清華大學 Tsinghua University

Paper Sharing

Zhu Liu

2023.01.06

Outline

Home / Archives / Vol. 35 No. 15: AAAI-21 Technical Tracks 15 /
AAAI Technical Track on Speech and Natural Language Processing II

XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation

Leveraging Word-Formation Knowledge for Chinese Word Sense Disambiguation

Hua Zheng,* Lei Li,* Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, Yang Liu†
Department of Computer Science and Technology, Peking University
Key Lab of Computational Linguistics (MOE), Peking University
{zhenghua,daidamai,chendeli,tianyu0421}@pku.edu.cn
lilei@stu.pku.edu.cn
{xusun,liuyang}@pku.edu.cn

XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation

Tommaso Pasini[*♠], **Alessandro Raganato**[*♡], **Roberto Navigli**[♠]

[♠]Sapienza NLP Group, Computer Science Department, Sapienza University of Rome
[♡]Department of Digital Humanities, University of Helsinki, Finland
[♠]{pasini,navigli}@di.uniroma1.it
[♡]alessandro.raganato@helsinki.fi

AAAI 2021; Cited by 31

https://sapienzanlp.github.io/xl-wsd/

Roberto Navigli

2007, Mar. ~ Now

Faculty

BabelNet | Multilinguality | Natural Language Understanding | Semantics

Full professor and head of the Sapienza NLP Group, conducting research in multilingual Natural Language Understanding, including Word Sense Disambiguation, Semantic Role Labeling and Semantic Parsing in multiple languages.

Federico Martelli

2019, Feb. ~ Now

PhD Student

Knowledge Acquisition | Multilinguality | Sense Clustering | Word Sense Disambiguation

Ph.D. student in Natural Language Processing. On a mission to overcome linguistic barriers through innovative technology.

Simone Conia

2019, Jul. ~ Now

PhD Student

Concept Representation | Multilinguality | Semantic Role Labeling

Natural Language Understanding: is it possible to learn this power?

Pere-Lluís Huguet Cabot

2020, Nov. ~ Now

PhD Student

Common sense reasoning

Multi-task learning | Multilinguality

Early Stage Researcher interested in how to learn from any text, such as unsupervised and semi-supervised approaches. Also interested in explainable and ethical AI. Currently working on multilingual Relation and Entity extraction.

Simone Tedeschi

2021, Nov. ~ Now

PhD Student

Entity Linking | Idioms

Named Entity Recognition

PhD student interested in the automatic creation of training data, semantics and multilinguality.

Riccardo Orlando

2021, Nov. ~ Now

PhD Student

MLOps | Multilinguality

Semantic Role Labeling

PhD student interested in Multilingual NLP and its applications.

Stefan Bejgu

2021, Nov. ~ Now

PhD Student

Multilinguality | Semantics

Sentence Representations

PhD student interested in teaching multilinguality and semantics to machines.

Lorenzo Proietti

2022, Nov. ~ Now

PhD Student

Interpretability | Multilinguality

Semantic Parsing

PhD student interested in Semantic Parsing and its use in other NLP tasks.

数据集建构

MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation)

Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

评估框架

XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation

Proc. of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)

F. Martelli, N. Kalach, G. Tola, R. Navigli

SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation

Proc. of the 15th International Workshop on Semantic Evaluation

其他应用

B. Scarlini, T. Pasini, R. Navigli

SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation

Proc. of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020), New York, USA, 7-12th February, 2020.

T. Pasini, F. M. Elia, R. Navigli

Huge Automatically Extracted Training-Sets for Multilingual Word Sense Disambiguation.

Proceedings of the Language Resources and Evaluation Conference (18)

模型方法

R. Orlando, S. Conia, F. Brignone, F. Cecconi, R. Navigli

AMuSE-WSD: An All-in-one Multilingual System for Easy Word Sense Disambiguation

Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing

L. Procopio, E. Barba, F. Martelli, R. Navigli

MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation

Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence

F. Scozzafava, M. Maru, F. Brignone, G. Torrisi, R. Navigli

Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation

Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2020)

T. Pasini

The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation

Proc. of the 29th International Joint Conference on Artificial Intelligence (IJCAI-20)

E. Barba, L. Procopio, N. Campolungo, T. Pasini, R. Navigli

MuLaN: Multilingual Label propagatioN for Word Sense Disambiguation

Proc. of the 29th International Joint Conference on Artificial Intelligence (IJCAI-20)

J. Camacho-Collados, C. Delli Bovi, A. Raganato, R. Navigli

SenseDefs: a multilingual corpus of semantically annotated textual definitions - Exploiting multiple languages and resources jointly for high-quality Word Sense Disambiguation and Entity Linking

Language Resources and Evaluation

Motivation

- Word Sense Disambiguation (WSD) is the task of associating words in context with their possible meanings contained in a pre-defined sense inventory.
- A unified evaluation [Raganato et al, 2017] in English WSD has facilitated the fast development of models for this task.
- Multilingual WSD is lack of reliable benchmarks, even though with:

  1) automatically sense-annotated corpora for different languages;

  2) language-specific WordNet-like resources;

  The gold standards use diverse data formats and outdated, or even unavailable, inventories of senses -> hard to compare among systems.
- Transformer-based architectures and Pretrain-finetune paradigm.

