UC San Diego

Importance-Weighted LLM Fine-Tuning for Relation Extraction

CSE 261 / DSC 253: Advanced Data-Driven Text Mining

Tino Trangia, Teresa Lee, Raunak Sengupta

Problem Statement

Relation Extraction: Detecting and classifying semantic relationships between entities in text.

- Requires sufficient quantities of high-quality labeled training data (scarce, costly)
- Real-world data distribution ≠ training distribution (distribution gap)

Why is this important?

- Expert annotation makes large-scale RE infeasible in many domains (medical, legal, etc.)
- Weak supervision enables scale but introduces noise

UC San Diego

Research Objective

Paper: ATLANTIS: Weak-to-Strong Learning via Importance Sampling (Liu et al., ACL 2025)

- Real distribution is impossible to calculate directly, so estimate it using the probability gap between a small base model and a fine-tuned reference model
- Let p* be the optimal distribution, pbL be the large model's base distribution, pbS be the small model's base distribution, and pr the reference model's distribution

$$\frac{p^*(y|x)}{p_b^L(y|x)} \propto \frac{p_r(y|x)}{p_b^S(y|x)} \longrightarrow \frac{p^*(y|x)}{p_r(y|x)} \propto \frac{p_b^L(y|x)}{p_b^S(y|x)}$$

Goal

- Implement ATLANTIS from scratch and apply it to sentence relation extraction
- Investigate whether it can improve fine-tuned model performance in the presence of LLM-generated weak labels
- Benchmark extraction performance compared to the standard SFT baseline

ATLANTIS

- Using the three models we pre-compute importance weights for each training example
- Train the large model and multiply loss by the cached weights
- Weights bias the optimization direction toward influential examples

Algorithm 1 ATLANTIS

Input p_b^L, p_b^S, p_r, $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, max training steps M, learning rate α

Output p^*

$W \leftarrow \{\frac{p_b^L(y_i|x_i)}{p_b^S(y_i|x_i)} p_r(y_i|x_i)\}_{i=1}^N$

$p_{\theta_0} \leftarrow p_b^L$

for $m = 1$ to M **do**

  $\mathcal{B} \leftarrow next(\mathcal{D})$

  $W_B \leftarrow next(W)$

  $\mathcal{L}(\theta_{m-1}) \leftarrow -\sum_{i \in \mathcal{B}} \frac{W_B^i}{|\mathcal{B}|} \log p_{\theta_{m-1}}(y_i|x_i)$

  $\theta_m \leftarrow \theta_{m-1} - \alpha \frac{\partial \mathcal{L}(\theta_{m-1})}{\partial \theta_{m-1}}$

end for

$p^* \leftarrow p_{\theta_M}$

SemEval-2010 Task 8 Dataset

Characteristics

- Designed for sentence-level classification of mutually exclusive semantic relations
- 8K training samples, 2.7K test samples
- Entity pairs are tagged in training input, corresponding to a relation type
- 19 relation classes (including “other”), direction-collapsed to 10
- Manually annotated relation set

Gold vs. weakly labeled

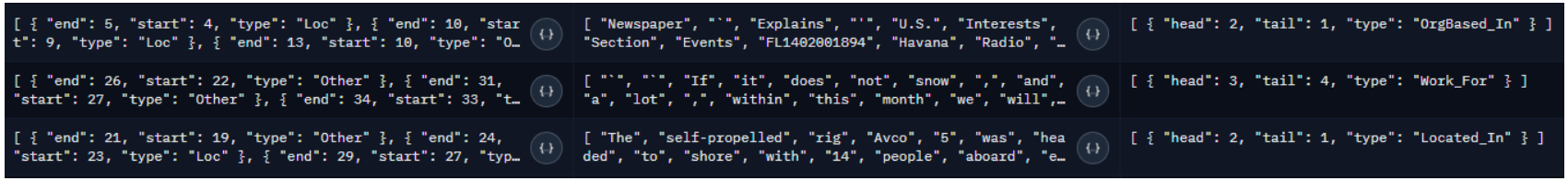
- Fine-tuning on both the original training set and on a weakly labeled training set to simulate real-world use of weak supervision
- Adjustable noise rate from 20-50%

