Related Named Entities Classification in the Economic-Financial Context

Daniel De Los Reyes[1], Allan Barcelos[1], Renata Vieira[2], Isabel H. Manssour[1]
[1]Pontifical Catholic University of Rio Grande do Sul, PUCRS.
School of Technology. Porto Alegre, Brazil.
[2]CIDEHUS, University of Évora, Portugal.
{daniel.reyes, allan.silva}@edu.pucrs.br, renatav@uevora.pt, isabel.manssour@pucrs.br

Abstract

The present work uses the Bidirectional Encoder Representations from Transformers (BERT) to process a sentence and its entities and indicate whether two named entities present in a sentence are related or not, constituting a binary classification problem. It was developed for the Portuguese language, considering the financial domain and exploring deep linguistic representations to identify a relation between entities without using other lexical-semantic resources. The results of the experiments show an accuracy of 86% of the predictions.

1 Introduction

In the context of the financial market, the news bring information regarding sectors economy, industrial policies, acquisitions and partnerships of companies, among others. The analysis of this data, in the form of financial reports, headlines and corporate announcements, can support personal and corporate economic decision making (Zhou and Zhang, 2018). However, thousands of news items are published every day and this number continues to increase, which makes the task of using and interpreting this huge amount of data impossible through manual means.

Information Extraction (IE) can contribute with tools that allow the monitoring of these news items in a faster way and with less effort, through automation of the extraction and structuring of information. IE is the technology based on natural language, that receives text as input and generates results in a predefined format (Cvitaš, 2011). Among the tasks of the IE area, it is possible to highlight both Named Entity Recognition (NER) and Relation Extraction (RE). For example, it is possible to extract that a given organization (first entity) was purchased (relation) by another organization (second entity) (Sarawagi, 2008).

A model based on the BERT language model (Devlin et al., 2018) is proposed to classify whether a sentence containing a tuple entity 1 and entity 2 (e1,e2), expresses a relation among them. Leveraging the power of BERT networks, the semantics of the sentence can be obtained without using enhanced feature selection or other external resources.

The contribution of this work is in building an approach for extracting entity relations for the Portuguese language on the financial context.

The remainder of this work is organized as follows. Section 2 presents news processing for the Competitive Intelligence (CI) area. Section 3 presents the related work. Section 4 provides a detailed description of the proposed solution. Section 5 explains the experimental process in detail, followed by section 6, which shows the relevant experimental results. Finally, section 7 presents our conclusions, as well as future work.

2 Competitive Intelligence and News Processing

Some of the largest companies in the financial segment have a Competitive Intelligence (CI) sector where information from different sources is strategically analyzed, allowing to anticipate market trends, enabling the evolution of the business compared to its competitors. This sector is usually formed by one or more professionals dedicated specifically to monitor the movements of the competition.

In a time of competitiveness that is based on knowledge and innovation, CI allows companies to exercise pro-activity. The conclusions obtained through this process allow the company to know if it really remains competitive and if there is sustainability for its business model. CI can provide some advantages to companies that use it, such as: minimizing surprises from competitors, identify-

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 8–15
April 19, 2021 © Association for Computational Linguistics

ing opportunities and threats, obtaining relevant knowledge to formulate strategic planning, understanding the repercussions of their actions in the market, among others.

The process of capturing information through news still requires a lot of manual effort, as it often depends on a professional responsible for carefully reading numerous news about organizations to highlight possible market movements that also retain this knowledge. It is then estimated that a system, that automatically filters the relations between financial market entities, can reduce the effort and the time spent on these tasks. Another benefit is that this same system can feed the Business Intelligence (BI) systems and, thus, establish a historical database with market events. Thus, knowledge about market movements can be stored and organized more efficiently.

3 Related Work

ER is a task that has been the subject of many studies, especially now when information and communication technologies allow the storage of and processing of massive data.

