Cross-lingual entailment detection

Gil Leibovici

Submitted as final project report for the NLP course, Reichman university, 2025

1 Introduction

Natural Language Inference (NLI) tests whether a hypothesis can be inferred from a premise—predicting one of three labels: entailment, contradiction, or neutral. While NLI has been widely studied in English many real-world applications require reasoning across languages. Cross-lingual NLI (XNLI) probes whether models trained on one language can generalize to others, and whether multilingual or translation-based pipelines can support inference when the premise and hypothesis are written in different languages.

This project focuses on English→Hebrew cross-lingual entailment: given an English premise and a Hebrew hypothesis, predict the NLI label. I chose this setting for three reasons. First, it reflects realistic scenarios (e.g., cross-border content moderation, multilingual QA and support, or news analytics) where evidence and claims appear in different languages. Second, Hebrew adds interesting modeling challenges: a distinct script, rich morphology, clitics, and data scarcity compared to English. Third, from an educational standpoint, the task is a compact way to compare families of multilingual techniques—translate-test/translate-train pipelines, multilingual encoders, and sentence-embedding approaches.

1.1 Related Work

Cross-lingual NLI benchmarks: XNLI [Conneau et al., 2018] extends MultiNLI [Williams et al., 2018] to 15+ languages and remains the standard benchmark for evaluating cross-lingual inference.

Multilingual pretraining and zero-shot transfer: Multilingual BERT (mBERT) [Devlin et al., 2019] demonstrated surprising zero-shot cross-lingual transfer without parallel data. Subsequent models improved scale and language coverage - XLM with translation language modeling [Conneau and Lample, 2019] and XLM-R with massive CommonCrawl pretraining [Conneau et al., 2020]. These models are commonly finetuned on English NLI and evaluated zero-shot on other languages, including Hebrew.

Sentence embeddings for multilingual NLI: Language-agnostic sentence encoders like LASER [Artetxe and Schwenk, 2019] and LaBSE [Feng et al., 2020] produce aligned representations across many languages and support simple classifiers (e.g., MLPs or distance-based decision rules) for cross-lingual tasks. SBERT [Reimers and Gurevych, 2019] introduced sentence-level contrastive training for English. multilingual SBERT [Reimers et al., 2020] extends this via knowledge distillation or parallel corpora, enabling efficient inference with decent cross-lingual alignment.

2 Methodology

I chose 3 cross-lingual models fine-tuned on NLI tasks as the baseline models and tried to improve their performances with different methods. The work combined data analysis, preprocessing, baseline evaluation, machine translation strategies, multilingual fine-tuning, embedding-based methods, prompting, and ensembling. Below I describe each stage in detail.

2.1 Dataset

I used HebNLI [HebArabNlpProject, 2024], released under the HebArabNlpProject. This dataset is the first large-scale resource for Hebrew NLI. It is derived from MultiNLI, where the original English premises and hypotheses were machine-translated into Hebrew using Google Gemini. Each example preserves the triplet structure of entailment, contradiction, or neutral labels.

2.2 Models

I evaluated cross-lingual models fine-tuned on NLI tasks, each derived from different base multilingual architectures:

- MayaGalvez/bert-base-multilingual-cased-finetuned-nli: based on mBERT.
- MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7: based on mDeBERTa-v3.
- joeddav/xlm-roberta-large-xnli: based on XLM-RoBERTa large [Conneau et al., 2020].

I also experimented with google/flan-t5-large (prompt-based sequence-to-sequence), multilingual SBERT architectures for representation learning and dicta-il/dictabert a Hebrew-specific BERT model, pre-trained on large-scale Hebrew corpora, which I fine-tuned on our dataset to strengthen the ensemble's classifier coverage of Hebrew morphology and vocabulary.

2.3 Data Analysis and Preprocessing

Label and genre distributions were analyzed to confirm balance across entailment, contradiction, and neutral. Sentence length distributions revealed heavy tails in Hebrew translations. Preprocessing steps included removal of NaNs, duplicate pairs, and outliers with excessively long translated sentences.

2.4 Data Translation and Evaluation of Fidelity

In addition to the original English-premise / Hebrew-hypothesis dataset, I created two translated variants to support downstream experiments:

- English-premise / English-hypothesis, obtained by translating Hebrew hypotheses into English.
- Hebrew-premise / Hebrew-hypothesis, obtained by translating English premises into Hebrew.

