Outline

Recent Trends in Word Sense Disambiguation: A Survey

Michele Bevilacqua[1] , **Tommaso Pasini**[2] ,
Alessandro Raganato[3] and **Roberto Navigli**[1]
[1]Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome
[2]Department of Computer Science, University of Copenhagen
[3]Department of Digital Humanities, University of Helsinki
michele.bevilacqua@uniroma1.it, tommaso.pasini@di.ku.dk
alessandro.raganato@helsinki.fi, roberto.navigli@uniroma1.it

ESC: Redesigning WSD with Extractive Sense Comprehension

Edoardo Barba[1] **Tommaso Pasini**[2,*] **Roberto Navigli**[1]
[1]Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome
[2]Department of Computer Science, University of Copenhagen
{barba,navigli}@di.uniroma1.it
tommaso.pasini@di.ku.dk

- 汇报时间：2022.03.29

Recent Trends in Word Sense Disambiguation: A Survey

Michele Bevilacqua[1] , **Tommaso Pasini**[2] ,
Alessandro Raganato[3] and **Roberto Navigli**[1]
[1]Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome
[2]Department of Computer Science, University of Copenhagen
[3]Department of Digital Humanities, University of Helsinki
michele.bevilacqua@uniroma1.it, tommaso.pasini@di.ku.dk
alessandro.raganato@helsinki.fi, roberto.navigli@uniroma1.it

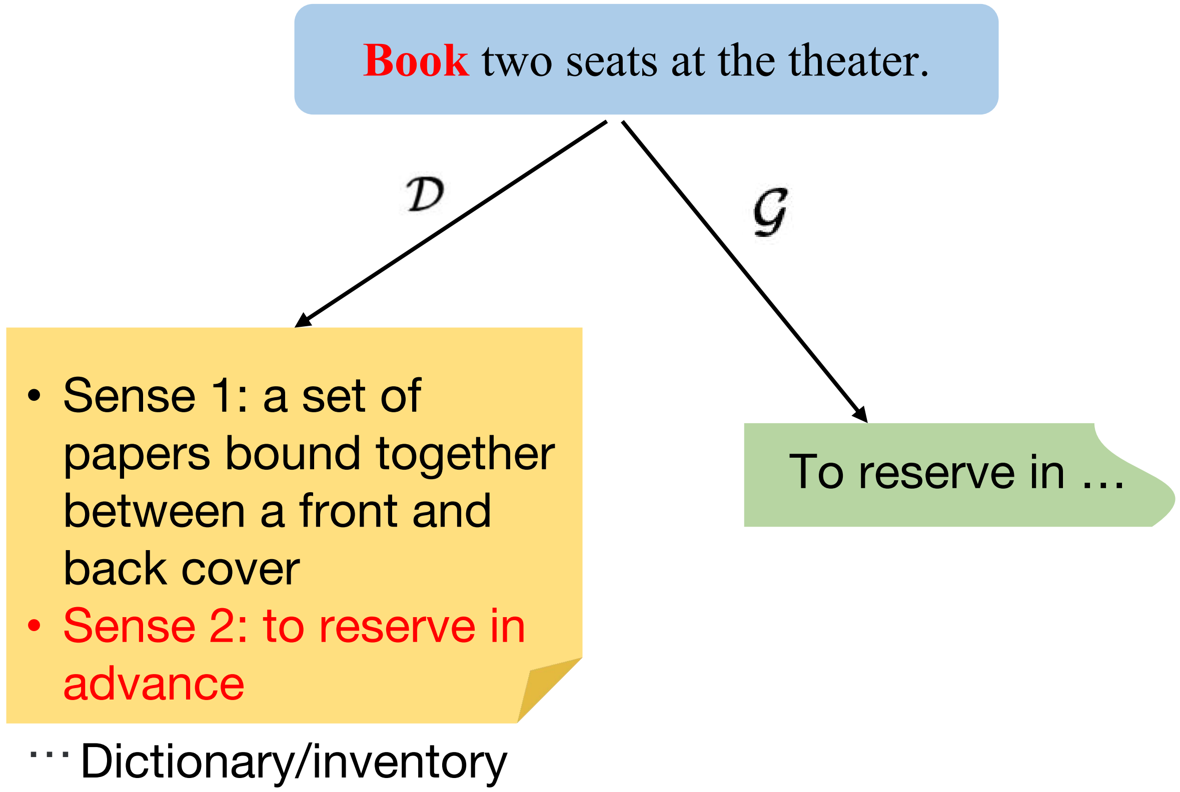
IJCAI-21 Survey Track; Cited by 14

What is WSD?

- Word Sense Disambiguation (WSD) aims at making explicit the semantics of a word in context by:

  1) identifying the most suitable meaning from a predefined sense inventory. (*discriminate*)

  Or, 2) generating it. (*generative*)



Why lexical ambiguity?

Different meanings (ambiguous?) for the <u>same</u> lexical form:

- Polysemous words (多义词)

  - (Historically) related meanings, e.g., “pet chicken” v. “roast chicken”

  - Often in the same entry of a dictionary. 84%, and 37% have five or more senses [Rodd et al., 2004].

- Homonymous words (同形/音异义词)

  - Unrelated meanings, e.g., “tree bark” (树皮) v. “dog bark” (狗吠).

