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

Progress Report – Year 1
2020-2021

Emanuel Matos

University of Aveiro
Portugal
September - 2021

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]. Despite their potential, creation of such systems for new domains is not simple and, in general, development has been limited to developers with in-depth knowledge of the area. As extraction methods improve, the amount of information also increases, making more and more difficult an effective access by humans. Currently, the number of Named Entity Recognizers (NER) that are publicly available have not yet been systematically used in synergically adapted works. Despite a good number of databases, for our initial phase we resort to a dataset created from texts collected from a wiki in Internet. After, and together with this dataset, we started to develop a concept where this text would pass through several NERs, the result of which will serve as training data for new Machine Learning NERs, such as a state-of-the-art BERT-based NER.

With this scenario in mind, our initial proposal also considers the development of an easy to use application to access the information alongside the evolution of the state-of-the-art in Open Information Extraction. An easy to use application to access information is key for the democratization of the access to knowledge. Within what we were expected to accomplish, we succeeded in the vast majority of points. But we have refer some situations that impacted the speed of both production and research. The COVID-19 that got more severe than expected, my return to Brazil due to family problems, my livelihood (full time job), all these items led me to reduce my working time.

The main objective of the PhD is to investigate new methodologies that allow the democratization of the exploitation of information on any basis in an environment focused on Tourism.

This report describes the work developed during the 1st year of the MAP-i Doctoral Program.

Chapter 2

Developments

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

- Beginning of the implementation of the assembly for automatic annotation, based on the premises that we have listed in the definition of the work on the pre-thesis. In parallel we have started structuring the NER's to be included in the initial version of the assembly. More precisely we have adjusted Linguakit that was the first NER to be used. We have elaborated on some basic functions in order to have a set that can be treated in a simpler way, but not with low quality. Linguakit is based on pre-annotations and lists.

- In Nov/2020 we have identified the second NER, which was extracted from the ALLENnlp system. This NER, despite being based on neural networks (NN), provides results in a format easy to interpret and integrate with other NERs, as it adopts the BIO (short for beginning, inside, outside) system. Already in Dec/2020, with the 2 systems adjusted to work in Windows, both were integrated into Python scripts, a relevant step for the creation of an initial Proof of Concept. Working at the junction of these 2 NER's we have already visualized the need for a third NER for "tiebreaker". DBPedia was chosen as the basis for the creation on a third NER.

- Jan/2021, we have put Dbpedia-based NER to work. This took much longer than the others due to interaction via Internet, and its way of processing took much longer than the initial expectation.

- Feb/2021, with the 3 NER's running on separate data, we started debugging and improving the pipeline developed. We have conducted a comparative evaluation and development of a decision module capable of implementing different strategies on the outputs of the 3 individual NERs (ex: Winner Take All).

- In Mar/2021 we identified and started the process of extraction and cleaning data for a new Dataset for Tourism, adopting as source Wikivoyage.

- After April, and until the end of June 2021, the work was concentrated in the alignment of individual NERs output, evolution of decision strategies - particularly development of a strategy providing a binary output regarding being an ENTITY or not-, and evaluation of the concept. The new dataset was used to test the first proof-of-concept system and the results published in a conference paper in Jul/2021. This evaluations had as secondary result the production of automatically annotated data that can foster development of Machine Learning based NERs. During the process of preparing the paper, an extra effort was made in the continuous update of the state-of-the-art in NERs and NERs for Portuguese.

- In August/2021, in parallel to the continuation of processing of Wikivoyage texts to increase as much as possible the automatically annotated dataset, we started the tests with a new NER, a BERT-base one.

- This September, to make possible evaluation with established gold standards, we are creating the necessary conditions to use HAREM [1] datasets to evaluate our BERT-based NER, in order to have some comparative evaluation.

The next section presents additional information on 2 of the developments:

2.1 State-of-the-art update

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:

[1] https://www.linguateca.pt/HAREM/

Rule-based NER – covering both systems based in patterns and in lists (the so-called Gazetteers).

Surface Patterns – Surface patterns NER are usually implemented using regular expressions, or regex for shorthand, which are sequences of characters that specify search patterns. They are easy to implement but are quite sensitive to errors as they can break patterns and thus cause detection failures. Also, improvements come with increasing complexity as rules can interact, and their order can be meaningful, which makes more difficult to manage the rule set [5]. One of the main challenges of creating handcrafted rules is that it can be very time consuming to compile a comprehensive set of rules when target entities do not have well-defined and mandatory surface patterns. It can also be difficult to port the solution to other application contexts. For entities with well-defined surface patterns this approach is often easy to implement and provides reliable results.

