Named Entity Recognition Approaches Applied to Legal Document Segmentation

F. X. B. da Silva[1], G. M. C. Guimaraes[1], R. M. Marcacini[2], A. L. Queiroz[1], Vinicius R. P. Borges[1], T. P. Faleiros[1], L. P. F. Garcia[1]

[1] Universidade de Brasília, Brazil
{felipe.barbosa, gabriel.ciriatico}@aluno.unb.br
{andreiqueiroz, viniciusrpb, thiagodepaulo, luis.garcia}@unb.br
[2] Universidade de São Paulo, Brazil
ricardo.marcacini@icmc.usp.br

Abstract. Document Segmentation is a method of dividing a document into smaller parts, known as segments, which share similarities that allow machines to distinguish between them. It might be useful to classify these segments, making it a problem with two steps: (*i*) the extraction of the segments; and (*ii*) the annotation of these segments. The Named Entity Recognition problem's goal is to identify and classify entities within a text, having also to deal with those two questions: extraction and classification. In this study, we tackle the problem of Document Segmentation and the annotation of these segments through NER approaches, using CRF, CNN-CNN-LSTM and CNN-biLSTM-CRF models. The study is focused on Brazilian legal documents, proposing a data set of 127 annotated Portuguese texts from the Official Gazette of the Federal District, published between 2001 and 2015. The experiments were made using word-based and sentence-based models, with CRF sentence-based model showing the best results.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: Legal documents, Named Entity Recognition, Segmentation

1. INTRODUCTION

Among legal documents published by States, government gazettes are some of the most important ones, providing updated information on a nearly daily basis. Its text is usually divided into segments, defined by topics such as contracts, public service hiring, normative acts issued by presidents etc.[1] This type of document can benefit from tools to automatically extract information through Machine Learning (ML), since it is susceptible to changes that make traditional approaches such as regular expression searches inadequate.

In Brazil, this is a specially important source of information, and it has been published by the Brazilian Government since 1862, without interruption. The full text of the following is mandatory for publication in the Federal Official Gazette: laws, amendments to the Constitution, legislative decrees, and other acts resulting from the legislative process; treaties, agreements, covenants; decrees, provisional measures, and other normative acts issued by the President of the Republic; among others [Passos 2002].

This kind of gazette exists in every Brazilian state or district, with the same information (adapted to local specificities). Since its foundation, in 1960, the Federal District Government publishes its official

[1] https://portal.in.gov.br/

gazette, the Official Gazette of the Federal District (OGFD) [2]. Editions published during the first seven years are not available on the Internet, and the editions published between October 1967 and April 2020 can be downloaded only in PDF format, without any segmentation between the different topics in the document. Since May 2020, OGFDs can be found on the Internet in text format, divided into segments called acts.

Our goal here is to use ML models to extract and annotate segments from the OGFDs, since the data is not structured and regular expressions searches are unable to perform satisfactorily [Luz de Araujo et al. 2020]. These models need to deal with two questions: (*i*) document segmentation; and (*ii*) segmentation annotation. Traditionally, these two problems are dealt with separately: first, a document segmentation model is applied; and then, a text classifier is applied to classify the segments. This study examines the usefulness, in terms of accuracy, of tackling both problems at the same time, given that it is easier to classify a segment, and also easier to segment a text already classified [Barrow et al. 2020].

To deal with these two problems, we resort to Named Entity Recognition (NER) models. The NER problem is to identify entities (a segment composed of words) in a text and classify them. At its core, it has the same two steps as our problem: extraction and classification of words. This paper shows that document segmentation can benefit from NER techniques, although it requires changes since NER usually deals with smaller texts than document segmentation. The emphasis will be on ML models, from classical techniques, such as Conditional Random Field (CRF) [Lafferty et al. 2001], to more advanced approaches, such as Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber 1997] and Convolutional Neural Network (CNN) [LeCun et al. 1989].

This work has two main contributions to literature. First, we provide a data set of 127 segmented documents annotated by humans, in a field where the majority of the available data sets are smaller and automatically annotated. And second, multiple CRF-based techniques for NER will be evaluated on the extraction of text segments and compared, contributing to the literature on how NER models can be used in the text segmentation task.

