NLP_UNED at eHealth-KD Challenge 2019

Deep Learning for Named Entity Recognition and Attentive Relation Extraction

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

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

Abstract. This paper describes the approach presented by the NLP_UNED team in the *eHealth Knowledge Discovery* challenge of the IberLEF 2019 competition. Our proposal is based on the use of deep neural networks for performing keyphrase detection and attention-based networks for extracting relationships between those keyphrases. Our experiments show promising results especially in the Relation Extraction subtask, offering the second best results among all participant systems.

Keywords: Deep learning · Neural networks · Attention systems · Biomedical domain · Named Entity Recognition · Relation Extraction

1 Introduction

The biomedical domain is one of the most important research fields at the moment when it comes to Natural Language Processing (NLP). Successful identification and extraction of valuable data in an automatic way is a crucial process considering the huge amount of information available in this particular domain. Clinical notes, medical reports or biomedical research papers are just some of the multiple types of documents in which medical information can be found.

In that context, the *eHealth Knowledge Discovery Challenge* [15], carried out within the *Iberian Languages Evaluation Forum* (IberLEF 2019), offers an opportunity for the development of NLP systems that may help in this search for useful data in biomedical information. The task aims for the correct identification and classification of keyphrases in sentences extracted from biomedical documents written in the Spanish language, and the search for meaningful relationships between those keyphrases.

In this paper, we present the deep learning-based system *DeepNER+ARE*, designed for considering the challenge as a sequential pipeline divided into two phases. In the first phase, a Named Entity Recognizer (NER) is used for detecting and classifying keyphrases, while the relationships between them are extracted in a second step through the use of an Attention-based Relation Extractor (ARE).

The paper is structured as follows: Section 2 offers an overview on systems developed for similar tasks. Section 3 briefly describes the characteristics of the task. The proposed system is presented in Section 4, and the results obtained in the challenge are shown in Section 5. Finally, some conclusions and future lines of work are discussed in Section 6.

2 Background

Many approaches can be found in the literature addressing the identification and classification of keyphrases and the extraction of relationships between them, both regarding general NLP [19,14,12] and its application in the biomedical domain. The proposed challenge itself is closely related to task 3 (*eHealth Knowledge Discovery*) of TASS 2018 [11], in which systems addressed 3 subtasks: keyphrase identification, keyphrase classification and relation extraction. The two best performing systems were based on deep learning solutions: one of them made use of bidirectional Long Short-Term Memory (Bi-LSTM) layers combined with Conditional Random Field (CRF) classifiers, although it only offered results for the detection and classification of keyphrases [20]. The other system considered the task as a whole and developed a strategy for jointly classifying keyphrases and relationships, through the use of Convolutional Neural Networks (CNN) [13]. Convolutional layers were also used by another system, although only for the relation extraction subtask [17].

Other approaches have also been considered by participant teams in the task: in [8], morphological analysis and the biomedical knowledge base *Unified Medical Language System* (UMLS) [1] are combined, only for the keyphrase extraction subtask. A classic statistical NLP pipeline is combined in another system with machine learning techniques such as CRF classifiers and logistic regression model for offering results for all the subtasks [18].

In a similar way to that presented in [20], our system also proposes the use of Bi-LSTM layers, but is able to offer results for both phases of the task. Moreover, it incorporates the use of attention layers [21] in the second step, which help the model to focus on specific parts of the input for improving the relationship detection.

Beyond this task, the use of deep learning techniques for entity and relation extraction is being widely explored in the biomedical domain during the last years. Different works can be found in the literature that exploit these techniques for either extracting general entities and relationships [4,9] or addressing specific types such as drugs and adverse effects [7] or rare diseases and dissabilities [3].

3 Task Proposal

In this section the most important characteristics of the *eHealth Knowledge Discovery* challenge are presented. Further details about the task can be found in [15].

3.1 Subtask A: Identification and Classification of Keyphrases

The first subtask aims for finding and classifying relevant pieces of information (words or sequences of words) within a sentence extracted from a biomedical document. Once the span of those keyphrases have been detected, the systems must classify them as belonging to one of the following classes: *Concept* (a general term or idea), *Action* (a term that process or modifies other concepts), *Predicate* (a term that represents a function or filter over a set of elements) or *Reference* (a term that refers to a concept).

3.2 Subtask B: Detection of Semantic Relations

The second subtask proposes the classification of possible relationships between the keyphrases. Types of relationships are categorized as *general*, *conceptual* (both of them involving the four different keyphrase classes), *action roles* (involving only actions) and *predicate roles* (involving only predicates). Each category contains a set of classes, for a total of 13 different relationships.