Zhang (Zhang et al., 2017) proposes to incorporate the position of words and entities into an approach employing combinations of N-grams for extracting relations. Presenting a different methodology to extract the relations, Wu (Wu and He, 2019) proposed to use a pre-trained BERT language model and the entity types for RE on the English language. In order to circumvent the problem of lack of memory for very large sequences in convolutional networks, some authors (Li et al., 2018; Florez et al., 2019; Pandey et al., 2017) have adopted an approach using memory cells for neural networks, Long short-term memory (LSTM). In this sense, Qingqing's Li work (Li et al., 2018) uses a Bidirectional Long Short-Term Memory (Bi-LSTM) network, which are an extension of traditional LSTMs, for its multitasking model, and features a version with attention that considerably improves the results in all tested datasets. Also using Bi-LSTM networks, Florez (Florez et al., 2019) differs from other authors in that it uses types of entities and the words of the entities being considered for a relation in addition to using information such as number of entities and distances, measured by the number of words and phrases between the pair of entities. The entry of the Bi-LSTM layer is concatenation of words and relations, with all words between the candidate entities (included), provided by a pre-trained interpolation layer. Yi (Yi and Hu, 2019) proposes to join a BERT language model and a Bidirectional Gated Recurrent Unit (Bi-GRU) network, which is a version of Bi-LSTM with a lower computational cost. Finally, they train their model based on a pre-trained BERT network, instead of training from the beginning, to speed up coverage.

Some works (Qin et al., 2017; GAN et al., 2019; Zhou and Zhang, 2018) use attention mechanisms to improve the performance of their neural network models. Such mechanisms assist in the automatic information filtering step that helps to find the most appropriate sentence section to distinguish named entities. Thus, it is possible that even in a very long sentence, and due to its size being considered complex, the model can capture the context information of each token in the sentence, being able to concentrate more in these terms the weights of influence. Pengda Qin (Qin et al., 2017) proposes a method using Bi-GRU with an attention mechanism that can automatically focus on valuable words, also using the pairs of entities and adding information related to them.

Tao Gan (GAN et al., 2019) also addresses RE with an attention method to capture important parts of the sentence and for that, it uses an LSTM attention network for entities at the subsequent level. In this way, he focuses more on important contextual information between two entities. Zhou (Zhou and Zhang, 2018) also implement a model based on RNN Bi-GRU with an attention mechanism to focus on the most important assumptions of the sentences for the financial market.

Despite having great importance, the financial domain, specifically, has been little explored in the literature. The authors at (Zhou and Zhang, 2018) created a corpus collecting 3000 sentence records manually from the main news sites, which was used to recognize the entity and extract relations such as learning and training as a whole.

Most studies present RE solutions for English texts, and, in this way, it is also possible to identify a larger number of data sets in this language. There are few data sets available in the Portuguese language, such as the Golden Collection HAREM, which is widely used in the literature (Chaves, 2008; Cardoso, 2008; Collovini et al., 2016). HAREM is a joint assessment event for the Portuguese language, organized by Linguateca

(Santos and Cardoso, 2007). Its objective is to evaluate recognizing systems of NE (Santos and Cabral, 2009). The Golden Collection (GC) is a subset of the HAREM collection, being used for the task of evaluating the systems that deal with Recognition of Named Entities.

The lack of this type of resource forces researchers to develop their own research corpus. In most cases, it is necessary to first create a set with the sentences and write them down when the classification is supervised to proceed with the RE task. Besides, the lack of public data sets also makes it difficult to fairly compare related work, as well as requires more time and effort from the researcher.

It is possible to observe that there are works that discuss the task of extracting relations between NE and that already employ machine learning techniques for this purpose. However, although we found some works for the RE task, few of them are suitable for the Portuguese language, and none of them are related to the financial context. Considering other languages, The work of Zhou (Zhou and Zhang, 2018) was the only one that came closest to our goals. However, there is a gap in the literature for works that address such tasks using deep learning techniques and Portuguese as the main language, especially in the financial-economic context as addressed in this work.

4 Architecture

In this section, we present our BERT-based model in detail. As shown in Figure 1, it contains three parts: (1) Input layer; (2) BERT layer; and (3) Output layer, which is composed of a Sigmoid activation function and two neurons that represent the classes to be predicted.

The input layer consists of a BERT encoder used for input sentence tokenization and produces a tuple of arrays (token, mask, sequence ids), which were used as input to the second layer that is the Portuguese BERT language model (Souza et al., 2020) [1] from Huggingface python package [2] (Wolf et al., 2020). Figure 2 illustrates the input layer of the proposed model. The entry consists of (1) the original sentence with the mentioned entities and (2) the entities to be verified concatenated. A special token [cls] and a token [sep] are added at the beginning and end of the input string respectively, as mentioned in the original BERT implementation (Devlin et al., 2018).

