

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 glosses and a concise bilingual prompt. The main method, ExpandNet, 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.

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 and passed directly to the model, ensuring consistent dictionary-style definitions across synsets.

We used the Mistral-7B model 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 Chinese word that best matches this definition. Provide only the 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 describing the country Brazil), 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 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 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 min2] /(slang) shitizen/commoner/hoi pol-

loi/' 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'. This illustrates how our cleaned dictionary converts noisy glosses into usable single-word English–Chinese pairs for alignment. Using this cleaned dictionary for alignment helped reduce noise from multiword or heavily annotated glosses and provided more reliable lexical correspondences for dbalign.

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