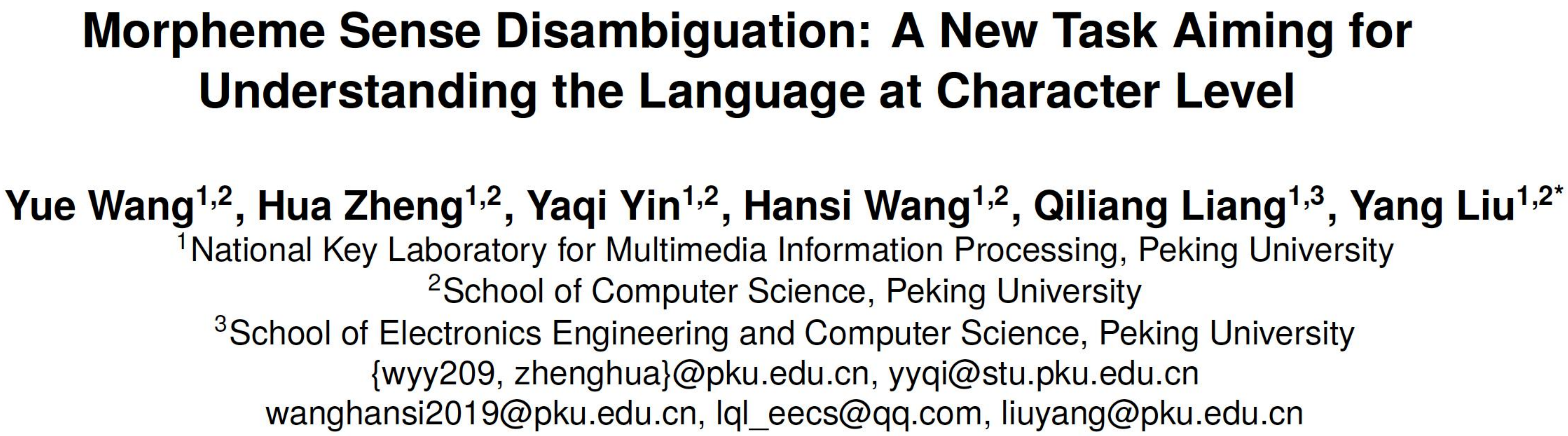
Paper Sharing

Zhu Liu

2024.05.30

Paper Overview

- Yue Wang, Hua Zheng, Yaqi Yin, Hansi Wang, Qiliang Liang, and Yang Liu. 2024. Morpheme Sense Disambiguation: A New Task Aiming for Understanding the Language at Character Level. (LREC-COLING 2024), pages 11605–11618, Torino, Italia. ELRA and ICCL.
- Simone Conia and Roberto Navigli. 2022. Probing for Predicate Argument Structures in Pretrained Language Models. ACL 2022, pages 4622–4632, Dublin, Ireland. Association for Computational Linguistics.



Morpheme Sense Disambiguation: A New Task Aiming for Understanding the Language at Character Level

Yue Wang[1,2], Hua Zheng[1,2], Yaqi Yin[1,2], Hansi Wang[1,2], Qiliang Liang[1,3], Yang Liu[1,2*]
[1]National Key Laboratory for Multimedia Information Processing, Peking University
[2]School of Computer Science, Peking University
[3]School of Electronics Engineering and Computer Science, Peking University
{wyy209, zhenghua}@pku.edu.cn, yyqi@stu.pku.edu.cn
wanghansi2019@pku.edu.cn, lql_eecs@qq.com, liuyang@pku.edu.cn

LREC-Coling 2024

https://github.com/COOLPKU/MSD_task

The Lab

- Yang Liu (刘扬), Associate Professor in Institute of CL
- Chinese Lexical Semantics: Word Structure and Morpheme
- Leveraging Word-Formation Knowledge for Chinese Word Sense Disambiguation (EMNLP'21 Findings)
- Construction of Chinese Semantic Word-Formation and its Computing Applications (CCL'22)
- Decompose, Fuse and Generate: A Formation-Informed Method for Chinese Definition Generation (NAACL'21)
- LREC-Coling'24
- Chinese Morpheme-informed Evaluation of Large Language Models
- Morpheme Sense Disambiguation: A New Task Aiming for Understanding the Language at Character Level

LREC-Coling’24

- 20-25 May, 2024; Torino (Italia)
- Papers related to *Word Sense Disambiguation:*

1. ContrastWSD: Enhancing Metaphor Detection with Word Sense Disambiguation Following the Metaphor Identification Procedure

2. Labeling Results of Topic Models: Word Sense Disambiguation as Key Method for Automatic Topic Labeling with GermaNet

3. Word Sense Disambiguation as a Game of Neurosymbolic Darts

4. Language Pivoting from Parallel Corpora for Word Sense Disambiguation of

5. Historical Languages: A Case Study on Latin

6. Sense of the Day: Short Timeframe Temporal-Aware Word Sense Disambiguation

7. Ukrainian Visual Word Sense Disambiguation Benchmark

Other interesting papers

- A Construction Grammar Corpus of Varying Schematicity: A Dataset for the Evaluation of Abstractions in Language Models
- A Canonical Form for Flexible Multiword Expressions
- Analyzing the Understanding of Morphologically Complex Words in Large Language Models
- Annotating Chinese Word Senses with English WordNet: A Practice on OntoNotes Chinese Sense Inventories
- Are Large Language Models Good at Lexical Semantics? A Case of Taxonomy Learning
- A Study on How Attention Scores in the BERT Model Are Aware of Lexical Categories in Syntactic and Semantic Tasks on the GLUE Benchmark
- Building a Broad Infrastructure for Uniform Meaning Representations
- Has It All Been Solved? Open NLP Research Questions Not Solved by Large Language Models

Two keynotes

https://lrec-coling-2024.org/keynote-and-invited-speakers/

- Large language models and human cognition (by Roger Levy, MIT’s Computational Psycholinguistics Laboratory)
- Knowledge in LLM Era: Actuality, Challenge, and Potentiality (by Juanzi Li, THU)

Background (Challenges)

- Words as a unit of tasks

  Tasks like: intrinsic (word sense similarity) and extrinsic (WSD)

  Challenges: Limited coverage of the word inventory

- Characters as a unit of computation (tokenization) for Chinese

  Word-based tokenization faces the issues from vast lexicon and out-of-dictionary (OOV) words

Background (Solutions)

- Sememe Prediction

  A manually curated set of atomic semantics used to define words

  Hownet (with around 2800 sememes): 减肥->变形状+瘦

  Cons: Subjective and Uncertain

- Morpheme-based Tasks

  In Chinese, morphemes are the smallest semantic and sound-bearing units.

  Pros: Objective, natural, effective, and easier

  Other related tasks has proven the effectiveness of morpheme features

Related Work

- Word Sense Disambiguation

Incorporating definitional, relational, formational, conceptual knowledges

- Sememe Prediction

Hownet sememes, effective in tasks: Word Similarity, WSD, Sentiment Analysis

- Chinese Morpheme-Related Resources

Four resources, including morphemes and word-formation

- Chinese Morpheme-Enhanced Methods

Word Embedding, accuracy, OOV words

Contributions

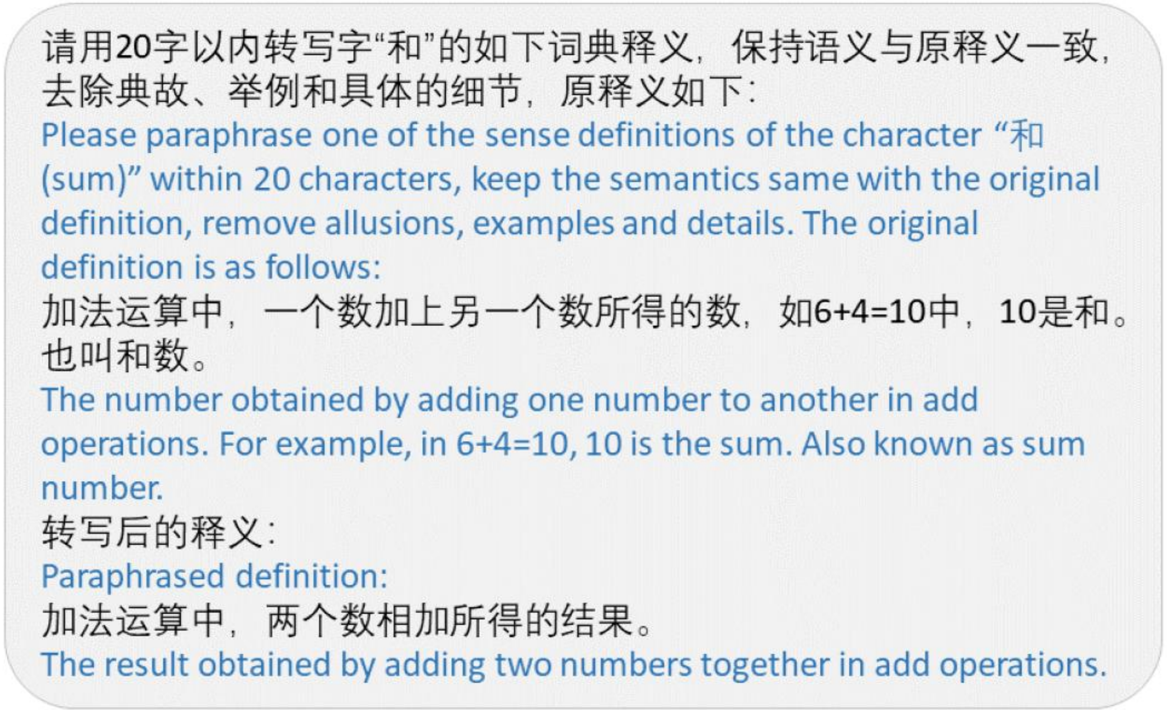
- To propose a task of Morpheme sense disambiguation (MSD) and build two annotated datasets.
- To implement two baseline models for MSD
- To apply Morpheme senses into other downstream tasks

Resources

- Morpheme inventory
- Morpheme-Annotated Datasets (MorTxt)
- Morpheme-Annotated Datasets (MorWord)

Morpheme Inventory

- Objective: from morphemes to POS and senses
  - To extract morpheme sense and PoS from CCD (现代汉语词典)
  - ChatGPT further paraphrase and simplify the senses
  - Three mother-tongue reviewers manually check the senses
  - Multi-character morphemes, e.g. 葡萄



请用20字以内转写字“和”的如下词典释义，保持语义与原释义一致，去除典故、举例和具体的细节，原释义如下：

Please paraphrase one of the sense definitions of the character “和 (sum)” within 20 characters, keep the semantics same with the original definition, remove allusions, examples and details. The original definition is as follows:

加法运算中，一个数加上另一个数所得的数，如6+4=10中，10是和。也叫和数。

The number obtained by adding one number to another in add operations. For example, in 6+4=10, 10 is the sum. Also known as sum number.

