Generating Concept Lexicalizations via Dictionary-Based Cross-Lingual Sense Projection

David Basil, Chirooth Girigowda, Bradley Hauer, Sahir Momin, Ning Shi, Grzegorz Kondrak

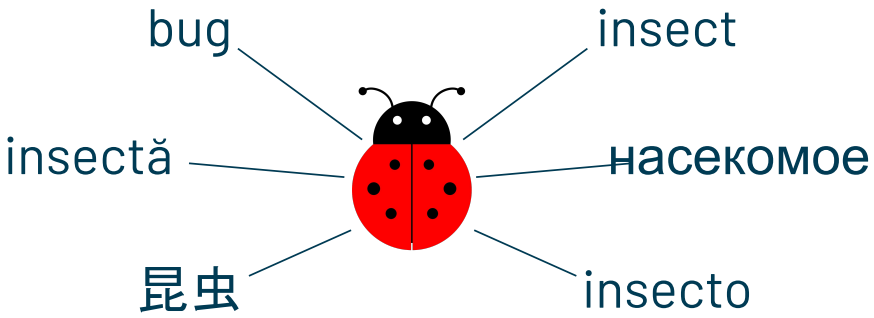
Alberta Machine Intelligence Institute (Amii)
Department of Computing Science
University of Alberta, Edmonton, Canada

amii

Lexical Concepts

A **lexical concept** is a single idea that a word or phrase can express.

A lexical concept is language-independent, and can be expressed by a number of words in a number of languages.



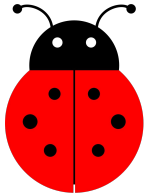
In our formulation, we assume that every content word expresses exactly **one concept**.

Multilingual Wordnets

A **wordnet** is a database where a concept is defined by the set of lemmas that can express that concept.

Wordnets (especially **multilingual wordnets**, which contain lemmas from multiple languages) are very useful.

bug, glitch, error, fault



bug, insect, pest

We assume that an English wordnet is available to us, and seek to expand such a wordnet to other languages.

Sense Generation

Expanding a wordnet involves creating **word senses**.

A sense is a unique meaning of a word.

Wordnets represent a sense as a pairing of a lemma with a concept.

(bug, )

(bug, )

(glitch, )

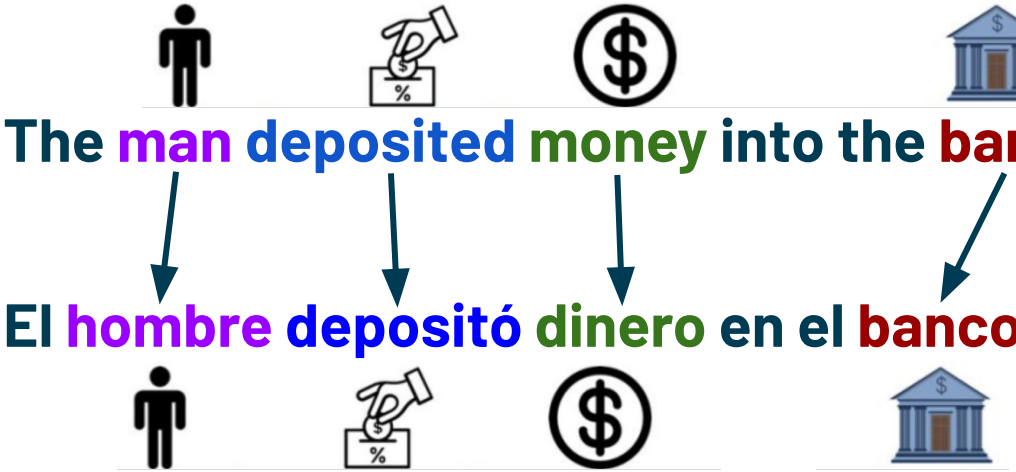
(insect, )

Sense generation therefore means linking existing concepts to new lemmas.

In our case, we want to link these concepts to lemmas in new languages.

Projection Framework

Projection allows information available in some languages to be gleaned in other languages.

