Creation of a Pseudodictionary to Minimize the Ambiguity

Luca Barra

Department of Computer Science, University of Turin, Turin, Italy
luca.barra@edu.unito.it

Abstract. This paper investigates the construction of cross-lingual pseudodictionaries designed to minimize lexical ambiguity through optimal translation pair selection. Building on the observation that historical, cultural, and geographic factors create varying degrees of sense alignment between languages, I propose a method that systematically selects word pairs (L_1-L_k) maximizing the Ambiguity Reduction Score (ARS). My approach leverages these naturally occurring asymmetries in semantic overlap to create a resource that preferentially includes translation pairs demonstrating the strongest sense alignment. The resulting pseudodictionary provides empirically optimized word mappings that significantly reduce polysemy. It has applications in machine translation and word sense disambiguation.

Keywords: NLP · Languages · Semantic.

1 Introduction

1.1 The Key Idea

Consider a word $x \in L_1$ with N senses and its translation $y \in L_2$ with M senses (where $M \leq N$). We can bind x to y to create a new lexical unit called *pseudoword*, denoted $x-y$. This new unit $x-y$ inherits only the common senses between the two starting word resulting in a word that is less ambiguous than either x or y individually. Formally:

- Let $\mathcal{S}_x$ be the set of senses of x in L_1
- Let $\mathcal{S}_y$ be the set of senses of y in L_2
- The senses of the pseudoword $x \cdot y$ are $\mathcal{S}_{x \cdot y} = \mathcal{S}_x \cap \mathcal{S}_y$

$$|\mathcal{S}_{x-y}| \leq \min(|\mathcal{S}_x|, |\mathcal{S}_y|)$$

In particular, it frequently holds that:

$$|\mathcal{S}_{x \cdot y}| \ll |\mathcal{S}_x| \quad \text{and} \quad |\mathcal{S}_{x \cdot y}| \ll |\mathcal{S}_y|$$

1.2 Ambiguity Reduction Score (ARS)

A simple approach to measure the reduction of ambiguity of a pseudoword $x - y$ is to use a score like:

$$\text{AmbiguityReduction}(x, y) = \frac{|\mathcal{S}_x| + |\mathcal{S}_y| - 2 \cdot |\mathcal{S}_x \cap \mathcal{S}_y|}{|\mathcal{S}_x| + |\mathcal{S}_y|}$$

where $|\mathcal{S}_x|$ is the set of sense in x, $|\mathcal{S}_y|$ is the set of sense in y and $|\mathcal{S}_{x-y}$ is the set of common sense between x and y. Taking a step further we can generalize the formula to take N languages:

$$\text{AmbiguityReduction} = \frac{\sum_{i=1}^{N} |\mathcal{S}_i| - N \cdot \left|\bigcap_{i=1}^{N} \mathcal{S}_i\right|}{\sum_{i=1}^{N} |\mathcal{S}_i|}$$

where $|\bigcap_{i=1}^{N} \mathcal{S}_i| \cdot N$ are the sense shared in all the languages and $\sum_{i=1}^{N} |\mathcal{S}_i|$ are the total number of sense across all the languages.

This score ranges between 0 and 1:

- A score of 0 ($N \cdot \left|\bigcap_{i=1}^{N} \mathcal{S}_i\right| \to 1$) indicates no reduction in ambiguity, all languages share exactly the same set of senses.
- A score close to 1 ($N \cdot \left|\bigcap_{i=1}^{N} \mathcal{S}_i\right| \to 0$) indicates maximum disambiguation, there is little or no overlap between the senses of words across languages.

2 The Pseudodictionary

2.1 The Ambiguity between Different Languages

The foundational hypothesis of this work posits that lexical compatibility between languages operates at the word-level rather than being a language-wide property. This manifests as significant variance in ARS potential between specific word pairs across different languages:

- The word form "Fiore" (IT) and the word form "Blume" (DE) have an ARS greater than "Fiore" and "Flower" (EN).
- The word form "Carne" (IT) and the word form "Meat" (EN) have an ARS Score greater than "Carne" and "Fleish"(DE).

We can use this fact to construct a pseudodictionary that contains only pairs of words with the maximum ARS. This can be done with the following passage:

- Let $w \in L_1$.
- For each language $L_j \in L_2, \ldots, L_n$:
  - Retrieve all potential translations $T_j = t_{j1}, \ldots, t_{jk}$ in L_j.
  - For each translation candidate $t_{ji} \in T_j$:
    - Extract all senses from w and t_{ji}.
    - Compute ARS for each pair.
    - Record the maximal ARS for this word pair.
- Identify the optimal pair (w, t_{ji}) that achieves the highest ARS across all languages.

2.2 How to Choose a Translation

A key challenge in this approach is selecting optimal translations. A potential approach is to:

- Evaluate all possible translations of x across target languages.
- Select the pairing that maximizes the ARS (tab. 1).

Table 1. Example with some translation (EN-IT). The pseudoword Room-Camera is better than the pseudoword Room-Chamber.

