

SINAI at eHealth-KD Challenge 2020: Combining Word Embeddings for Named Entity Recognition in Spanish Medical Records

Pilar López-Úbeda^a, José M. Perea-Ortega^b, Manuel C. Díaz-Galiano^a, M. Teresa Martín-Valdivia^a and L. Alfonso Ureña-López^a

^aSINAI research group, University of Jaén, Spain

^bUniversity of Extremadura, Spain

Abstract

This paper describes the system presented by the SINAI research group to the eHealth-KD challenge at IberLEF 2020. Two main subtasks for knowledge discovery in Spanish medical records were defined: entity recognition and relationship extraction. In the Natural Language Processing (NLP) field, Named Entity Recognition (NER) may be formulated as a sequence labeling problem where the text is treated as a sequence of words to be labeled with linguistic tags. Since current state-of-the-art approaches for sequence labeling typically use Recurrent Neural Networks (RNN), our proposal employs a BiLSTM+CRF neural network where different word embeddings are combined as an input to the architecture. Thus we could test the performance of different types of word embeddings for the NER task in Spanish medical records: own-generated medical embeddings, contextualized non-medical embeddings, and pre-trained non-medical embeddings based on transformers. The obtained results for the entity recognition task achieved the highest F1-score among all the participants, while those obtained for the relationship extraction task show that the proposed approach needs to be further explored.

Keywords

Natural Language Processing, Named Entity Recognition, Relation Extraction, Word Embeddings, Deep Learning, eHealth, Knowledge Discovery, Spanish language

1. Introduction

The volume of digitized data available from the healthcare domain is increasing exponentially. In this scenario, Natural Language Processing (NLP) techniques are becoming essential to discover new knowledge in clinical texts, where valuable information about symptoms, diagnosis, treatments, or drug use is included. Plenty of research has been carried out for clinical text processing for text written in English, but not that much for under-resourced languages such as Spanish. The IberLEF eHealth-KD [1] shared task comes to meet this challenge since two years ago [2, 3], providing different NLP tasks to automatically extract a variety of knowledge from electronic health documents written in Spanish. In 2020, eHealth-KD proposes two subtasks

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: plubeda@ujaen.es (P. López-Úbeda); jmperea@unex.es (J.M. Perea-Ortega); mcdiaz@ujaen.es (M.C. Díaz-Galiano); maite@ujaen.es (M.T. Martín-Valdivia); laurena@ujaen.es (L.A. Ureña-López)

ORCID: 0000-0003-0478-743X (P. López-Úbeda); 0000-0002-7929-3963 (J.M. Perea-Ortega); 0000-0001-9298-1376 (M.C. Díaz-Galiano); 0000-0002-2874-0401 (M.T. Martín-Valdivia); 0000-0001-7540-4059 (L.A. Ureña-López)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

related to capturing the semantic meaning of health-related sentences: entity recognition (subtask A), whose goal is to identify all the entities in a document and their types; and relation extraction (subtask B), which seeks to identify all relevant semantic relationships between the entities recognized [1].

In recent years, deep learning-based methods have been widely used for some NLP-related tasks, such as Named Entity Recognition (NER) or Part-of-Speech (PoS) tagging. These tasks are commonly approached as a sequence labelling problem, where the text is treated as a sequence of words to be labeled with linguistic tags. Current state-of-the-art approaches for sequence labeling propose the use of Recurrent Neural Networks (RNNs) to learn useful representations automatically since they facilitate modeling long-distance dependencies between the words in a sentence. These networks usually rely on word embeddings, which represent words as vectors of real numbers. There are different types of word embeddings (classical [4, 5], character-level [6, 7], or contextualized [8, 9]) that are commonly pre-trained over very large corpora to capture latent syntactic and semantic similarities between words. A novel type of word embeddings called *contextual string embeddings* is proposed by Akbik et al. [10], which essentially model words as sequences of characters, thus contextualizing a word by their surrounding text and allowing the same word to have different embeddings depending on its contextual use.