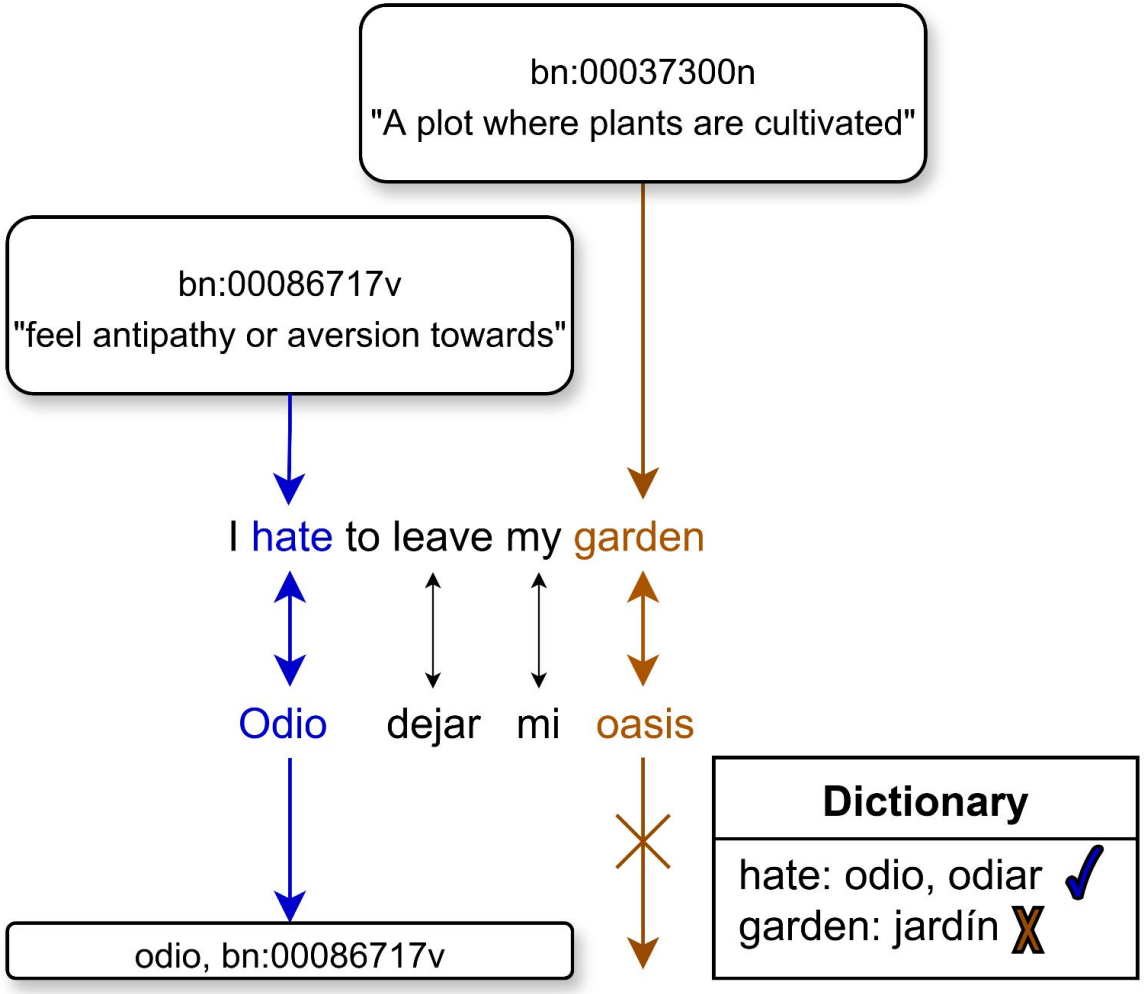


A sentence is translated, and lexical information is **projected** from one language to another via word alignment.

A **concept** can be projected from one word to another to generate a **sense**.

Our Pipeline

1. A sense-annotated English sentence is **translated**.
2. Word **alignments** are found.
3. Alignments not supported by a **bilingual dictionary** are discarded.
4. Sense tags are **projected** across valid alignments, creating senses in the target language.



Projection Filtering

Previous projection-based methods have used corpus-level statistics to decide whether a word alignment is valid for projection.

We use a dictionary instead, which:

1. Allows for performance to be **independent of the size of a corpus**.

2. Ensures all projections are **linguistically justified**.

We assume implicitly that when a word and its translation are **synonyms**, they express the **same concept**.

Dictionary-Based Aligner (DBAlign)

To achieve high-quality alignment, we combine a pre-trained base aligner with our dictionary.

