清華大學 Tsinghua University

Paper & Work Sharing

Zhu Liu

2022.09.08

Outline

- Background
- One paper about evaluation of WSD
- Master thesis work

WSD in NAACL'22

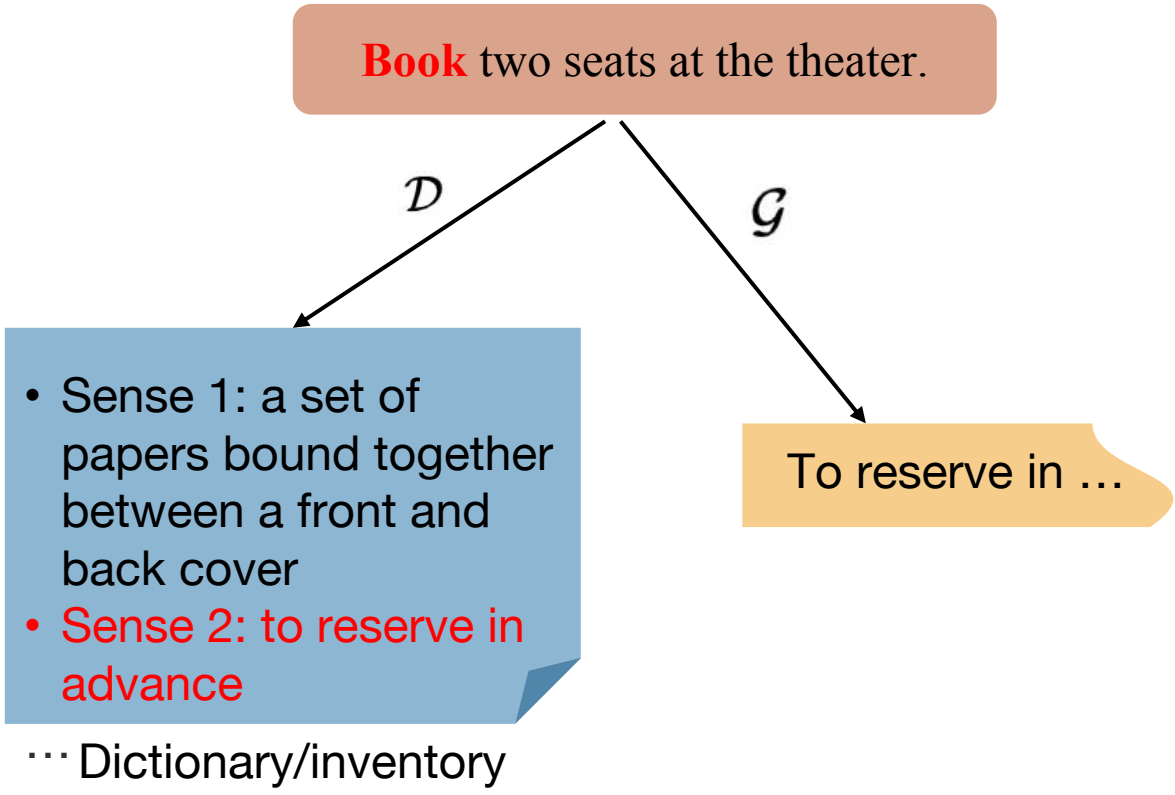
- NAACL'22 was held from July 11 to July 13.

1.Reducing Disambiguation Biases in NMT by Leveraging Explicit Word Sense Information Niccolò Campolungo, Tommaso Pasini, Denis Emelin, Roberto Navigli
2.WiC = TSV = WSD: On the Equivalence of Three Semantic Tasks Bradley Hauer, Grzegorz Kondrak
3.Global Entity Disambiguation with BERT Ikuya Yamada, Koki Washio, Hiroyuki Shindo, Yuji Matsumoto
4.MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation) S. Tedeschi and R. Navigli

https://2022.naacl.org/program/accepted_papers/

Word Sense Disambiguation

- Given a target word in a context, choose the best sense from an inventory (dictionary).
- Discriminative (as classification or retrieval) v. Generative (as structured sequence prediction)
- Corpora v. Knowledge
- Data-driven v. Knowledge-based
- Rule-based v. Statistics-based [Navigli, 2009] v. DL-based [Bevilacqua et al., 2021]