This paper describes the system presented by the SINAI team for both subtasks of the eHealth-KD 2020: entity recognition and relation extraction. Our group has experience in NER task in the biomedical domain using different methodologies such as traditional machine learning [11], RNNs [12] and unsupervised machine learning [13, 14], among others. For this challenge, our proposal is based on RNNs or, more precisely, the bidirectional variant of Long Short Term Memory along with a stacked Conditional Random Fields decoding layer (BiLSTM+CRF) [15, 6]. This specialized architecture is chosen to approach NER because it facilitates the processing of arbitrary length input sequences and enables the learning of long-distance dependencies, which is particularly advantageous in the case of the focused NER task. Besides, our approach proposes the combination of different types of word embeddings by concatenating each embedding vector to form the final word vectors. In this way, the probability of recognizing a specific word (entity) in a text should be increased as different types of representation of that word are combined. For the case of contextual string embeddings, since they are robust in face of misspelled words, we suppose they could be highly suitable for NER in clinical texts.

In the next section, we describe the architecture of our deep learning-based model along with the different word embeddings used as an input for such a model. Furthermore, we explain how the learning was performed by using the datasets provided by the eHealth-KD shared task. Section 3 reports the results obtained for the different evaluation scenarios and, finally, we conclude with a brief discussion in Section 4 and the conclusions in Section 5.

2. System Description

The approach used to address both subtasks of the eHealth-KD challenge (NER and relation extraction) is based on deep learning by implementing the RNN proposed by Huang et al. [15]. Specifically, we have used a Bidirectional Long-Short Term Memory (BiLSTM) with a sequential Conditional Random Field layer (CRF). BiLSTM is an extension of traditional LSTM

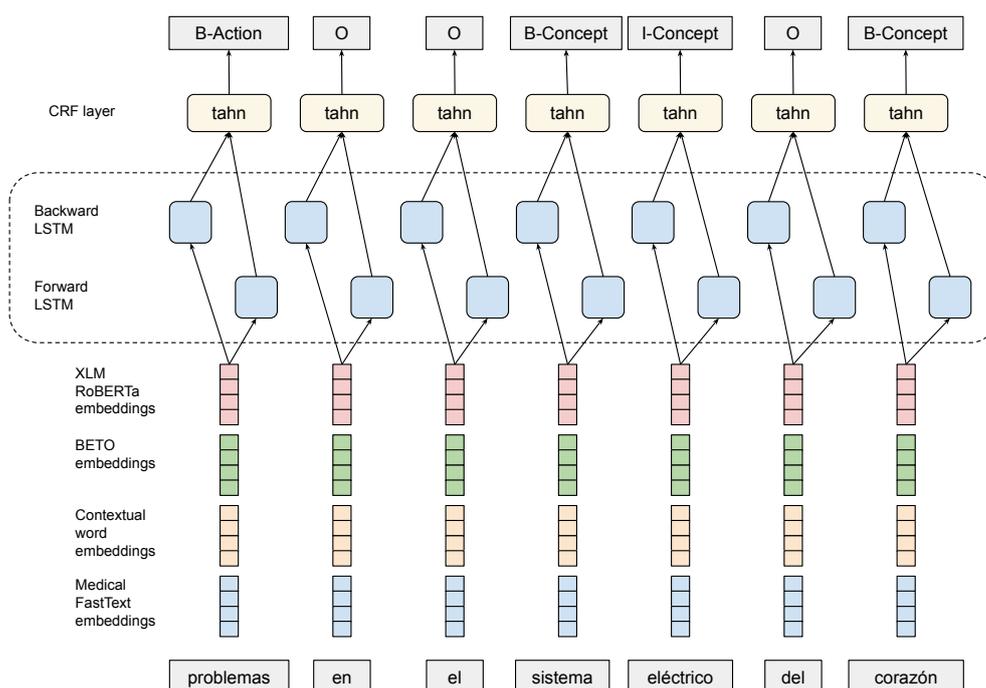


Figure 1: Proposed approach based on a BiLSTM+CRF neural network that uses a combination of different word embeddings as an input layer.