The third layer of the model architecture is identified as the output layer. This layer is fully connected with a tangent activation function. The output of this layer is propagated to a new fully connected layer, with a Sigmoid activation function, whose characteristic is the mapping of input values to 0 or 1. In this model, these values represent non-relation and relation, respectively. As shown in Figure 1, this layer still has two output neurons, which indicate the respective classes to be predicted by the model. In the end, we added a dropout layer with a 0.1 rate to avoid model overfitting, which happens when the model memorizes the training data and thereby loses the power of generalization.

5 Experiments

The purpose of this section is to verify the proposed model performance thought experiments on the financial domain corpus. The proposed study follows the classic methodology of Knowledge Discovery in Databases (KDD) (Fayyad et al., 1996), which contains 5 phases that range from data collection to the evaluation of the results.

The following subsections aim to indicate how each step of the methodology was applied in the context of our work. Subsection 5.1 refers to the Selection step and seeks to indicate what data will be used during the experiments for the RE task. Subsection 5.2 addresses the Pre-processing step, indicating procedures for quality checking, cleaning, correction, or removal of inconsistent or missing data. Subsection 5.3 reports the Transformation phase, where the transformation processes applied to the data set in the context of our work are explored. Subsection 5.4 brings the penultimate phase, of Mining, where the data mining process is presented. Finally, the last phase of the methodology is presented in the subsection 5.5, which consists of evaluating the performance of the model applied on top of the data that were not used in the training or mining phase.

5.1 Selection

As indicated in section 3, there was no evidence of open data sets in the context of extracting relations in the financial field for the Portuguese language. Therefore, for this work, a corpus was created with 3,288 tuples annotated manually. These tuples originate from more than 4,000 paragraphs of financial

[1] Available at https://simpletransformers.ai/

[2] Available at https://github.com/huggingface/transformers

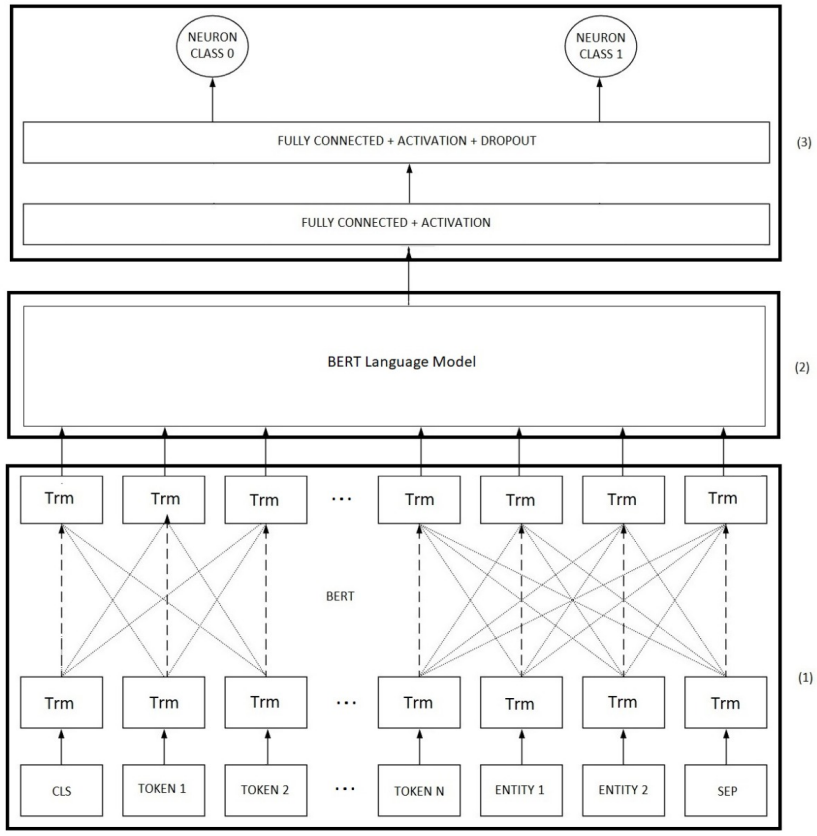

Figure 1: Complete model architecture with its 3 layers: (1) Input layer; (2) BERT layer; (3) Output layer.