Main Approaches

Knowledge-based

- External inventories, like WordNet, BabelNet
- Independent from labeled training data (thus, unsupervised)
- Graph-based method

Hybrid supervised and knowledge-based

Pro:
Well-generalized; inclusive; robust

Con:
Time-consuming; unrelated to task; (sometimes) noisy; Inadequate performance

Pro:
Precise (similar distribution to test data); lightweight; time-saving

Con:
Generalized risks; Knowledge acquisition bottleneck; annotated bias

Supervised

- Annotated corpora, like SemCor.
- How to define the task matters. (That decides how to use the supervisory signal)
- Classification/retrieval/ extractive/generative

[Bevilacqua et al., 2021]

Evaluation

- How to compare different WSD systems fairly?
- How to measure the result of a WSD system?
  - (Discriminative) Output: confidence value (one-hot) $p(w_i)$ distributed on N_i possible senses for each target word w_i
  - Only accuracy is enough as a multi-class classification?

Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison

Alessandro Raganato, Jose Camacho-Collados and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
{raganato,collados,navigli}@di.uniroma1.it

EACL 2017; Cited by 290

http://lcl.uniroma1.it/wsdeval/

Motivation

- Lack of a reliable evaluation framework

1) Evaluation datasets differ in format, construction guidelines and underlying sense inventory.
2) Different training corpus and preprocessing.
3) Different dictionaries (coarse v. fine)

- Contributions

1) A unified benchmark for a fair comparison
2) Analysis on the effect of unlabeled corpus

Standardization of WSD datasets

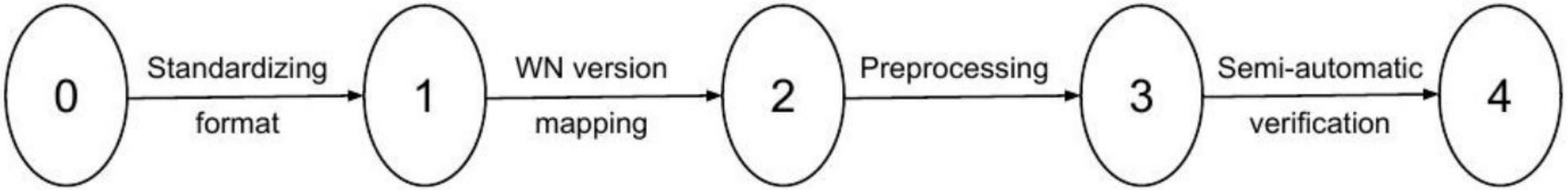
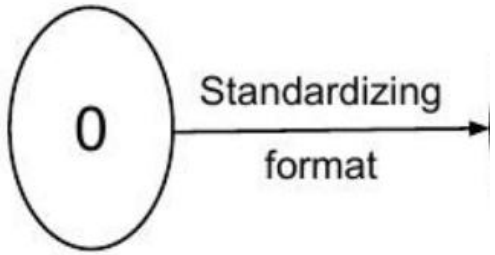


Figure 1: Pipeline for standardizing any given WSD dataset.

