Contextual Representations and Semi-Supervised Named Entity Recognition for Portuguese Language

Pedro Vitor Quinta de Castro[1], Nádia Félix Felipe da Silva[1], and Anderson da Silva Soares[1]

[1]Universidade Federal de Goiás, Goiânia GO 74690-900, Brazil
I.pedrovitorquinta@inf.ufg.br, II.nadia@inf.ufg.br,
III.anderson@inf.ufg.br

Abstract. Named Entity Recognition is a Natural Language Processing task which is difficult to adapt across different domains. In this work, we propose a Semi-Supervised approach using Deep Learning models in order to support three different domains for the Portuguese language: general, police and medical. We perform the self-training of a model with an architecture based on a Bidirectional Long Short-Term Memory network with a Conditional Random Fields sequential classifier, using five Portuguese corpora. The word representations of the proposed model are contextual and provided by ELMo's language model. The results achieve a competitive performance in the IberLEF evaluation forum.

Keywords: Natural Language Processing · Named Entity Recognition · Deep Learning · Neural Networks · Portuguese Language

1 Introduction

Information Extraction (IE) is the process of obtaining structured data from sources which can not be interpreted directly by machines, like texts [23]. This is particularly important considering the amount of textual information which is exchanged every minute on the internet [34]. Named Entity Recognition (NER) is the Natural Language Processing (NLP) task which focus on identifying and classifying named entities from this unstructured textual information, making them interpretable and accessible to different communication channels.

When dealing with multiple domains, a NER prediction model needs to be able to handle not only the difference of lexicon between them, but also the difference of morphological features. This adds an additional layer of complexity to this task, requiring a more scalable model to perform well in this challenge.

This paper describes our participation in IberLEF (Iberian Languages Evaluation Forum), Task 1: Named Entity Recognition [31]. We present a system based on different deep learning architectures for both NER model and word representations. We propose a semi-supervised training in order to deal with the different domains targeted by the evaluation.

2 Related Work

The first Deep Learning architectures to be applied in NER models were based on CNNs [5, 32], and later on Recurring Neural Networks (RNN)[9, 11, 4, 17, 22]. The reason why Deep Learning models perform well on NLP tasks is because they learn latent features from words, as well as the interactions between them, during the training of specific tasks, such as NER.

Collobert et al. [5] proposed a model based on a Multilayer Perceptron with a convolutional layer, and the following works for NER were mostly based on bidirectional LSTMs, with a few differences between them. Huang et al. [11] used a biLSTM-CRF network with manually selected features, combined with features from SENNA [5] word embeddings. Chiu and Nicols [4] used a biLSTM model without the CRF layer for classification, and had their best results with character level features extracted from a CNN layer, concatenated with SENNA embeddings. Lample et al. [17] and Ma and Hovy [22] used similar approaches based on biLSTM-CRF models, with the difference that [17] used a biLSTM to extract character level features, combined with Word2Vec [24] representations, while [22] used a CNN to extract the character level features, that were combined with GloVe [29] embeddings. These works show that biLSTM-CRF networks became a standard architecture for NER models (as well as for other NLP sequential classification tasks). Following works focused on representation of the words, instead of the actual NER model. Language models have been the primary architecture for contextualized word representations.

Peters et al. [30], Devlin et al. [7] and Akbik et al. [1] developed different architectures for contextual word representations based on bidirectional language models and evaluated their performance on the NER task (as well as on other NLP tasks). Both papers, [30] and [1] used a biLSTM-CRF baseline NER model for evaluating their representation models, while [7] evaluated his model by adding a neural layer to the language model, performing the NER classification with it. The ELMo (Embeddings from Language Model) representations from [30] are provided by the biLM language model, which is based on 2 biLSTM networks, with 2 layers each, and the model's input is a character level representation provided by a CNN network. In another way, [7] created BERT, a language model based on the Transformer [36] architecture, which is based only on the neural mechanism of attention. The author from [1] created a language model on character level, in a way that his objective was not to predict words, but characters. The architecture of his CharLM model is also based on a biLSTM network. Table 1 lists the models presented on this section with their respective F-Score performance on the English benchmark from CoNLL-2003 [35].

