Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4625–4630
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Embeddings for Named Entity Recognition in Geoscience Portuguese Literature

Bernardo Consoli[1], Joaquim Santos[1], Diogo Gomes[2], Fabio Cordeiro[2], Renata Vieira[1] and Viviane Moreira[3]
[1]Pontifical Catholic University of Rio Grande do Sul,
[2]Petrobras Research and Development Center,
[3]Federal University of Rio Grande do Sul
{Porto Alegre[1,3], Rio de Janeiro[2]}, Brazil
{bernardo.consoli, joaquim.santos}@acad.pucrs.br, {diogo.gomes, fabio.cordeiro}@petrobras.com.br,
renata.vieira@pucrs.br, viviane@inf.ufrgs.br

Abstract

This work focuses on Portuguese Named Entity Recognition (NER) in the Geology domain. The only domain-specific dataset in the Portuguese language annotated for Named Entity Recognition is the GeoCorpus. Our approach relies on Bidirecional Long Short-Term Memory - Conditional Random Fields neural networks (BiLSTM-CRF) - a widely used type of network for this area of research - that use vector and tensor embedding representations. We used three types of embedding models (Word Embeddings, Flair Embeddings, and Stacked Embeddings) under two versions (domain-specific and generalized). We originally trained the domain specific Flair Embeddings model with a generalized context in mind, but we fine-tuned with domain-specific Oil and Gas corpora, as there simply was not enough domain corpora to properly train such a model. We evaluated each of these embeddings separately, as well as we stacked with another embedding. Finally, we achieved state-of-the-art results for this domain with one of our embeddings, and we performed an error analysis on the language model that achieved the best results. Furthermore, we investigated the effects of domain-specific versus generalized embeddings.

Keywords: Geology, Named Entity Recognition, Word Embeddings, Stacked Embeddings

1. Introduction

Named Entity Recognition (NER) is a task within the field of Natural Language Processing (NLP) that deals with the identification and categorization of Named Entities (NE) in a given text. Resolutions to this task usually focus on a conventional set of categories, with the most commonly used being the Person, Location and Organization categories (dos Santos and Guimarães, 2015).

The Geology domain, commonly defined as the science that studies the origin, history, life, and structure of the Earth, must be described by other than the conventional Named Entity Recognition categories. While the Person, Location, and Organization categories can indeed be found within corpora of this domain, they, and other conventional categories, do not account for the vast majority of Named Entities found within such texts.

There is still a scarcity of research focusing on this domain, on the other hand the domain has a potentially large quantity of non-conventional Named Entities categories, in fact, yet to be defined. There is little consensus in literature as to what these categories should be to represent the core of this domain. In this work we consider the subdomain of Brazilian Sedimentary Basins, and defined the categories accordingly.

For the Named Entity Recognition problem, in general, Neural Networks (NN) coupled with Word Embeddings (WE) are usually present in the best evaluated systems (Lample et al., 2016; dos Santos and Guimarães, 2015; Santos et al., 2019). Word Embeddings are part of a set of language modeling techniques that aim to provide mathematical representations of natural language as multidimensional vector spaces. These vector spaces enable the use of mathematical abstractions to determine conceptual relations between the words of a language. Finally, since they can be trained in an unsupervised manner simply by feeding them a large raw text input, they are very simple to create.

Another Neural Network-based approach that deserves to be investigated for use in most Natural Language Processing task is the Flair Embeddings (Flair) language model (Akbik et al., 2018). Based on the same ideas first presented in the context of Word Embeddings, Flair Embeddings take into account not only word sequences, but also the character sequence distribution. It also incorporates training in both the forward (that is, left-to-right word and character sequences) and backward (right-to-left word and character sequences) directions. These innovations led Flair authors to call their models *contextual string embeddings*, in order to highlight the fact that they take sentence-level context and polysemy into account when calculating word vectors.

