ICML | 2020
Thirty-seventh International
Conference on Machine Learning

PoKED: A Semi-Supervised System for Word Sense Disambiguation

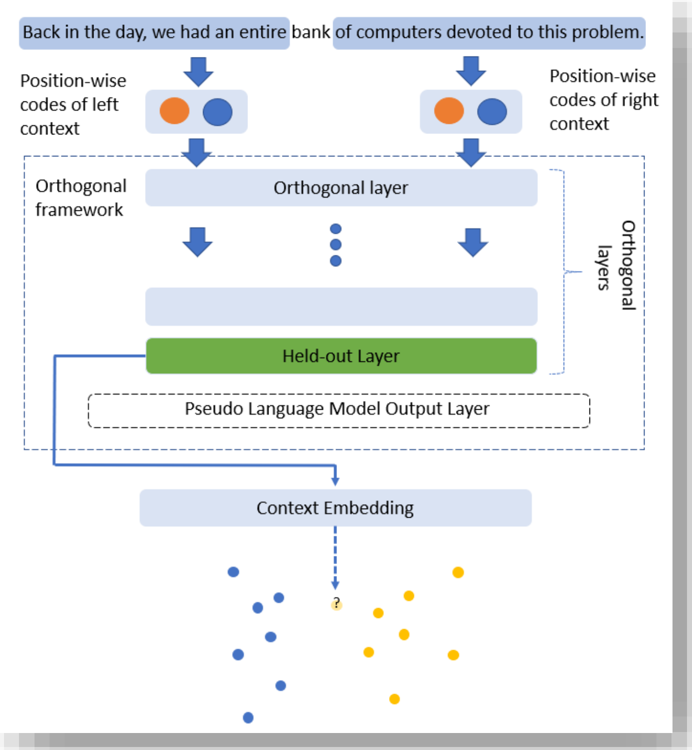
Feng Wei, Uyen Trang Nguyen
EECS, York University, Canada

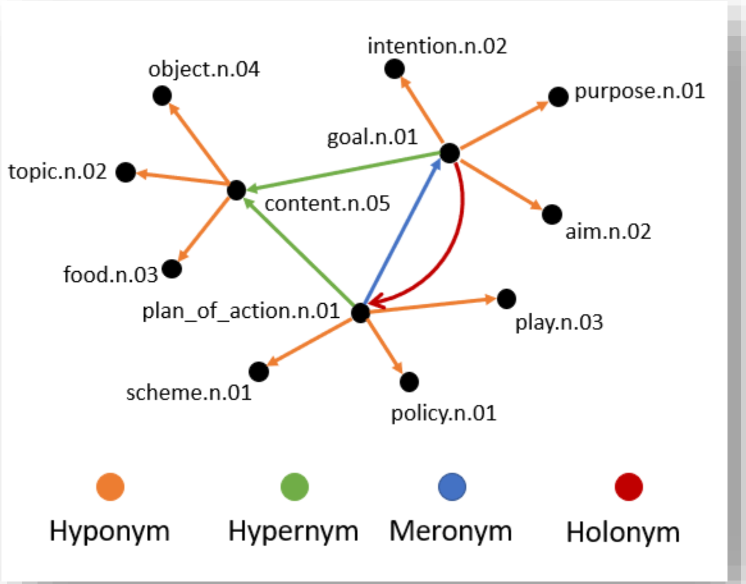
July 12 - 18, 2020 @ ICML 2020

Objective

- How position-wise embedding (unsupervised) could help the downstream WSD task.

- How information from descriptive linguistic knowledge graphs (WordNet) can be incorporated into neural network architectures to solve and improve the linguistic WSD task.





Contributions & Highlights

- Propose a semi-supervised neural system named Position-wise Orthogonal Knowledge-Enhanced Disambiguator (PoKED), supporting attention-driven, long-range dependency modeling.
- Incorporate position-wise encoding into an orthogonal framework and applies a knowledge-based attentive neural model to solve the WSD problem.
- Propose to use the semantic relations in the WordNet, by extracting semantic level inter-word connections from each document-sentence pair in the WSD dataset.
- PoKED achieves better performance than state-of-the-art knowledge-based WSD systems on standard benchmarks.

Human Semantic Knowledge

Human semantic knowledge is essential to WSD.

Document is a hypernym of *information*, or *information* is a hyponym of *document*.

Document: *This document is a summary of the European Public Assessment Report (EPAR). It explains how the Committee for Medicinal Products for Human Use (CHMP) assessed the studies performed, to reach their recommendations on how to use the medicine. If you need more information about your medical condition or your treatment, read the Package Leaflet (also part of the EPAR) or contact your doctor or pharmacist. If you want more information on the basis of the CHMP recommendation, read the Scientific Discussion (also part of the EPAR).*

...