For Portuguese language, the first work that used a Deep Learning approach was from Dos Santos and Guimarães [32], who adapted the architecture from [5]

Work	Benchmark	F-Score	Year
Akbik et al. [1]	CoNLL-2003	**93.09%**	2018
Devlin et al. (*BERT Large*) [7]	CoNLL-2003	92.80%	2018
Devlin et al. (*BERT Base*) [7]	CoNLL-2003	92.40%	2018
Peters et al. [30]	CoNLL-2003	92.22%	2018
Quinta de Castro et al. [3]	HAREM-Sel	**76.27%**	2018
	HAREM-Tot	**70.33%**	
Da Costa e Paetzold [6]	HAREM-Tot	69.14%	2018
Chiu and Nichols [4]	CoNLL-2003	91.62%	2016
Ma and Hovy [22]	CoNLL-2003	91.21%	2016
Lample et al. [17]	CoNLL-2003	90.94%	2016
Huang et al. [11]	CoNLL-2003	90.10%	2015
Dos Santos e Guimarães [32]	HAREM-Sel	71.23%	2018
	HAREM-Tot	65.41%	
Collobert et al. [5]	CoNLL-2003	89.59%	2011

Table 1. NER models using Deep Learning architectures for English and Portuguese languages, both evaluated using the CoNLL script [35]. The English language results are reported on the CoNLL-2003 [35] benchmark, and the Portuguese ones are reported on the HAREM [33] benchmark.

and proposed CharWNN. For this work, besides using character level features from CNN, the authors also used word embeddings that were pre-trained using the Word2Vec tool [38]. Da Costa and Paetzold [6] and Quinta de Castro et al. [3] used a BiLSTM-CRF architecture with minor differences between them. [6] concatenated character level features from a BiLSTM network with FastText [13] word embeddings, prior to passing this concatenation through another BiLSTM network. [3] used a similar approach from [17] and concatenated the character level features from a BiLSTM network with the representations of a second BiLSTM, which processed pre-trained Wang2Vec [20] embeddings.

3 Proposed Model

In this work, we propose a system based on different deep learning architectures, similar to that was used by [30]: a Bidirectional Long Short-Term Memory (BiLSTM)[10] NER model with a Conditional Random Fields (CRF)[16] sequential classifer; fed by the contextual word representations from an ELMo [30] language model, combined with character level representations from a Convolutional Neural Network (CNN) [8, 18]. Our system differs from [30] in the way that we do not use pre-trained word embeddings, and we use two different ELMo models, one for the general domain of Portuguese language, and one for the police domain.

The ELMo embeddings are obtained using the biLM (bidirectional Language Model) [30] architecture. This architecture is based on 2 BiLSTM networks, each of them responsible for one direction in the bidirectional language model: one for keeping a representation while making predictions in the forward direction of the text and one for the reverse direction. The first layer from the biLM model produces character level features from the training words using two CNNs, one for each direction of the text, each of them with 2048 convolutional filters. They produce a representation with a total dimension of 4096, which is fed to the first BiLSTM layer of the biLM model. Each layer of the model (the CNN and the two BiLSTMs) projects the input it receives to a vector of dimension 1024. These 3 projections represent the ELMo embeddings which are produced by the biLM model. The size of the biLM training vocabulary determines the amount of words that will be predicted in the Softmax layer of the model, as shown in figure 1.

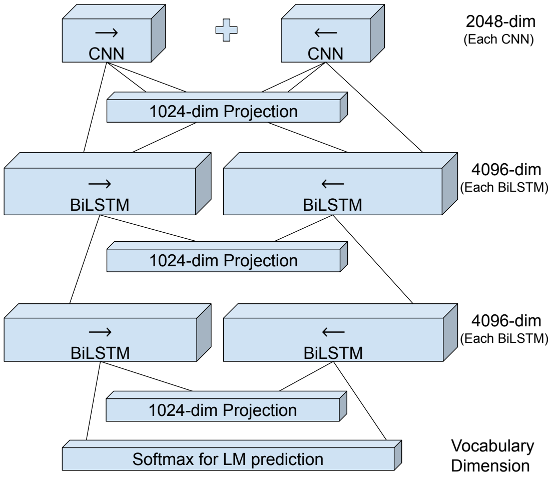

Fig. 1. Layer representations of the biLM architecture and their connections between layers and projections. Note that the arrows → e ← in the LSTM layers indicate the direction of the objective function from the bidirectional language model, not the direction of the LSTM networks, which are also bidirectional. Each 2-layer BiLSTM network used in this scheme works as a unidirectional language model, and their composition provides bidirectionality to the whole language model.

