EA-MT Entity-Aware Machine Translation in English to Chinese

Saturday 7th December, 2024

Jou-Yi Lee
jlee1039@ucsc.edu

Abstract

The primary objective of this project is to tackle the challenge of translating a given input sentence from the source language, English, into the target language, Chinese. This task focuses on addressing the accurate translation of named entities, which may be rare, ambiguous, or unknown to traditional machine translation systems.

The goal is to develop a machine translation system capable of effectively handling such named entities within input sentences, ensuring that their translation into the target language is precise.

1 Introduction

Named entity translation poses significant challenges in machine translation due to contextual ambiguity and linguistic nuances. For example, the term *"Apple"* may refer to a fruit or a technology company, depending on the context, making accurate translation difficult without sufficient contextual clues.

Multi-word entities, such as organizational or product names, further complicate translation. Literal translations often fail to convey accurate meanings, as seen with terms like *"Department of Justice"*, which require culturally appropriate equivalents to maintain semantic accuracy.

To address these issues, this study utilizes the **IWSLT 2017 English-to-Chinese** dataset, focusing on named entity translation. Named entities are identified using **SpaCy**'s NER capabilities and integrated into a machine translation pipeline based on the **mBART** architecture. By leveraging pre-trained and fine-tuned weights, this approach aims to improve the accuracy and consistency of complex entity translations, ensuring contextual fidelity in cross-lingual tasks.

2 IWSLT Dataset

The International Conference on Spoken Language Translation (IWSLT) 2017 dataset involves English-to-Chinese translations and provides a valuable resource for real-world conversational examples. For instance:

"Thank you so much, Chris. And it's truly a great honor to have the opportunity to come to this stage twice; I'm extremely grateful."
"非常谢谢，克里斯。的确非常荣幸能有第二次站在这个台上的机会，我真是非常感激。"

In IWSLT, English sentence lengths are more dispersed compared to Chinese sentences. Chinese sentence lengths are concentrated in shorter intervals because Chinese is a character-based language.

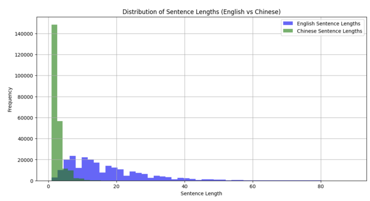

图 1: Distribution of Sentence Lengths (English vs. Chinese)

2.1 Linguistic Structural Differences

2.1.1 Character-Based vs. Word-Based Languages:

Chinese, as a character-based language, treats each character as the smallest semantic unit. In contrast, English, as a word-based language, requires additional auxiliary words to form complete sentences. For instance:

English: *"The dog is on the mat."* (6 words)
Chinese: "狗在垫子上。" (5 characters)

2.1.2 Differences in Word Order and Syntax:

English generally follows a fixed Subject-Verb-Object (SVO) structure, while Chinese exhibits more flexibility in word order, allowing sentence structure to vary depending on the intended emphasis. However, this flexibility in Chinese may lead to challenges during translation, especially when attempting to map flexible Chinese structures to the fixed SVO patterns of English. Furthermore, Chinese allows for reordering of elements to emphasize different parts of a sentence:

Standard: "我喜歡這本書。" (Subject-first: *"I like this book."*)

Emphasis: "這本書我喜歡。" (Emphasizing *"this book"*)

2.1.3 Sentence Complexity

In the Figure(2) we can see the complex sentence type appears most frequently. Complex sentences are typically sentences with one independent clause and one or more dependent clauses, often involving subordinating conjunctions like "because," "although," or "since.". Compound sentences are the second most frequent type it consist of two or more independent clauses connected by coordinating conjunctions such as "and," "but," or "or."

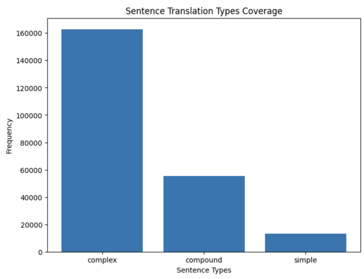

图 2: The Translation Coverage Types in IWSLT Dataset

The high frequency of complex sentences mean the translation model to handle subordinate clauses effectively.

