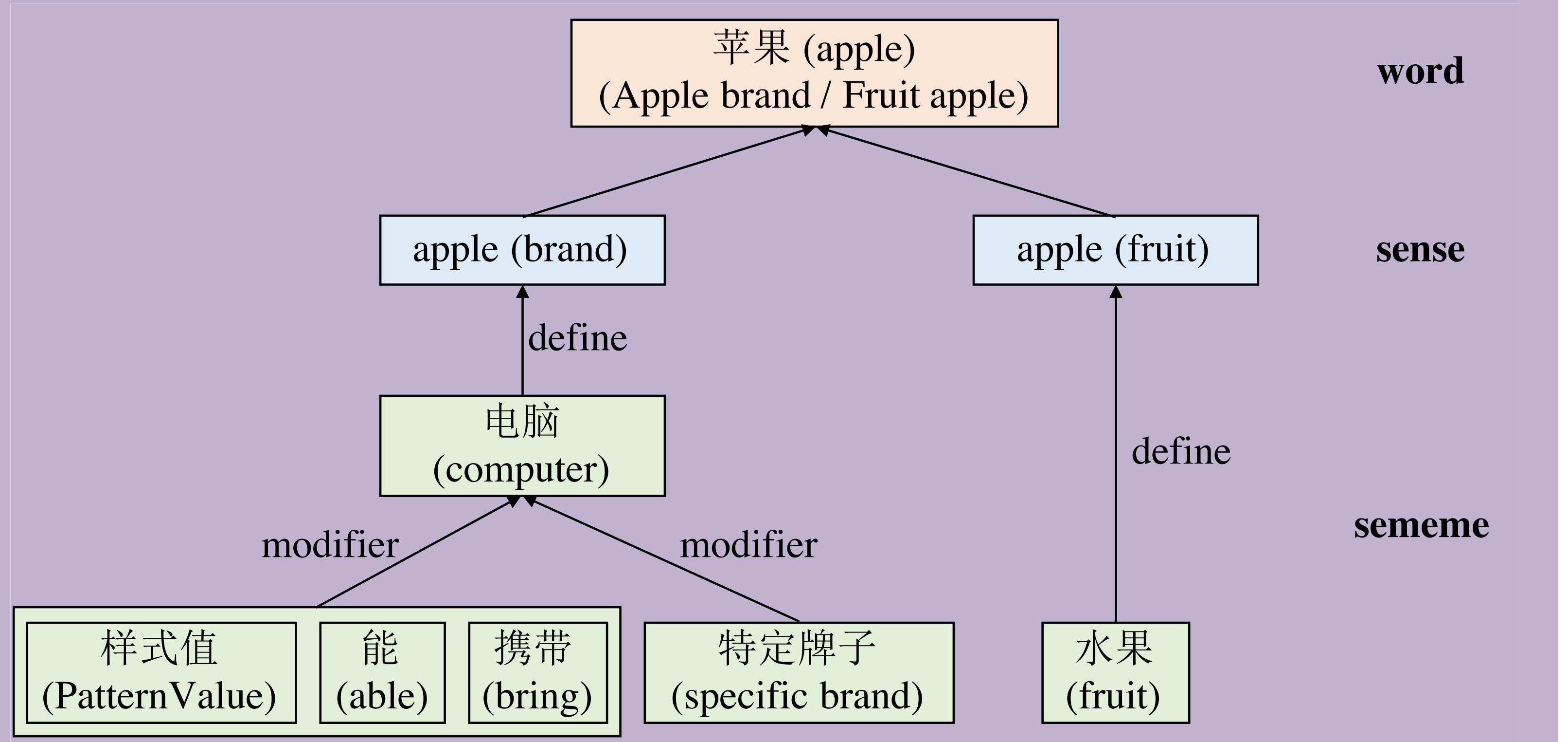
Incorporating Chinese Characters of Words for Lexical Sememe Prediction

Huiming Jin[1]*, Hao Zhu[2]*, Zhiyuan Liu[2], Ruobing Xie[3], Maosong Sun[2], Fen Lin[3], Leyu Lin[3]

[1] Beihang University [2] Tsinghua University [3] Wechat, Tencent

huimingj@cmu.edu, zhuhao15@mails.tsinghua.edu.cn, liuzy@tsinghua.edu.cn, xrbsnowing@163.com

Backgrounds



- **Sememes**: minimum semantic units.
- The meanings of concepts can be composed by a finite number of sememes.
- Linguists build knowledge bases to annotate words with sememes manually.
- HowNet (Dong and Dong, 2006) is a classical widely-used sememe KB.
  - 100,000 common words in Chinese and English and 2,000 sememes.
  - Each word is represented as a tree-like sememe structure.