转写后的释义：

Paraphrased definition:

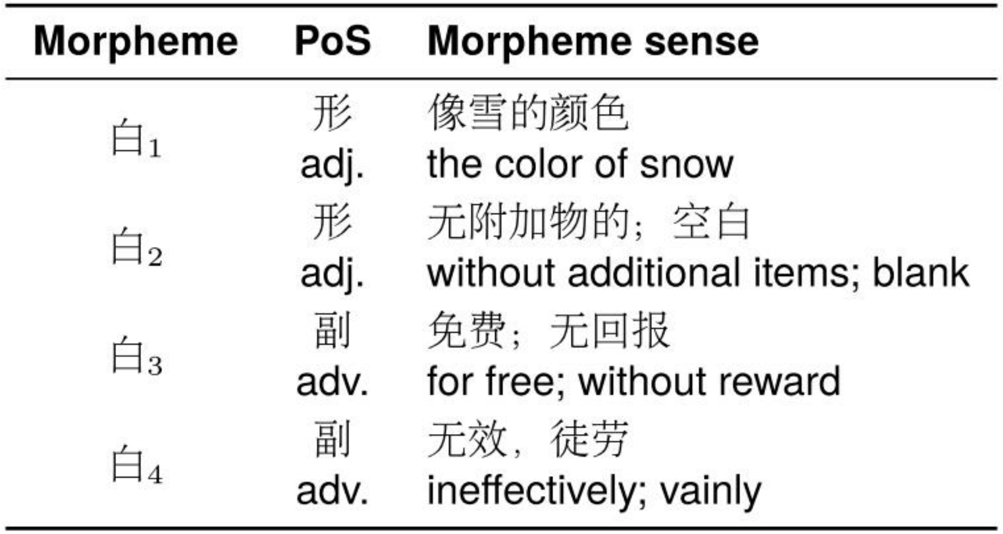
加法运算中，两个数相加所得的结果。

The result obtained by adding two numbers together in add operations.

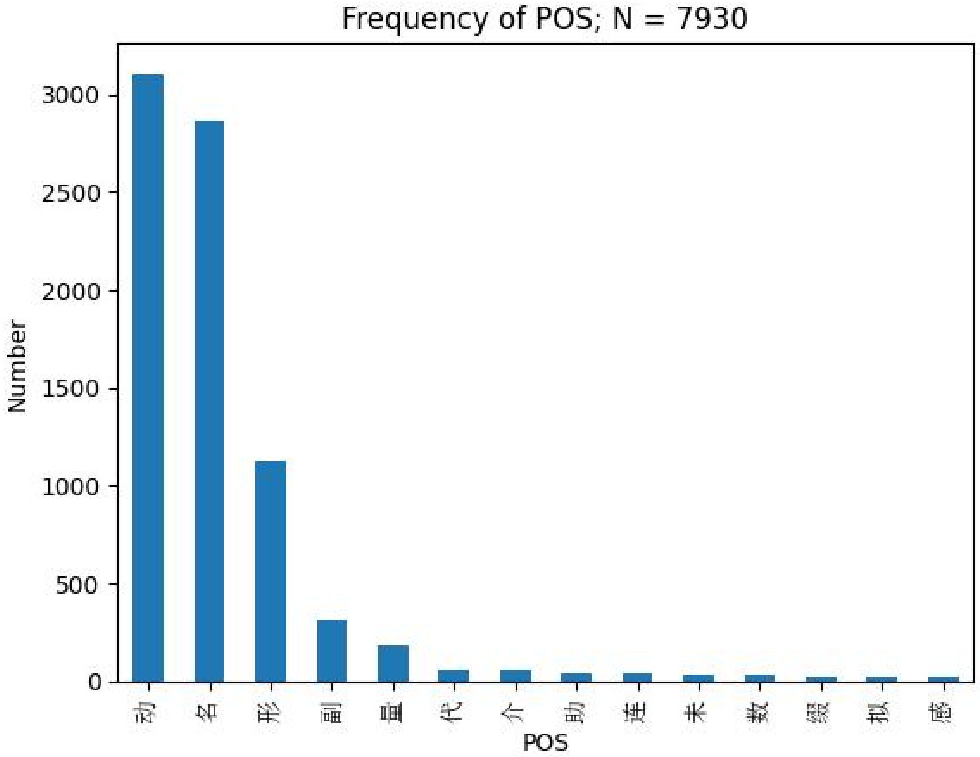
Figure 2: A sample prompt and its result for paraphrasing one of the sense definitions of ’和’, which is "和$_{11}$" in CCD.

Statistics of MorInv

- 20,586 morphemes for 8516 characters.
- Only 7930 morphemes are listed in the CODE

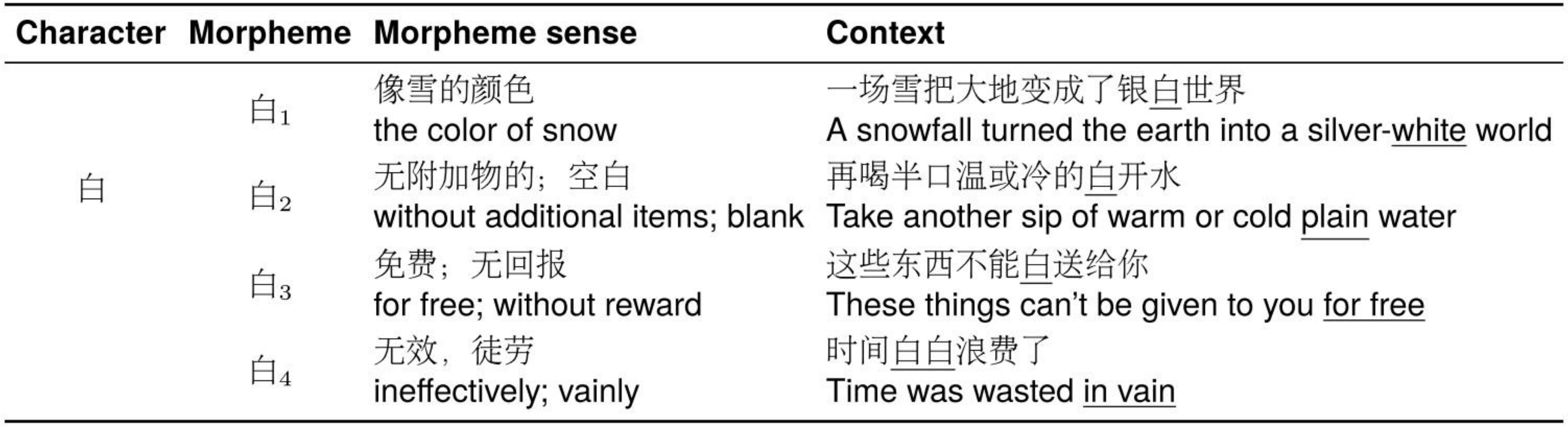


Morpheme	PoS	Morpheme sense
白$_1$	形 adj.	像雪的颜色 the color of snow
白$_2$	形 adj.	无附加物的；空白 without additional items; blank
白$_3$	副 adv.	免费；无回报 for free; without reward
白$_4$	副 adv.	无效，徒劳 ineffectively; vainly



MorTxt

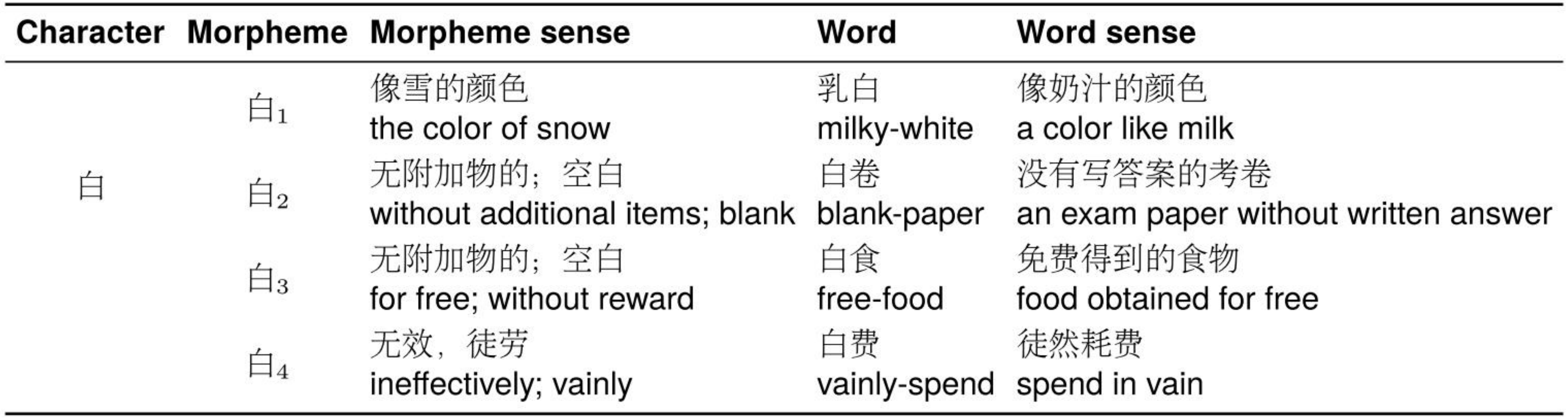
- MorTxt contains (1) target character; (2) context; (3) morpheme and its sense.
- Sources: Two previous work by his lab [Zheng et al. 2021; 2021a]
- Final: 27080 entries, totaling 10567 morphemes for 3240 polysemous characters