that improves the model performance on sequence classification problems [16]. On the one hand, the main goal of BiLSTM is to split the neurons of a regular LSTM into two directions, one for positive time direction (forward states), and another for negative time direction (backward states). Using two time directions, input information from the past and future of the current time frame can be used to better understand the context of the medical report. On the other hand, a CRF layer is represented by lines that connect consecutive output layers, so that allows us to efficiently use past and future tags to predict the current tag, which is similar to the use of past and future input features via a BiLSTM network.

Recurrent Neural Networks generally use an embedding layer as an input, which makes it possible to represent words and documents using a dense vector representation. Thus the position of a word within the vector space is learned from the text allowing words that are used in a similar way to have a similar representation. In order to fit the text input into the deep neural network structure, we combine different types of word embeddings: own-generated biomedical word embeddings for Spanish, contextual word embeddings, and pre-trained word embeddings based on transformers. Figure 1 shows the proposed architecture based on a BiLSTM+CRF neural network that uses a combination of different word embeddings as an input layer.

2.1. Word Embeddings

Different word embeddings have been combined to form the input layer to the proposed BiLSTM+CRF neural network:

2.1.1. Medical FastText embeddings

Although there are available biomedical word embeddings for Spanish [17, 18, 19], we have tried to generate new ones from existing corpora related to the biomedical domain in Spanish. For this purpose, firstly we extracted the Spanish corpus from MeSpEN [20]. Then, extra information in Spanish from different clinical information websites such as Mayo Clinic [21], World Health Organization [22], and WebMD [23] was added to the former corpus. The pre-processing carried out to train the word embeddings consisted of converting the text to lowercase, removing the URLs, and removing the multi-lines. Finally, FastText [24] was used to perform the training by applying the following setup: skip-gram model, 0.05 for the learning rate, size of 300 for the word vectors, 10 for the number of epochs, and 5 for the minimal number of word occurrences.

2.1.2. Contextual Word Embeddings

Contextual word embeddings are considered powerful embeddings because they capture latent syntactic-semantic information that goes beyond standard word embeddings [10]. These embeddings are based on character-level language modeling and their use is particularly advantageous when the NER task is approached as a sequential labeling problem. For our experiments, we have used the *pooled contextualized embeddings* proposed by Akbik et al. [25] for NER. In that work, the authors propose an approach in which a pooling operation is applied to distill a global word representation from all contextualized instances of that word, thus producing evolving word representations that change over time as more instances of the same word are observed in the data.

2.1.3. Pre-trained Word Embeddings Based on Transformers

Finally, we have combined pre-trained word embeddings based on *transformers* [26], a popular attention mechanism used in transfer learning approaches. Transfer learning [27] is based on applying the knowledge learned from previous tasks to a new task domain. This kind of embeddings are commonly pre-trained over very large corpora to capture latent syntactic and semantic similarities between words. For our experiments, we have used two specific pre-trained word embeddings: BETO [28], which follows a BERT model trained on a big Spanish corpus, and XLM-RoBERTa [29], which were generated by using a large multilingual language model trained on 2.5 TB of filtered CommonCrawl data.

2.2. Training Phase

The task organizers provided several datasets (training, validation and test) to allow the participants to train their systems properly. Besides, an unreviewed dataset named "ensemble" that contained the submissions from past editions, and a dataset called "transfer" from the Spanish

Los	O	Los	O
tumores	B-Concept	médicos	O
malignos	B-Concept	diagnostican	B-entails
de	O	talasemias	O
glándulas	B-Concept	mediante	O
suprarrenales	I-Concept	pruebas	I-entails
son	O	de	I-entails
poco	O	sangre	I-entails
comunes	B-Concept	.	O
.	O		

Figure 2: Example of BIO tagging scheme for entity and relationship recognition.

version of Wikinews were provided as well. For scenarios 1, 2 and 3, we used the training, ensembled and transfer (800 + 3000 + 100 sentences respectively) datasets for training, while the development set (200 sentences) was used to validate our system. However, for scenario 4, the training, ensembled and development datasets (800 + 3000 + 200 sentences, respectively) were used for training, while the transfer set (100 sentences) was used to validate the system.

