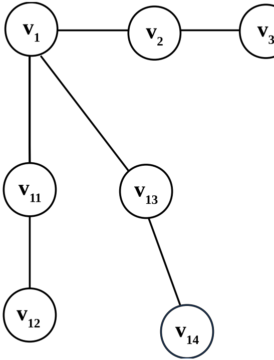
EDS-MEMBED

Multi-sense embeddings based on enhanced distributional semantic structures via a graph walk over word senses [1]



Eniafe Festus Ayetiran*, Petr Sojka, Vít Novotný

ayetiran@mail.muni.cz, sojka@fi.muni.cz, witiko@mail.muni.cz

Faculty of Informatics, Masaryk University

March 16, 2021

1. Introduction

- Word embeddings are popular, but conflate word meanings, a.k.a. *polysemy*.
- Multi-sense embeddings (M-SE) tackle polysemy by modeling word senses, not words.
- Existing approaches to M-SE include:

Multi-prototype [2, 3] Unsupervised, based on clustering, and

Sense-inventory-based [4, 5] Semi-supervised, based on knowledge bases, our approach.

Problems

- Sense-annotated corpora are small, a.k.a. *the knowledge acquisition bottleneck*.
- Existing approaches focus on word similarity task, not word sense disambiguation.

Our Contributions

- We propose augmenting sense-annotated corpora using semantic relations.
- We propose new semantic similarity measures for word sense disambiguation.

2. Related work

Existing approaches to M-SE include:
Multi-prototype [6, 2, 3, 7, 8], and
Sense-inventory-based [9, 4, 10, 11, 12, 13, 14, 5, 15, 16].
Existing NLP tasks for M-SE include:

Word similarity [17, 2, 3, 9, 4, 10, 11, 13, 7, 14, 5, 15],

Word sense disambiguation [9, 11, 13, 16],

Relational similarity [4],

Analogical reasoning [10, 7],

Sense clustering [13, 14],

Domain labeling [13],

Synonym recognition [15], and

Outlier detection [15].

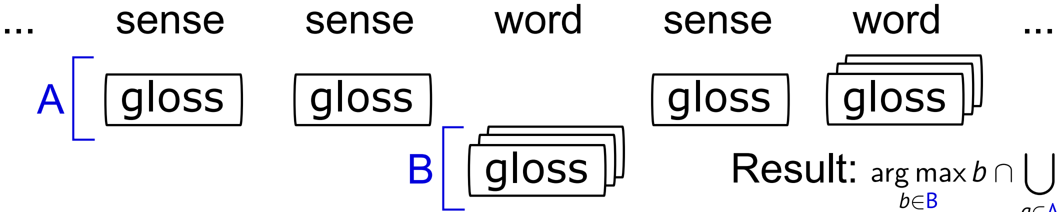
3. & 4. Preprocessing I

- For training our M-SE, we use the following sense-annotated corpora:

Semcor [18] 362 English texts comprising over 200,000 words, and

WordNet Gloss Tags (WNGT) [19] Semi-automatic sense annotations of *all WordNet glosses*.

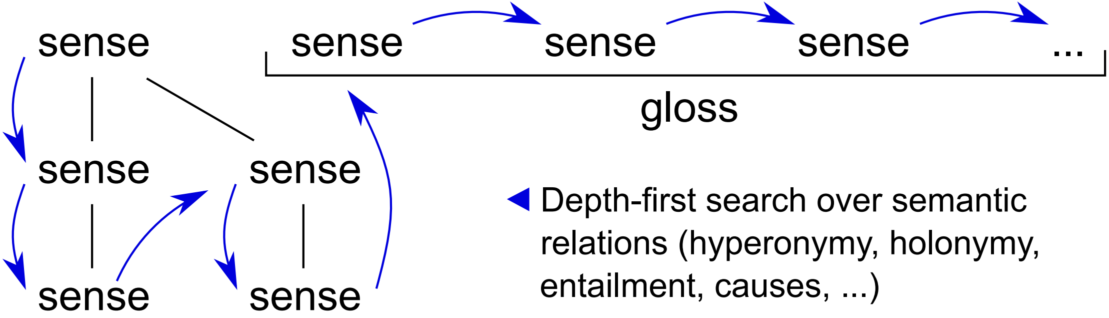
- First, we remove stopwords and we disambiguate unannotated words by maximum overlap of target and context sense glosses from WordNet using a Lesk algorithm [20]:



Result: $\arg\max_{b \in B} b \cap \bigcup_{a \in A} a$

3. & 4. Preprocessing II

- Second, we train M-SE using word2vec [21] and we use our word sense disambiguation algorithm (to be described later) to improve the disambiguation from first step. (MEMBED)
- Third, we augment WNGT using semantic relations from WordNet: (EDS-MEMBED)



◀ Depth-first search over semantic relations (hyperonymy, holonymy, entailment, causes, ...)

- See [1, Section 4.2] for hyperparameters and [1, tables 7 and 8] for their ablation study.

3. & 4. Preprocessing III

4.1 Augmentation Example

- In WNGT, the sense *permit.v.01* is defined by the gloss *consent to, give permission*.
- We remove the stopword *to* and disambiguate the gloss to *accept.v.03 give.v.09 permission.n.01*.
- Using WordNet, we expand *permit.v.01* to its hypernym *accept.v.03*.
- Using WordNet, we expand *permit.v.01* to its hyponyms:
  1. *admit.v.02*
  2. *admit.v.03*
  3. *allow.v.10*
  4. *authorize.v.01*
  5. *digest.v.03*
  6. *furlough.v.02*
  7. *give.v.40*
  8. *legalize.v.01*
  9. *privilege.v.01*
  10. *trust.v.02*
- We produce the following augmented sequence:
  permit.v.01 accept.v.03 admit.v.02 admit.v.03 allow.v.10 authorize.v.01 digest.v.03 furlough.v.02 give.v.40 legalize.v.01 privilege.v.01 trust.v.02 accept.v.03 give.v.09 permission.n.01
- Before training M-SE, we convert the senses to ⟨WordNet offset, part of speech⟩:
  00802318v 00797697v 02502536v 02449847v 00802946v 00803325v 00668099v 00748803v 00748972v 02481436v 02453692v 02481819v 00797697v 01629403v 06689297n

5. Experiments I

- We use the following word–word similarity measures:

AvgSim(w, w') [6] `np.mean(`$W \cdot W'^T$`)`, where W, W' are ℓ_2-normalized M-SE matrices for words w, w' and $\cdot$ is dot product, which gives us all pairwise cosine similarities,

MaxSim(w, w') [6] `np.max(`$W \cdot W'^T$`)`, where W, W' as above,

AvgSimC(w, w', c) [6] `np.mean(`$(W \cdot W'^T) \odot S$`)`, where $S = (W \cdot \bar{C}) \otimes (W' \cdot \bar{C})$, C is an M-SE matrix for context words c, $\bar{C} =$ `np.mean(C, axis=0)` is mean context M-SE, $\odot$ is Hadamard product, $\otimes$ is outer product, and W, W' as above,

MaxSimC(w, w', c) [6] `np.max(`$(W \cdot W'^T) \odot S$`)`, $S = (W \cdot \bar{C}) \otimes (W' \cdot \bar{C})$, where W, W', C as above,

globalSim(w, w') [3, 5] $\bar{W} \cdot \bar{W}'$, where W, W' as above.

- We propose the following sense–word similarity measures for word sense disambiguation:

AvgSimS(s, w') `np.mean(`$\vec{s} \cdot W'^T$`)`, where $\vec{s}$ is an M-SE for sense s and W' as above,

SumAvgSimS$(s, \mathbf{w}')$ $\sum_{w' \in \mathbf{w}'}$ AvgSimS(s, w'), where $\mathbf{w}'$ is a set of context words,

MaxSimS(s, w') `np.max(`$\vec{s} \cdot W'^T$`)`, where $\vec{s}, W'$ as above,

SumMaxSimS$(s, \mathbf{w}')$ $\sum_{w' \in \mathbf{w}'}$ MaxSimS(s, w'), where $\mathbf{w}'$ as above.