Our work focuses on Named Entity Recognition for the geology domain in the Portuguese language. This is an important domain for the Oil and Gas industry. Named Entity Recognition is an important step in the retrieval of textual information, making this work very relevant to the domains in question. Additionally, Portuguese literature in this area was limited, and this work seeks to expand upon it. Our GitHub page[1] contains the resources we created for this work, including a modified version of GeoCorpus, created by Amaral (2017).

[1] `https://github.com/jneto04/geocorpus`

The remainder of this work is organized as follows: Section 2. describes previous works on Named Entity Recognition and the use of embeddings in particular to the geology domain; Section 3. describes our experimental resources: embeddings models and Named Entity Recognition annotated corpus; Section 4. describes the neural network used for Named Entity Recognition; Section 5. describes the experiments and their results; and Section 6. describes our conclusions and planned future work.

2. Related Work

Amaral (2017) proposed a Named Entity Recognition system specific to the Geology domain, regarding the Brazilian Sedimentary Basin subdomain. In that work, she created the first Portuguese Named Entity annotated corpus in this domain. She reported a final F1 score of 54.33% as the best result with a CRF classifier.

Nooralahzadeh et al. (2018) presented Oil and Gas domain-specific Word Embeddings models. Their tests showed that domain-specific embeddings can be worthwhile even when the corpus used for their training is considerably smaller than that which is available to the general-domain counterpart.

Following that, Qiu et al. (2019) proposed an attention-based BiLSTM-CRF neural network for use in Named Entity Recognition specific to the geoscience domain for the Chinese language. Their approach leveraged attention mechanisms to enforce tagging consistency across whole documents, and used word2vec and Glove Word Embeddings models trained on the Chinese Wikipedia to add semantic knowledge to the model. Their results were comparable to other state-of-the-art systems, with their best model achieving a F1 of 91.47%.

In that same direction, Santos et al. (2019) preented state-of-the-art results for the general-domain Portuguese corpus using a BiLSTM-CRF Named Entity Recognition. The system is based on stacked Flair embeddings and traditional Word Embeddings, which outperformed the existing baseline by about 6 percentage points in terms of F1.

Gomes et al. (2018) generated the first set of domain-specific Portuguese word embeddings models for the Oil and Gas domain. They trained the models on a corpus composed of Petrobras' Geotechnical Bulletins as well as several thesis sponsored by Brazil's National Agency of Petroleum, Natural Gas and Biofuels (*Agência Nacional de Petróleo, Gás Natural e Biocombustíveis*).

Similarly to what Qiu et al. (2019) did for Chinese, our approach was to apply domain-specific Portuguese Word Embeddings (Gomes et al., 2018) to the Named Entities GeoCorpus (Amaral, 2017), based on the BiLSTM-CRF Named Entity Recognition system (Santos et al., 2019). This is the first experiment of this kind considering the proposed language and domain.

3. Resources

This work required the use of language resources that encompass both the Geology domain, our main focus, as well as a general domain. These resources include Portuguese embeddings models and a Named Entity annotated corpus for the Geology domain. The annotated textual corpus we used contains several domain-specific categories uncommon in general-domain corpora. To reflect this, we used a domain-specific Word Embeddings model to compare with a more common, general-domain model. A Flair Embeddings model was also considered in order to assess its context-sensitive capabilities. These resources are explained in more detail next.

3.1. General domain Word Embeddings

Word Embeddings aim to assign a mathematical representation to each term in a vocabulary, reportedly being able to capture semantic and syntactic similarities from the context they occur, considering a textual dataset (Hartmann et al., 2017; Mikolov et al., 2013). These techniques are based on the distributional hypothesis (Sahlgren, 2008) and provide a continuous n-dimensional vector for each word, in such a way that related words are assigned to a nearby position in the vector-space and their similarity can be measured in terms of their cosine distance (Mikolov et al., 2013).