Character	Morpheme	Morpheme sense	Context
白	白$_1$	像雪的颜色 the color of snow	一场雪把大地变成了银白世界 A snowfall turned the earth into a silver-white world
	白$_2$	无附加物的；空白 without additional items; blank	再喝半口温或冷的白开水 Take another sip of warm or cold plain water
	白$_3$	免费；无回报 for free; without reward	这些东西不能白送给你 These things can't be given to you for free
	白$_4$	无效，徒劳 ineffectively; vainly	时间白白浪费了 Time was wasted in vain

MorWrd

- Entry: (1) target character; (2) a word containing it; (3) morpheme sense; (4) word sense
- Source: [Zheng et al. 2021a]
- Final MorWord contains 98065 entries, totaling 11,874 morphemes for 4,974 polysemous characters



Character	Morpheme	Morpheme sense	Word	Word sense
白	白$_1$	像雪的颜色 the color of snow	乳白 milky-white	像奶汁的颜色 a color like milk
	白$_2$	无附加物的；空白 without additional items; blank	白卷 blank-paper	没有写答案的考卷 an exam paper without written answer
	白$_3$	无附加物的；空白 for free; without reward	白食 free-food	免费得到的食物 food obtained for free
	白$_4$	无效，徒劳 ineffectively; vainly	白费 vainly-spend	徒然耗费 spend in vain

Experiments (MorTxt)

- Two baselines: BEM and ChatGPT
- BEM: a bi-encoder model to close the representation between morpheme sense and corresponding character.

$$\mathcal{L}(c, s_i) = -\,\mathsf{sim}(r_{c_*}, r_{s^i}) + log\sum_{j=0}^{|S_c|} exp(sim(r_{c_*}, r_{s^j})).$$

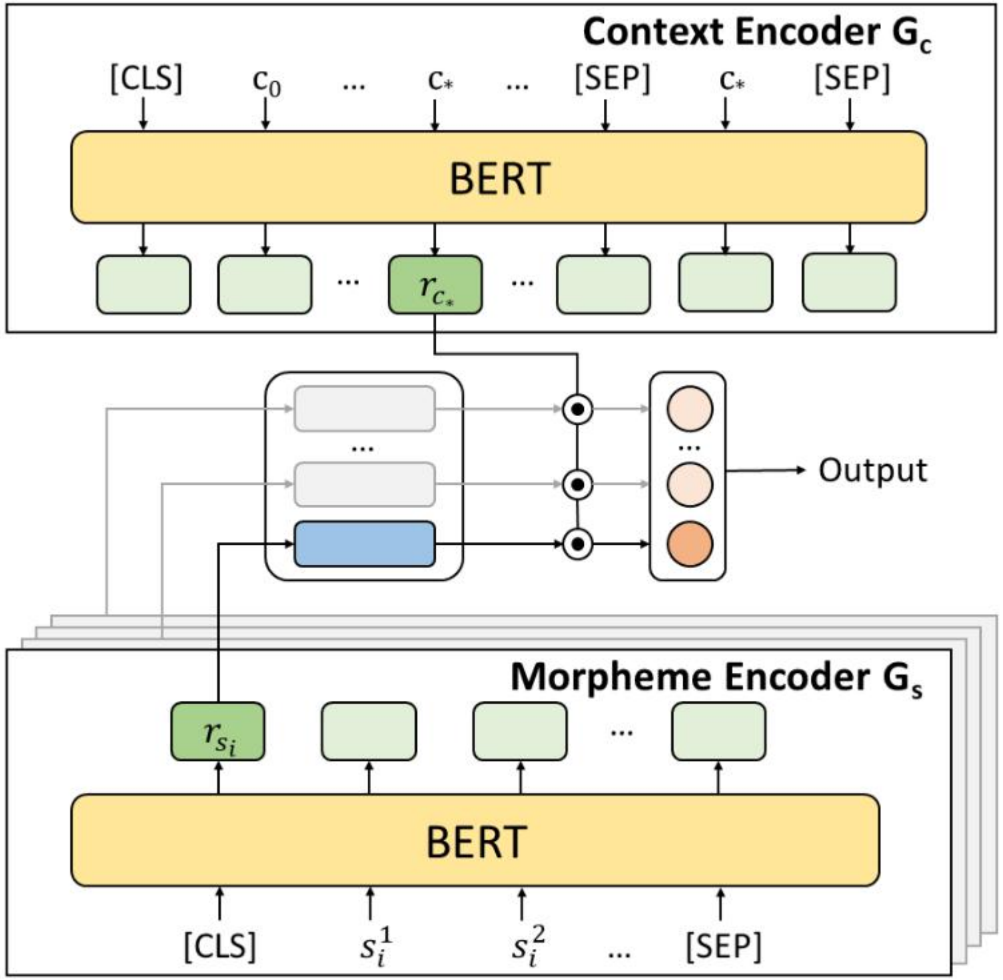
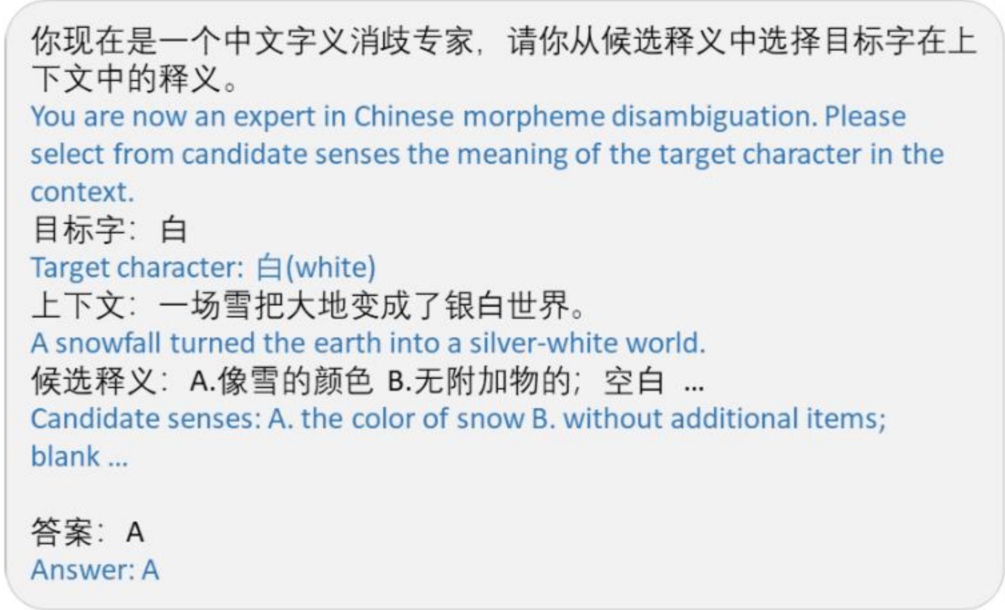


Figure 3: An Illustration of the BEM baseline of in-text MSD.

Experiments

- ChatGPT
- Prompt Engineering
  - best from 10 in val. dataset
- Exact & Fuzzy Matching

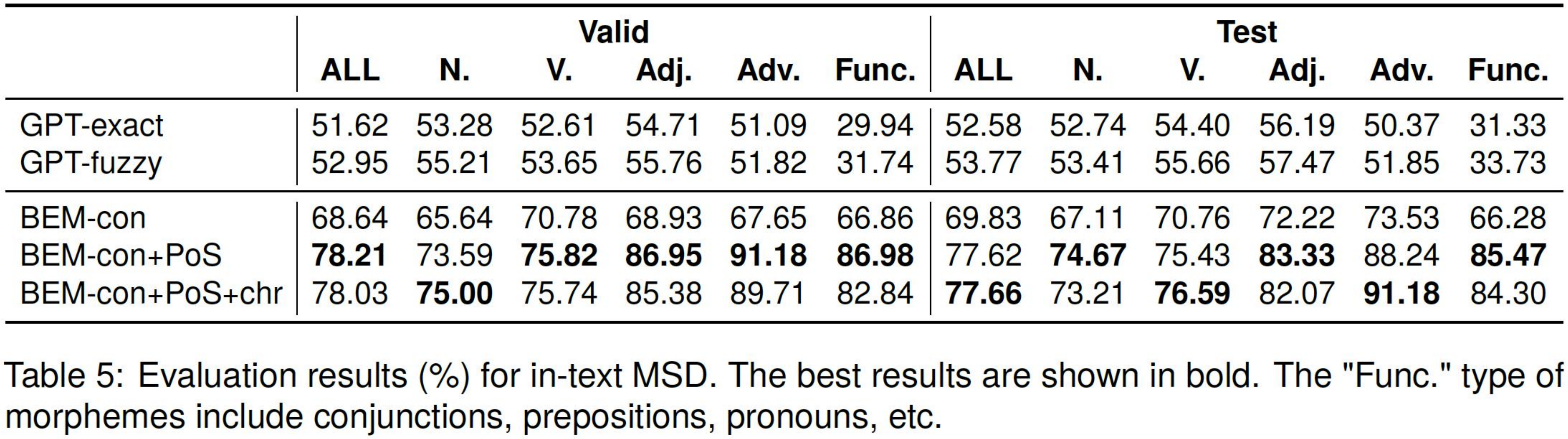


你现在是一个中文字义消歧专家，请你从候选释义中选择目标字在上下文中的释义。
You are now an expert in Chinese morpheme disambiguation. Please select from candidate senses the meaning of the target character in the context.
目标字：白
Target character: 白(white)
上下文：一场雪把大地变成了银白世界。
A snowfall turned the earth into a silver-white world.
候选释义：A.像雪的颜色 B.无附加物的；空白 ...
Candidate senses: A. the color of snow B. without additional items; blank ...