For all the scenarios, each sentence was first tokenized. For the entity recognition task, the provided annotations were encoded by using the BIO tagging scheme. Thus each token in a sentence was labeled with O (non-entity), B (beginning token of an entity), or I (inside token of an entity). This scheme is the most popular in the NER task although it presents problems when the entity contains discontinuous tokens. Figure 2 shows an example of the encoded annotation carried out for both tasks.

For the relationship extraction task, we decided to generate the same neural network to train each of the 13 semantic relationship defined for the task.

Therefore the only input linguistic features used for the training phase were the tokens (original words) of each sentence and the generated BIO tags.

3. Experimental Setup and Results

For generating the different learning models we employed Flair as a NLP framework [30]. We used the default hyperparameter setting in Flair with the following configuration: learning rate as 0.1, batch size as 32, dropout probability as 0.5 and maximum epoch as 150. All experiments were performed on a single Tesla-V100 32 GB GPU with 192 GB of RAM.

The SINAI team submitted 3 runs for each proposed scenario, where each run represents a different combination of word embeddings, as described in Section 2:

- Run 1: Medical FastText embeddings + Contextual word embeddings.
- Run 2: Medical FastText embeddings + Contextual word embeddings + BETO embeddings.
- Run 3: Medical FastText embeddings + Contextual word embeddings + BETO embeddings + XLMRoBERTa embeddings.

Table 1

Evaluation results obtained by the SINAI team at IberLEF2020’s eHealth-KD.

Scenario	Run	Precision	Recall	F1-score
1	1	0.6515	0.3106	0.4207
	2	0.6451	0.3005	0.41
	3	0.6446	0.3014	0.4107
2	1	0.8303	0.8094	0.8197
	2	0.8432	0.7932	0.8174
	3	0.8446	0.8067	0.8252
3	1	0.6145	0.325	0.4252
	2	0.6271	0.3654	0.4617
	3	0.6124	0.3615	0.4547
4	1	0.6366	0.1685	0.2664
	2	0.6263	0.1813	0.2812
	3	0.611	0.1795	0.2775

The metrics defined by the eHealth-KD challenge to evaluate the submitted experiments are those commonly used for some NLP tasks such as NER or text classification, namely precision, recall, and F1-score. Table 1 shows the results obtained by the SINAI team for each scenario and run submitted.

As shown in Table 1, the results obtained for scenario 4 are low (28.12% of F1-score), although this behavior may be due to the limited amount of documents available to train an alternative domain. For scenario 3 (subtask B), we achieved our best F1-score with run 2 (46.17%), which means that the XLMRoBERTa embeddings did not provide valuable information during the learning phase. Regarding scenario 1 (main evaluation), we achieved a relatively poor F1-score of 42.07%, which is somehow in line with the behavior obtained for scenario 3, where the use of the pre-trained embeddings based on transformers have not improved the performance. Finally, it should be noted that the F1-score achieved by our run 3 submitted for scenario 2 (82.52%) outperformed the rest of the competition runs for that scenario among all the participants.

4. Discussion

Based on the obtained results, we found that the combination of word embeddings generally provides an extra value for the learning phase of the neural network. In the case of the entity recognition (subtask A, scenario 2), the combination of all embeddings (run 3) does improve all the metrics obtained when fewer types of embeddings are combined. It should be noted that our starting point (run 1) combines embeddings that were generated from the biomedical domain and for which we already obtained a high F1-score. Therefore it can be considered positive to have slightly improved those results when other non-biomedical embeddings were added.