The popularization of Word Embeddings for Natural Language Processing (NLP) tasks yielded several promising results, as reported by studies into several deep learning algorithms (Tshitoyan et al., 2019; Young et al., 2018; Camacho-Collados and Pilehvar, 2018; Goldberg, 2016). With that in mind, we selected the best performing Word Embeddings model from Santos et al. (2019) as our general-domain model. This was the 300-dimensional Word2Vec Skip-Gram model from Interinstitutional Center for Computational Linguistics - São Paulo University (NILC-USP), available on their website[2].

3.2. Oil and Gas Word Embeddings

Despite some public pre-trained embedding vectors being already available for Portuguese (Bojanowski et al., 2017; Hartmann et al., 2017; Santos et al., 2019), the highly technical oil and gas vocabulary presents a challenge to Natural Lan-guage Processing applications, in which some terms may assume a completely different meaning compared to the general-context domain. Therefore, there are consistent evidences that generating embedding models from a domain-specific corpus can significantly increase the quality of their semantic representation and, hence, the performance of NLP applications on specialized downstream tasks on the same domain (Gomes et al., 2018; Nooralahzadeh et al., 2018; Lai et al., 2016). As stated by Tshitoyan et al. (2019), the domain-specificity of the corpus is crucial to determine the quality of the embeddings and their utility for domain-specific tasks.

In order to provide neural networks with word vector representations suitable for the Geology domain-specific vocabulary in Portuguese, we used the public set of Word Embeddings provided by Gomes et al. (2018)[3]. Considering the results as reported by the authors, we focused on the word2vec 100-dimensional skip-gram models, which presented the most consistent results in their qualitative evaluations.

[2] http://nilc.icmc.usp.br/embeddings

[3] https://github.com/diogosmg/wordEmbeddingsOG

3.3. Flair Embedding Models

In this work, we used the FlairBBP (Santos et al., 2019) pre-trained model. It was trained on three corpora: BlogSet-BR (dos Santos et al., 2018), a Brazilian Portuguese web blog text corpus with 2.7 billion tokens; brWaC (Filho et al., 2018), comprising 3 billion tokens from Brazilian Portuguese texts retrieved through a web crawling process; and ptwiki-20190301[4], a Wikimedia Foundation data dump in Portuguese with about 162 million tokens.

We further trained this model with about 95,454 Portuguese sentences (for a total of 2,276,554 tokens) belonging to the Oil and Gas domain (of which the sedimentary basins subdomain is a part of), from Petrobras' Geoscience bulletins[5]. The goal was to test the effect of the addition of a more focused domain vocabulary to a generalised domain model in a domain-specific Named Entity Recognition task. Neural Networks were thus trained and tested using both the original FlairBBP, as well as our domain-enhanced model we refer to as FlairBBP$_{GeoFT}$.

3.4. Stacked Embeddings

Flair brings the possibility of stacking embeddings. Stacked Embeddings (SE) are what the authors of Flair (Akbik et al., 2018) call the addition of the dimensionality of Word Embeddings on top of those calculated by Flair, in effect appending the dimensions calculated by the Word Embeddings algorithm to the Flair model. This adds another layer of meaning to the embeddings, and generally yield improvement in the results if both the Flair and Word Embeddings models are of good quality. In this work, we symbolize a Stacked Embeddings by placing the "+" sign between two embeddings to show that they have been stacked. An example would be "GeoWE+FlairBBP", a Stacked Embeddings composed of the GeoWE and FlairBBP models.

3.5. Named Entity Recognition Corpus

The Named Entity annotated corpus used in this work is called GeoCorpus-2. It is a revised version of GeoCorpus (Amaral, 2017). The texts that make up this corpus are all in Portuguese, and were extracted from theses, dissertations, research articles, and Petrobras' geoscience bulletins. GeoCorpus-2's main changes from the original are the following: we removed duplicated sentences and put it in the CoNLL format, as opposed to the original XML format.

Table 1 presents the Named Entities categories, the shortened forms which will henceforth be used to refer to each category, the number of Named Entities in each category, and their distribution in the training, validation and testing datasets. We chose these categories with the help of experts, through a series of interviews, and considering the subdomain of Brazilian Sedimentary Basin. The question in mind was what would be the important categories for geologists in finding information in document collections. More details on the annotation process is given in Amaral (2017).