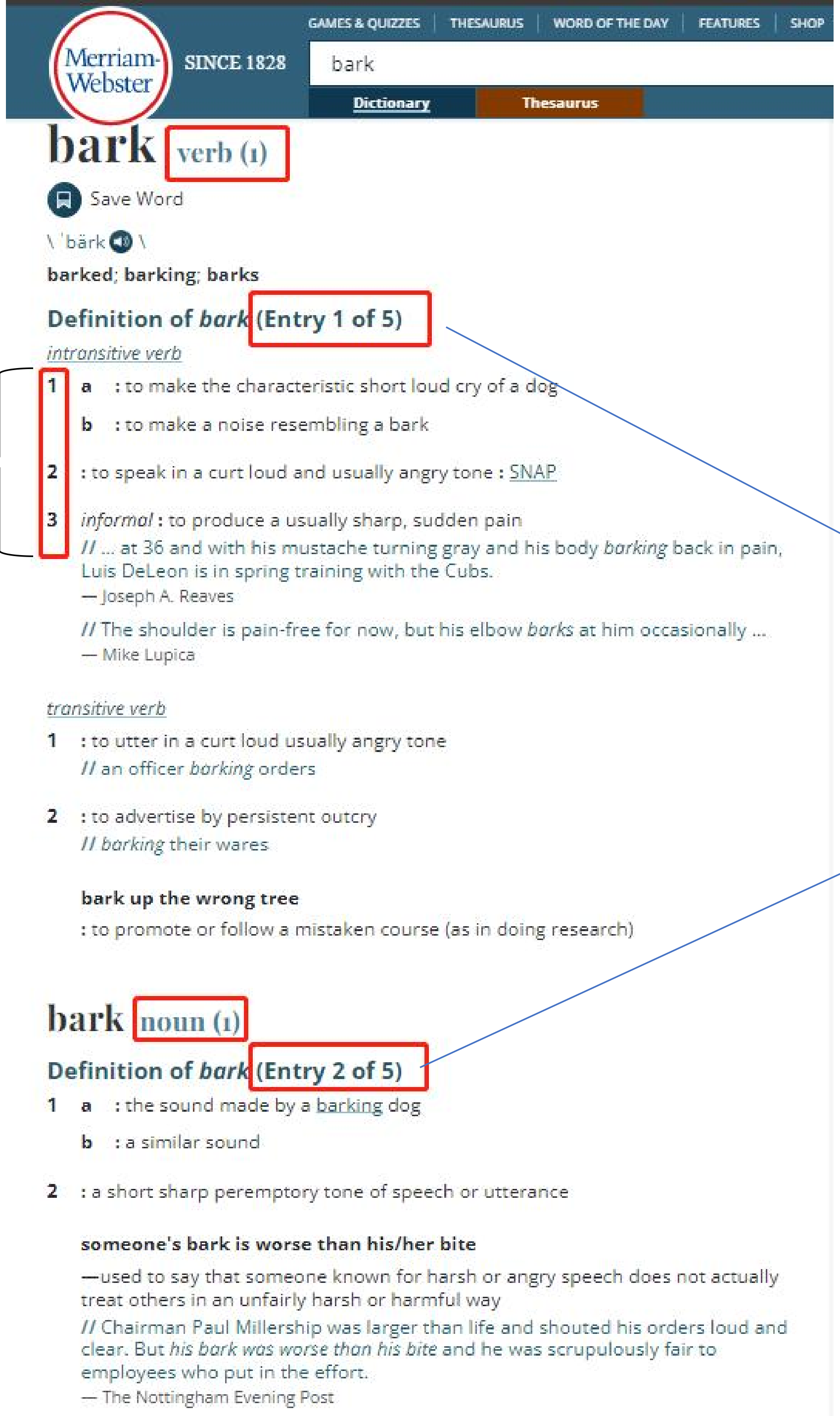
  - Often in the separate headwords (entries). 7.4%.

- Other factors, e.g., tones, pronunciation…

Lexeme: <lemma, pos>
词素: <词目, 词类>

多义词

同形异义词


GAMES & QUIZZES | THESAURUS | WORD OF THE DAY | FEATURES | SHOP

Merriam-Webster SINCE 1828

bark

Dictionary Thesaurus

bark verb (1)

Save Word

\ ˈbärk \

barked; **barking**; **barks**

Definition of *bark* (Entry 1 of 5)

intransitive verb

1 a : to make the characteristic short loud cry of a dog

b : to make a noise resembling a bark

2 : to speak in a curt loud and usually angry tone : SNAP

3 *informal* : to produce a usually sharp, sudden pain

// ... at 36 and with his mustache turning gray and his body *barking* back in pain, Luis DeLeon is in spring training with the Cubs.
— Joseph A. Reaves

// The shoulder is pain-free for now, but his elbow *barks* at him occasionally ...
— Mike Lupica

transitive verb

1 : to utter in a curt loud usually angry tone
// an officer *barking* orders

2 : to advertise by persistent outcry
// *barking* their wares

bark up the wrong tree
: to promote or follow a mistaken course (as in doing research)

bark noun (1)

Definition of *bark* (Entry 2 of 5)

1 a : the sound made by a barking dog

b : a similar sound

2 : a short sharp peremptory tone of speech or utterance

someone's bark is worse than his/her bite

—used to say that someone known for harsh or angry speech does not actually treat others in an unfairly harsh or harmful way

// Chairman Paul Millership was larger than life and shouted his orders loud and clear. But *his bark was worse than his bite* and he was scrupulously fair to employees who put in the effort.
— The Nottingham Evening Post

Overview

- Resources: What do we have?
- Methods: knowledge or data-driven?
- Evaluation

Resources for WSD

WSD is a knowledge-intensive task:

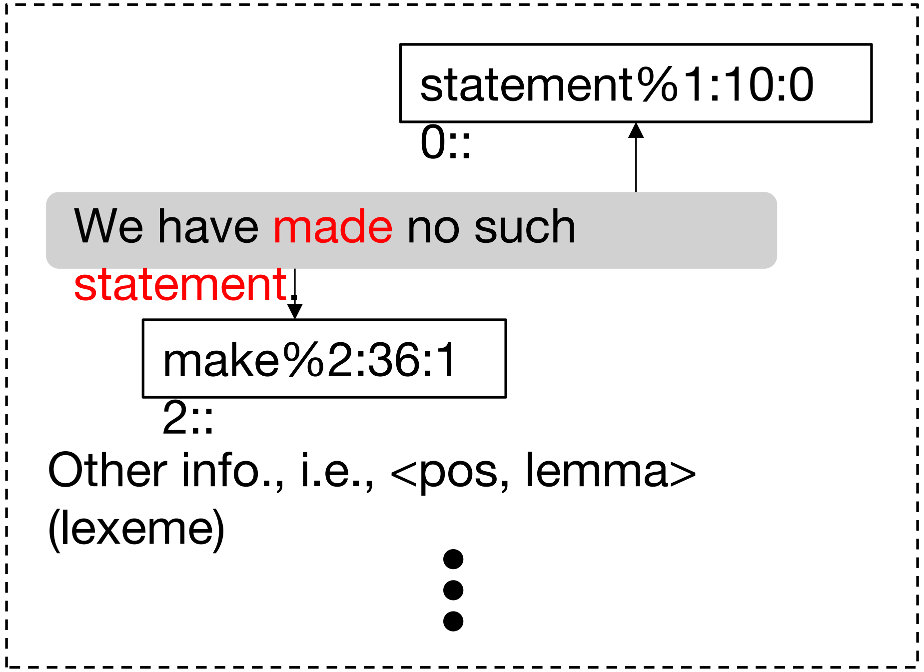
- Sense inventories

  reference computational lexicons which enumerate possible meanings.