Contributions

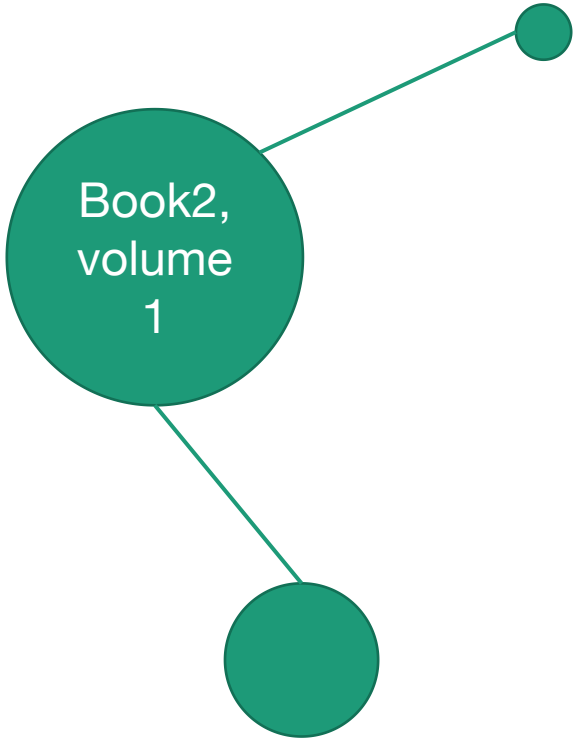
- A multilingual WSD test suite in 18 languages from six language families, comprising 99,450 gold annotations in total, new automatically-produced training data for non-English languages and a unified multilingual inventory of concepts.
- An extension of the fine-grained English WSD framework [Raganato et al, 2017] by including new training, development and testing data as well as a coarse-grained evaluation dataset.
- Extensive experiments (large pre-trained multilingual language models)

Resources

Dictionary

- WordNet [Miller 1998]: synonym-based; manually-curated; English

WordNet



synset

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: book Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **book** (a written work or composition that has been published (printed on pages bound together)) *"I am reading a good book on economics"*
- S: (n) **book**, volume (physical objects consisting of a number of pages bound together) *"he used a large book as a doorstop"*
- S: (n) record, record book, **book** (a compilation of the known facts regarding something or someone) *"Al Smith used to say, `Let's look at the record'"; "his name is in all the record books"*
- S: (n) script, **book**, playscript (a written version of a play or other dramatic composition; used in preparing for a performance)
- S: (n) ledger, leger, account book, book of account, **book** (a record in which commercial accounts are recorded) *"they got a subpoena to examine our books"*
- S: (n) **book** (a collection of playing cards satisfying the rules of a card game)
- S: (n) **book**, rule book (a collection of rules or prescribed standards on the basis of which decisions are made) *"they run things by the book around here"*
- S: (n) Koran, Quran, al-Qur'an, **Book** (the sacred writings of Islam revealed by God to the prophet Muhammad during his life at Mecca and Medina)
- S: (n) Bible, Christian Bible, **Book**, Good Book, Holy Scripture, Holy Writ, Scripture, Word of God, Word (the sacred writings of the Christian religions) *"he went to carry the Word to the heathen"*
- S: (n) **book** (a major division of a long written composition) *"the book of*

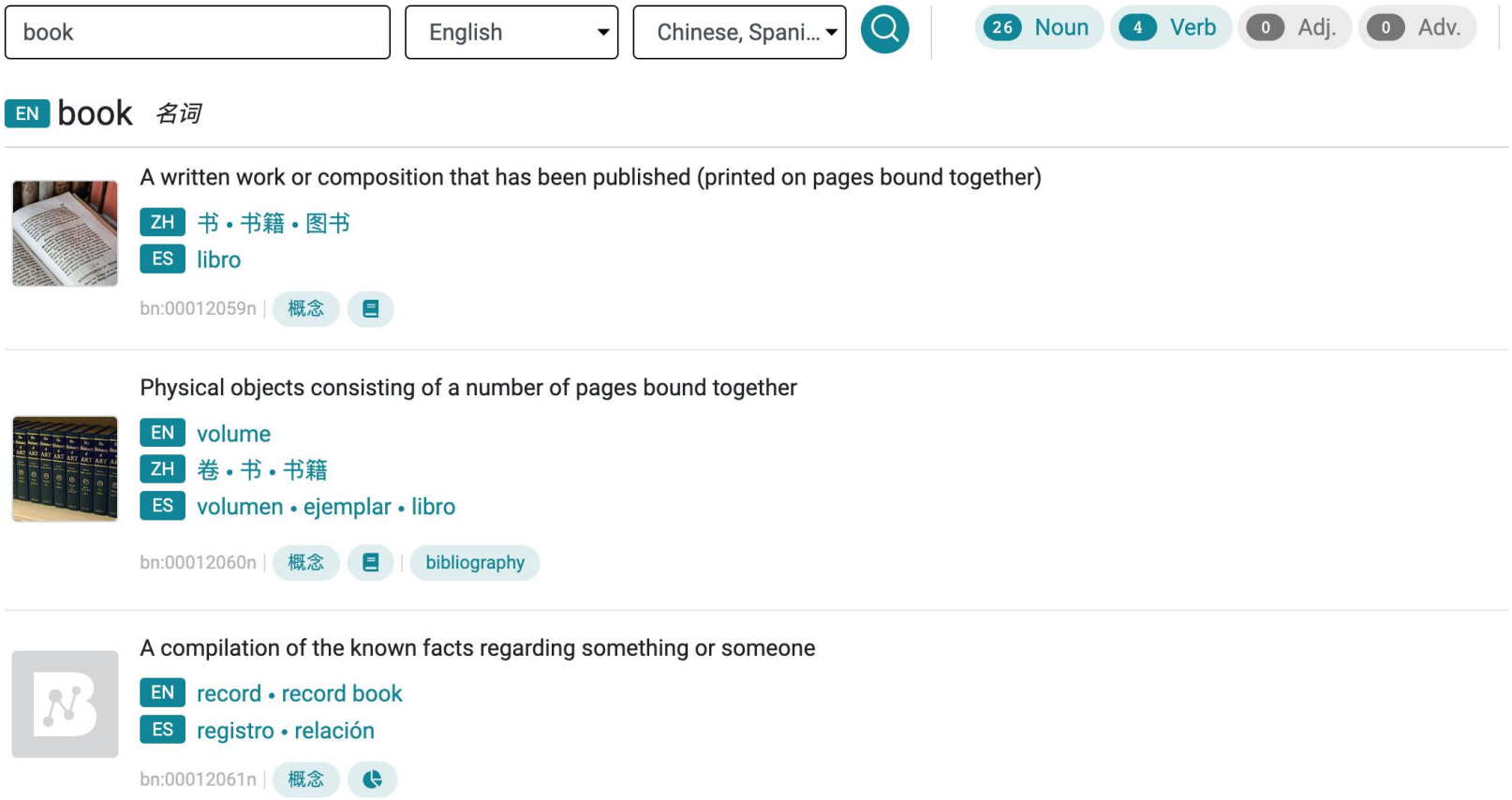
Resources

Dictionary

- WordNet [Miller 1998]: synonym-based; manually-curated; English
- Translation-based multilingual wordnets
- BabelNet [Navigli and Ponzetto 2012]: multilingual encyclopedic dictionary

BabelNet

- (multilingual) Synset-based
- Even Mutimodal!
- Coordinate with Wordnet
- Not every concept maps to every language.



