Fine-Tuning mBERT for Icelandic PoS Tagging and Integrating Multilingual NLP Tools

Valgarð Guðni Oddsson

Abstract

This project explores the fine-tuning of the mBERT multilingual language model for Icelandic Part-of-Speech (PoS) tagging using the MIM-Gold dataset. The fine-tuned model achieves a high accuracy of 97.8%, effectively tagging Icelandic text with detailed grammatical information, including word class, gender, number, case, article, subcategory, mood, and person. To enhance usability, a Vue.js frontend and a FastAPI backend were developed, enabling users to input either Icelandic or English sentences. The system supports multilingual functionality by leveraging AWS Translator, allowing English sentences to be translated into Icelandic for tagging. Additionally, the backend integrates GreynirCorrect to provide grammar and spelling suggestions for Icelandic text and includes a custom-trained classifier for language detection. This comprehensive solution bridges the gap between advanced NLP tools and end-user applications, making Icelandic PoS tagging accessible to a broader audience while demonstrating the potential of fine-tuned multilingual models for low-resource languages.

1 Introduction

Natural Language Processing (NLP) is a field of artificial intelligence that focuses on enabling machines to understand and interpret human language. Part-of-Speech (PoS) tagging is one of the fundamental tasks in NLP, involving the assignment of grammatical tags to words in a sentence. In this report, we present the development and evaluation of a PoS tagging system for both English and Icelandic. The system also incorporates additional components such as spelling and grammar suggestions and multilingual translation capabilities.

The primary goal of this work was to create an NLP pipeline that can accurately predict PoS tags for a given sentence, suggest spelling and grammar improvements, and translate sentences between English and Icelandic. The system is built using a variety of state-of-the-art tools and techniques, including a pre-trained language model, a language classifier, and integration with AWS translation services.

In the following sections, we describe the methodology used to build and fine-tune the system, the results of its performance on a test dataset, the challenges encountered during development, and potential avenues for future work and improvement.

2 Methods

2.1 Dataset and Preprocessing

The dataset used to finetune the pretrained model was a preprocessed version of the Part-of-Speech tagged Icelandic corpus, the Icelandic Gold Standard; MIM-Gold(Loftsson et al., 2018). The preprocessed version contains one sentence per line, with a "/" between tokens and tags. Each tag contains information about the word, including its word class, gender, case, number, person, and a few more outlined by the file `MIM_GOLD_DESCRIPTION_EN_tagset.pdf` found in the project. This was a good dataset as it contained a very large number of sentences, about 58,000, which ensured that the finetuned model had enough data to train and test on.

The dataset was further preprocessed to ensure it could be handled by the mBERT model. The preprocessing involved the following steps:

1. Each line in the dataset was parsed to separate words and their associated PoS tags. The dataset was transformed into a structured format where each sentence was represented as a dictionary containing two lists:
   - `sentence`: a list of words in the sentence.

    - `tags`: a list of corresponding PoS tags.

2. The unique PoS tags in the dataset were extracted to create mappings for the fine-tuning process. This produced the dictionary `tag2id`, which maps each tag to a unique integer ID for training, and the dictionary `id2tag`, which reverses this mapping.

3. Sentences were tokenized using the mBERT tokenizer, which splits text into sub-word tokens. To ensure that PoS tags aligned correctly with the tokenized output, each token was associated with the corresponding PoS tag or marked with a `-100` for tokens such as padding or sub-word fragments that should be ignored during training. This alignment was achieved by comparing word indices generated during tokenization to the original sentence structure.

4. A custom PyTorch class, `PosDataset`, was implemented to manage the training data. This class leveraged the functions used to carry out the previous steps in order to tokenize each sentence and align its tags, and then returned tokenized inputs and the corresponding label sequence.

These preprocessing steps ensured that the data was properly structured and aligned, enabling the mBERT model to handle them properly during training and evaluation.