- Annotated corpora

  a subset of words (*instance*) are tagged with one or more possible meanings drawn from the given inventory.

corpor
a



Resources for WSD

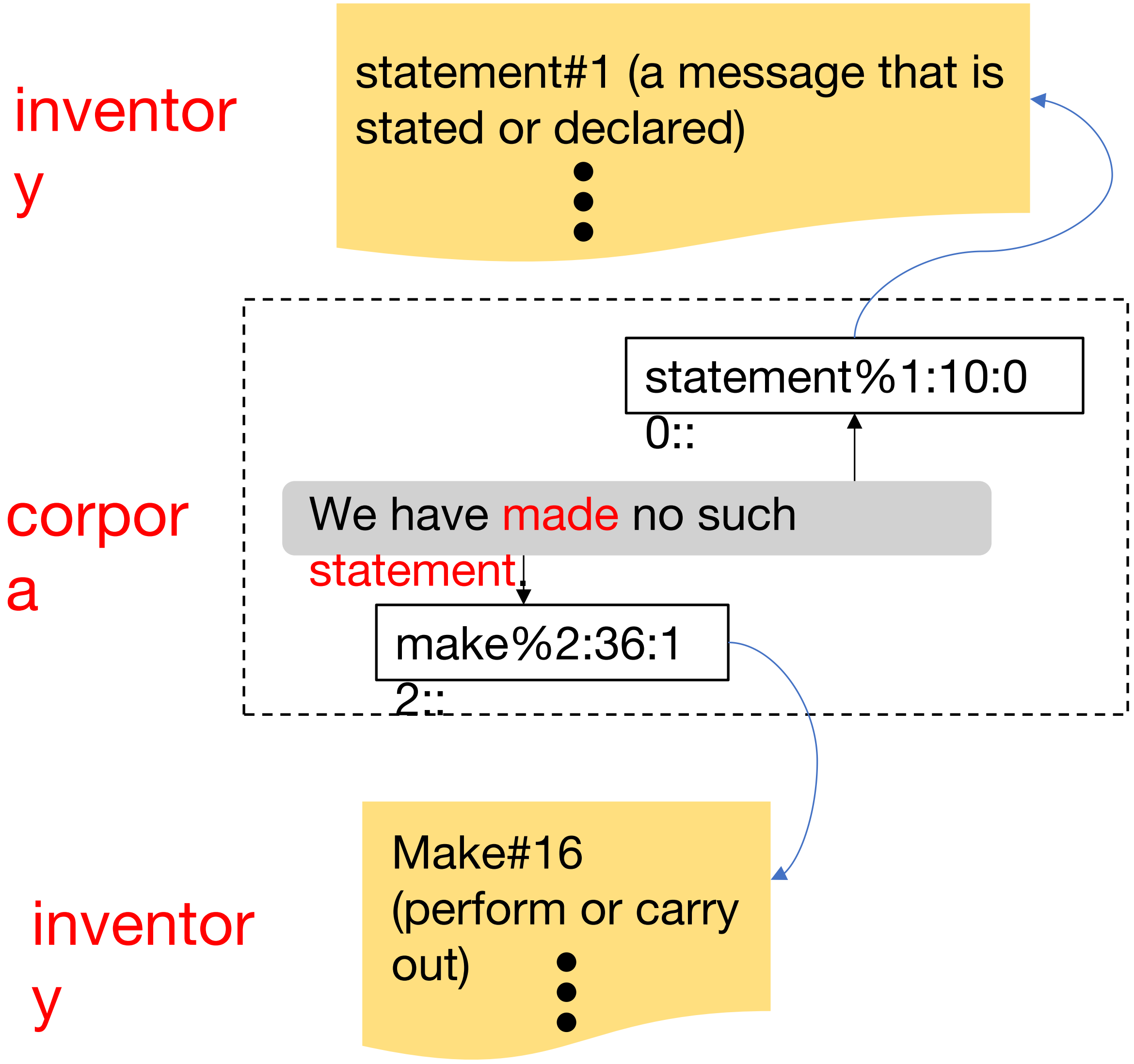
WSD is a knowledge-intensive task:

- Sense inventories

  reference computational lexicons which enumerate possible meanings.

- Annotated corpora

  a subset of words (*instance*) are tagged with one or more possible meanings drawn from the given inventory.



Sense Inventories: WordNet

- A large, manually-curated lexicographic database of English and the *de facto standard inventory* for WSD.
- WordNet was first created by psychology professor Miller in Princeton University in 1985. [Miller et al., 1990]
- Structured as a graph:

1) Node: synsets (同义词集): groups of contextual synonyms. (lemma)
+ gloss (brief definition) + examples + flag …
  - Consist of common nouns, verbs, adjectives and adverbs…
  - "form-meaning pair" thesaurus

2) Edge: hypernymy (is-a, 上下级关系): armchair→chair→furniture
meronymy (part-of, 部分从属): seat/leg—chair