Resources

Dictionary

- WordNet [Miller 1998]: synonym-based; manually-curated; English
- BabelNet [Navigli and Ponzetto 2012]: multilingual encyclopedic dictionary

Sense-annotated corpus

- SemCor [Miller et al., 1993]
- Data from Senseval and SemEval competitions

1) Mainly English -> unified framework [Raganato et al, 2017]
2) Multilingual benchmarks, but only a few, and outdated

XL-WSD

A multilingual WSD benchmark

- Sense Inventory
- Gold Standards (Test and Dev set)
- Training Data
- Statistics

Sense Inventory

- Subset of the BabelNet 4.0, where the synset meets the condition:

It contains at leat one sense from the Princeton WordNet 3.0.

- 117,659 BabelNet synsets
- Reason for the constraint: a shared inventory; to transfer more easily
- Limits: may not a faithful equivalent of a dictionary in other languages.
- For each sysnet (meaning), it defines (sense, lemma, pos) triplet pair.

Gold Standards (test)

Language-specific WordNet datasets

- Basque (巴斯克语), Bulgarian (保加利亚语), Catalan (加泰罗尼亚语), Chinese, Croatian (克罗地亚语), Danish, Dutch, Estonian (爱沙尼亚语), Galician (加利西亚语), Hungarian, Japanese, Korean, and Slovenian (斯洛文尼亚语)
- Automatic annotations from examples
- Given synset, pos and lemma, match the target word in the example.
- Synset mapping: Mutilingual WN -> PWN -> BabelNet

Gold Standards

- SemEval Datasets
- Italian and Chinese: SemEval-10 Task 17
- French, German, Italian and Spanish datasets: SemEval-13 Task 12
- Italian and Spanish: SemEval-15 Task 13

- English: Original framework + SemEval-10 Task 17; SemEval07 Task 7 (Coarse)
- Wordnet mapping (data cleaning); 8-2 splitting

Training Data

- SemCor (SC)
- Princeton WordNet Gloss Corpus (WNG)
- Translated corpora (T-SC+WNG)

1) 15 non-English languages

2) Machine translation model: Opus-MT

3) Translated sentences with target words and senses annotated by looking up the multilingual sysnet.

4) Silver (noisy) training data: not perfect, but i) automatic; 2) maintained sense distribution; 3) high coverage.

Statistics

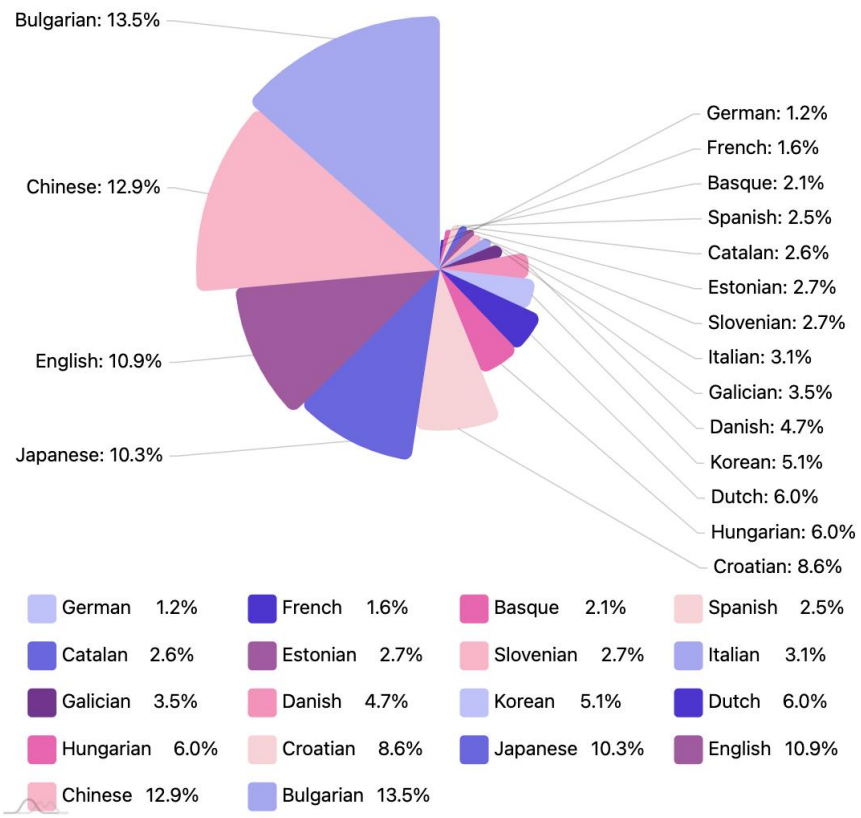
Language	Word Types			Polysemous Words			Word-Type Polysemy			Instances			Unique Synsets		
	Train	Test	Dev	Train	Test	Dev	Train	Test	Dev	Train	Test	Dev	Train	Test	Dev
English-Fine	106906	2882	330	24658	2199	308	1.458	3.689	6.209	840471	8062	455	117653	3469	361
English-Coarse	-	980	-	-	750	-	-	4.255	-	-	1816	-	-	2190	-
Basque	12503	771	304	5294	525	253	2.331	3.224	4.543	197309	1580	395	16604	1423	388
Bulgarian	12413	2450	1413	2412	1325	839	1.304	1.670	1.938	148479	9968	2493	12600	2658	1517
Catalan	18603	1276	428	8378	1107	384	2.291	3.940	4.981	331757	1947	487	25624	1767	479
Chinese	-	1786	1173	-	1402	955	-	2.638	3.045	-	9568	2392	-	2687	1524
Croatian	6882	4389	1416	1161	1652	675	1.268	1.244	1.758	94575	6333	1584	6739	4543	1449
Danish	15822	2623	816	3324	1318	428	1.338	1.722	1.950	234681	3502	876	16707	2693	817
Dutch	28351	2935	985	9121	2122	766	1.711	2.356	3.067	305692	4400	1100	30490	2716	950
Estonian	10460	1615	460	1768	917	281	1.246	1.815	2.091	132240	1999	500	10462	1852	490
French	17850	549	203	5978	339	130	1.585	2.413	2.744	252756	1160	289	21510	584	213
Galician	8390	1244	486	3799	773	349	2.079	2.219	2.852	247379	2561	641	11821	1474	548
German	16213	421	154	2332	166	64	1.203	1.639	1.864	184952	862	214	16437	417	155
Hungarian	13234	3491	1022	2908	1931	625	1.367	1.842	2.346	161119	4428	1107	13297	4285	1103
Italian	23773	985	385	9540	758	316	2.021	3.790	4.569	385248	2278	561	29869	1212	475
Japanese	1008	4338	1538	581	2390	1001	2.516	1.871	2.460	23217	7602	1901	1141	5964	1755
Korean	-	1886	740	-	920	408	-	1.373	1.815	-	3796	950	-	1452	683
Slovenian	7577	104	87	1296	93	81	1.245	3.519	3.954	128395	2032	509	7705	243	172
Spanish	22020	847	329	11784	696	270	2.811	4.955	5.435	393539	1851	452	32151	1103	422

