Bridging The Gap Between BabelNet and HowNet: Unsupervised Sense Alignment and Sememe Prediction

Xiang Zhang, Ning Shi, Bradley Hauer, Grzegorz Kondrak

{xzhang23, ning.shi, bmhauer, gkondrak} @ ualberta.ca

Introduction

Sememes play an important role in many NLP applications. Compared to glosses, sememes are simply a set of words and can be easily encoded using word embedding and parsed by a computer. Therefore its application spans a wide range of NLP subfields.

However, their incorporation into natural language processing has been limited by a lack of available sememe resources for commonly used sense inventories, the only large-scale sememe KB HowNet is limited to only two languages: English and Chinese. Previous work on automatically predicting sememes for BabelNet concepts has depended on large human-labeled data.

- Rather than attempting to predict sememes directly, as in prior work, we instead attempt to align BabelNet concepts and HowNet senses.
- Our method has four stages, each relaxing some constraint until an alignment is found, as shown in the graph.


(a)   (b)
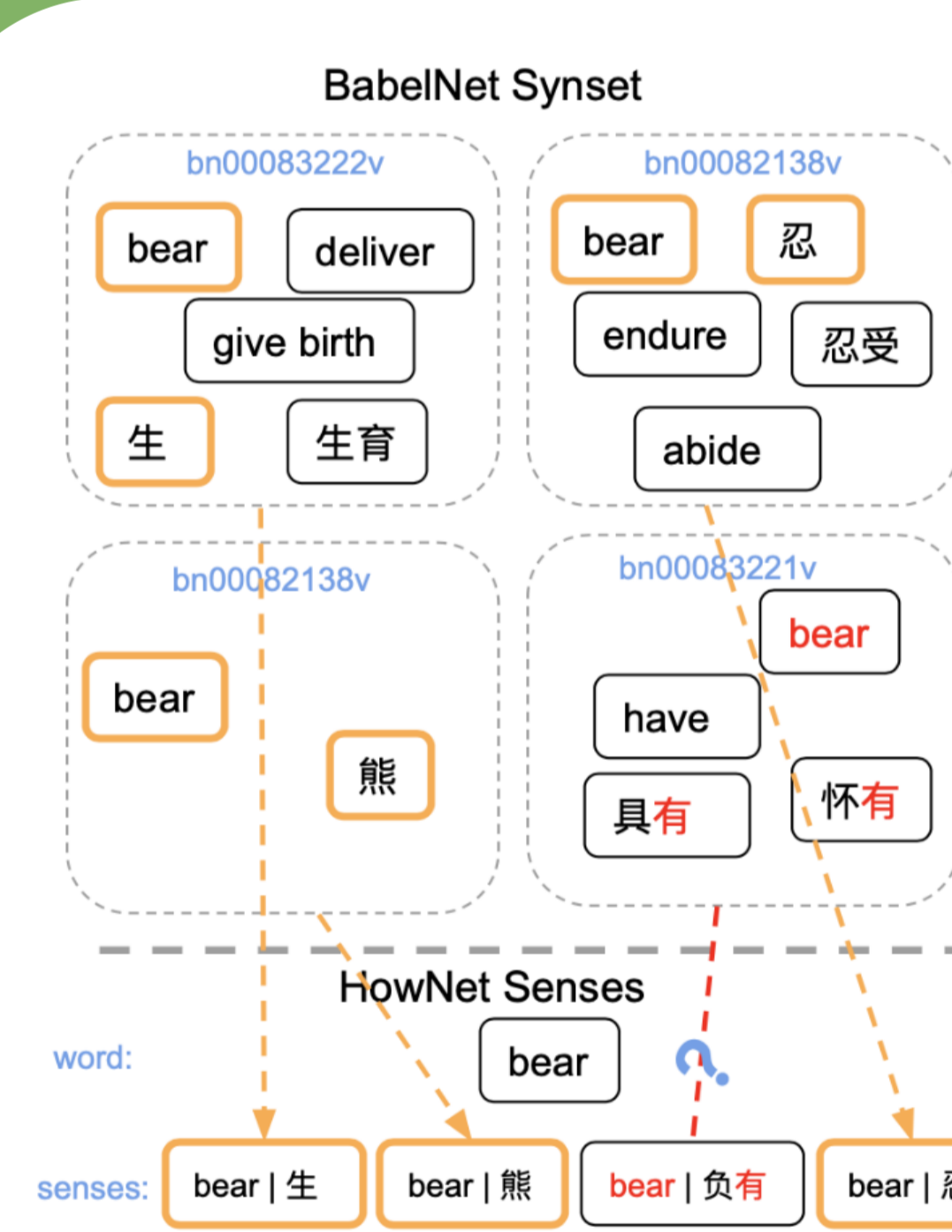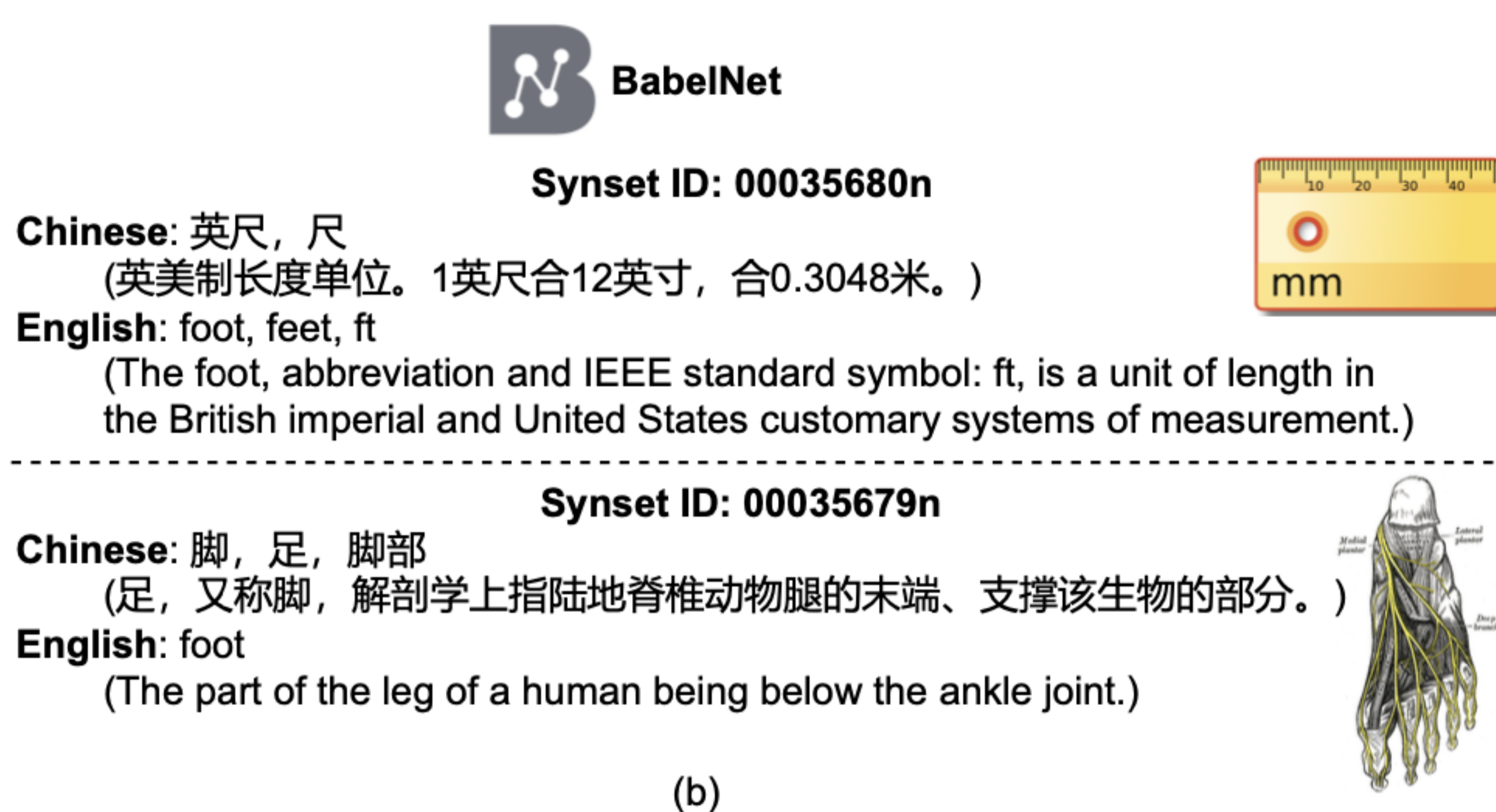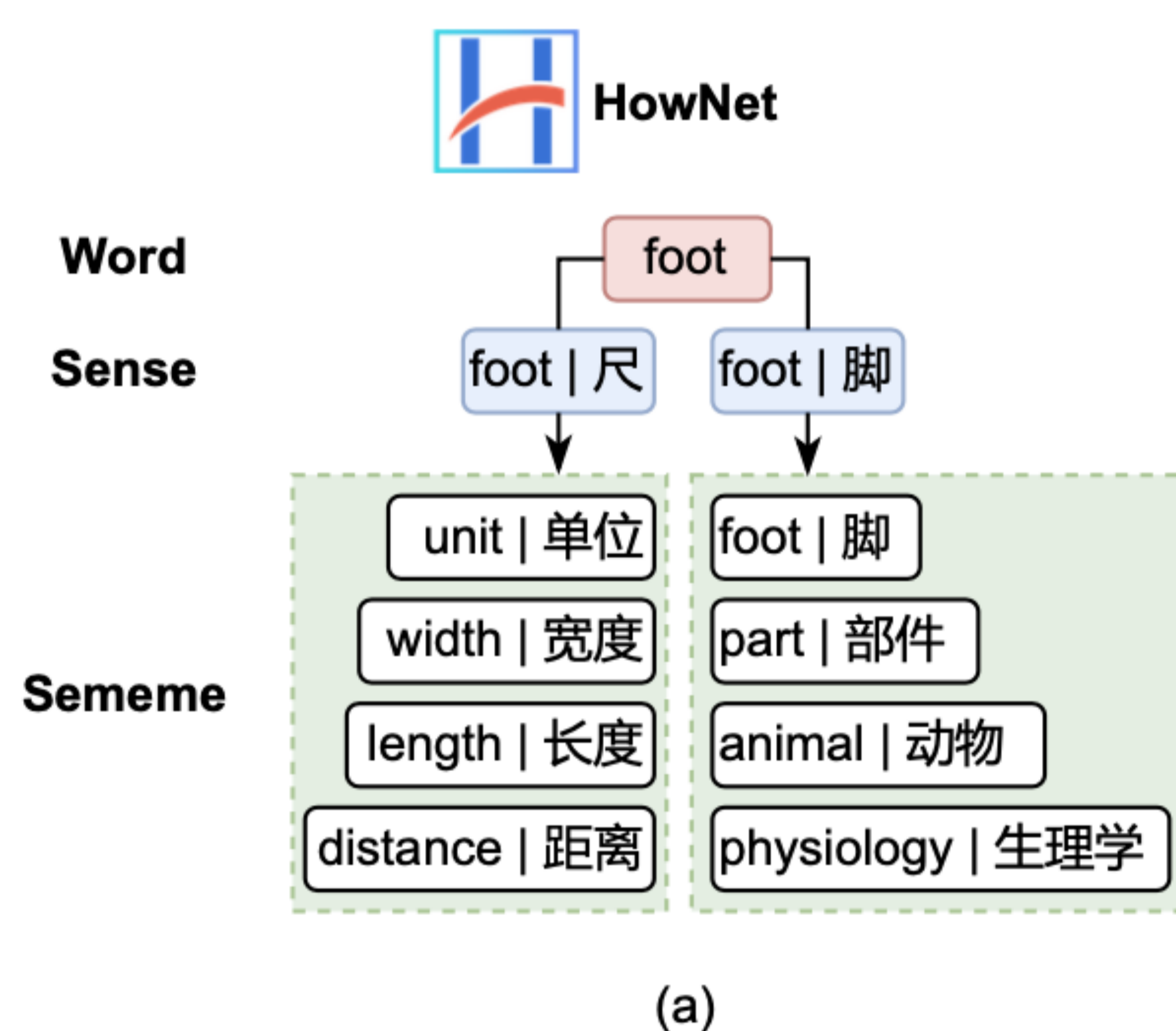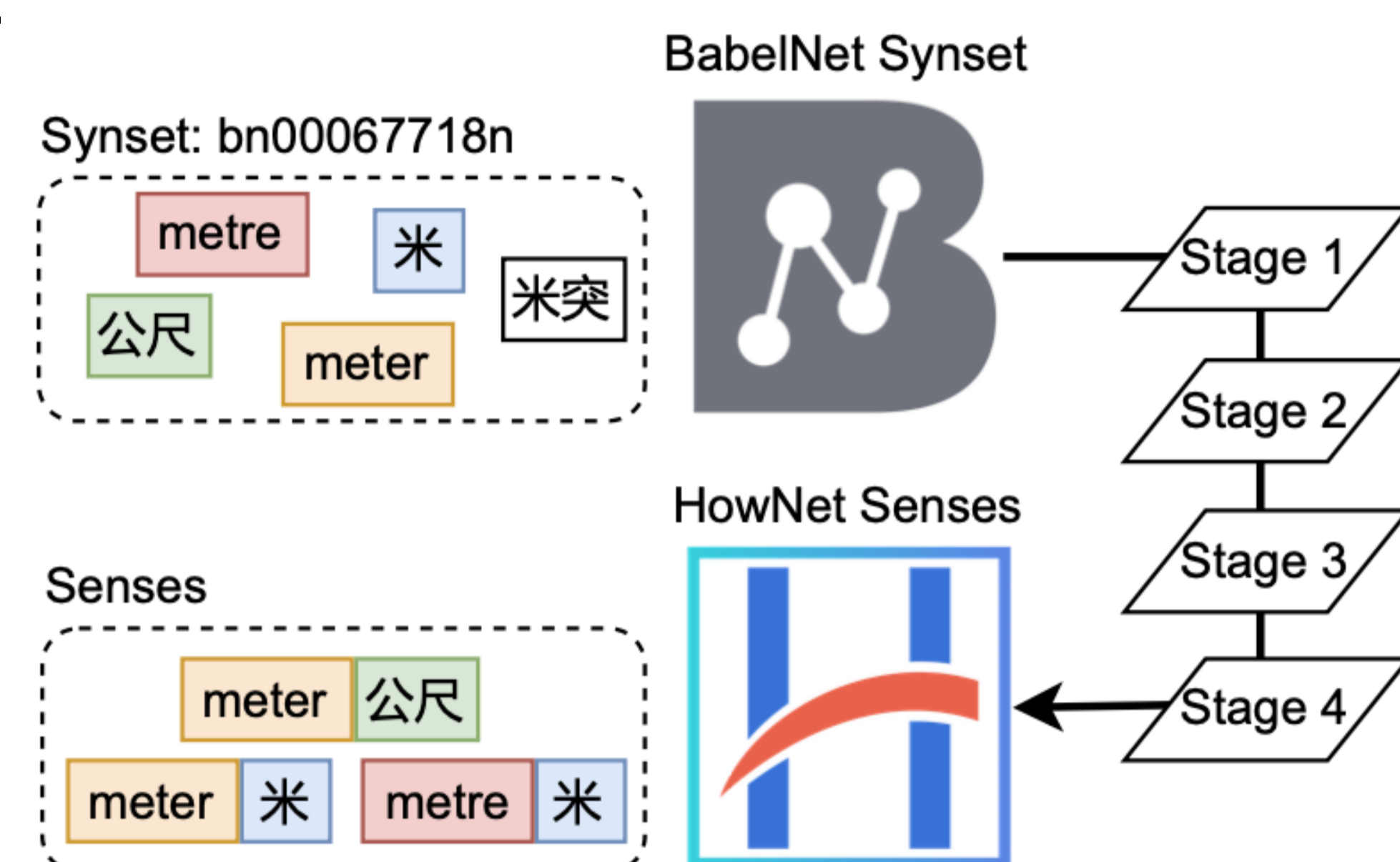
Stage one - Exact Match: Based on the well-known observation that distinct senses of a word may translate differently. Aim at high precision rather than high coverage.