Sememe Prediction

SP with Word Embeddings (SPWE)

- Applies the ideas of collaborative filtering.
- Recommends the sememes of similar words.

SP with Sememe Embeddings (SPSE)

- Maps sememe and word embeddings into the same low-dimensional space.
- Measures the distances between words and sememes as the scores to recommend.

Huiming's Homepage

Hao's Homepage

Our Code

Zhiyuan's Homepage

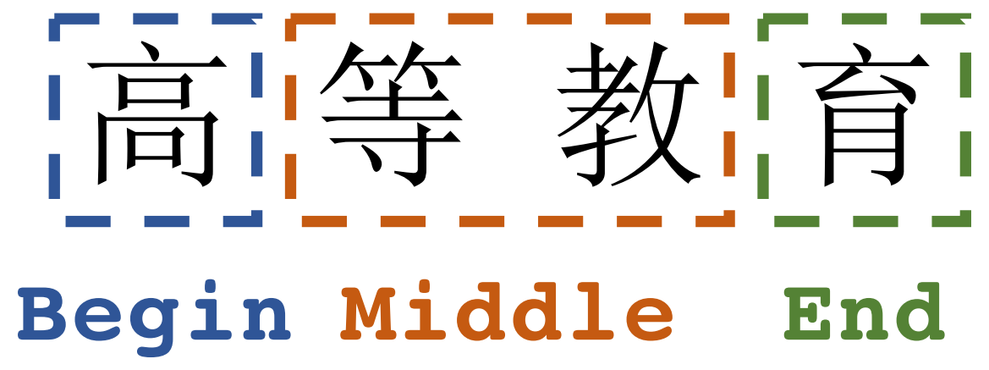
Ruobing's Homepage

The 56th Annual Meeting of the Association for Computational Linguistics

Methodology

- Character-enhanced Sememe Prediction (CSP).
- Ensemble of two parts:
  - Using internal information or character-level information (*internal* models) — SPWCF and SPCSE.
  - Using external information or corpora (*external* models) — the existing methods.

SP with Word-to-Character Filtering (SPWCF)



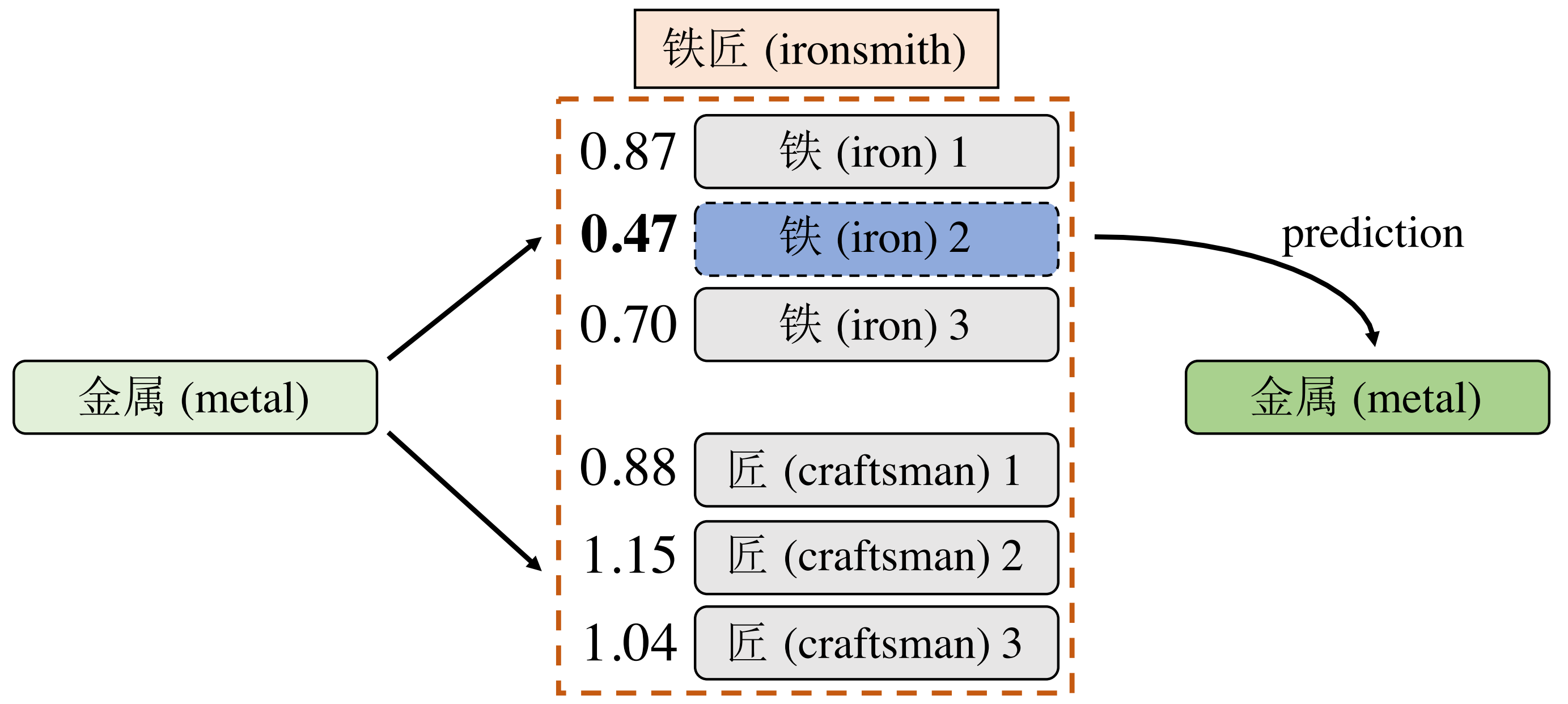
- The positional score function:

$$P_p(s_j|c) \sim \frac{\sum_{w_i \in W \wedge c \in \pi_p(w_i)} \mathbf{M}_{ij}}{\sum_{w_i \in W \wedge c \in \pi_p(w_i)} |S_{w_i}|}$$

- The final score function:

$$P(s_j|w) \sim \sum_{p \in \{B,M,E\}} \sum_{c \in \pi_p(w)} P_p(s_j|c)$$

SP with Character and Sememe Embeddings (SPCSE)



- Selects the most representative character embedding to represent a word:

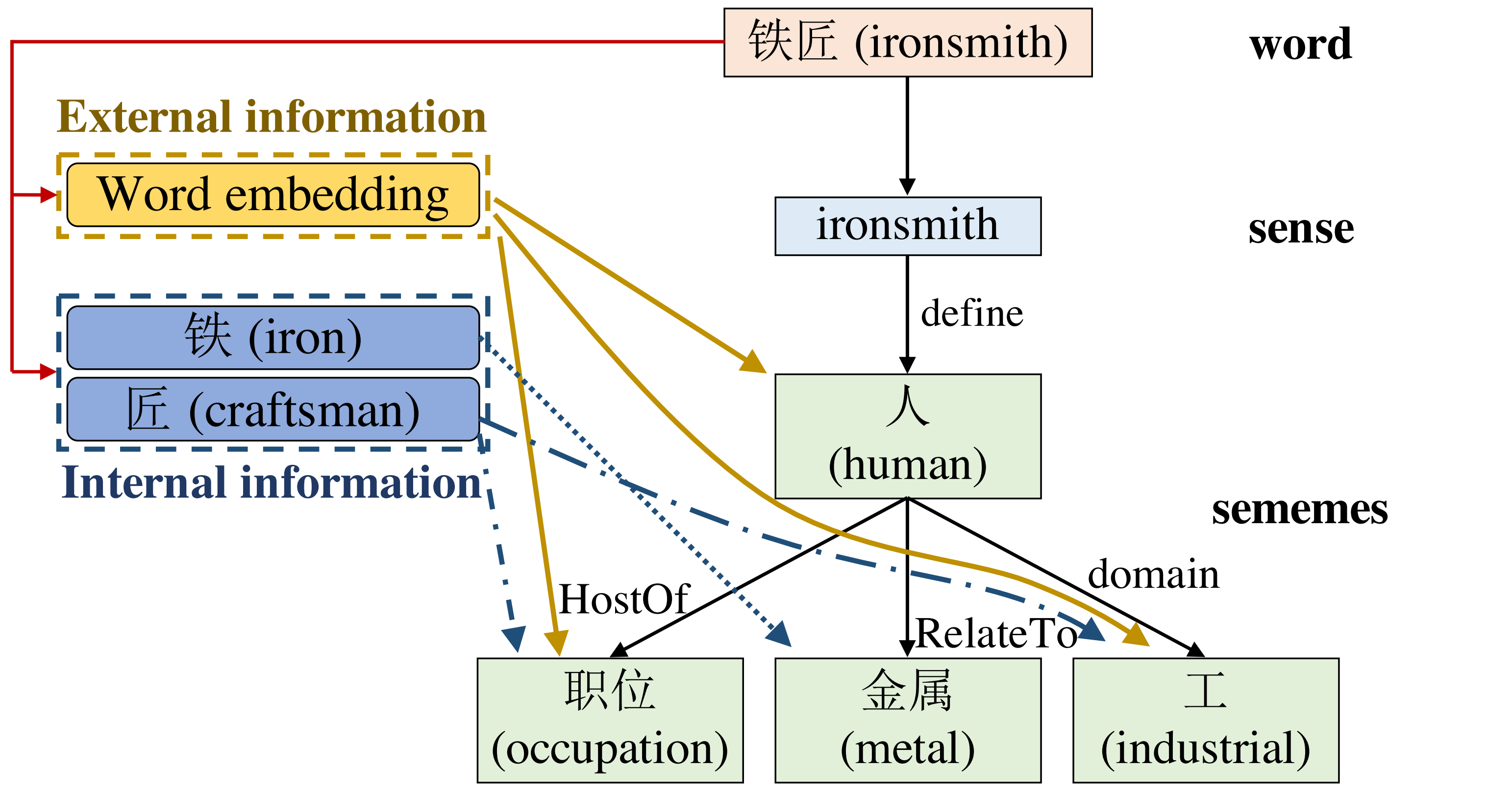
$$\hat{k}, \hat{r} = \arg\min_{k,r} \left[1 - \cos(\mathbf{c}_k^r, (\mathbf{s}'_j + \bar{\mathbf{s}}'_j))\right]$$