3.3 Datasets

The *eHealth-KD* corpus has been published by the organizers of the task, containing 700 (600 training and 100 development) different biomedical-related sentences. The whole test dataset contains 8,800 sentences.

3.4 Evaluation

Three different scenarios are proposed for the evaluation of the participant systems, which is carried out in terms of precision, recall and F1-measure. Metrics are computed in terms of correct, incorrect and partial matches for both the classification of keyphrases and relationships, and also missing and spurious matches are considered in the classification of keyphrases.

Scenario 1 considers the whole pipeline of the task, and hence can be seen as the main evaluation, in which systems receive raw sentences as input and must output the detected keyphrases and their assigned labels, as well as the relationships between keyphrases and their labels. Scenario 2 only evaluates keyphrase detection and classification from raw sentences, and Scenario 3 only evaluates relationship detection and classification considering raw sentences and labelled keyphrases as input.

The complete test dataset composed of 8,800 sentences is used in Scenario 1, and participants are asked to provide their solution for all the sentences, although only 100 of them have been used for the actual evaluation of the systems. Regarding Scenarios 2 and 3, the test datasets contain 100 sentences each.

4 System Description

The *DeepNER+ARE* system is divided into two separate sequential subsystems for addressing each of the subtasks of the challenge pipeline.

4.1 Keyphrase recognition and classification

The keyphrase detection phase has been addressed by using a subsystem which consists of a pre-processing phase, where input data is adapted and prepared, a supervised deep learning model, and a post-processing step for solving systematic errors through hand-crafted rules.

Pre-processing The corpus has been pre-processed and re-annotated following the BILOU annotation scheme [16]. Some simplification has been applied for avoiding hops and overlappings that can be found in the keyphrases of the corpus. Finally, in order to avoid conflicts with the offset of the different annotations, a tokenization process based on blank space splitting has been applied. A total of 14 classes resulting from all the possible combinations of the initial entity types and the BILOU annotations have been generated.

Features In this section we present the different attributes considered to be the input of the deep learning stack:

- **Words:** A representation based on pre-trained word embeddings has been used. The word vectors presented in [2] have been selected due to the richness of the sources from which they were generated and to their high recall. These vectors have a total of 300 dimensions and gather around 1,000,653 unique tokens.
- **Part-of-speech:** This feature has been considered due to its importance in general NLP tasks, and particularly to its connection and similarity with the proposed classes to which each keyphrase should be related. The PoS-Tagging model used was the one provided by the CoreNLP [10] library for Spanish. An embedding representation of this feature is learned during training, resulting in 25-dimensional vectors.
- **Casing:** This feature satisfies the need to minimize the impact of the simplification process applied to complex expressions found in the different instances. This is achieved by modeling each term with an additional 8-position one-hot vector which represents different cases: term ending in comma or in dot, uppercased first letter or uppercased term, digits within the term, etc.

Deep Learning model The model implemented for keyphrase detection, as shown in Figure 1, consists of a Bi-LSTM layer followed by two Dense layers. Inputs of the architecture are represented by vectors C_x, P_x and W_x, which represent casing information, POS-tag embedding and word embedding, respectively. The last dense layer corresponds to the output layer.

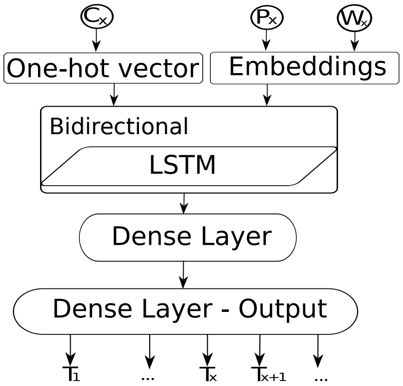

Fig. 1. Architecture of the proposed model for keyphrase extraction and classification.

- **Bi-LSTM:** LSTMs [6] are proven to offer good performance in sequential NLP tasks. This layer responds to the need to process each term according to its context. Each LSTM is configured with 150 neurons and a ReLU [5] as an activation function. In order to avoid over-adjustment, dropouts of 0.5 and recurrent dropouts of 0.3 have been applied.
- **Dense (middle):** This layer has been added in order to simplify the information generated by the previous layers, thus reducing the solution space in subsequent layers.
- **Dense (output):** The output layer is configured with 14 neurons and a softmax activation function.