Stage two - Partial Match: A less strict match than exact match. Aims at improving coverage rate. Based on the fact that Chinese words of similar meanings share same characters
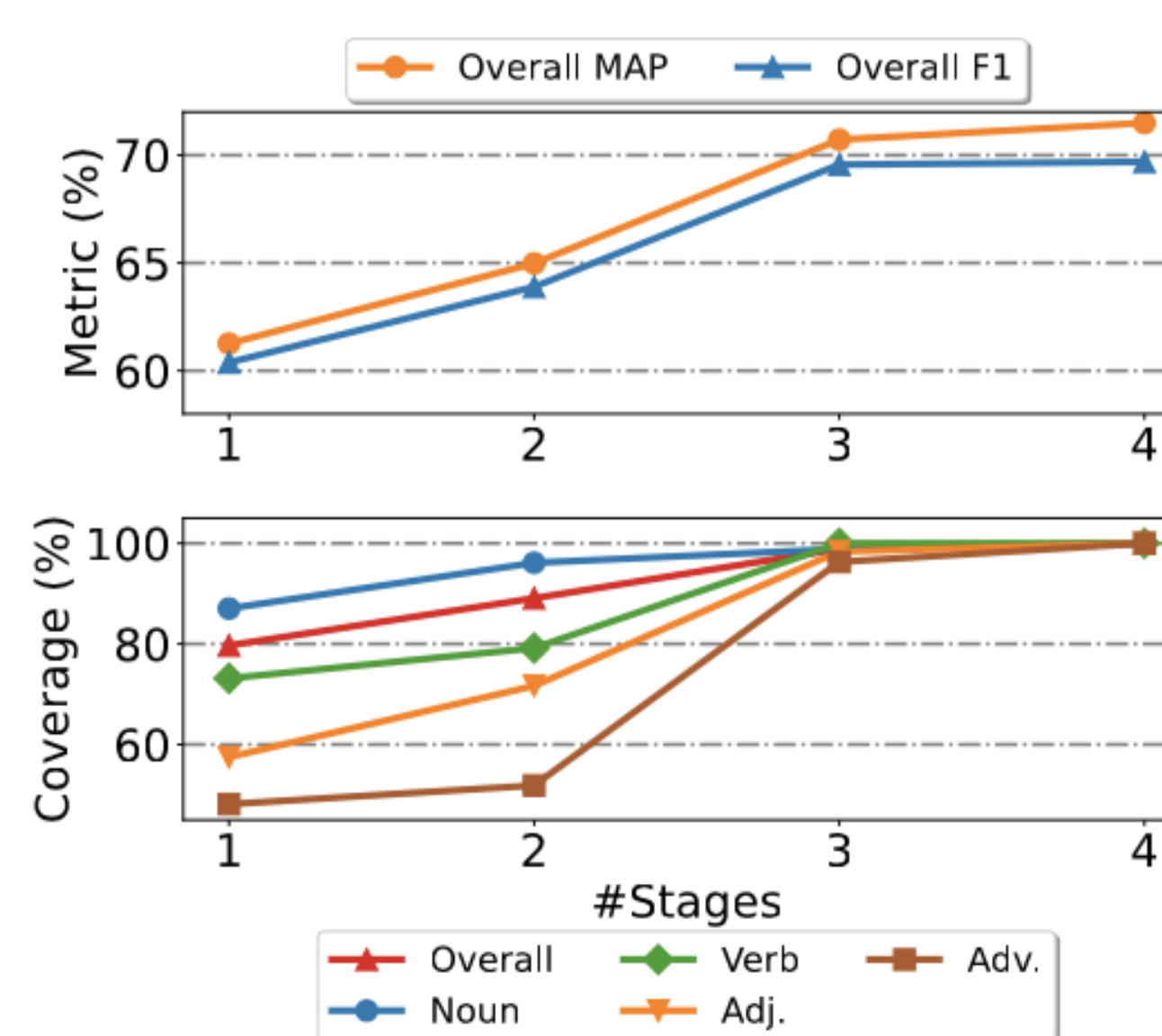
Stage Three and Four - Sense information Match and ProperName: Using synonym information and sememe information to match the rest of the synsets, The non-matched senses are then ProperName