3 Methodology - Flowchart

The process of integrating entity-aware translation is divided into several steps to manage the complexities of accurately translating named entities from English to Chinese. In this project, we

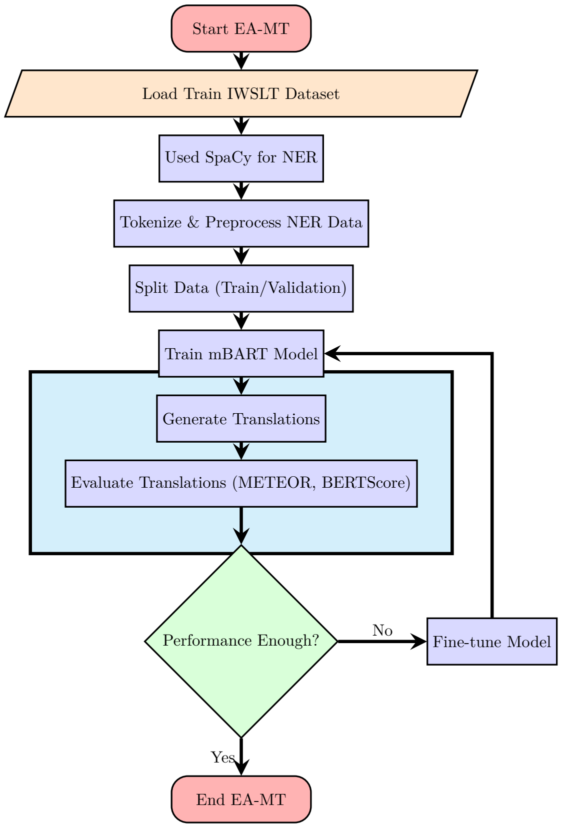

图 3: Programming Flowchart for Translation Evaluation Process

utilized SpaCy to extract named entities from English sentences and customized the entity annotation format. Next, the pre-annotated entity data was used as input for the mBART model. During training, the mBART model was fine-tuned with a focus on high-frequency entities such as `CARDINAL`, `DATE`, and `PERSON`. After translation, the entity annotations in the translated results were restored to readable text, ensuring that the output was both fluent and accurate. Multiple evaluation metrics, including METEOR and ROUGE, were employed to assess the model's performance in translating entities and maintaining contextual accuracy.

4 Named Entity Recognition

Named entities can be conceptualized as specific instances of broader entity categories (for example, "Apple" is an instance of the category `Organization`). A named entity refers to a real-world object or concept that can be identified by a proper name, such as a person, location, organization, or product. Recognizing named entities is crucial for tasks such as machine translation, where the accurate identification and handling of named entities can significantly enhance translation precision.

4.1 SpaCy: `en_core_web_trf`

SpaCy provides pre-trained models for multiple languages. For this study, we selected the `en_core_web_trf` model to identify all named entities in the IWSLT dataset. This model leverages the Transformer architecture, which significantly improves the accuracy of named entity recognition, particularly in contexts with complex linguistic structures.

4.2 Named Entity Tags and Descriptions

Using SpaCy, we employed the `.label_` attribute to extract named entities and implemented a custom Python function, `mark_entities()`, to format the extracted entities into a structured representation of the form `«{ent.label}:{ent.text}>`, for example, `«ORG:Apple»`. This explicit tagging approach enhances the model's ability to recognize and handle named entities during the translation process. By explicitly marking entities, the translation model gains a better understanding of their semantic significance, reducing the likelihood of misinterpretation or loss of meaning during translation.

4.3 Named Entity Tags Distributiion in IWSLT Dataset

In Figure(4), `CARDINAL` (numerical data) and `DATE` (dates) are occurring most frequently. This highlights the importance of numbers and dates in the dataset, such as references to years, statistics, or expressions of time. `PERSON` (names of people) and `GPE` (geopolitical entities like cities or countries) closely follow. These frequent entity types are vital for the model to learn and recognize entities effectively, which significantly improves the translation's accuracy.

Figure(7) shows the distribution of entity tags identified from the IWSLT dataset. Most entities fall into ten primary categories. This further demonstrates that optimizing the model's Entity-Aware translation capabilities requires focused training on high-frequency types, such as `CARDINAL` and `DATE`.