2.2 Model Fine-Tuning

The mBERT model was fine-tuned on the MIM-Gold dataset for Part-of-Speech (PoS) tagging in Icelandic. Fine-tuning was performed using the Hugging Face Trainer API (Wolf et al., 2020), with the following configuration:

- **Evaluation strategy**: Per epoch
- **Learning rate**: 3×10^{-5}
- **Batch size**: 8 for training, 2 for evaluation
- **Gradient accumulation steps**: 2
- **Number of epochs**: 3
- **Weight decay**: 0.01
- **Mixed-precision training (fp16)**: Enabled
- **Gradient Checkpointing**: Enabled

Fine-tuning was conducted on a system with the following specifications:

- **CPU**: Intel i9-12900K
- **GPU**: NVIDIA RTX 3090
- **RAM**: 64 GiB

The model's performance was evaluated using the following metrics:

- **Accuracy**: The proportion of correctly predicted tags.
- **Macro Precision**: The average precision across all tag classes.
- **Macro Recall**: The average recall across all tag classes.

Fine-tuning encountered challenges related to CUDA memory limitations. Initially, the batch size for training and evaluation was set to 16. However, during metric computation after each training epoch, the GPU ran out of memory, which was believed to be caused by too large a portion of the dataset being loaded onto the GPU. To test this hypothesis, fine-tuning was done using only 10% of the dataset, which resulted in no CUDA memory issues. However, the model trained on only 10% of the dataset did not perform as desired, so a workaround had to be found.

To address the issue:

1. **Reducing the batch size**: Training and evaluation batch sizes were reduced, but this alone did not resolve the memory overflow.

2. **Modifying `compute_metrics`**: The primary cause was identified as the `compute_metrics` function in the Trainer class, which consumed excessive GPU memory during evaluation. After removing this function, the full dataset could be used for training without memory issues.

With the resolution of the CUDA memory issue, the training batch size was restored to 8 to maintain stability while optimizing throughput. The fine-tuning process could then be successfully completed, resulting in a model with a high degree of accuracy and robust tagging capabilities.

2.3 Additional Components

In addition to the fine-tuned mBERT model for Icelandic PoS tagging, the system includes three key components to enhance functionality:

- **Language Classifier**: A custom language classification module was developed to determine whether an input sentence was in Icelandic or English. The model was trained using a pipeline with the Universal Declaration of Human Rights (UDHR) dataset. The training process involved:

  1. Generating synthetic sentences by randomly selecting between 3 to 18 words from the UDHR word list for each language, creating a diverse dataset with 5,000 sentences per language.
  2. Splitting the dataset into training and testing subsets, with 90% of the data used for training and 10% for testing.
  3. Fitting the pipeline to the training data using scikit-learn's `train_test_split` and shuffle methods to ensure a robust and randomized training process.

  The trained classifier enabled automatic detection of the input language, allowing the system to determine whether translation was required before PoS tagging.

- **Grammar and Spelling Suggestions**: The system integrates the GreynirCorrect library (version 3.4.7) (Team, 2023) for generating grammar and spelling suggestions. This component leverages the `check_single` method to analyze Icelandic sentences and provide corrections. To address compatibility issues, the `islenska` module was downgraded to version 1.0, as recommended by the GreynirCorrect project maintainers. This ensured stable functionality despite outdated documentation and bugs in newer versions of the GreynirCorrect library.

- **Multilingual Translation**: To support English-speaking users, the system incorporates AWS's translation service through the boto3 Python package (Services, Year). The `translate_text` method from the AWS translate client was used to perform real-time translation of English input sentences into Icelandic, as well as translation of each word in Icelandic sentences to English. This integration extended the system's utility by allowing users to submit English sentences, which were automatically translated before processing for PoS tagging and grammar suggestions.

These additional components collectively enhanced the usability and versatility of the system, allowing it to handle multilingual input, provide grammatical feedback, and cater to both Icelandic and English speakers.

2.4 System Integration

The system was implemented with a FastAPI backend and a Vue.js frontend, enabling simple and user-friendly interaction with the PoS tagging model and supporting features.

- **Backend**: The backend is built using FastAPI and provides a single RESTful endpoint that processes both Icelandic and English sentences. Upon receiving a sentence, the backend performs the following tasks:

  1. **Language Prediction**: A custom function uses the trained language classifier to determine whether the input sentence is in Icelandic or English.
  2. **Translation**: If the input is identified as English, another function utilizes AWS's `translate_text` method to translate the sentence into Icelandic. After the English sentence has been translated (or if the original sentence was Icelandic), each word in the Icelandic sentence is also translated to provide per-word translations to accompany the PoS information.
  3. **PoS Tagging**: A dedicated function generates PoS tag information for the Icelandic sentence using the fine-tuned mBERT model. The output includes grammatical attributes such as word class, gender, number, case, and more.
  4. **Grammar and Spelling Suggestions**: The GreynirCorrect library is employed to provide suggestions for improving grammar and spelling in the Icelandic text.