Main Result
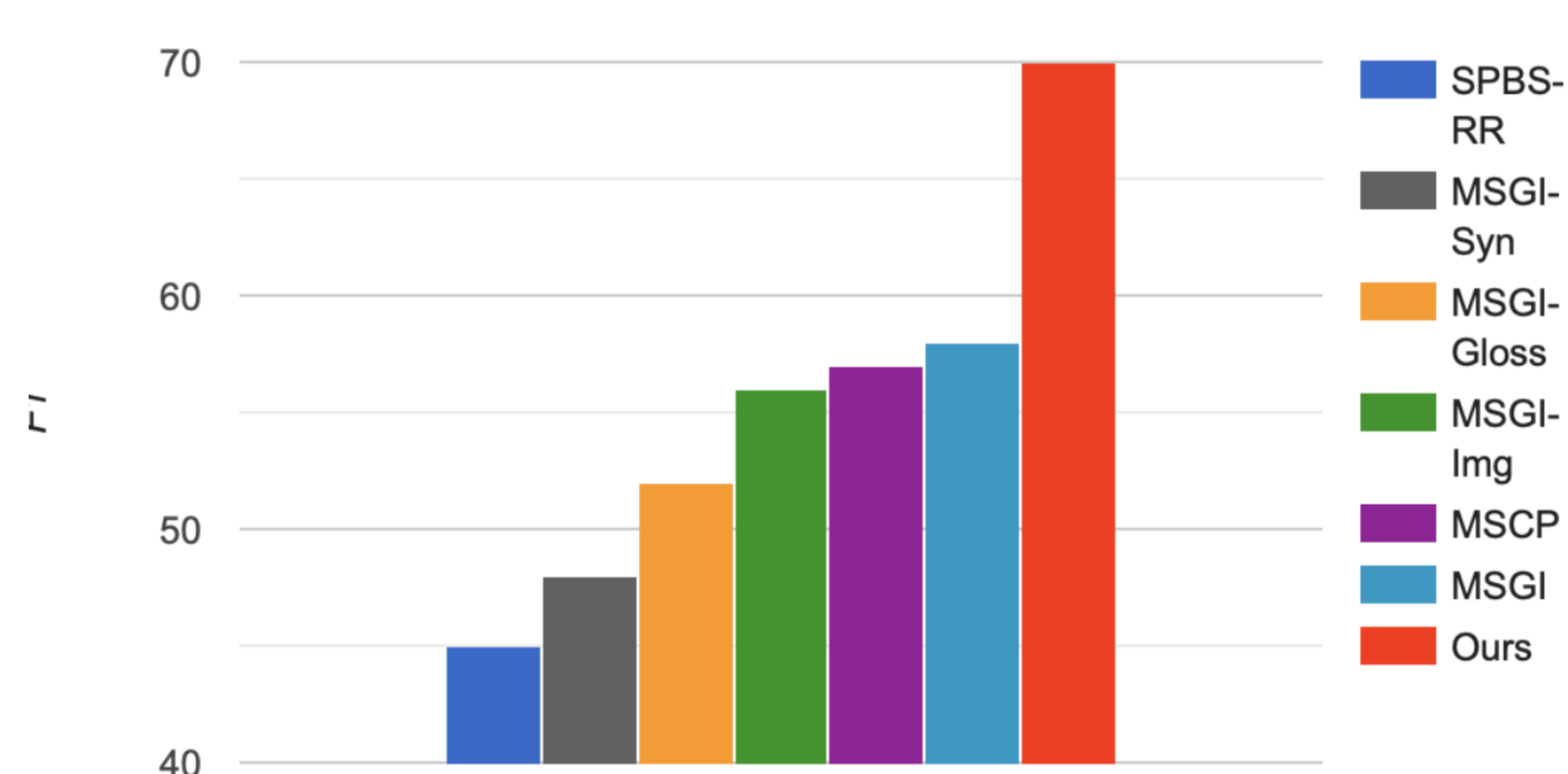
We achieve the state-of-the-art result on sememe prediction, outperforming all previous supervised methods

Performance over Stages

- Stage 1 is sufficient to cover roughly 80% of synsets, and achieve roughly 60% F1.

- Adding stages 2 and 3 greatly increases the coverage of adjectives and adverbs.

- Stage 4 is shown to provide marginal improvement