Some of the classes, such as chronological names, one could guess that quite accurate results would be achieved with simple lookup in dictionaries. However as we could see in our collection, and in fact should be expected from textual input, there are many small variations in the spelling of names. For instance, whereas "Pré-Cambriano" is the official and in fact the most frequent form, we found in at least other 6 variations, that includes not only gender and number but also drops of hifen and diacritics.

Considering that we have 13 different categories, each with different possible spelling variations, we considered that an annotation process mirroring what is really out there in document samples, for posterior machine learning, was the best way to deal with the problem.

4. Neural Network

Our neural network is a BiLSTM-CRF. It was originally used for Named Entity Recognition in English and German, but it is inherently language agnostic (Akbik et al., 2018). In brief, the network receives a character sequence as an initial input for the Character Language Model layer, where pre-trained embedding models are stored. The output of this layer are embeddings for each token, as calculated by the forward and backward pre-trained models. The embedding for each token is inputted separately into the Sequence Labeling Model layer, which adds word-level and character-level features to it, resulting in newly created vectors. These vectors are then used by the Conditional Random Fields in the Sequence Labeling layer to tag each individual token. The version trained for Portuguese comes from Santos et al. (2019), which is available on GitHub [6].

5. Experiments

We organized the experiments into eight tests, each involving a different solo embedding or stacked embedding. We evaluated all these tests by the CoNNL-2002 script (Sang, 2002). First, we extrinsically evaluated the performance of both the general-domain (W2V-SKPG) and the Geology domain (GeoWE) Word Embeddings model using them as the language model for the BiLSTM-CRF NN. We did the same for the general-domain Flair model (called FlairBBP), and for the Geology-enhanced Flair model (called FlairBBP$_{GeoFT}$). Finally, the last experiment involved the stacking of each Flair model with each Word Embeddings model, which resulted in four more embeddings models, and thus four more tests. Table 2 presents the results for each of these experiments.

According to the results, the lowest F-measure was the one achieved by GeoWE alone. In fact, whenever Flair models were stacked with these Word Embeddings, the resulting stacked embedding had worse performances than the Flair embeddings on their own. This extrinsic result means that GeoWE is not well suited to Named Entity Recognition in its current state. This was somewhat expected, though not to the degree found, and is thought to be due to the size and pre-processing performed upon the Word Embeddings

[4] https://dumps.wikimedia.org/ptwiki/20190301/

[5] http://publicacoes.petrobras.com.br/portal/revista-digital/pt_br/pagina-inicial.htm

[6] https://github.com/jneto04/ner-pt

Table 1: NE quantity in GeoCorpus-2 by Category

Category (Shortened)	**Quantity**	**Train**	**Test**	**Validation**
Eon (EON)	286	206	60	20
Era (ERA)	324	235	69	20
Period (PRD)	628	464	125	41
Epoch (EPC)	647	478	134	35
Age (AGE)	756	566	157	33
Siliciclastic Sedimentary Rock (sedSLCT)	738	543	150	45
Carbonate Sedimentary Rock (sedCARB)	240	173	50	17
Chemical Sedimentary Rock (sedCHEM)	5	3	1	1
Organic-rich Sedimentary Rock (sedORGN)	22	15	5	2
Brazilian Sedimentary Basin (BSN)	240	168	58	14
Basin Geological Context (BSNctx)	260	188	56	16
Lithostratigraphic Unit (LSTGunt)	574	425	107	42
Miscellaneous (MISC)	736	543	156	37
Total	**5456**	**4007**	**1128**	**321**