Rules Two types of rules have been applied to the output of the deep learning architecture in order to perform systematic error correction. The first set of rules is oriented to correcting frequent errors by extending or reducing the scope of a detected keyphrase, or modifying its type, according to casing and POS-tag information. On the other hand, the second set of rules aims to ensure that the final output of the system correctly follows the output BILOU format.

Equations 1 and 2 show examples of rules that can be applied in each of the aforementioned cases, respectively:

$$T_1(O)\ T_2(B|Action)\ T_3(L|Concept) \Rightarrow T_1(O)\ T_2(B|Action)\ T_3(L|Action) \quad (1)$$

$$T_1(O)\ T_2(I)\ T_3(L) \Rightarrow T_1(O)\ T_2(B)\ T_3(L) \quad (2)$$

Equation 1 shows term T_1, labeled as not belonging to an entity (O), term T_2, labeled as the beginning term (B) of an Action entity, and term T_3 labeled as the last term (L) of a Concept entity. In this case, the applied rule transforms the last entity type from Concept to Action, for it to match the type of entity beginning in T_2.

Equation 2, on the other hand, adapts the output to the expected BILOU format: term T_2 is labelled as intermediate term (I) of an entity, while the previous term T_1 is neither an intermediate nor a beginning term. The following term T_3 is the last term of the entity. Hence, for the output to make sense term T_2 must be relabeled as beginning term, so the final entity is composed of T_2 and T_3.

4.2 Attentive Relation Extraction

A different deep learning stack has been developed for the second subtask, devoted to the extraction meaningful relationships between keyphrases. This stack is also based on a Bi-LSTM layer but is enriched by the addition of an attention layer. Pre-processing is also needed in this step for correctly preparing the input data.

Features Some supplementary information has been added to the input features of the model, apart from those features already mentioned in the first subtask (word embeddings, casing and POS-tagging):

- **Entities:** In order to represent the entities that form part of each relationship, the four different entity types (action, concept, predicate and reference) have been represented using embeddings generated during the training phase. The resulting vectors have 25 dimensions and encode both the BILOU annotation and the type of each term of an entity.
- **Dependency graph:** Considering that the relationships to be identified are of a semantic nature (for instance *is-a*, *causes* or *domain*), the information obtained by performing semantic parsing over the sentence and extracting its dependency graph could be very valuable for the main aim of the subtask. This dependency graph has also been generated using the CoreNLP library. Specifically, the information modeled represents the lowest level of relationship that is offered by the graph. Both the direction of the relationship (related term) and the type of relationship are modeled. Both the related term and the type of dependency are mapped using One-hot vectors.

Deep Learning model The proposed model is based on the architecture presented in [21]. This model extends the original approach by including the previously mentioned NLP-based features. As Figure 2 shows, the model makes use of a Bi-LSTM layer followed by an attention layer which considers the output of the previous layer as a whole and merges word information into a higher level vector that attempts to represent the attention relationships at sentence level.

The output of the attention layer is then directed to an output dense layer with all the possible classes a relationship between two terms can classified into. The input is modeled using W_x, P_x and E_x embeddings on one hand, representing word, POS-tag and entity information respectively, and D_x, T_x and C_x One-hot vectors on the other hand, which contain information regarding dependency between terms, type of dependency and casing information respectively.

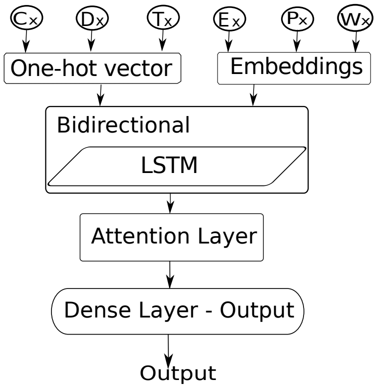

Fig. 2. Architecture of the proposed model for relation extraction.

5 Results

In this section we show results obtained in the *eHealth-KD* competition, as well as some results concerning different configurations of the DeepNER+ARE system. Tables 1, 2 and 3 show the task results for all the participating teams in each of the proposed scenarios.

Team	**F1**	**P**	**R**
TALP	0.6394	0.6506	0.6286
coin_flipper	0.6218	0.7454	0.5334
LASTUS-TALN	0.5816	0.7740	0.4658
NLP_UNED	**0.5473**	**0.6561**	**0.4695**
Hulat-TaskAB	0.5413	0.7734	0.4163
UH-MAJA-KD	0.5189	0.5644	0.4802
lsi2_uned	0.4934	0.7397	0.3702
IxaMed	0.4869	0.6896	0.3763
baseline	*0.4309*	*0.5204*	*0.3677*
Hulat-TaskA	0.4309	0.5204	0.3677
VSP	0.4289	0.4551	0.4056

Table 1. Results from Scenario 1 (whole pipeline), ordered by F1-Measure. Bold highlights our results and italic marks baseline results. F1 stands for F1 measure, P for precision and R for recall.