The system as described above has its greatest application in an arrayed <e1>configuration</e1> of antenna <e2>elements</e2>.	3 Component-Whole(e2,e1)
The <e1>child</e1> was carefully wrapped and bound into the <e2>cradle</e2> by means of a cord.	18 Other
The <e1>author</e1> of a keygen uses a <e2>disassembler</e2> to look at the raw assembly code.	11 Instrument-Agency(e2,e1)
A misty <e1>ridge</e1> uprises from the <e2>surge</e2>.	18 Other
The <e1>student</e1> <e2>association</e2> is the voice of the undergraduate student population of the State University of New York at Buffalo.	12 Member-Collection(e1,e2)

CoNLL2004 Dataset

Characteristics

- Designed for joint named entity recognition and relation extraction
- 922 training samples, 288 test samples
- Contains dictionaries with entity positions and relation type/positions
- 5 relation classes
- We parse into a natural sentence and split samples with multiple entity/relation pairs into separate training instances
- Curated and human annotated from news articles



Models

google/flan-t5-base · Hugging Face

Hyperparameters

- Full fine-tuning
- 8 epochs
- 3e-4 learning rate
- Batch size 32

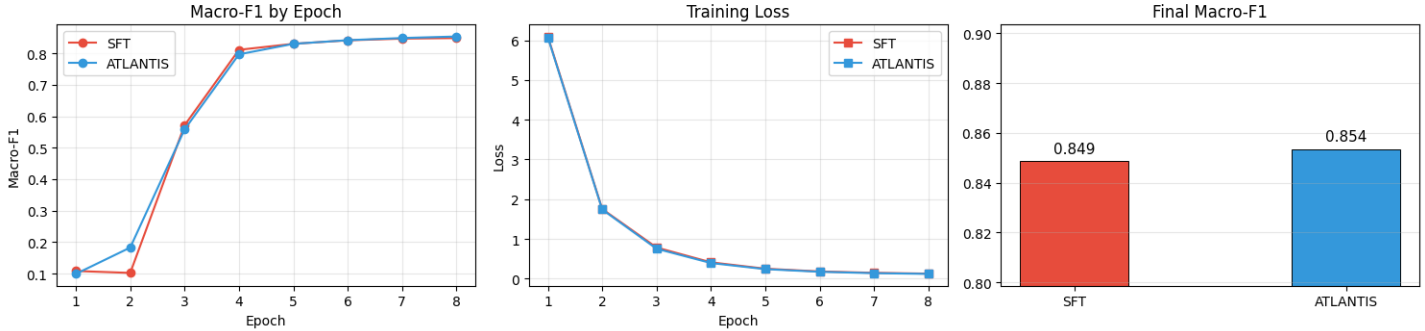
Qwen/Qwen2-1.5B · Hugging Face

Hyperparameters

- Full fine-tuning
- 3 epochs
- 2e-5 learning rate
- 0.5/1.5B configuration: Batch size 8, gradient accumulation 2
- 1.5/7B configuration: Batch size 1, gradient accumulation 16