The structure of this paper is as follows. Section 2 will cover the existing background on text segmentation and NER and other correlated subjects. Section 3 will describe the methodology used in this research. Section 4 will include a presentation and discussion of the results obtained. Finally, Section 5 will conclude this work and present the final remarks.

2. RELATED WORK

2.1 Document Segmentation

Document segmentation is a method of dividing a document into smaller parts, typically known as segments. Segments can be split into tokens, and each token can be categorized as a word, phrase, topic, sentence, or any unit of information that represents a subset of the document. Each segment unit has its relevant meaning, and its meaning is closely related to the sequence of tokens.

Dcument segmentation models use coherence to detect different segments in a text [Barrow et al. 2020]. Term co-occurrences were used in the TextTiling algorithm [Hearst 1997]. Bayesian methods were used successfully within the BayesSeg algorithm [Eisenstein and Barzilay 2008], among others [Riedl and Biemann 2012]. Another popular approach to this problem uses unsupervised algorithms, like GraphSeg [Glavaš et al. 2016], where a graph is built using sentences as nodes and semantic similarity is represented by an edge.

Newer models have been proposed, outperforming these traditional approaches. [Koshorek et al. 2018] present, based on a data set of Wikipedia articles, a neural model consisting of a hierarchy of

[2] `https://www.dodf.df.gov.br/`

two sub-networks, both based on the LSTM architecture. The model generalizes well to natural text not previously seen.

[Glavaš and Somasundaran 2020] propose the Coherence-Aware Text Segmentation (CATS) model, which produces state-of-the-art segmentation performance on a collection of benchmark data sets. It is a multi-task learning model, based on a neural architecture consisting of two hierarchically connected Transformer networks, which couples the sentence-level segmentation goal with the coherence goal that differentiates correct sentence sequences from corrupt ones. The model can successfully segment texts in languages not seen in training. The model has also been proven efficient in zero-shot language transfer experiments.

[Aumiller et al. 2021] present a segmentation approach that can predict the topical coherence of sequential text segments spanning multiple paragraphs, effectively segmenting a document and providing a more balanced representation for downstream applications. The approach is based on transformer networks and structural text segmentation, formulated as topical change detection and performing a series of independent classifications that allow efficient tuning on task-specific data.

2.2 Segmentation as a NER Problem

The NER task is a sequence labeling problem where each word in a sentence must be classified into one of many predefined categories. Therefore, NER data sets come with word-level annotations, assigning each word its true label. This annotation is usually done with Inside-Outside-Beginning (IOB) tagging, where a tag O indicates a token that belongs to no chunk, and I-X indicates that the word is within a specific chunk (X might be any classification).

NER can be applied in the most diverse domains, for many different purposes. Among the most common applications of NER, one can highlight information extraction, question and answer, text summarization, information retrieval, text translation, text mining etc. [Goyal et al. 2018]. It was used to extract information from historical manuscripts without the need for an intermediate transcription step [Toledo et al. 2019]. A parallel computing model based on Deep Attention was developed, evaluated using the generic English (CoNLL2003) and Chinese (MSRA) databases [Liu et al. 2019].

Segmentation can be seen as a NER problem when it is considered to have two steps: to find segments in a text; and to classify these segments. Those two steps are the steps required to tackle the named entity recognition problem. Successful attempts have been made to combine NER and document segmentation algorithms and techniques to improve the segmentation task results.

[Fragkou 2015] used the [Choi 2000] data set (the original and a modified version, with manual and automatic annotations) to test two different named entity algorithms, the Affinity Propagation algorithm and the MinCutSeg algorithm. The first one considers the measures of similarity between pairs of data points, while the second deals with segmentation as a graph task, measuring the similarity between partitions. The output of these NER algorithms was then used as input to document segmentation algorithms, combining both techniques. [Xu et al. 2013] proposed a joint segmentation and named entity recognition model, where the segmentation of a document was made simultaneously with the entity recognition of its content.

[Tepper et al. 2012] went beyond making joint models using NER and document segmentation, using NER models and techniques to segment documents. The segmentation was done over a clinical records data set, where the sections were also classified into several categories. The data set was explored in two different approaches: a one-step approach, where the segmentation was done together with the segmentation classification; and a two-step approach, where the segmentation was first done and then the classification. The accuracy of the two approaches was similar, showing the validity of the use of the one-step approach.