- Learns the sememe embeddings with the loss function:

$$\mathcal{L} = \sum_{w_i \in W, s_j \in S} \left(\mathbf{c}_{\hat{k}}^{\hat{r}} \cdot (\mathbf{s}'_j + \bar{\mathbf{s}}'_j) + \mathbf{b}_{\hat{k}}^c + \mathbf{b}''_j - \mathbf{M}_{ij}\right)^2 + \lambda' \sum_{s_j, s_q \in S} \left(\mathbf{s}'_j \cdot \bar{\mathbf{s}}'_q - \mathbf{C}_{jq}\right)^2$$

- The score function of a word $w = c_1 ... c_{|w|}$:

$$P(s_j|w) \sim \mathbf{c}_{\hat{k}}^{\hat{r}} \cdot (\mathbf{s}'_j + \bar{\mathbf{s}}'_j)$$



Experiments

- Sememe selection: 1,400.
- Corpus for embeddings: Sogou-T.
- Word embeddings: GloVe.
- Character embeddings: Cluster-based Character Embeddings (Chen et al., 2015).

Method	MAP
SPSE	0.411
SPWE	0.565
SPWE+SPSE	0.577
SPWCF	0.467
SPCSE	0.331
SPWCF + SPCSE	**0.483**
SPWE + fastText	0.531
CSP	**0.654**

word frequency occurrences	⩽ 50 8537	51– 100 4868	101 – 1,000 3236	1,001 – 5,000 2036	5,001 – 10,000 663	10,001 – 30,000 753	>30,000 686
SPWE	0.312	0.437	0.481	0.558	0.549	0.556	0.509
SPSE	0.187	0.273	0.339	0.409	0.407	0.424	0.386
SPWE + SPSE	0.284	0.414	0.478	0.556	0.548	0.554	0.511
SPWCF	0.456	0.414	0.400	0.443	0.462	0.463	0.479
SPCSE	0.309	0.291	0.286	0.312	0.339	0.353	0.342
SPWCF + SPCSE	0.467	0.437	0.418	0.456	0.477	0.477	0.494
SPWE + fastText	0.495	0.472	0.462	0.520	0.508	0.499	0.490
CSP	**0.527**	**0.555**	**0.555**	**0.626**	**0.632**	**0.641**	**0.624**

words	models	Top 5 sememes
钟表匠 (clockmaker)	internal	人**(human)**, 职位**(occupation)**, 部件(part), 时间**(time)**, 告诉**(tell)**
	external	人**(human)**, 专(ProperName), 地方(place), 欧洲(Europe), 政(politics)
	ensemble	人**(human)**, 职位**(occupation)**, 告诉**(tell)**, 时间**(time)**, 用具**(tool)**
奥斯卡 (Oscar)	internal	专**(ProperName)**, 地方(place), 市(city), 人(human), 国都(capital)
	external	奖励**(reward)**, 艺**(entertainment)**, 专**(ProperName)**, 用具(tool), 事情**(fact)**
	ensemble	专**(ProperName)**, 奖励**(reward)**, 艺**(entertainment)**, 著名(famous), 地方(place)

Take-home Message

- Models using only internal information could make good predictions.
- Integrating with methods incorporating character information could improve prediction performance and especially frequency robustness.
- Our CSP framework achieves the state of the art on sememe prediction.