Extraction and Access to Information in Natural Language for Non-Developers - Democratizing Information

Progress Report – 2021/2022
Partial Time

Emanuel Matos

MAPi DOCTORAL PROGRAMME IN COMPUTER SCIENCE

University of Aveiro
Portugal
September - 2022

Contents

Chapter 1

Introduction

Automatic extraction of information from natural language sources has many applications, including Business Intelligence [1], Forensics [2], Medicine [3] and question answering systems [4]. Seeking to work new concepts and diversity, both in terms of the environment, various datasets, such as the technical issue and experiments, the year of 2021 was very interesting. Even with the limitation of working part-time and being in Brazil, we were able to immerse ourselves in a world of new algorithms and evaluate diversity possibilities, as well as have technical production that allowed us to publish and present at an international conference. Not only that, the results obtained during this process of experiments led us to develop a pipeline that will be the basis for continuing the development of the work.

This report describes the work developed during the 2nd year Thesis work [1].

[1] I'm a part-time student, this 2nd year corresponds to the first half of the 2nd year of Thesis.

Chapter 2

Developments

The main developments during the second year of Thesis work were:

- Test the methods proposed and explored in 2020/2021 in new domains.
- Seek and find new Data-set and environment.
- Improve the process of NERs output combination.
- Publication of results.
- Testing and training several differents algorithms sounds like GPT3, GPT2, GPT-J and GEP-Neo.

The next section presents additional information on some of developments:

2.1 Testing an Training

In the search to know more about NER or about processes that would lead us to create NER with performance comparable to the state-of-the-art, considering the mentioned two main types, the most relevant approaches are:

GPT (Family) – Generative Pre-trained Transformer is an autoregressive language model that uses deep learning to produce human-like text.

The architecture is a standard transformer network (with a few engineering tweaks) with the unprecedented size of token-long context and parameters (requiring huge storage). The training method is ”generative pretraining”, meaning that it is trained to predict what the next token is. The model demonstrated strong few-shot learning on many text-based tasks.

GPT-3 – More recently development and deeper in token and pre-trainned. It is the third-generation language prediction model in the GPT-n series (and the successor to GPT-2) created by OpenAI, a San Francisco-based artificial intelligence research laboratory. GPT-3’s full version has a capacity of 175 billion machine learning parameters. GPT-3[5], which was introduced in May 2020, and was in beta testing as of July 2020, is part of a trend in natural language processing (NLP) systems of pre-trained language representations.

The quality of the text generated by GPT-3 is so high that it can be difficult to determine whether or not it was written by a human, which has both benefits and risks. Thirty-one OpenAI researchers and engineers presented the original May 28, 2020 paper introducing GPT-3. In their paper, they warned of GPT-3’s potential dangers and called for research to mitigate risk. David Chalmers[6], an Australian philosopher, described GPT-3 as ”one of the most interesting and important AI systems ever produced.”

GPT-NEO – Introducing GPT-Neo[7], an open-source Transformer model with only 2.7 Billion parameters, also notes that the largest GPT Neo is almost equivalent to the smallest GPT-3, which resembles GPT-3 both in terms of design and performance. Will be possible train this model from scratch using a mesh-TensorFlow library, a superb library for easy and efficient data and model parallelism to help with distributed support. These models have tons of data to train on and lots of parameters; hence parallelism is vital here. This means that you’ll be running different segments of your training simultaneously rather than doing it one after another. This is completely independent of different batches.

GPT-J – GPT-J[7], was trained using a new library, Mesh-Transformer-JAX. The library uses Google’s JAX linear algebra framework, instead of a dedicated deep-learning framework such as TensorFlow.

Komatsuzaki[8] claims that GPT-J provides ”more flexible and faster inference than Tensorflow,” and developing the model took much less time than previous projects. Compared to the 2.7GB GPT-Neo model, GPT-J shows a 125% improvement in training efficiency.

NERs using Machine learning – Machine learning methods are more flexible to adapt to distinct contexts provided that exists enough data about the target context. Diverse machine learning methods have been applied to NER. They can be categorized in three main branches that have distinct needs of training data: (1) supervised learning, (2) unsupervised learning and (3) reinforcement learning .