[Barrow et al. 2020] presented a model capable of learning segmentation boundaries and segmenta-

tion level labels together at training time. Segment Pooling LSTM (S-LSTM) is a supervised model based on an LSTM architecture trained both to predict segment boundaries and pool over and classify segments. In support of joint training, an approach was developed to teach the model to recover from errors by aligning predicted segments and ground truth. This approach both segments the document and annotates its segments, but without using NER models.

3. METHODOLOGY

The experiments performed in this research had the main task of correctly extracting segments from a group of OGFDs. To accomplish this, a new data set was made, resulting in 127 manually annotated documents, publicly available. Word-based and sentence-based models' capacity of extracting these segments was evaluated. In word-based models, each word from the document has an individual label in IOB format, while in the sentence-based models a single label was assigned to each sentence from the document (also using the IOB format).

3.1 Data set

The OGFDs texts are composed of 3 sections: Section I, where laws, government decrees and ordinances are published; Section II, where there is information about civil services, such as retirement and allowance; and Section III, where financial information can be found. The data set used in this paper focuses only on Section II.

A total of 127 gazettes had the second section extracted, whose subsections, also called acts, were then manually labeled. The acts usually begin with specific words that allow the annotator to identify where it starts. A total of 9058 acts were annotated, divided into 12 types. Subsections of the documents that were not part of any act type were removed, such as the beginning and the footer of the documents. The documents were published between 2001 and 2015.

Act type	**Number of acts**
Abono de Permanênica (Permanence Allowance)	134
Cessão (Cession)	265
Exoneração Comissionado (Dismissal of Commissioned Position)	2009
Exoneração Efetivo (Dismissal of Effective Position)	241
Nomeação Comissionado (Nomination of Commissioned Position)	2314
Nomeação Efetivo (Nomination of Effective Position)	46
Retificação Comissionado (Rectification of Comissioned Appointment)	181
Retificação Efetivo (Rectification of Effective Appointment)	1235
Reversão (Reversal)	58
Substituição (Substitution)	2352
Tornado Sem Efeito Apo (Rendered Ineffective Retirement Acts)	20
Tornado Sem Efeito Exo/Nom (Rendered Ineffective Dismissal or Nomination Acts)	250
Total	9058

Table I: Total of labeled acts available in the data set. In parentheses, there is the English translation to those acts.

3.2 Word-based and sentence-based models

The word-based models were CRF, CNN-CNN-LSTM [Shen et al. 2017], and CNN-biLSTM-CRF [Ma and Hovy 2016]. The CRF model was generated using the *sklearn-crfsuite* library and the *lbfgs* algorithm, with the input features of each word being: (*i*) the word itself, in lower case; (*ii*) whether or not the word is a title or in upper case, amount of digits in the word, and; (*iii*) all the previous

[2]The data set can be accessed at: `https://github.com/UnB-KnEDLe/persoseg-corpus`
[2]`https://github.com/TeamHG-Memex/sklearn-crfsuite`

items repeated for nearby words. The CNN-CNN-LSTM used 50 channels in the first CNN to create character-level embeddings, 800 channels in the second CNN to create word-level embeddings, and 200 as the size of the hidden layer of the LSTM. It also used a pre-trained GloVe [Pennington et al. 2014] embedding with 50 dimensions, a fixed learning rate of 10^{-3}, and the *Adam* optimization algorithm. The CNN-biLSTM-CRF had a similar configuration, with the main difference being that it did not create word-level embeddings.

For the sentence-based models, the word-based models mentioned in the previous section were adapted to work with sentence labels rather than word labels. The CNN-CNN-LSTM and CNN-biLSTM-CRF architectures had this adaptation done after the CNN layer, which generates an embedding for the words. This embedding for words was then converted to an embedding for the sentence using a single LSTM and taking the final cell state. For the CRF architecture, the adaptation was done by combining the features of only the first four words and the last three words of each sentence, discarding the rest. The features obtained for each of these words were the same as in the word-based version of the model.

