Solution

1. Base cased BERT was used, tuned.

2. 5 classes instead of 3 (B-, I- MOUNTAIN, O),

classes - (B- I- MOUNTAIN, B- I- ELEVATION, O)

3. Trained for 10 epochs, used weighted cross-entropy loss, didn't do any sophisticated hyperparameter tuning.

4. Training dataset consisted of 10k synthetic mount sentences generated using Llama 3.1(check notebooks/data_related), 10k non mount sentences

- arond 2k samples invalidation set of basically same sythetic data that isn't close to real world (if we're not talking about books etc.) (99% accuracy was achieved).

- Achieved 87% f-score (macro, not weighted) on real world test set from 50 hard samples manually sampled and annotated from subreddit r/alpinism.

How to make better, how i would do it if i start over

- If i started over:

   1. Make a list of additional questions for each task (especially CV task).

   2. Spend more time for NER trying to manually find and annotate real-world samples.

   3. Don't try to learn all needed theory about architectures i'm gonna use.(especially CV DL solutions for matching confused me (papers))

How to make better:

    1. Better data (manual sampling (or sophisticated parsing that may introduce bias), annotation)

    2. Tuned for NER models

    3. Custom losses