Deep Learning approach for Negation Cues Detection in Spanish

Aplicación Basada en Deep Learning para Identificación de Claves de Negación en Castellano

Hermenegildo Fabregat[1], **Juan Martinez-Romo**[1−2],**Lourdes Araujo**[1−2]
[1]Universidad Nacional de Educación a Distancia (UNED)
[2]IMIENS: Instituto Mixto de Investigación
{gildo.fabregat, lurdes, juaner}@lsi.uned.es

Abstract: This paper describes the negation cues detection model presented by the UNED group for task 2 (*Task 2: Negation cues detection*) of the NEGES workshop collocated in the SEPLN congress (Sevilla, 2018). This task deals with negation cues detection in Spanish reviews in domains such as cars, music and books. In order to deal with the extraction of both semantic and syntactic patterns and the extraction of contextual patterns, we have proposed a model based on the combination of some dense neural networks and one Bidirectional Long Short-Term Memory (Bi-LSTM). The evaluation is divided by domains and using an inter-domain average we have obtained acceptable results.
Keywords: Negation detection, negation cues, Deep Learning, Bi-LSTM

Resumen: Este artículo describe el modelo propuesto por el grupo UNED para la tarea 2 (*Task 2: Negation cues detection*) del *workshop* NEGES, asociado a la conferencia SEPLN (Sevilla, 2018). Esta tarea trata la detección de "señales o claves" de negación en castellano, centrando la atención en comentarios de dominios tales como coches, musica y libros. Con el fin de extraer patrones tanto sintácticos como semánticos además de patrones basados en información contextual, el modelo esta basado en el uso de varias redes neuronales junto a una *LSTM* (*Long Short-Term Memory*) bidireccional. Estando la evaluación de la tarea dividida en función del dominio de los comentarios, los resultados medios obtenidos durante la evaluación han sido aceptables.
Palabras clave: Detección de negación, claves de negación, Deep Learning, Bi-LSTM

CEUR-WS.org/Vol-2174/paper5.pdf

1 Introduction

To understand the meaning of a sentence through the use of the natural language processing techniques it is necessary to take into account that a sentence can express a negated fact. In some languages such as English, detection and processing of negation is a recurrent working area. 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 (Reitan et al., 2015; Chowdhury and Lavelli, 2013). NegEx (Chapman et al., 2001) is one of the most popular algorithms for negation detection in English. The use of this algorithm for other languages has been addressed by some recent works, such as Chapman et al. (2013) (French, German and Swedish), Skeppstedt (2011) (Swedish) and Cotik et al. (2016) (Spanish) which also explore other syntactic approaches based on rules derived from PoS-tagging and dependency tree patterns for negation detection in Spanish.

The proposal of the task 2 of NEGES workshop (Jiménez-Zafra et al., 2018a) focuses on the detection of negated cues in Spanish. For this purpose the organizers facilitate the corpus SFU ReviewSP-NEG (Jiménez-Zafra et al., 2018b) 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. The organizers have presented the corpus divided in three sets: training, development and

Proceedings of NEGES 2018: Workshop on Negation in Spanish, pages 43-48
Seville, Spain, September, 18, 2018

test. As can be seen in the figure 1, the corpus was presented using the format CoNLLL (Hajič et al., 2009).

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 - - -

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

Each line corresponds to a token, where an empty line is the end of a sentence and each column represents an annotation about a specific term (for instance, column one contains the name of the domain file and columns three and four contain word and lemma). Column eight onwards shows the annotations related to negation. If the sentence has no negations, column eight has a value "***" and there are no more columns. Otherwise, the notation for each negation is provided in three columns. The first column contains the word that belongs to the negation cue. The second and third columns contain "-".

This work is organized as follows: Section 2 contains both the description of the proposed model and the description of the features and resources used. In section 3 we report and discuss the results obtained during the evaluation stage. And finally, in section 4 conclusions and future work are presented.

2 Proposed model

Inspired by the model presented by Fancellu, Lopez, and Webber (2016), the problem is addressed as a sequence labeling task.