The data set used for training was the DODF Corpus, with 5-fold cross-validation. For the deep learning models, 20% of the training set was separated and used as a validation set. Early stopping was also used for both CNN-CNN-LSTM and CNN-biLSTM-CRF, stopping the training iteration after 40 epochs without a reduction of loss in the validation set.

3.3 Evaluations

After training, each model was evaluated on the testing set, labeling every word (by the word-based models) or every sentence (by the sentence-based models). The final score of a model was defined as the weighted average of the f1-score for the types B, I and E, with the weight being the frequency of each type in the set. The label O was excluded to avoid skewing the final score, as this label is much more frequent than the others, and a high score on this label does not indicate that the model has learned anything.

Additionally, each experiment was conducted using a 5-fold split, alternating the train and test sets. As a result, each experiment has 5 final scores, from which a mean and a standard deviation were calculated. The mean score is the metric used to compare these models, while the standard deviation indicates the impact of the train/test set selection on the results. A Nemenyi test was also performed to compare the performance of different model architectures.

4. EXPERIMENTAL RESULTS

4.1 Word-based models

The f1-scores obtained by the word-based models for each act type can be seen in Table II. The CRF model performed better in almost every act type, with the only exceptions being Permanence Allowance e Rendered Ineffective Dismissal or Nomination Acts. CNN-CNN-LSTM and CNN-biLSTM-CRF performed slightly worse.

The low overall performance for the act types Rendered Ineffective Retirement Acts, Reversal and Nomination of Effective Position can be partially explained by the fact that they have very few examples in the data set. Act types with the highest amount of documents (more than a thousand) all achieved more than 90% accuracy in all models. These results indicate that having exposure to more examples might be particularly important in this task.

Act type\Model	CNN-CNN-LSTM	CNN-biLSTM-CRF	CRF
Permanence Allowance	$\mathbf{0.736 \pm 0.141}$	0.53 ± 0.443	0.197 ± 0.394
Cession	0.956 ± 0.016	0.966 ± 0.021	$\mathbf{0.991 \pm 0.007}$
Dismissal of Commissioned Position	0.976 ± 0.005	0.986 ± 0.005	$\mathbf{0.995 \pm 0.001}$
Dismissal of Effective Position	0.84 ± 0.093	0.922 ± 0.028	$\mathbf{0.976 \pm 0.014}$
Nomination of Commissioned Position	0.976 ± 0.014	0.992 ± 0.007	$\mathbf{0.997 \pm 0.001}$
Nomination of Effective Position	0.0 ± 0.0	0.0 ± 0.0	$\mathbf{0.923 \pm 0.071}$
Rectification of Commissioned Appointment	0.79 ± 0.043	0.778 ± 0.081	$\mathbf{0.864 \pm 0.029}$
Rectification of Effective Appointment	0.912 ± 0.044	0.96 ± 0.011	$\mathbf{0.962 \pm 0.006}$
Reversal	0.0 ± 0.0	0.0 ± 0.0	$\mathbf{0.07 \pm 0.14}$
Substitution	0.984 ± 0.008	0.992 ± 0.007	$\mathbf{0.998 \pm 0.002}$
Rendered Ineffective Retirement Acts	0.0 ± 0.0	0.0 ± 0.0	$\mathbf{0.127 \pm 0.253}$
Rendered Ineffective Dismissal or Nomination Acts	0.95 ± 0.021	$\mathbf{0.99 \pm 0.006}$	0.979 ± 0.013
Average	0.676 ± 0.051	0.677 ± 0.032	$\mathbf{0.757 \pm 0.078}$

Table II: Results of word-based models.

Figure 1 shows the average precision and recall for each model and act type combination. The black dotted lines represent f1-score isolines: the leftmost one represents an f1-score of 0.1, the next one represents an f1-score of 0.2, and so on. The f1-score value of 1 would be represented by a single dot in the point (1.0, 1.0) of the graph. Each color represents a model, and each one has 12 points in total in the graph, one for each act type.

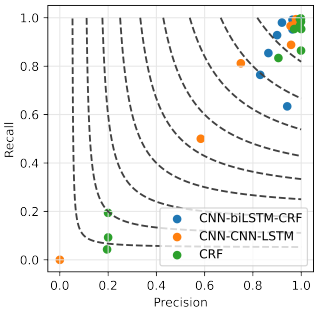

Fig. 1: Results obtained by word-based models in terms of their precision and recall.

