concept of opinion covers all the expressions attributed to different sources by the author of the text, including those in which the source transmits an objective content. We identify four relevant elements for the opinion:

- the predicate: expression that indicates the presence of an opinion (verbs like *opinar/say, rechazar/reject*; nouns like *opinión/opinion, rechazo/rejection* and source indicators like *según, de acuerdo con / according to*),
- the source: opinion holder,
- the topic: explicit subject on which the opinion is expressed,
- the message: content of the opinion.

In our analysis the predicate is the central element of the opinion and the remaining elements are its arguments.

In example (1) we show the different elements of the opinion using the following notation: underlined source, predicate in bold, topic in italics and message shaded in gray.

> (1)
> Consultado *sobre la lentitud de los procesos judiciales uruguayos*, Carranza **respondió**: "Hay una situación de un muy alto número de presos sin condena, hay que agilizar los procesos".
>
> [*Consulted about the slowness of the Uruguayan judicial processes, Carranza said, "There is a situation of a very high number of unsentenced prisoners, we must speed up processes."*]

Most instances of opinions in texts do not contain all the defined elements. The topic, for instance, is not very common. The source is sometimes absent, mainly when it can be recovered from context. In Sect. 5.4 we present a module for recovering missing sources.

4 Opinion Predicate Lexicon

Automatic opinion recognition is built around the identification of an opinion-introducing predicate. A predicate lexicon is essential to identify these elements.

The lexicon we built contains 100 verbs and 55 nouns, mostly extracted from our development corpus. The rules system evaluation (presented in section 4.1) showed that the lexicon had good coverage (91 %) on the evaluation corpus.

For each lexicon entry a type is assigned, according to a model we defined for predicates, where the syntactic and semantic properties of the predicates and their arguments are described. These properties make it possible to identify the source, the topic and the message within the syntactic structures in which the predicates occur.

The model focuses on predicates, mostly verbal predicates, and the other opinion elements (source, topic and message) are arguments for these predicates. It is specified in the language of Typed Feature Structures (TFS) [13].

The type system defined for verbal predicates is based on the hierarchically organized semantic classes of ADESSE and each subtype inherits properties or restrictions specified by TSF structures. In Figure 1 we show the hierarchical organization for opinion verbs. Semantic classes in bold belong to ADESSE classification.

We have defined one semantic property (semantic orientation, with possible values positive, negative or neutral) and several syntactic-semantic properties:

- The semantic role of the grammatical subject of the opinion verb, which can take the values *source* (OV_SS) or *topic* (OV_ST), resulting in the first binary branching of the tree.
- The possibility of accepting the topic of the opinion as an object complement of the verb (OV_SS_NO_NEU: evaluation, acceptation and some sensation verbs) or of not accepting it (OV_SS_BEL_COM: belief and communication verbs). In addition, OV_SS_BEL_COM are neutral and OV_SS_NO_NEU have a semantic orientation positive (OV_SS_POS) or negative (OV_SS_NEG).
- The possibility of accepting a subordinate construction containing the opinion message (OV_SS_BEL_REP: belief and reported speech verbs) or of not accepting it (OV_SS_TALK: verbs like "talk").
- The possibility of topicalisation for the opinion topic, usually with the preposition *sobre /about* (OV_SS_BEL_REP: belief and reported speech verbs).

Our model also includes nominal predicates for which the main feature we specified is the opinion element introduced by the Spanish preposition *de*: in some cases this preposition introduces the source (*la delcaración del presidente / the president statement*), in other cases it can introduce the source or the topic (*el anuncio del presidente / the president's announcement* or *el anuncio de su llegada / the announcement of his arrival*).