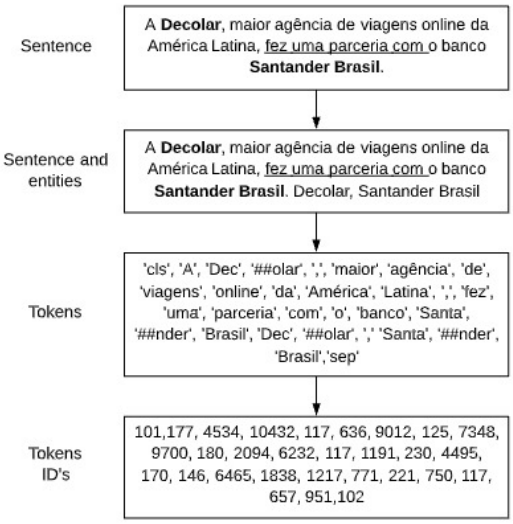

Figure 2: Examples of data transformations in the input layer of the model. The entities to be evaluated appear in bold, and the text that represents the semantic relation between them is underlined.

market news, provided by a partner company that collected them in various communication vehicles such as financial market websites, newspapers, and corporate balance sheets. Sentences that include co-referral are also removed because co-reference treatment would require additional processing.

5.2 Pre-processing

The next step concerns the data pre-processing and cleaning. This step occurs through the manual process of spelling correction of each sentence. Acronyms are also extended, as well as the standardization of different ways of indicating the same named entity.

The standardization can be done manually, but in a real work scenario, this task becomes massive and can be automated by creating a base of named entities and their acronyms. Thus, it is possible to elaborate a process that validates the acronyms contained in the sentence and replace them with their extensions or even with an approach that focuses on only a few specific entities informed by the CI

analyst himself.

The data cleaning process is also done manually, where special characters and acronyms that follow the description itself are removed. Sentences containing less than 4 tokens will also be removed, as they can be considered irrelevant to the context of the approach. At the end of this cleaning step, just over 2500 sentences are filtered.

In this same phase, the identification of named entities will also occur, through a single NER tool, called SpaCy [3], ensuring that the same criterion was used for all sentences.

The named entities in question are those related to the categories person, location, and organization. The focal point is information about the organizations, as well as its relations with other organizations, persons, and locations.

After identifying all named entities, sentences that have less than 2 entities are discarded. At the end of this new disposal, the corpus consists of 1292 unique sentences that move on to the next stage.

5.3 Transformation

With the identification of the Named Entities in the previous phase, a combination of all the entities present in the sentence is made and a triple (sentence, entity, entity) is formed for each combination, which can generate several records for the same sentence. After this creation of records with the combination of entities, manual annotation of records that have a semantic relation between the highlighted named entities is made manually.

After the end of the manual annotation of the relations between the entities, the corpus consists of 3288 records. Of this total, 1485 (45%) are positive tuples, that is, it contains a relation between the highlighted entities, and 1803 (55%) are negative tuples, where there is no relation between the entities. Finally, the two named entities are concatenated at the end of the sentence. The data set is available at https://github.com/DanielReeyes/relation-extraction-deep-learning.

The relation annotating process did not consider the past defined classes or relations. A positive tuple is considered when there is any semantic relation between two named entities of the categories defined in 5.1. Here are some examples of positive annotated tuples that contain relation between named entities of type organization:

- A **Abraço** é uma **Instituição Particular de Solidariedade Social**.
- A **Caixa** é controladora do **Pan** , ao lado do BTG , com 32,8% do negócio.
- A **Havanna** fecha parceria com o **Santander** para inaugurar um novo modelo de negócios.
- A partir de agora , a **NET** está na **Claro**.

As sentences are naturally composed of words and characters, then the transformation step in the present study also consists of transforming the tokens into numerical representations by the BERT encoder. As stated in past sections, the special tokens [CLS] and [SEP] are also added and encoded properly on each sentence, finalizing the composition of the input layer.

5.4 Mining

The predictive task is characterized by the search for a behavioral pattern that can predict the behavior of a future entity (Fayyad et al., 1996). The corpus data are randomly divided into two parts, 80% of which are used for training the model and 20% for testing. The part for the test is still divided equally into 2, where they are used as validation and test sets to test the generalization of the model. The first set is used so that the algorithm can search for this particular pattern in the data concerning the relation label. Thus, after the training stage where the model can recognize this pattern, it is possible to apply it to the validation data and later on the test set, simulating a real environment. In this step, the original balance level is also maintained in all sets created, being able to rule out that the model contains any bias to learn a certain type of complexity.