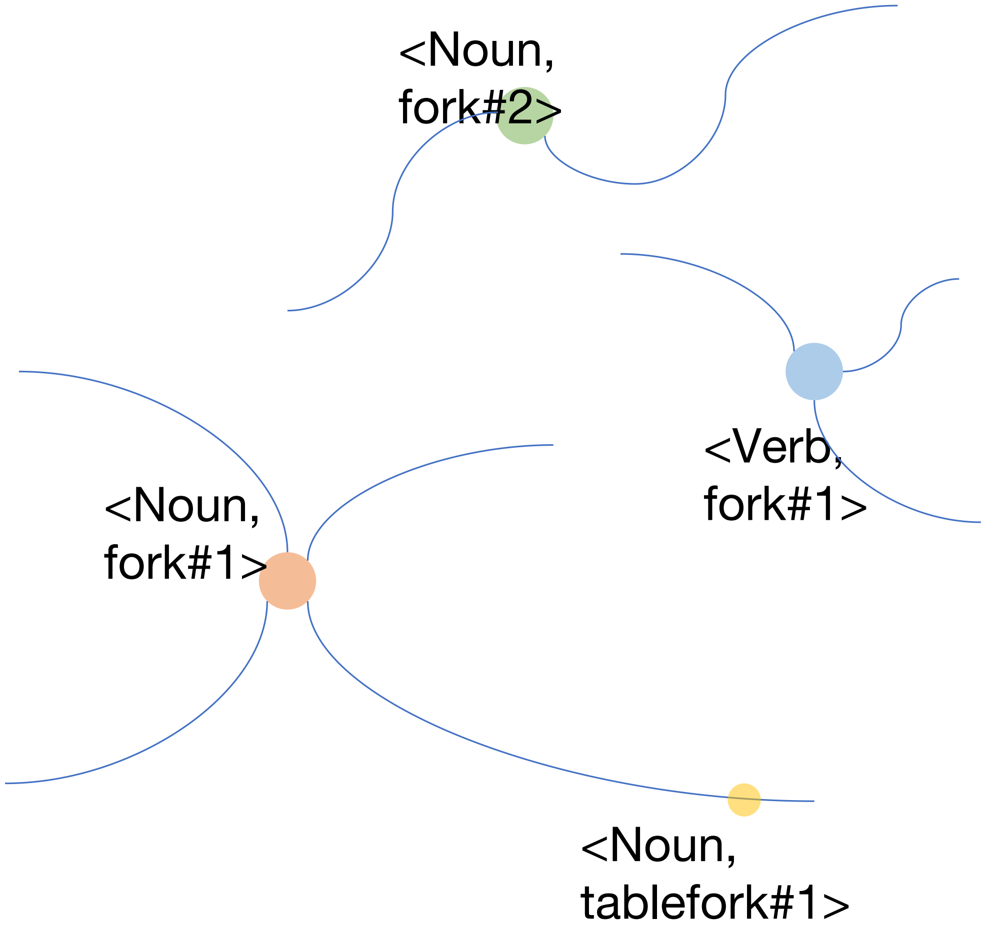
Sense Inventories: WordNet

Noun

- (4){03388794} <noun.artifact>[06] S: (n) **fork#1 (fork%1:06:00::)** (cutlery used for serving and eating food)
- (2){00389200} <noun.act>[04] S: (n) branching#1 (branching%1:04:00::), ramification#1 (ramification%1:04:00::), **fork#2 (fork%1:04:00::)**, forking#2 (forking%1:04:00::) (the act of branching out or dividing into branches)
- (1){13937280} <noun.shape>[25] S: (n) **fork#3 (fork%1:25:00::)**, crotch#1 (crotch%1:25:00::) (the region of the angle formed by the junction of two branches) *"they took the south fork"; "he climbed into the crotch of a tree"*
- (1){03389013} <noun.artifact>[06] S: (n) **fork#4 (fork%1:06:02::)** (an agricultural tool used for lifting or digging; has a handle and metal prongs)
- {05605191} <noun.body>[08] S: (n) crotch#2 (crotch%1:08:00::), **fork#5 (fork%1:08:00::)** (the angle formed by the inner sides of the legs where they join the human trunk)

Verb

- {01582189} <verb.contact>[35] S: (v) pitchfork#1 (pitchfork%2:35:00::), **fork#1 (fork%2:35:00::)** (lift with a pitchfork) *"pitchfork hay"*
- {01121306} <verb.competition>[33] S: (v) **fork#2 (fork%2:33:00::)** (place under attack with one's own pieces, of two enemy pieces)
- {00329612} <verb.change>[30] S: (v) branch#2 (branch%2:30:00::), ramify#3 (ramify%2:30:00::), **fork#3 (fork%2:30:00::)**, furcate#1 (furcate%2:30:00::), separate#13 (separate%2:30:04::) (divide into two or more branches so as to form a fork) *"The road forks"*
- {00141734} <verb.change>[30] S: (v) **fork#4 (fork%2:30:01::)** (shape like a fork) *"She forked her fingers"*



http://wordnetweb.princeton.edu/perl/webwn

Sense Inventories: Others

- BabelNet [Navigli and Ponzetto, 2012]:
  multilingual, similar structure to WordNet.
  (500 languages & 20M synsets vs. ~117K for WordNet)

- HowNet: sememe-based (义原), 2K sememes and uses them to annotate over 100K Chinese and English words.

https://openhownet.thunlp.org/about_hownet

Sense-annotated Corpora

- Data for Training

(1) SemCor [Miller et al., 1994] is the largest *manually* sense-annotated corpus annotated with WordNet senses.

(2) OMSTI [Taghipour and Ng, 2015a] is an automatically constructed corpus based on WordNet 3.0 inventory.

- Data for Testing

(1) Senseval-2: WordNet 1.7 based.

(2) Senseval-3: editorial, news story and fiction

(3) SemEval-07 task 17: smallest, based on WordNet 2.1.

(4) SemEval-13 task 12: Wordnet 3.0, only nouns

(5) SemEval-15 task 13: Wordnet 3.0, biomedical, mathematics/computing and social issues

Sense-annotated Corpora - Statistics

	#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

Dataset limitations

- Wordnet has fine-grained annotations, causing the glosses non-orthogonal. E.g., two glosses for *change* are *a thing that is different* and *a different or fresh set of clothes.*
- Training Corpus: Most Frequent Sense Bias (0-shot or few-shot for some senses)
- [Personal] It fails to capture the uncertainty, since the sense of one word given a context is certain, in other words, *not* ambiguous. (I am not *sure* what this word mean in the context.)

Main Approaches

Knowledge-based

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

Pro:
Generalized; inclusive; robust; no need to labeling

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

Just think of it as reciting a dictionary before your IELT test…

Main Approaches

Pro:
Precise (as a target); lightweight; time-saving

Con:
Epistemic uncertainty (less generalized);
Labeling/training time consuming; Annotated bias

Supervised

- Annotated corpora, like SemCor.
- Multi-label
- How to define the task matters. (That decides how to use the supervisory signal)

Just think it as preparing your IELT test only by old exercises 4-15…

Main Approaches

Knowledge-based

Hybrid supervised and knowledge-based

Supervised

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

Pro:
Generalized; inclusive; robust

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

Pro:
Precise (as a target); lightweight; time-saving

Con:
Epistemic uncertainty (less generalized);
Labeling/training time consuming; Annotated bias

- Annotated corpora, like SemCor.
- Multi-label
- How to define the task matters. (That decides how to use the supervisory signal

Knowledge-based WSD

Method	Algorithm	Corpus	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*

(Purely) Supervised WSD

- Annotations: SemCor: [w1, w2, …<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*

Purely Data-Driven WSD (token-level)