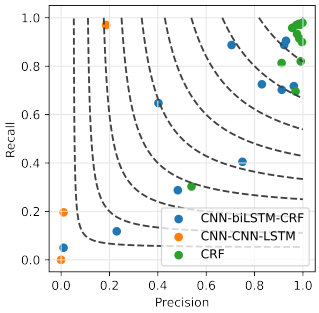

Fig. 2: Results obtained by sentence-based models in terms of their precision and recall.

For the most part, the models had similar values of precision and recall. The largest difference was for the CNN-biLSTM-CRF model with the act type Permanence Allowance, which had a precision score of 0.942 and a recall score of 0.634. All significant differences (above 0.1) in precision and recall were caused by models that had higher precision than recall, indicating a comparatively high number of false positives.

4.2 Sentence-based models

Table III shows the f1-scores obtained by the sentence-based version of the models. The CNN-CNN-LSTM failed to learn for most act types, and the CNN-biLSTM-CRF models performed worse than their word-based counterparts overall. It might be that the algorithm used for segmentation increased the difficulty of the task rather than lowering it, due to separating segments that were important for proper classification. The CRF model, however, still achieved high scores and had an average f1-score significantly higher than any of the word-based models. Once again, it is likely that the CRF models had an advantage due to their access to capitalization features.

Act type\Model	**CNN-CNN-LSTM**	**CNN-biLSTM-CRF**	**CRF**
Permanence Allowance	0.000 ± 0.000	0.780 ± 0.166	$\mathbf{0.948 \pm 0.013}$
Cession	0.000 ± 0.000	0.918 ± 0.043	$\mathbf{0.979 \pm 0.005}$
Dismissal of Comissioned Position	0.000 ± 0.000	0.155 ± 0.268	$\mathbf{0.967 \pm 0.011}$
Dismissal of Effective Position	0.018 ± 0.022	0.018 ± 0.030	$\mathbf{0.945 \pm 0.014}$
Nomination of Comissioned Position	0.312 ± 0.010	0.775 ± 0.147	$\mathbf{0.977 \pm 0.008}$
Nomination of Effective Position	0.000 ± 0.000	0.760 ± 0.097	$\mathbf{0.895 \pm 0.042}$
Rectification of Comissioned Appointment	0.000 ± 0.000	0.485 ± 0.300	$\mathbf{0.846 \pm 0.019}$
Rectification of Effective Appointment	0.000 ± 0.000	0.903 ± 0.031	$\mathbf{0.953 \pm 0.002}$
Reversal	0.000 ± 0.000	0.332 ± 0.372	$\mathbf{0.803 \pm 0.044}$
Substitution	0.000 ± 0.000	0.735 ± 0.368	$\mathbf{0.987 \pm 0.002}$
Rendered Innefective Retirement Acts	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{0.360 \pm 0.137}$
Rendered Innefective Dismissal or Nomination Acts	0.000 ± 0.000	0.492 ± 0.154	$\mathbf{0.957 \pm 0.014}$
Average	0.028 ± 0.003	0.529 ± 0.165	$\mathbf{0.885 \pm 0.026}$

Table III: Results of sentence-based models.

Figure 2 shows the average precision and recall for each model and act type combination. The largest difference was for the CNN-CNN-LSTM model with the Nomination of Commissioned Position act type, which achieved a recall of 0.970 and a precision of 0.186, indicating multiple false negatives. The differences in precision and recall were overall higher for the sentence-based models than for the word-based models.

To evaluate the statistical significance of the experimental results, the Friedman and the Nemenyi post-hoc statistical tests with 95% of confidence level were applied to compare the predictive performances of the word-based and sentence-based models. The CRF architecture had the best performance, followed by CNN-CNN-LSTM for the word-based models and CNN-biLSTM-CRF for the sentence-based models. Since all rank differences were higher than CD, they are all considered to be statistically significant.

5. CONCLUSIONS

OGFDs represent an important source of information about the Federal District Government's activities. Its content is usually composed of acts that fit in different categories. In this paper, the performance of three different models was compared for the task of document segmentation in 12 act types: a classical CRF architecture, CNN-CNN-LSTM and CNN-biLSTM-CRF. While these models were made for taking words as inputs, they were here adapted to sentence-based versions and also compared to their word-based counterparts.

