Extending a Deep Learning Approach for Negation Cues Detection in Spanish

Hermenegildo Fabregat[1,3], **Andres Duque**[2,3,4], **Juan Martinez-Romo**[1,3,4], and **Lourdes Araujo**[1,3,4]

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos
[2]Departamento de Sistemas de Comunicación y Control
[3]Universidad Nacional de Educación a Distancia (UNED)
[4]Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS)
gildo.fabregat@lsi.uned.es, aduque@scc.uned.es
juaner@lsi.uned.es, lurdes@lsi.uned.es

Abstract. This paper describes the negation cue detection system presented by the NLP_UNED group for Task A (*Tarea A: Detección de claves de negación*) of the second edition of NegES workshop [9]. The task deals with negation cues detection in Spanish reviews in domains such as cars, music and books. The proposed system is an extension of the one proposed in the previous edition by the UNED team. This system consists of a deep learning architecture and the application of a set of rules. The deep learning architecture is based on the use of a Bi-LSTM to process contextual information. The purpose of applying a stack of rules is to correct frequent classification errors. The results obtained improve the performance achieved by our team in the previous edition and are highly competitive compared to the rest of participants, placing second in the global ranking of this edition.

Keywords: Negation detection · negation cues · Deep Learning · Bi-LSTM · based-rules system

1 Introduction

The automatic processing of negation is a very important task in natural language processing as it allows us to identify negated facts. It is a very interesting field of study if we consider the influence of the negation in tasks such as sentiment analysis and relationship extraction [15, 4]. NegEx [2] is one of the most popular algorithms in the study of negation in English. The use of this algorithm for other languages has been addressed by some recent work, such as Chapman et al. [3] (French, German and Swedish), Skeppstedt [16] (Swedish) and Cotik et al. [5] (Spanish) which also explore other syntactic approaches based on rules derived from PoS-tagging (Part-of-speech) and dependency tree patterns for negation detection in Spanish.

The proposal for Task A of the NegEs workshop is the same as that presented in the previous edition [10] and focuses on the detection of Spanish negation cues. For this purpose the organizers facilitate the corpus SFU ReviewSP-NEG [11] which consists of 400 reviews related to 8 different domains (cars, hotels, washing machines, books, cell phones, music, computers and movies), 221866 words and 9455 sentences, out of which 3022 sentences contain at least one negation structure. In the same way as last year, the organizers have presented the same corpus divided into three sets, training, development and test. According to the information provided by the organizers, the corpus division was carried out randomly, ensuring 33 reviews per domain in training, 7 per domain in development and 10 per domain in test. As can be seen in Figure 1, the corpus was presented using the format CoNLL [7]. For this task, two different systems/teams participated in the 2018 edition of the NegES workshop. On one hand, the model proposed by the UPC team, based on the use of a Conditional Random Field (CRF). This model was trained with some casing features such as "word contains punctuation" and "information about n-grams of up to 6 words before the observed word" among others [13]. The second one, presented by the UNED team, consisted of a Deep Learning architecture based on Bi-LSTM and neural networks [6]. Like the previous one, the model of the UNED team used casing information, however, this information was represented by an embedding generated during training phase. The system proposed by the UPC team obtained the best results.

```
hoteles 21 1 Y y cc coordinating - - -
hoteles 21 2 no no rn negative no - -
hoteles 21 3 hay haber vmip3s0 main - - -
hoteles 21 4 en en sps00 preposition - - -
hoteles 21 5 la el da0fs0 article - - -
hoteles 21 6 habitación habitación ncfs000 common - - -
hoteles 21 7 ni ni rn negative ni - -
hoteles 21 8 una uno di0fs0 indefinite - - -
hoteles 21 9 triste triste aq0cs0 qualificative - - -
hoteles 21 10 hoja hoja ncfs000 common - - -
```

Fig. 1: Corpus SFU ReviewSP-NEG - Annotation format.

This work is organized as follows: Section 2 contains both the description of the proposed system and the description of the features and resources used. In

Section 3 we report and discuss the results obtained during the evaluation stage. Finally, in Section 4 conclusions and future work are presented.

2 Proposed system

The proposed system consists of two components: a deep learning model and a post-processing phase. On one hand, the proposed deep learning model is a revision of the model proposed by Fabregat et al. [6]. The proposed model incorporates a deep learning sub-architecture for character-level term processing [17] and makes use of a One-hot vector to represent term casing information. On the other hand, the proposed system considers the application of a post-processing phase based on the use of a stack of rules (regular expressions) to correct some frequent errors.

