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

Dept. Computing Science, University of Alberta
Alberta Machine Intelligence Institute (Amii)

UNIVERSITY OF ALBERTA

amii

Lexical Knowledge Bases: BabelNet

- Sense Inventory: List of senses for each word

  Most popular knowledge bases (KBs): WordNet (English), BabelNet (Multilingual)

  WordNet and BabelNet group senses into synonym sets (*synsets*)

- Example of sense definitions in BabelNet:

Bn:00008363n

EN **bank**

ZH 岸 • 河边 • 河岸

gloss: Sloping land (especially the slope beside water).

Bn:00008364n

EN **bank**

ZH 银行 • 存放款金融机构 • 銀行 • 银行业

gloss: A financial institution that accepts deposits and channels the money into lending activities

Lexical Knowledge Bases: HowNet

- Problems with BabelNet in Chinese:

  Poor coverage for Chinese Words

  Parsing glosses can be difficult for machines

- HowNet:

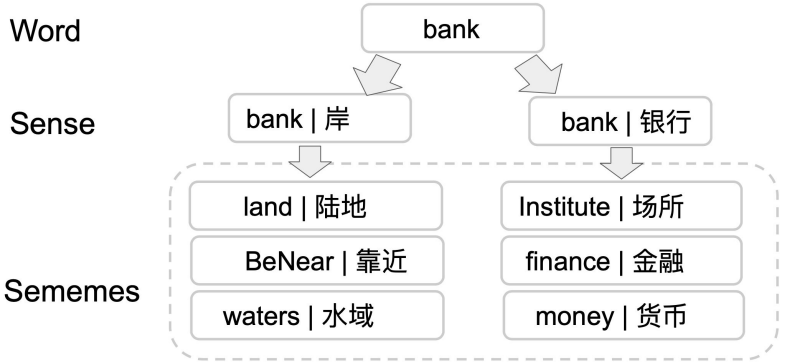
  A sememe based sense inventory

  No synsets

  Represents each word sense by a set of sememes

  Better coverage for Chinese words

  Each sense is associated with one English word and one Chinese word



The Importance of Sememes

- Easy to encode with many applications:

  Word Sense Disambiguation (WSD), Representation Learning, Language Modelling, Text Matching, to name a few.

- Supported by linguistic primitives hypothesis

However:

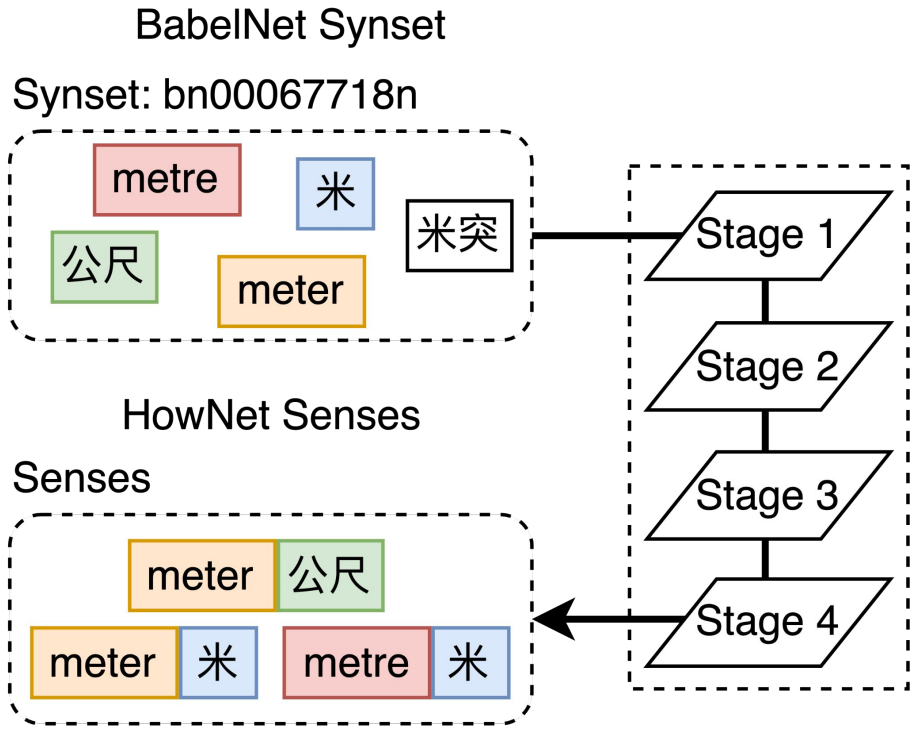
- KBs such as BabelNet does not have sememes, limiting the potential usage of sememes on downstream applications
- Prediction based methods of adding sememes perform poorly and lack explainability