The proposed model has been implemented using Python's Keras library (Chollet and others, 2015) with TensorFlow backend and it is a supervised approach which uses the following embedded features: words, lemmas, PoS-tagging and case-tagging. Both words and lemmas are encoded using a pre-trained Spanish word embedding (Cardellino, 2016) and both PoS-tagging and casing embedding models have been implemented using two Keras Embedding Layer[1] initialized using a random uniform distribution. In order to avoid any cascade error we used both lemmas and PoS-tagging provided in the corpus.

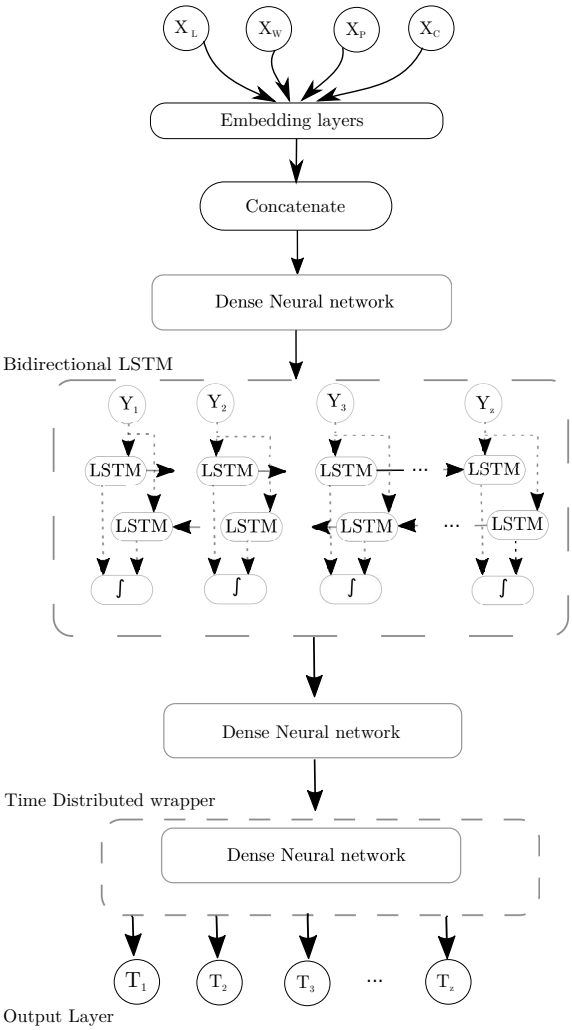

Figure 2: Architecture of the proposed model, where X_L and X_W (L: Lemma, W: Raw word) are the encoded word inputs and X_P and X_C are the encoded inputs representing the PoS-tagging and casing information. Bi-LSTM inputs (Y_x) are the concatenated embedded features of each word. In the output layer, T_x represents the assigned tag.

The casing embedding matrix is a hot-one encoding matrix of size 8 which was

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

calculated for each input token making use the following encoder dictionary: { 0: Input token is numerical - 1: - 2: - 3: Initial character is upper case - 4: Input token is mainly numerical - 5: Contains at least one digit - 6: Other case }.

In order to ensure that the words presented in the corpus, which are linked by an underscore such as "ya_que" are not being left out of the embedding, we have carried out a preprocessing step to divide these expressions according to the number of underscores that these expressions have. To standardize the sentences to a common length, after dividing expressions with more than one term, a padding of up to 200 positions has been applied. To label the targets, we follow the standard IOB labeling scheme (Ramshaw and Marcus, 1999). 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 kind of entity considered. For example:

Del (*O*) *buffet* (*O*) *del* (*O*) *desayuno* (*O*) *no* (*B*) *puedo* (*O*) *opinar* (*O*) *ya_que* (*B*) *no* (*I*) *lo* (*O*) *incluia* (*O*) *nuestro* (*O*) *regimen* (*O*) . (*O*)

Figure 2 shows the proposed model architecture. The first layer is a densely connected hidden layer (Dense neural network), which has as activation function the hyperbolic tangent function (tanh). This layer takes as input the concatenation of the different embeddings. The output of the first layer is connected to an LSTM (Long Short-Term memory) enveloped in a bidirectional wrapper (forward and backward processing network). For each network, this second layer uses a hidden state for processing data from the current step taking into account information of previous steps. In the next layer and connected to the output layer, another dense hidden layer has been used to reduce the complexity of the bidirectional LSTM output. To avoid possible over-fitting we have applied a dropout factor of 0.25 to the output of this dense layer. Finally, another dense hidden layer, using the softmax activation function, calculates the probabilities of all tags for each word in a sentence. The most probable label is the one selected as the final tag.