2.1 Deep Learning model

The proposed deep learning model (Figure 2) uses the following embedding features: words, PoS-tagging and characters. Words are encoded using a pre-trained Spanish word embedding [1] and both PoS-tagging and character embedding models have been implemented using two Keras Embedding Layers[1] initialized using a random uniform distribution. The part-of-speech model used is the one provided in the corpus. In addition, a One-hot vector has been used to represent casing information. Since the corpus contains information extracted from Internet websites and written by users without the supervision of a corrector or similar, it is necessary to keep in mind that some of the terms in the corpus may not be present in the word embedding dictionary, or their PoS-tags could have been incorrectly assigned. The use of universal representations such as the character embedding model and the casing vector satisfies the need to represent information that cannot be collected through experience-based representation models. On the one hand, in order to process the sequence of characters that compose a word we have represented these using an embedding representation and we have applied a set of convolutions to extract the most relevant information from each sequence of characters. On the other hand, the casing vector allows us to represent information that has been deleted from each term to match an element of the word embedding dictionary. This representation can be used to indicate to the model some scenarios such as a “term ending in a comma” or that a “term contains numbers” among others.

Using the training set, the model has been trained during a total of 50 epochs. We have tried not to give preference to any category or domain by integrating the training data of each category in a single set. Pre-trained resources and model’s parameters are the following:

- Pre-trained English Word Embedding dimension: 300
- Embeddings dimension (Characters / PoS-tagging): 50 / 50

[1] https://keras.io/layers/embeddings/

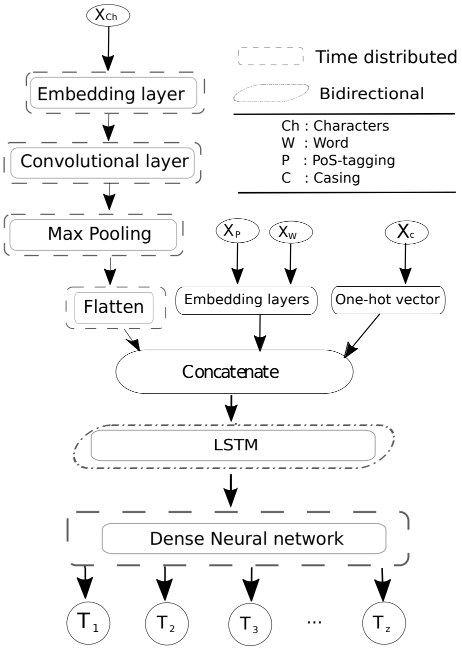

Fig. 2: Architecture of the proposed model, where Ch and W (Ch: Character sequence, W: Raw word) are the encoded word inputs and P is the encoded input representing the PoS-tagging using an embedding representation. Casing information (C) is encoded using a One-hot vector based representation. Bi-LSTM inputs are the concatenated features of each word. In the output layer, T_x represents the assigned tag.

- One-hot dimension (Casing): 8
- Bi-LSTM output dimension: 150 (per each LSTM)
- Conv1D (kernel_size / filter): 3 / 30
- Batch size / Model optimizer: 32 / Adam [12]

In order to avoid any possible over-adjustment of the model over the training set, some dropouts have been applied in both the character level and term level processing.

- Conv1D output dropout: 0.5
- LSTM-Dropout: 0.5
- LSTM Recurrent Dropout: 0.25

Finally, we used the standard IOB labeling scheme [14] to label the targets. The first cue of a negation phrase is denoted by B (Begin) and the remaining cues, if any, with I (Inside). O (Out) indicates that the word does not correspond to any considered kind of entity.

2.2 Post-processing

After the deep learning architecture, the proposed system includes a post-processing phase for the correction of frequent errors detected in the development phase. The stack of rules applied in this phase has consisted of a total of 11 rules focused mainly on processing those cases where the role of a trigger is modified by elements such as conjunctions and modifiers such as *tan* or *siquiera* among others. Some of the rules applied are:

- The expression “sin embargo” does not correspond to a negation cue.
- After “,”, “?”, “¿”, “!”, “¡” or “;” any discovered cue is part of another scope.
- The term “no” in expressions such as “no tan”, “hasta que no” or “no [me|le|te] [ha|han|has|he|...] [Verb] [Quantifier]” doesn’t correspond to a negation cue.

All the applied rules have been generated after analyzing the model errors in the development set, trying to avoid possible generalization errors. Even so, although some of the rules developed behave effectively in this corpus, these may not properly represent the complexity of the problem.

3 Evaluation

This section describes the results obtained by the proposed system. The evaluation has been carried out taking into account the evaluation criteria proposed by the organizers.