  The backend returns a structured JSON object containing:

  - The PoS-tagged and per-word translated Icelandic sentence.
  - Grammar and spelling suggestions.
  - The predicted language of the input.

  This modular architecture ensures flexibility and maintainability while accommodating all system features.

- **Frontend**: The frontend was developed using Vue 3, Vite, and Pinia for state management, with Bootstrap for styling. It serves as a user-friendly interface that facilitates interaction with the backend. Key features include:
  - **Input Field**: A text field allows users to input either Icelandic or English sentences.
  - **Integration with Backend**: Upon submission, the frontend sends the input sentence to the FastAPI backend via an HTTP request.
  - **Results Display**: The response is parsed and displayed, showing:
    - The PoS-tagged Icelandic sentence with detailed grammatical attributes and English translation.
    - Grammar and spelling suggestions for the Icelandic text.
    - The predicted language of the input.

  The frontend simplifies user interaction, making it intuitive to explore the functionality of the PoS tagger and the additional features.

This modular integration of the backend and frontend ensures the system is both flexible and user-friendly, providing comprehensive PoS tagging and additional language-related functionalities.

3 Results

3.1 Model Performance

The fine-tuned mBERT model achieved excellent performance on the PoS tagging task for Icelandic text. Training was conducted on the full dataset, comprising 58,412 sentences, while testing was performed on a held-out 10% subset (5,842 sentences). The model achieved the following metrics during evaluation:

- **Accuracy**: 0.978
- **Macro Precision**: 0.936
- **Macro Recall**: 0.923

These metrics highlight the model's ability to accurately tag Icelandic sentences across a diverse set of grammatical features, including word class, gender, number, case, mood, and others. While precision and recall are slightly lower than accuracy, they demonstrate the model's capability to generalize effectively across all tag categories, even those with fewer examples.

3.2 System Functionality

The system was tested end-to-end, with a focus on validating the integration of its various components. The key results are as follows:

1. **Language Detection and Translation**:
   - The language classifier reliably identified Icelandic and English sentences, achieving an accuracy of 96.8%.
   - English inputs were successfully translated into Icelandic using the AWS translator; however, only sentence-level translations were generally accurate. During per-word translation from Icelandic to English, the translation often returned incorrect interpretations due to ambiguity and lack of context, or sometimes no translation at all. This component was the least performing out of the system's features and has the highest potential for improvement.

2. **Grammar and Spelling Suggestions**:
   - The GreynirCorrect library provided somewhat accurate and context-aware grammar and spelling suggestions for Icelandic sentences. Spelling suggestions were particularly useful, often catching incorrect spellings and providing correct alternatives. However, incorrect grammar was often missed and not corrected, leading to a similar level of dissatisfaction as with the per-word translation component.

3. **PoS Tagging Output**:
   - As mentioned previously, the fine-tuned model performed very well and consistently produced detailed annotations for Icelandic sentences, tagging each word with the correct attributes.

4. **Example Output**: To illustrate the system functionality, consider the following example:
   - **Input (English)**: "Basketball is one of my favorite sports and I am an active player."
   - **Translated to Icelandic**: "Körfubolti er ein af mínum uppáhalds íþróttum og ég er virkur leikmaður."
   - **Tagged Output**:
     - Körfubolti (noun, masculine, singular, nominative)
     - er (verb, indicative, active, 3rd person, singular, present)
     - ein (numeral, genitive)
     - af (adverb, governs case)
     - mínum (pronoun, possessive pronoun, feminine, plural, dative)
     - uppáhalds (noun, neutral, singular, genitive)
     - íþróttum (noun, feminine, plural, dative)
     - og (conjunction)
     - ég (pronoun, personal pronoun, 1st person, singular, nominative)
     - er (verb, indicative, active, 1st person, singular, present)
     - virkur (adjective, masculine, singular, nominative, strong declension, positive)
     - leikmaður (noun, masculine, singular, nominative)

These results demonstrate the system's overall effectiveness, with the model performing excellently on PoS tagging and the integrated components working well, though there are areas (such as translation and grammar correction) that could be further improved.

