Language Transfer in Named Entity Recognition

Rayane Bouafia & Armin Pousti

McGill University

Abstract

This paper presents a comprehensive study of five Named Entity Recognition (NER) models, ranging from traditional statistical methods to modern neural approaches, and evaluates their performance across seven languages: English, French, Chinese, Arabic, Persian, Swahili, and Finnish. We conduct two sets of experiments: (1) baseline performance evaluation on monolingual datasets, and (2) Few-Shot Learning (FSL) to explore the impact of transfer learning from high-resource to low-resource languages. Our findings offer insights into the effectiveness of different models in diverse linguistic contexts.

1 Introduction

This semester we read papers that dealt with NER where there was language transfer which was greatly intriguing to us and motivated us to explore this concept further. This paper investigates the capabilities of five distinct NER models across seven languages, divided into high-resource (English, French, Chinese) and low-resource (Arabic, Persian, Swahili, Finnish) groups. The primary objective is to evaluate the effectiveness of these models in multilingual contexts and to explore the potential of transfer learning to enhance performance in resource-scarce languages. The study is structured into two experiments. First, a baseline evaluation was conducted to assess the performance of each model on individual monolingual datasets. Second, we applied Few-Shot Learning (FSL) by fine-tuning models pretrained on high-resource languages using limited data (5%, 10%, and 20%) from low-resource languages. This aimed to measure the adaptability of transfer learning techniques and the extent to which pretrained multilingual representations can bridge linguistic resource gaps

Our results demonstrate that modern transformer-based models, particularly DistilBERT, excel in both baseline and few-shot scenarios, significantly outperforming traditional methods like HMM and Decision Trees. LSTM-CRF showed moderate improvements with transfer learning. These findings highlight the transformative potential of pretrained multilingual embeddings and the challenges inherent in achieving robust multilingual NER.

which employs a probabilistic framework to model the relationships between words (observations) and their corresponding tags (states) through start, transition, and emission probabilities. The model uses the Viterbi algorithm to decode the most likely sequence of tags for a given input, making it an interpretable method that leverages both prior and contextual dependencies. The goal is to demonstrate how this simple yet effective probabilistic model can serve as a baseline for NER tasks.

We also examine a hybrid model combining a Bidirectional Long Short-Term Memory (BiLSTM) network and a Conditional Random Field (CRF), designed to capture contextual dependencies while ensuring valid label sequences through the CRF layer. This model enhances predictive accuracy for structured prediction tasks, such as NER, where the relationships between neighboring labels are essential. The CRF refines the BiLSTM outputs, enforcing consistency in the predicted label sequence.

Additionally, we explore the use of Brown Clustering, a technique that groups words into clusters based on frequency distributions. By representing tokens at the cluster level rather than individually, we reduce complexity and improve efficiency in NER tasks. Experimental results demonstrate that while smaller cluster sizes tend to achieve better performance, there is a trade-off between cluster granularity and prediction consistency.

To further enhance NER performance, we implement a Decision Tree classifier that uses handcrafted features, such as capitalization and word length, to identify entities. Despite its simplicity, the Decision Tree provides a valuable baseline for multilingual NER tasks, offering an interpretable model that can be adapted for more complex systems.

Finally, we fine-tune a multilingual DistilBERT model for NER, balancing computational efficiency with performance. By freezing most of the model's layers and reducing the sequence length, we optimize resource usage while still achieving moderate performance. This experiment highlights the feasibility of fine-tuning lightweight transformer models for multilingual NER tasks with limited computational resources.

Together, these methods present a comprehensive approach to NER, combining probabilistic models, neural networks, and efficient clustering techniques to address the challenges of sequence labeling tasks in natural language processing.

2 Related Work

In one implementation a sliding-window technique, paired with a pre-trained BERT model for sequence labeling, is used for Named Entity Recognition (NER) in clinical notes, focusing on medication spans [1]. This model divides input into overlapping 512-token subsequences with a stride of 128, applying the BILOU scheme for token classification. Aggregated predictions are made using averaging techniques, and the model integrates span-based and question-answering systems for enhanced accuracy. While our model, based on HMM and LSTM-CRF, offers greater interpretability and computational efficiency, BERT's sliding-window method is better for longer sequences and complex context in clinical data but may struggle with fixed-token length constraints.