答案：A
Answer: A

Figure 4: A sample prompt for in-text GPT baseline.

Results and analysis (MorTxt)



	Valid						Test					
	ALL	**N.**	**V.**	**Adj.**	**Adv.**	**Func.**	**ALL**	**N.**	**V.**	**Adj.**	**Adv.**	**Func.**
GPT-exact	51.62	53.28	52.61	54.71	51.09	29.94	52.58	52.74	54.40	56.19	50.37	31.33
GPT-fuzzy	52.95	55.21	53.65	55.76	51.82	31.74	53.77	53.41	55.66	57.47	51.85	33.73
BEM-con	68.64	65.64	70.78	68.93	67.65	66.86	69.83	67.11	70.76	72.22	73.53	66.28
BEM-con+PoS	**78.21**	73.59	**75.82**	**86.95**	**91.18**	**86.98**	77.62	**74.67**	75.43	**83.33**	88.24	**85.47**
BEM-con+PoS+chr	78.03	**75.00**	75.74	85.38	89.71	82.84	**77.66**	73.21	**76.59**	82.07	**91.18**	84.30

Table 5: Evaluation results (%) for in-text MSD. The best results are shown in bold. The "Func." type of morphemes include conjunctions, prepositions, pronouns, etc.

- Best Performance
- GPT < BEM, especially in *function words*

Wrong cases

- MFS bias: most frequent sense bias (46% error)
  - 钻钱眼儿：眼1-eye 眼2-hole；The model prefers the second
  - [N] It lacks an MFS baseline

- The word is ambiguous/polysemous itself.
  - 水分：1-water content 2-exaggeration; 水 is wrongly predicted as water
  - [N] But the correct answer? [Non-compositionality of Morphemes]

- Some morpheme senses are derived from others, or mixed
  - 咖啡豆 豆3-bean；豆4-something look like a bean; 囡1-kid；囡2-little girl
  - [N] Uncertainty of morpheme senses

Experiments(MorWrd)

- BEM & ChatGPT
- Word Sense Information

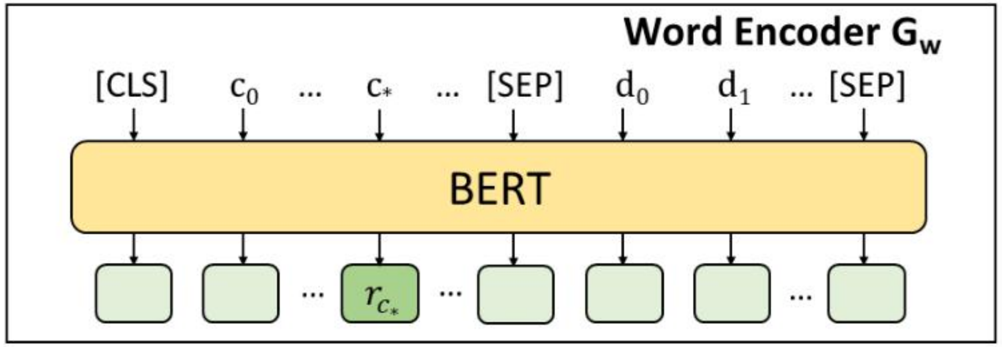
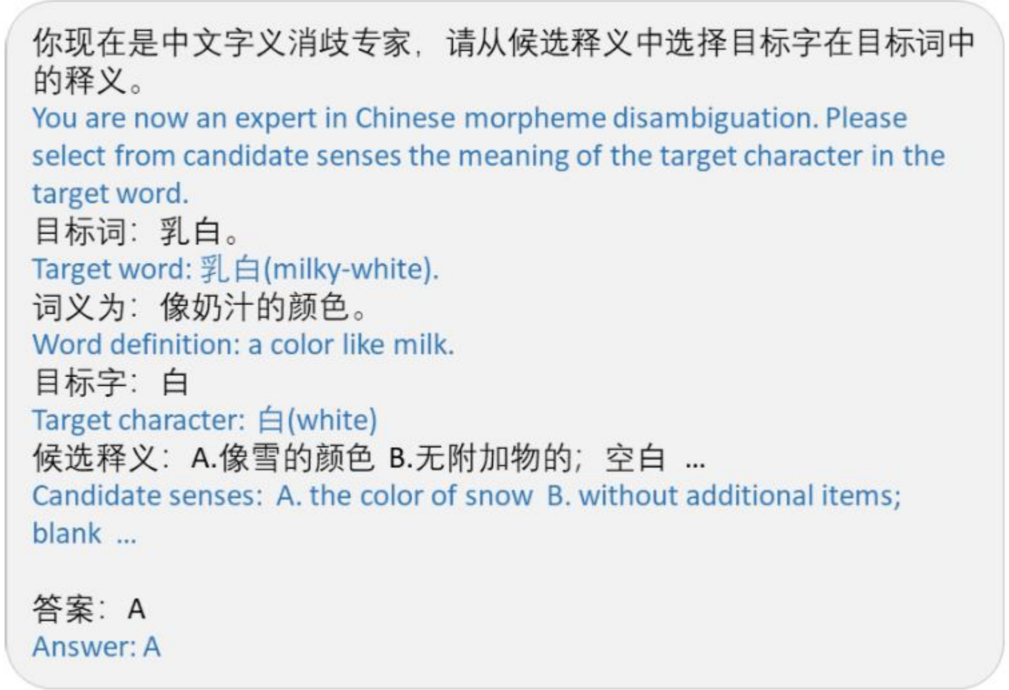


Figure 5: Word encoder G_w of the BEM baseline of in-word MSD.



你现在是中文字义消歧专家，请从候选释义中选择目标字在目标词中的释义。

You are now an expert in Chinese morpheme disambiguation. Please select from candidate senses the meaning of the target character in the target word.

目标词：乳白。

Target word: 乳白(milky-white).

词义为：像奶汁的颜色。

Word definition: a color like milk.

目标字：白

Target character: 白(white)

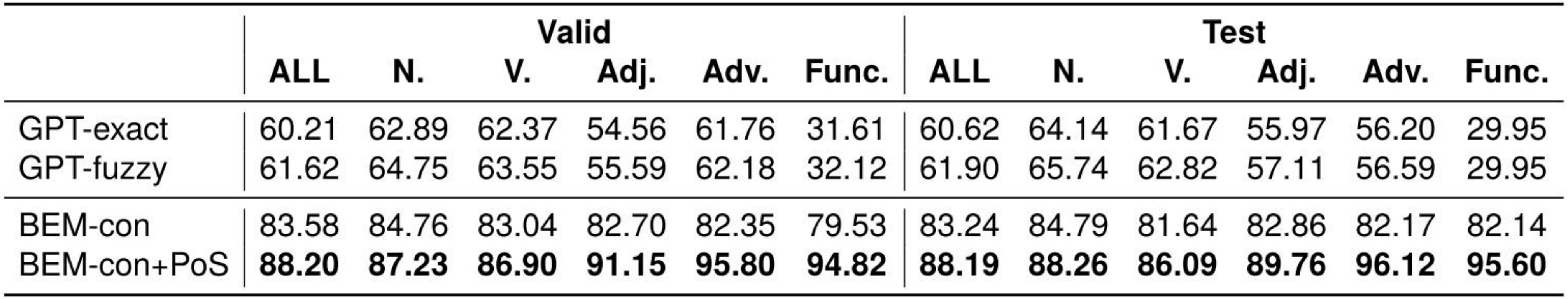
候选释义：A.像雪的颜色 B.无附加物的；空白 ...

Candidate senses: A. the color of snow B. without additional items; blank ...

答案：A

Answer: A

Results



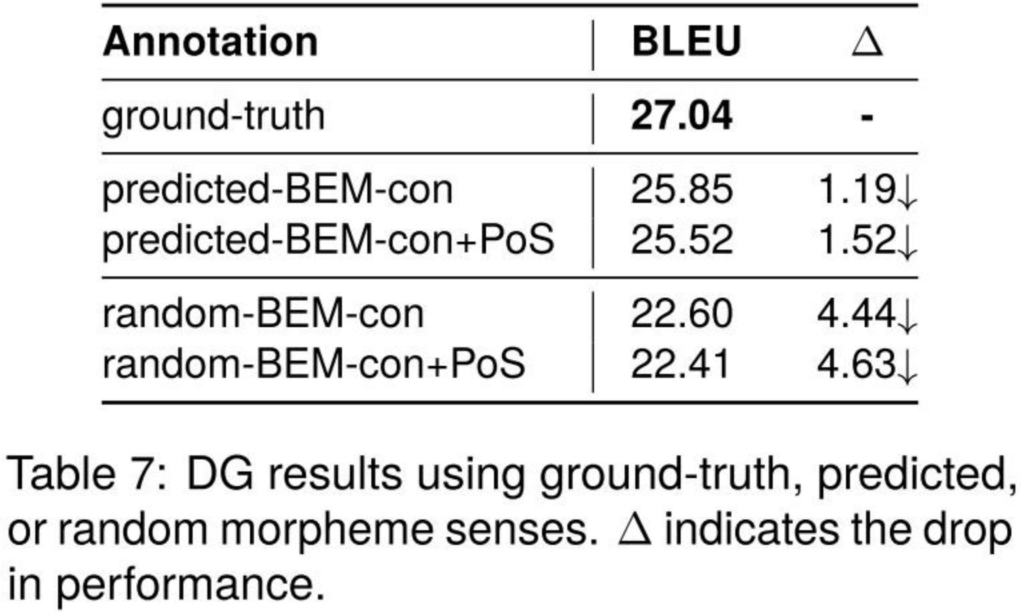
	Valid						Test					
	ALL	N.	V.	Adj.	Adv.	Func.	ALL	N.	V.	Adj.	Adv.	Func.
GPT-exact	60.21	62.89	62.37	54.56	61.76	31.61	60.62	64.14	61.67	55.97	56.20	29.95
GPT-fuzzy	61.62	64.75	63.55	55.59	62.18	32.12	61.90	65.74	62.82	57.11	56.59	29.95
BEM-con	83.58	84.76	83.04	82.70	82.35	79.53	83.24	84.79	81.64	82.86	82.17	82.14
BEM-con+PoS	**88.20**	**87.23**	**86.90**	**91.15**	**95.80**	**94.82**	**88.19**	**88.26**	**86.09**	**89.76**	**96.12**	**95.60**

Table 6: Evaluation results (%) for in-word MSD. The best results are shown in bold. The "Func." type of morphemes include conjunctions, prepositions, pronouns, etc.