To further assess the quality of these translations, I employed LaBSE (Language-agnostic BERT Sentence Embedding) [Feng et al., 2020]. Each original sentence and its translation were transformed into embeddings, and cosine similarity has been evaluated between them. The average similarity score across all sentence pairs served as a measure of translation fidelity, providing a quantitative check on how much semantic meaning was preserved during translation.

To implement the translation, I relied on the Helsinki-NLP OPUS-MT models, which are open-source, reproducible, and widely adopted for academic NLP research. Specifically, I used Helsinki-NLP/opus-mt-en-he for English → Hebrew translation in the translate-premise pipeline, as it provides an efficient Transformer-base model trained on the OPUS parallel corpora. For the Hebrew → English direction, I selected Helsinki-NLP/opus-mt-tc-big-he-en, a Transformer-big variant offering improved handling of Hebrew's morphological richness and yielding higher-quality translations. These choices balance translation accuracy with computational efficiency.

2.5 Baselines

- Direct cross-lingual evaluation: Models were evaluated directly on English-premise / Hebrew-hypothesis pairs without translation.
- Translate-premise pipeline: Models were evaluated on the Hebrew-premise / Hebrew-hypothesis pairs, enabling Hebrew-only evaluation.
- Translate-hypothesis pipeline: Models were evaluated on the English-premise / English-hypothesis pairs, enabling English-only evaluation.

These baseline evaluations served as a way to measure the base performance of the models across different input configurations. By comparing results across the three settings, I was able to examine whether the models showed a preference for a particular language representation of the dataset (English vs. Hebrew) and to analyze how translation affected performance relative to direct cross-lingual inference.

2.6 Fine-Tuning

Each cross-lingual model was further fine-tuned with the English-premise / Hebrew-hypothesis structure. Training was conducted in Google Colab on T4 GPU, with each hyperparameter configuration requiring approximately 1–1.5 hours.

- Optimization used AdamW, with grid search over learning rates, batch sizes, and epochs.
- Gradient accumulation was applied to simulate larger effective batch sizes within limited GPU memory. In all training procedures gradient_accumulation_steps set to 2, meaning effective batch size = 2*batch size.
- Mixed-precision training (fp16) was enabled to reduce memory consumption and accelerate training.
- TensorBoard was used to monitor validation loss and macro-F1, guiding hyperparameter selection.

2.7 Ensemble Classifier

To improve robustness, I combined models' predictions via majority voting. In case of ties, a confidence-based tie-break was applied, selecting the label with the highest probability score across models. I constructed two ensemble classifiers:

First classifier (heterogeneous setup) included three models, each operating on a different language setup:

- The best-performing fine-tuned model on the English-premise / Hebrew-hypothesis configuration was evaluated directly on the same dataset.
- The second-best model was evaluated on the English-premise / English-hypothesis dataset (using the baseline, not the fine-tuned cross-lingual version).

- A Hebrew-only classifier was added by fine-tuning dicta-il/dictabert on the Hebrew-premise / Hebrew-hypothesis dataset, making it capable of handling purely Hebrew inputs.

Second classifier (homogenous setup):

- Composed of the two best-performing models trained on the English-premise / Hebrew-hypothesis dataset.

Both ensembles were designed to test whether combining models across different language alignments (first classifier) or focusing solely on optimized cross-lingual models (second classifier) yields better generalization and robustness in entailment prediction.

2.8 Prompt-Based Methods

I explored prompting with flan-T5-large. Flan-T5 is an instruction-tuned extension of T5 [Raffel et al., 2020], meaning it was trained not only to perform text-to-text tasks but also to follow natural-language prompts. This makes it particularly well-suited for zero-shot and few-shot evaluation without additional fine-tuning. Flan-T5 is primarily English-centric, therefore, I translated Hebrew hypotheses into English to leverage its strengths effectively.

- Zero-shot setting: Directly prompting with natural language instructions.
- Few-shot setting: Adding a small set of labeled premise–hypothesis pairs as in-context examples. The numbers of few-shot examples were dynamically configured, and the examples were taken from the training set.

Additional safeguards were applied during prompting:

- Max input tokens was set with buffer from the limit of the tokenizer to avoid truncation of the prompt in the prompt tokenization process.
- Constrained decoding was applied by computing the negative log-likelihood (NLL) of each candidate label given the prompt, i.e., NLL(label | prompt). This approach makes the model more robust to formatting drift in free generation output format and prevents it from producing invalid outputs.