Sentence: *This document is a summary of the European Public Assessment Report (EPAR).*

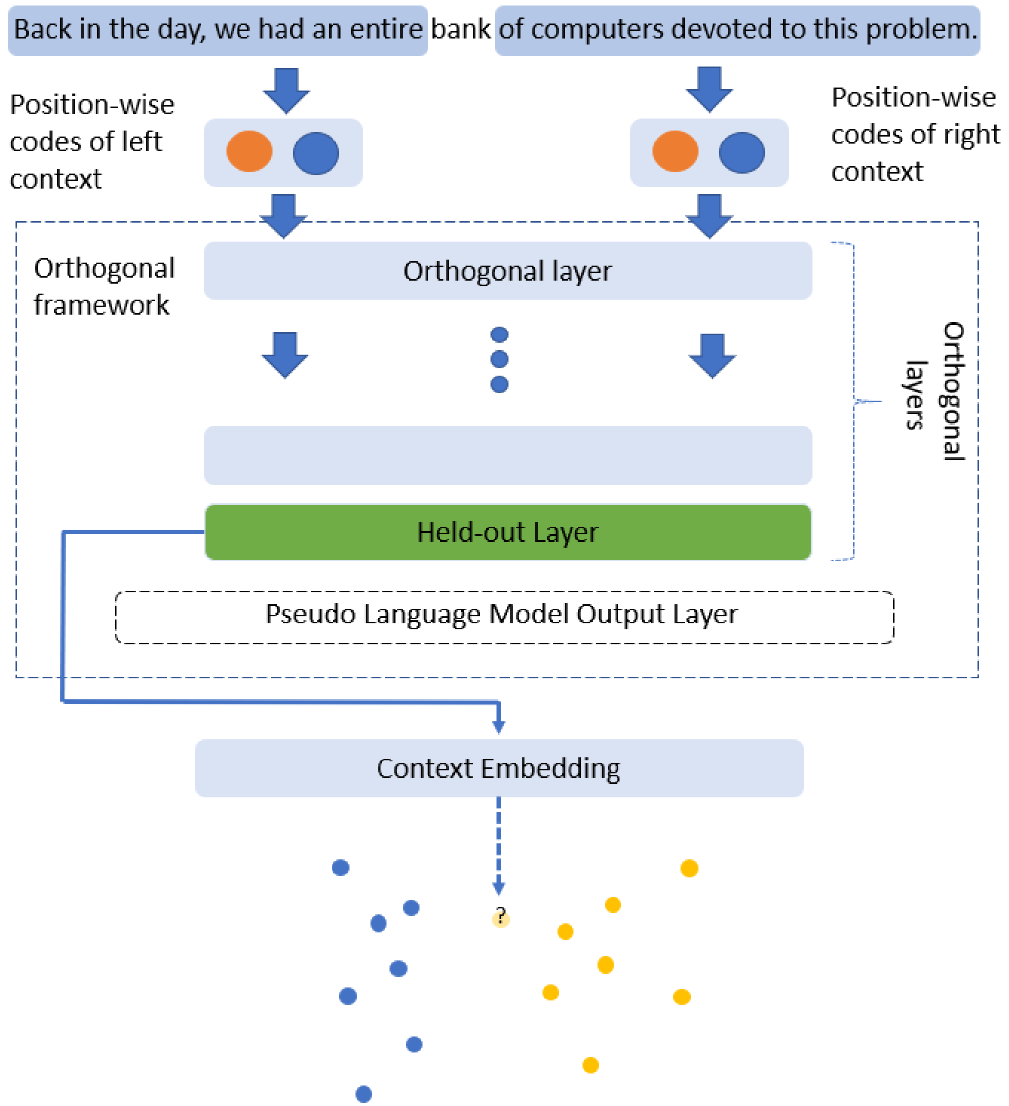
Answer: *<noun.communication>[10] S: (n) document.01 (document%1:10:00::), written document.01 (written_document%1:10:00::), papers.01 (papers%1:10:00::) (writing that provides information (especially information of an official nature))*

SemEval-15 dataset

PoNet (Unsupervised Language Model)

- Humans decide the sense of a polyseme by firstly understanding its occurring context [Harris, 1954].

- Two stages: PoNet to abstracts context as embeddings; KED to classify over pre-trained context embeddings.



Position-wise Encoding

$S = \{w_1, \cdots, w_N\}$: A sequence of *N* words from vocabulary *V*

$$\boxed{z_n = \alpha \cdot z_{n-1} + e_n}$$

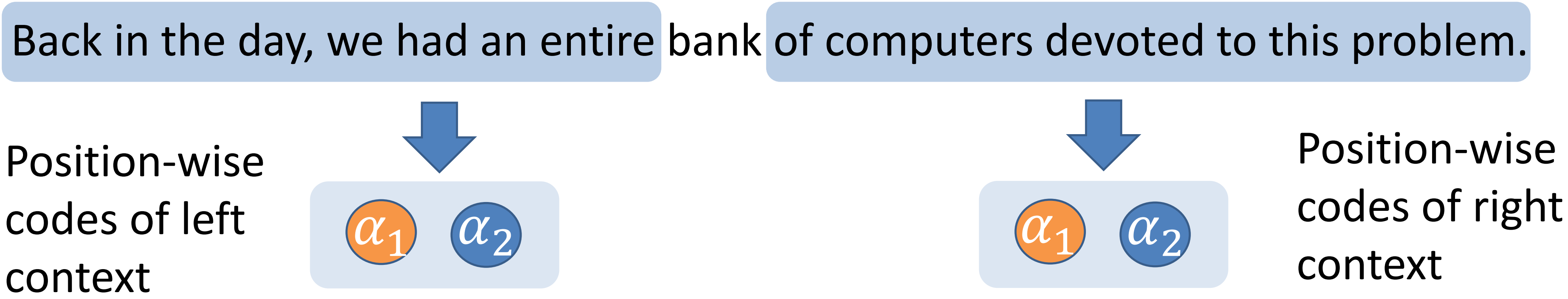
[Watcharawittayakul et al., 2018; Wei et al., 2019]

