Bitext Processing: Translation, Alignment, and Word Sense Disambiguation

1 Word sense disambiguation

1.1 Method description

Preprocessing Data We create a length-sorted substitution dictionary for multi-word expressions and special characters.

AMUSE API Integration Preprocessed data are uploaded in JSON format (`'lang': 'EN'`), returning tokens with BabelNet synset IDs.

Token Alignment and Extraction We align API output with source tokens and extract those with gold sense tags.

Evaluation We use `evaluate_wsd.py` to compare output against `se13.key.txt` annotations.

1.2 Accuracy

The overall F1-score of 0.576 (947 correct, 697 incorrect).

1.3 Examples of typical or interesting errors

Polysemous Word Disambiguation: Words with multiple senses (e.g., ”plan”, ”world”) contribute significantly to errors.

Preprocessing-Induced Errors: Token replacement occasionally converts multi-word expressions (e.g., ’greenhouse gases’) into ’?’ with synset ID 0.

Domain-Specific Terminology: Professional vocabulary from specialized domains (e.g., ’Washington’, ’Technology’) poses significant disambiguation challenges.

1.4 Reflection

Strengths: Easy implementation with low computational cost.

Weaknesses: Preprocessing errors and limited contextual awareness result in 57.6% accuracy.

Limitations: API dependency prevents fine-tuning and tokenization mismatches create performance ceilings.

2 Translation

2.1 Method description

Initialize the Google Translator We configure GoogleTranslator with `source='auto'` and `target='zh-CN'`.

Preprocessing Data Sentences are batched (batch size=20) and preprocessed by removing leading/trailing whitespace and normalizing multiple spaces.

JSON Format Output Translation results are saved in JSON format with `"src"` (English), `"mt"` (Chinese), and `"status"` fields.

CometKiwi Evaluation CometKiwi evaluates the JSON file, returning quality scores for each sentence pair.

2.2 Evaluation results

The dataset contains 301 sentences with a system score of 0.75 (range: 0.172-0.893). Quality distribution: 128 high (≥0.8), 93 good (0.7-0.8), 51 moderate (0.6-0.7), and 29 low (<0.6).

2.3 Error Analysis

Two main error categories were identified:

Complete Translation Failures: Sentences with scores below 0.2 (e.g., 264, 41) returned untranslated text due to complexity and API errors.

Proper Noun Translation Issues: Sentences scoring 0.4-0.6 left proper nouns untranslated or inconsistently transliterated.

2.4 Reflection

Strengths: Achieves 0.750 system score with effective handling of straightforward sentences.

Weaknesses: Performance degrades on terminology-heavy content (9.6% scored below 0.6).

Limitations: Lacks domain-specific tuning and struggles with proper noun transliteration.

3 Word Alignment

3.1 Brief Summary

First, we load the source sentences and their Chinese translations into two separate files. Since Chinese text lacks word boundaries, perform word segmentation on the Chinese translations before alignment to enable proper tokenization. Then use SimAlign to align the source sentences with the Chinese translations.

3.2 Setup of the Method

Hyperparameters we tried

1. model: `bert-base-multilingual-cased`
2. distortion: `0`
3. null align: `1`
4. token type: `bpe`
5. matching methods: `a (argmax)`
6. num-test-sents: `None`
7. batch size: `100`
8. log: `false`
9. device: `cpu`
10. add probs: `false`
11. layer: `8`

3.3 Evaluation and Error Analysis

We found some successes and errors in the alignment output, here are some examples:

1. In the 60th sentence, “10” and “10” are aligned correctly, “foreign” and “外国” are aligned correctly, and “companies” and “公司” are aligned correctly.

2.

Complete Translation Failures

Sentences with CometKiwi scores below 0.2 (e.g., sentences 264, 41) were not translated at all, returning the original English text. These failures occurred due to:

- Sentence complexity and excessive length
- API processing errors

Proper Noun Translation Issues

Sentences scoring 0.4-0.6 contained numerous proper nouns and technical terms that the system struggled to handle appropriately. Common issues included:

- Leaving proper nouns untranslated in Chinese output

- Inconsistent handling of technical terminology

- Uncertainty between transliteration and semantic translation approaches

3.4 Reflection

The overall translation quality achieved a CometKiwi system score of 0.750, indicating a good quality for most of the sentences.

Where Google Translate Worked Best: The translator works best when the sentence is straightforward and easy to understand.

Where Google Translate Worked Worst: The translator works worst when the sentence is with a lot of terminology and difficult to understand.