- Punctuation tokens are ignored.
- True positives are counted when the system produces negation elements exactly as they are in the gold set.
- Partial matches are not counted as False Positive (FP), only as False Negative (FN).
- False negatives are counted either by the system not identifying negation elements present in the gold set, or by identifying them partially.
- False positives are counted when the system produces a negation element not present in the gold set.

Using the development set to evaluate, Table 1 shows the improvement obtained before and after taking into account the post-processing phase. Results show that improvements are obtained in all domains except in “Computers” where both precision and recall decrease in a remarkable way. The reason for this behavior comes from a particular rule that is applied incorrectly in some cases, being more evident in this domain. The applied rule incorrectly considers that any enumeration of expressions with some negation cue and separated by “,” correspond to different scopes. However, due to its overall positive influence it has been finally maintained within the set of rules. Since no solution could be identified, it was decided to include it in the set due to the global improvement

Table 1: Development set - Evaluation per domain. Comparison of the results obtained by the proposed system, with and without rules: NLP_UNED - 2019 with rules (NLP_UNED - 2019 no rules).

Domain	Precision	Recall	F-measure
Cars	95.35% (86.36%)	87.23% (80.85%)	91.11% (83.51%)
Hotels	98.00% (93.88%)	80.33% (75.41%)	88.29% (83.64%)
Washing machines	97.37% (97.30%)	80.33% (80.00%)	88.29% (87.81%)
Books	93.28% (92.54%)	86.81% (86.11%)	89.93% (89.21%)
Phones	96.94% (94.95%)	87.16% (86.24%)	91.79% (90.39%)
Music	96.00% (85.19%)	92.31% (88.46%)	94.12% (86.89%)
Computers	91.62% (92.00%)	84.62% (88.46%)	88.00% (90.20%)
Films	94.68% (92.78%)	80.91% (81.82%)	87.26% (86.96%)

it implies. The difference in the micro-average F1 values for both cases (**89.67% with rules and 87.88% without rules**) shows that the addition of a post-processing phase provides remarkable improvements.

Once we have validated the model, it has been compared with the model presented by the UNED team in the last edition of the workshop. As Table 2 shows, for both with and without rules, the improvements are quite noticeable especially in recall. This improvement may be due both to the addition of the character-based representation model and to the simplification of the casing model (in the previous model this casing model was represented as an embedding).

On the other hand, using the test set, as can be seen in Table 3 the results of our proposed system are compared with those of CLiC team. Although our proposed system does not improve the overall results obtained by the CLiC team, in most cases our system obtains better results in terms of precision, with the exception of the "car" domain. Regarding recall, the greatest differences between these two systems are found in domains such as "Books", "Computers" and "Phones". After an analysis of these results, one of the possible conclusions is that our system does not behave properly when processing certain unseen samples.

Table 4: Test set - Evaluation per domain: Full comparison.

Domain	Precision	Recall	F1
CLiC	89.67%	79.40%	84.09%
NLP_UNED	91.82%	75.98%	82.99%
IBI	91.22%	72.16%	80.50%
Aspie	18.80%	28.34%	22.58%

Table 2: Development set - Evaluation per domain. Comparison of the results obtained in this edition and those obtained in the previous edition: NLP_UNED - 2019 (UNED - 2018).

Domain	Precision	Recall	F-measure
Cars	95.35% (88.37%)	87.02% (80.85%)	91.11% (84.44%)
Hotels	98.00% (90.62%)	80.33% (47.54%)	88.29% (62.36%)
Washing machines	97.37% (96.88%)	80.33% (68.89%)	88.29% (80.52%)
Books	93.28% (91.00%)	86.81% (63.19%)	89.93% (74.59%)
Phones	96.94% (94.20%)	87.16% (59.63%)	91.79% (73.03%)
Music	96.00% (85.19%)	92.31% (88.46%)	94.12% (86.79%)
Computers	91.62% (84.62%)	84.62% (63.46%)	88.00% (72.53%)
Films	94.68% (93.33%)	80.91% (63.64%)	87.26% (75.68%)

Table 3: Test set - Evaluation per domain. Comparison of the results obtained by our system (NLP_UNED) with those obtained by the best system proposed (CLiC): NLP_UNED - 2019 (CLiC).