WORD	1-OF-K
w_0	1000000
w_1	0100000
w_2	0010000
w_3	0001000
w_4	0000100
w_5	0000010
w_6	0000001

ANY SEQUENCE	Position-wise Encoding
w_6	$0,0,0,0,0,0,1$
w_6, w_4	$0,0,0,0,1,0,\alpha$
w_6, w_4, w_5	$0,0,0,0,\alpha,1,\alpha^2$
w_6, w_4, w_5, w_0	$1,0,0,0,\alpha^2,\alpha,\alpha^3$
w_6, w_4, w_5, w_0, w_5	$\alpha,0,0,0,\alpha^3,1+\alpha^2,\alpha^4$
$w_6, w_4, w_5, w_0, w_5, w_4$	$\alpha^2,0,0,0,1+\alpha^4,\alpha+\alpha^3,\alpha^5$

Position-wise Encoding

- Generate augmented encoding codes by concatenating two codes using two different forgetting factors.
- Represent both short-term and long-term dependencies.
- Maintain the sensitivity to both nearby and faraway context.

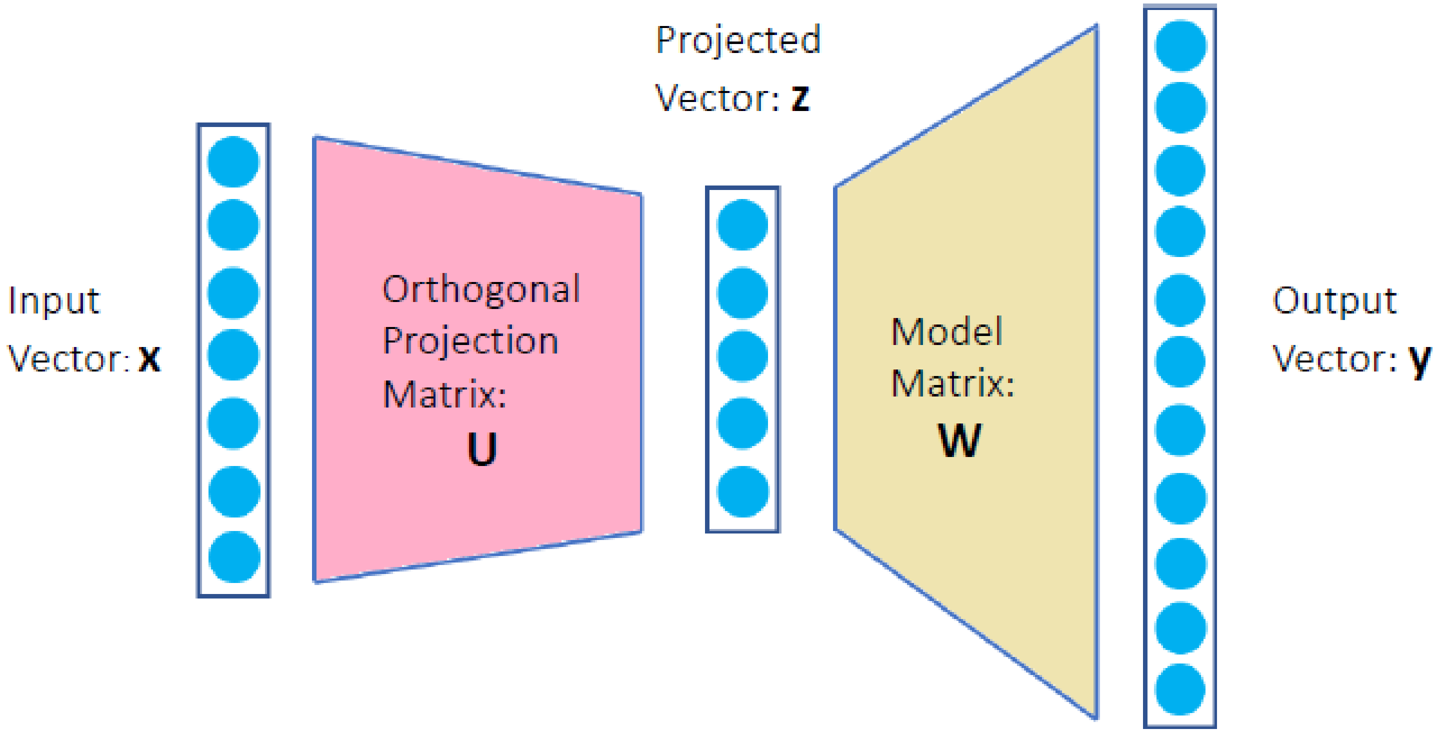


$$z_n = \boxed{\alpha} \cdot z_{n-1} + e_n$$

Orthogonal Framework

- Introduce a linear orthogonal projection to reduce the dimensionality of the raw high-dimension data and then uses a finite mixture distribution to model the extracted features.

- Each hidden layer can be viewed as an orthogonal model being composed of the feature extraction stage and data modeling stage.

