Leveraging Probabilistic Graph Models in Nested Named Entity Recognition for Polish

Jędrzej Jamnicki
Wrocław University of Science and Technology Wrocław, Poland
jedrzej.jamnicki@pwr.edu.pl

Abstract

This paper presents ongoing work on leveraging probabilistic graph models, specifically conditional random fields and hidden Markov models, in nested named entity recognition for the Polish language. NER is a crucial task in natural language processing that involves identifying and classifying named entities in text documents. Nested NER deals with recognizing hierarchical structures of entities that overlap with one another, presenting additional challenges. The paper discusses the methodologies and approaches used in nested NER, focusing on CRF and HMM. Related works and their contributions are reviewed, and experiments using the KPWr dataset are conducted, particularly with the BiLSTM-CRF model and Word2Vec and HerBERT embeddings. The results show promise in addressing nested NER for Polish, but further research is needed to develop robust and accurate models for this complex task.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP) that involves identifying and classifying named entities, such as names of people, organizations, locations, and more, within text documents. The ability to accurately extract and categorize these entities plays a crucial role in various NLP applications, including information retrieval, question answering, text summarization, and machine translation. NER serves as the foundation for understanding the semantics and context of textual data, enabling more advanced language understanding systems.

In the realm of NER, there exists a more intricate and challenging variant known as Nested NER. While traditional NER focuses on identifying individual named entities within a sentence, Nested NER deals with the recognition of nested or hierarchical structures of entities that overlap with one another. This added complexity arises when named entities, such as organizations or locations, encompass other entities within them, such as person names or specific addresses. Nested entities refer to entities that are embedded within other entities. For instance, consider the entity „John Smith" Here entity can include two additional entities inside of it – „John" and „Smith" which can be labeled as first and last names. The goal of Nested NER is to accurately extract these overlapping entities while preserving their hierarchical relationships, allowing for a more nuanced understanding of the information contained within the text.

I will examine the methodologies and approaches employed in this task, focusing on probabilistic graph models, such as conditional random fields (Lafferty et al., 2001) and hidden Markov models (Eddy, 1996), which have proven effective in capturing the contextual dependencies and relationships between entities. These methods are further described in the sections 2 and 3.

2 Conditional Random Fields

Conditional Random Fields (CRF) are probabilistic models used for structured prediction tasks, particularly in the field of natural language processing. They are often employed in tasks such as sequence labeling, named entity recognition, part-of-speech tagging, and speech recognition.

In the context of nested entity recognition, CRF play a significant role in identifying and labeling hierarchical or nested entities within a text. Nested entity recognition involves identifying and labeling both the outer and inner entities correctly. This task is challenging because the boundaries of nested entities can overlap, making it difficult to determine the correct labeling. CRF address this challenge by considering the dependencies and correlations among neighboring words and labels in a sequence.

Proceedings of the Student Research Workshop associated with RANLP-2023, pages 64–67, Varna, Bulgaria, Sep 4–6, 2023.
https://doi.org/10.26615/issn.2603-2821.2023_007

3 Hidden Markov Models

Hidden Markov Models (HMM) are statistical models widely used in various fields, including natural language processing. They are particularly useful in sequence labeling tasks, such as nested entity recognition, where the goal is to identify and classify named entities within a text.

In the context of NER, HMM are often employed for the task of Nested Named Entity Recognition, which involves identifying named entities that are hierarchically structured and nested within each other.

The basic idea behind HMM is to model the underlying structure of a sequence of observations and their corresponding labels. In the case of NER, the observations are the words or tokens in a text, and the labels represent different named entity categories. HMM assume that the underlying labels (states) generating the observations (emissions) form a Markov chain, where the current state depends only on the previous state. One common approach is to use a layered HMM, where each layer corresponds to a level of nesting. The innermost layer represents the most specific entities, and as we move outward, the layers represent progressively broader entities.

During the training phase, the model learns the transition probabilities between different states (labels) based on the training data. It also learns the emission probabilities, which represent the likelihood of observing a particular word given a certain state. These probabilities are estimated using techniques such as the maximum likelihood estimation or the Viterbi algorithm. The Viterbi algorithm is often employed to efficiently compute the most probable label sequence.

4 Related Works

(Shen et al., 2003) leveraged the HMM and integrated various features, including simple deterministic features, morphological features, part-of-speech (POS) features, and semantic trigger features, to recognize flat entities. They presented a simple algorithm to solve the abbreviation problem and a rule-based method to deal with the cascaded phenomena.