The BiLSTM-CRF architecture used in this work is the same from the AllenNLP *framework* [2], following a parameterization similar to the one described in [30] for the NER task. The CNN network used for producing character level features from words used embeddings with dimension 16 and 128 convolutional filters of size 3, with the ReLU [12, 26] activation function. The BiLSTM network used for encoding the words has 2 layers, with 200 hidden units each. Figure 2 shows the dimensionality of the word representations obtained from the CNN and the 2 ELMo embeddings used. The 2 ELMo we use were trained in two separate domains: for the general Portuguese domain we used a Portuguese Wikipedia [37] dump, and for the police domain we used a 1.6 billion word *corpus* created from public documents from Brazil's Labor Courts [15]. The Portuguese ELMo model we trained is publicly available at https://allennlp.org/elmo. For the IberLEF evaluation, we performed the fine tuning of this ELMo in this combined dataset, following [30].

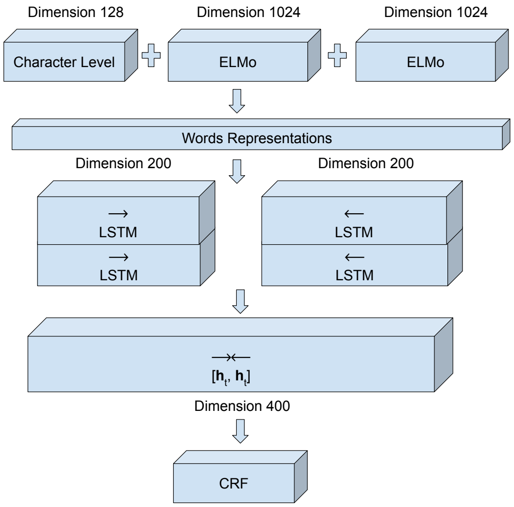

Fig. 2. Representation of words in the proposed architecture

4 Experimental Setup and Results

For the Portuguese NER task, IberLEF specified the evaluation of models in three different domains: general, police and clinical. For the specific domains only person names (PER category) are annotated, while the general domain dataset is annotated with 5 different categories: person, place (PLC), organization (ORG), value (VAL) and time (TME). The following public *corpora* were used for the model proposed in this work: WikiNER [27], LeNER-Br [21], HAREM I [33] and MiniHAREM [28] golden collections, and Paramopama [14]. We also used a private legal *corpus* provided by the Datalawyer company, consisting of 76 annotated documents from the Brazilian Labor Court. The only dataset annotated with all five categories is HAREM. These *corpora* have the following categories annotated in them:

- HAREM: Place, Organization, Person, Time, Value, Abstraction, Work, Event, Thing and Other;
- LeNER-Br: Legal Case, Law, Place, Organization, Person and Time;
- Paramopama: Place, Organization, Person and Time;
- WikiNER: Place, Miscellaneous, Organization and Person;
- Datalawyer: Function, Legal Basis, Place, Organization, Person, Court, Settlement Value, Pleed Value, Conviction Value, Court Costs and District.

Since only the HAREM datasets contains all the categories needed for the IberLEF evaluation, we adopted a semi-supervised approach training for an initial NER model to perform the self-training of the final model. This training had the following procedure:

1. For each one of the datasets, we ignored all the entities that were not annotated as one of the 5 relevant categories for this evaluation. Their annotation was removed;
2. We merged the datasets from HAREM, LeNER-Br and Paramopama, and randomly split them into training, validation and test sets;
3. The resulting datasets from the previous step were used to train a NER model for bootstrapping Time and Value annotations for the datasets that didn't contain these categories;
4. The bootstrap model was used to annotate:
   - 4.1. Time and Value entities in the WikiNER dataset;
   - 4.2. Value entities in the LeNER-Br dataset;
   - 4.3. Value entities in the Paramopama dataset;
   - 4.4. Time and Value entities in the Datalawyer dataset.
5. The resulting boostrapped *corpora* were merged and split into training, validation and test sets;
6. The resulting datasets from the previous step were used to train the final NER model that was submitted to the IberLEF evaluation.