- Similar trend with MorTxt
- Better than MorTxt (word sense information can help MSD)
- Better than Sememe Prediction (69.19 [Other work using word sense] vs. 83.24)

Applications - Definition Generation

- How can morpheme information help DG
- Definition Generation: generate its definition given a morpheme and its sense



Annotation	**BLEU**	Δ
ground-truth	**27.04**	-
predicted-BEM-con	25.85	1.19↓
predicted-BEM-con+PoS	25.52	1.52↓
random-BEM-con	22.60	4.44↓
random-BEM-con+PoS	22.41	4.63↓

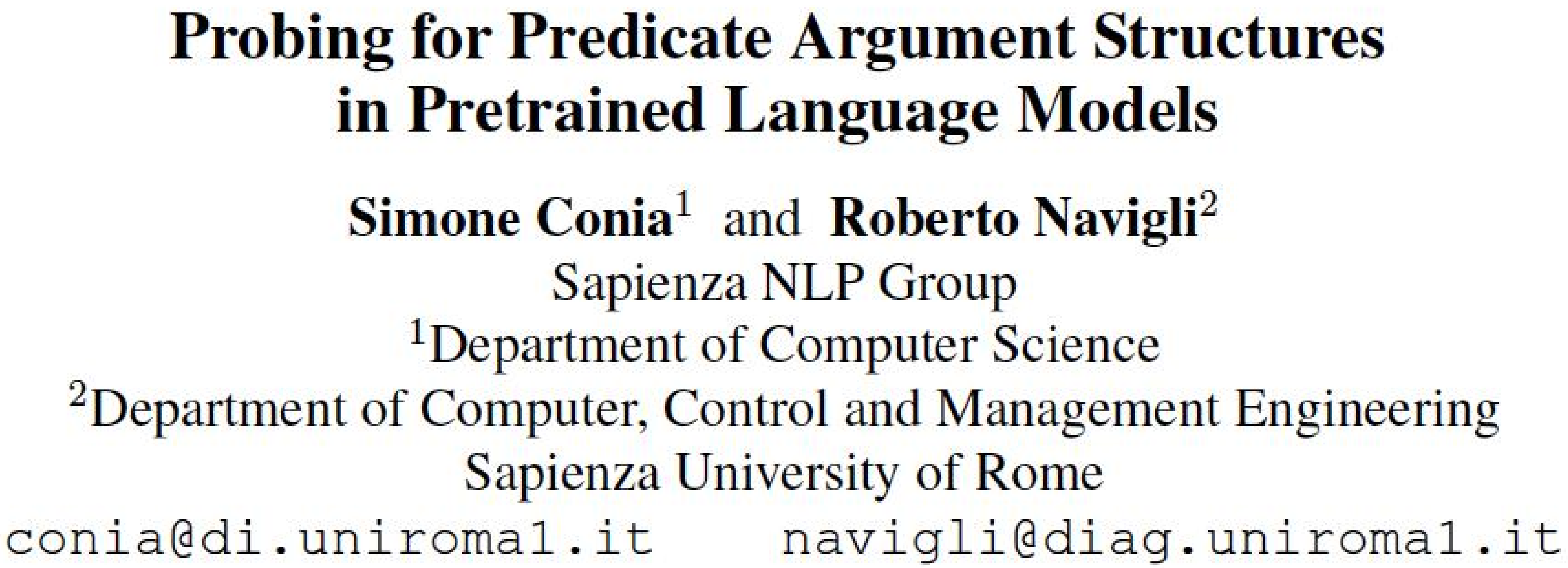
Table 7: DG results using ground-truth, predicted, or random morpheme senses. Δ indicates the drop in performance.

Conclusion

- To propose a new task of MSD with typical subtasks: in-text and in-word.
- Two baseline models for evaluation and application

- Morpheme senses are more natural and necessary for Chinese, but...
  - Different levels of compositionality. 白人 - 白酒 - 白菜/白茶？
  - More polysemous than a word. 家：语言学家；酒家；阴谋家
  - More sources of uncertainty: Windows can be limited into a word, but may be not enough. 山脚

Reference



Probing for Predicate Argument Structures in Pretrained Language Models

Simone Conia[1] and **Roberto Navigli**[2]
Sapienza NLP Group
[1]Department of Computer Science
[2]Department of Computer, Control and Management Engineering
Sapienza University of Rome
`conia@di.uniroma1.it` `navigli@diag.uniroma1.it`

ACL 2022 Cited by 20

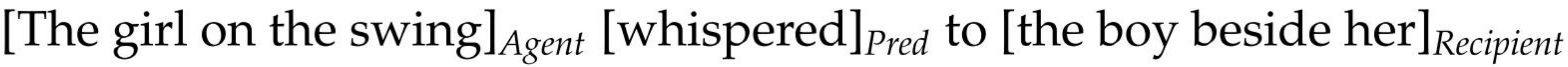
https://github.com/SapienzaNLP/srl-pas-probing

Sapienza NLP - mainly by S. Conia

- Exploring Non-verbal Predicates in Semantic Role Labeling: Challenging and Opportunities (Findings ACL 2023)
- SRL4E - Semantic Role Labeling for Emotions: A Unified Evaluation Framework (ACL'22)
- Semantic Role Labeling Meets Definition Modeling: Using Natual Language to Describe Predicate-Argument Structures (EMNLP'22)
- UniteD-SRL: A Unified Dataset for Span- and Dependency-Based Multilingual and Cross-Lingual Semantic Role Labeling (EMNLP'21)
- InVeRo-XL: Making Cross-Lingual Semantic Role Labeling Accessible with Intelligible Verbs and Roles (EMNLP'21)
- Generating Senses and RoLes: An End-to-End Model for Dependecy- and Span-based Semantic Role Labeling (IJCAI'21)
- Unifying Cross-Lingual Semantic Role Labeling with Heterogeneous Linguistic Resources (NAACL'21)
- Bridging the Gap in Multilingual Semantic Role Labeling: a Language-Agnostic Approach (COLing'20)
- InVeRo: Making Semantic Role Labeling Accessible with Intelligible Verbs and Roles (EMNLP'20)
- VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling (EMNLP'19)

Background

- Semantic Role Labeling (SRL)
  - *Who did What to Whom, Where, When and How*
  - Agent, Patient, Location,Temporal, and Manner, etc



[The girl on the swing]$_{Agent}$ [whispered]$_{Pred}$ to [the boy beside her]$_{Recipient}$

- Cross-Lingual SRL has gained impressive results when finetuned from Pretrained Language models
- Less work on whether, how and where the exact PLM encodes knowledge of SR. [KG 2020, Tenny 2019]
- [KG 2020, Tenny 2019] regards SRL as a atomic task, rather than multisteps

Contributions

- To probe PLMs for PASs (Predicate argument structures)
- Discriminate nominal and verbal PASs
- Crosslingual similarity
- Integrating knowlege into current SRL and improve the results

Related Work

- Probing pretrained language models

  BERTology, simple NNs to probe, drawbacks

- Probing techniques for SRL

  Middle layer [Tenny et al., 2019], max-pooling or weighted average [T 2020], Different linguistic ontologies?[KG, 2020]

- Recent advances in SRL

  PLM (built on top of PLMs) + GCNs/syntax ...

  Lack of the inner of PLM

Probing

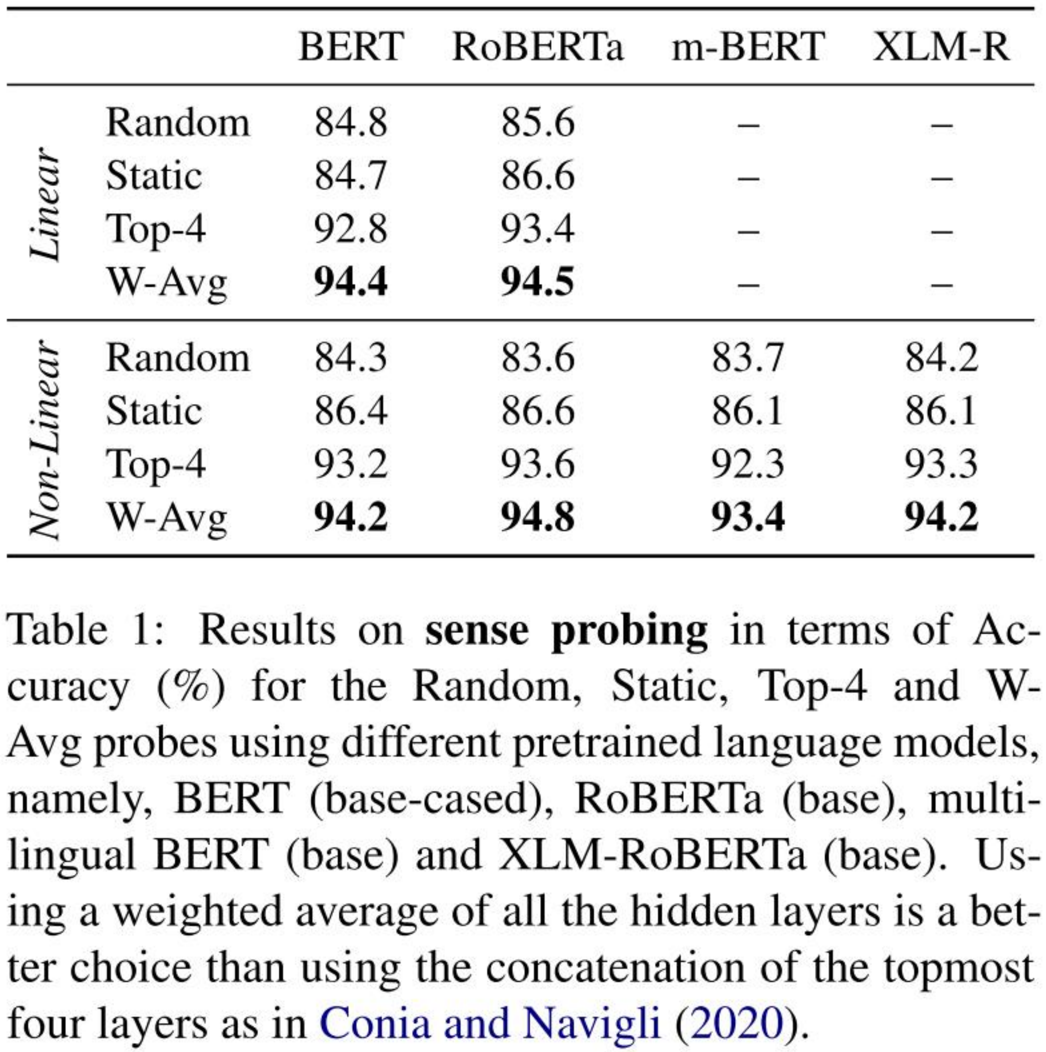
- Four different subtasks
  - Predicate identification: an action or event (not necessarily verb)
  - Predicate sense disambiguation: different meanings or frames
  - Argument identification: “semantically” related argument
  - Argument classification: what semantic roles
- predicate senses are tightly coupled to their possible rolesets
  - *He loved everything about her* with a frame experiencer_focused_emotion
  - FrameNet: {Experiencer, Content, ..., Degree}
  - English PropBank {ARG0 (lover), ARG1 (loved)}
  - VerbAtlas {Experiencer, Stimulus, ..., Cause}