(Alex et al., 2007) introduced three models based on CRF which can reduce the nested NER problem into one or more sequence tagging problems. They separately built inside-out and outside-in layered CRF for addressing nested NER, both of which can use the current guesses as to the input to the next layer. They also cascaded separate CRF of each entity category by using output from the previous CRF as features of the subsequent CRF, yielding the best performance in their work.

(Ju et al., 2018) proposed a novel neural model to identify nested entities by dynamically stacking flat NER layers. Each flat NER layer is based on a state-of-the-art (SoTA) flat NER model that captures sequential context representation – BiLSTM, that feeds the output further to the cascaded CRF layer.

(Shibuya and Hovy, 2020) proposed a method where each named entity type output from BiLSTM is being handled by multiple CRF independently. As a result, contributed to handling situations where the same mention span in assigned multiple entity types. Their method allowed them to recognize not only outermost named entities but also inner nested ones. Used decoding method iteratively recognizes entities from outermost ones to inner ones in an outside-to-inside way.

For Polish, there is a system for the NER task called PolDeepNer2 (Marcinczuk and Radom, 2021). It is based on a pre-trained language model of the Transformer type. It has the ability to detect nested NER by extending the set of possible label classes to include classes representing overlapping annotation types. The solution was trained and tested on a dataset available as part of the PolEval 2018 (Wawer and Malek, 2018) competition. A noticeable problem of such a solution is that with numerous label classes (and this is the collection we are dealing with in this work), the number of class combinations that can overlap grows very quickly. For example, in the case of the set we are analyzing, assuming that we are only analyzing one degree nesting we will get 13,285 classes, and assuming that we are analyzing possible double nesting it will already be 1,062,721.

5 Experiments

The purpose of the experiments is to identify the best model in terms of prediction accuracy for the Polish language, which is challenging due to its rich inflectional system, compound words, ambiguity, and context sensitivity. The methods will be tested on a larger set for the nested NER task, which is several times bigger than the current best-known corpora in terms of class size.

The solution presented in (Shibuya and Hovy, 2020) scored the highest F1-score in nested NER task benchmark corpora such as ACE2005 (Walker and Consortium, 2005) and GENIA (Kim et al., 2003). Therefore, it will be adopted as the first to the Polish corpus.

5.1 Corpus

Well known corpus in the domain of the Polish language is The National Corpus of Polish (Tomaszczyk et al., 2012) which was not used during the experiments due to the insufficient number of classes (14) that makes it impossible to test methods on a multi-class, fine-grained collection.

The experiment's dataset, named KPWr (Broda et al., 2012), has been divided into three sets: train, dev, and test, as displayed in table 1. This particular dataset comprises only Polish text and has been sourced from platforms such as *Wikipedia*, *Wikinews*, and information portals under a Creative Commons license. Contrary to ACE2005 (7) or GENIA (36), KPWr includes as many as 120 fine-grained classes, for example, first and last names, cities, countries, districts, postal codes, and many others. It is essential to note that the dataset's class frequency is imbalanced, making it even more challenging.

5.2 BiLSTM-CRF

The algorithm discussed in (Shibuya and Hovy, 2020) underwent testing using two embedding sources: Word2Vec (Piasecki et al., 2017) and HerBERT (Mroczkowski et al., 2021). The vector lengths of the HerBERT and Word2Vec models were 768 and 100 respectively. Moreover, the HerBERT model considers context, while Word2Vec does not.

In figure 1, you can see the curves of F1 values that have been tracked throughout the training epochs. Surprisingly, a smaller and non-contextual embedding model does perform better in this comparison with a value of F1 at *67.94%* on the test set and recall, precision at values of 77.91%, 60.23% respectively. HerBERT, on the other hand, performed as follows: F1 - *61.11%*, recall - 48.56%, and precision - 82.42%. Nonetheless, results should be repeated a few times and confirmed with statistical tests.

6 Future Work

Future work will involve training and evaluation of other methodologies for nested NER from the literature. The analysis of the experimental results will focus on the prediction accuracy of nested entities given the degree of nesting. Core benchmark corpora do not include as many classes as KPWr, therefore, analyzing the algorithms used on the NNE (Ringland et al., 2019) set, which is closer to the KPWr set given the number of classes (112) and nesting entities, may prove valuable results.