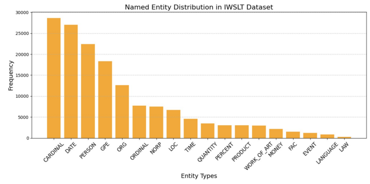

图 4: NER Bar Chart in the IWSLT Dataset

5 Model Design — mBART

The **mBART** model, developed by Facebook AI (now Meta AI), is a Transformer-based encoder-decoder architecture designed for multilingual sequence-to-sequence tasks. This study utilizes the `mbart-large-50-many-to-many-mmt` model, a pre-trained multilingual translation system fine-tuned for named entity annotations to improve translation accuracy. The model's tokenizer is customized to process entity markers, leveraging contextual information for enhanced performance.

5.1 mbart-large-50-many-to-many-mmt

The `mbart-large-50-many-to-many-mmt` model supports translation across 50 languages, including English and Chinese. Pretrained using the ML50 benchmark dataset, it demonstrates exceptional performance in both high- and low-resource language pairs. According to the paper *Multilingual Denoising Pre-training for Neural Machine Translation*, mBART achieved up to a 12 BLEU score improvement for low-resource languages and 9.5 BLEU for unsupervised tasks.

Built on a 12-layer Transformer encoder-decoder architecture with 680 million parameters, mBART excels in multilingual and low-resource translation tasks, even with reduced parallel corpora. Tested on datasets like TED Talks (IWSLT) and Europarl, it showcases state-of-the-art results, particularly in unsupervised machine translation.

6 EA-MT Process

The EA-MT process utilizes SpaCy for NER and the mBART model for En to Chinese translation on the IWSLT 2017 dataset. Named entities are tagged and preprocessed with special formats before being fed into the `mbart-large-50-many-to-many-mmt` model. The tokenizer encodes input data with truncation and padding, using a maximum sequence length of 256. The dataset is split into 90% training and 10% validation for evaluation after each epoch.

6.1 Training Parameter Settings

Training parameters are summarized below:

Parameter	Value
Learning Rate	1×10^{-5}
Batch Size (Train/Val)	4
Weight Decay	0.01
Epochs	20
Training Samples	231,266
Early Stopping Patience	3
Max Sequence Length	256
GPU	Google Colab (A100 GPU)
Total Training Time	>100 hours

表 1: Training Parameter Summary

Validation was performed after each epoch, with the model saved to prevent data loss. Logs were stored using `logging_dir` to track progress.

6.2 Trainer Initialization

The mBART model was initialized with the following components:

- **Datasets:** `train_dataset` and `eval_dataset`.
- **Tokenizer:** `MBart50Tokenizer`.
- **Callbacks:** Early stopping (`early_stopping_patience=3`) to halt training if validation performance plateaued for 3 consecutive epochs.
- **Model Saving:** Model weights were saved after each epoch to a specified directory.

6.3 Model Training and Validation Loss

Figure 5 shows the loss trends:

- **Training Loss:** Decreased from 0.20 to 0.0707 over 20 epochs, with a noticeable decline after the 14th epoch.
- **Validation Loss:** Declined steadily from 0.20 to 0.16 in the first 8 epochs, stabilizing around 0.14 with slight fluctuations after the 10th epoch.

6.4 Future Improvements

Future work could include:

- Early stopping based on dynamic validation performance.
- Enhanced regularization (e.g., increased weight decay, dropout, data augmentation) to prevent overfitting and improve generalization.

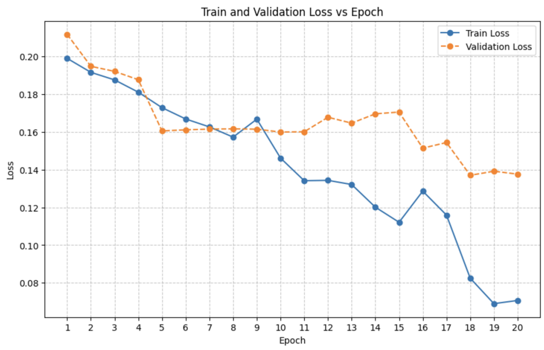

图 5: Train and Validation Loss vs Epoch

7 EA-MT Evaluation Process

Evaluation process we implemented an evaluation framework using metrics such as **METEOR** and **BERTScore** to assess translation quality, while also considering synonym relationships.

