“Do CoNLL-2003 Named Entity Taggers Still Work Well in 2023?”

This paper investigates the generalization of named entity recognition (NER) models trained on the CoNLL-2003 dataset, which was collected in 1996, to modern data. The authors create a new test set called CoNLL++ using news articles from 2020, matching the format and style of CoNLL-2003. They evaluate over 20 different NER models trained on CoNLL-2003 and compare their performance on both datasets.

Key Findings:
No evidence of widespread overfitting: The performance improvements observed on the original CoNLL-2003 test set over the years are not primarily due to overfitting. Instead, the observed performance drops on CoNLL++ are mainly attributed to temporal drift.
Transformer-based models generalize better: Models like RoBERTa and T5 demonstrate good generalization, even when fine-tuned on decades-old data. BiLSTM models with Flair and ELMo embeddings show more significant performance drops, indicating that Transformer architectures have a significant advantage in terms of generalization.
Model size matters: Larger models generally exhibit better generalization. This suggests that larger models have more capacity to capture temporal changes in language.
Pre-training corpus matters: The time period of the pre-training corpus significantly affects generalization. Training Flair and ELMo embeddings on a more recent corpus improves their performance on CoNLL++. Similarly, continued pre-training of RoBERTa with temporally closer data leads to better generalization.
Amount of fine-tuning data affects generalization: While both RoBERTa and Flair benefit from more fine-tuning data, the effect is more pronounced for Flair.
Implications:

The CoNLL-2003 dataset remains relevant in 2023, but it is crucial to consider generalization capabilities.
Transformer-based models with large parameter counts and pre-trained on recent corpora demonstrate superior generalization in NER.
The temporal drift phenomenon requires further investigation to develop strategies for mitigating its impact on NLP models.

Limitations:
The exact time period of the pre-training corpora for some models was unavailable.
The analysis on test reuse was conducted on a single new train/dev/test split.
The study does not account for domain shifts that may occur due to emerging text types.

Overall:
This study provides valuable insights into the factors affecting NER model generalization. It highlights the importance of temporal alignment between pre-training and test data and underscores the need for ongoing research to address temporal drift in NLP.

“Domain Adaption of Named Entity Recognition to Support Credit Risk Assessment “

This paper explores the use of Named Entity Recognition (NER) for extracting credit risk attributes from financial documents like loan agreements, which are crucial for financial institutions to assess their risk and allocate capital.

The authors face a common challenge in NER: the need for large amounts of labeled data for training. To address this, they propose a domain adaptation approach using a combination of out-of-domain data (CoNLL-2003 English data) and a small amount of in-domain data (a manually annotated dataset of public financial agreements).

Here are the key findings:
Out-of-domain data alone is not sufficient: Directly applying a NER model trained on the CoNLL-2003 data to financial documents yielded poor performance, highlighting the importance of domain-specific information.
In-domain data significantly improves performance: Training on a combination of CoNLL-2003 data and a small amount of in-domain financial data yielded a significant improvement in NER accuracy.
Purely in-domain training is best: A model trained exclusively on the annotated financial agreements achieved the best performance, indicating that domain context is crucial for this task.
The authors conclude that while using a small amount of in-domain data can improve NER performance, purely in-domain training offers the best results. This study emphasizes the significance of domain-specific data in NER tasks and suggests potential strategies for adapting NER models to financial domains.

The paper also contributes to the research community by making their annotated financial agreement dataset publicly available for further research.
Future directions include:
Expanding the entity types to encompass more credit risk-specific attributes like values and dates.
Conducting extrinsic evaluation of the NER model's output within a real-world credit risk assessment scenario.
Identifying additional features relevant to credit risk assessment beyond regulatory requirements.

FiNER-ORD: Financial Named Entity Recognition Open Research Dataset

This paper introduces FiNER-ORD, a new, manually annotated dataset for Financial Named Entity Recognition (NER). FiNER-ORD is a significant contribution because:

1. Addressing Limitations of Existing Datasets:

Current financial NER datasets like CRA are limited by skewed entity distributions and small size.

General NER datasets like CoNLL-2003 lack the specific focus on financial language and entities.

2. High-Quality and Scalable Dataset:

FiNER-ORD is built from a large collection of English financial news articles, ensuring diversity and relevance.

Manual annotation by two annotators with a 96.85% agreement ensures accuracy.

3. Benchmarks and Performance Analysis:

The paper benchmarks various pre-trained language models (PLMs) and large language models (LLMs) on FiNER-ORD.

Results show RoBERTa-base outperforms other models, including FinBERT designed for finance.

Zero-shot LLMs like GPT-4 are less effective than fine-tuned PLMs for this specific task.

4. Transfer Learning Study:

FiNER-ORD outperforms CRA and CoNLL when used for transfer learning, highlighting its superior representation of financial language.

5. Importance of Finance-Specific Datasets:

The paper argues that finance-specific datasets like FiNER-ORD are crucial for improving NLP models in the financial domain.

6. Ethical Considerations:

The paper acknowledges geographic and linguistic biases in the dataset and addresses the environmental impact of model training.

Limitations and Future Directions:
The dataset is limited to English language news articles, and future work could include other languages and document types.
More entity classes like 'product' and relationship types could be added to improve downstream applications.
Overall, this paper is a valuable contribution to the field of financial NLP, offering a high-quality dataset that can advance the development of domain-specific models and research.