Table 1: Statistics of the training, test and development sets comprised in XL-WSD. The Train column refers to SC+WNG for English and to T-SC+WNG for all the other languages.

Statistics

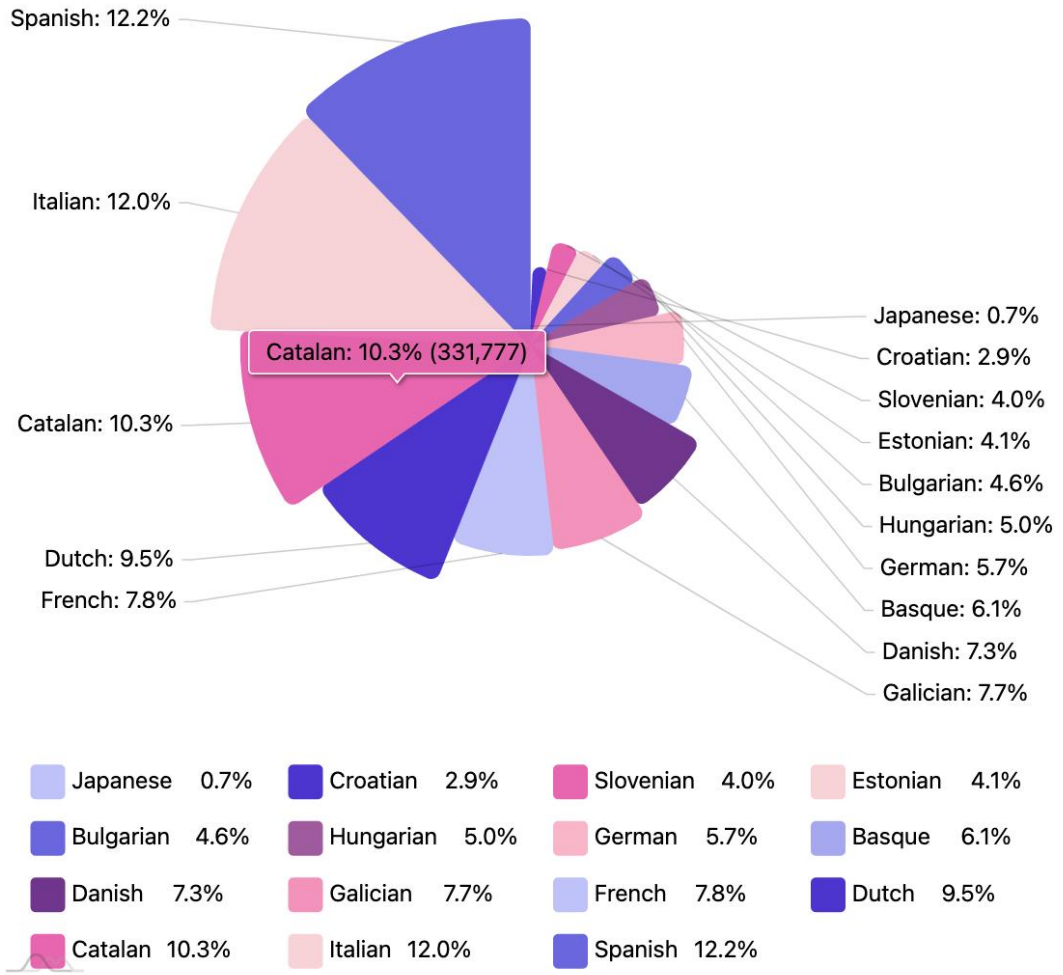
Test Sets (Gold)

Number of test instances by language.



Training Sets (Silver)

Number of training instances by language.



Experiment

Models

- Multilingual pre-trained models: XLMR-Base, XLMR-Large, M-BERT
- Language-specific versions of Bert: L-BERT

Evaluation measure

- F1 score

Data

- English: SemCor and WordNet Gloss (SC+WNG)
- Other languages: T-SC+WNG
- One baseline: MULAN datasets: pair sentences in a corpus of raw texts by a multilingual contextualized word representations.

