Semi-Automated Construction of Sense-Annotated Datasets for Practically any Language

Jai Riley, Bradley Hauer, Nafisa Sadaf Hriti, Guoqing Luo, Amirreza Mirzaei, Ali Rafiei, Hadi Sheikhi, Mahvash Sivashpour, Mohammad Tavakoli, Ning Shi, Grzegorz Kondrak

Introduction

Word **S**ense **D**isambiguation **(WSD)**: task of identifying the meaning of a word in context.

The man deposited money into the ***bank.***

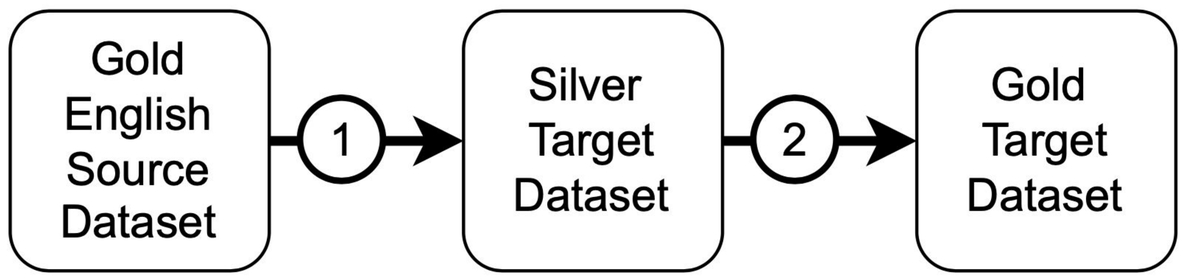
High-quality (Gold) sense-annotated datasets are essential for evaluating WSD systems but are often ***restricted to high-resource languages or limited in scope.***

Outline

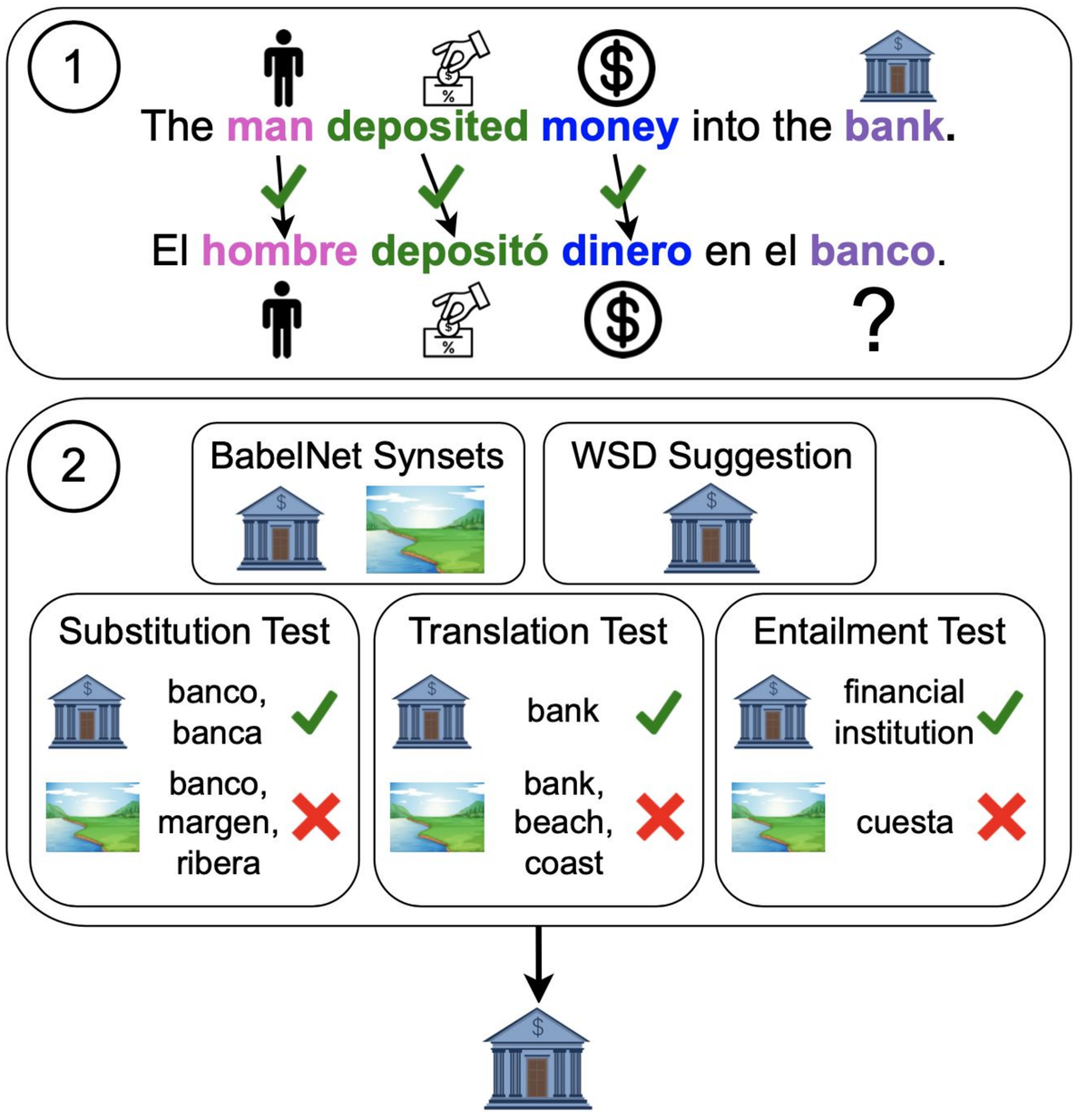
Our ***language-agnostic pipeline*** for creating sense-annotated datasets has two parts**:**