Our Method: Knowledge Base Alignment

Input : A BabelNet synset

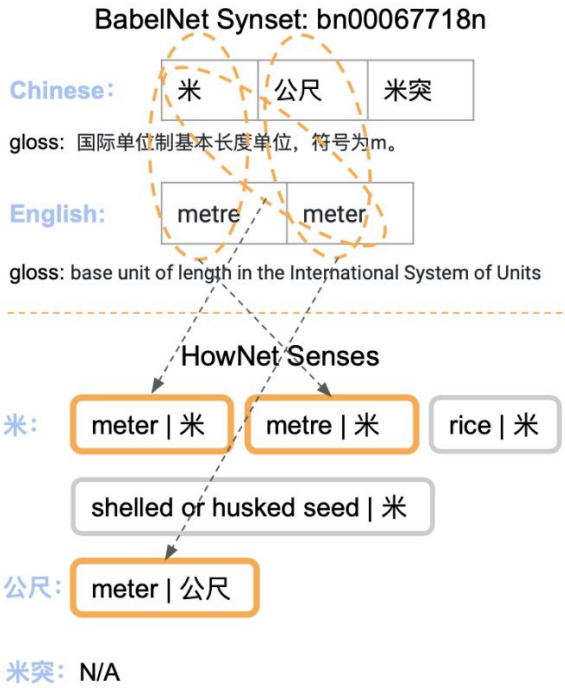
Output: HowNet senses

To align the synset to the HowNet senses representing the same meaning.



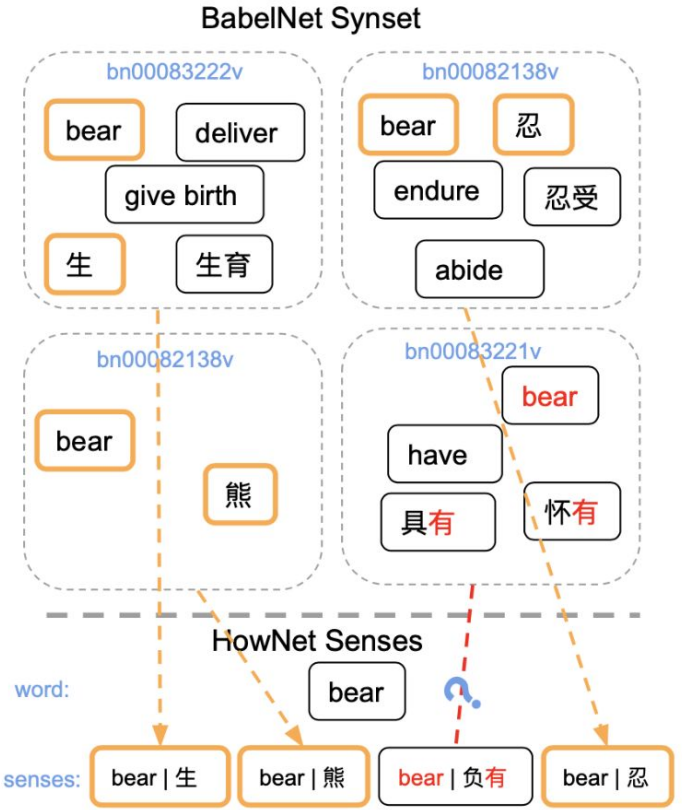
Stage One: Exact Match

- Based on well-known observation that distinct senses of a word may translate differently
- Aim at high precision rather than high coverage
- Although strict, it is correct almost all the time since an English-Chinese translation pair usually disambiguates the sense



Stage Two: Partial Match

- A less strict version of exact match
- Aim at improving coverage
- Key idea: Chinese words that have same or similar meanings share the same characters