The supervised learning methods use a training set (a corpus) that was already manually labeled by experts. The unsupervised learning method consumes an untrained data set and extract patterns from it, contrary to the supervised method this one doesn’t need a labeled training set. The reinforcement learning method uses agents to learn polices[9] that can be used to label an untrained data set. These agents are trained using a reward system. Machine learning methods that were successfully applied to NER over the years include:

Transfer learning – Transfer learning reuses pre-trained models in order to perform a different task. It’s very popular as it makes possible training deep neural networks with small amounts of data. In NER it was used, for example, to develop NERs for novel types of entities [10].

Transformers – Transformers [11], introduced in 2017, are a deep learning model based on the attention mechanism designed to handle sequential input data, such as natural language. Unlike RNNs, Transformers do not require data to be processed in order allowing much more parallelization and, because of that, training with huge datasets. This created the conditions for the development of pre-trained systems such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) [12]. Transformers demonstrated their superior efficiency in the recognition of named entities and in a variety of other classification tasks. A variety of state-of-the-art NER systems were developed adopting BERT for different domains and languages (ex: [13]).

In my actual stage of the research, was possible used the BERT process, the choise was between two types of BERT:

- `https://tinyurl.com/ac42ev5v`
- `https://tinyurl.com/35tjmh8n`

Both in website pages, where we did the implementation and its comparative evaluation with the BERT implementation. By strategy and more domain about the implementation of BERT, chosen to use the first item to development process and so we are creating the continuation of our pipeline. Now using as a basis our adaptive creation of NER, passing through BERT having a model to evaluate the performance of our current pipeline.

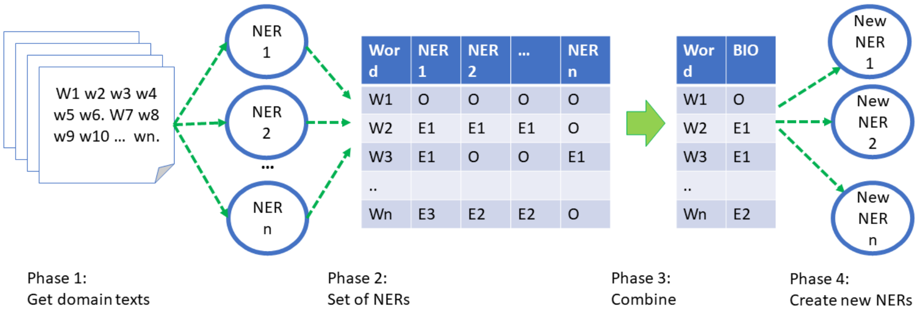
2.2 Ensemble of NERs

Named Entity Recognition (NER) is an essential step for many natural language processing tasks, including Information Extraction. Despite recent advances, particularly using deep learning techniques, the creation of accurate named entity recognizers continues a complex task, highly dependent on annotated data availability. To foster existence of NER systems for new domains it is crucial to obtain the required large volumes of annotated data with low or no manual labor.

In [14] it is proposed a system to create the annotated data automatically, by resorting to a set of existing NERs and information sources (DBpedia). The approach was tested with documents of the Tourism domain. Distinct methods were applied for deciding the final named entities and respective tags. The results show that this approach can increase the confidence on annotations and/or augment the number of categories possible to annotate.

The paper also presents examples of new NERs that can be rapidly created with the obtained annotated data. The annotated data, combined with the possibility to apply both the ensemble of NER systems and the new Gazetteer-based NERs to large corpora, create the necessary conditions to explore the recent neural deep learning state-of-art approaches to NER (ex: BERT) in domains with scarce or nonexistent data for training.

Fig. 2.1, presents the initial of the proposed process for automatic tagging of named entities by using and Ensemble of NERs, showing its 4 phases .



Word	NER 1	NER 2	…	NER n
W1	O	O	O	O
W2	E1	E1	E1	O
W3	E1	O	O	E1
..				
Wn	E3	E2	E2	O

Word	BIO
W1	O
W2	E1
W3	E1
..	
Wn	E2

Figure 2.1: Caption From [14]

The Fig 2.2 present our update and new strategy to apply for development the pipeline.