These results clearly imply that the proposed system *DeepNER+ARE* is able to obtain really promising results particularly in the subtask that aims for the detection and classification of relationships between keyphrases previously found. The system ranks in second place for this phase (subtask B, scenario

Team	F1	P	R
TALP	0.8203	0.8073	0.8336
LASTUS-TALN	0.8167	0.7997	0.8344
UH-MAJA-KD	0.8156	0.7999	0.8320
Hulat-TaskA	0.7903	0.7706	0.8111
coin_flipper	0.7873	0.7986	0.7763
Hulat-TaskAB	0.7758	0.7500	0.8034
NLP_UNED	**0.7543**	**0.8069**	**0.7082**
lsi2_uned	0.7315	0.7817	0.6873
IxaMed	0.6825	0.6567	0.7105
baseline	*0.5466*	*0.5129*	*0.5851*
VSP	0.5466	0.5129	0.5851

Table 2. Results from Scenario 2 (subtask A), ordered by F1-Measure. Bold highlights our results and italic marks baseline results. F1 stands for F1 measure, P for precision and R for recall.

Team	F1	P	R
TALP	0.6269	0.6667	0.5915
NLP_UNED	**0.5337**	**0.6235**	**0.4665**
VSP	0.4933	0.5892	0.4243
coin_flipper	0.4931	0.7133	0.3768
IxaMed	0.4356	0.5195	0.3750
UH-MAJA-KD	0.4336	0.4306	0.4366
LASTUS-TALN	0.2298	0.1705	0.3521
baseline	*0.1231*	*0.4878*	*0.0704*
Hulat-TaskAB	0.1231	0.4878	0.0704
Hulat-TaskA	0.1231	0.4878	0.0704
lsi2_uned	0.1231	0.4878	0.0704

Table 3. Results from Scenario 3 (subtask B), ordered by F1-Measure. Bold highlights our results and italic marks baseline results. F1 stands for F1 measure, P for precision and R for recall.

3). Regarding the whole evaluation, the *DeepNER+ARE* system obtains the fourth position in terms of F1-Measure, while in scenario 2 (subtask A), ranks in seventh place. However, in terms of precision our system is also able to offer the second best performance (over 80%) in this subtask.

The performance of our system is consistent with the complexity of the networks used in each of the phases: for the first subtask, the neural network contains just a Bi-LSTM layer for accurately processing sequential textual information. On the other hand the network used for the second step of the pipeline adds an attention-based layer which is able to improve precision and raise up the F1 measure. This attention mechanism allows the network to merge word-level features into a sentence-level feature vector, which eventually helps the model to focus on specific parts of the input. Furthermore, the use of the graph extracted from the dependency parsing over the input sentence also adds valuable prior information to the network about the possible semantic relationships that can be found in the sentence.

In order to illustrate this behaviour, Table 4 shows the results obtained by our system on the development set provided by organizers in subtask B (relation extraction), as we added dependency parsing information and the attention layer to the original base configuration. This base configuration only considered embeddings, POS-tagging and letter case information as input, as well as the output of subtask A (detected keyphrases and their classes).

System	F1	P	R
Base	0.5918	0.676	0.5263
Base+D	0.6162	0.6836	**0.5608**
Base+D+A	**0.6284**	**0.7396**	0.5463

Table 4. Results (development set) from incremental configurations of our system in subtask B: base system (Base), base system with dependency parsing information (Base+D), and base system, dependency information and attention layer (Base+D+A). F1 stands for F1 measure, P for precision and R for recall. Bold highlights best F1, precision and recall results.

As we can observe, F1-measure globally increases as we add both dependency parsing information and the attention layer to the system. In particular, when the dependency graph is also considered as input, both precision and recall increase, which indicates that this additional information allows the system to find more meaningful relationships. The use of an attention layer, despite causing a small decrease in recall, achieves a higher precision increase (less but more accurate relationships are found) which results in a better F1 measure.

6 Conclusions and Future Work

In this paper we have described our system *DeepNER+ARE*, and its performance in the *eHealth Knowledge Discovery* challenge of the IberLEF 2019 competition. The proposed system is divided into two phases which make use of deep neural networks for addressing the two subtasks of the challenge: detection and classification of keyphrases in biomedical texts, and relation extraction between those keyphrases. The main contribution of our method is the combined use of semantic parsing information and attention-based techniques in the network that performs relation extraction. This contribution is reflected particularly in the results that the system obtains in the relation extraction subtask, in which it is ranked in second position out of 10 participant teams.

