Part-Of-Speech Tagging and Named Entity Recognition with Probabilistic Graphical Models

Clemens Biehl (clemens.biehl@stud.tu-darmstadt.de)[1], Daniel Wehner (daniel.wehner@stud.tu-darmstadt.de)[1], and Fabian Otto (fabian.otto@stud.tu-darmstadt.de)[1]

[1]Technische Universität Darmstadt

***Abstract*—Natural language processing is an increasingly important task, where Part-of-Speech and Named Entity Recognition are important steps for complex semantic procssing or similar tasks. Therefore, this project evaluates different probabilistic graphical models – Naïve Bayes, Hidden Markov Models and Conditional Random Fields – for those two tasks and additionally compares the performance of different handcrafted features as well as GloVe word embeddings. Another objective is to provide a better explanation for the assignments of different POS tags and named entities by a graphical model, which might improve the transparency and the comprehensibility of the model and the ability to create more complex language models.**

I. Introduction

Natural Language Processing (NLP) is the application of computational techniques for the automatic analysis and representation of human language. This field becomes increasingly interesting as the number of available resources on the web is increasing and millions of websites and documents can be accessed and processed. Making use of this large amount of unstructured data to enable and improve computational natural language understanding is therefore a task worth investigating [1].

The tasks of part-of-speech tagging as well as named entity recognition are non-trivial, since natural language expressions can be ambiguous in multiple ways. For instance a word can have multiple parts of speech and a term like *Florence* might refer to either a named entity *PERSON* or *LOCATION* [2, cf. pages 167-169].

The goal of this project is to compare the different approaches of manual feature engineering and word embeddings as well as different probabilistic graphical models on the tasks of part-of-speech tagging and named entity recognition. The Groningen Meaning Bank of University of Groningen is used as a training corpus [3].

II. Literature Review

A. Natural Language Processing Pipeline

Solving a natural language processing (NLP) problem necessitates a pipeline – a sequence of tasks which have to be dealt with in order. Part-of-Speech (POS) tagging is one key part in almost every NLP pipeline. It can e. g. be used to facilitate the identification of named entities or constituents and dependencies. Figure 1 depicts such an NLP pipeline.

As can be seen in Figure 1 POS tagging is applied at an early stage of the pipeline which emphasizes its very importance. Almost all tasks in NLP rely on POS tagging making its performance crucial for the performance of the entire pipeline.

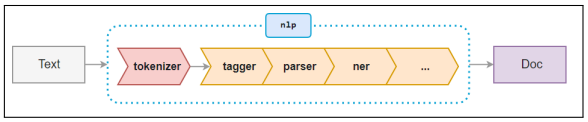

Figure 1. Natural Language Processing consists of several steps. These steps can be arranged in an NLP pipeline.

B. Part-Of-Speech Tagging

POS tagging refers to the assignment of the part of speech, such as noun, plural-noun, verb, etc. to each token in a text. There are several rule-based and stochastic algorithms to perform this disambiguation of word categories. While most words in large text corpora are nouns, accurate part-of-speech tagging still is a challenging task, because words can be ambiguous, i. e. can have different parts of speech. A way to distinguish between the different possible parts of speech for a particular word occurrence is to incorporate the context of this word, i. e. the surrounding tokens, as well as morphological features such as affixes [2, cf. pages 167-170].

State-of-the-art POS taggers reach a word-level accuracy of 97% to 98%, whereas the sentence-level performance of the Stanford PCFG tagger on the Penn Tree Bank (version 3) is only around 32% [4] [5] [6].

C. Named Entity Recognition

Named Entity Recognition (NER) is the task of identifying spans of text that constitute proper names and classifying these as a particular type of entity such as *PERSON*, *LOCATION* or *DURATION*. A model which seeks to solve this task needs to deal with ambiguities arising from the fact that some words (e. g. *Florence*) could refer to multiple entities (e. g. *PERSON*, *LOCATION*). Again, the context of a named entity can be a useful feature for NER and help to disambiguate between multiple entity types in such cases [2, cf. pages 761-765].

Deep learning based methods for NER reach an F1 measure of about 0.91 [6].

D. Naïve Bayes

Naïve Bayes serves as a baseline for the performance in both POS tagging and NER. The implementation of the *scikit-learn* framework will be used [7].