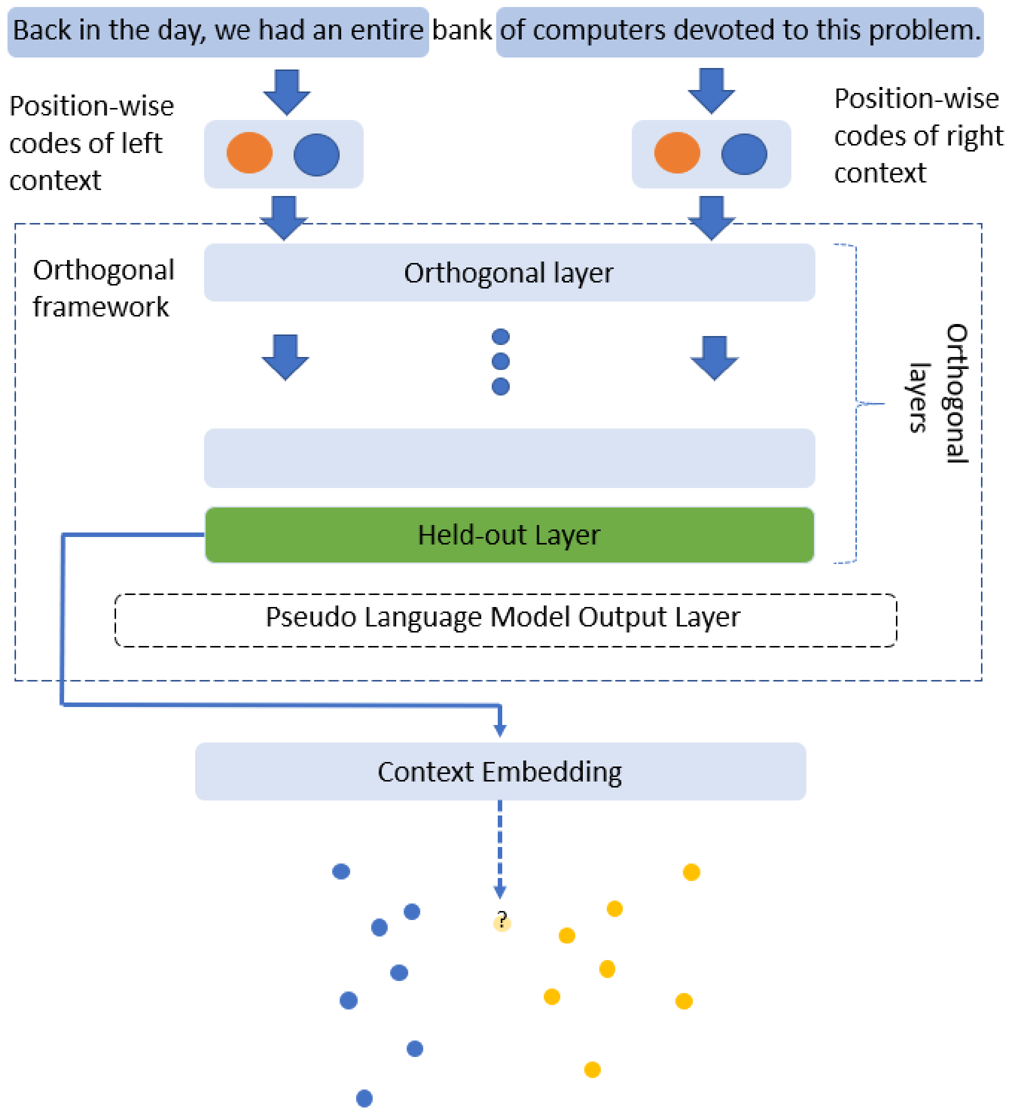


$$p(z) = \sum_{k=1}^{K} \pi_k \cdot f_k(z|\theta_k)$$

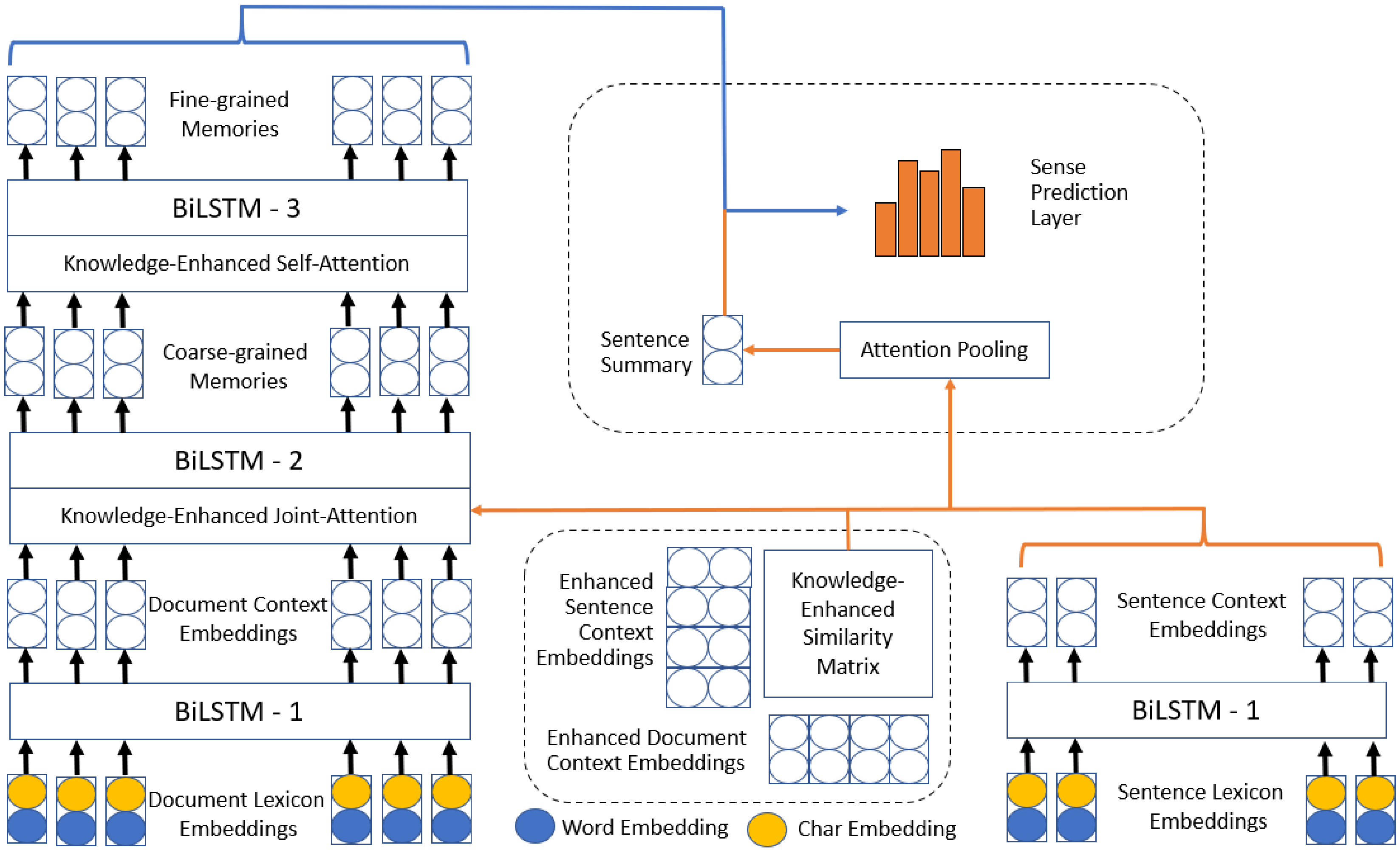
[Zhang et al., 2016; Wei et al., 2020]

Context Embeddings

Held-out layer are retained as context embeddings, which provides an effective representation of the surrounding context of a given target word.