Token-level classification

Training:

$$
\begin{aligned}
E_c &= \text{Embed}(c) & E_c &= \text{Embed}(c)\\
H_{c,w} &= \text{FFN}(E_{c,w}) & H_{c,w} &= \text{Transformer}(E_c)_w\\
P_{c,w} &= \text{Softmax}(H_{c,w}O) & P_{c,w} &= \text{Softmax}(H_{c,w}O)
\end{aligned}
$$

$P_{c,w} \in \mathbb{R}^N$ shows the probability of N possible senses

Test:

$$\hat{s} = \underset{s' \in V^{(w)}}{\text{argmax}}\, P_{c,w,s'}$$

[Hadiwinoto et al., 2019], [Bevilacqua and Navigli, 2019], [Vial et al., 2019]

Purely Data-Driven WSD

1-nn vector-based (retrieval based)

Training:

$$v^{(c,w)} = \text{Embed}(c)_w$$
$$v^{(s)} = \frac{1}{|\mathcal{D}^{(w,s)}|} \sum_{c' \in \mathcal{D}^{(w,s)}} \text{Embed}(c')_w$$

sense embeddings: averaging the contextual vectors of instances within the training set with the same sense.

Test:

$$\hat{s} = \underset{s' \in V^{(w)}}{\text{argmax}}\ \text{sim}_{\cos}(v^{(c,w)}, v^{(s')})$$

[Peters et al., 2018]

Supervised WSD Exploiting Glosses

- From one-hot to linguistic sequence.

1) Token-level classification → Sequence-level classification/matching/retrieval [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 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 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 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 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]

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]

Evaluation

	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)

What's next?

- New challenging test sets, e.g., OOD sense distribution. (domain shift is common for Web text and evolving languages). Note that sentence-level and knowledge-based methods offer zero-shot capabilities due to more data available.
- Multilingual WSD. (dataset, evaluation, specific issues…)
- How to employ them in downstream tasks, like NMT, QA and so on. [One simple observation is that not every word needs disambiguation given a clear context; And some words (like metaphors) do not even appear at a Wordnet-like dictionary.]
- How to interpret? Does it really capture the sense, even if it breaches the glass ceiling?

What’s next?

- New challenging test sets, e.g., OOD sense distribution. (domain shift is common for Web text and evolving languages). Note that sentence-level and knowledge-based methods offer zero-shot capabilities due to more data available.
- Multilingual WSD. (dataset, evaluation, specific issues…)
- How to employ them in downstream tasks, like NMT, QA and so on. [One simple observation is that not every word needs disambiguation given a clear context; And some words (like metaphors) do not even appear at a Wordnet-like dictionary.]
- How to interpret? Does it really capture the sense, even if it breaches the glass ceiling?

ESC: Redesigning WSD with Extractive Sense Comprehension

Edoardo Barba[1] **Tommaso Pasini**[2,*] **Roberto Navigli**[1]
[1]Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome
[2]Department of Computer Science, University of Copenhagen
{barba,navigli}@di.uniroma1.it
tommaso.pasini@di.ku.dk

NAACL-21; Cited by 11

Introduction

- WSD definition:

Multi-label classification

Formulation:
a very large vocabulary of discrete senses
Limitations:

- Poor generalization (sense only defined by occurrences in the training data).
- Unexplored lingual cues.
- Not flexible

Introduction

Context-Gloss Pairs of the target word [research]	Label
[CLS] Your research ... [SEP] systematic investigation to ... [SEP]	Yes
[CLS] Your research ... [SEP] a search for knowledge [SEP]	No
[CLS] Your research ... [SEP] inquire into [SEP]	No
[CLS] Your research ... [SEP] attempt to find out in a ... [SEP]	No

- WSD definition:

Multi-label classification → Gloss-aware classification

Formulation:
a very large vocabulary of discrete senses
Limitations:
- Poor generalization (sense only defined by occurrences in the training data).
- Unexplored lingual cues.
- Not flexible

Formulation:
integrating sense definition
Limitations:
- Not *all* candidate definitions are utilized together at once.

→ Limit the capability and generalization.

Introduction

- WSD definition:

Multi-label classification	→	Gloss-aware classification

Multi-label classification

Formulation:
a very large vocabulary of discrete senses

Limitations:
- Poor generalization (sense only defined by occurrences in the training data).
- Unexplored lingual cues.
- Not flexible

Gloss-aware classification

Formulation:
integrating sense definition

Limitations:
- Not *all* candidate definitions are explicitly utilized.

→ Limit the capability and generalization.

ESC: integrating all the candidates

Introduction: ESC

- New frame: Extractive Sense Comprehension (ESC) is inspired by the Extractive Reading Comprehension in the field of Question Answering.
- Formulation:

1) Input: a sentence with a target word and all its possible sense definitions.

2) Output: the location of the text span for the correct meaning.

- Advantages:

1) Generalization: efficient few-shot learning

2) Flexibility: It can scale effectively across different lexical resources.

Related Work

Mechanism	Method-based	Resources	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*

Method

Input:

$$m = <s>\ w_1 \dots <t>\ \hat{w}\ </t> \dots w_n\ </s>$$
$$w_1^{d_1} \dots w_{|d_1|}^{d_1} \dots w_1^{d_k} \dots w_{|d_k|}^{d_k}\ </s>$$