The adjustment of hyper-parameters of the BERT used was due to the combination of all values indicated by Jacob Devlin in (Devlin et al., 2018), in addition to the standard values for the Simple Transformers library model. In this work, Jacob used most of the hyper-parameters with default values except for the lot size, learning rate, and the number of training epochs. The dropout rate was always maintained at 0.1. Thus, the values tested for this task were:

- **Batch Size:** 16, 32;

[3] https://spacy.io/

Hyper-parameter	Value
Batch Size	32
Learning Rate	5e-5
Epochs	4

Table 1: Combination of hyper-parameters that presented better results.

Set	Samples	Positive Class Distribution (%)	Positive Samples
Original	3288	45.16	1485
Training	2630	45.17	1188
Validation	329	45.28	149
Test	329	45.98	148

Table 2: Sample composition of each data set used in the experiments.

- **Learning Rate (AdamW):** 5e-5, 3e-5, 2e-5;
- **Epochs:** 2, 3, 4, 5.

In the end, we did a total of 24 experiments with all the possible combinations of the above described parameters. After analyzing the results, the model with the values was selected according to Table 1.

5.5 Evaluation

To evaluate the model, metrics such as Accuracy, Recall, Precision, and F1-Measure were provided. According to Table 2, each set maintained the original imbalance of the data set according to the target variable, in this case, indicating whether or not there is a relation between the entities assessed. In this way, the model is evaluated for the ability to indicate whether a given pair of entities contained in a sentence has a relation or not, configuring a binary classification problem, whose positive class refers to entities that have a semantic relation.

6 Results

After the training stage of the model, it was applied to the test data set. In this evaluation step, the model obtained reasonable results, achieving an overall accuracy and F-Measure of 86%. An important observation to make is that results are also good when it comes to the target class, that is, when the label is positive, as can be seen in Table 3.

As indicated in Section 3, the vast majority of studies present RE solutions for texts in English or a domain other than finance. Thus, it is difficult to compare the results of the proposed method with state-of-the-art approaches.

Metric	Positive	Negative	General
Recall	0,8993	0,8389	0,8662
Precision	0,8221	0,9096	0,8699
F-Measure	0,8590	0,8728	0,8665
Accuracy	-	-	0,8663

Table 3: Precision, Recall and F-Measure calculated for each class and Accuracy and general F-Measure of the model.

Nevertheless, it is shown that the proposed model was able to recognize patterns and indicate when two entities are semantically related in the same sentence in the financial domain.

The process of finding the best parameters for BERT is time-consuming as the predictions made by the network. The time might not be a constraint to using the RE task model applied to the context of the financial domain considering that this demand does not require the processing time to be real-time.

We believe that if the data set is increased with more samples, the model may have a performance gain. Also, we can notice that the data set has a small unbalanced distribution rate, with a greater number of negative samples.

This imbalance can help explain the difference in precision and F-measure between the positive and negative class indicated in Table 3, where it is possible to see that the model gets more right when the tested entities had no relation in the sentence.

Regarding Recall, the study indicates that, even with the imbalance of the data, the proposed model achieved a very good performance of approximately 90% when it comes to the positive class (it has a relation). That is, when it really belongs to the positive class, in approximately 90% of the cases, it identifies correctly.

It is also possible to carry out tests with adjustments of more hyper-parameters such as loss function, optimizers, among others. In addition to adjustments to the hyper-parameters of the approach, more contextual information of the samples can be added, such as the type of the named entity, whether it is an organization, person, or place, and scope adopted for the task being worked on. In this way, it is possible to delimit the types of relations between 2 entities, excluding, for example, an acquisition relation between two entities of the person type.

7 Conclusion and Future works

The present work proposed an approach to extract relations between named entities, in the financial-economic context, based on the Portuguese BERT language model, to our best knowledge, different from what is already in the literature. Thus, it provides an insight into the use of pre-trained deep language models for extracting relations for the Portuguese language financial market.

From the related work section, it is possible to verify that there is little research on the technology for extracting the relation between named entities for the financial domain, for the Portuguese language. This domain lacks practical solutions, given a large amount of information in the financial field, and manual analysis becomes difficult to meet the needs and make full use of that information.