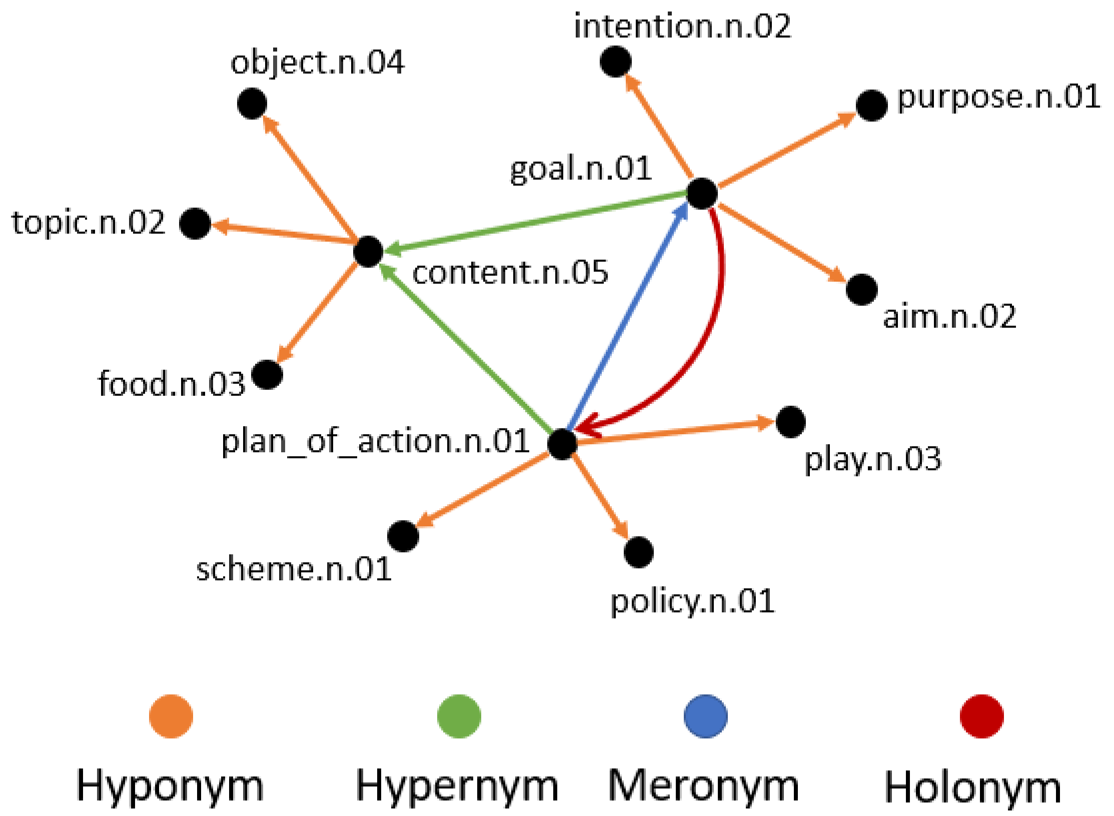


KED (Supervised Knowledge-based Attentive Model)



Data Enrichment with WordNet

For each word ω in a document-sentence pair, obtain a set Z_{ω} which contains the positions of the document words that ω is semantically connected to.