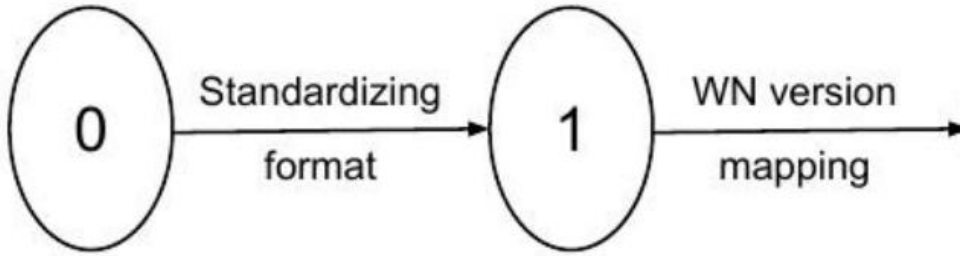
Corpus format




```
<sentence id="d000.s000">
<wf lemma="this" pos="DET">This</wf>
<instance id="d000.s000.t000" lemma="document" pos="NOUN">document</instance>
<wf lemma="be" pos="VERB">is</wf>
<wf lemma="a" pos="DET">a</wf>
<instance id="d000.s000.t001" lemma="summary" pos="NOUN">summary</instance>
<wf lemma="of" pos="ADP">of</wf>
<wf lemma="the" pos="DET">the</wf>
<instance id="d000.s000.t002" lemma="european" pos="ADJ">European</instance>
<instance id="d000.s000.t003" lemma="public" pos="ADJ">Public</instance>
<instance id="d000.s000.t004" lemma="assessment" pos="NOUN">Assessment</instance>
<instance id="d000.s000.t005" lemma="report" pos="NOUN">Report</instance>
<wf lemma="(" pos=".">(</wf>
<wf lemma="epar" pos="NOUN">EPAR</wf>
<wf lemma=")" pos=".">)</wf>
<wf lemma="." pos=".">.</wf>
</sentence>
```

Formalized as an XML file with a unique instance pointer.

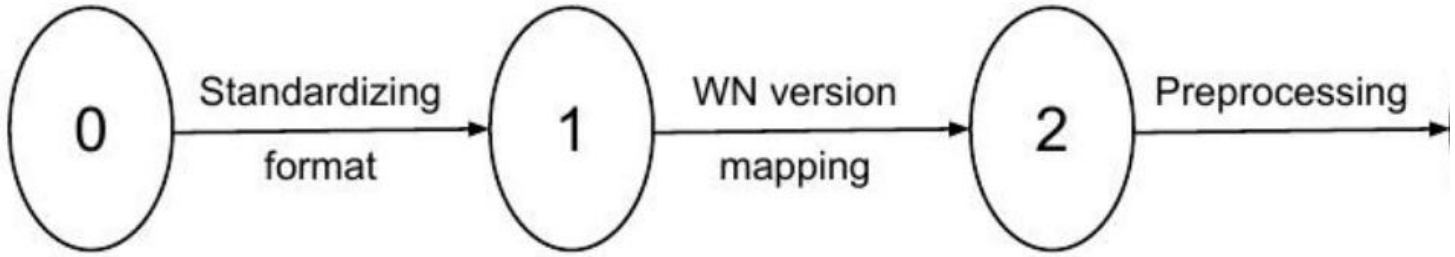
Knowledge format



Mapped original WordNet version to 3.0 (current version) semiautomatically:

1) Mapping senses with 100% confidence
2) The remaining sense annotation is checked manually.
3) Removal all annotations of auxiliary verbs.

Preprocessing



StanfordCoreNLP toolkit for PoS tagging and lemmatization

Standardization of WSD datasets

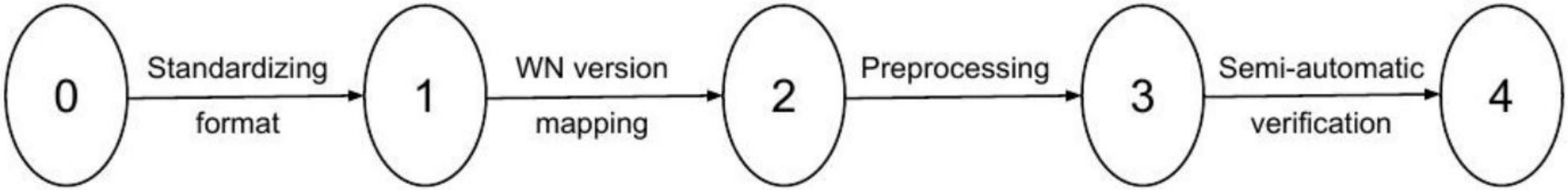


Figure 1: Pipeline for standardizing any given WSD dataset.