In Chinese NER, lexical information is integrated into a BERT model using a multi-task learning framework to reduce noise from external lexicons [2]. A ranking model scores lexicon-matched words, which are then used in character-level sequence labeling with multi-head attention and a CRF layer. Unlike our model, which directly models entity relationships, this approach benefits from external lexicons and multi-task learning, providing additional semantic insights, particularly in domain-specific tasks.

Another NER model uses domain-adversarial training and multi-task learning for automotive domain NER, incorporating bilingual Korean and English datasets to enhance domain generalization [3]. While our model is more flexible and domain-agnostic, this model is particularly effective for domain-specific tasks, leveraging domain-invariant features and addressing word spacing errors.

In conclusion, our model offers simplicity and efficiency as a baseline for NER, while advanced methods like BERT's sliding windows [1], lexical integration [2], and domain-adversarial training [3] excel in domain-specific contexts and leveraging additional semantic resources.

3 Methods

3.1 Dataset

We used the publicly available WikiAnn dataset, which supports multilingual NER in a consistent BIO format, with the following label map: "O": 0,"B-PER": 1,"I-PER": 2,"B-ORG": 3,"I-ORG": 4,"B-LOC": 5,"I-LOC": 6. We selected a varied range of languages. Which we then categorized into high-resource (English, French, Chinese) and low-resource (Arabic, Persian, Swahili, Finnish) groups. The dataset is preprocessed with methods tailored to the specific requirements of each model to ensure optimal performance. For HMM, Tokens are encoded into numerical IDs based on a constructed vocabulary from training data. Unseen tokens are replaced with an <UNK> token, and sequences are padded to a maximum length of 50 tokens with <PAD> tokens. Tag labels are similarly encoded, with padding using the label O. LSTM-CRF uses similar preprocessing. For the Decision Tree Classifier, handcrafted features are extracted for each token, such as token length, capitalization, numeric nature, prefixes, suffixes, and surrounding context tokens. These features are converted into numerical vectors using a feature mapping constructed from the training data. For DistilBERT, tokens are tokenized using the DistilBERTTokenizerFast, which aligns the labels with the tokenized outputs using the word-piece tokenization technique. Labels corresponding to subword tokens or special tokens are set to -100 to ignore them during training. For Brown Clustering, tokens are grouped into clusters based on unigram frequency counts derived from the training dataset, and each cluster is assigned the most frequent tag observed in the training data.

3.2 Models

3.2.1 HMM

A Hidden Markov Model (HMM) is implemented for sequence labeling tasks, specifically Named Entity Recognition (NER), in a structured and probabilistic manner. The implementation models the relationships between words (observations) and their corresponding tags (states) using start, transition, and emission probabilities. During training, the HMM calculates these probabilities from labeled data, and it uses the Viterbi algorithm during inference to decode the most likely sequence of tags for unseen input. This approach is valuable because HMMs provide an interpretable and probabilistic framework for modeling sequences, effectively leveraging both prior and contextual dependencies. The goal is to demonstrate how a simple, yet effective probabilistic framework can achieve reasonable performance for sequence labeling tasks, serving as either a baseline or a complementary method in our project.

3.2.2 LSTM-CRF

This implementation presents a hybrid model combining a BiLSTM and a Conditional Random Field (CRF) for sequence labeling tasks, commonly applied in natural language processing tasks like Named Entity Recognition (NER). The BiLSTM captures contextual dependencies in input sequences by encoding both past and future information into token representations, while the CRF layer models the dependencies between output tags to ensure valid label sequences. This combination is particularly useful for structured prediction problems where the relationships between neighboring labels (e.g., beginning, inside, and outside of entities) are critical. The CRF layer refines the outputs of the BiLSTM, enforcing label sequence consistency through transition scores and facilitating efficient decoding using the Viterbi algorithm. The goal of this implementation is to improve predictive accuracy for tasks requiring both contextual understanding and structured output. The modularity of the framework allows it to be extended with other embedding or contextual representation methods, making it a flexible and effective approach for sequence labeling.