Data Enrichment with WordNet

Algorithm Extract semantic level inter-word connections from each document-sentence pair

---

procedure EXTRACT(D,S)
Input: Given a document D and a relevant sentence S.
Output: Return the extraction results on D and S
  for each document word d_i in D **do**
    $Z_{d_i} \leftarrow \{j \in \{1, ..., n\} \setminus \{i\} : (\Phi_{d_i} \bigcup \bar{\Phi}_{d_i}) \bigcap \Phi_{d_j} \neq \emptyset\}$ ▷ Obtain the extraction results Z_{d_i}
  for each sentence word s_i in S **do**
    $Z_{s_i} \leftarrow \{j \in \{1, ..., n\} \setminus \{i\} : (\Phi_{s_i} \bigcup \bar{\Phi}_{s_i}) \bigcap \Phi_{d_j} \neq \emptyset\}$ ▷ Obtain the extraction results Z_{s_i}

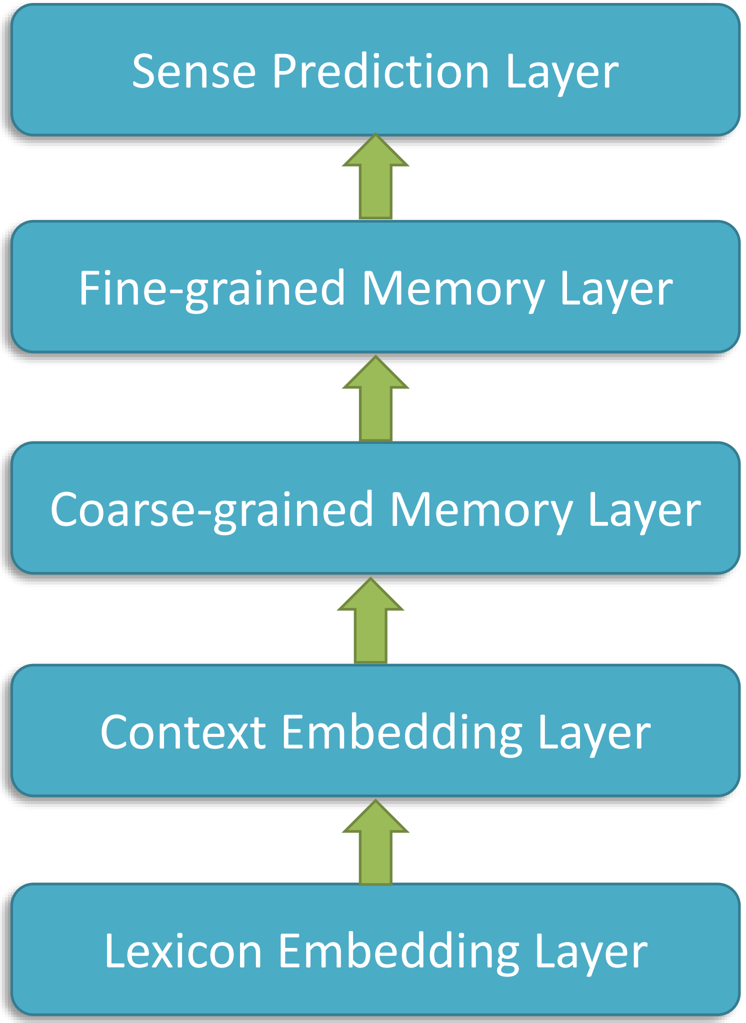
Φ_w : directly-involved synsets

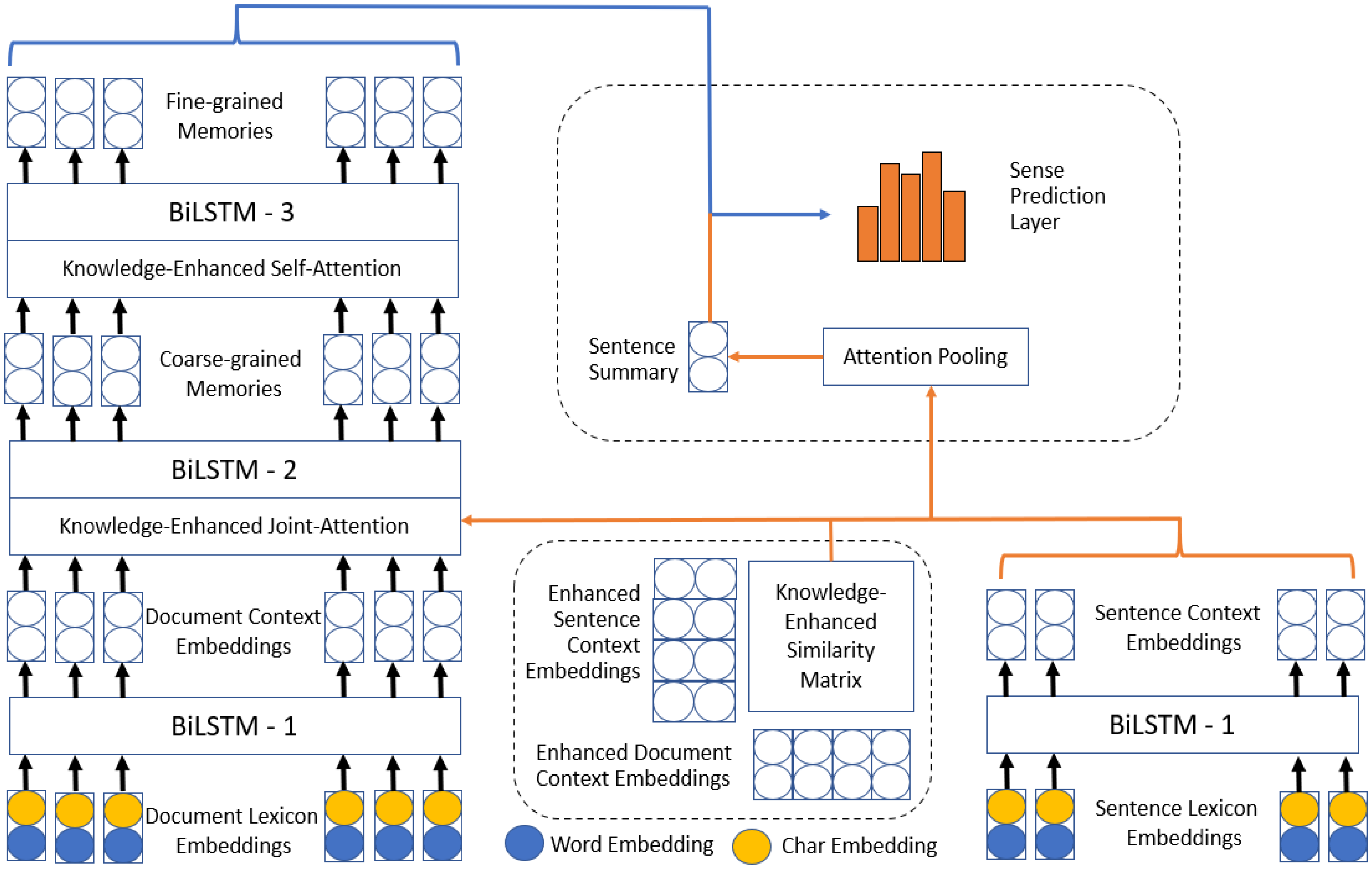
$\bar{\Phi}_w$: indirectly-involved synsets

keratin.n.01 → feather.n.01 → bird.n.01 → parrot.n.01

substance holonym | part holonym | hyponym

KED (Supervised Knowledge-based Attentive Model)





Experiments and Results

	Model	S2	S3	SE07	SE13	SE15	ALL	N	V	A	R
semi-sup.	*MFS*	64.7	65.4	53.9	62.9	66.6	64.1	68.1	49.5	74.1	80.6