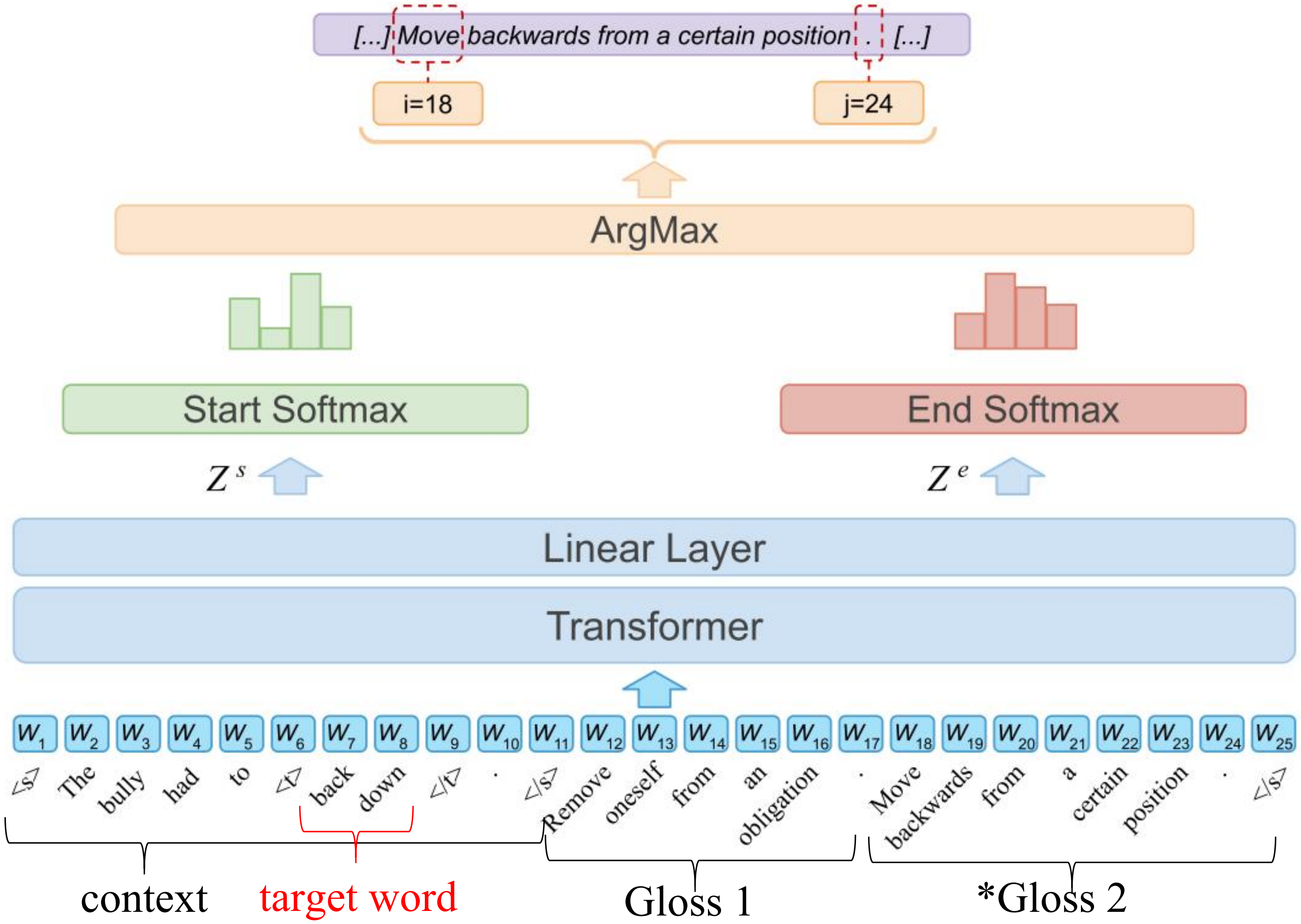
(Training) Output:

$$H = \texttt{transformer}(m)$$
$$Z = W^T H + b$$
$$Z^s = \begin{bmatrix} Z_{11} & \dots & Z_{1l} \end{bmatrix}$$
$$Z^e = \begin{bmatrix} Z_{21} & \dots & Z_{2l} \end{bmatrix}$$

where, $H \in \mathbb{R}^{f \times l}$
$W \in \mathbb{R}^{f \times 2}$

Loss:

$$\mathcal{L}_s = -Z^s_{i*} + log \sum_{v=1}^{l} exp(Z^s_v)$$
$$\mathcal{L}_e = -Z^e_{j*} + log \sum_{v=1}^{l} exp(Z^e_v)$$



Method

Input:

$$m = <s>\ w_1 \ldots <t>\ \hat{w}\ </t> \ldots w_n\ </s>$$
$$w_1^{d_1} \ldots w_{|d_1|}^{d_1} \ldots w_1^{d_k} \ldots w_{|d_k|}^{d_k}\ </s>$$

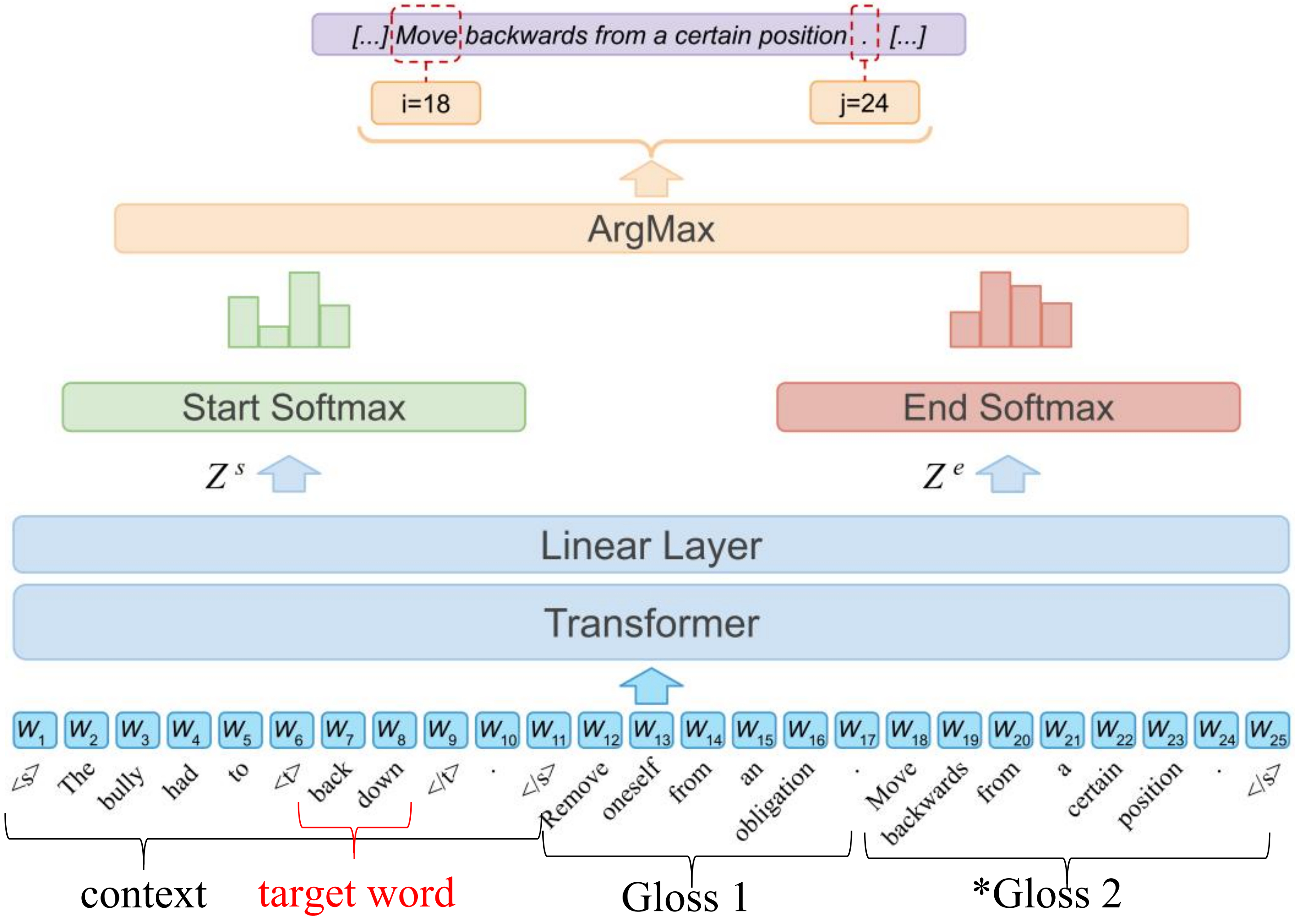
Test:

$$\texttt{output} = \arg\max_{(i,j)} P(w_i, w_j)$$

$$P(w_i, w_j) = P(w_i = start \mid Z^s) \times P(w_j = end \mid Z^e)$$

$$P(w_u = start \mid Z^s) = \frac{exp(Z_u^s)}{\sum_{v=1}^{l} exp(Z_v^s)}$$

$$P(w_u = end \mid Z^e) = \frac{exp(Z_u^e)}{\sum_{v=1}^{l} exp(Z_v^e)}$$



Method: Rebalancing MFS bias

- It may still be biased towards the most frequent sense (MFS) regardless of its contextualization.
- Inspired by negative sampling technique, the author adds k frequent definitions (Gloss Noise) that are not related to the target word.
- The k unrelated glosses are sampled from the following multinomial distribution:

$$p(d_i) = \frac{f_{d_i}}{\sum_{j=1}^{|D|} f_{d_j}}$$

Where k is sampled from a Poisson distribution with λ=1 （expectation is 1).
It keeps the discrepancy between training and testing as small as possible.

Standard WSD Evaluation

- Standard data: SemCor for training; SE07, SE2, SE3, SE13, SE15 for testing.
- New constructed data for generalization testing:

	Instances with …
MFS	Most frequent senses
LFS	Least frequent senses
0-lex	Unseen lexeme in the training set
0-lex-def	Unseen <lexeme, gloss/definition> in the training set
0-def	Unseen gloss in the training set

water #1, #2, #3, #4, #5, #6

water #1

water #1+H2O#1

Note: 1) lexeme = <lemma, part of speech>.
2) 0-def includes definitions shared by different synonyms (synset), with more seen cases than 0-lex-def.

Results (F1) - Performance

BERT and BART are as feature extractors of classifiers

		Dev Set	Test Sets				Concatenation of all Datasets				
	Model	SE07	SE2	SE3	SE13	SE15	Nouns	Verbs	Adj.	Adv.	ALL
Baselines	MFS SemCor	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5
	BERT_{base}	68.6	75.9	74.4	70.6	75.2	75.7	63.7	78.0	85.8	73.7
	BART_{large}	63.5	75.0	72.2	69.3	74.2	74.0	61.6	76.9	86.1	72.2
Prior work	EWISE‡	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
	GLU	68.1	75.5	73.6	71.1	76.2	—	—	—	—	74.1
	LMMS††	68.1	76.3	75.6	75.1	77.0	—	—	—	—	75.4
	SVC	—	—	—	—	—	—	—	—	—	75.6
	GlossBERT†	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	87.4	77.0
	ARES††	71.0	78.0	77.1	78.7	75.0	80.6	68.3	80.5	83.5	77.9
	EWISER‡	71.0	78.9	**78.4**	78.9	79.3	81.7	66.3	81.2	85.8	78.3
	BEM†	74.5	79.4	77.4	79.7	81.7	81.4	68.5	83.0	**87.9**	<u>79.0</u>
Ours	$\text{ESCHER}_{\text{No-GN}}$†	75.0	80.5	76.9	81.1	83.0	83.0	68.5	81.9	86.1	79.7
	ESCHER†	**76.3**	**81.7**	77.8	**82.2**	**83.2**	**83.9**	**69.3**	**83.8**	86.7	<u>**80.7**</u>

Results (F1) – Generalization

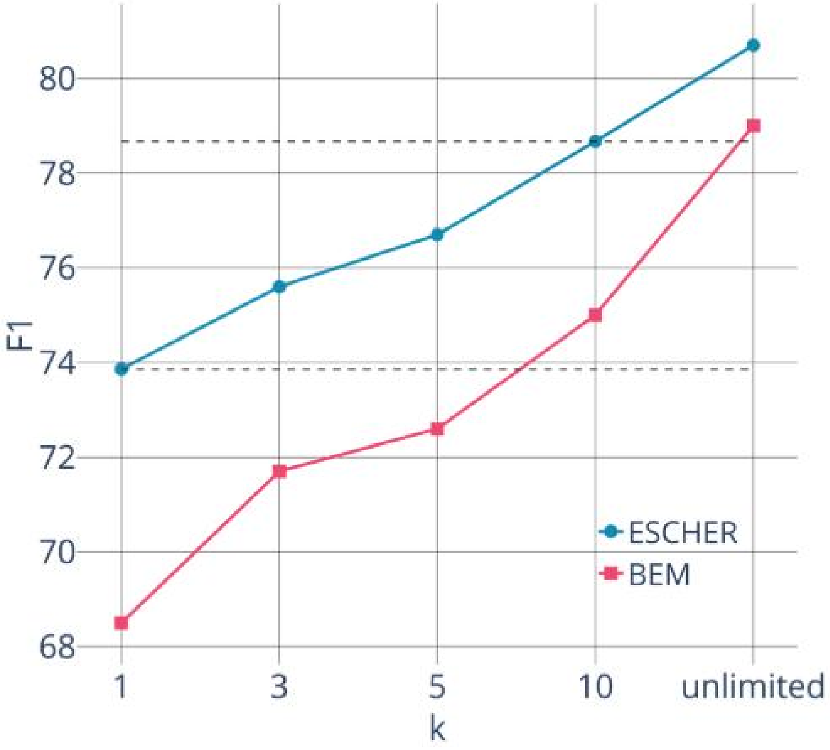
Model	MFS	LFS	0-lex	0-lex-def	0-def
BEM	**94.7**	52.1	91.2	67.1	68.2
ESCHER$_{\text{No-GN}}$	93.7	52.8	94.5	74.3	76.4
ESCHER	93.7	**55.7**	**95.1**	**75.0**	**76.8**