5. Experiments II

5.3 Word Sense Disambiguation I

- We disambiguate word w in context $\mathbf{w}'$ as $\arg\max_{s \in w} f(s, \mathbf{w}'), f \in \{\text{SumAvgSimS, SumMaxSimS}\}$.
- For words with no context or negative arg max, we assign the first WordNet sense.
- We evaluate on six datasets, see [1, Section 5.3] for details:
  1. Senseval-2 (SE2) [22]
  2. Senseval-3 English all-words task (SE3) [23]
  3. SemEval-2007 task 17 (SE07-17) [24]
  4. SemEval-2007 task 07 (SE07-07) [25]
  5. SemEval-2013 task 12 (SE13-12) [24]
  6. SemEval-2015 task 13 (SE15-13) [25]
- In our comparison, we distinguish M-SE-based and non-M-SE-based systems.
- In our comparison, we also distinguish four kinds of approaches:

Unsupervised Based on large unannotated corpora, worst performance,

Knowledge-based Based on sense inventories and lexical resources, our approach,

Semi-supervised Based on small sense-annotated corpora and large unannotated corpora,

Supervised Based on small sense-annotated corpora, best performance.

5. Experiments III

5.3 Word Sense Disambiguation II

Table 3

F_1 (%) of EDS-MEMBED in comparison with MEMBED and state-of-the-art embeddings-based WSD systems on the benchmark datasets. Best results for each approach in bold.

Approach	System	SE2	SE3	SE07-17	SE13-12	SE15-13	SE07-07
Supervised	IMS+Word2Vec(OMSTI)	68.3	68.2	**59.1**	–	–	–
	IMS + S-product	66.8	**73.6**	–	–	–	–
Semi-Supervised	Chen et al.(Model S2C)	–	–	–	–		**82.6**
Knowledge-based	+LS	58.4	59.4	–	–	–	–
	Chen et al.(Model S2C)	–	–	–	–	–	75.8
	VecLesk(baroni c)	–	–	–	–	58.0	75.3
	VecLesk(glove)	–	–	–	–	59.0	73.0
	MEMBED$_{AvgS}$	65.7	58.8	47.5	64.2	69.2	78.6
	MEMBED$_{MaxS}$	66.1	58.9	48.4	64.4	70.3	78.9
	EDS-MEMBED$_{AvgS}$	66.3	**60.4**	46.2	65.6	**72.4**	78.5
	EDS-MEMBED$_{MaxS}$	**66.6**	59.9	47.0	**66.7**	71.9	**79.1**

5. Experiments IV

5.3 Word Sense Disambiguation III

Table 4

F_1 (%) of EDS-MEMBED in comparison with MEMBED and state-of-the-art WSD systems on the benchmark datasets. Best results for each approach in bold. † indicates values not given by authors but deduced [48].