	Babelfy	67.0	63.5	51.6	66.4	70.3	65.5	68.6	49.9	73.2	79.8
	ppr_{w2w}	68.8	66.1	53.0	**68.8**	70.3	67.3	-	-	-	-
	WSD-TM	69.0	66.9	55.6	65.3	69.6	66.9	69.7	51.2	76.0	**80.9**
	WSDG	68.9	65.5	54.5	67.0	**72.8**	67.2	70.4	51.3	75.7	80.6
	$PoKED_{\alpha}$ *(ours)*	68.8	**67.4**	54.6	65.2	70.5	67.0	70.1	50.8	75.5	79.7
	$PoKED_{\beta}$ *(ours)*	69.5	67.0	55.8	67.3	**72.8**	67.4	70.5	51.4	**76.2**	80.8
	$PoKED_{\gamma}$ *(ours)*	**69.8**	67.1	**56.0**	67.5	72.7	**67.7**	**70.8**	**51.6**	**76.2**	80.6
sup.	*IMS*	70.9	69.3	61.3	65.3	69.5	68.9	70.5	55.8	75.6	82.9
	IMS_{w2v}	72.2	70.4	62.6	65.9	71.5	70.1	71.9	56.6	75.9	84.7
	$Yuan_{LSTM}$	73.8	71.8	63.5	69.5	72.6	71.5	-	-	-	-
	$Raganato_{BLSTM}$	72.0	69.1	64.8	66.9	71.5	69.9	71.5	57.5	75.0	83.8
	GAS	72.2	70.5	-	67.2	72.6	-	-	-	-	-
	fastSense	73.5	73.5	62.4	66.2	73.2	-	-	-	-	-
	GLU-LW	75.5	73.4	68.5	71.0	76.2	-	-	-	-	-
	GlossBERT	76.5	73.4	69.2	75.1	79.5	-	-	-	-	-

Experiments and Results

Ablation Study on Knowledge-Enhancement

Model	S2	S3	SE07	SE13	SE15
PoKED_{α}	68.8	67.4	54.6	65.2	70.5
- w/o knowledge-enhancement	64.3	63.5	50.2	61.4	65.1
Performance Drop (%)	▼-4.5	▼-3.9	▼-4.4	▼-3.8	▼-5.4