Table 2: Experiment results by embedding model

Embedding Model		**PRE**	**REC**	**F1**
Word Embeddings	GeoWE	73.31%	42.38%	53.71%
	W2V-SKPG	80.27%	64.18%	71.33%
Flair Embeddings	FlairBBP	85.97%	80.41%	83.10%
	FlairBBP$_{GeoFT}$	86.03%	82.45%	84.20%
Stacked Embeddings	GeoWE+FlairBBP	86.87%	72.16%	78.84%
	W2V-SKPG+FlairBBP	86.78%	81.47%	84.04%
	GeoWE+FlairBBP$_{GeoFT}$	86.35%	81.29%	83.74%
	W2V-SKPG+FlairBBP$_{GeoFT}$	**86.63%**	**82.71%**	**84.63%**

creation corpus. There are not many documents in Portuguese pertaining to the Geology domain, most of which are kept in the PDF format. Beyond that, a good portion of the PDFs required Optical Character Recognition in order to successfully extract their information. Errors inherent to this method meant that pre-processing had to be particularly stringent in order to result in a useful language model, which meant even less of the already limited domain corpus could be used in the model.

The Flair models performed well, with the Geology-enhanced version achieving an F-measure 1.1% higher than the original. The most interesting results, however, come from the Stacked Embeddings. The general-domain W2V-SKPG model, that yielded decent results by itself, managed to slightly enhance the results of the Flair models when stacked with them. In fact, the W2V-SPG+FlairBBP$_{GeoFT}$ model achieved the best results, with an F-measure of 84.63%. This result is 0.43% higher than FlairBBP$_{GeoFT}$ by itself.

Finally, even though results for GeoWE stacked embeddings were lower overall, some interesting observations can be made about them. amed Entity Recognition results achieved with the GeoWE+FlairBBP (F1 of 78.84%) and GeoWE+FlairBBP$_{GeoFT}$ (F1 of 83.74%) stacked embeddings were quite different, despite W2V-SKPG stacked models achieving similar F1 measures to each other. Stacking GeoWE with the Geology-enhanced FlairBBP$_{GeoFT}$ resulted in an F1 growth of 4.9%, whereas W2V-SKPG+FlairBBP$_{GeoFT}$ achieved an F1 only 0.59% superior to W2V-SKPG+FlairBBP. Though it is possible that we are merely reaching the best possible results using the GeoCorpus NER training and testing corpus, it is also possible that creating an enhanced version of GeoWE will further improve the results, beyond what was found when stacking the general-domain Word Embeddings model.

Overall, it outperformed the results reported by Amaral (2017) on this corpus (*i.e.* an F1 of 54.33% with a CRF classifier), but since she evaluated her results through cross-validation and not the CoNNL-2002 script, this is not a strict comparison.

Table 3 presents the results for our best model for the individual GeoCorpus categories. Most categories have F-measures within about 10 percentage points of the mean (84.63%), with a few clear outliers. The first is "*sedCHEM*", with and F1 of 0%, and "*sedORGN*", with an F1 of 100%. Both categories have very few instances in the dataset (5 and 22, respectively), which leads to the conclusion that the system was simply unable to learn how to label "*sedCHEM*" instances, and did not have enough "*sedORGN*" instances to properly test the labeling process.

The reasons for the outlying results of categories "*BSNctx*" and "*sedCARB*", with F-measures of 65.38% and 72.53% respectively, are more nuanced. From an error analysis focusing on these, we surmised that these categories are some of the most diverse in terms of token variety, whilst also being among the least represented in the corpus. Underrep-

Table 3: Individual Category Results for W2V-SKPG+FlairBBP$_{GeoFT}$

CATEG	PRE	REC	F1
Overall	86.63%	82.71%	84.63%
EON	98.25%	93.33%	95.73%
EPC	96.32%	97.76%	97.04%
ERA	89.06%	82.61%	85.71%
AGE	87.65%	94.90%	91.13%
MISC	72.32%	51.92%	60.45%
PRD	93.18%	98.40%	95.72%
BSN	79.25%	72.41%	75.68%
BSNctx	70.83%	60.71%	65.38%
sedCARB	80.49%	66.00%	72.53%
sedORGN	100.00%	100.00%	100.00%
sedCHEM	0.00%	0.00%	0.00%
sedSLCT	89.19%	84.67%	86.39%
LSTGunt	82.61%	88.79%	85.59%

resented categories with high scores tended to be more uniform, and are composed mostly of uni-grams, which made it easier for the system to learn from fewer samples.