Results

- English Benchmark
- Multilingual Evaluation

1) Zero-shot setting
Pre-trained multilingual models matter (model or data?)
2) Language-specific setting (Italian and Spanish, more WNG)

	∅-Shot (SC+WNG)			Language-Specific (MULAN)		Language-Specific (T-SC+WNG)		Knowledge-Based		
Dataset	XLMR-Large	XLMR-Base	M-BERT	XLMR-Large	L-BERT	XLMR-Large	L-BERT	SyntagRank	Babelfy	MCS
English-Fine	76.28	74.50	72.40	-	-	76.28	**76.77**	69.96	64.09	63.37
English-Coarse	91.30	91.02	89.70	-	-	91.30	**91.57**	83.78	82.54	80.23
Basque	**47.15**	43.80	42.41	-	-	41.96	43.04	42.91	36.65	32.72
Bulgarian	**72.00**	71.59	68.78	-	-	58.18	57.85	61.10	60.39	58.16
Catalan	**49.97**	47.77	47.35	-	-	36.00	36.98	43.98	36.65	27.17
Chinese	**51.62**	49.77	48.99	-	-	-	-	41.23	34.94	29.62
Croatian	**72.29**	72.13	70.65	-	-	63.15	62.89	68.35	63.75	62.88
Danish	**80.61**	79.18	76.04	-	-	78.67	76.41	72.93	71.33	64.33
Dutch	**59.20**	58.77	56.64	-	-	57.27	56.64	56.00	44.27	44.61
Estonian	**66.13**	64.82	64.33	-	-	50.78	51.23	56.31	49.62	46.87
French	**83.88**	82.33	81.64	81.98	80.78	71.38	71.12	69.57	67.41	59.31
Galician	66.28	64.79	**68.07**	-	-	56.18	56.95	67.56	64.17	60.85
German	83.18	82.13	80.63	**83.29**	82.13	73.78	73.78	75.99	77.84	75.99
Hungarian	67.64	**68.38**	65.24	-	-	52.60	52.17	57.98	51.99	47.29
Italian	77.66	76.73	76.16	74.10	73.88	**77.70**	75.68	69.57	64.22	52.77
Japanese	**61.87**	61.46	60.34	-	-	50.55	50.16	57.46	51.91	48.71
Korean	**64.20**	63.65	63.37	-	-	-	-	50.29	51.95	52.48
Slovenian	**68.36**	66.34	62.16	-	-	51.13	49.66	52.25	35.38	36.71
Spanish	75.85	76.55	74.66	73.47	74.77	**77.26**	74.88	68.58	64.07	55.65
Micro AVG	**65.66**	64.82	62.84	-	-	-	-	57.68	52.85	49.31

Table 3: F1 scores of supervised and knowledge-based approaches as well as language-specific BERT models (L-BERT) and the Most Common Sense (MCS) baseline on the test splits. As for the ∅-Shot columns, models are trained and tuned in English only and tested in all the other languages. As for the Language-Specific columns, models are trained, tuned and tested on either MULAN or T-SC+WNG language-specific datasets. The Micro AVG row shows the micro F1 across all languages but English.

Discussion

- XLMR-Large still performs worse in some languages, Basque, Catalan, and Chinese.

  From ML learning perspective: less proportion of test instances appeared much in training set.

Discussion

- Results on each POS tag
- Verbs are harder except for

1) Bulgarian, Dutch (less polysemous)
2) Catalan (加泰罗尼亚语) (less verbal instances)

Dataset	ALL	NOUN	VERBS	ADJ	ADV
English-Fine	76.28	77.92	65.74	81.47	86.71
English-Coarse	91.30	92.72	88.64	89.55	91.75
Basque	47.15	47.15	-	-	-
Bulgarian	72.00	70.69	86.04	74.07	-
Catalan	49.97	49.28	54.84	52.89	-
Chinese	51.62	57.92	45.47	47.01	84.48
Croatian	72.29	71.85	70.37	85.03	-
Danish	80.61	80.32	79.66	83.63	-
Dutch	59.20	56.08	63.56	-	-
Estonian	66.13	68.81	49.66	74.63	68.14
French	83.88	83.88	-	-	-
Galician	66.28	71.43	-	65.97	-
German	83.18	83.18	-	-	-
Hungarian	67.64	70.41	50.41	-	-
Italian	77.66	77.91	71.89	81.58	77.27
Japanese	61.87	67.87	52.72	56.39	71.29
Korean	64.20	64.47	46.43	-	-
Slovenian	68.36	68.34	-	-	-
Spanish	75.85	76.72	66.83	77.88	85.00

Table 4: F1 scores breakdown of XLMR-Large model on the zero-shot setting by POS tags.

Conclusion

- XL-WSD, a large-scale evaluation benchmark for WSD in 18 different languages: 34 new gold datasets; 15 silver datasets for training
- Extend previous English evaluation benchmark
- Strong baseline models (multilingual bert)

Leveraging Word-Formation Knowledge for Chinese Word Sense Disambiguation

Hua Zheng[*], **Lei Li**[*], **Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, Yang Liu**[†]
Department of Computer Science and Technology, Peking University
Key Lab of Computational Linguistics (MOE), Peking University
{zhenghua,daidamai,chendeli,tianyu0421}@pku.edu.cn
lilei@stu.pku.edu.cn
{xusun,liuyang}@pku.edu.cn

EMNLP 2021 short paper; Cited by 2

https://github.com/TobiasLee/FormBERT

中文简介

孙栩，北京大学信息学院研究员、博士生导师，并担任新体制长聘副教授。2010年于日本东京大学获得计算机博士学位。先后在日本东京大学、微软公司美国雷蒙德研究院、美国康奈尔大学、香港理工大学担任研究职位。研究方向为自然语言处理和机器学习，特别是自然语言生成、面向语言的深度学习。在ACL、ICML、NIPS、ICLR、EMNLP、COLING等发表多篇论文。获得香港求是科技基金会“求是杰出青年学者奖”、中国电子学会科学技术奖一等奖、COLING 2018最佳论文奖。

Liu, Yang

Associate Professor

Research Interests: Chinese information processing, language knowledge engineering

Office Phone: 86-10-6276 5835-205

Email: liuyang@pku.edu.cn

Liu, Yang is an associate professor in the Department of Computer Science and Technology, School of EECS. He obtained his B.Sc. from Shaanxi Normal University in year 1993, and M.Sc./L.D. from Peking University in year 2000/2003 respectively. His research interests include Chinese Information Processing, Language Knowledge Engineering, etc.

- Work related to researches and construction of the global WordNets
- Work related to knowledge representation of Chinese lexical semantics

Motivation

- To leverage word-formation knowledge to enhance Chinese WSD.
- For Chinese words, word formation rules/categories are more eminent than Indo-European languages.
- Words with diff meanings.