3.2.3 DistilBERT

This implementation focuses on fine-tuning a multilingual Named Entity Recognition (NER) model using a pre-trained DistilBERT model with minimal resource requirements. The approach involves combining datasets from the seven languages, tokenizing the inputs while aligning word-level labels to token-level labels, and fine-tuning only the classifier head of the DistilBERT model, leaving the rest of the model's parameters frozen. The freezing of layers and reducing the sequence length to 30 tokens were deliberate decisions aimed at optimizing computational efficiency. Despite these constraints, training for a single epoch still required nearly two hours, underscoring the high computational cost of multilingual NER tasks. The results show modest performance, with an F1-score of 0.237 and an accuracy of 67.8%. The limited performance can be attributed to the constraints, as freezing most of the model and using a short sequence length reduces the capacity to learn complex language patterns. However, this approach demonstrates the feasibility of fine-tuning lightweight transformer models for multilingual NER tasks with constrained resources and highlights opportunities for improvement through extended training, larger sequence lengths, or selective layer unfreezing.

3.2.4 Decision Tree Classifier

An NER system using a Decision Tree classifier was implemented to identify and classify entities such as names, locations, and organizations in text. The system utilized handcrafted features derived from each token, including attributes like capitalization, length, prefixes, suffixes, and contextual information from surrounding words. These features were numerically encoded to train a supervised machine learning model on a multilingual dataset. The Decision Tree classifier was chosen for its simplicity and interpretability, providing a lightweight solution for NER tasks without the computational demands of more complex models.

The evaluation of the model demonstrated moderate performance, with precision, recall, and F1-score all measured at 0.5324. These results indicate that while the model could reasonably distinguish entities from non-entities, there is room for improvement in capturing more nuanced patterns in the data. The significance of this approach lies in its adaptability to multilingual text and its use as a baseline system. Despite its limitations, the model offers a clear and interpretable starting point for entity recognition tasks, which can be extended or enhanced using more sophisticated approaches, such as neural networks or contextual embeddings, to achieve better accuracy and performance. This system is particularly valuable for projects requiring foundational entity extraction capabilities with straightforward implementation and analysis.

3.2.5 Brown Clustering

This experiment explored the use of a simplified Brown Clustering algorithm to group words based on their frequency distributions, with the goal of improving efficiency in Named Entity Recognition (NER). The clustering technique divides words into predefined clusters, enabling tokens to be represented at the cluster level instead of individually, thereby reducing complexity. Each cluster was assigned the most frequent NER tag observed in the training data, and this cluster-level information was used for predicting tags during evaluation.

The results of the experiment, conducted with varying numbers of clusters, showed that smaller cluster sizes generally achieved higher precision, recall, and F1-scores. For example, with 5 clusters, the system achieved a precision of 0.356, recall of 0.597, and F1-score of 0.446. As the number of clusters increased, the F1-score slightly declined, with 50 clusters yielding a precision of 0.350, recall of 0.591, and F1-score of 0.439. This pattern indicates a trade-off between cluster granularity and prediction consistency: fewer clusters allow for broader generalizations, improving recall, while more clusters potentially capture finer distinctions at the cost of introducing more errors.

This approach is significant because it offers a faster, resource-efficient alternative to token-level tagging by leveraging word clusters, making it suitable for systems with limited computational resources or for applications that prioritize speed over high accuracy. The experiment demonstrates how Brown Clustering can serve as a baseline for NER tasks, providing insights into how word frequency patterns can simplify entity recognition while maintaining reasonable performance levels.

3.3 Experiments

3.3.1 Experiment 1: Establishing Baselines

For each language, the dataset is split into training, validation, and test sets. Each model is trained and evaluated solely on the monolingual dataset. This experiment establishes the standalone capability of each of the 5 models for NER tasks across different languages.