A model of classification of relations between named entities based on BERT was proposed, which replaces explicit linguistic resources, required by previous methods. This approach uses the information from the sentence and the concatenated entity pair, which allows more than one entry to be sent since a sentence can have N pairs of named entities. Therefore, the adopted approach allows the sentence and the pair of entities to be inferred to be sent separately.

The results demonstrate that the approach used can bring satisfactory results, reaching an accuracy of 86%. During the discussion of results, some adjustments were made to try to improve accuracy, such as testing other combinations of hyper-parameters and also the increase in the corpus. However, the development of memory improvements and optimizations are still in need, especially in the training period, due to the complexity of the pre-trained BERT model.

As a natural continuation of this work, we will proceed with tests with other combinations of hyper-parameters as indicated in Section 6. To try to reduce the chance of the model being surprised with some non-standard samples, new data will be annotated and added to the research corpus. Thus, the model can be trained with a greater amount of data and a greater diversity of data patterns.

As a continuity, a second model will also be developed, with sequential classification, so that it is possible to highlight the parts of the sentences that represent or describe the relation between the named entities verified. To achieve this goal, this second model will be trained only with the tuples that contain the annotated relation. Thus, the output of the model proposed in this work will be the input of the sequential classifier model.

Acknowledgments

This work was partially funded by the Portuguese Foundation for Science and Technology, project UIDB/00057/2020.

References

Nuno Cardoso. 2008. Rembrandt-reconhecimento de entidades mencionadas baseado em relaçoes e análise detalhada do texto. *quot; Encontro do Segundo HAREM (Universidade de Aveiro Portugal 7 de Setembro de 2008)*.

Marcírio Chaves. 2008. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem. *quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008*.

Sandra Collovini, Gabriel Machado, and Renata Vieira. 2016. A sequence model approach to relation extraction in portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1908–1912.

A Cvitaš. 2011. Relation extraction from text documents. In *2011 Proceedings of the 34th International Convention MIPRO*, pages 1565–1570. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37. GS Search.

Edson Florez, Frederic Precioso, Romaric Pighetti, and Michel Riveill. 2019. Deep learning for identification of adverse drug reaction relations. In *Proceedings of the 2019 International Symposium on Signal Processing Systems*, pages 149–153.

TAO GAN, YUNQIANG GAN, and YANMIN HE. 2019. Subsequence-level entity attention lstm for relation extraction. In *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, pages 262–265. IEEE.

Qingqing Li, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, Jian Wang, Liang Yang, Kan Xu, and Yijia Zhang. 2018. A multi-task learning based approach to biomedical entity relation extraction. In *2018 IEEE International Conference on*

Bioinformatics and Biomedicine (BIBM), pages 680–682. IEEE.

Chandra Pandey, Zina Ibrahim, Honghan Wu, Ehtesham Iqbal, and Richard Dobson. 2017. Improving rnn with attention and embedding for adverse drug reactions. In *Proceedings of the 2017 International Conference on Digital Health*, pages 67–71.

Pengda Qin, Weiran Xu, and Jun Guo. 2017. Designing an adaptive attention mechanism for relation classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4356–4362. IEEE.

Diana Santos and Luís Miguel Cabral. 2009. Gikiclef: Crosscultural issues in an international setting: asking non-english-centered questions to wikipedia. In *quot; In Francesca Borri; Alessandro Nardi; Carol Peters (ed) Cross Language Evaluation Forum: Working notes for CLEF 2009 (Corfu 30 Setembro-2 Outubro) Springer*. Springer.

Diana Santos and Nuno Cardoso. 2007. Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área.

Sunita Sarawagi. 2008. *Information extraction*. Now Publishers Inc.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pretrained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2361–2364.

Rongli Yi and Wenxin Hu. 2019. Pre-trained bert-gru model for relation extraction. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, pages 453–457.

Qin Zhang, Jianhua Liu, Ying Wang, and Zhixiong Zhang. 2017. A convolutional neural network method for relation classification. In *2017 International Conference on Progress in Informatics and Computing (PIC)*, pages 440–444. IEEE.

Zhenyu Zhou and Haiyang Zhang. 2018. Research on entity relationship extraction in financial and economic field based on deep learning. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pages 2430–2435. IEEE.