A fine grained evaluation of these systems can be defined in terms of comparing the response

Table 2

Evaluation results obtained by the SINAI team in terms of comparing the response of the system against the golden annotation.

Scenario	Run	Task A					Task B		
		correct	incorrect	partial	missing	spurious	correct	missing	spurious
1	1	250	15	21	288	46	75	431	108
	2	250	18	19	287	44	65	441	107
	3	250	18	19	287	44	66	440	108
2	1	436	30	28	62	48	x	x	x
	2	424	27	34	71	38	x	x	x
	3	432	24	33	67	42	x	x	x
3	1	x	x	x	x	x	169	351	106
	2	x	x	x	x	x	109	330	113
	3	x	x	x	x	x	188	332	119
4	1	348	33	21	808	48	47	1150	140
	2	350	20	25	815	46	74	1123	182
	3	345	27	26	812	50	74	1123	185

of the system against the golden annotation [31]. Then, the evaluation of our system considering these different categories of errors is shown in Table 2.

As shown in Table 2, our system performs well in the first task (entity recognition) and scenario 2, but it has some performance shortcomings for the relation extraction task. For future work, we should focus on finding out why the system is not correctly capturing the relationships between concepts. Finally, we would also like to highlight the behavior of our system in the alternative scenario (scenario 4), in which it could not be able to recognize entities that were not related to the medical domain. We believe that it could be due to the use of word embeddings specifically pre-trained in that domain.

5. Conclusions

This paper presents the participation of the SINAI research group in the eHealth-KD challenge at IberLEF 2020 (Iberian Languages Evaluation Forum), where two main NLP tasks involving the Spanish language were defined: entity recognition and relationship extraction. Several challenge scenarios were also proposed where different NLP subtasks were evaluated.

Our proposal follows a deep learning-based approach for Named Entity Recognition (NER) in Spanish health documents. It is focused on the use of a BiLSTM+CRF neural network where different word embeddings are combined as an input to the architecture. Then this neural network is trained by using the annotated datasets provided by the organization, which were previously tokenized and tagged by using the BIO scheme. Our main goal was to prove the performance of different types of word embeddings for the NER task in the medical domain:

own-generated medical embeddings, contextualized non-medical embeddings, and pre-trained non-medical embeddings based on transformers. The obtained results for the entity recognition task achieved the highest evaluation score among all the participants, achieving 82.52% of F1-score, 84.46% of precision and 80.67% of recall. However, our proposal revealed certain weaknesses for the relationship extraction task, any of the existing relationships in the gold test have not been captured by our system which means a loss in recall, we need analyze in detail if the problem lies in the BIO notation of the relationship because using this notation when the system finds a beginning of relationship it is associated with the end of the closest relationship annotated, so the approach used will need to be further explored.

For future work, we will study the performance of using more linguistic features such as Part-Of-Speech tags as an input in the neural network, as well as the use of ontologies related to the biomedical domain and other types of word embeddings. Furthermore, we will try to improve the extraction of relationships by implementing another neural network that captures in a better way the relationships between concepts.

Acknowledgments

This work has been partially supported by LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government, Junta de Extremadura (GR18135) and Fondo Europeo de Desarrollo Regional (FEDER).

References

- [1] A. Piad-Morffis, Y. Gutiérrez, H. Cañizares-Díaz, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, Spain, September, 2020., 2020.
- [2] A. Piad-Morffis, Y. Gutiérrez, S. Estévez-Velarde, Y. Almeida-Cruz, A. Montoyo, R. Muñoz, Analysis of eHealth knowledge discovery systems in the TASS 2018 Workshop, *Procesamiento de Lenguaje Natural (2019)*. doi:10.26342/2019-62-1.
- [3] A. Piad-Morffis, Y. Gutiérrez, J. P. Consuegra-Ayala, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth knowledge discovery challenge at IberLEF 2019, in: *CEUR Workshop Proceedings*, 2019.
- [4] J. Pennington, R. Socher, C. D. Manning, GloVe: Global vectors for word representation, in: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014. doi:10.3115/v1/d14-1162.
- [5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: *Advances in Neural Information Processing Systems*, 2013. arXiv:1310.4546.
- [6] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF, in: *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016. doi:10.18653/v1/p16-1101. arXiv:1603.01354.