Approach	System	SE2	SE3	SE07-17	SE13-12	SE15-13	SE2+SE3+SE07-17+SE13-12+SE15-13					SE07-07
							Nouns	Verbs	Adj.	Adv.	ALL	
Supervised	$LMMS_{2348}$(BERT)	76.3	75.6	68.1	75.1	77.0	–	–	–	–	75.4	–
	HCAN	72.8	70.3	–	68.5	72.8	72.7	58.2	77.4	84.1	71.1	–
	LSTMLP	73.8	71.8	63.5	69.5	72.6	†73.9	–	–	–	†71.5	83.6
	GAS	72.2	70.5	–	67.2	72.6	72.2	57.7	76.6	85.0	70.6	–
	BERT	75.5	73.6	68.1	71.1	76.2	–	–	–	–	74.1	–
	GLOSSBERT	77.7	75.2	72.5	76.1	80.4	79.3	66.9	78.2	86.4	77.0	–
	SemCor+WNGC	**79.7**	**77.8**	**73.4**	**78.7**	**82.6**	**81.4**	**68.7**	**83.7**	**85.5**	**79.0**	**90.4**
	MFS	65.6	66.0	54.5	63.8	67.1	67.7	49.8	73.1	80.5	65.5	78.9
Knowledge-based	KEF	**69.6**	66.1	**56.9**	68.4	72.3	**71.9**	**51.6**	74.0	80.6	**68.0**	–
	Babelfy	67.0	63.5	51.6	66.4	70.3	68.6	49.9	73.2	79.8	65.5	–
	UKB	68.8	66.1	53.0	**68.8**	70.3	–	–	–	–	67.3	–
	WSD-TM	69.0	**66.9**	55.6	65.3	69.6	69.7	51.2	**76.0**	**80.9**	66.9	–
	WN 1st sense	66.8	66.2	55.2	63.0	67.8	67.6	50.3	74.3	**80.9**	65.2	78.9
	$MEMBED_{AvgS}$	65.7	58.8	47.5	64.2	69.2	66.7	47.2	69.3	73.7	62.9	78.6
	$MEMBED_{MaxS}$	66.1	58.9	48.4	64.4	70.3	66.8	48.7	68.8	75.7	63.3	78.9
	$EDS\text{-}MEMBED_{AvgS}$	66.3	60.4	46.2	65.6	**72.4**	68.2	48.4	69.9	74.9	64.2	78.5
	$EDS\text{-}MEMBED_{MaxS}$	66.6	59.9	47.0	66.7	71.9	68.0	49.2	70.3	76.0	64.4	**79.1**

5. Experiments V

5.4 Word Similarity I

- For words w, w', we compute similarity as $f(w, w'), f \in \{\text{AvgSim, MaxSim, globalSim}\}$.
- For words w, w' in context c, we use $f(w, w', c), f \in \{\text{AvgSimC, MaxSimC, globalSim}\}$.
- We evaluate on five datasets, see [1, Section 5.4] for details:
  1. RG65, MC28, MEN [26, 27, 28]
  2. SimLex999 [29]
  3. WordSim-353 [30]

Table 5
Spearman's correlation coefficient ($\rho \times 100$) of models on the RG65, MC28, and MEN datasets. Best performing models in bold.

Model	RG65			MC28			MEN		
	AvgSim	MaxSim	globalSim	AvgSim	MaxSim	globalSim	AvgSim	MaxSim	globalSim
GloVe	–	–	82.9	–	–	83.6	–	–	–
Retro	–	–	84.2	–	–	–	–	–	75.9
Word2Vec	–	–	75.4	–	–	–	–	75.0	–
SENSEMBED	87.1	89.4	–	–	88.0	–	**80.5**	77.9	–
MSSA	82.8	87.8	85.9	84.5	88.8	87.5	78.5	74.4	79.5
DeConf	–	**89.6**	–	–	–	–	–	78.6	–
SW2V	–	74.0	–	–	–	–	–	76.0	–
MEMBED	64.4	85.0	83.6	70.2	86.3	81.7	55.2	67.9	58.3
EDS-MEMBED	77.1	**89.6**	86.9	83.2	**89.6**	88.4	63.3	68.4	64.3

5. Experiments VI

5.4 Word Similarity II

Table 6
Spearman's correlation coefficient ($\rho \times 100$) of models on SimLex999 and WordSim-353. Best performing models in bold.