Zero-shot datasets

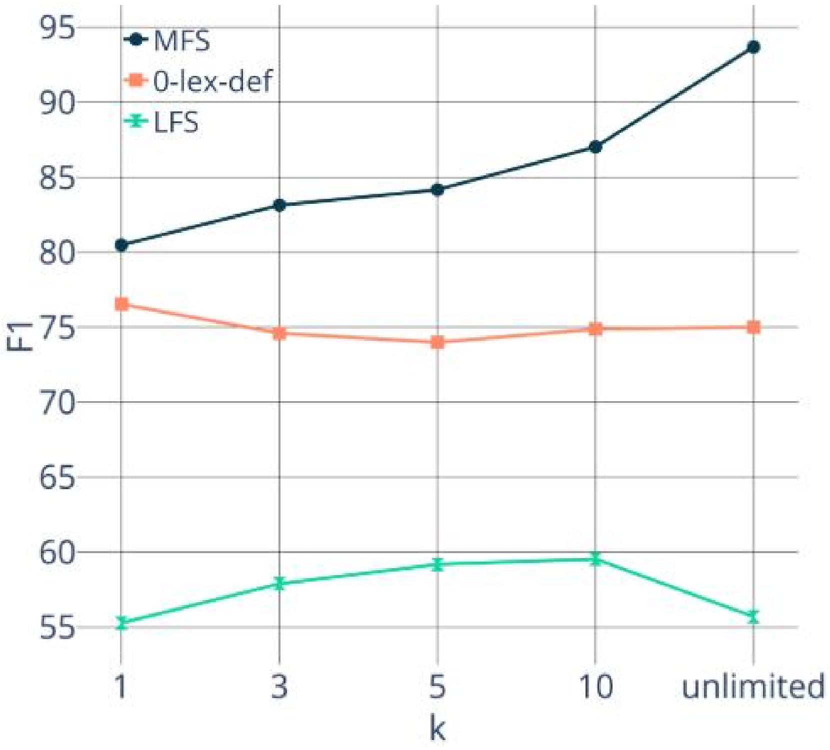
Results (F1) – Few-shot scenario

k: number of training instances （i.e., annotated word) per sense

k	Instances
1	33,206
3	64,814
5	83,068
10	109,751
unlimited	226,036



(a) ALL dataset.



(b) LFS and 0-lex-def datasets.

Flexibility – Merging Multiple Knowledge Bases

Additional corpus: Oxford Dictionary
Statistics: Exp. Polysemy #{appeared senses} over #{all possible senses}

	Dataset	Polysemy	Exp. Polysemy	#Senses	#Instances
WordNet	SemCor	6.88	0.76	33,362	226,036
	SE07	8.48	0.29	375	455
	ALL	5.87	0.54	3,669	7,611
Oxford	Oxford_{train}	3.81	0.98	79,105	555,695
	Oxford_{dev}	6.69	0.68	33,197	78,550
	Oxford_{test}	6.79	0.76	37,714	151,306

Table 3: Statistics for training, development and test corpora annotated with two inventories: WordNet (top) and Oxford (bottom).

Model	SE07	ALL	OX_{dev}	OX_{test}
BEM_S	74.5	79.0	61.5	61.7
ESCHER_S	76.3	80.7	67.6	67.9
BEM_{OT}	56.9	67.2	84.2	84.3
ESCHER_{OT}	60.7	70.3	86.3	86.3
BEM_{S+OT}	74.9	78.8	85.0	85.2
ESCHER_{S+OT}	**77.8**	**81.5**	**87.6**	**87.7**

Table 4: Comparison of ESCHER and BEM when using different training sets, i.e., SemCor (BEM_S and ESCHER_S), Oxford_{train} (BEM_{OT} and ESCHER_{OT}) and their concatenation (BEM_{s+OT} and ESCHER_{S+OT}).

Error Analysis

- Most frequent sense bias
- Insufficient context

Phenomenon: ESCHER mistakes most often appear in sentences with an average length of 27 tokens, roughly 5 tokens less than that in ALL (32)

Annotation bias: annotators consider each instance in the context of the documents instead of a sentence.

- WordNet sense granularity

Phenomenon: A considerable overlap with domains of the correct sense and predicted sense for the misclassified instance. (0.49 > random classifier 0.27) [via CSI inventory]

Bias: WordNet has a fine-grained annotations which is highly correlated.

Error Analysis

- Most frequent sense bias
- Insufficient context

Phenomenon: ESCHER mistakes most often appear in sentences with an average length of 27 tokens, roughly 5 tokens less than that in ALL (32)

Annotation bias: annotators consider each instance in the context of the documents instead of a sentence.

- WordNet sense granularity

Phenomenon: A considerable overlap with domains of the correct sense and predicted sense for the misclassified instance. (0.49 > random classifier 0.27) [via CSI inventory]

Bias: WordNet has a fine-grained annotations which is highly correlated.

Conclusion

- A new formulation for WSD – Extractive Sense Comprehension (ESC)
- More efficient use of the training data (1.7 higher than SOTA)
- Better generalization for zero-shot evaluation and few-shot learning.
- More flexible to scale across different inventories and combine them effectively.

- Note: another work *ConSec* expanding ESC has been published on EMNLP'21

Reference

[Rodd et al., 2004] Jennifer M Rodd, M Gareth Gaskell, and William D Marslen-Wilson. 2004. Modelling the effects of se- mantic ambiguity in word recognition. Cognitive science, 28(1):89–104.

[Miller et al., 1990] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to WordNet: An on-line lexical database. International journal of lexicography, pages 235–244, 1990.

[Navigli and Ponzetto, 2012] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence, pages 217–250, 2012.

[Miller et al., 1994] George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In Proceedings of the workshop on Human Language Technology, pages 240–243. Association for Computational Linguistics.

[Taghipour and Ng, 2015a] Kaveh Taghipour and Hwee Tou Ng. 2015a. One million sense-tagged instances for word sense disambiguation and induction. CoNLL 2015, pages 338-344.