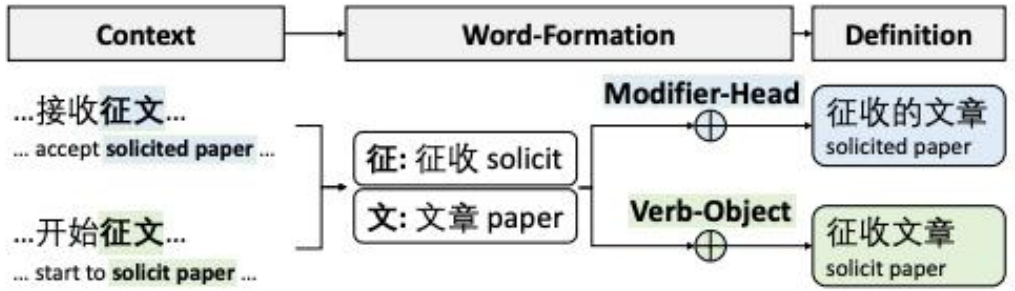


Figure 1: The contexts indicate that the word "征文" holds two senses constructed by different word-formations, which can be used to enhance WSD.

Word-Formation

- From Prof. Zhu, Chinese words are formed mainly in three ways: 重叠、附加、复合

Word-Formation	Explanation	Example	%
联合 (Parallel)	$morph_1$ and $morph_2$ are similar, contrasting or complementary.	文体 (literary-physics)	34.40
定中 (Modifier-Head)	$morph_1$ modifies $morph_2$ (noun).	引文 (cited-paper)	18.72
述宾 (Verb-Object)	$morph_1$ operates on $morph_2$.	发文 (publish-paper)	14.66
单纯 (Single Morpheme)	The word is a single morpheme.	葡萄 (grape)	9.09
状中 (Adverb-Verb)	$morph_1$ modifies $morph_2$ (verb).	博引 (widely-cite)	5.81
连谓 (Verb-Consequence)	$morph_2$ is the consequence of $morph_1$.	休息 (stop-rest)	4.09
后缀 (Suffixation)	$morph_2$ is the suffix of $morph_1$.	花头 (trick-∅)	3.61
前缀 (Prefixation)	$morph_1$ is the prefix of $morph_2$.	老师 (∅-teacher)	3.47
述补 (Verb-Complement)	$morph_2$ is the action follows $morph_1$.	压低 (press-down)	2.50
重叠 (Overlapping)	$morph_1$ and $morph_2$ are the same.	白白 (vainly-vainly)	1.15
主谓 (Subject-Predicate)	$morph_1$ is the subject of $morph_2$.	眼花 (eyesight-dim)	1.13
介宾 (Preposition-Object)	$morph_1$ is a preposition, $morph_2$ is an object.	凭空 (from-nowhere)	0.49
方位 (Entity-Position)	$morph_1$ is an entity, $morph_2$ is a position.	期中 (semester-mid)	0.41
数量 (Number-Quantifier)	$morph_1$ is a number, $morph_2$ is a quantifier.	一点 (one-dot)	0.28
复量 (Quantifier-Quantifier)	Both $morph_1$ and $morph_2$ are quantifiers.	千米 (kilo-meter)	0.11
名量 (Noun-Quantifier)	$morph_2$ is the quantifier of $morph_1$.	花朵 (flower-bud)	0.07

Table 6: Descriptions of the total 16 word-formations. ∅ denotes the affix and % denotes the instance percentage. The first and the third columns are in the format of "Chinese characters (English translation)". We give a simple explanation in the second column to describe the relation between two characters, which functions as a guideline to the annotators.

Contribution

- To construct a large-scale **F**ormation-**i**nformed **C**hinese **L**exical **S**ample WSD dataset (**FiCLS**)
- To propose a model FormBERT to explicitly incorporate word-formations into sense disambiguation.
- To design a word-formation predictor module to predict word-formations for unannotated data.

FiCLS Dataset

- Each FiCLS entry consists of (1) a word, (2) a sense definition, (3) a word-formation and (4) a context sentence.
- Dictionary: 5th edition of the Contemporary Chinese Dictionary (CCD)

More complete and native than HowNet sememe and Chinese WordNet

FiCLS

- CCD contains 62,241 words, of which 22.32% are polysemous.
- 7,064 polysemous words (20,382 senses) with use cases (short sentences)
- Data augmentation

(distributional hypothesis)

- 145,964 entries in total, where

the average length and number of

contexts per sense are 53.04, 7.16

- Human check -> 121,655 entries

Word-Formation Annotations

- 16 Chinese word-formations
- Annotators are professors and postgraduates
- Each entry is cross-validated by three independent annotators and reviewed by one if not agreed
- ITA: 92.61

Word-Formation	Example	%
Parallel	文体 (literary-physics)	34.40
Modifier-Head	引文 (cited-paper)	18.72
Verb-Object	发文 (publish-paper)	14.66
Adverb-Verb	博引 (widely-cite)	9.09
Single Morpheme	葡萄 (grape)	5.81

Table 1: Top 5 word-formations and examples. % denotes the instance percentage.

BERT

- Formation: sentence-level binary classification task:

triplet (w,c,d) with word, context definition is matched or not. [GlossBert, Huang et al, 2021]

$$p(y \mid w, c, d) = f(\mathbf{h}),$$
$$\hat{y} = \arg\max_{y} p(y \mid w, c, d).$$