Model	SimLex999			WordSim-353		
	AvgSim	MaxSim	globalSim	AvgSim	MaxSim	globalSim
GloVe	–	–	–	–	–	75.9
Retro	–	–	–	–	–	61.2
Word2Vec	–	39.0	–	–	–	–
SENSEMBED	–	–	–	**77.9**	71.4	–
Huang et al.	–	–	–	64.2	–	22.8
Pruned-TF–IDF	–	–	–	–	–	73.4
MSSG	–	–	–	68.6	–	69.1
MSSA	46.9	38.5	43.9	73.0	66.2	73.7
DeConf	–	51.7	–	–	–	–
SW2V	–	47.0	–	–	71.0	–
Chen et al.	–	43.0	–	–	–	–
MEMBED	37.0	47.6	33.2	45.6	54.0	44.6
EDS-MEMBED	48.0	**53.7**	46.4	59.5	59.6	58.2

5. Experiments VII

5.6 Computational Complexity

Table 9

Training data size, training time, dimension of the models and the hardware environment.

Model	Data size (tokens)	Training time	Dimension	Hardware
Huang at al.	990 million	168 h	50	–
MSSG	990 million	6 h	300	–
SENSEMBED	3 billion	–	400	–
MSSA	540 million	–	1000	–
Word2Vec	783 million	24 h	300	–
GloVe	840 billion	85 m/6 billion	300	Intel CPU @ 2.1 GHz $\times$ 32
MEMBED	1,302,549	12 m:19 s	100	Intel CPU @ 1.60 GHz $\times$ 4, 11.7 GiB RAM
EDS-MEMBED	1,590,998	16 m:26 s	200	Intel CPU @ 1.60 GHz $\times$ 4, 11.7 GiB RAM

Conclusion

- Word embeddings suffer from polysemy.
- Multi-sense embeddings (M-SE) solve polysemy, but suffer from lack of training data.
- We propose an approach that covers all word senses using WordNet.
- We also adapt existing word similarity measures to word sense disambiguation.
- Our approach is *efficient* and *competitive* on word similarity and word sense disambiguation.

Thank You for Your Attention!

Bibliography I

[1] Eniafe Festus Ayetiran et al. “EDS-MEMBED: Multi-sense embeddings based on enhanced distributional semantic structures via a graph walk over word senses”. In: *Knowledge-Based Systems* (2021), p. 106902. ISSN: 0950-7051. DOI: `http://dx.doi.org/10.1016/j.knosys.2021.106902`.

[2] Eric H. Huang et al. “Improving Word Representations via Global Context and Multiple Word Prototypes”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea*. ACL, 2012, pp. 873–882.

[3] Ignacio Iacobacci et al. “SENSEMBED: Learning Sense Embeddings for Word and Relational Similarity”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Beijing, China. Volume 1: Long Papers*. ACL, 2015, pp. 95–105.

Bibliography II

[4] Mohammad Taher Pilehvar and Nigel Collier. "De-Conflated Semantic Representations". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP, Austin, Texas, USA*. 2016, pp. 1680–1690.

[5] Lev Finkelstein et al. "Placing search in context: the concept revisited". In: *ACM Transactions on Information Systems* 20.1 (2002), pp. 116–131.

[6] Xinxiong Chen et al. "A Unified Model for Word Sense Representation and Disambiguation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, Doha, Qatar*. 2014, pp. 1025–1035.

[7] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *Transactions of Association of Computational Linguistics* 5 (2017), pp. 135–146.

Bibliography III

[8] Massimiliano Mancini et al. “Embedding Words and Senses Together via Joint Knowledge-Enhanced Training”. In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada*. 2017, pp. 100–111.

[9] Tao Chen et al. “Improving Distributed Representation of Word Sense via WordNet Gloss Composition and Context Clustering”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing ACL, Beijing, China. Volume 2: Short Papers*. ACL, 2015, pp. 15–20.

[10] Sascha Rothe and Hinrich Schütze. “AutoExtend: Combining Word Embeddings with Semantic Resources”. In: *Computational Linguistics* 43.3 (2017), pp. 593–617.

Bibliography IV

[11] Kurt D. Bollacker et al. “Freebase: a collaboratively created graph database for structuring human knowledge”. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada*. 2008, pp. 1247–1250.

[12] Birgit Hamp and Helmut Feldweg. “GermaNet - a Lexical-Semantic Net for German”. In: *Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, ACL*. 1997, pp. 9–15.

