

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

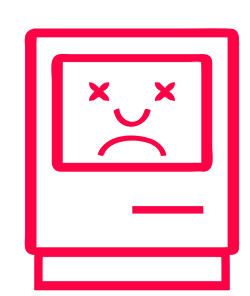
David Basil, Chirooth Girigowda, Bradley Hauer, Sahir Momin, Ning Shi, Grzegorz Kondrak

{dbasil1, gkondrak}@ualberta.ca

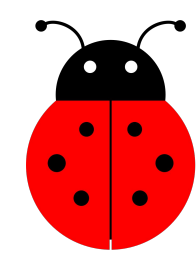
Introduction

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

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



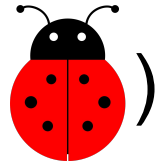
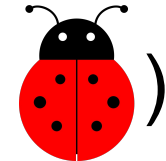
bug,
glitch,
error,
fault



bug,
insect,
pest

A **sense** is a unique meaning of a word.

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

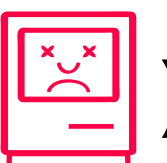
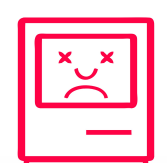
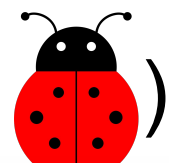
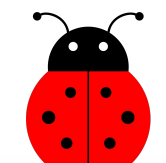
(glitch, ) (bug, ) (insect, ) (bug, )

Task

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

Wordnets can be expanded to additional languages by **generating senses** in the target language.

This can be done by associating **existing concepts** with **target-language lemmas**.

(fehler, ) (bogue, ) (insectă, ) (昆虫, )

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.
2. **Dictionary Expansion**: Add dictionary-proposed links that do not conflict with Step 1.
3. **Residual Alignment**: Add remaining base aligner links that do not conflict with the existing set.

Projection Framework

Projection allows information available in a certain language to be **projected** into other, usually lower-resource, languages.

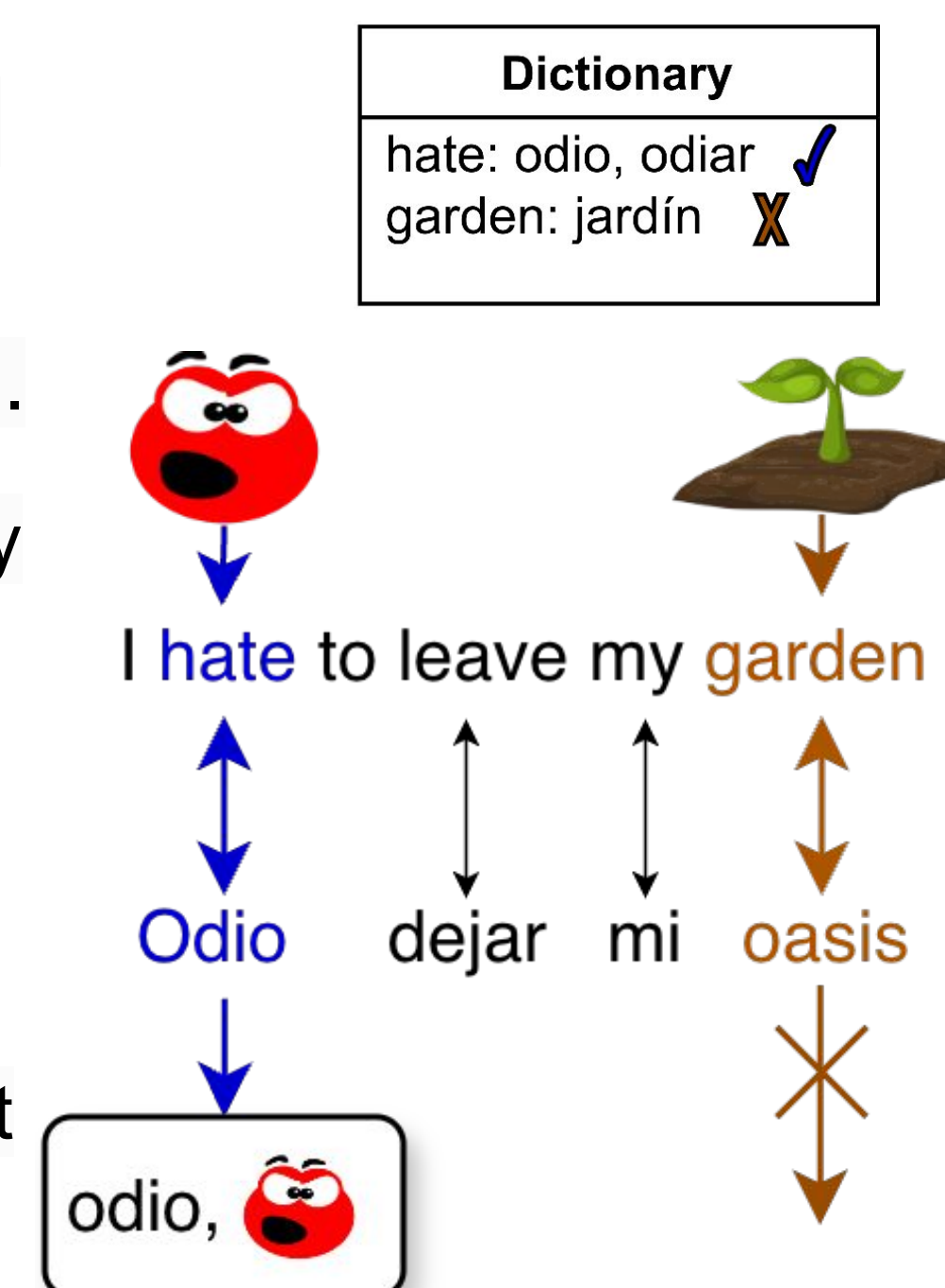
We project **concept** tags to generate target-language senses.