Example zero-shot prompt:

```
Decide the relation between the premise and hypothesis.

Output EXACTLY one of: entailment, contradiction, neutral.

Premise: Just go on with what you’re doing now.

Hypothesis: Continue with the current task you have

Label:
```

2.9 Multilingual SBERT Training

I trained a multilingual SBERT model with SoftmaxLoss, aligning embeddings of English premises and Hebrew hypotheses for entailment classification. This approach provided cross-lingual sentence-level semantic representations tailored to NLI.

Multilingual SBERT is particularly useful in this setting for several reasons:

- Sentence-level representations: Unlike vanilla BERT, which is designed for token-level outputs, SBERT produces semantically meaningful sentence embeddings. This is well-suited for NLI, where the task depends on the global relationship between two sentences rather than fine-grained token alignment.
- Cross-lingual adaptability: When trained with pairs of English premises and Hebrew hypotheses, multilingual SBERT learns to project them into a shared embedding space, allowing semantic similarity and label classification across languages.

I selected SoftmaxLoss as the training objective because it directly optimizes multi-class classification (entailment, contradiction, neutral) rather than only learning similarity scores. Moreover, compared to contrastive losses (e.g., triplet loss), SoftmaxLoss is more straightforward for supervised tasks with explicit categorical labels such as NLI.

2.10 Code

All experiments were implemented in Google Colab and executed on a T4 GPU. The codebase is modular: models, hyperparameters and few-shot settings can be easily redefined and run again.

2.11 Technical Challenges

- Resource constraints: Training and evaluation were limited to Colab T4 GPU, restricting batch sizes and the number of experiments. This also limited the number of hyperparameter combinations that could be explored when searching for the best model.
- Checkpointing: Because long training runs occasionally crashed due to Colab runtime limits or disconnections, checkpoints had to be saved and resumed, ensuring that progress was not lost and training could continue from the same point.
- Translation caching: To avoid repeatedly incurring translation costs and time, all translated sentences were cached and reused across runs, ensuring reproducibility and efficiency.
- Data quality assurance: Data cleaning (removal of NaNs, duplicates, and extreme outliers) needed to be performed carefully to avoid accidentally discarding valid examples while still ensuring a high-quality dataset for training and evaluation.
- Trial of training joeddav/xlm-roberta-large-xnli as multilingual SBERT encountered out of memory issues with GPU and couldn't finish the training process.

3 Experimental Results

3.1 Evaluation Metrics

For model evaluation, both Accuracy and Macro-F1 were used, along with per-label precision scores and confusion matrices for detailed error analysis.

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} F1_c$$

Where:

- C = number of classes
- $F1_c$ = F1 score of class c

This averaging gives equal weight to all classes, regardless of label frequency, which is important in a 3-class setup (entailment, contradiction, neutral).

Per-label precision also computed to better understand model biases.

A confusion matrix was generated for each model, illustrating how often examples from one class were misclassified as another. This provided insights into systematic errors (e.g., models overpredicting "neutral").

3.2 Results

The average cosine similarity scores of the translations showed that the sentences semantics were largely preserved during translations, although not perfectly:

Translation	Average Cosine Similarity
Hebrew Hypothesis -> English Hypothesis	0.8878
English premise -> Hebrew premise	0.8762

Table 1: Average cosine-similarity scores of translations

For convenient purposes I'll denote the models I evaluated with the following shortcuts:

- mBERT custom = MayaGalvez/bert-base-multilingual-cased-finetuned-nli
- mDeBERTa custom = MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7
- XLM-RoBERTa custom = joeddav/xlm-roberta-large-xnli

Model	Premise	Hypothesis	Accuracy	Macro F1
mBERT custom	EN	HE	0.6131	0.6035
mBERT custom	EN	EN	0.8484	0.8474
mBERT custom	HE	HE	0.6923	0.6923
mDeBERTa custom	EN	HE	0.8721	0.8715
mDeBERTa custom	EN	EN	0.8925	0.8921
mDeBERTa custom	HE	HE	0.8597	0.8591
XLM-RoBERTa custom	EN	HE	0.8733	0.8723
XLM-RoBERTa custom	**EN**	**EN**	**0.9185**	**0.9178**
XLM-RoBERTa custom	HE	HE	0.8631	0.8621