We plan to address improvements in the keyphrase extraction as a future lines of work, especially studying how more valuable syntactic and semantic information can be added to the network that performs keyphrase identification, and also how systematic post-processing rules can be automatically extracted from the obtained results. The detection of multi-span and nested keyphrases is also an interesting research line which may lead to increasing the number of keyphrases correctly detected. Regarding relation extraction, more work should be done on modeling the input of the network, in order to feed it with additional and more complex information obtained from the dependency graph, and also on the improvement of the attention mechanisms.

Acknowledgments

Funding: 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. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
2. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (March 2016), `https://crscardellino.github.io/SBWCE/`
3. Fabregat, H., Araujo, L., Martinez-Romo, J.: Deep neural models for extracting entities and relationships in the new rdd corpus relating disabilities and rare diseases. Computer methods and programs in biomedicine **164**, 121–129 (2018)
4. Gligic, L., Kormilitzin, A., Goldberg, P., Nevado-Holgado, A.: Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. arXiv preprint arXiv:1901.01592 (2019)
5. Hahnloser, R.H., Sarpeshkar, R., Mahowald, M.A., Douglas, R.J., Seung, H.S.: Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature **405**(6789), 947 (2000)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
7. Li, F., Zhang, M., Fu, G., Ji, D.: A neural joint model for entity and relation extraction from biomedical text. BMC bioinformatics **18**(1), 198 (2017)
8. López-Ubeda, P., Dıaz-Galiano, M.C., Martın-Valdivia, M.T., Urena-López, L.A.: Sinai en tass 2018 task 3. clasificando acciones y conceptos con umls en medline. In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018) co-located with 34nd SEPLN Conference (SEPLN 2018): Sevilla, Spain, September 18th, 2018. vol. 2172 (2018)
9. Lv, X., Guan, Y., Yang, J., Wu, J.: Clinical relation extraction with deep learning. IJHIT **9**(7), 237–248 (2016)
10. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)

11. Martínez Cámara, E., Almeida Cruz, Y., Díaz Galiano, M.C., Estévez-Velarde, S., García Cumbreras, M.Á., García Vega, M., Gutiérrez, Y., Montejo Ráez, A., Montoyo, A., Muñoz, R., et al.: Overview of tass 2018: Opinions, health and emotions (2018)
12. Martinez-Romo, J., Araujo, L., Duque Fernandez, A.: S em g raph: Extracting keyphrases following a novel semantic graph-based approach. Journal of the Association for Information Science and Technology **67**(1), 71–82 (2016)
13. Medina Herrera, S., Turmo Borras, J.: Joint classification of key-phrases and relations in electronic health documents. In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018) co-located with 34nd SEPLN Conference (SEPLN 2018): Sevilla, Spain, September 18th, 2018. pp. 83–88. CEUR-WS. org (2018)
14. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 1003–1011. Association for Computational Linguistics (2009)
15. Piad-Morffis, A., Gutiérrez, Y., Consuegra-Ayala, J.P., Estevez-Velarde, S., Almeida-Cruz, Y., Muñoz, R., Montoyo, A.: Overview of the ehealth knowledge discovery challenge at iberlef 2019 (2019)
16. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. pp. 147–155. CoNLL '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1596374.1596399
17. Suárez-Paniagua, V., Segura-Bedmar, I., Martınez, P.: Labda at tass-2018 task 3: Convolutional neural networks for relation classification in spanish ehealth documents. In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018) co-located with 34nd SEPLN Conference (SEPLN 2018): Sevilla, Spain, September 18th, 2018. vol. 1510 (2018)
18. Vivaldi Palatresi, J., Rodríguez Hontoria, H.: Tass2018: Medical knowledge discovery by combining terminology extraction techniques with machine learning classification. In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018) co-located with 34nd SEPLN Conference (SEPLN 2018): Sevilla, Spain, September 18th, 2018. pp. 89–95. CEUR-WS. org (2018)
19. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: Kea: Practical automated keyphrase extraction. In: Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, pp. 129–152. IGI Global (2005)
20. Zavala, R.M.R., Martınez, P., Segura-Bedmar, I.: A hybrid bi-lstm-crf model for knowledge recognition from ehealth documents. In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018) co-located with 34nd SEPLN Conference (SEPLN 2018): Sevilla, Spain, September 18th, 2018. vol. 2172 (2018)
21. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 207–212 (2016)