None of the existing annotations was removed or overriden during the bootstrapping of the datasets. Only words that prior to this process had no category associated to them were classified as either Time or Value, according to the bootstrap model.

4.1 Models Evaluation

Prior to submitting the NER model with word representations from 2 ELMo and a CNN (henceforth referred to as 2xELMo+CNN), we performed the training of two other models, with different types of word representation: (i) ELMo+CNN and (ii) ELMo+CNN+Wang2Vec [19]. These two models use only the general domain ELMo. We performed the training of these three models using the same configuration, and performed an additional evaluation of them in the following datasets: MiniHAREM, test datasets from Datalawyer Company and LeNER-Br, and the full datasets from Paramopama and WikiNER. For all of them, except MiniHAREM, we evaluated both variants: with and without bootstrapped Time and Value entities. The best model with the best F-Score was ELMo+CNN+Wang2Vec, followed by 2xELMo+CNN.

We also evaluated the three models in all nine datasets (MiniHAREM, Datalawyer, LeNER-Br, Paramopama and WikiNER, with these last four being evaluated in the original dataset, and the bootstrapped dataset). The 2xELMo+CNN had the best results for the MiniHAREM dataset, as well as for the datasets in the police domain (Datalawyer and LeNER-Br datasets). ELMo+CNN had the best results for Paramopama and WikiNER. After grouping these evaluation results by model, the best mean F-Score was from the 2xELMo+CNN variant. Since 2xELMo+CNN performed better in the police domain (which is relevant for the IberLEF evaluation), we chose this model for the task evaluation.

Table 2 presents the results obtained from the IberLEF evaluation. We point out that the only *corpus* we did not use from HAREM to train our models was the one from HAREM II [25], which is the one used in the general domain evaluation. We also did not have any access to any type of clinical documents or embeddings, so our model contained no type of adaptation for this specific domain.

Corpus	Category	Precision	Recall	F-Score
Police Dataset	Person	86.14%	92.82%	89.35%
Clinical Dataset	Person	32.47%	51.02%	39.68%
General Dataset (SIGARRA + HAREM II)	Overall	63.11%	51.69%	56.83%

Table 2. Results from the IberLEF evaluation, for the 3 different domains.

5 Concluding Remarks

For the Portuguese NER task of the Iberian Languages Evaluation Forum, we experimented with different systems based on deep learning architectures, for both NER model and word representations. For the NER model we used the BiLSTM-CRF architecture, which became a reference for sequential classification NLP tasks. For word representations we experimented with character level features from Convolutional Neural Networks, Wang2Vec pre-trained word embeddings, and the ELMo embeddings from a biLM language model. We evaluated different models with different types of word representations in 5 different *corpora*, and submitted a system based on 2 different ELMo, combined with character level features. Our model was trained in a semi-supervised scenario, in order to account for the lack of certain types of categories in the used *corpora*.

Our main contribution is the use of ELMo embeddings for the Portuguese NER task, which have not been reported so far in the related literature. Our pre-trained ELMo model is publicly available at https://allennlp.org/elmo.

For future work, instead of training a single NER model with different ELMo representations for different domains, we will experiment with an ensemble of different models, each one trained separately in a different domain.