Gazetteers – A gazetteer is a geographical dictionary or directory. In NLP context, the term gazetteer was further extended and now means a list of items that often include organizations, people names, alongside geographical entities such as cities or landmarks. Approaches using gazetteers can be as simple as matching candidate portions of text against the lists and having the decision just based on the existence in the list. More sophisticated methods of using gazetteers include using them as triggers in which a keyword can be used to find an entity (for example, Ms. can be used to identify that the next statement is a person), This approach is easy and can get rather good results. Unfortunately, the creation and maintenance of the lists can be a hard and tedious process and it also has problems with ambiguity (ex: gate can be an object or a name of a person depending on the context).

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 [6] 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:

Hidden Markov Model (HMM) – HMM is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobservable, hidden, states. There is another process, visible, whose behavior depends on the underlying hidden process. The goal is to learn about the hidden process by observing the visible one. NER is considered as a classification problem where the named entity class is the hidden part and the textual form the visible one. The goal is to decide which word is part of some name or not part of any name, and only one label can be assigned to a word in each context. Therefore, the model assigns to every word either one of the desired classes or a label representing none of the desired classes.

Support Vector Machines (SVM) – SVM are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Support vectors are data points that are closer to the decision boundaries that help classify the data points. In NER, SVM classifies each word, using its context, to one of the classes that represent region information and named entity's semantic class.

Conditional Random Field (CRF) – In CRF, the prediction implements dependencies between predictions. In NER, a linear chain CRF the decision if a word belongs to a named entity class, or not, depends only the word itself and on the information of the previous word. The CRF approach utilizes conditional probability to train untrained data using a trained data set.

Recursive Neural networks – Most neural-based models for NER are based on some sort of Long Short Term Memory (LSTM). LSTM are recursive neural networks in which the hidden layer updates are replaced by

purpose-built memory cells. As a result, they may be better at finding and exploiting long range dependencies in data [7, 8]. Bidirectional LSTM are amongst the best performers and in these, word and character embeddings are passed through a left-to-right LSTM and a right-to-left LSTM. The outputs are combined to produce a single output layer. In the simplest method, this layer can then be directly passed onto a softmax that creates a probability distribution over all NER tags, and the most likely tag is chosen.

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 [9].

Transformers – Transformers [10], 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) [11]. 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: [12]).

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 [13] 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 overview of the proposed process for automatic tagging of named entities by using and Ensemble of NERs, showing its 4 phases.

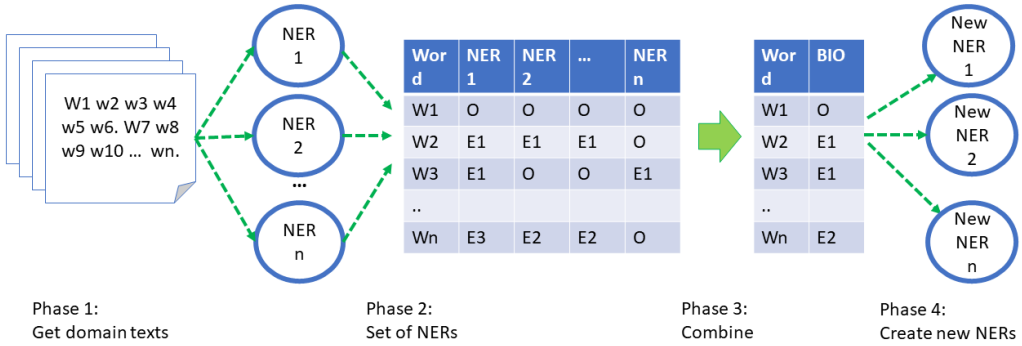

Figure 2.1: Caption From [13]

2.3 Dataset(s) for 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 [14], 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 [15], 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.

2.4 BERT-based NER trained with automatically annotated data

Another challenge we start 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.

Chapter 3

Results

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

3.1 Publications

3.1.1 Published

Based in the work of this first year, it was possible a publication in SLATE Conference in July 2021, entitled "Towards Automatic Creation of Annotations to Foster Development of Named Entity Recognizers" [13], having as authors Emanuel Matos, Mário Rodrigues, Pedro Miguel, António Teixeira.

Abstract: 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 this paper 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. This 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.

Link: `https://doi.org/10.4230/OASIcs.SLATE.2021.11`

3.1.2 In preparation

The work regarding the development of the BERT-based NER with automatically annotated train data is the basis for an article in preparation. We plan to submit it to PROPOR 2022, having submission deadline the end of September/21.

3.2 Software

We also started the development of the software "Tools for non-programmers" [1], still in an embryonic state, and joining several parts (scripts) that we used to implement the first proof-of-concept to make the process of processing new texts and perform entity recognition as simple as possible.

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. 4.2).

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

Chapter 4

Future work

Our goal continues to be the creation of "Tools for non-programmers"[1]. These software tools will combine parts already created with others not yet develop (ex: CAIT). To advance towards our goal the next steps are essential:

- **Improve the NERs integrating the Ensemble:** 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:** 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 [16] ENTITY/NO-ENTITY tagging.