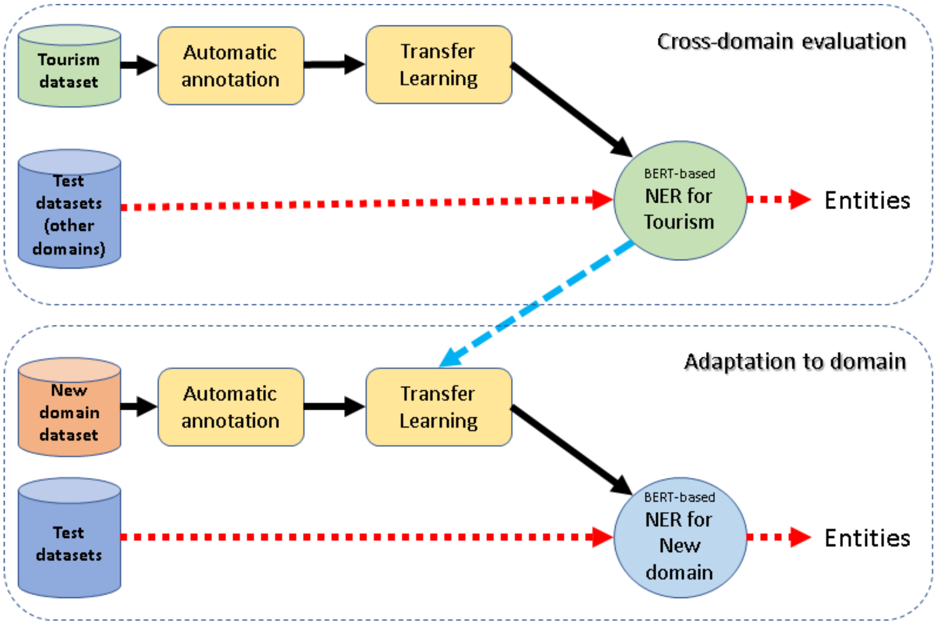


Figure 2.2: New process [15]

2.3 Dataset(s)

2.3.1 Tourism Domain

For our initial proof-of-concept the area of Tourism was selected. A set of sources were selected manually, and the text scrapped using the Scrapy library for retrieving documents and BeautifulSoup library for getting document data.

The option for Tourism domain resulted from our previous work in this domain [16]. Selected by the high potential of automatic information extraction to provide relevant information to several domain stakeholders (e.g., hotel managers).

The main source selected was Wikivoyage [17], with hundreds of texts regarding countries, regions, cities, tourist attractions, etc. Based on the tool's ease of use concept, the manual extraction option, without advanced features, was implemented to adjust expectations regarding the development of the work and evaluate the capture properties of possible entities without advanced techniques.

2.3.2 LeNER Legal texts dataset

2021 was an experiment year, we use a LeNER Dataset[18]. LeNER-Br is a Portuguese language dataset for named entity recognition applied to legal documents. LeNER-Br consists entirely of manually annotated legislation and legal cases texts and contains tags for persons, locations, time entities, organizations, legislation and legal cases. To compose the dataset, 60 legal documents from several Brazilian Courts were collected. Courts of superior and state levels were considered, such as Supremo Tribunal Federal, Superior Tribunal de Justiça, Tribunal de Justiça de Minas Gerais and Tribunal de Contas da União. In addition, four legislation documents were collected, such as ”Lei Maria da Penha”, giving a total of 70 documents.

2.3.3 Paramopama

Extends the PtBR version of WikiNER corpus, revising incorrect assigned tags in order to improve corpus quality, also extend the corpus size and provide proper evaluation[19]. This dataset has a total of 240,755 words tagged as part of an entity, considering 4 types (PERSON, LOCATION, ORGANIZATION and TIME).

2.4 Exploration of state-of-the art Algorithms

Throughout the year, an active search for state-of-art developments was performed. A selection of promising approaches was further tested and evaluated. The most relevant are briefly mentioned next.

2.4.1 BERT

They are based in BERT [25] implementation by Tobias Sterbak [26] using the Transformers package by Huggingface [27], Keras and TensorFlow.

Last year we started to address is to train ML systems using data annotated automatically. Not just expressing data in the format(s) accepted by existing systems but also doing so in a way non-developers feel comfortable with.

Aiming to mitigate the problems highlighted above, the following main objectives were adopted for the ongoing work on this topic:

1. Develop processes to simplify the creation of NER systems for new domains, starting by the creation of the needed annotated data;

2. Make NER deployment as easy as possible in order to be used by non specialists, contributing to breaking existing usage barriers thus fostering wider adoption of such systems.

2.4.2 GPT Family

Was evaluated the possibility of using GPT3, GPT2, GPT-J and GPT-Neo. As the volume of parameters and GPU's required for each of the algorithms to work, in addition to the machine that currently processes/processed BERT, we seek alternatives without cost, because one of the basic issues of this work is to be democratic, and with cost this means restriction. We still use Google's Collaboration in its lighter paid version to try to process and work with some new algorithm.

