Eigennamen- und Zitaterkennung auf dem deutschsprachigen Rechtstext-Korpus LER[1] mit Transformermodellen in FLAIR[2]

Stefan Krüger

19. Januar 2023

1 Motivation

Die named-entity recognition (NER) steht im deutschsprachigen Domänenraum der Rechtstexte mehreren nicht trivialen Aufgaben gegenüber. Zunächst gibt es keine einheitliche Norm über die Gestaltung möglicher named entities (NE). Des weiteren werden einige der typischen NE's vor der Veröffentlichung teilanonymisiert[3]. Im Bereich des Natural language processing (NLP) gibt es state-of-the-art Ansätze (fine-tuned transformer-based models)[4], die o.g. Probleme besser bewältigen könnten als herkömmliche Methoden. Domänenunabhängig entstehen aktuell viele Tools, API's und Frameworks, die diese neuen Ansätze beinhalten. Eines dieser Frameworks, FLAIR, wird in dieser Arbeit benutzt.

2 Ziel der Arbeit

Wie gut ist die Performanz vortrainierter und auf dem LER Datensatz fine-tuned Transformermodelle?

Dazu werden drei verschiedene vortrainierte transformerbasierte Modelle auf dem Datensatz fine-tuned und miteinander verglichen. Abschließend wird in einer Datenanalyse geklärt:

Welche Muster lassen sich zwischen den Modellen in den Ergebnisdaten finden?

3 Related Work

In der NER findet man state-of-the-art Ansätze und Datensätze, die aufzeigen, dass dieser Forschungszweig in diversen Domänen viel Aufmerksamkeit erfahren hat. Der Öffentlichkeit zugänglich existieren für manche Domänen Benchmark-Datensätze, an denen neue Forschungsansätze evaluiert und präsentiert werden. So wurde 2018 auf dem deutschsprachigen Datensatz Conll03 von Akbik et al.[5] mit contextual string embeddings ein neuer state-of-the-art F1-Score von 0.8832 erzielt, später von Schweter et al.[4] mit einem fine-tuned Transformermodell überholt mit einem F1-Score von 0.8834.

Untersuchungen dieser Art gab es lange nicht für die deutschsprachige Rechtsdomäne (vgl. [3]). Leitner et al. entwickeln 2019 einen hierfür entsprechenden annotierten Datensatz[1] mit 6-grobkörnigen und 19-feinkörnigen NE's. Dabei werden verschiedene BiLSTM-CRF-Modelle verglichen, wobei das character embedding-basierte Modell die beste Performanz erzielt (F1-Score von 0.9595 auf den grobkörnigen und 0.9546 auf den feinkörnigen NE's)[3].
Dieser Datensatz und FLERT[4] sind 2022 Grundlage für die Arbeit von R. Erd[6]. Er evaluiert auf diesem Datensatz einen fine-tuned Transformer- sowie ein BiLSTM-CRF-Modell mit FLAIR und kommt zu dem Schluss, dass Data Augmentation die Leistung von sowohl rekurrenten als auch von transformerbasierten NER-Modellen in ressourcenarmen Umgebungen effektiv verbessern kann.

4 Vorgehen

Einleitend werden Probleme herkömmlicher NER-Methoden in der deutschen Rechtstext-Domäne angeschnitten. Dem gegenüberstehend wird der Ansatz des kontextuellen Inhaltes auf Dokumentebene bei der Bestimmung von NE's von FLERT erläutert und in dieser Arbeit angewendet.[4]

Der Datensatz ist in einer einzigen Datei[7] verfügbar. Die einzelnen Dokumente liegen kontexterhaltend sequenziell annotiert vor. Die für das spätere fine tunen mit FLAIR erfolderlichen Daten train, dev und test werden hieraus erzeugt.