Experiments and Results

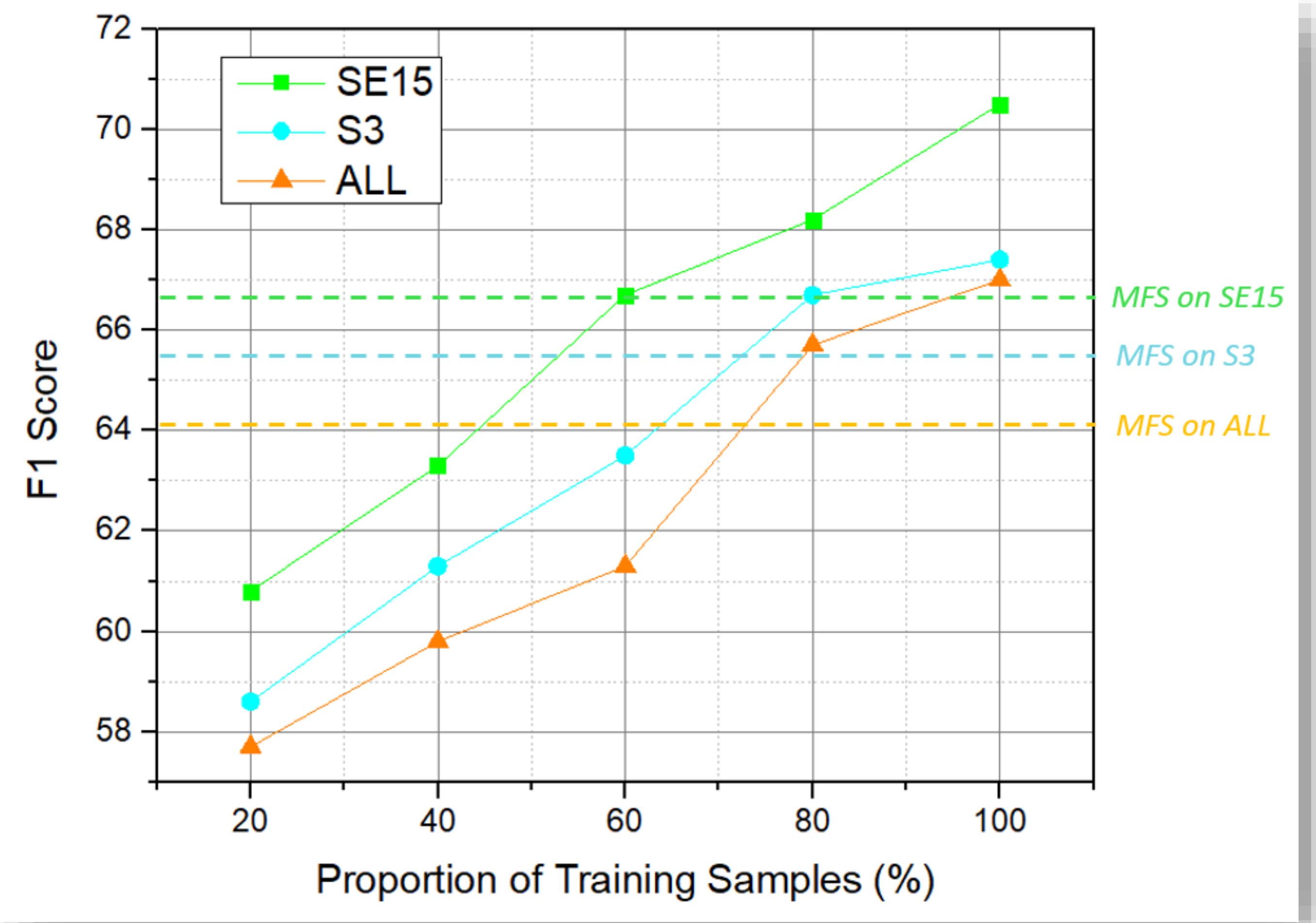
Effectiveness of General Knowledge Extraction

τ	#average	S2	#average	S3	#average	SE07	#average	SE13	#average	SE15
0	0.51	67.1	0.47	65.6	0.30	52.7	0.29	63.2	0.36	67.1
1	0.92	67.7	0.84	66.2	0.51	53.4	0.41	63.6	0.68	68.4
2	1.47	68.1	1.32	66.5	1.93	53.9	0.93	64.1	2.03	69.6
3	2.35	68.4	3.06	66.9	**3.28**	**54.6**	1.58	64.7	**3.97**	**70.5**
4	**3.89**	**68.8**	**3.97**	**67.4**	3.86	54.0	**3.67**	**65.2**	4.87	69.9
5	5.79	68.3	4.28	66.8	4.77	53.2	4.22	64.5	5.52	69.3
6	6.22	68.0	5.45	66.4	6.02	52.6	5.19	63.4	6.45	68.2

#average: average number of inter-word connections per word.
Bold font: best performance.

Experiments and Results

Quantitative Analysis of the Hunger for Data



MFS baseline: the Most Frequent Sense heuristic computed on SemCor corpus on each dataset.

YORK U
UNIVERSITÉ
UNIVERSITY

ICML | 2020
Thirty-seventh International Conference on Machine Learning

PoKED: A Semi-Supervised System for Word Sense Disambiguation

Feng Wei, Uyen Trang Nguyen
EECS, York University, Canada

Thank You