For the word-based models, the CRF model performed significantly better than the remaining two, achieving an average f1-score of 0.757. Although the CNN-CNN-LSTM and CNN-biLSTM-CRF had similar averages (0.676 and 0.677, respectively), the Nemenyi test indicates that the CNN-CNN-LSTM performed better. For the sentence-based models, the CNN-CNN-LSTM architecture was unable to learn any act type properly, and had an average f1-score of just 0.028. The CNN-biLSTM-CRF had an average of 0.529, also lower than its word-based version. The sentence-based CRF, however, achieved the best score among all models, with an average f1-score of 0.885. Access to capitalization features may have given the CRF architecture an advantage over the other two.

6. ACKNOWLEDGMENT

The authors would like to thank FAPDF project KnEDLe (grant 07/2019).

REFERENCES

Aumiller, D., Almasian, S., Lackner, S., and Gertz, M. Structural text segmentation of legal documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ICAIL '21. Association for Computing Machinery, New York, NY, USA, pp. 2–11, 2021.

Barrow, J., Jain, R., Morariu, V., Manjunatha, V., Oard, D., and Resnik, P. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pp. 313–322, 2020.

Choi, F. Y. Y. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*. NAACL 2000. Association for Computational Linguistics, USA, pp. 26–33, 2000.

Eisenstein, J. and Barzilay, R. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pp. 334–343, 2008.

Fragkou, P. Use of named entity recognition and co-reference resolution tools for segmenting english texts. In *Proceedings of the 19th Panhellenic Conference on Informatics*. PCI '15. Association for Computing Machinery, New York, NY, USA, pp. 331–336, 2015.

Glavaš, G., Nanni, F., and Ponzetto, S. P. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, Berlin, Germany, pp. 125–130, 2016.

Glavaš, G. and Somasundaran, S. Two-level transformer and auxiliary coherence modeling for improved text segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (05): 7797–7804, apr, 2020.

Goyal, A., Gupta, V., and Kumar, M. Recent named entity recognition and classification techniques: A systematic review. *Comput. Sci. Rev.* vol. 29, pp. 21–43, 2018.

Hearst, M. A. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23 (1): 33–64, 1997.

Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation* 9 (8): 1735–1780, 11, 1997.

Koshorek, O., Cohen, A., Mor, N., Rotman, M., and Berant, J. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pp. 469–473, 2018.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289, 2001.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1 (4): 541–551, 1989.

Liu, X., Yang, N., Jiang, Y., Gu, L., and Shi, X. A parallel computing-based deep attention model for named entity recognition. *The Journal of Supercomputing* 76 (2): 814–830, sep, 2019.

Luz de Araujo, P. H., de Campos, T. E., and Magalhães Silva de Sousa, M. Inferring the source of official texts: Can svm beat ulmfit? In *Computational Processing of the Portuguese Language*, P. Quaresma, R. Vieira, S. Aluísio, H. Moniz, F. Batista, and T. Gonçalves (Eds.). Springer International Publishing, Cham, pp. 76–86, 2020.

Ma, X. and Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf, 2016.

Passos, E. Doing legal research in brazil, 2002.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543, 2014.

Riedl, M. and Biemann, C. TopicTiling: A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*. Association for Computational Linguistics, Jeju Island, Korea, pp. 37–42, 2012.

Shen, Y., Yun, H., Lipton, Z., Kronrod, Y., and Anandkumar, A. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Vancouver, Canada, pp. 252–256, 2017.

Tepper, M., Capurro, D., Xia, F., Vanderwende, L., and Yetisgen-Yildiz, M. Statistical section segmentation in free-text clinical records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, pp. 2001–2008, 2012.

Toledo, J. I., Carbonell, M., Fornés, A., and Lladós, J. Information extraction from historical handwritten document images with a context-aware neural model. *Pattern Recognition* vol. 86, pp. 27–36, 2019.

Xu, Y., Wang, Y., Liu, T., Liu, J., Fan, Y., Qian, Y., Tsujii, J., and Chang, E. I. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. *Journal of the American Medical Informatics Association* 21 (e1): e84–e92, 08, 2013.