Make sure that the sense annotations match the lemma and the PoS tag

WSD datasets

- Evaluation: the Senseval/SemEval competitions: held from 1998
- (Training) corpus: SemCor (manually) & OMSTI (automatically)

	#Docs	#Sents	#Tokens	#Annotations	#Sense types	#Word types	Ambiguity
Senseval-2	3	242	5,766	2,282	1,335	1,093	5.4
Senseval-3	3	352	5,541	1,850	1,167	977	6.8
SemEval-07	3	135	3,201	455	375	330	8.5
SemEval-13	13	306	8,391	1,644	827	751	4.9
SemEval-15	4	138	2,604	1,022	659	512	5.5
SemCor	352	37,176	802,443	226,036	33,362	22,436	6.8
OMSTI	-	813,798	30,441,386	911,134	3,730	1,149	8.9

#Sense types: number of unique sense (sense vocabulary)
Ambiguity: number of candidate senses on average

WSD systems

- Supervised

1) IMS [Zhong and Ng, 2010]: SVM over a set of conventional WSD features
2) IMS + embeddings: word embedding (trained on unlabeled data)
3) Context2Vec [Melamud et al., 2016]: Neural networks
4) Baseline: MFS (lower bound)

- Knowledge-based

1) Lesk [Lesk, 1986]: based on overlap between the definitions of a given sense and the context of the target word
2) UKB-based: Personalized Page Rank algorithm and Random walk
3) Babelfy [Moro et al., 2014]: integrating World knowledge, such as Wikipedia.
4) Baseline: Wordnet first sense (lower bound)

Metrics

- F1 score

$$Prec = \frac{N_{correct}}{N_{annotated}} \quad Rec = \frac{N_{correct}}{N_{targets}}$$

$$F1 = \mathrm{GeoAvg}(Prec, Rec)$$

- When system annotates all the instances, $F1 = Prec = Rec = Acc$.
- [Personal] While F1 score is the de facto metric, it has some flaws:

1) Micro v. Macro [Maru, et al, 2022]: micro-averaged overestimates MFS bias
2) F1 is a hard match, Roc curve can be a softer operator [Cohn, 2003]
3) Lack of uncertainty quantification

Results

Ambiguity: 5.4, 6.7, 8.5, 4.9, 5.5

	Tr. Corpus	System	Senseval-2	Senseval-3	SemEval-07	SemEval-13	SemEval-15
Supervised	**SemCor**	IMS	70.9	69.3	61.3	65.3	69.5
		IMS+emb	71.0	69.3	60.9	**67.3**	71.3
		IMS$_{-s}$+emb	**72.2**	**70.4**	**62.6**	65.9	71.5
		Context2Vec	71.8	69.1	61.3	65.6	**71.9**
		MFS	65.6	66.0	54.5	63.8	67.1
		Ceiling	*91.0*	*94.5*	*93.8*	*88.6*	*90.4*
	SemCor + OMSTI	IMS	72.8	69.2	60.0	65.0	69.3
		IMS+emb	70.8	68.9	58.5	66.3	69.7
		IMS$_{-s}$+emb	**73.3**	**69.6**	61.1	66.7	70.4
		Context2Vec	72.3	68.2	**61.5**	**67.2**	**71.7**
		MFS	66.5	60.4	52.3	62.6	64.2
		Ceiling	*91.5*	*94.9*	*94.7*	*89.6*	*91.1*
Knowledge	**-**	Lesk$_{ext}$	50.6	44.5	32.0	53.6	51.0
		Lesk$_{ext}$+emb	63.0	63.7	**56.7**	66.2	64.6
		UKB	56.0	51.7	39.0	53.6	55.2
		UKB_gloss	60.6	54.1	42.0	59.0	61.2
		Babelfy	**67.0**	63.5	51.6	**66.4**	**70.3**
		WN 1st sense	66.8	**66.2**	55.2	63.0	67.8

Table 2: F-Measure percentage of different models in five all-words WSD datasets.

Results

	Nouns	Verbs	Adj.	Adv.	All
#Instances	4,300	1,652	955	346	7,253
Ambiguity	4.8	10.4	3.8	3.1	5.8

Table 3: Number of instances and ambiguity level of the concatenation of all five WSD datasets.