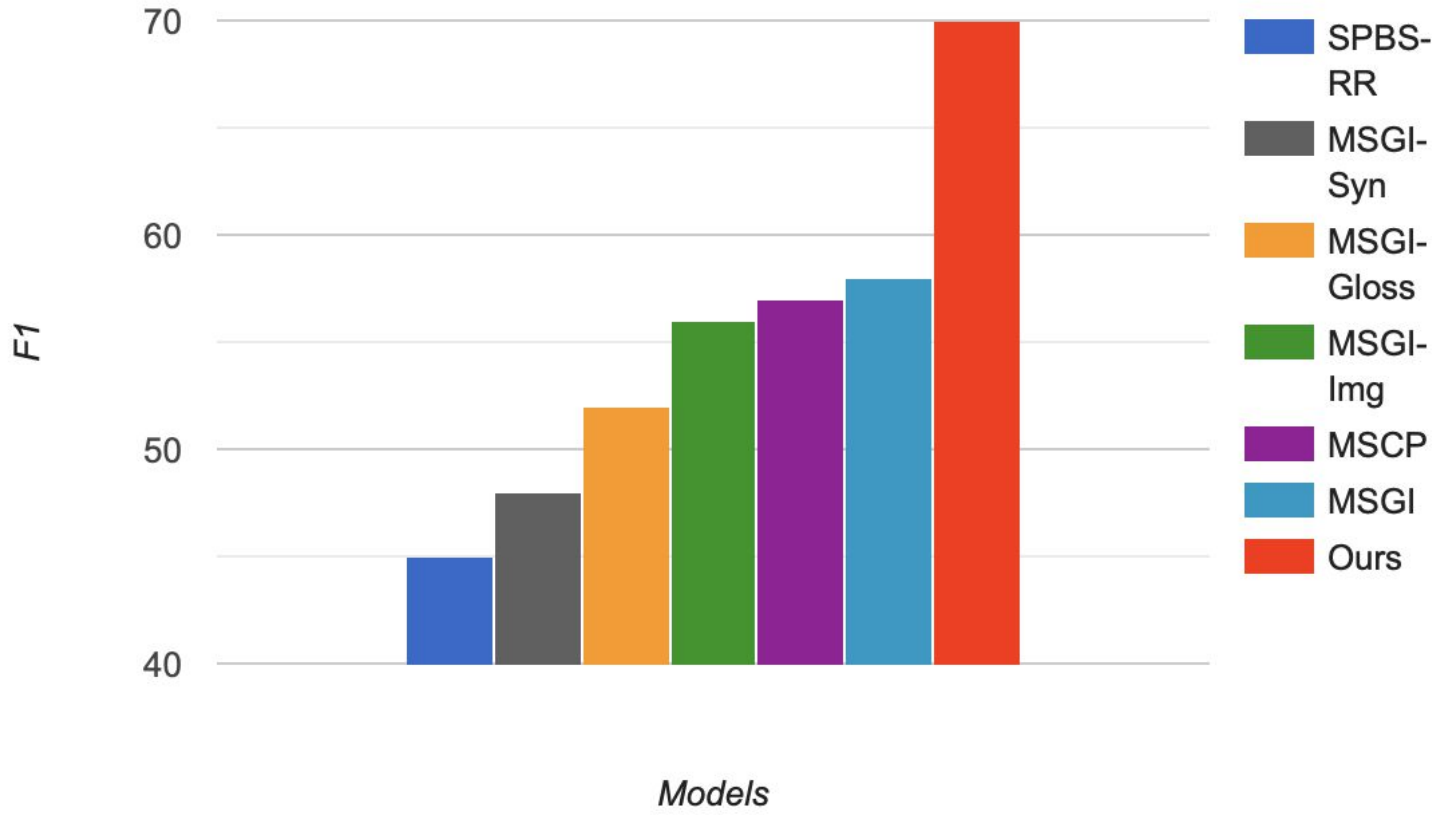
Stage Three: Sense Information Matching

- Aim to cover most of the left over senses with a coarse alignment rule
- BabelNet hypernyms often match HowNet sememes

Stage Four: Proper Name Matching

- Cover any left over synsets: Map them to HowNet's "proper name" sense

Main Results on BabelSememe Dataset



Conclusion

- We present a novel unsupervised method for aligning BabelNet and HowNet
- State-of-the-art results on sememe prediction
  - We outperform supervised systems
- Future work: leverage sense alignment for other semantic tasks, including WSD