Mit dem in der FLAIR-Dokumentation angegebenen Link[8] der von FLAIR unterstützen TransformerWordEmbeddings[9] eignen sich drei der folgenden Modelle eher als andere:

- bert-large-cased

- bert-base-german-cased
- bert-base-german-dbmdz-cased
- RoBERTa-large
- XLM-RoBERTa
- distilbert-base-german-cased

Hiervon werden ein multilinguales Modell sowie zwei auf deutschsprachigen Texten vortrainierte Modelle ausgewählt, z.B. RoBERTa[10], BERT[11], DistilBERT und mit dem NLP-framework FLAIR fine-tuned. Manche dieser Modelle existieren jeweils als cased bzw. uncased und sind auf huggingface[8] frei zugänglich.

Für das fine-tunen wird der Ansatz von Schweter et al.[4] übernommen, wobei die oben genannten Transformermodelle jeweils unter Verwendung eines linear layer für die word level prediction[11] fine-tuned werden. Dazu werden python-Scripte erstellt, die anschließend auf dem im Institut zur Verfügung stehenden GPU-Server ausgeführt werden.

Mit diesem Vorgehen wird auch der Einsatz von FLAIR zum Lösen von NER auf dem Datensatz verdeutlicht.

Vor der Evaluation wird die F1-Metrik im Bereich von NER erläutert. Mit den drei Originalmodellen und den drei fine-tuned Modellen wird NER jeweils auf LER gelöst, und mit der micro-basierten F1-Metrik deren Performanz untereinander evaluiert und das fine-tunen bewertet.

Unmittelbar vor der Datenanalyse wird der Datensatz genauer erläutert: Welche NE's sind vorhanden, wie viele davon sind (teil)anonymisiert und zu welchem Prozentsatz ist dies geschehen. Die Analyse beantwortet die zweite Forschungsfrage. Hierfür wird die Menge der tagged token sowohl graphisch als auch anhand von Beispielsätzen untersucht. Dabei wird z.B. nach folgenden Mustern in den Ergebnisdaten gesucht:

- RoBERTa erkennt nicht anonymisierte Personen besser als BERT, DistilBERT erkennt jedoch die anonymisierten Personen am besten.
- RoBERTa und BERT haben bezogen auf die Locations eine große Schnittmenge an false positive tagged token. Das dritte Modell hat dieses Problem gar nicht und labelt die Locations meist richtig.

Literatur

[1] Elena Leitner, Georg Rehm und Julián Moreno-Schneider. „A Dataset of German Legal Documents for Named Entity Recognition“. In: (29. März 2020). arXiv: 2003.13016v1 [cs.CL].

[2] Alan Akbik u. a. „FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP“. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, Juni 2019, S. 54–59. DOI: 10.18653/v1/N19-4010. URL: https://aclanthology.org/N19-4010.

[3] Elena Leitner. *Eigennamen- und Zitaterkennung in Rechtstexten*. Potsdam, Feb. 2019. URL: https://raw.githubusercontent.com/elenanereiss/Legal-Entity-Recognition/master/docs/Leitner_LER_BA.pdf.

[4] Stefan Schweter und Alan Akbik. „FLERT: Document-Level Features for Named Entity Recognition“. In: (13. Nov. 2020). arXiv: 2011.06993v2 [cs.CL].

[5] Alan Akbik, Duncan Blythe und Roland Vollgraf. „Contextual String Embeddings for Sequence Labeling“. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, S. 1638–1649. URL: https://aclanthology.org/C18-1139.

[6] Robin Erd. *Data augmentation for named eneity recognition in the German legal domain*. Online. Abgerufen am: 2023-01-14. URL: https://www.db-thueringen.de/receive/dbt_mods_00055103.

[7] Leitner. *LER all data in one file*. Abgerufen am: 2023-01-13. URL: https://raw.githubusercontent.com/elenanereiss/Legal-Entity-Recognition/master/data/ler.conll.

[8] huggingface. Online. Abgerufen am: 2023-01-13. URL: https://huggingface.co/transformers/v2.3.0/pretrained_models.html.

[9] Alan Akbik und community. *Transformer Word Embedding documentation*. Online. Abgerufen am: 2023-01-14. URL: https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/TRANSFORMER_EMBEDDINGS.md.

[10] Yinhan Liu u. a. „RoBERTa: A Robustly Optimized BERT Pretraining Approach“. In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[11] Jacob Devlin u. a. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. In: *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/n19-1423.