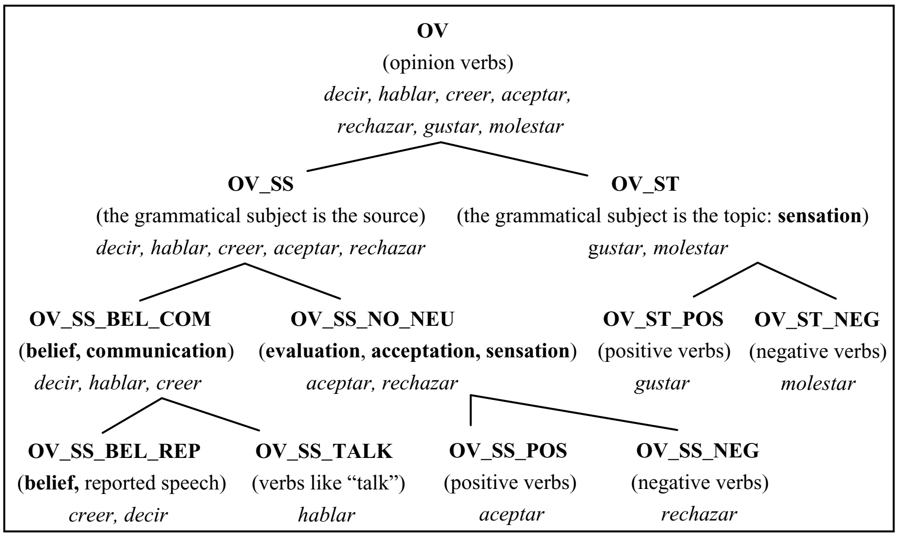

Fig. 1. Opinion Verbs: Types Hierarchy, some examples are shown for each class[4]

5 An Automatic Opinion Identification Tool

5.1 A Rule-Based System

We developed a rule-based system, based on contextual rules [22], that uses the predicate lexicon for the identification of the opinions and their elements. Example (2) shows the system output:

(2)

```
<opinion so="pos">
        <source so="neu">Mujica</source>
        <predicate so="pos">respaldó</predicate>
        <topic so="neu">importante inversión minera</topic>
</opinion>
[Mujica supports a major mining investment.]
```

We defined five rule modules: one for each opinion element (predicate, source, topic and message) and a final module for the whole opinion. For each element, except for the message, a semantic orientation (so) value (neutral, positive or negative) is assigned. Finally, the semantic orientation of the whole opinion is calculated, based on the values of the elements. A more detailed description of the rule-based system was provided in [15], [16].

[4] decir/say, hablar/talk,speak, creer/believe, aceptar/accept, rechazar/reject, molestar/annoy, gustar/please,like.

The rule system was evaluated on a 13,000 token corpus, including 300 opinion instances. Table 1 shows exact and partial results.

Table 1. Results obtained with the rule-based system

	Predicate	Source	Topic	Message	Opinion
Precision (exact)	92%	81%	67%	65%	52%
Precision (partial)	92%	93%	96%	95%	94%
Recall (exact)	91%	63%	45%	58%	42%
Recall (partial)	91%	72%	62%	84%	77%
F-measure (exact)	91.5%	71%	54%	61%	47%
F-measure (partial)	91.5%	81%	75%	89%	85%

There are several elements which are sometimes partially recognized, especially for the topic and the message which are usually longer than the predicate and source.

Results for the predicate show that the lexicon has good coverage, although there remains some place for improvement by adding new predicates (recall for predicate recognition is 91%).

5.2 Machine Learning Experimentation

In order to improve the results for source recognition, we applied a machine learning process, based on the Conditional Random Fields (CRF) model. CRF [10], a sequential discriminative probabilistic model, has proved to be successful in various applications of Natural Language Processing, such as nominal group segmentation, named entity identification and morphological tagging [19]. It has also been used in Opinion Mining for source recognition in English texts [5, 6] and to classify subjective sentences in English and Chinese texts [12].

We treat the problem of source recognition as a sequential classification problem, where we estimate the conditional probability of a sequence of output values (the class of each lexical unit) given an input sequence (observations).

We generated a manually annotated corpus of 40,000 tokens: 30,000 for training and 10,000 for testing. The training corpus contains 486 sources and the testing corpus contains 158 sources.

For training, we used morpho-syntactic attributes (word, lemma, POS-tag, number, gender) and some special attributes that indicate which verbs, nouns and prepositions belong to the opinion predicate lexicon. The corpus includes an output attribute, based on the B-I-O notation, indicating whether a word is the beginning (B) of a source or interior (I) to a source. The value O is assigned to non source words. We performed several experiments [16], varying the way of combining the attributes, the number of elements to consider before and after the current element, and the use of bigrams for output values. The best results obtained were 66% in recall and 92% in precision.

To compare the two systems, we carried out a new evaluation of the rules system, just for source recognition, assuming all predicates were recognized, and using the same test corpus that we used for the CRF system evaluation. The rules system reaches the best results for recall, 73% rules / 66% CRF, while CRF is better for precision, 85% rules / 92% CRF (exact measures).