The model has been trained with data from all the categories and this process has been limited to 25 epochs in order to avoid possible over-fitting. We have evaluated during the training phase for each epoch the generated model using the script provided by the organizers (Morante and Blanco, 2012) and the development set and we have observed that, for most of the domains, 20th epoch are enough to reach the best results (Figure 3).

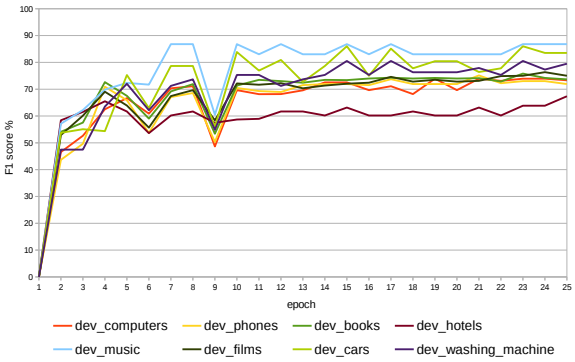

Figure 3: Training phase, temporal evaluation for each domain using development set.

Pre-trained resources parameters and model's hyper-parameters are the following:

- Pre-trained English Word Embedding dimension: 300
- Embeddings dimension (Casing / PoS-tagging): 8 / 50
- Hidden Dense units (output dimension / activation function): 200 / tanh
- LSTM output dimension: 300
- Dropout (for each dense unit): 0.25
- Batch size / Model optimizer: 32 / AdaGrad (Duchi, Hazan, and Singer, 2011)

Once the model has been set and it has a stable and similar performance for all categories, the model has been re-trained with the data of the development set.

3 Evaluation

In this section we describe the obtained results, taking into account the following 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 gold.

Domain	Precision	Recall	F-measure
Cars	44.74 %	72.34 %	55.29 %
Hotels	51.32 %	63.93 %	56.94 %
Washing machines	55.36 %	68.89 %	61.39 %
Books	53.11 %	65.28 %	58.57 %
Phones	54.62 %	65.14 %	59.42 %
Music	43.59 %	65.38 %	52.31 %
Computers	38.57 %	51.92 %	44.26 %
Films	50.00 %	59.09 %	54.17 %

Table 1: Baseline - Evaluation per domain: development set

Domain	Precision	Recall	F-measure
Cars	94.23 % (88.37 %)	72.06 % (80.85 %)	81.67 % (84.44 %)
Hotels	97.67 % (90.62 %)	71.19 % (47.54 %)	82.35 % (62.36 %)
Washing machines	92.00 % (96.88 %)	66.67 % (68.89 %)	77.31 % (80.52 %)
Books	79.52 % (91.00 %)	66.27 % (63.19 %)	72.29 % (74.59 %)
Phones	93.33 % (94.20 %)	73.68 % (59.63 %)	82.35 % (73.03 %)
Music	92.59 % (85.19 %)	57.47 % (88.46 %)	70.92 % (86.79 %)
Computers	– (84.62 %)	– (63.46 %)	– (72.53 %)
Films	86.26 % (93.33 %)	69.33 % (63.64 %)	76.87 % (75.68 %)

Table 2: Evaluation per domain: test set (development set)

- Partial matches are not counted as FP, only as FN.
- False negatives are counted either by the system not identifying negation elements present in gold, or by identifying them partially.
- False positives are counted when the system produces a negation element not present in gold.

In order to carry out a study of the performance of the presented system, it has been compared with a baseline based on a lookup of a filtered list of terms extracted from the training set. To take into account the scope of the negation, the sentences have been divided according to the following delimiters: "." - "," - ";". The list of terms has been tunned in order to improve the results obtained through this baseline. Table 1 shows the results obtained using the baseline (evaluating it with the development set) and table 2 shows the results obtained using the proposed approach. As can be seen, table 2 presents two scores for each evaluation metric (precision, recall and f-measure). These scores correspond to the evaluation of the system using the development set during the training phase and to the evaluation of the system carried out by the organizers using the unannotated test set. Due to an error submitting the system output, there are no test results for the computer category. On the one hand, the results obtained in a preliminary analysis (development set) show that the proposed system significantly improves the results obtained by the baseline. On the other hand, as shown in table 2, the difference between recall and precision is very remarkable. Taking into account that we have not generated a specific model for each domain, due to the needs of the system presented, the differences between precision and recall observed during the evaluation of the test set may indicate, among other things, that the system has some over-fitting and is adjusting to very recurrent patterns or that there are expressions that have not been processed correctly (for example, there may be expressions that are not correctly included in the word embedding used). On the other hand, the fall of the recall value in the music domain is notable, comparing the results of the test and training.