7.1 Tokenization Issues

The initial **METEOR** score was 0.0.Because **METEOR** tool is primarily designed for English, and its support for Chinese is limited. In English, word boundaries are clearly defined, but in Chinese are characters base(no explicit spaces between each characters). This lack of word boundaries caused the **METEOR** score to be abnormally low. To address this issue, we used **Jieba**, an open-source segmentation tool specifically designed for Chinese. **Jieba** splits Chinese sentences into words or phrases, enabling effective tokenization for NLP tasks. For example:
Chinese without tokenization: " 这是一本书" → [" 这是一本书"]
Jieba tokenization → [" 这", " 是", " 一本", " 书"]

7.2 Punctuation Format Inconsistencies

In Chinese text, punctuation is typically not spaced from words, whereas **METEOR** may be

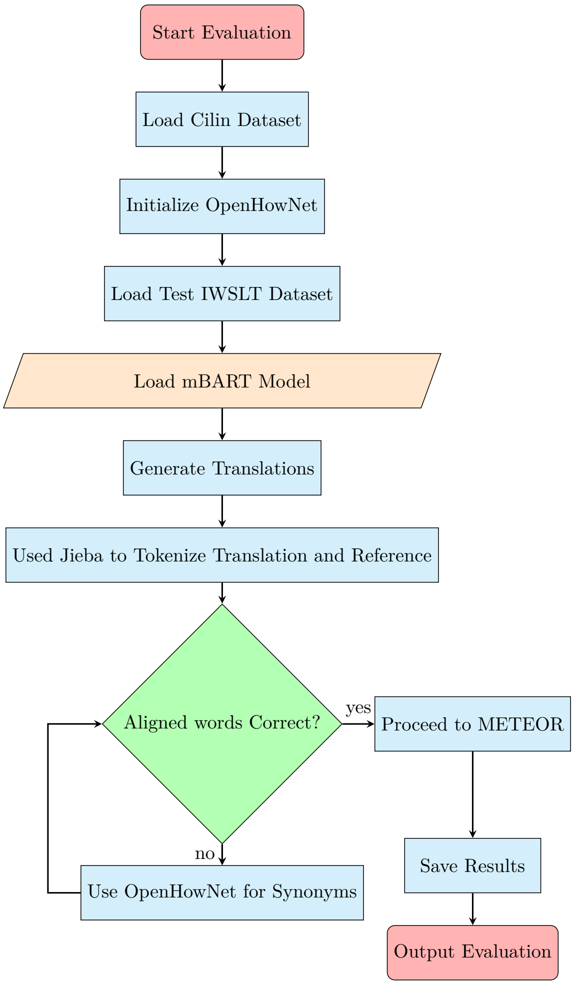

图 6: Evaluation Workflow for Translation Using METEOR and OpenHowNet

affected by such differences. To resolve this, we used regular expressions to separate punctuation from words and remove unnecessary spaces in the custom **preprocess_text** function to ensure consistent formatting.

7.3 Lack of Built-in Synonym Support for Chinese

The METEOR Eval is for English relies on lexical databases (such as **WordNet**) for lemmatization and synonym matching. However, **METEOR** does not support Chinese synonym databases, which means that synonym matching cannot be performed automatically. Therefore, semantic similarity for Chinese requires additional handling. To address this, we introduced **Tongyici Cilin (Harbin Institute of Technology (HIT) Synonym Wordnet)** as the core resource for synonym matching and constructed a Chinese synonym dataset using it.

7.4 Tongyici Cilin HIT Synonym Wordnet)

HIT Synonym Wordnet is a widely used lexical resource for Chinese natural language processing (NLP) developed by the HIT. It is similar to the English **WordNet**, and it effectively expands the coverage of synonyms that **METEOR** lacks for Chinese. Synonym Wordnet five levels:

- Level 1: Top-level categories ("Person", "Object").
- Level 2: Subcategories under each top-level category,("Person" have "Human Behavior"...).
- Level 3: More specific subcategories.
- Level 4: Specific semantic units.
- Level 5: Individual words or phrases.