(1) Automatic Silver Dataset Creation: translation, alignment, and filtering

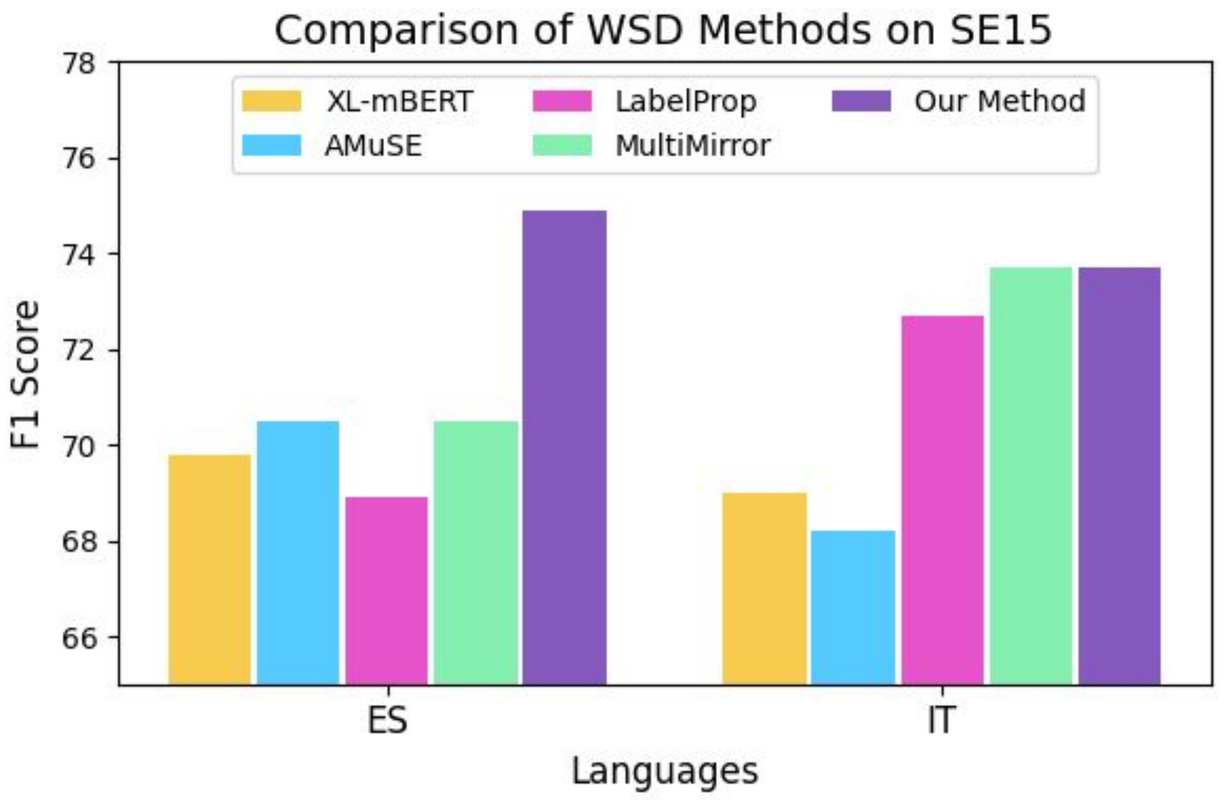
(2) Manual Gold Annotation Procedure: WSD system suggestions, BabelNet synsets, and three tests