Results

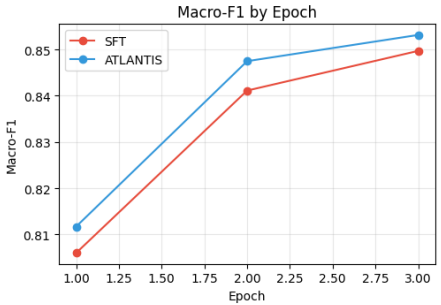
Flan-T5 on Semeval

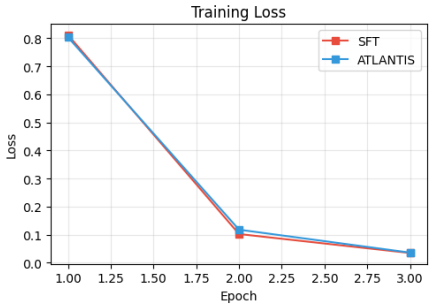


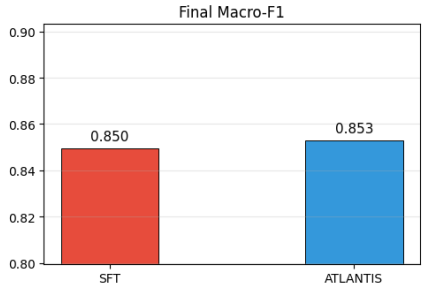
ATLANTIS importance weighting can be made compatible with both encoder-decoder and decoder-only models for relation extraction fine-tuning