Because the gold standard has not been published, we have not been able to perform an exhaustive analysis of the recognition mistakes made evaluating with the test set. However, some of the detected errors during the training phase related to the obtained recall,

correspond to situations in which the model has not been able to recognize some multiword expressions related to a negation such as "a_no_ser_que" and "no_hay_mas_que".

4 Concluding Remarks

The detection of negation cues is an important task in the natural language processing area. In this field we present a deep learning model for detection of negation cues inspired in named entity recognition architectures and negation scope detection models. This model achieves high performance without any sophisticated features extraction process and although the model has some weaknesses in terms of coverage, the results are acceptable and comparable with those obtained by the UPC-TALP team (average results, 91.47 % precision, 82.17 % recall and 86.44 % F-measure).

As a future work, based on the low recall obtained we will explore others regularization methods such as the use of some regularization function (Cogswell et al., 2015) and we will explore some model modifications such as the addition of a semantic vector representation for the whole sentence and the use of a CRF-based layer instead of the current dense based output layer. Finally, the study of the patterns generated by the current model can lead to the creation of a rule-based auxiliary model for the re-labeling of negation beginning cues (label B). If we take into account that the model has been trained using non-handcrafted features, the results obtained indicate that the system is capable of achieving more competitive levels of precision and recall.

Acknowledgments

This work has been partially supported by the projects EXTRECM (TIN2013-46616-C2-2-R), PROSA-MED (TIN2016-77820-C3-2-R), and EXTRAE (IMIENS 2017).

References

Cardellino, C. 2016. Spanish billion words corpus and embeddings.

Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301 – 310.

Chapman, W. W., D. Hilert, S. Velupillai, M. Kvist, M. Skeppstedt, B. E. Chapman, M. Conway, M. Tharp, D. L. Mowery, and L. Deleger. 2013. Extending the NegEx lexicon for multiple languages. *Studies in health technology and informatics*, 192:677.

Chollet, F. et al. 2015. Keras. `https://github.com/fchollet/keras`.

Chowdhury, M. F. M. and A. Lavelli. 2013. 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*, pages 765–771.

Cogswell, M., F. Ahmed, R. B. Girshick, L. Zitnick, and D. Batra. 2015. Reducing Overfitting in Deep Networks by Decorrelating Representations. *CoRR*, abs/1511.06068.

Cotik, V., V. Stricker, J. Vivaldi, and H. Rodríguez Hontoria. 2016. 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*, pages 156–165. Association for Computational Linguistics.

Duchi, J., E. Hazan, and Y. Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Fancellu, F., A. Lopez, and B. Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 495–504.

Hajič, J., M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, et al. 2009. 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*, pages 1–18. Association for Computational Linguistics.

Jiménez-Zafra, S. M., N. P. Cruz-Díaz, R. Morante, and M. T. Martín-Valdivia. 2018a. Resumen de la Tarea 2 del Taller NEGES 2018: Detección de Claves de Negación. In *Proceedings of NEGES 2018: Workshop on Negation in Spanish*, volume 2174, pages 35–41.

Jiménez-Zafra, S. M., M. Taulé, M. T. Martín-Valdivia, L. A. Ureña-López, and M. A. Martí. 2018b. 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.

Morante, R. and E. Blanco. 2012. * SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 265–274. Association for Computational Linguistics.

Ramshaw, L. A. and M. P. Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*. Springer, pages 157–176.

Reitan, J., J. Faret, B. Gambäck, and L. Bungum. 2015. Negation scope detection for twitter sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 99–108.

Skeppstedt, M. 2011. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. In *Journal of Biomedical Semantics*, volume 2, page S3. BioMed Central.