6 Acknowledgements

Thanks to Datalawyer (https://www.datalawyer.com.br/) for the financial support and for providing the legal dataset used for training the submitted model. This work was developed in Deep Learning Brazil research group. Our researches are sponsored by Copel Energy Distribution, Data-H Artificial Intelligence, CyberLabs Artificial Intelligence, Americas Health and iFood Food Delivery.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: COLING 2018, 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018)
2. AllenNLP: An open-source nlp research library, built on pytorch. (2018), https://allennlp.org/, [Online; accessed 06-July-2019]
3. Quinta de Castro, P.V., Félix Felipe da Silva, N., da Silva Soares, A.: Portuguese named entity recognition using lstm-crf. In: Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., Paetzold, G.H. (eds.) Computational Processing of the Portuguese Language. pp. 83–92. Springer International Publishing, Cham (2018)
4. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics **4**, 357–370 (Dec 2016), https://www.aclweb.org/anthology/Q16-1026
5. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**, 2493–2537 (Nov 2011), http://dl.acm.org/citation.cfm?id=1953048.2078186
6. da Costa, P., Paetzold, G.H.: Effective sequence labeling with hybrid neural-crf models. In: Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., Paetzold, G.H. (eds.) Computational Processing of the Portuguese Language. pp. 490–498. Springer International Publishing (2018)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
8. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics **36**(4), 193–202 (Apr 1980), https://doi.org/10.1007/BF00344251
9. Graves, A., rahman Mohamed, A., Hinton, G.E.: Speech recognition with deep recurrent neural networks. CoRR **abs/1303.5778** (2013)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997), http://dx.doi.org/10.1162/neco.1997.9.8.1735
11. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. CoRR **abs/1508.01991** (2015)
12. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision. pp. 2146–2153 (2009)
13. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (Apr 2017)
14. Júnior, C.M., Macedo, H., Bispo, T., Santos, F., Silva, N., Barbosa, L.: Paramopama: a Brazilian-Portuguese Corpus for Named Entity Recognition. Tech. rep., Universidade Federal de Sergipe (2015)
15. de Justiça, C.N.: Processo judicial eletrônico (pje) (2019), http://www.cnj.jus.br/tecnologia-da-informacao/processo-judicial-eletronico-pje, [Online; accessed 06-July-2019]
16. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc. (2001)

17. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. Association for Computational Linguistics (2016)
18. Le Cun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Proceedings of the 2Nd International Conference on Neural Information Processing Systems, pp. 396–404. NIPS'89, MIT Press, Cambridge, MA, USA (1989), http://dl.acm.org/citation.cfm?id=2969830.2969879
19. Ling, W., Dyer, C., Black, A., Trancoso, I.: Extension of the original word2vec using different architectures. URL: https://github.com/wlin12/wang2vec
20. Ling, W., Dyer, C., Black, A.W., Trancoso, I.: Two/too simple adaptations of Word2Vec for syntax problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1299–1304. Association for Computational Linguistics (2015)
21. Luz de Araujo, P.H., de Campos, T.E., de Oliveira, R.R.R., Stauffer, M., Couto, S., Bermejo, P.: Lener-br: a dataset for named entity recognition in brazilian legal text. In: International Conference on the Computational Processing of Portuguese (PROPOR). pp. 313–323. Lecture Notes on Computer Science (LNCS), Springer, Canela, RS, Brazil (September 24-26 2018)
22. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1064–1074. Association for Computational Linguistics (2016)
23. Maynard, D., Bontcheva, K., Augenstein, I.: Natural language processing for the semantic web. Synthesis Lectures on the Semantic Web: Theory and Technology **6**(2), 1–194 (2016)
24. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR (2013), http://arxiv.org/abs/1301.3781
25. Mota, C., Santos, D. (eds.): Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca (2008), http://www.linguateca.pt/LivroSegundoHAREM/, iSBN: 978-989-20-1656-6
26. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. pp. 807–814. ICML'10, Omnipress (2010)
27. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from wikipedia. Artificial Intelligence **194**, 151–175 (2013)
28. Nuno Cardoso: Harem e miniharem: Uma análise comparativa (7 2006)
29. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics (2014)
30. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics (Jun 2018)

31. Sandra Collovini, Joaquim Santos, B.C.J.T.R.V.P.Q.M.S.D.B.C.R.G., Xavier, C.C.: Portuguese named entity recognition and relation extraction tasks at iberlef 2019 (2019)
32. dos Santos, C., Guimarães, V.: Boosting named entity recognition with neural character embeddings. In: Proceedings of the Fifth Named Entity Workshop. pp. 25–33. Association for Computational Linguistics, Beijing, China (Jul 2015)
33. Santos, D., Cardoso, N.: Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca (November 2007), http://www.linguateca.pt/LivroHAREM/, iSBN: 978-989-20-0731-1
34. Schultz, J.: How much data is created on the internet each day? URL: https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/
35. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. pp. 142–147. CONLL '03, Association for Computational Linguistics (2003)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)
37. Wikipédia: Wikipédia - a free encyclopedia (2019), https://www.wikipedia.org/, [Online; accessed 06-July-2019]
38. Word2vec: Tool for computing continuous distributed representations of words. (2013), https://code.google.com/archive/p/word2vec/, [Online; accessed 06-July-2019]