Results

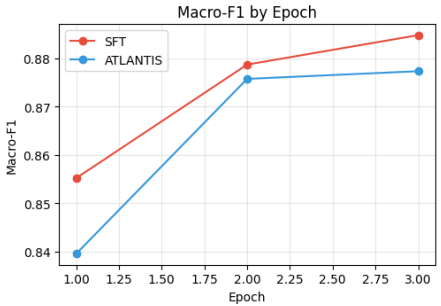
Qwen2-1.5B on Semeval

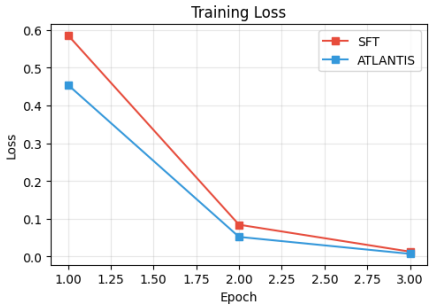


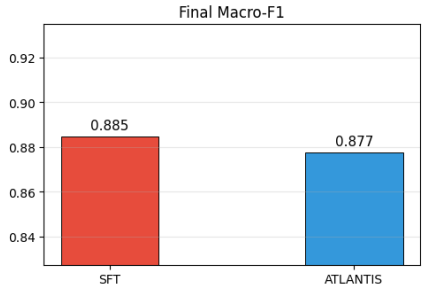




Qwen2-7B on Semeval

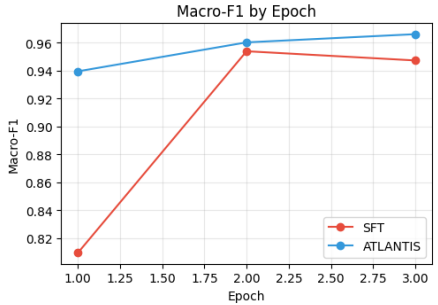


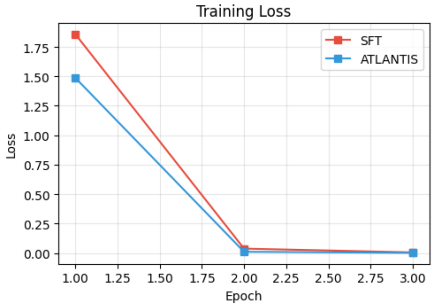


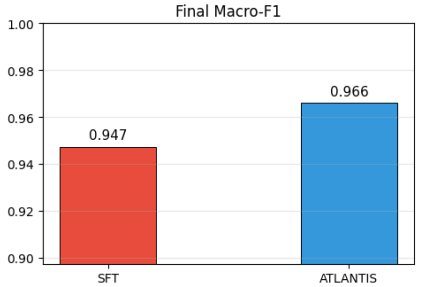


Results

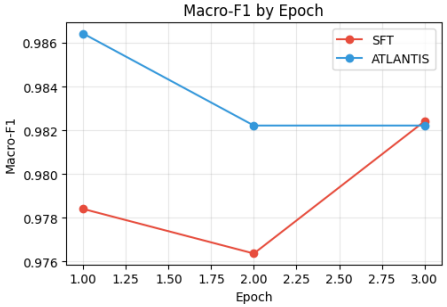
Qwen2-1.5B on CoNLL2004

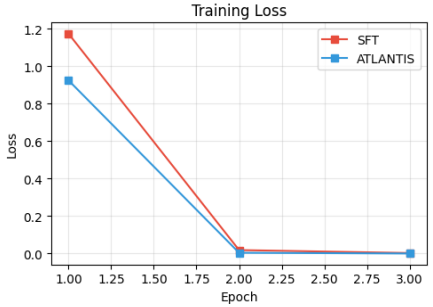


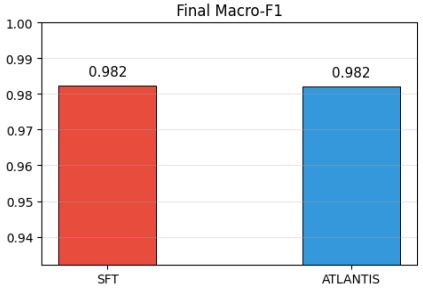




Qwen2-7B on CoNLL2004

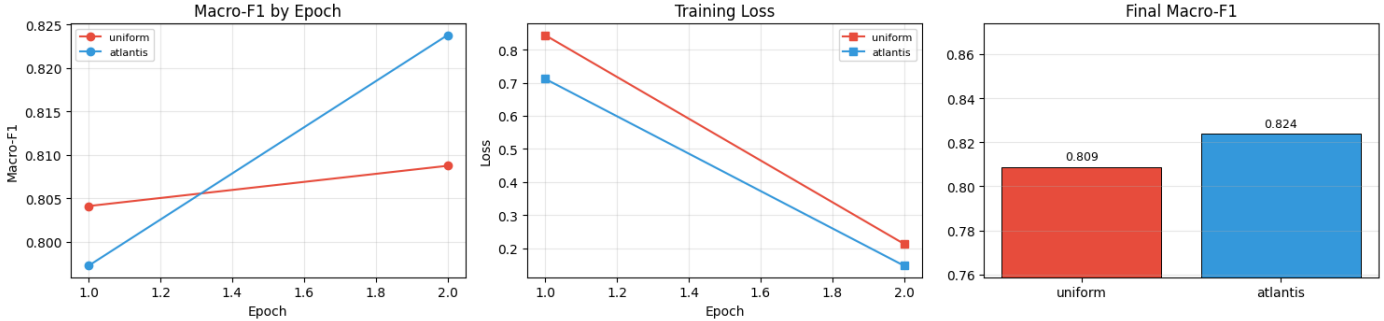






Results

Qwen2-1.5B on weakly labeled (w/ 20% noise) Semeval



- Importance weighting may yield slight improvements to relation extraction performance, but the effect is sensitive to experiment settings
- Standard SFT remains competitive where clean data is available, but zero/few-shot relation extraction methods remain highly desirable

Limitations

- Setting mismatch:
  - Original ATLANTIS was designed for general instruction tuning, with official instruction-tuned models as reference and evaluation on mixed benchmarks
  - Instead, we apply it to a much narrower task with a small fine-tuning dataset
- Datasets are evaluated on annotated entity pairs only, bypassing the entity detection component of full RE pipelines and inflating absolute F1 relative to expected numbers
  - Can evaluate on a more diverse set benchmarks (e.g. ADE, TACRED, NYT, etc)
- Lack of SOTA comparisons: we compare the relative improvements of the method against SFT, but we do not measure absolute performance
  - Can be combined with other methods (e.g. data selection) to demonstrate a realistic end-to-end model
  - Can conduct ablation studies across noise rates, model sizes, weight clipping, etc.
- Lack of document-level relation extraction support (e.g. with Re-DocRED dataset)

Thank you!
Questions?