Probing Tasks

- Predicate Sense Probing
  - To predict sense s from contextual representation x_p
- Roleset probing
  - To predict the semantic role set {r_1,r_2,...,r_n} from x_p
- Four choices of x_p
  - Random: a random weight of language models (control exp.)
  - Static: pre-layer before BERT
  - Top-4: concatenation of topmost 4 hidden layers (C&N, 21)
  - W-Avg: weighted average of all the hidden layers (learned weights)
- Linear and non-linear Probing on part of CoNLL-2009 shared task

Probing Results

- Sense Probing
- Random is quite well (MFS bias)
- Non-Linear > Linear
- Context > Static (even less than random)
- Full layers with learned weights are the best way

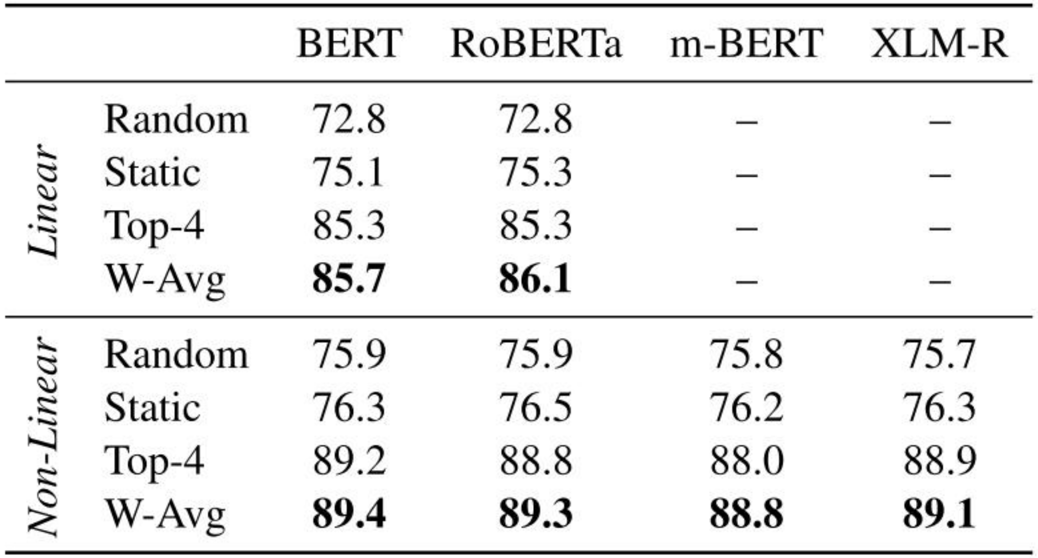


		BERT	RoBERTa	m-BERT	XLM-R
Linear	Random	84.8	85.6	–	–
	Static	84.7	86.6	–	–
	Top-4	92.8	93.4	–	–
	W-Avg	**94.4**	**94.5**	–	–
Non-Linear	Random	84.3	83.6	83.7	84.2
	Static	86.4	86.6	86.1	86.1
	Top-4	93.2	93.6	92.3	93.3
	W-Avg	**94.2**	**94.8**	**93.4**	**94.2**

Table 1: Results on **sense probing** in terms of Accuracy (%) for the Random, Static, Top-4 and W-Avg probes using different pretrained language models, namely, BERT (base-cased), RoBERTa (base), multilingual BERT (base) and XLM-RoBERTa (base). Using a weighted average of all the hidden layers is a better choice than using the concatenation of the topmost four layers as in Conia and Navigli (2020).

Probing Results

- Roleset probing
- Similar trends:
- Good Random (bias: predicates always have ARG0 and ARG1 - agentive and patientive proto-roles)
- Better Non-linear, but fails to prove its non-linearity in the model, given the control test
- Probing: learn vs. extract

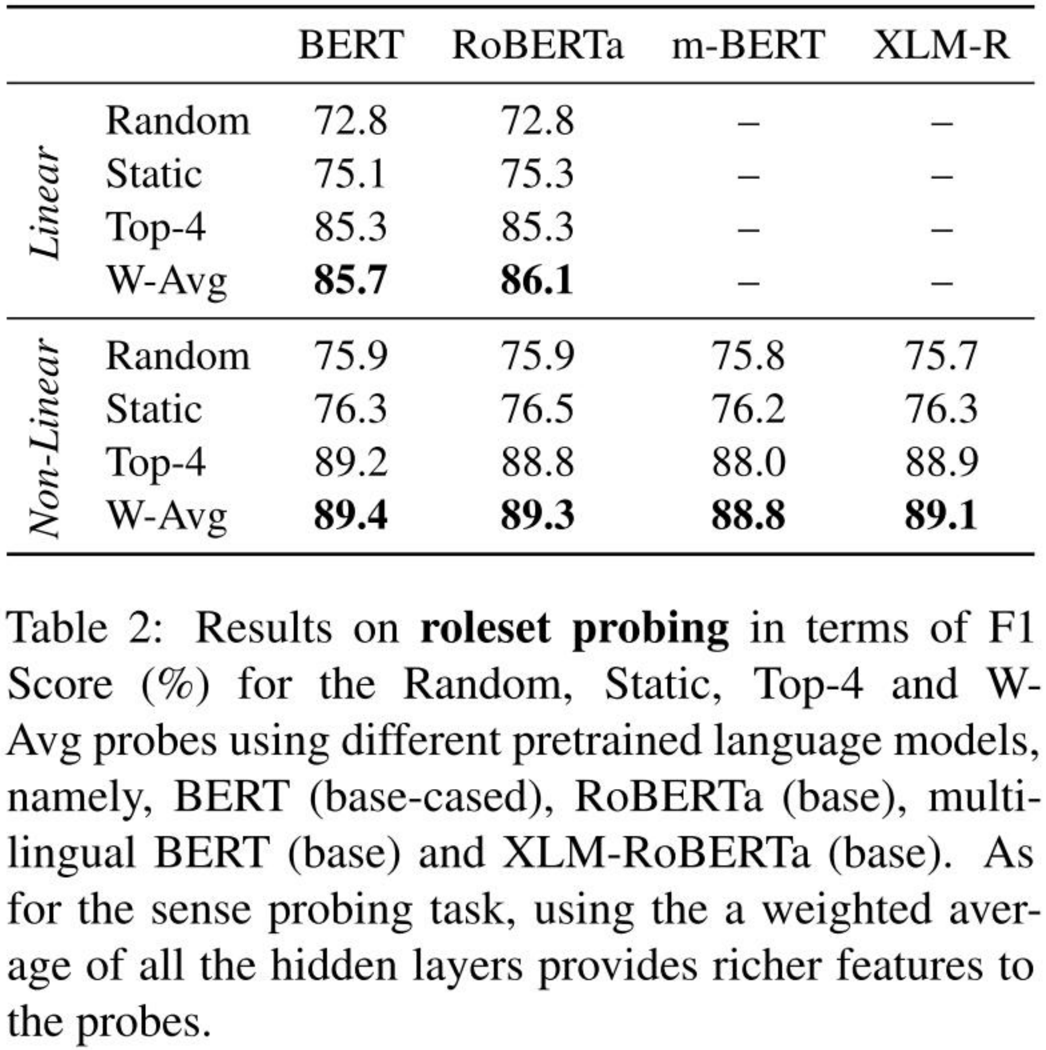


		BERT	RoBERTa	m-BERT	XLM-R
Linear	Random	72.8	72.8	–	–
	Static	75.1	75.3	–	–
	Top-4	85.3	85.3	–	–
	W-Avg	**85.7**	**86.1**	–	–
Non-Linear	Random	75.9	75.9	75.8	75.7
	Static	76.3	76.5	76.2	76.3
	Top-4	89.2	88.8	88.0	88.9
	W-Avg	**89.4**	**89.3**	**88.8**	**89.1**

Table 2: Results on **roleset probing** in terms of F1 Score (%) for the Random, Static, Top-4 and W-Avg probes using different pretrained language models, namely, BERT (base-cased), RoBERTa (base), multilingual BERT (base) and XLM-RoBERTa (base). As for the sense probing task, using the a weighted average of all the hidden layers provides richer features to the probes.