- [7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, 2016. doi:10.18653/v1/n16-1030. arXiv:1603.01360.
- [8] M. E. Peters, W. Ammar, C. Bhagavatula, R. Power, Semi-supervised sequence tagging with bidirectional language models, in: ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2017. doi:10.18653/v1/P17-1161. arXiv:1705.00108.
- [9] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2018. doi:10.18653/v1/n18-1202. arXiv:1802.05365.
- [10] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.
- [11] P. L. Úbeda, M. C. D. Galiano, M. T. Martín-Valdivia, L. A. U. Lopez, Using machine learning and deep learning methods to find mentions of adverse drug reactions in social media, in: Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task, 2019, pp. 102–106.
- [12] P. L. Úbeda, M. C. D. Galiano, L. A. U. Lopez, M. T. Martín-Valdivia, Using Snomed to recognize and index chemical and drug mentions, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 115–120.
- [13] P. López-Ubeda, M. C. Díaz-Galiano, M. T. Martín-Valdivia, L. A. Urena-López, Sinai en tass 2018 task 3. clasificando acciones y conceptos con umls en medline, Proceedings of TASS 2172 (2018).
- [14] P. López-Úbeda, M. C. Díaz-Galiano, A. Montejo-Ráez, M.-T. Martín-Valdivia, L. A. Ureña-López, An Integrated Approach to Biomedical Term Identification Systems, Applied Sciences 10 (2020) 1726.
- [15] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).
- [16] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (1997) 2673–2681.
- [17] F. Soares, M. Villegas, A. Gonzalez-Agirre, M. Krallinger, J. Armengol-Estapé, Medical word embeddings for Spanish: Development and evaluation, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 124–133. URL: <https://www.aclweb.org/anthology/W19-1916>. doi:10.18653/v1/W19-1916.
- [18] S. Santiso, A. Casillas, A. Pérez, M. Oronoz, Word embeddings for negation detection in health records written in Spanish, Soft Computing (2019). doi:10.1007/s00500-018-3650-7.
- [19] I. Segura-Bedmar, P. Martínez, Simplifying drug package leaflets written in Spanish by using word embedding, Journal of Biomedical Semantics (2017). doi:10.1186/

s13326-017-0156-7.

- [20] M. Villegas, A. Intxaurre, A. Gonzalez-Agirre, M. Marimon, M. Krallinger, The MeSpEN resource for English-Spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations, LREC MultilingualBIO: Multilingual Biomedical Text Processing (Malero M, Krallinger M, Gonzalez-Agirre A, eds.) (2018).
- [21] Mayo clinic, 1998-2020. URL: <https://www.mayoclinic.org/es-es>.
- [22] Organización mundial de la salud, 2020. URL: <https://www.who.int/es>.
- [23] Webmd - better information. better health., 2005-2020. URL: <https://www.webmd.com/>.
- [24] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.
- [25] A. Akbik, T. Bergmann, R. Vollgraf, Pooled contextualized embeddings for named entity recognition, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 724–728.
- [26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
- [27] S. Thrun, Is learning the n-th thing any easier than learning the first?, in: Advances in neural information processing systems, 1996, pp. 640–646.
- [28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [29] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).
- [30] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, pp. 54–59.
- [31] N. Chinchor, B. Sundheim, MUC-5 evaluation metrics, in: Proceedings of the 5th conference on Message understanding, Association for Computational Linguistics, 1993, pp. 69–78.