The GPT-Neo was the one that came closest to the possibility of working, because in its lightest version of processing i.e. with 125mega was the one that under test it was possible to use with a ready example.

Chapter 3

Results

This chapter highlights the main results obtained so far. They integrate presentations, publications and software developments.

3.1 Publications and Presentations

During the period 2 submissions to international conferences were made, one already accepted and presented, the second waiting revision results.

3.1.1 PROPOR 2022

In last report, in section 3.1.2 was mentioned a work regarding the development of the BERT-based NER with automatically annotated train data is the basis for an article in preparation. We submitted it to PROPOR 2022, with the title "Named Entity Extractors for New Domains by Transfer Learning with Automatically Annotated Data"[20], accepted, published as full paper in a book of Springer LNCS series and presented by the author.

Abstract: Named entity recognition (NER) tasks imply token-level labels. Annotating documents can be time-consuming, costly, and prone to human error. In many real-life scenarios, the lack of labeled data has become the biggest bottleneck preventing NER being effectively used in some domains and with some natural languages, with negative impacts in the quality of some tasks. To overcome the

barrier of the lack of annotated data for new application domains in some natural languages, we propose a method that uses the output of an ensemble of NER's to automatically annotate the data needed to train a Bidirectional Encoder Representations from Transformers (BERT) based NER for Portuguese. The performance was assessed using MiniHAREM dataset with promising results. For domain relevant classes such as LOCAL, F1, Precision and Recall above 50% were obtained when training only with automatically annotated data.

3.1.2 IberSpeech (submitted)

In this year, we prepared and submitted a paper called "Assessing Transfer Learning and automatically annotated data in the development of Named Entity Recognizers for new domains" at Iberspeech Conference in Granada, whose will be in November 2022[15].

Abstract: With recent advances Deep Learning, pretrained models and Transfer Learning, the lack of labeled data has become the biggest bottleneck preventing use of Named Entity Recognition (NER) in more domains and languages. To relieve the pressure of costs and time in the creation of annotated data for new do- mains, we proposed recently automatic annotation by an ensem- ble of NERs to get data to train a Bidirectional Encoder Rep- resentations from Transformers (BERT) based NER for Por- tuguese and made a first evaluation. Results demonstrated the method has potential but were limited to one domain. Having as main objective a more in-depth assessment of the method ca- pabilities, this paper presents: (1) evaluation of the method in other domains; (2) assessment of the generalization capabilities of the trained models, by applying them to new domains without retraining; (3) assessment of additional training with in-domain data, also automatically annotated. Evaluation, performed us- ing the test part of MiniHAREM, Paramopama and LeNER Portuguese datasets, confirmed the potential of the approach and demonstrated the capability of models previously trained for tourism domain to recognize entities in new domains, with better performance for entities of types PERSON, LOCAL and ORGANIZATION.

3.2 Software

Despite the option made for not developing at this stage the “Tools for non-programmers” [1], several scripts were developed and existing ones improved to contemplate the specifities of new domains. As is still very manual the process of joining scripts, we believe that before making it more “friendly” to the end user, we still have much to develop and validate, so we consider the overall system in an initial, embryonic, stage, that can be considered a delay relative to our initial plan (see Fig. 4.1).

3.3 Comments

The publications planned for the year 22/23 were achieved as we anticipated in the planning.

Advances made in 2022 such as the evaluation of some state-of-the-art algorithms and various dataset environments, this evaluation being a detailed work relating performance and cost and the previous processing carried out in a different environment, makes us optimistic about the effective reach of the next one M2 milestone (M2 expected to be delivered at the end of 2023 - New NER rating concepts).

Deviation regarding the initial idea of making a initial system with basic modules that would be later improved. Work is now focused on a more bottom-up approach, developing modules till good performance individualy.

[1] ”Tools for Dummies” in the initial proposal.

Chapter 4

Future work

Taking in consideration our results, our goal for 2023 was changed, but the main focus remained, that is, to study and create the building blocks to automate the capture and tagging of entities (NER), as we realize that small portions developed bring technological advances and possibilities to demonstrate the growth of the entire solution. Once we have a more robust developed, we can re-evaluate create a full pipeline and building the software. These software tools will combine parts already created with others not yet develop (ex: CAIT). To advance towards our goal the next steps still continued essential to develop the work:

- **Improve the NERs integrating the Ensemble:** This task has been maintained since the previous year, as this point should be prioritized this year. Particularly improving the performance of DBpedia-based NER, both in augmenting its speed (ex: with a local Virtuoso server) and improving the decision of Tags to keep from the DBpedia query results. Also the exploration of translation to increase the number of entities annotated has been started.