Word L_1	Translation L_2	Common Senses	Ambiguity Reduction Score
Room	Camera	Room (Architecture) Room (Novel) The Room (2003) The Room (2015)	0.500
Chamber	Camera	Chamber (firearms) Bedroom Room Legislative chamber Chambers (law) Chamber (Architecture) The Chamber (Song) The Chamber (1996)	0.176

3 Result

3.1 Limitations

A significant limitation of this approach stems from the scarcity of comprehensive multilingual semantic resources. Even when using BabelNet, one of the most extensive multilingual encyclopedias, several challenges persist:

- **Asymmetrical sense alignment**: Semantic mappings often work in only one direction (e.g., Italian → English but not English → Italian).
- **Incomplete coverage**: Many words lack proper cross-lingual sense mappings.
- **Under-representation**: the languages less used have little representation.

4 Future Prospect

References

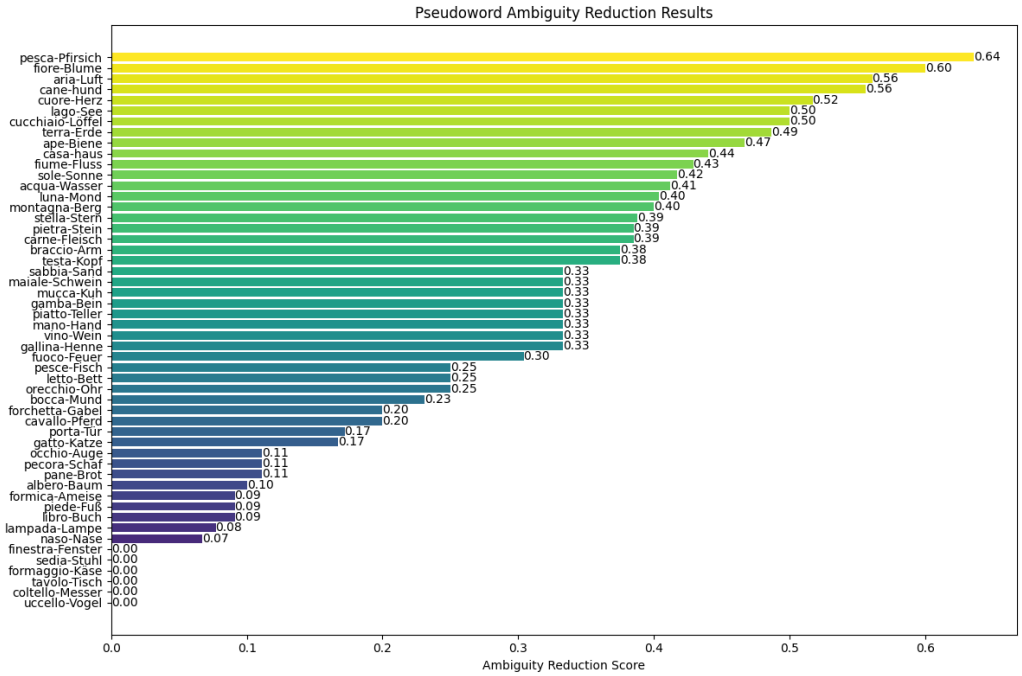

Fig. 1. Ambiguity Reduction Score for IT-DE.

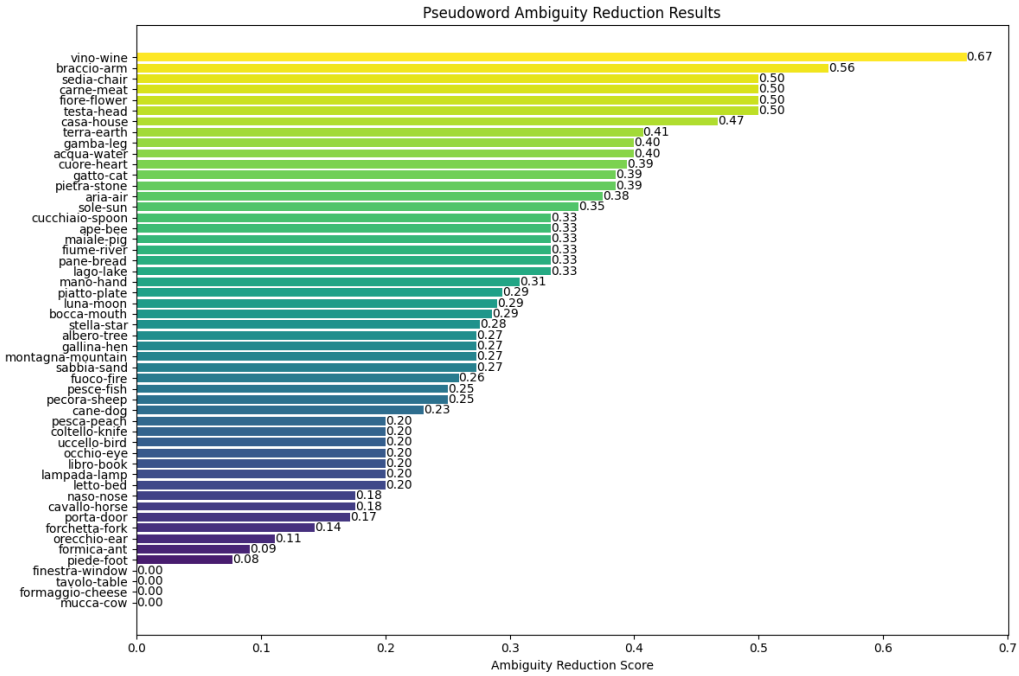

Fig. 2. Ambiguity Reduction Score for IT-EN.