- **Improve the process of NERs output combination**.

- **Test the system in new domains**.

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

- **Integrate the newly obtained NERs in an Information Extraction pipeline**.

4.1 Workplan Update

In figure 4.1 is presented the initial worplan, presented last year, and figure 4.2 the proposed revision to the workplan.

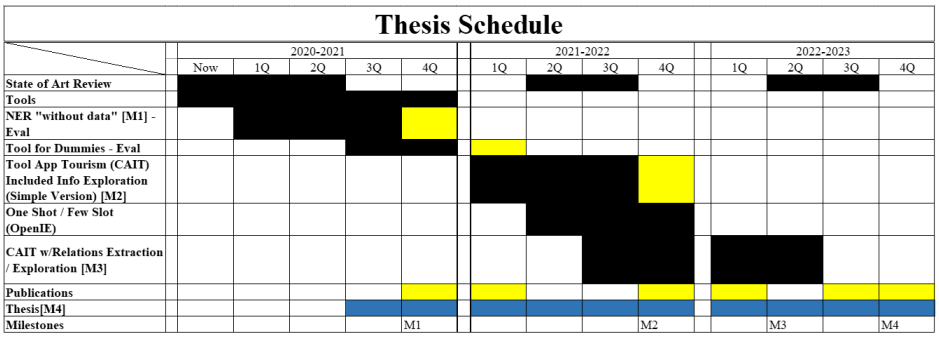

Figure 4.1: Timetable from Pre-Thesis

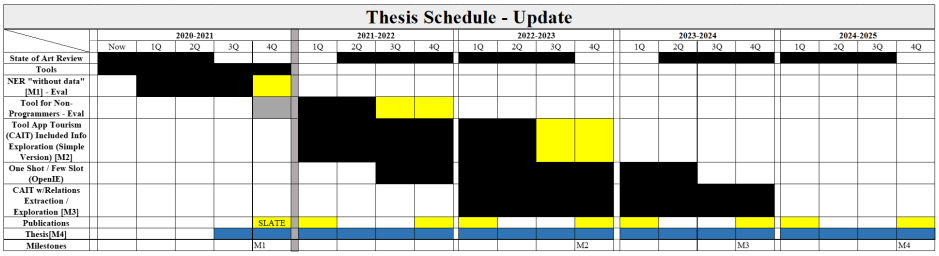

Figure 4.2: Timetable Thesis Schedule - Update.

We don't see for now reason for change in the main tasks integrating the workplan.

Due to work outside the PhD and personal issues, requiring a change in dedication to the PhD from Full-Time to Part-Time, thus reconciling research, the tasks duration and start dates were updated accordingly.

Despite the extension of the time required, I consider this can be opportunity for synergy with my professional work and increases the possibility to complement the Research work with the pragmatism of the industry.

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 - which is the 1st milestone is like the delivery of the “NER without data registration”, this delivery will take place at the end of the 1st year, that is, until 20/21 Sep. - OK.
- M2 - will be the delivery of the CAIT App in a simpler version around 4Q on ~~21/22~~ 22/23.
- M3 - Will already be the delivery of CAIT in a version with Exploration of other relationships within the domain in ~~2Q to 22/23~~ 4Q to 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 will be in ~~September 2023~~ September 2025 .

Chapter 5

Conclusion

Aligned with our objectives and plan for 2020-2021, we were able to create a proof-of-concept for a NER without annotated data, succeeding in our first Milestone (M1) essentials, despite the negative effects of the pandemic situation.

The initial results revealed that combining the outputs of the NERs ensemble have potential to both extend the set of entities (limited in systems as AllenNLP) and contribute to reduce the difficulties distinguishing some entities (ex. AllenNLP and Linguakit present high degree of disagreement for PERSON and LOCATION).

With the combination strategy providing a 1-class annotation, designated as ENTITY, the proposed system could detect an interesting set of Named Entities in texts related to Tourism that were used to create initial versions of new Gazetteer-based NERs. The other combination strategy, WTA, is useful to create higher confidence annotations for a set of entities, such as those handled by Linguakit.

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] Tobias Ek, Camilla Kirkegaard, Håkan Jonsson, and Pierre Nugues. Named entity recognition for short text messages. *Procedia - Social and Behavioral Sciences*, 27:178–187, 2011.

[6] Pablo Gamallo, Marcos Garcia, César Piñeiro, Rodrigo Martínez-Castaño, and Juan Pichel. Linguakit: A big data-based multilingual tool for linguistic analysis and information extraction. 10 2018.

[7] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[8] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

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

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

[11] 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.

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

[13] 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.

[14] 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.

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

[16] 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.