- Verb bears much more ambiguity

Results

- Training corpus
- KB v. supervised
- Verbs
- Baselines
- MFS bias (71%-75%)
- 80% ITA ceiling

	Tr. Corpus	System	Nouns	Verbs	Adjectives	Adverbs	All
Supervised	**SemCor**	IMS	70.4	56.1	75.6	82.9	68.4
		IMS+emb	71.8	55.4	**76.1**	82.7	69.1
		IMS$_{-s}$+emb	**71.9**	56.9	75.9	**84.7**	**69.6**
		Context2Vec	71.0	**57.6**	75.2	82.7	69.0
		MFS	67.6	49.6	73.1	80.5	64.8
		Ceiling	*89.6*	*95.1*	*91.5*	*96.4*	*91.5*
	SemCor + OMSTI	IMS	70.5	**56.9**	76.8	82.9	68.8
		IMS+emb	71.0	53.3	77.1	82.7	68.3
		IMS$_{-s}$+emb	**72.0**	56.5	76.6	**84.7**	**69.7**
		Context2Vec	71.7	55.8	**77.2**	82.7	69.4
		MFS	65.8	45.9	72.7	80.5	62.9
		Ceiling	*90.4*	*95.8*	*91.8*	*96.4*	*92.1*
Knowledge	-	Lesk$_{ext}$	54.1	27.9	54.6	60.3	48.7
		Lesk$_{ext}$+emb	**69.8**	**51.2**	51.7	80.6	63.7
		UKB	56.7	39.3	63.9	44.0	53.2
		UKB_gloss	62.1	38.3	66.8	66.2	57.5
		Babelfy	68.6	49.9	73.2	79.8	**65.5**
		WN 1st sense	67.6	50.3	**74.3**	**80.9**	65.2

Table 4: F-Measure percentage of different models on the concatenation of all five WSD datasets.

- Ceiling (Upper bound): proportion of test data appearing in the training set.

Later works

	Kind	System	ALL	S2	S3	S7	S13	S15
KB		[Scozzafava *et al.*, 2020, SyntagRank]	71.7	71.6	72.0	59.3	72.2	75.8
		[Wang and Wang, 2020, SREF_{KB}]	**73.5**	72.7	71.5	61.5	76.4	79.5
Vector-based 1-nn		[Loureiro and Jorge, 2019, LMMS]	75.4	76.3	75.6	68.1	75.1	77.0
		[Berend, 2020]	76.8	77.9	77.8	68.8	76.1	77.5
		[Scarlini *et al.*, 2020b, ARES]	**77.9**	78.0	77.1	71.0	77.3	83.2
		[Conia and Navigli, 2020, Conception]	76.4	77.1	76.4	70.3	76.2	77.2
		[Luan *et al.*, 2020]	76.4	77.2	77.1	69.2	76.1	77.2
		[Scarlini *et al.*, 2020a, SensEmBERT]	-	-	-	-	78.7	-
		[Wang and Wang, 2020, SREF]	77.8	78.6	76.6	72.1	78.0	80.5
Token Classifier		[Hadiwinoto *et al.*, 2019, GLU]	74.1	75.5	73.6	68.1	71.1	76.2
		[Vial *et al.*, 2019, SVC]	76.7	76.5	77.4	69.5	76.0	78.3
		[Kumar *et al.*, 2019, EWISE]	71.8	73.8	71.1	67.3	69.4	74.5
		[Blevins and Zettlemoyer, 2020, BEM]	**79.0**	79.4	77.4	74.5	79.7	81.7
		[Calabrese *et al.*, 2020a, EViLBERT]	75.1	-	-	-	-	-
		[Bevilacqua and Navigli, 2020, EWISER]	78.3	78.9	78.4	71.0	78.9	79.3
		[Conia and Navigli, 2021]	77.6	78.4	77.8	72.2	76.7	78.2
Seq. Classif.		[Huang *et al.*, 2019, GlossBERT]	77.0	77.7	75.2	72.5	76.1	80.4
		[Bevilacqua *et al.*, 2020, Generationary]	76.7	78.0	75.4	71.9	77.0	77.6
		[Yap *et al.*, 2020]	78.7	79.9	77.4	73.0	78.2	81.8
		[Barba *et al.*, 2021, ESCHER]	**80.7**	81.7	77.8	76.3	82.2	83.2

- Metric: F1 score
- Upper bound ~80% (By inter-annotator agreement) (uncertainty)

Conclusion & Future work

- A whole evaluation framework to fairly compare WSD systems

  http://lcl.uniroma1.it/wsdeval
- To exploit large amounts of unlabeled corpus (->pretrained model)
- To automatically constructing sense-annotated corpus (->active/semi-supervised learning)
- Multilingual WSD evaluation (XL-WSD [Pashini, et al, 2021])

[Personal] Anything else for evaluation?