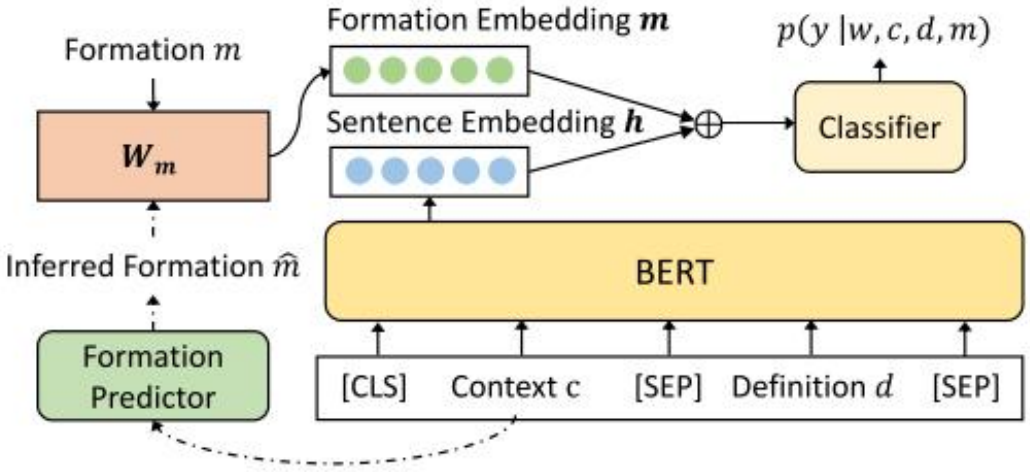


Figure 3: Illustration of the proposed FormBERT with FP. The dashed line indicates that, during inference, the inferred formation based on the context will be exploited to generalize to scenarios without formation.

FormBERT

- To learn a formation embedding m* for each formation type.

$$p(y \mid w, c, d, m^*) = f(\mathbf{h} + \mathbf{m}^*).$$

- An auxiliary formation prediction task

$$p(m \mid w, c) = g(w, c),$$
$$\hat{m} = \arg\max_m p(m \mid w, c),$$
$$\mathcal{L}_{\text{fp}} = -\log p(m^* \mid w, c).$$

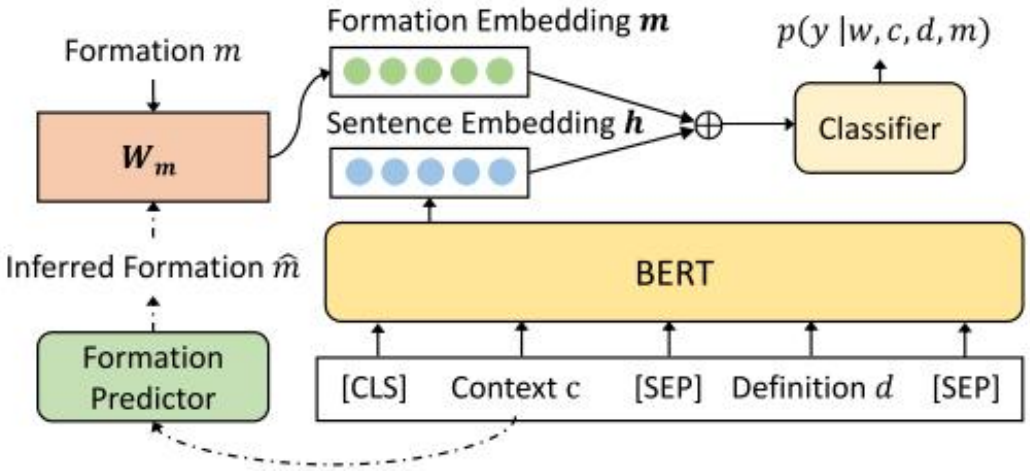


Figure 3: Illustration of the proposed FormBERT with FP. The dashed line indicates that, during inference, the inferred formation based on the context will be exploited to generalize to scenarios without formation.

Experiment - Dataset

Split	#Words	#Senses	#Entries	Context Length	Definition Length
Train	6,989	18,615	95,698	52.32	8.88
Valid	4,004	7,368	12,500	52.45	8.92
Test	3,930	7,307	12,500	52.45	8.83

Table 2: Statistics of FiCLS. The length is calculated as the average number of Chinese characters.

[NOTE] I have received the sense-annotated corpus but without CCD dictionary.

Results

(1) Word formation rule plays a role

(2) FormBERT w/ FP still performs comparable results.

[NOTE] BEM outperformed GlossBERT in the English setting.

(3) ADV is harder (not in wordnet-based corpus)

Method	Valid	Test				
		Noun	Verb	Adj.	Adv.	All
MFS	34.39	35.23	34.49	33.25	36.65	34.99
BERT	71.21	74.68	71.10	72.05	64.29	71.78
GLU	71.24	74.80	70.89	71.60	63.79	71.65
GlossBERT	84.55	82.94	81.95	82.59	81.88	84.51
BEM	72.06	73.32	72.58	74.64	66.22	72.17
FormBERT	**87.34**	**88.74**	87.07	**88.59**	81.41	87.35
FormBERT w/ FP	87.33	88.71	**87.67**	88.52	**83.07**	**87.62**

Table 3: Evaluation results (F1) on FiCLS. Best results are shown in **bold**. FormBERT w/ FP denotes FormBERT using the formation predictor without annotated word-formations.

Analysis

- Generalizability of FP: classification of WFR of 92.80.
- FormBERT in low-resource settings

Method	LFD	MFD	Zero-shot	Few-shot
GlossBERT	83.89	85.15	76.69	84.53
BEM	63.23	86.58	48.54	65.11
FormBERT	85.81	89.60	82.42	86.01
FormBERT w/ FP	**85.93**	**90.01**	**82.65**	**86.25**

Table 5: Evaluation results (F1) on the MFD, LFD, zero-shot and few-shot subsets of the test set.

Conclusion

- To enhance Chinese WSD with word-formation knowledge
- A large-scale formation-informed dataset
- FormBERT

Thinking…

- How does morphology help WSD?
- Does POS information matter much for Chinese WSD?

Reference

- [Raganato et al, 2017] Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison.
- [Miller 1998] Miller, G. A.; Leacock, C.; Tengi, R.; and Bunker, R. T. 1993. A semantic concordance. In Proc. of Human Language Technology.
- [Navigli and Ponzetto 2012] BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence.
- [GlossBert, Huang et al, 2021] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In EMNLP- IJCNLP, pages 3509–3514.