Finally, the "*MISC*" category, with an F-measure of 60.45%, is the last outlier. This category includes all Named Entities that do not fit into other categories, but were considered relevant by the annotators. This happens because the domain is Sedimentary Basins, and as such is very broad, with very little relation between individual Named Entities. We believe that this category is simply too broad for the limited number of training instances we possess. However, we must see further the elements that were considered relevant in the domain to analyse whether some form of unity may indicate the existence of another category.

6. Conclusion

This work tested a Bidirecional Long Short-Term Memory - Conditional Random Fields (BiLSTM-CRF) Named Entity Recognition (NER) system on a Geology dataset. We tested several embeddings, and stacked embeddings, in order to find the combination that helped the system achieve the best results. Out of all language model combinations used in this work, our Named Entity Recognition architecture trained on the the Geology domain achieved its best results using general-domain Word Embeddings and a geology-enhanced general-domain Flair model, with an F1 of 84.63%. This architecture was able to achieve state-of-the-art results for the GeoCorpus dataset. It outperformed results by Amaral (2017), the only previous work using this corpus (F1 of 54.33% with a CRF classifier), still recalling that it considered a different validation process. Since GeoCorpus is the first Portuguese dataset for the Named Entity Recognition task, we cannot present other comparative evaluations. Even so, the results of our best model are comparable to state-of-the-art Named Entity Recognition systems for general-domain Portuguese texts.

In the process of our work, the corpus was converted to CONLL format and is now available. It must go through refinement and enrichment, as we have yet highly under covered classes, and also the classification itself is going through an analysis. It is the first of its kind for Portuguese, yet there are not many for other languages. So far, we are aware of Qiu et al. (2019) for Chinese.

As seen in the results presented in Section 5., the GeoWE embeddings alone did not perform well in this task. Our future work will involve the development of an enhanced version of GeoWE, created from more robust corpora, as well as the training of embedding models using Devlin et al. (2019)'s BERT and Peters et al. (2018)'s ELMO architectures.

Acknowledgments

This work was partially funded by CNPq, CAPES Finance Code 001, and Petrobras.

References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Amaral, D. O. F. (2017). *Reconhecimento de entidades nomeadas na Área da geologia: bacias sedimentares brasileiras*. Ph.D. thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *TACL*, 5:135–146.

Camacho-Collados, J. and Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *J. Artif. Intell. Res.*, 63:743–788.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

dos Santos, C. N. and Guimarães, V. (2015). Boosting named entity recognition with neural character embeddings. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics*, pages 25–33.

dos Santos, H. D. P., Woloszyn, V., and Vieira, R. (2018). Blogset-br: A brazilian portuguese blog corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Filho, J. A. W., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *J. Artif. Intell. Res.*, 57:345–420.

Gomes, D., Cordeiro, F., and Evsukoff, A. (2018). Word embeddings em português para o domínio específico de Óleo e gás. In *Proceedings of Rio Oil & Gas Expo and*

Conference 2018, Rio de Janeiro. Instituto Brasileiro de Petróleo, Gás e Biocombustíveis.

Hartmann, N., Fonseca, E. R., Shulby, C., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131.

Lai, S., Liu, K., He, S., and Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*.

Nooralahzadeh, F., Øvrelid, L., and Lønning, J. T. (2018). Evaluation of domain-specific word embeddings using knowledge resources. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Marilyn A. Walker, et al., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Qiu, Q., Xie, Z., Wu, L., Tao, L., and Li, W. (2019). Bilstm-crf for geological named entity recognition from the geoscience literature. *Earth Science Informatics*.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.

Sang, E. F. T. K. (2002). Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning*.

Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., and Vieira, R. (2019). Assessing the impact of contextual embeddings for portuguese named entity recognition. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems*, pages 437–442.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 08.