- **Addition of NERs to the Ensemble:** This task was kept but with a different focus, we have the idea of treating post-general training as UNICLASS. Add other NERs to the ensemble: For example, specialized NERs for specific classes (ex: time) or using other sources of information (ex: YAGO and Wikipedia) and ontologies.

- **Use tagged data to train NERs:** Continue and extend the work on training ML systems, starting with a NER based on BERT [21] ENTITY/NO-ENTITY tagging.

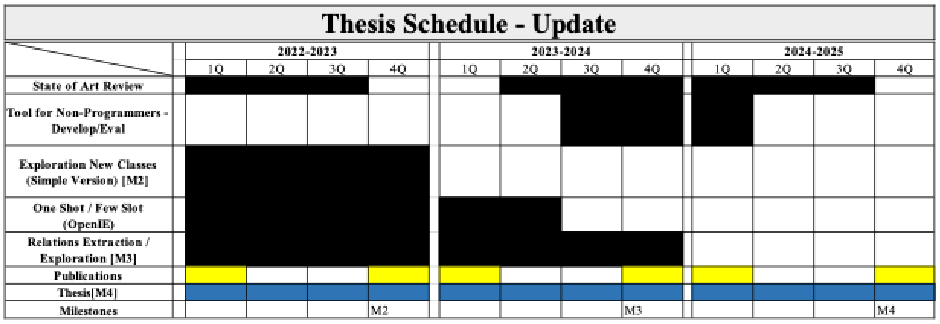
- **Improve the process of NERs output combination**. Implement more robust word/phrase alignment for NER assessment.

- **Integrate the newly obtained NERs in an Information Extraction base line**. At this point, we will develop a block that allows us, after complete post-processing, to use the results to extract some information from the text evaluated by the NERs.

Based on what happened in 2021/2022, working part-time, we believe it is possible to deliver everything we planned for this year, as the experience together with the support of supervisors created a balance that makes sense to maintain for the next period. With this, we reorganized the work plan that will be seen in the next chapter.

4.1 Updated Workplan

In figure 4.1 is presented the update worplan.



Thesis Schedule - Update

	2022-2023				2023-2024				2024-2025			
	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q
State of Art Review	■	■	■			■	■	■	■	■	■	
Tool for Non-Programmers - Develop/Eval							■	■	■			
Exploration New Classes (Simple Version) [M2]	■	■	■	■								
One Shot / Few Slot (OpenIE)	■	■	■	■	■	■						
Relations Extraction / Exploration [M3]	■	■	■	■	■	■	■	■				
Publications	🟨			🟨	🟨			🟨	🟨			🟨
Thesis[M4]	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Milestones				M2				M3				M4

Figure 4.1: Updated Workplan for next years.

We adjusted the main tasks to adhere more relevantly to the work plan.

Due to work outside the PhD and personal issues, still continued a working in a Part-Time, thus reconciling research, the tasks duration and start dates were updated accordingly.

Despite the extension of the time required, 2022 showed us that the balance between academic life and the professional side makes a lot of sense, as they are synergistic, as long as we have resilience in everyday life.

4.1.1 Milestones - Update

Each Milestone refers to a delivery and then will be necessary an update too , next the product or report as follows:

- M1 - delivery of the “NER without data registration”, planned to the end of the 1st complete year of Thesis work. Was achieved a bit later, in the final months of 2021.

- M2 - New conceptions of NER classifications planned for 4th quarter of 2022/2023.

- M3 - Exploration of other relationships and Information within the domain in the 4th Quarter of 23/24.

- M4 - must be the delivery of the document with the details of how the App was developed, the technology shipped and the results obtained the delivery planned for September 2025 .

Chapter 5

Conclusion

Comparing the results, we had a slice distance from the proposal, but not interfering in the proposed study context, on the contrary, we believe that this new vision was only possible through the studies that took place during the year. The proof of this new direction is right, we had the publication and presentation of PROPOR and the submission to IberSpeech in Granada.