[13] Ben Athiwaratkun et al. “Probabilistic FastText for Multi-Sense Word Embeddings”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia. Volume 1: Long Papers*. ACL, 2018, pp. 1–11.

Bibliography V

[14] Ignacio Iacobacci and Roberto Navigli. “LSTMEmbed: Learning Word and Sense Representations from a Large Semantically Annotated Corpus with Long Short-Term Memories”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL, Florence, Italy. Volume 1: Long Papers*. 2019, pp. 1685–1695.

[15] George A. Miller et al. “A Semantic Concordance”. In: *Proceedings of the ARPA Workshop on Human Language Technology*. 1993, pp. 303–308.

[16] Helen Langone et al. “Annotating WordNet”. In: *Proceedings of the Workshop Frontiers in Corpus Annotation@HLT-NAACL 2004, Boston, MA, USA*. 2004, pp. 63–69.

[17] Arvind Neelakantan et al. “Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, Doha, Qatar*. ACL, 2014, pp. 1059–1069.

Bibliography VI

[18] Loïc Vial et al. "UFSAC: Unification of Sense Annotated Corpora and Tools". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation LREC, Miyazaki, Japan*. Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA), 2018, pp. 1027–1034.

[19] Eniafe Festus Ayetiran and Kehinde Kayode Agbele. "An Optimized Lesk-Based Algorithm for Word Sense Disambiguation". In: *Open Computer Science* 8.1 (2018), pp. 165–172.

[20] Joseph Reisinger and Raymond J. Mooney. "Multi-Prototype Vector-Space Models of Word Meaning". In: *Human Language Technologies: Proceedings of Conference of the North American Chapter of the Association of Computational Linguistics, Los Angeles, California, USA*. ACL, 2010, pp. 109–117.

Bibliography VII

[21] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2013. URL: `https://arxiv.org/abs/1301.3781`.

[22] Michael Lesk. "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone". In: *Proceedings of the 5th ACM-SIGDOC Conference*. ACM, 1986, pp. 24–26.

[23] Sameer Pradhan et al. "SemEval-2007 Task-17: English Lexical Sample, SRL and All Words". In: *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic*. Ed. by Eneko Agirre et al. ACL, 2007, pp. 87–92.

Bibliography VIII

[24] Roberto Navigli et al. "SemEval-2007 Task 07: Coarse-Grained English All-Words Task". In: *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic*. Ed. by Eneko Agirre et al. ACL, 2007, pp. 30–35.

[25] Roberto Navigli et al. "SemEval-2013 Task 12: Multilingual Word Sense Disambiguation". In: *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA*. Ed. by Mona T. Diab et al. ACL, 2013, pp. 222–231.

[26] Manaal Faruqui et al. "Retrofitting Word Vectors to Semantic Lexicons". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA*. ACL, 2015, pp. 1606–1615.

Bibliography IX

[27] Joseph Reisinger and Raymond J. Mooney. “A Mixture Model with Sharing for Lexical Semantics”. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, MIT Stata Center, Massachusetts, USA*. ACL, 2010, pp. 1173–1182.

[28] Eniafe Festus Ayetiran. “An index-based joint multilingual/cross-lingual text categorization using topic expansion via BabelNet”. In: *Turkish Journal of Electrical Engineering & Computer Sciences* 28.1 (2020), pp. 224–237. DOI: `10.3906/elk-1901-140`.

Bibliography X

[29] Ignacio Iacobacci et al. "Embeddings for Word Sense Disambiguation: An Evaluation Study". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016, pp. 897–907. DOI: 10.18653/v1/p16-1085. URL: https://doi.org/10.18653/v1/p16-1085.

[30] Zhi Zhong and Hwee Tou Ng. "It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text". In: *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, System Demonstrations*. The Association for Computer Linguistics, 2010, pp. 78–83. URL: https://www.aclweb.org/anthology/P10-4014/.

MUNI

FACULTY

OF INFORMATICS