Probing Results

- W-Avg consistently outperforms Top-4
- It suggest that we need all the layers.



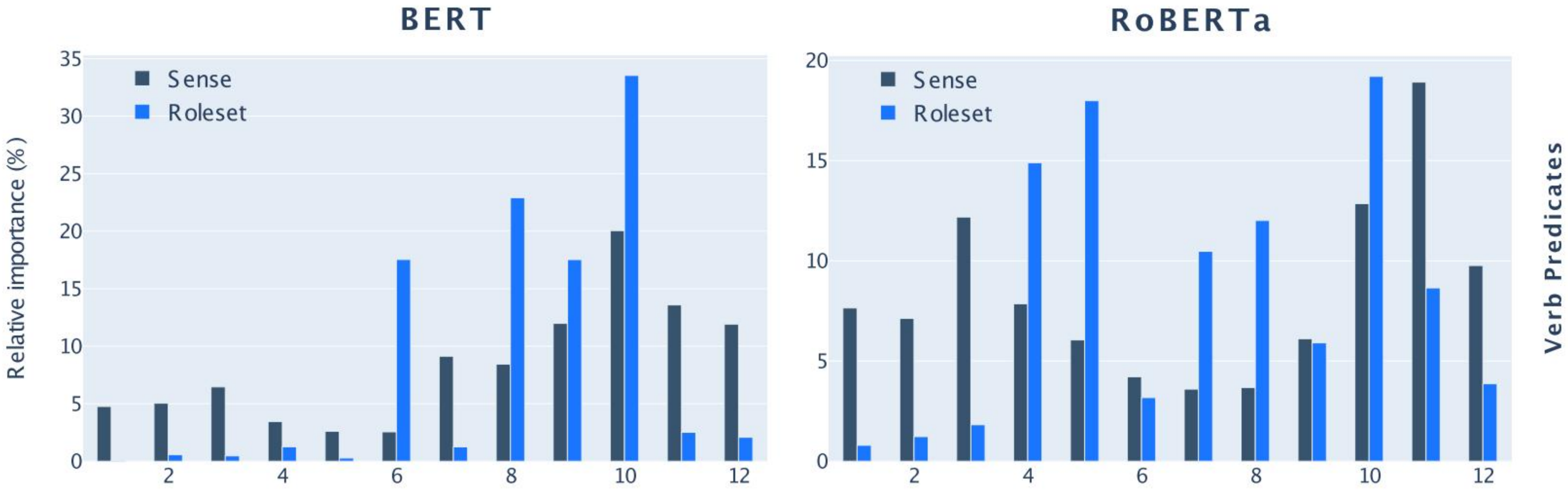
		BERT	RoBERTa	m-BERT	XLM-R
Linear	Random	72.8	72.8	–	–
	Static	75.1	75.3	–	–
	Top-4	85.3	85.3	–	–
	W-Avg	**85.7**	**86.1**	–	–
Non-Linear	Random	75.9	75.9	75.8	75.7
	Static	76.3	76.5	76.2	76.3
	Top-4	89.2	88.8	88.0	88.9
	W-Avg	**89.4**	**89.3**	**88.8**	**89.1**

Table 2: Results on **roleset probing** in terms of F1 Score (%) for the Random, Static, Top-4 and W-Avg probes using different pretrained language models, namely, BERT (base-cased), RoBERTa (base), multilingual BERT (base) and XLM-RoBERTa (base). As for the sense probing task, using the a weighted average of all the hidden layers provides richer features to the probes.

Other experiments

- On the correlation between senses and rolesets
- Do PLMs distribute sense and roleset features similarly over their inner layers?
- W-Avg has learned the weights for the layers.
- We can compare the distribution of the weights

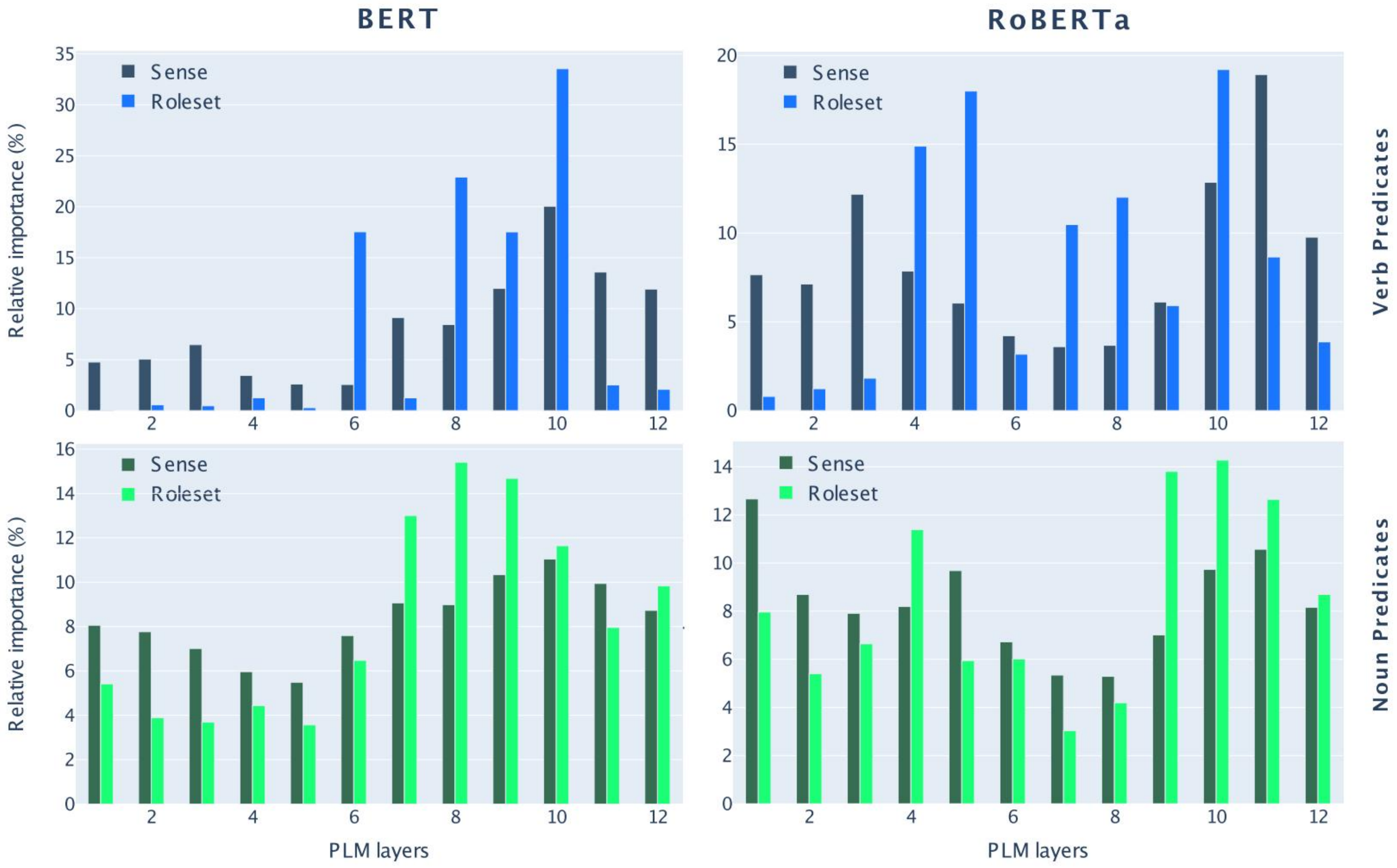
Results

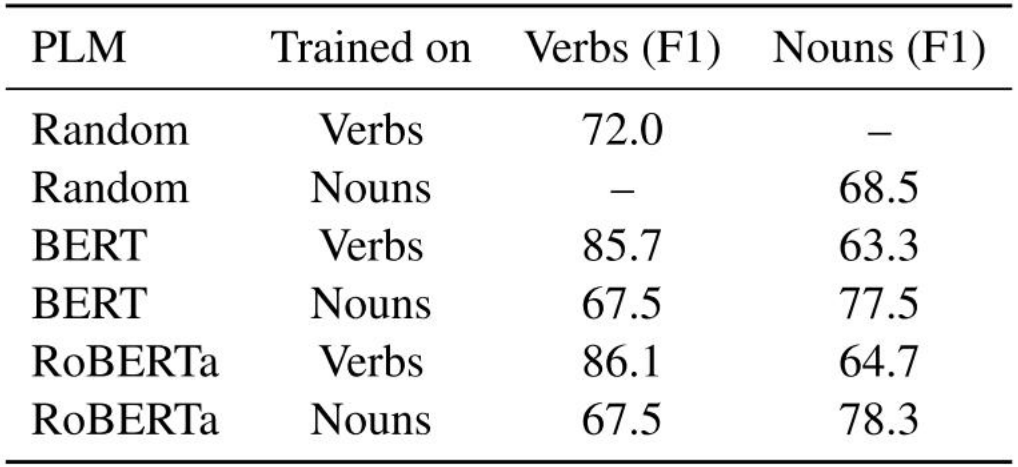


- Not similar distributions: more uniform vs. more concentrated

Verbal and Norminal Predicates

- norminal predicates, like writer, worker..
- They show different similarity trends
- Acc probing shows the difficulty to zero-shot transfer from each other.