In line with our overall goals, the plan for 2023 to 2025 has been adjusted, we will focus more on development from part to part and then, if possible, we will create an aggregator of all parts. With the combination strategy providing a single class annotation, designated as ENTITY, from now on we will have two major challenges, the first to create a model, or several, that identify each class/ENTITY in isolation, and the second to start modeling. to extract information automatically.

The results obtained between 2020 and 2021 revealed that the combination of outputs from the set of NERs has potential to extend the entity set regardless of the environment that was trained. So for us it makes sense to continue investing time and work in the search for a model that helps in the automatic annotation of Entities and later in Information Extraction.

With the current combination strategy providing a 1-class annotation, designated as ENTITY, we are going to tweak and study the creation of a model that is capable of detecting one class at a time, but on a recurring basis, thus carrying all the interesting aspects or important classes that were detected in the first scan of the current processing.

The research and experiments carried out between 2021 and 2022 were necessary and effective in developing both the validation of the concept that the future of OpenIA involves automatic annotation and the need for the “democratization” of knowledge, that is, realizing that OpenIA does not mean FreeIA, this undermines the basic concept of democratization. That said, we feel the need to deal with algorithms that can be developed within possibilities without financial or very low cost, because then we will have more people with the possibility to apply OpenIA.

Bibliography

[1] K. Sintoris and K. Vergidis. Extracting business process models using natural language processing (NLP) techniques. In *2017 IEEE 19th Conference on Business Informatics (CBI)*, volume 01, pages 135–139, 2017.

[2] Flora Amato, Giovanni Cozzolino, Vincenzo Moscato, and Francesco Moscato. Analyse digital forensic evidences through a semantic-based methodology and nlp techniques. *Future Generation Computer Systems*, 98:297–307, 2019.

[3] Antonio Moreno Sandoval, Julia Díaz, Leonardo Campillos Llanos, and Teófilo Redondo. Biomedical term extraction: Nlp techniques in computational medicine. *IJIMAI*, 5(4):51–59, 2019.

[4] Hiral Desai, Mohammed Firdos Alam Sheikh, and Satyendra K Sharma. Multi-purposed question answer generator with natural language processing. In *Emerging Trends in Expert Applications and Security*, pages 139–145. Springer, 2019.

[5] Robert Dale. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, 2021.

[6] Michael Casteluccio. A written test for artificial general intelligence. *Strategic Finance*, 102(5):53–54, 2020.

[7] Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10, 2022.

[8] Stella Biderman and Edward Raff. Neural language models are effective plagiarists. *arXiv preprint arXiv:2201.07406*, 2022.

[9] Pablo Gamallo, Marcos Garcia, César Pineiro, Rodrigo Martinez-Castano, and Juan C Pichel. Linguakit: a big data-based multilingual tool for linguistic analysis and information extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244. IEEE, 2018.

[10] Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, and Timothy Baldwin. Named entity recognition for novel types by transfer learning, 2016.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[12] A. Patel and A.U. Arasanipalai. *Applied Natural Language Processing in the Enterprise: Teaching Machines to Read, Write, and Understand*. O'Reilly Media, Incorporated, 2021.

[13] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese Named Entity Recognition using BERT-CRF, 2020.

[14] Emanuel Matos, Mário Rodrigues, Pedro Miguel, and António Teixeira. Towards automatic creation of annotations to foster development of named entity recognizers. In *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[15] Emanuel Matos, Mário Rodrigues, Pedro Miguel, and António Teixeira. Assessing transfer learning and automatically annotated data in the development of named entity recognizers for new domains. In *IberSpeech-2022*, 2022(submitted).

[16] António Teixeira, Pedro Miguel, Mário Rodrigues, José Casimiro Pereira, and Marlene Amorim. From web to persons - providing useful information on hotels combining information extraction and natural language generation. In *Proc. IberSpeech*, Lisbon, November 2016.

[17] Wikivoyage. `https://pt.wikivoyage.org/`.

[18] Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *PROPOR*, LNCS. Springer, 2018.

[19] C Mendonça Júnior, Hendrik Macedo, Thiago Bispo, Flávio Santos, Nayara Silva, and Luciano Barbosa. Paramopama: a brazilian-portuguese corpus for named entity recognition. *Encontro Nac. de Int. Artificial e Computacional*, 2015.

[20] Emanuel Matos, Mário Rodrigues, and António Teixeira. Named entity extractors for new domains by transfer learning with automatically annotated data. In *International Conference on Computational Processing of the Portuguese Language*, pages 288–298. Springer, 2022.

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