- A system with uncertainty quantification
- The source of uncertainty:

1) Aleatoric (data): language vagueness; annotation errors; inter-annotation disagreement (~20% [Chklovski, 2003])

2) Epistemic (model): Different models; OOD

- The tool of UQ

1) Probability Theory, Information Theory

2) Fuzzy Logic [Kazemi, 2021]

Reference

[Navigli, 2009] Navigli, Roberto. "Word sense disambiguation: A survey." *ACM computing surveys (CSUR)* 41.2 (2009): 1-69.

[Bevilacqua et al., 2021] Bevilacqua, Michele, et al. "Recent trends in word sense disambiguation: A survey." *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc, 2021.

[Zhong and Ng, 2010] Zhi Zhong and Hwee Tou Ng. 2010. It Makes Sense: A wide-coverage Word Sense Disambiguation sys- tem for free text. In Proceedings ofthe ACL System Demonstrations, pages 78–83.

[Melamud et al., 2016] Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context em- bedding with bidirectional lstm. In Proceedings of CONLL.

[Lesk, 1986] Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings ofthe 5th Annual Conference on Systems Documen- tation, Toronto, Ontario, Canada, pages 24–26.

[Moro et al., 2014] Andrea Moro, Alessandro Raganato, and Roberto Nav- igli. 2014. Entity Linking meets Word Sense Dis- ambiguation: a Unified Approach. Transactions of the Association for Computational Linguistics (TACL), 2:231–244.

[Pashini, et al, 2021] Pasini, Tommaso, Alessandro Raganato, and Roberto Navigli. "XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 15. 2021.

[Chklovski, 2003] Cohn, Trevor. "Performance metrics for word sense disambiguation." *Proceedings of the Australasian Language Technology Workshop 2003*. 2003.

[Kazemi, 2021] Kazemi, Parham, and Hossein Karshenas. "Fuzzy Word Sense Induction and Disambiguation." *IEEE Transactions on Fuzzy Systems* (2021).

Q & A

THANK YOU

Sapienza NLP

http://nlp.uniroma1.it/

Knowledge-based WSD

Method	Algorithm	Knowledge	Language
SyntagRank [Scozzafava et al., 2020]	Personalized PageRank algorithm	WordNet portion of BabelNet; WNG	Multiple languages
SREF_KB [Wang and Wang, 2020]	vector-based approach	WordNet	English only

Other methods:
random walks [Agirre et al., 2014, UKB], clique approximation [Moro et al., 2014, Babelfy], or game theory [Tripodi and Navigli, 2019].

(Purely) Supervised WSD

- Annotations: SemCor: word, context, sense>
- How to define the task?

Mechanism	Method-based	Input	Output
Discriminative	(Multi-label) classification-based	Sense id (one-hot)	Sense id by logits
	Retrieval-based	All Glosses/senses	Sense id by similarity
	Span Extraction	All Glosses/senses	<Start id, End id>
Generative	Sequential generation	Gloss/sense	Sense *itself*

Supervised WSD Exploiting Glosses

- From one-hot to linguistic sequence.

1) Token-level classification → Sequence-level classification [Huang et al., 2019; Yap et al., 2020]

2) 1nn-approach (retrieval based):

i) to concatenate gloss vector to the original sense vector.

SensEmBERT [Scarlini et al., 2020a], ARES [Scarlini et al., 2020b], SREF [Wang and Wang, 2020]

ii) to learn an aligned training text and sense representations.