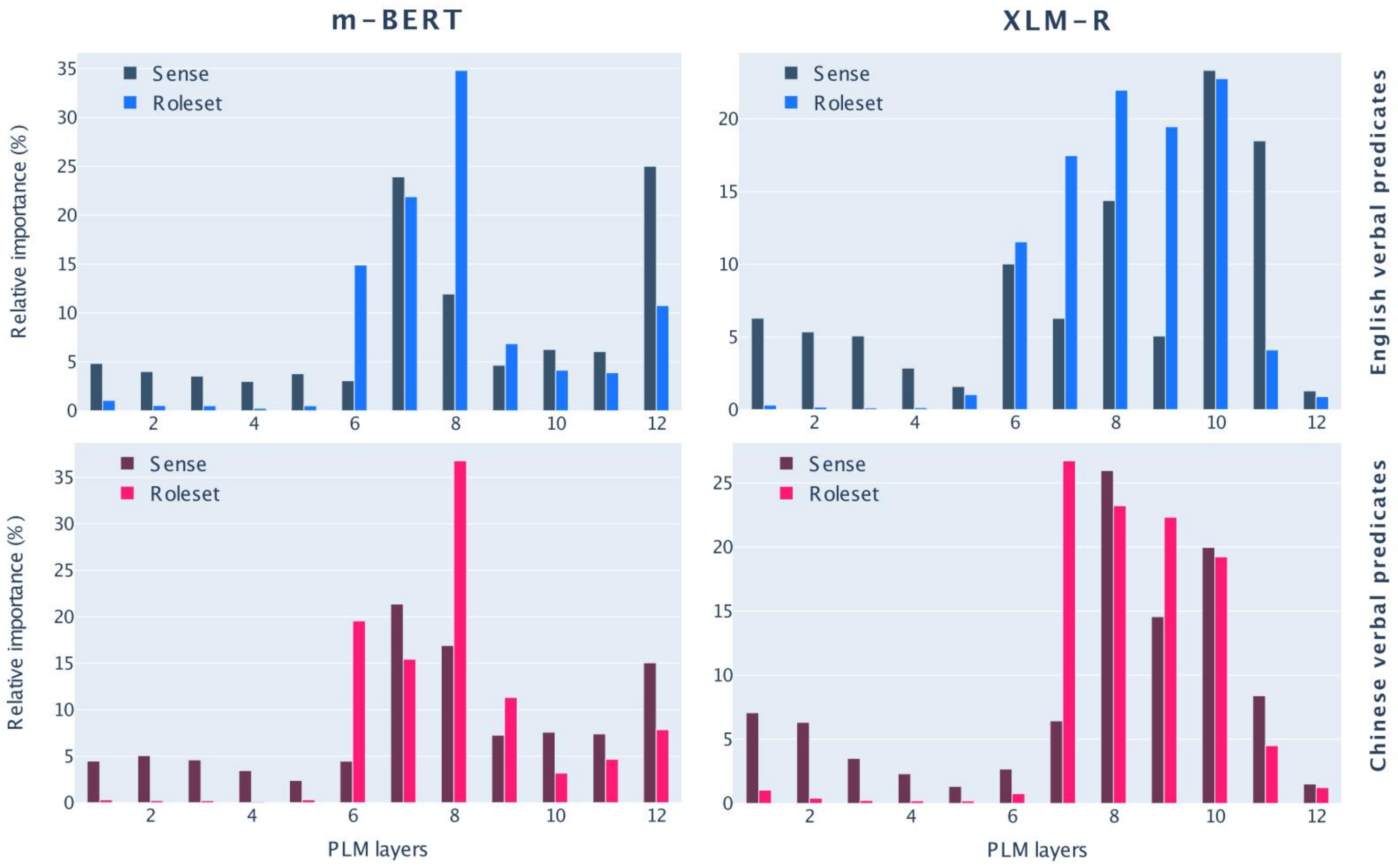




PLM	Trained on	Verbs (F1)	Nouns (F1)
Random	Verbs	72.0	–
Random	Nouns	–	68.5
BERT	Verbs	85.7	63.3
BERT	Nouns	67.5	77.5
RoBERTa	Verbs	86.1	64.7
RoBERTa	Nouns	67.5	78.3

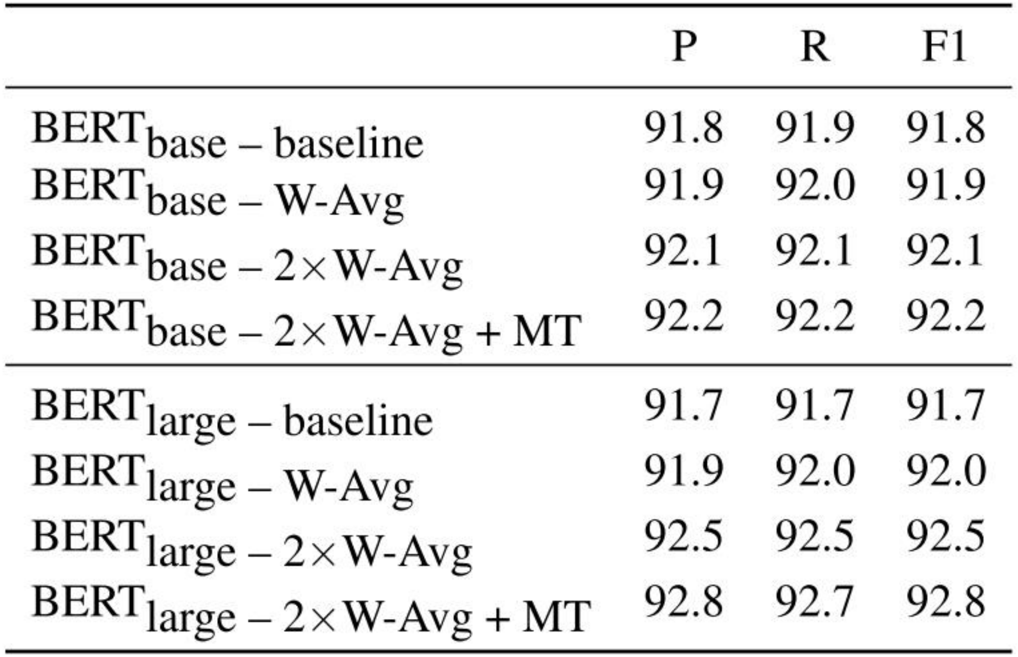
Universality across languages

- English and Chinese have similar trends



Integrating Predicate-Argument Structure Knowledge

- Task: SRL based on [Conia and Navigli 2020]
  - Enhancing SRL models
  - A shared weighted average score
  - Two different weights
  - secondary task to predict rolesets in a multi-task learning fashion.



	P	R	F1
BERT$_{\text{base}}$ – baseline	91.8	91.9	91.8
BERT$_{\text{base}}$ – W-Avg	91.9	92.0	91.9
BERT$_{\text{base}}$ – 2×W-Avg	92.1	92.1	92.1
BERT$_{\text{base}}$ – 2×W-Avg + MT	92.2	92.2	92.2
BERT$_{\text{large}}$ – baseline	91.7	91.7	91.7
BERT$_{\text{large}}$ – W-Avg	91.9	92.0	92.0
BERT$_{\text{large}}$ – 2×W-Avg	92.5	92.5	92.5
BERT$_{\text{large}}$ – 2×W-Avg + MT	92.8	92.7	92.8

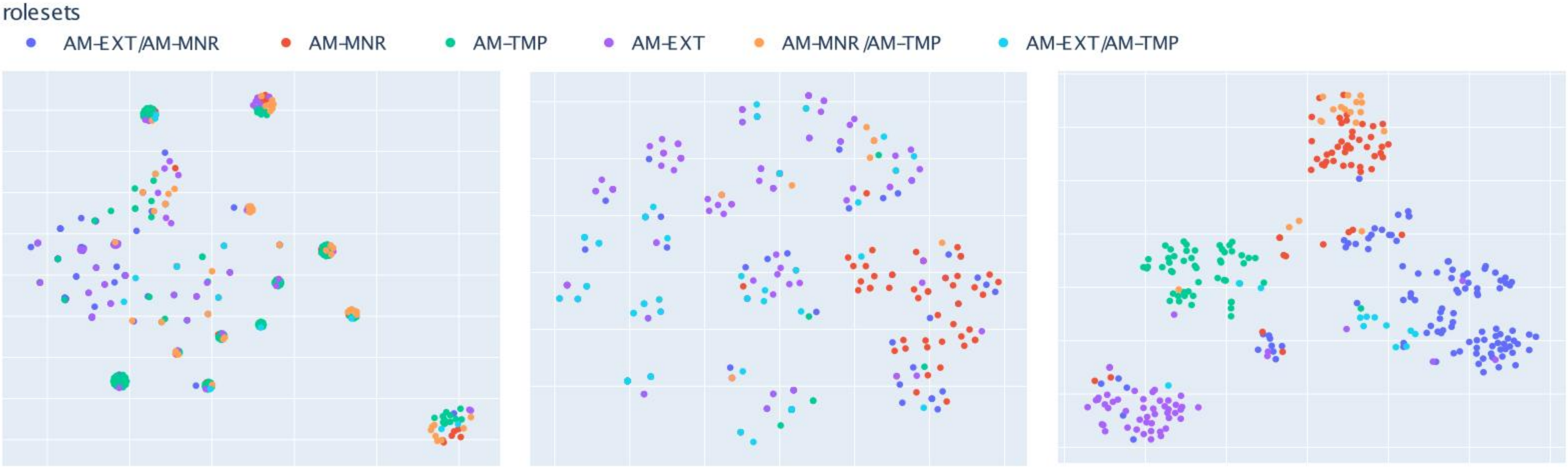


Figure 3: t-SNE visualization of the representations for the predicate *close*. Different colors represent different rolesets, even though some rolesets are partially overlapping (e.g. {AM-EXT, AM-MNR} and {AM-EXT, AM-TMP}). From left to right: predicate representations from the baseline SRL model which is completely unaware of rolesets (left); predicate representations from an SRL model that can use two different weighted averages to create different representations for predicate senses and their arguments (center); predicate representations from an SRL model that is tasked to explicitly identify rolesets through a secondary learning objective in a multi-task fashion (right).

Conclusion

- To probe PMS for PASs: two different core subtasks (senses and rolesets)
- different PLMs encode their features across significantly different layers (by weighted scores)
- verbal and nominal predicates and their PASs are represented differently
- current multilingual language models encode PASs similarly across two very different languages, English and Chinese

Q & A

THANK YOU

Note