4 Discussion

4.1 Model Performance Comparison

The fine-tuned mBERT model for Icelandic PoS tagging achieved an accuracy of 97.8%, comparable to results reported for the same task. For instance, a GitHub repository (Jónsson, 2019) also claims 97.8% accuracy for a large PoS tagging model on the MIM-Gold dataset. This similarity in performance demonstrates that the current approach, despite leveraging a general-purpose model like mBERT, is capable of achieving competitive results in specialized language tasks.

Furthermore, the high accuracy, coupled with the macro precision and macro recall of 93.6% and 92.3% respectively, reflects the model's ability to generalize well across diverse linguistic constructs. This suggests that the dataset used for training and evaluation provided sufficient representation of Icelandic grammatical features, enabling robust tagging across categories.

4.2 System Functionality

While the PoS tagging component performed exceptionally well, other system functionalities revealed notable limitations:

1. **Grammar and Spelling Suggestions**: The GreynirCorrect module, while effective at identifying some simple spelling errors, struggled with grammar-related issues. Many grammatical mistakes in test cases were left unaddressed, limiting the tool's utility for users seeking comprehensive language feedback. This lack of performance highlights the need for integrating or developing a more robust grammar correction component for future iterations.

2. **Translation**: The AWS translation service delivered good results for whole-sentence translations, reliably converting English input to Icelandic in most cases. However, word-by-word translation often failed due to the inherent ambiguity of isolated words and the lack of context. Common issues included incorrect or nonsensical translations, as well as instances where no translation was provided at all. This limitation significantly impacts the user experience, especially for English-speaking users relying on accurate word meanings for language learning.

4.3 Key Limitations

The primary limitations of the system are as follows:

- **Grammar Suggestion**: The system's inability to consistently identify grammar errors limits its value as a comprehensive language tool, particularly for language learners who might prioritize grammatical accuracy.

- **Word-by-Word Translation**: The failure to account for context during word-by-word translation renders this feature unreliable and potentially confusing for users. This is a critical shortcoming, as accurate translations for individual words are essential for aiding English-speaking users in understanding Icelandic grammar and vocabulary.

4.4 Future Directions

Addressing these limitations offers several directions for future improvements:

- **Enhanced Grammar Correction**: Integrating a more sophisticated grammar correction model, possibly leveraging fine-tuned transformers, could dramatically improve this aspect of the system.

- **Context-Aware Translation**: Developing a translation model that incorporates sentence-level context when translating individual words could significantly enhance this component and provide much better usability for all users.

4.5 Strengths and Contributions

Despite its limitations, the system represents a step forward in Icelandic language processing. The integration of accurate PoS tagging, grammar correction, and translation functionalities into a unified platform is a valuable contribution to the field, and provides a strong foundation for iterative enhancements and future development.

5 Conclusion

In this report, we presented a comprehensive evaluation of a system designed to perform Part-of-Speech (PoS) tagging, grammar and spelling suggestions, and translation for Icelandic text. The system leverages a fine-tuned mBERT model for PoS tagging, achieving an accuracy of 97.8%, comparable to similar results on the MIM-Gold dataset. However, while the PoS tagging componet demonstrated high performance, the grammar and spelling suggestions, as well as word-by-word translation revealed some significant limitations. The grammar correction was under performing, often missing grammatical issues, while word-by-word translation failed due to ambiguity and lack of context.

Despite these challenges, the overall system shows promise, especially in the context of PoS tagging and full-sentence translation. The work lays a solid foundation, and future developments could focus on integrating more advanced grammar correction models and improving context-aware word translation to make the system more useful to English-speaking users and language learners.

References

Haukur Páll Jónsson. 2019. Pos. Accessed: 2024-11-10.

Hrafn Loftsson, Eiríkur Rögnvaldsson, Sigrún Helgadóttir, Jökull H. Yngvason, Kristján Friðbjörn Sigurðsson, Steinunn Valbjörnsdóttir, Brynhildur Stefánsdóttir, Jón Friðrik Daðason, and Starkaður Barkarson. 2018. Mim-gold: A dataset for icelandic part-of-speech tagging. Accessed: 2024-11-10.

Amazon Web Services. Year. Amazon translate. Accessed: Date.

Greynir Team. 2023. Greynircorrect: Spelling and grammar correction for icelandic. Accessed: 2024-11-10.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, et al. 2020. Transformers: State-of-the-art natural language processing. Accessed: YYYY-MM-DD.