Domain	Precision	Recall	F-measure
Cars	94.83% (94.92%)	80.88% (82.35%)	87.30% (88.19%)
Hotels	93.62% (87.50%)	74.58% (71.19%)	83.02% (78.51%)
Washing machines	94.34% (92.98%)	72.46% (76.81%)	81.96% (84.13%)
Books	84.02% (80.59%)	81.35% (72.62%)	87.66% (76.57%)
Phones	88.37% (87.76%)	66.67% (75.44%)	76.00% (81.13%)
Music	95.38% (94.44%)	71.26% (78.16%)	81.57% (85.53%)
Computers	94.12% (90.48%)	79.01% (93.83%)	85.91% (92.12%)
Films	89.86% (88.67%)	81.60% (81.60%)	85.53% (84.99%)

Finally, table 4 shows a comparison of the average results obtained in this edition of the task A of NegEs workshop. As can be seen, the obtained results are very competitive, although as we mentioned above, the average recall of the proposed system is low compared with the obtained precision. However, this fact is repeated in most of the reported systems. These results may indicate that systems do not work properly with some kinds of instances unseen during the training phase. A more detailed analysis would be necessary to draw other conclusions.

3.1 Error analysis

After analyzing the output provided by the system for the development set we found that many of the errors detected by the UNED team with this new attempt were minimized, especially those cases where there are typos in the text. The errors detected in multi-term expressions and enumerations have also been reduced. Some of the detected errors in this new version of the system are derived from scenarios not covered by the set of rules or negation triggers not seen during training phase.

4 Concluding Remarks

Due to the importance of automatic negation processing in the field of natural language processing, the detection of negation cues is a very important task. In this paper we present a revision of a system based on deep learning for the detection of negation cues in Spanish. In summary, the system consists of a model based on Deep Learning and a post-processing phase based on the application of a stack of rules. Considering the results shown, the system achieves a quite remarkable improvement compared to its predecessor and results are comparable to those obtained by the rest of the participants, both in this edition and in the previous one.

Taking into account the results obtained, our future work focuses on improving the stack of rules in order to solve the errors detected in the development phase. Another improvement could be to explore the replacement of the output layer of the Deep Learning model by a CRF. The combination of Bi-LSTM and CRF has achieved very interesting results in similar tasks [8].

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the projects PROSA-MED (TIN2016-77820-C3-2-R) and EXTRAE (IMIENS 2017).

References

1. Cardellino, C.: Spanish billion words corpus and embeddings (2016)
2. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics **34**(5), 301 – 310 (2001)
3. Chapman, W.W., Hilert, D., Velupillai, S., Kvist, M., Skeppstedt, M., Chapman, B.E., Conway, M., Tharp, M., Mowery, D.L., Deleger, L.: Extending the negex lexicon for multiple languages. Studies in health technology and informatics **192**, 677 (2013)
4. Chowdhury, M.F.M., Lavelli, A.: Exploiting the scope of negations and heterogeneous features for relation extraction: A case study for drug-drug interaction extraction. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 765–771 (2013)
5. Cotik, V., Stricker, V., Vivaldi, J., Rodríguez Hontoria, H.: Syntactic methods for negation detection in radiology reports in Spanish. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, BioNLP 2016: Berlin, Germany, August 12, 2016. pp. 156–165. Association for Computational Linguistics (2016)

6. Fabregat, H., Martinez-Romo, J., Araujo, L.: Deep Learning Approach for Negation Cues Detection in Spanish. In: NEGES 2018: Workshop on Negation in Spanish: Seville, Spain: September 19-21, 2018: proceedings book. pp. 43–48 (2018)
7. Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M.A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., et al.: The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task. pp. 1–18. Association for Computational Linguistics (2009)
8. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
9. Jiménez-Zafra, S.M., Cruz Díaz, N.P., Morante, R., Martín-Valdivia, M.T.: NEGES 2019 Task: Negation in Spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (2019)
10. Jiménez-Zafra, S.M., Díaz, N.P.C., Morante, R., Martín-Valdivia, M.T.: NEGES 2018: Workshop on Negation in Spanish. Procesamiento del Lenguaje Natural **62**, 21–28 (2019)
11. Jiménez-Zafra, S.M., Taulé, M., Martín-Valdivia, M.T., Ureña-López, L.A., Martí, M.A.: Sfu review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. Language Resources and Evaluation **52**(2), 533–569 (2018)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Loharja, H., Padró, L., Turmo Borras, J.: Negation cues detection using CRF on Spanish product review texts. In: NEGES 2018: Workshop on Negation in Spanish: Seville, Spain: September 19-21, 2018: proceedings book. pp. 49–54 (2018)
14. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Natural language processing using very large corpora, pp. 157–176. Springer (1999)
15. Reitan, J., Faret, J., Gambäck, B., Bungum, L.: Negation scope detection for twitter sentiment analysis. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 99–108 (2015)
16. Skeppstedt, M.: Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. In: Journal of Biomedical Semantics. vol. 2, p. S3. BioMed Central (2011)
17. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in neural information processing systems. pp. 649–657 (2015)