E. Hidden Markov Models

A Hidden Markov Model (HMM) is a directed probabilistic graphical model which models a system of randomly changing hidden states and assumes the Markov property, i. e. that the probability of a particular state depends only on the previous state [2, cf. pages 211-212]:

$$p(q_i|q_1, \ldots, q_{i-1}) = p(q_i|q_{i-1})$$

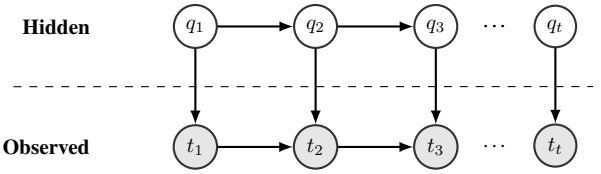

Figure 2. An example of a Hidden Markov Model

In [8] the authors use Brant's TnT tagger [9] for POS tagging which uses Hidden Markov Models (HMM). Based on the error the tagger makes they additionally learn transformation rules which should correct the errors. The authors report an F1-Score of 79.66 % (without error correction) and 80.74 % with error correction.

F. Conditional Random Fields

Conditional Random Fields (CRF) are closely related to HMMs. They are undirected probabilistic graphical models [10] which take into account the labels of neighbouring samples [11]. It is a discriminative approach which models the conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$ and refrains from modeling the marginal $p(\boldsymbol{x})$. [11]

The authors of [12] make use of a conditional random field ('CRF++, Yet another CRF package') in order to perform POS tagging for the language Hindi. The authors of the paper trained the model on 21.000 words and report an accuracy of 82.67 %.

Named entity recognition represents another NLP problem which can be addressed by CRFs [13]. The authors report results on the CoNLL-2003 named entity recognition shared task consisting of tagged news articles. They achieved an F1-score of 84.04 % (for the English language) and 68.11 % (for the German language). The classes were as follows: PERSON, *LOCATION*, *ORGANIZATION* and *MISC*.

III. Methodology and Objective

A. General Objective

This project seeks to accurately identify POS and NE tags by using different probabilistic graphical models. Further, this project utilizes two distinct feature sets – manually engineered features (1) and GloVe word embeddings (2) [14] – in combination with a Naïve Bayes and a CRF. Additionally a *Hidden Markov Model* (HMM) is evaluated given the token sequence as input.

For computing the Naïve Bayes model, the *scikit-learn* framework[1] will be used, the python-crfsuite[2] is providing fast computation for CRF models and the Hidden Markov model is based on the implementation from nltk[3] as it already provides an interface for processing natural language.

All approaches are evaluated with Precision, Recall and F1 measure as harmonic mean between Precision and Recall and compared to up-to-date benchmarks.

B. Research questions

- Q1 How do CRF and HMM compare to the baseline of Naïve Bayes?
- Q2 Which handcrafted features are suited best for the different models in order to reach a high accuracy for POS tagging and NER?
- Q3 How do handcrafted features compare to word embeddings?
- Q4 Can graphical models with well-engineered features beat deep learning benchmarks?

IV. Planned Execution

05.11 - 26.11.:
Data preparation and first results for Naïve Bayes.

26.11 - 23.12.:
Evaluation of different feature sets and comparison with word embeddings as well as graphical models for POS problem.

07.01 - 28.01.:
Evaluation of different feature sets and comparison with word embeddings as well as graphical models for NER problem.

28.01 - 28.02.:
Smaller follow up experiments, preparing presentation and visualizations.

References

[1] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research [review article]," *IEEE Computational Intelligence Magazine*, vol. 9, pp. 48–57, May 2014.

[2] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1st ed., 2000.

[3] J. Bos, V. Basile, K. Evang, N. Venhuizen, and J. Bjerva, "The groningen meaning bank," in *Handbook of Linguistic Annotation* (N. Ide and J. Pustejovsky, eds.), vol. 2, pp. 463–496, Springer, 2017.

[4] V. Jatav, R. Teja, S. Bharadwaj, and V. Srinivasan, "Improving part-of-speech tagging for NLP pipelines," *CoRR*, vol. abs/1708.00241, 2017.

[5] C. D. Manning, "Part-of-speech tagging from 97linguistics?," in *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'11, (Berlin, Heidelberg), pp. 171–189, Springer-Verlag, 2011.

[1] http://scikit-learn.org/stable/modules/naive_bayes.html
[2] https://python-crfsuite.readthedocs.io/en/latest/
[3] https://www.nltk.org/_modules/nltk/tag/hmm.html

[6] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *CoRR*, vol. abs/1708.02709, 2017.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] P. Awasthi, D. Rao, and B. Ravindran, "Part of speech tagging and chunking with hmm and crf," 01 2006.

[9] T. Brants, "Tnt: A statistical part-of-speech tagger," in *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, (Stroudsburg, PA, USA), pp. 224–231, Association for Computational Linguistics, 2000.

[10] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, pp. 267–373, Apr. 2012.

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

[12] A. Pvs and G. Karthik, "Part-of-speech tagging and chunking using conditional random fields and transformation based learning," 2006.

[13] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, (Stroudsburg, PA, USA), pp. 188–191, Association for Computational Linguistics, 2003.

[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.