3.3.1 Experiment 2: Few-shot Learning for Transfer

Experiment 2 investigates the effectiveness of transfer learning to improve NER performance for low-resource languages. This approach is particularly relevant because low-resource languages often lack extensive labeled datasets, making direct training insufficient. Models are trained on combined datasets from high-resource languages (English, French, and Chinese). This step aims to leverage the rich annotated data and linguistic structures of these languages to learn generalizable representations. Models pretrained in step 1 are fine-tuned on small subsets (5%, 10%, and 20%) of low-resource language datasets. This step tests whether the pretrained models can adapt to new languages using minimal data. Both LSTM-CRF and HMM are fine-tuned by retraining their parameters on the low-resource datasets. For Decision Tree, the feature mapping is updated using the low-resource dataset to account for domain-specific features. For DistilBERT, The model is fine-tuned with the sampled datasets using its transformer architecture. The relevance of this experiment lies in its potential to address the resource scarcity problem in NER for underrepresented languages. By pretraining on high-resource languages and fine-tuning with limited data, the models can leverage cross-lingual similarities and generalize better for low-resource languages. This setup provides insights into the transferability of NER models and the minimum data requirements for achieving acceptable performance

4 Results

This experiment produced 14 tables of data. We have the experiment 1 results, the experiment 2 pretraining results, and then the experiment 2's few

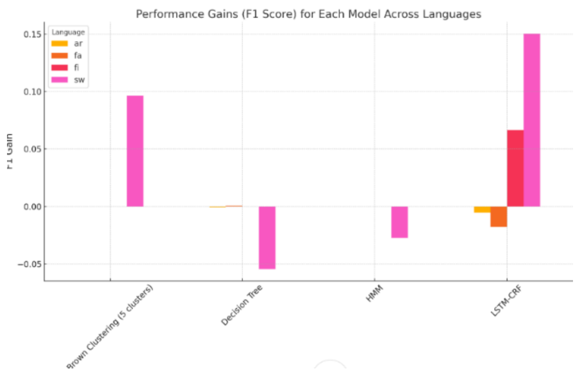

Figure 1: Performance Gains in F1 for each model across languages.

shot percentage results for each language (5%, 10% and 20%, times 4 languages which is 12 tables). We need to analyze results across models, across languages and across Few Shot percentages.

DistilBERT had the best performance across languages, with the closest contender being LSTM-CRF, with up to 15% improvement in F1, making it the most reliable (not online) model for scenarios with a moderate amount of labeled data. HMM and Decision Tree models offered competitive performance, particularly in low-resource settings, but their absolute scores lagged behind deep-learning-based models like LSTM-CRF. Languages with complex morphology, such as Arabic and Persian, benefitted significantly from LSTM-CRF's ability to learn contextual and sequential patterns. Simpler structures like Swahili favored simpler models like Decision Tree, which leveraged feature-based learning effectively.

5 Discussion and Conclusion

In conclusion, DistilBERT and LSTM-CRF are best for accomplishing knowledge transfer across languages. We had many options for experiment 2. We could've done Zero-shot Transfer where we train models only on high-resource languages and evaluate on low-resource ones without fine-tuning. We could've done Sequential Transfer Learning and a few more. We tried implementing bootstrapping and have a 6th model but it was clear that it would require a lot of work to get appropriate seed entities because we got abysmal performance from bootstrapping, ultimately deciding that it would not be a wise choice for a multilingual NER model. This paper is easy to expand upon: we can add more models to test. We can also add experiments, as discussed above, we could've done Zero-shot Transfer and Sequential Transfer Learning and a few more.

6 Statement of Contribution

The workload of this project was evenly divided. Rayane Bouafia was responsible for implementing the LSTM-CRF, HMM, and experiment design for all the models, while Armin Pousti implemented the Brown clustering, decision trees, and DistilBERT models. Both individuals collaborated on writing the report and conducting research on related papers about NER.

References

1. Tian, X., Bu, X. and He, L. 2023 . Multi-task learning with helpful word selection for lexicon-enhanced Chinese NER. In *Applied Intelligence : The International Journal of Research on Intelligent Systems for Real Life Complex Problems, 53(16), pp. 19028–19043*. Available at: https://doi.org/10.1007/s10489-023-04464-0.

2. Tsujimura, T. et al. 2023. Contextualized medication event extraction with striding NER and multi-turn QA. In *Journal of Biomedical Informatics, 144*. Available at: https://doi.org/10.1016/j.jbi.2023.104416.

4. Park, C., Jeong, S. and Kim, J. 2023. ADMit: Improving NER in automotive domain with domain adversarial training and multi-task learning. In *Expert Systems With Applications, 225*. Available at: https://doi.org/10.1016/j.eswa.2023.120007.

5. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

6. WikiANN. *Datasets at Hugging Face.* https://huggingface.co/datasets/wikiann