7.4.1 Example Analysis

- Aa: Top-level category, meaning "Person".
- 01A01: Subcategory specific class of people.
- Synonym relation: The words " 人" and " 人物" are semantically identical or very similar.

7.5 Insufficient Coverage in Tongyici Cilin

Although we used **Tongyici Cilin**, the coverage of synonyms in this resource is limited, and it does not include all possible synonyms. Therefore, we also incorporated **OpenHowNet**, using the **calculate_word_similarity** method to compute semantic similarity between unaligned tokens, thus supplementing the inadequacy of **Tongyici Cilin** and improving synonym matching.

7.6 Integration of Synonym Wordnets with METEOR

By using **Tongyici Cilin** in combination with **OpenHowNet**'s semantic similarity calculation, we were able to consider synonym matching when calculating **METEOR** scores. During the evaluation process, for unaligned tokens, we check whether they can be matched with synonyms by semantic similarity and include them in the aligned tokens.

8 Evaluation Metrics

We used METEOR and BERTScore to evaluate machine translation quality.

8.1 METEOR Calculation

METEOR (Metric for Evaluation of Translation with Explicit ORdering) balances precision and recall for a holistic assessment. METEOR include:

- **Tokenization and Alignment:** Tokenize hypothesis and reference sentences. Align tokens based on exact matches, stems, synonyms, or paraphrases.

- **Harmonic Mean (F-score):**

$$F_{\text{mean}} = \frac{(1+\beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R}, \quad \beta = 3$$

- **Fragmentation Penalty:**

$$\text{Pen} = \gamma \cdot \left(\frac{\text{Chunks}}{\text{Matches}}\right), \quad \gamma = 0.5$$

- **Final Score:**

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Pen})$$

8.1.1 Enhancements for Chinese Translation

METEOR's English optimization was extended for Chinese using:

- **Synonym Lexicon:** The HIT dataset provided a Chinese synonym lexicon for semantic matching (e.g., 好 and 优良).

- **OpenHowNet:** Integrated semantic similarity scores for non-exact matches, aligning tokens with similarity above a threshold (e.g., 0.8).

8.2 BERTScore

BERTScore evaluates semantic similarity using contextual embeddings from BERT. Cosine similarity measures alignment between generated and reference text:

$$\text{cosine\_similarity}(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

where x and y are word embedding vectors.

8.3 Overall BERTScore and NER Accuracy

The **EA-MT: Entity-Aware Machine Translation** model achieved strong results in English-to-Chinese translation: **NER** component achieved a high accuracy of **81.30%**, demonstrating the model's effectiveness in identifying named entities crucial for translation accuracy.

Metric	Description and Score
Precision	80.08% of the generated content aligns with the reference.
Recall	79.33% of the reference content is captured.
F1 Score	0.7965 - A balanced measure of precision and recall.

表 2: BERTScore Metrics for Translation Performance

8.4 Conclusion

With an NER accuracy of 81.30%, the EA-MT system demonstrates robust entity recognition capabilities, which are crucial for preserving semantic integrity during translation. Combined with a METEOR score of 0.6649 and high BERTScore metrics (Precision: 0.8008, Recall: 0.7933, F1: 0.7965), the system excels in maintaining semantic similarity and ensuring proper word order alignment. The slightly higher Precision score reflects its ability to deliver accurate translations, minimizing errors, while the balanced F1 score indicates strong overall coverage of reference translations. In this Project, the EA-MT system performed superiorly in English-to-Chinese machine translation, particularly in scenarios where named entity recognition plays a critical role.

9 Task Distribution

9.1 Codeing Distribution

Task	Jou-Yi Lee (%)	Yifei Gan (%)	Camellia Bazargan (%)
Find Dataset	60	?	?
Research	80	?	?
Data cleaning	90	?	?
mBART Model	80	?	?
METEOR Eval	80	?	?
Experiment	80	?	?

表 3: Codeing Distribution by Team Members

9.2 Overall Distribution

Task	Jou-Yi Lee (%)	Yifei Gan (%)	Camellia Bazargan (%)
Presentation	40	30	30
Code Contribution	80	?	?
Slide Preparation	80	?	?
Total Distribution	80	?	?