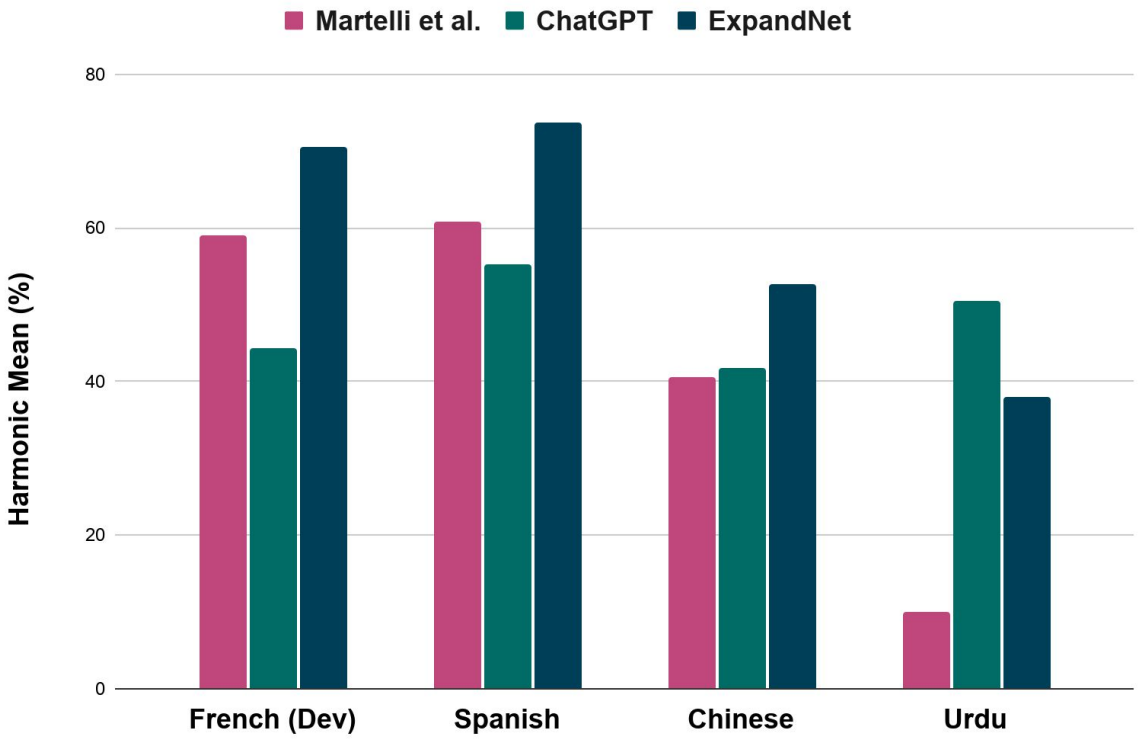
The algorithm follows a three-step priority sequence:

1. **Intersection**: Prioritize links where the base aligner and dictionary agree (**High Precision**).

2. **Dictionary Expansion**: Add dictionary-proposed links that do not conflict with Step 1 (**Lexically Supported**).

3. **Residual Alignment:** Add remaining base aligner links that do not conflict with the existing set (**Maximizing Coverage**).

Results



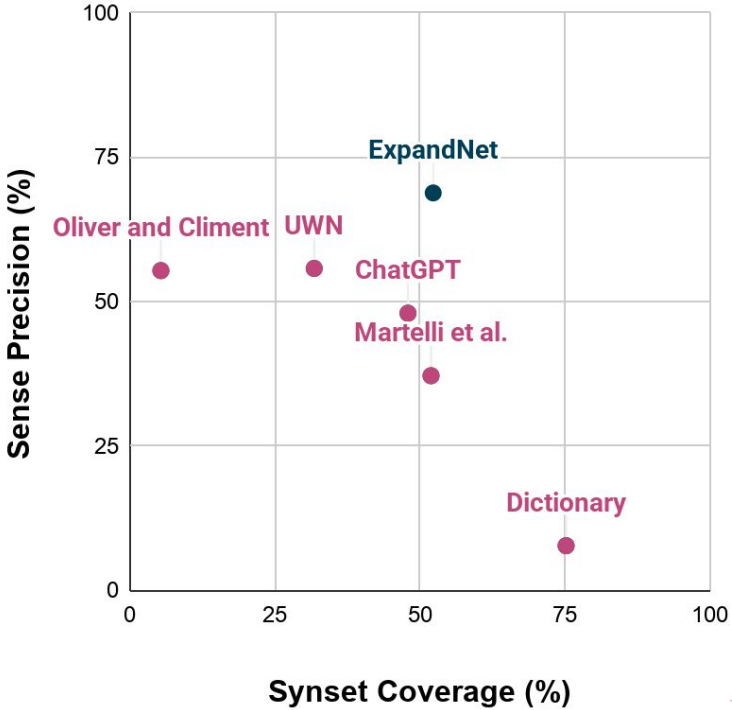
We compare ExpandNet to state-of-the-art projection and LLM models.