5.3 Combining the Rule-Based System with the CRF Classifier

An additional input attribute, based on the B-I-O notation, indicates whether a word was marked as a source or as part of a source by the rule-based system. Thus we obtained our third system: a combined system that inherits the benefits of each of the systems described above, reaching good results in precision, like CRF, and in recall, like the rule-based system. It is even one or two points higher on each measure compared to the best values of the original systems. This leads to an improvement of the F-measure (83%): 4 points on the rule-based system and 7 points on the CRF system.

We found that one of the advantages of the CRF system is the flexibility to include different elements in the sources, so that it achieves complete sources in some cases in which the rules system finds only partial sources (example 3).

(3)

Original text:	según[5] una denuncia efectuada por funcionarios del INAU …
English translation:	*according to a complaint made by officials of INAU …*
Expected annotation:	según [una denuncia efectuada por funcionarios del INAU]
Rules annotation:	según [una denuncia] efectuada por funcionarios del INAU
CRF annotation:	según [una denuncia efectuada por funcionarios del INAU]
CRF+rules annotation:	según [una denuncia efectuada por funcionarios del INAU]

On the other hand, the rule system performs better for sources of nominal predicates, which have a low frequency in the training corpus (example 4).

(4)

Original text:	en palabras[6] del economista Fernando Ribeiro …
English translation:	*in the words of economist Fernando Riberio …*
Expected annotation:	en palabras [del economista Fernando Ribeiro] …
Rules annotation:	en palabras [del economista Fernando Ribeiro] …
CRF annotation:	en palabras del economista Fernando Ribeiro …
CRF+rules annotation:	en palabras [del economista Fernando Ribeiro] …

5.4 Recovery of Omitted Sources and Co-reference Chains for Sources

As an additional improvement in the recognition of the opinion source, we count now with a module specialized in co-reference resolution [1], including the recovery of omitted sources, very frequent in Spanish due to the possibility of omitting the subject.

It is worth noticing that it is very common in news texts expressing opinions of politicians or governors, to write out the opinion spread in more than one sentence.

[5] In this case the predicate is *según / according to*.

[6] In this case the predicate is *palabras / words.*

For stylistic reasons, the source is not repeated in each subsequent sentence: it is rather omitted or spelled in a different form.

The algorithm deals with:

1- Recovery of omitted sources. This process is triggered by opinions with no recognized source, and is very similar to pronominal anaphora resolution.

2- Co-reference chains. For each opinion source, the system chooses between two possibilities: a. the source belongs to a previously started co-reference chain, b. the source initiates a new co-reference chain.

The method relies on the maximization of a function that ranks a source according to a scale of first mention features:

- Existence of proper nouns and appositions within the nominal group
- Indefinite determinant
- Definite determinant
- Demonstrative

If the function value is below the threshold for initiating a new chain, criteria for selecting an existing chain, based on morphological agreement and WordNet relations, are applied.

The co-reference module recovers 61% of omitted sources and achieves 84% of F-measure for the co-reference chain task. It is not easy to compare this number with similar work because of the difference in scenarios and languages, and even the proliferation of metrics for the co-reference task [4].

6 Conclusions

We report on a system for the automatic identification of source opinions in Spanish journalistic texts. To our knowledge, this is the first system for this task for the Spanish language. A set of linguistic resources has been generated and will be made publicly available: an opinion predicate lexicon, two annotated corpora and a software tool for the automatic recognition of the opinion elements. All these resources have been extensively tested. The combined system for source recognition achieves 83% of exact F-measure, this result being similar to those reported for other languages: 78.1% of partial F-measure for English [6], 78% of partial F-measure for Chinese [11] and 62.6% of exact F-measure for English [21]. A detailed comparison was not possible due to the difference in languages and scope of related work. Future directions of this work will focus on the identification of the theme within the message component, and in the inference of an affective orientation for opinions.