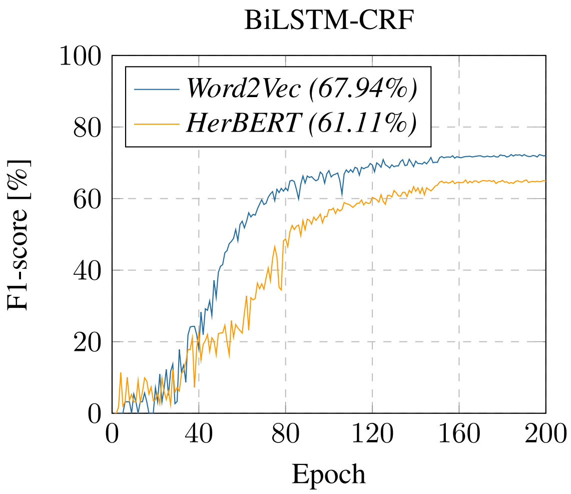

Figure 1: Comparison of *embedding methods (best on test set)* – F1-score over epochs during the training of BiLSTM-CRF

Due to the number of classes in the set for the Polish language, it would also be necessary to take into account the computational complexity of the solutions, which affects the processing time of the data, which should be taken into account in the case of mass text processing services.

The interesting direction will be the validation of prediction accuracy on samples strongly dependent on the given context. To achieve this, it would be necessary to collect such partly by rules based on the number of assigned classes for a particular token in the collection but it would also be worthwhile to select such samples manually by a linguist.

7 Conclusions

In this paper, I have focused on exploring the task of nested named entity recognition (NER) for the Polish language and investigated the use of probabilistic graph models, specifically conditional random fields (CRF) and hidden Markov models (HMM), to address this challenging problem.

I reviewed a few related works that have employed CRF and HMM for nested NER. These studies proposed various models and techniques, such as incorporating semantic features, cascading CRF layers, and leveraging neural models, to improve nested NER performance.

	Train	(%)	**Dev**	(%)	**Test**	(%)
# documents	1,424	(87)	100	(6)	113	(7)
# sentences	24,815	(86)	2,001	(7)	2,000	(7)
# tokens	392,351	(87)	27,318	(6)	30,316	(7)
# entities	28,882	(86)	2,498	(7)	2,219	(7)
- nested	8,049	(28)	772	(31)	526	(24)

Table 1: Statistics of the dataset used in the experiments – *KPWr*

To evaluate the effectiveness of these approaches for the Polish language, I conducted experiments using the KPWr dataset focusing on the BiLSTM-CRF model and comparing the performance of Word2Vec and HerBERT embeddings.

Overall, leveraging probabilistic graph models, such as CRF and HMM, shows promise for addressing nested NER in the Polish language. Further research and experimentation are needed to develop robust and accurate models for this complex task, which has important implications for various NLP applications.

References

Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Biological, translational, and clinical language processing*, pages 65–72.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a free corpus of Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3218–3222, Istanbul, Turkey. European Language Resources Association (ELRA).

Sean R Eddy. 1996. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl$_1$) : $i180 - -i182$.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Michal Marcinczuk and Jarema Radom. 2021. A single-run recognition of nested named entities with transformers. In *International Conference on Knowledge-Based Intelligent Information & Engineering Systems*.

Robert Mroczkowski, Piotr Rybak, Alina Wr 'oblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Maciej Piasecki, Arkadiusz Janz, Dominik Kaszewski, and Gabriela Czachor. 2017. Word embeddings for polish. CLARIN-PL digital repository.

Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. Nne: A dataset for nested named entity recognition in english newswire. *arXiv preprint arXiv:1906.01359*.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Effective adaptation of hidden markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*, pages 49–56.

Takashi Shibuya and Eduard Hovy. 2020. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620.

Barbara Lewandowska Tomaszczyk, Mirosław Bańko, Rafał Górski, Piotr Pęzik, and Adam Przepiórkowski. 2012. Narodowy korpus języka polskiego.

C. Walker and Linguistic Data Consortium. 2005. *ACE 2005 Multilingual Training Corpus*. LDC corpora. Linguistic Data Consortium.

Aleksander Wawer and E Malek. 2018. Results of the poleval 2018 shared task 2: Named entity recognition. In *Proceedings of the PolEval 2018 Workshop*, pages 53–62.