EWISE [Kumar et al., 2019], EWISER [Bevilacqua and Navigli, 2020], BEM [Blevins and Zettlemoyer, 2020]

3) Span extraction (location) problem: Barba et al. [2021, ESC & ESCHER]

4) Natural Language Generation (definition modeling): Bevilacqua et al. [2020]

Supervised WSD Exploiting Relations

How to exploit the graph structure of knowledge?

- Relations:

  Neighbor embeddings in WordNet. → Senses lack in SemCor [LMMS, 2019]

  WordNet hypernymy and hyponymy relations. → Refining prediction. [2020, SREF]

  Ancestor in the WordNet taxonomy → Reducing the output class number. [Vial et al. 2019]

  The full graph structure (GCNs) → Increasing more knowledge. [EWISER, 2020][Conia and Navigli 2021]

  Note: Token-level methods than sentence-level ones **more commonly** exploit relational knowledge.

- Other knowledges:

  BabelNet → Refining results by comparing them with NMT and BabelNet translations [Luan et al., 2020]

  BabelPic dataset → Adding visual modal [Calabrese et al., 2020b]

  Wikipedia and Web search contexts [Scarlini et al., 2020a; Scarlini et al., 2020b; Wang and Wang, 2020]

Senseval-1	1998	English, French and Italian		
Senseval-2	2001	Czech, Dutch, English, Estonian Basque, Chinese, Danish, English, Italian, Japanese, Korean, Spanish, Swedish Japanese	A S T	Double blind annotation by two linguistically trained annotators was performed on corpus instances, with a third linguist adjudicating between inter-annotator differences to create the "Gold Standard." 66.5% ITA (nouns and adj) 2283 sense annotations, including nouns, verbs, adverbs and adjectives
Senseval-3	2004	English, Italian Basque, Catalan, Chinese, English, Italian, Romanian, Spanish, Multilingual, Swedish Swedish	A LS SR	ITA: 72.5% (verbs: 67.8%, nouns: 74.9%, adjectives: 78.5%); Note: The typical ita of 70%-75% (Task 1) three documents from three different domains (editorial, news story and fiction), totaling 1850 sense annotations.
SemEval-1	2007	Cross languages Arabic Many tasks…	IR SL	(Task 17) 455 sense annotations for nouns and verbs only
SemEval-2010	2010		Task 14 17	
SemEval-13	2013		Task 12 (Navigli)	Wordnet 3.0; nouns only;thirteen documents from various domains
SemEval-15	2015		Task 13 (Navigil)	From three domains: biomedical, math/computing, social issues; 1022 sense annotations in four documents.
SemEval-2017	2017		MWS	Name change from this year?
SemEval-2021 Task2 [Web] [Github] [MyLink]	2021	Arabic, Chinese, English, French, Russian	Discriminative	Multilangual and cross-lingual settings; Sentence pair (context-inconsistent, one-to-one)
SemEval-2022 [Web]	2022	COmparing Dictionaries and WOrd Embeddings	E2D; D2E	

跨学科

- 定义？与其它相关概念的联系区别，例如：模糊（vagueness），不确定性（uncertainty），不具指（unspecified）
- 类型：polysemy & homonym & non-literal … ？
- 如何数学建模？
- Computational Lexical Ambiguity
  - 检测歧义词汇，并带有合理的“歧义程度”(不确定性衡量)。
  - 词义表征 & 评估
  - 词义识别 & 去歧义（WSD）& 评估
  - 词义建模 & 生成 （generation）& 评估
- 中文/跨语言？词义演变？

语言学

计算机科学

统计&数学

任务的必要性

- 词汇语义是自然语言处理任务的“先修任务”，它往往决定着如何分词，如何表征词汇（语义）等等。而词汇的表征往往是处理各个NLP任务的第一道大门。词汇的多义性使得表征这一任务更加困难。
- NLP任务的可解释性的一大方面是语义理解，而机器对于多义性词汇的理解程度则是语义理解的重要部分。
- 歧义性是所有语言的通用特征。

任务的可行性

- 可以结合很多语言学知识，例如在定义，评估等方面。多义性也是普通语言学的重要课题。
- 词汇多义为代表的解释性工作相对较少（？），更加看重模型比较和分析
- 所耗费资源相对较少。