CMPUT 497 Assignment 2: Filtering

Yuhui Sun and **Tianyang Ling**
University of Alberta
Edmonton, AB, Canada
{yuhui15,tling4}@ualberta.ca

1 Introduction

We implemented two systems for sense projection. The LLM baseline generated one Chinese lemma per English synset using WordNet(Fellbaum, 1998) glosses and a concise bilingual prompt. The main method, ExpandNet(of Alberta, 2024), followed the canonical pipeline with translation, dbalign alignment, and dictionary-based filtering to project English senses into Chinese contextually. While the baseline provides high coverage, it does not incorporate sentence context and therefore struggles with fine-grained distinctions. ExpandNet, in contrast, uses contextual alignment and dictionary filtering to achieve more precise token-level projections by removing unsupported alignment links. In our evaluation, the LLM baseline reached only 34.7% precision and 34.5% recall at the synset level. In contrast, ExpandNet achieved 75.6% precision and 38.4% recall, more than doubling precision while also improving recall.

In the semeval 2013 dataset, there are 1644 sense annotations for tokens and 773 types of sense annotations. In the dataset created by the LLM baseline, there are 1644 sense annotations for tokens and 773 types of sense annotations. And in the dataset created by the ExpandNet, there are 434 sense annotations for tokens and 394 types of sense annotations. For the projected sense annotations, there are 840 were verified by ExpandNet and 626 were refused. The pipeline can be parallelized and run efficiently on GPUs, making scaling to larger datasets feasible.

2 Baseline

We implemented an LLM-based baseline that generates one Chinese lemma for each gold annotation English synset. Glosses were retrieved from Princeton WordNet via NLTK(Bird et al., 2009) and passed directly to the model, ensuring consistent dictionary-style definitions across synsets.

We used the open-mistral-7b model(Jiang et al., 2023) accessed through its API, chosen for its fast inference and stable short-text outputs. Its lightweight structure made it practical for processing the full synset inventory.

Our prompt was: “You are a bilingual lexicon expert. Given a dictionary definition, produce the single word in Chinese that best matches this definition. Provide only the Chinese word without explanations!” This design encourages concise lexical outputs and avoids descriptive or multi-word answers.

We generated exactly one lemma per synset to match the gold lexicon format and to avoid instability in the model’ s output. When asked to produce multiple lemmas, the LLM often failed to follow the expected format, making it difficult to reliably separate the results with regular expressions. Restricting the model to a single token therefore ensured predictable output and simplified downstream processing.

Several errors illustrate the limitations of the LLM baseline. For the synset loss.n.01 (bn:00052050n, gloss: ’something that is lost’), the model produced the non-existent form ’shímiǎo’, likely due to the abstractness of the gloss and hallucination arising from the small model size. In another case, for brazil.n.01 (bn:00012786n, gloss: ’the largest Latin American country and the largest Portuguese speaking country in the world’), the model output ’Brasil’ instead of the correct Chinese lemma ’ 巴西’. This occurred because the concept is a proper noun, and the model tends to preserve foreign-language forms—especially Portuguese variants—when it cannot reliably map named entities into Chinese.

3 Method

For the translation stage, we continued using the same translation system as in Assignment 2,

namely the MyMemory(Trombetti, 2009) translator. In our previous work, MyMemory produced reasonably accurate and stable translations on the SemEval data, so reusing it here allowed us to keep the translation behaviour consistent across assignments. Since ExpandNet depends heavily on the quality of the initial English-to-Chinese translations, retaining a previously tested translator reduced variability and simplified debugging.

For alignment, we used our cleaned CEDICT-based dictionary rather than the raw CEDICT(Denisowski, 2023) resource employed in Assignment 2. In Assignment 2, we implicitly treated English glosses as if they were single English words, which we later realized was problematic. In this assignment, we first normalized the CEDICT glosses into standardized single-word English tokens and then paired them with their corresponding Chinese entries, producing a cleaner and more lexically coherent dictionary. For example, in the CEDICT entry 'P 民 P 民 [P min2] /(slang) shitizen/commoner/hoi polloi/' we take the simplified form P 民 as the Chinese value. The gloss string is split by '/' into segments such as '(slang) shitizen', 'commoner', and 'hoi polloi'. We remove any bracketed material like (slang), then discard non-lexical or multiword expressions such as 'hoi polloi'. The remaining glosses, 'shitizen' and 'commoner', are normalized into single English tokens and each is used as a dictionary key mapped to the value 'P 民'. Using this cleaned dictionary for alignment helped reduce noise from multi-word or heavily annotated glosses and provided more reliable lexical correspondences for dbalign(Sabet et al., 2020).

For the filtering stage, we used the same cleaned dictionary as in the alignment step. Because this dictionary was constructed by applying strict normalization rules to the English glosses and mapping them to Chinese words, it allowed us to retain only well-formed and semantically plausible translation pairs during filtering. Reusing the same resource for both alignment and filtering ensured internal consistency and helped remove spurious or noisy projections, thereby improving the accuracy of the projected Chinese senses.

4 Analysis

Of the senses annotated manually in 10 sentences selected, 30 (different annotations with the same BabelNet ID are counted individually) of them were correct and 14 of them could have been correct in a different context. Among the errors we found, there was an alignment error where "reductions" (instance id semeval2013.d000.s001.t008) was aligned with "减排" (translated as "reduce emissions" in English), which is clearly incorrect and should be aligned with "减". Similarly, aligning "global warming" (instance id semeval2013.d000.s004.t005) with " 变暖" (which translates to "warming" in English) is also incorrect; it should be aligned with " 全球变暖". In addition, there are translation errors. Translating "group" (which should be translated as "小组" in the context) as "集团" (which translates to "organization" in English) is not entirely accurate, although "group" can indeed be translated as "集团" in other contexts. To address the common recurring alignment mistakes mentioned above, errors typically occur during tokenization in the alignment process. We should adopt a better tokenizer to resolve this issue. For instance, using a tokenizer capable of splitting the previously mentioned "减排" into '减' and "排" tokens might help eliminate some alignment errors.

Compared to ExpandNet, some results from LLM may be slightly better （ExpandNet generally produces better results）but are not entirely accurate, such as projecting the senses of "group" to " 团体" and "reduction" to " 减小". Based on the result of automatic evaluation, ExpandNet is much more precise and reliable, while the LLM generates more translations but sacrifices accuracy. For building a quality bilingual lexicon, ExpandNet's high precision makes it clearly superior despite lower coverage.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Paul Denisowski. 2023. Cedict: Chinese–english dictionary. https://www.mdbg.net/chinese/dictionary?page=cedict. Accessed 2025.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Albert Q. Jiang et al. 2023. Mistral 7b.

University of Alberta. 2024. Expandnet: Cross-lingual lexical projection system (cmput 497 assignment framework). Course-provided software pipeline.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

Marco Trombetti. 2009. Mymemory: creating the world's largest translation memory. In *Proceedings of Translating and the Computer 31*, London, UK. Aslib.