References

1. Acerenza, F., Rabosto, M., Zubizarreta, M., Rosá, A., Wonsever, D.: Resolución de correferencias entre fuentes de opiniones en español. In: XXXVIII Conferencia Latinoamericana en Informática (to appear, CLEI 2012)

2. Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., Jurafsky, D.: Automatic Extraction of Opinion Propositions and their Holders. In: AAAI Spring Symposium on Exploring Attitude and Affect in Text, pp. 20–27. The AAAI Press, Menlo Park (2004)
3. Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., Jurafsky, D.: Extracting Opinion Propositions and Opinion Holders Using Syntactic and Lexical Cues. In: Shanahan, J., Qu, Y., Wiebe, J. (eds.) Computing Attitude and Affect in Text – Theory and Applications. The Information Retrieval Series, vol. 20, pp. 125–141. Springer, Heidelberg (2006)
4. Cai, J., Strube, M.: Evaluation Metrics for End-to-End Coreference Resolution Systems. In: 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2010), pp. 28–36. Association for Computational Linguistics, Stroudsburg (2010)
5. Choi, Y., Cardie, C., Riloff, E., Patwardhan, S.: Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005), pp. 355–362. Association for Computational Linguistics, Vancouver (2005)
6. Choi, Y., Breck, E., Cardie, C.: Joint Extraction of Entities and Relations for Opinion Recognition. In: 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 431–439. Association for Computational Linguistics, Stroudsburg (2006)
7. Choi, Y., Cardie, C.: Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In: 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), pp. 793–801. Association for Computational Linguistics, Stroudsburg (2008)
8. Kim, S.-M., Hovy, E.: Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Texts. In: 2006 Workshop on Sentiment and Subjectivity in Text (SST 2006), pp. 1–8. Association for Computational Linguistics, Stroudsburg (2006)
9. Krestel, R., Bergler, S., Witte, R.: Minding the Source – Automatic Tagging of Reported Speech in Newspaper Articles. In: 6th International Language Resources and Evaluation Conference (LREC 2008), pp. 2823–2828. ELRA (2008)
10. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields – Probabilistic Models for Segmenting and Labeling Sequence Data. In: 18th International Conference on Machine Learning (ICML 2001), pp. 282–289. ACM (2001)
11. Lu, B.: Identifying Opinion Holders and Targets with Dependency Parser in Chinese News Texts. In: Student Research Workshop at Human Language Technologies – 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT-SRWS 2010), pp. 46–51. Association for Computational Linguistics, Stroudsburg (2010)
12. Nakagawa, T., Inui, K., Kurohashi, S.: Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. In: 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), pp. 786–794. Association for Computational Linguistics, Stroudsburg (2010)
13. Pollard, C., Sag, I.A.: Information-Based Syntax and Semantics, Volume 1 – Fundamentals. CSLI Lecture Notes no. 13. Center for the Study of Language and Information (CSLI). University of Chicago Press, Stanford (1987)
14. Pouliquen, B., Steinberger, R., Best, C.: Automatic Detection of Quotations in Multilingual News. In: Recent Advances in Natural Language Processing (RANLP 2007), pp. 487–492 (2007)

15. Rosá, A., Wonsever, D., Minel, J.L.: Opinion Identification in Spanish Texts. In: Young Investigators Workshop on Computational Approaches to Languages of the Americas at Human Language Technologies – 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), pp. 54–61. Association for Computational Linguistics, Stroudsburg (2010)
16. Rosá, A.: Identificación de opiniones de diferentes fuentes en textos en español. PhD Thesis. Universidad de la República (Uruguay) / Université Paris Ouest Nanterre La Défense (France) (2011)
17. Ruppenhofer, J., Somasundaran, S., Wiebe, J.: Finding the Sources and Targets of Subjective Expressions. In: 6th International Language Resources and Evaluation Conference (LREC 2008), pp. 2781–2788. ELRA (2008)
18. Saurí, R.: A Factuality Profiler for Eventualities in Text. PhD dissertation. Brandeis University (2008)
19. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields. arXiv, p. arXiv:1011.4088v1 (2010)
20. Wiebe, J., Wilson, T., Cardie, C.: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation 39(2-3), 165–210 (2005)
21. Wiegand, M., Klakow, D.: Convolution Kernels for Opinion Holder Extraction. In: Human Language Technologies – 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), pp. 795–803. Association for Computational Linguistics, Stroudsburg (2010)
22. Wonsever, D., Minel, J.-L.: Contextual Rules for Text Analysis. In: Gelbukh, A. (ed.) CICLing 2001. LNCS, vol. 2004, pp. 509–523. Springer, Heidelberg (2001)