清華大學
Tsinghua University

Q & A

THANK YOU

Performance Summary

Type	Model	SE2	SE3	SE07	SE13	SE15	ALL_m	ALL_M*	N	V	A	ADV	Source
KB	SREF_KB	72.7	71.5	61.5	76.4	79.5	73.5		78.5	56.6	79.0	76.9	
KB	SyntagRank	71.6	72.0	59.3	72.2	75.8	71.7	64.1					
KB	WSDG*	68.7	68.3	58.9	66.4	70.7	67.7		71.1	51.9	75.4	80.9	
Corpus	SREF_sup	78.6	76.6	72.1	78.0	80.5	77.8		80.6	66.5	82.6	84.4	
Corpus+relation	MLWSD*	78.4	77.8	72.2	76.7	78.2	77.6		80.1	67.0	80.5	86.2	Conia & Navigli EACL'21
Corpus+relation+gloss+example	MLWSD+gloss/example	80.4	77.8	76.2**	81.8	83.3	80.2		82.9	70.3	83.4	85.5	
C+G	ESC	81.7	77.8	76.3	82.2	83.2	80.7		83.9	69.3	83.8	86.7	ESC, NAACL'21
C+G	GEN	77.8	73.7	68.8	78.3	77.6	76.3	70.7	79.8	63.3	80.1	84.7	GEN, EMNLP'20
Corpus+gloss	ARES	78.0	77.1	71.0	77.3	83.2	77/9	72.9	80.6	68.3	80.5	83.5	ARES, EMNLP'20
Corpus+gloss	BEM	79.4	77.4	74.5**	79.7	81.7	79.0	73.9	81.4	68.5	83.0	87.9	BEM, ACL'20
C+G+Relation	EWISER	78.9	78.4	71.0	78.9	79.3	78.3	73.3	81.7	66.3	81.2	85.8	EWISER, ACL'20
C+G+R+Example	EWISER_2	80.8	79.0	75.2	80.7	81.8	80.1		82.9	69.4	83.6	87.3	
C+G+Relation	EWISE, ACL'19	73.8	71.1	67.3**	69.4	74.5	71.8		74.0	60.2	78.0	82.1	EWISE, ACL'19
C+Relation+WNGC*	SVC, GWNC'19	79.4	78.1	71.4	77.8	81.4	78.5		80	68.6	82.8	85.5	SVC, GWNC'19
C	GLU	75.5	73.6	68.1**	71.1	76.2	74.1		-	-	-	-	GLU, EMNLP'19
C+G	GlossBert	77.7	75.2	72.5	76.1	80.4	77.0	71.3	79.8	67.1	79.6	87.4	EMNLP'19

WSDG*: one result from three in the paper.
MLWSD*: without formal model name provided
C+Relation+WNGC*: Note that here the model is different from that in IJCAI survey, which uses "SemCor, hypernyms".
Value** for SE07: SE07 is used as a development set.
ALL_M*: Macro-F1, source from (Nibbling)

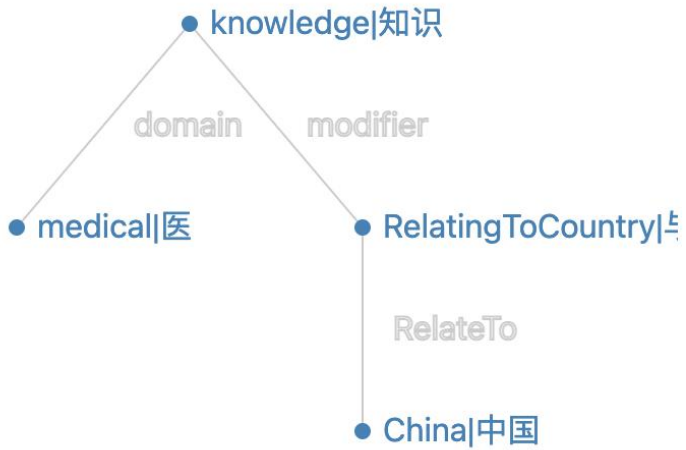
HowNet-based WSD

- Based on SemEval-2007 task 5
- 2969 instances for 36 target polysemous words (17 nouns and 19 verbs)

基于义原的定义

{knowledge|知识:domain={medical|医},modifier={RelatingToCountry|与特定国家相关:RelateTo="China|中国"}}

义原树演示



{'context': ['围绕', '《', '指导', '意见', '》', '的', '贯彻', '落实', ',', '结合', '中医药', '工作', '的', '实际', ',', '各地', '要', '加大', '力度', ',', '积极', '而', '稳妥', '地', '推进', '<target>', '医疗', '机构', '改革', '。'], 'part-of-speech': ['v', 'w', 'vn', 'n', 'w', 'u', 'v', 'v', 'w', 'v', 'n', 'vn', 'u', 'n', 'w', 'r', 'v', 'v', 'n', 'w', 'a', 'c', 'a', 'u', 'v', 'n', 'n', 'n', 'vn', 'w'], 'target_word': '中医', 'target_position': 25, 'target_word_pos': 'n', 'sense': {'知识', '中国', '与特定国家相关', '医'}}

https://github.com/thunlp/SememeWSD

HowNet-based WSD

- SemCo (Yang et al., 2001). This method utilizes the statistics on the co-occurrence of sememes of the target polysemous word and context to conduct WSD.
- SemEmbed (Tang et al., 2015). This method first learns sememe embeddings and further obtains sense embeddings, and then employs the embedding similarity between senses of the target word and the context for disambiguation.
- Dense (Ustalov et al., 2018). This model is originally designed for WordNet-based WSD, which first obtains sense embeddings from the word embeddings of the corresponding senses' synonyms and then selects the sense that has the closest embedding similarity with the context. In HowNet-based WSD, we regard the words whose one sense has the same sememes as the target sense as the synonyms.

Model	Nouns		Verbs		All	
	micro-F1	macro-F1	micro-F1	macro-F1	micro-F1	macro-F1
SemCo	38.11	26.53	27.29	21.03	31.47	23.63
SemEmbed	47.38	33.13	32.51	27.85	38.25	30.34
Dense	52.36	39.06	35.14	33.01	41.79	35.86
Random	37.24	34.83	20.54	20.72	26.98	27.38
Ours	**53.76**	**41.71**	**52.50**	**48.02**	**52.98**	**45.04**

Table 1: WSD results (%F1 scores) of all the models.

Chinese-translated WSD

- BOW (Sinica Bilingual Ontological Wordnet)
- SEW (Southeast University WordNet)
- CWN (Taiwan University WordNet) – test data
- COW (Chinese open wordnet)