Mountain NER: Potential Improvements

Current Results

Model Performance:

- **F1-score**: 0.9474
- **Precision**: 0.9184
- **Recall**: 0.9783
- **Training**: 3 epochs, BERT-base-cased

Dataset Statistics:

- Total sequences: 1,584
- Sequences with mountains: 226 (14.3%)
- Mountain entities: 232
- Mountain token ratio: 1.8%

Current approach uses BERT-base with custom BIO tagging and achieves strong performance on mountain recognition, but has limitations with complex mountain names and geographic contexts.

Proposed Improvements

1. Advanced Model Architecture

Replace BERT-base with RoBERTa-large or DeBERTa

- Benefits:
  - 3-5% F1 improvement on complex entities
  - Better handling of geographic context
  - Improved generalization to unseen mountain names

2. Multi-Entity Expansion

Add support for additional geographic entities

- New entity types: RIVER, CITY, COUNTRY, LAKE
- Benefits:
  - More comprehensive geographic understanding
  - Better contextual awareness for mountain detection
  - Enables relation extraction (mountains in countries)

3. Data Augmentation

Synthetic data generation and annotation expansion

- Techniques:
  - Back-translation (EN>FR>DE>EN)
  - Synonym replacement for non-mountain terms
  - Template-based sentence generation

- Benefits:
  - 20-30% more training data
  - Improved model robustness
  - Better handling of rare mountain names

4. Ensemble Approach

Combine multiple NER models

- Architecture:
  - BERT-base (current)
  - SpaCy transformer model
  - Flair embeddings
- Benefits:
  - 2-4% F1 improvement through voting
  - Error correction across different architectures
  - More consistent predictions

5. Geographic Context Integration

Incorporate external knowledge bases

- Sources:
  - GeoNames database
  - Wikipedia mountain lists
  - Geographic coordinates
- Benefits:
  - Validation of predicted mountain names
  - Disambiguation of similar names
  - Enhanced precision through verification

6. Multilingual Support

Extend to other languages

- Target languages: Spanish, French, German
- Approach:
  - XLM-RoBERTa base model
  - Cross-lingual transfer learning
- Benefits:
  - Broader application scope
  - Improved tourism/travel applications
  - Market expansion opportunities

Priority

1. **Data Augmentation** - Quick wins with synthetic data
2. **Model Ensemble** - Reliable performance boost

3. **Advanced Architecture** - RoBERTa-large migration