Evaluation Metrics:

- **Precision**: % of generated senses that are correct.
- **Coverage**: % of source concepts successfully projected.
- **Harmonic Mean**: The overall balance of Precision and Coverage.

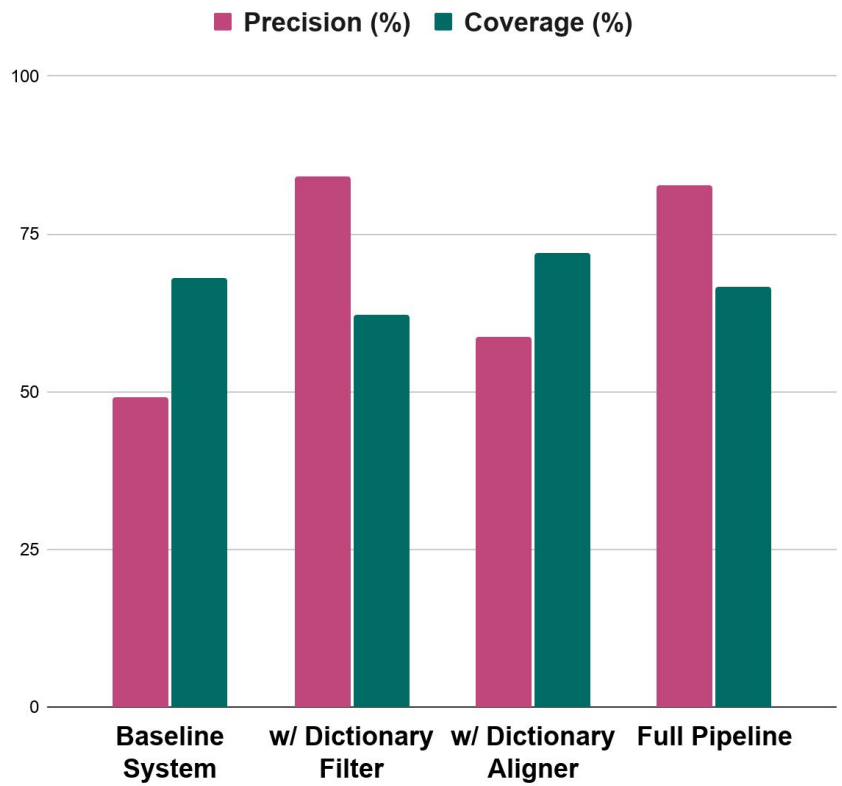
Analysis: Average Precision vs. Coverage Across All Languages

- **The Goal**: Maximize precision without sacrificing coverage.
- **Dictionaries**: High coverage with excessive noise.
- **Prior SOTA**: Moderate coverage, moderate precision.
- **ExpandNet**: Matches SOTA coverage while boosting precision significantly.



Ablation

- **Baseline:** Moderate coverage, poor precision
- **+ Filter Only:** Pronounced boost to precision, hurts coverage
- **+ Aligner Only:** Boosts coverage, precision remains low
- **Full Pipeline:** Results in the best overall performance



Conclusion

We have:

- **Introduced** ExpandNet, a sense generation system that uses a bilingual dictionary to generate and filter alignments.
- **Found** that the dictionary filter improves precision without relying on corpus-level statistics.
- **Revealed** that the dictionary-augmented word aligner successfully drives coverage.
- **Demonstrated** overall results that improve, on balance, over previous state-of-the-art methods and LLM prompting.

github.com/UAlberta-NLP/ExpandNet