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.

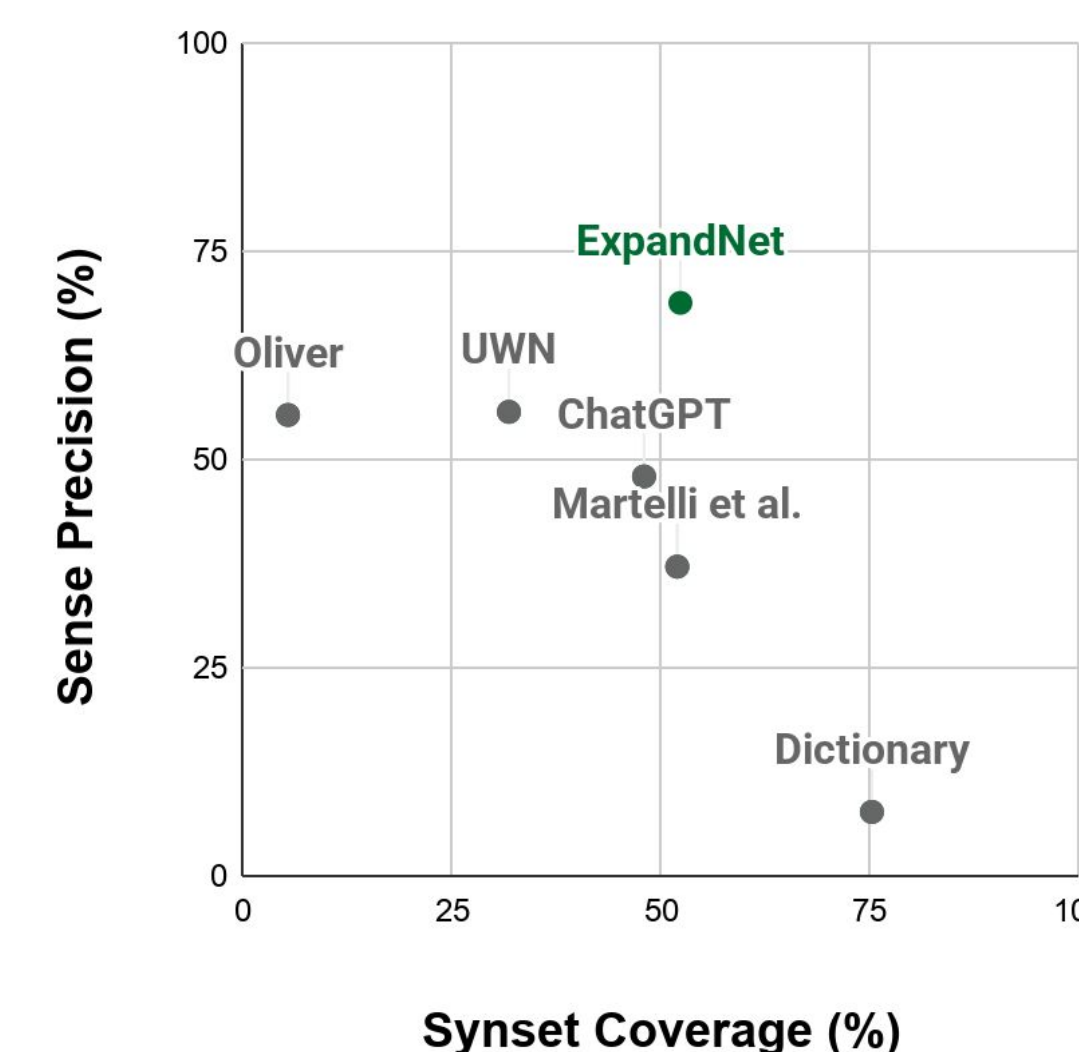


Results

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

Evaluation Metrics:

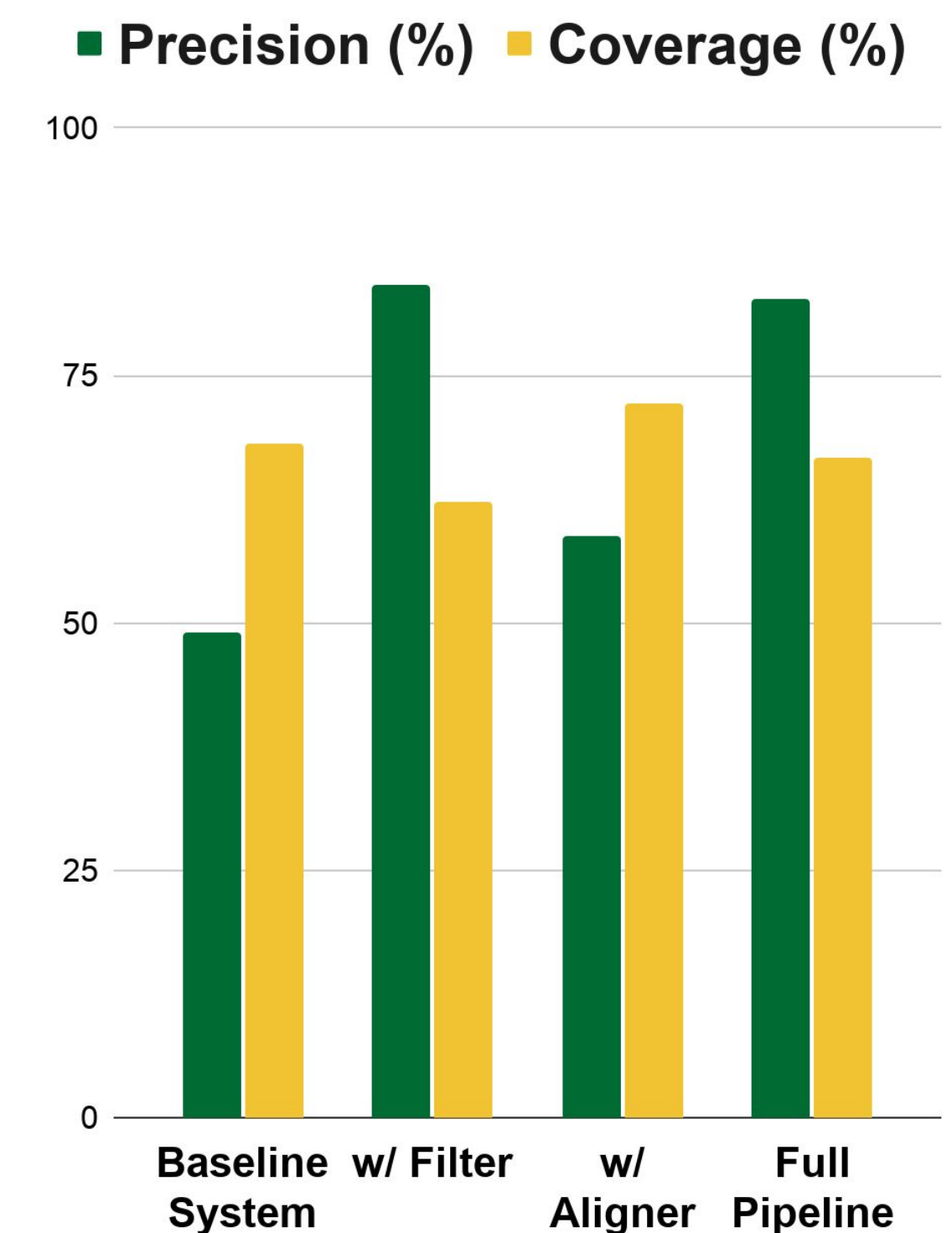
- Precision: % of generated senses that are correct.
- Coverage: % of source concepts successfully projected.
- Harmonic Mean: The overall balance of Precision and Coverage.



This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Alberta Machine Intelligence Institute (Amii).

Ablation

- **Baseline**: Moderate coverage, poor precision.
- **+ Dictionary Filter Only**: Pronounced boost to precision, hurts coverage.
- **+ Dictionary Aligner Only**: Boosts coverage, precision remains low.
- **Full Pipeline**: Results in the best overall performance (maximizes the harmonic mean of precision and coverage).



Conclusion

- **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.

Canadian AI 2026
25-29 May 2026, Vancouver, BC



AI/CRV
2026