From Silver to Gold

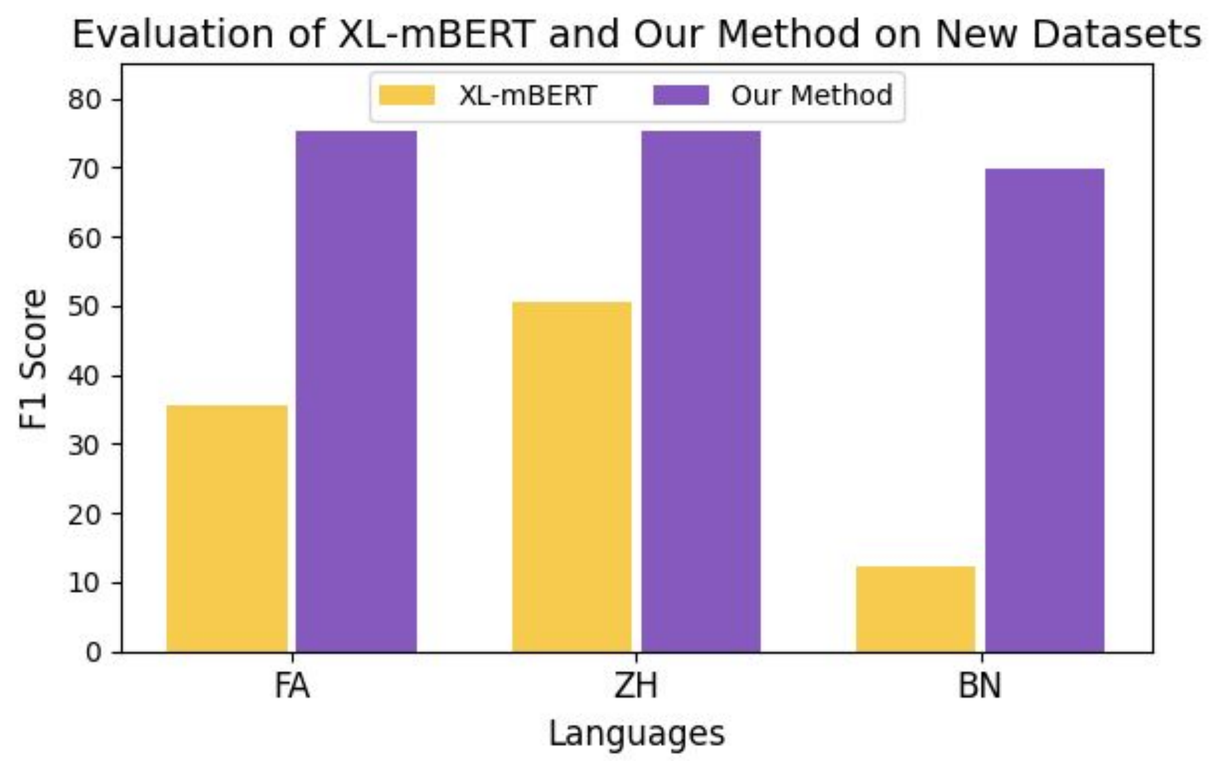


Results: IT and ES



*Our Method refers to part (1) of our pipeline using provided or verified gold translations

Results: FA, ZH, and BN



Summary

- Introduced a novel approach for automatically creating sense-annotated data for ***any language,***
- Designed an efficient annotation procedure that ***accelerates manual sense validation.***
- Created ***new parallel WSD datasets for Farsi, Chinese, and Bengali,*** verified by native speakers.
- Performed ***empirical validation*** of our method on both new and existing gold datasets, showing our method has ***competitive or superior performance*** on all datasets.

github.com/jai-riley/Sense-Projection

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii). @COLING 2025