表 4: Task Distribution by Team Members

10 References

1. Liu, Y., et al. (2020). "Multilingual Denoising Pre-training for Neural Machine Translation." In *Transactions of the Association for Computational Linguistics (TACL).*

2. Tang, Y., et al. (2020). "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning." In *Findings of the Association for Computational Linguistics: EMNLP 2020.*

3. Conneau, A., et al. (2020). "Unsupervised Cross-lingual Representation Learning at Scale." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).*

4. Liu, Z., et al. (2018). "Synonyms and Semantic Relations in HIT-SCIR." In *Proceedings of the 27th International Conference on Computational Linguistics (COLING).*

5. Qi, J., et al. (2020). "CiLin: A Chinese Lexical Semantics Resource for NLP." In *Journal of Chinese Information Processing (JCIP).*

6. Hugging Face and TensorFlow. (2017). "IWSLT 2017 English-to-Chinese Translation Dataset." Retrieved from https://www.tensorflow.org/datasets/community_catalog/huggingface/iwslt2017.

7. Xu, H., et al. (2019). "OpenHowNet: An Open Sememe-Based Lexical Knowledge Base for Chinese." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).*

11 Appendix

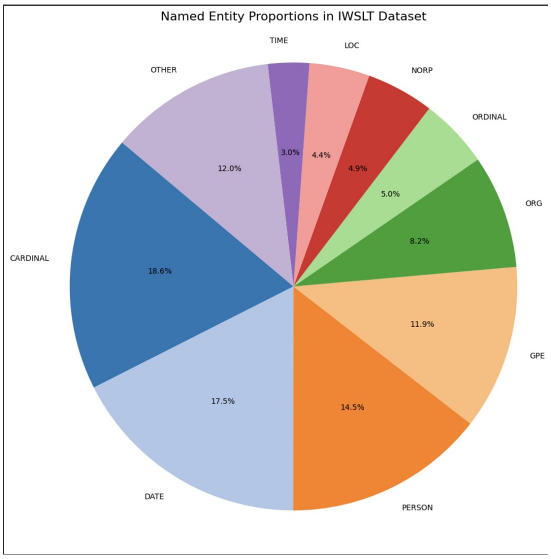

图 7: NER Pie Chart in the IWSLT Dataset

表 5: Epoch-wise Training and Validation Loss

Epoch	Train Loss	Validation Loss
1	0.1989	0.2116
2	0.1915	0.1947
3	0.1875	0.192
4	0.181	0.1876
5	0.1728	0.1605
6	0.1668	0.1611
7	0.1626	0.1614
8	0.1572	0.1616
9	0.1667	0.1614
10	0.1461	0.1599
11	0.1341	0.1600
12	0.1343	0.1677
13	0.1321	0.1646
14	0.1201	0.1695
15	0.1121	0.1705
16	0.1286	0.1514
17	0.1159	0.1543
18	0.0825	0.1370
19	0.0690	0.1391
20	0.0707	0.1376

Tag	Description	Example
PERSON	A person's name	"Barack Obama"
NORP	Nationalities, religious groups, or political groups	"American"
FAC	Buildings, landmarks, or facilities	"Eiffel Tower"
ORG	Organizations, companies, or institutions	"Google"
GPE	Countries, cities, or geopolitical regions	"France"
LOC	Non-GPE locations, such as mountain ranges or bodies of water	"Mount Everest"
PRODUCT	Objects, vehicles, or products	"iPhone"
EVENT	Named events	"Olympics"
WORK_OF_ART	Titles of books, songs, or artworks	"Mona Lisa"
LAW	Legal documents or clauses	"First Amendment"
LANGUAGE	Languages	"English"
DATE	Dates or time periods	"2024-01-01", "Wednesday"
TIME	Specific times	"2 PM", "noon"
PERCENT	Percentage expressions	"50%"
MONEY	Monetary amounts	"$20"
QUANTITY	Measurements	"10 kg"
ORDINAL	Ordinal numbers	"first", "2nd"
CARDINAL	Cardinal numbers	"one", "100"

表 6: Named entity tags, their descriptions, and examples in en_core_web_trf.