Table 2: Baseline evaluations across different premise–hypothesis configurations

Model	Epoch	Learning Rate	Batch Size	Accuracy	Macro F1
mBERT custom	1	2e-5	16	0.7853	0.7856
mBERT custom	**1**	**4e-5**	**16**	**0.7883**	**0.7886**
mDeBERTa custom	**1**	**2e-5**	**16**	**0.8699**	**0.8703**
mDeBERTa custom	1	4e-5	16	0.8669	0.8672
XLM-RoBERTa custom	**1**	**2e-5**	**16**	**0.8859**	**0.8862**
XLM-RoBERTa custom	1	4e-5	16	0.8699	0.8699

Table 3: Fine-tuned results on validation set

Model	Epoch	Learning Rate	Batch Size	Accuracy	Macro F1
mBERT custom	1	4e-5	16	0.8076	0.8067
mDeBERTa custom	1	2e-5	16	0.8981	0.8972
XLM-RoBERTa custom	**1**	**2e-5**	**16**	**0.9151**	**0.9142**

Table 4: Fine-tuned results on test set

Model	Accuracy	Macro F1
First classifier (heterogeneous)	0.9095	0.9088
Second classifier (homogeneous)	**0.9197**	**0.9189**

Table 5: Ensemble classifiers performances

Prompt evaluation with the English premise / Hebrew hypothesis yielded accuracy and macro F1 results lower than 0.4. The table presents the results of the evaluation on English premise / English hypothesis. The examples chosen in 3 and 6 few shots are such that there is an equal distribution of labels:

Model	Accuracy	Macro F1
Zero-shot	**0.8111**	**0.8011**
1-shot	0.8077	0.7990
2-shot	0.8043	0.7934
3-shot	0.8077	0.7961
6-shot	0.8088	0.7978

Table 6: Prompt-based evaluation with Flan-T5

Model	Epoch	Learning Rate	Batch Size	Accuracy	Macro F1
mBERT custom	1	2e-5	16	0.6809	0.6804
mDeBERTa custom	**1**	**2e-5**	**16**	**0.7036**	**0.702**

Table 7: Multilingual SBERT training results

4 Discussion

What I learned from the experiments:

- Direct cross-lingual evaluation is challenging. Out-of-the-box multilingual models provided reasonable performance, but their accuracy lagged behind monolingual rich languages setups, reflecting the difficulty of aligning English and Hebrew in a shared embedding space.
- Translating Hebrew hypotheses into English and applying strong English NLI models consistently improved results compared to direct cross-lingual inference. This shows that high-quality MT can serve as an effective bridge. However, translations introduced semantic drift, which limited overall gains.
- Fine-tuning multilingual models directly on the English-premise / Hebrew-hypothesis dataset substantially improved macro-F1 compared to zero-shot transfer, confirming the value of adapting to the specific cross-lingual setting.
- Ensembles add robustness. The homogenous ensemble classifier outperformed all other methods.
- Zero-shot and few-shot prompting with Flan-T5 was intuitive and required no training, but results were weaker than fine-tuned discriminative models.
- Training multilingual SBERT with SoftmaxLoss yielded lower results than the baseline models and pipelines. Maybe other training and loss techniques can yield better results (e.g, Multiple Negatives Ranking Loss).

Next steps – Future work should focus on improving translation quality by experimenting with models that can both rephrase and refine sentence accuracy. For example, the Dicta-Instruct Hebrew large language model [Research, 2023] can be applied to rephrase Hebrew sentences, thereby correcting errors and mitigating imperfect translations. In addition, further experimentation with different model combinations for the ensemble classifier may yield better overall performance.

Overall, the study demonstrates that a hybrid approach—combining fine-tuned multilingual models, translation-based pipelines, and ensemble methods—offers the most robust solution for English–Hebrew cross-lingual entailment detection.

References

Mikel Artetxe and Holger Schwenk. LASER: Language-agnostic sentence representations. In *EMNLP*, 2019.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *NeurIPS*, 2019.

Alexis Conneau et al. XNLI: Evaluating cross-lingual sentence representations. In *EMNLP*, 2018.

Alexis Conneau et al. Unsupervised Cross-lingual Representation Learning at Scale. *ACL*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

F. Feng et al. LaBSE: Language-agnostic bert sentence embedding. In *ACL*, 2020.

HebArabNlpProject. HebNLI: Hebrew natural language inference dataset. `https://huggingface.co/datasets/HebArabNlpProject/HebNLI`, 2024.

Colin Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. In *JMLR*, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.

Nils Reimers et al. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP*, 2020.

Dicta Research. Dicta-Instruct: Hebrew instruction-tuned llm. `https://www.dicta.org.il`, 2023.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, 2018.