Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation

Gábor Borbély[1], Márton Makrai[2], Dávid Nemeskey[3], András Kornai[3]

[1]Department of Algebra, Budapest University of Technology and Economics, `borbely@math.bme.hu`

[2]Institute for Linguistics, Hungarian Academy of Sciences, `makrai.marton@nytud.mta.hu`

[3]Institute for Computer Science, Hungarian Academy of Sciences, `{nemeskeyd,kornai}@sztaki.hu`

Semantic resolution (overview)

- motivation: 1 vector per concept (underlying word sense unit)
- usage splits words more finely than semantics, synonyms and near-synonyms end up in distant clusters (Reisinger and Mooney, 2010)
- two evaluation scenarios proposed + sanity check
  1. number of senses of each word, compared to monolingual dictionaries
  2. word translation from MSEs with the linear method (Mikolov et al., 2013b)
- different but related senses
  - homonymy: Russian *mir* 'world' and *mir* 'peace'
  - polysemy: Hungarian *nap* 'day' and *nap* 'sun'

Multi-sense embeddings (MSE)

- `huang`: spherical context clustering (Huang et al., 2012)
- `neela`: multi-sense skip-gram (Neelakantan et al., 2014)
- Tian et al. (2014)
- AdaGram (Bartunov et al., 2015)
- `jiweil`: Chinese Restaurant Process (Li and Jurafsky, 2015)

Monolingual dictionaries

- English
  - Collins-COBUILD (CED, Sinclair (1987)): semantic distinctions > POS
  - Longman dictionary (LDOCE, Boguraev and Briscoe (1989)) POS-level split > semantic split
- Hungarian (some processing from Miháltz (2010); Recski et al. (2016))
  - Comprehensive Dictionary of Hungarian (NSZ, Ittzés (2011)): semantics > POS
  - Explanatory Dictionary of Hungarian (EKSZ, Pusztai (2003)): POS > semantics
- low-resource simulated
  - no machine-readable monolingual dictionary
  - dictionary extracted from `OSub`, the OpenSubtitles parallel corpus (Tiedemann, 2012) automatically
  - number of the senses ≈ number of words it translates to, averaged among many languages.
  - unigram perplexity of the translations instead of their count – reduce noise in dictionary

Corpora

- en: UMBC Webbase (Han et al., 2013)
- hu: Webkorpusz (Halácsy et al., 2004)

Correlation with dictionaries

Resources compared	n	ρ
LDOCE vs CED	23702	0.266
EKSZ vs NSZ (b)	3484	0.648
`neela.30k` vs CED	23508	0.089
`neela.NP.6k` vs CED	23508	0.084
`neela.NP.30k` vs CED	23508	0.112
`neela.30k` vs LDOCE	21715	0.226
`neela.NP.6k` vs LDOCE	21715	**0.292**
`neela.NP.30k` vs LDOCE	21715	0.278
`huang` vs CED	23706	0.078
`huang` vs LDOCE	21763	**0.280**
`neela.s4` vs EKSZ	45401	0.067
`jiweil` vs EKSZ	32007	0.023
AdaGram vs EKSZ	26739	0.086
AdaGram.a05 vs EKSZ	26739	0.088
`neela.30k` vs `huang`	99156	0.349
`neela.NP.6k` vs `huang`	99156	**0.901**
`neela.NP.30k` vs `huang`	99156	0.413
`neela.s4` vs `jiweil`	283083	0.123
AdaGram vs `neela.s4`	199370	**0.389**
AdaGram vs `jiweil`	201291	0.140

Table 1: Word sense distribution similarity between various resources

Confounding factors

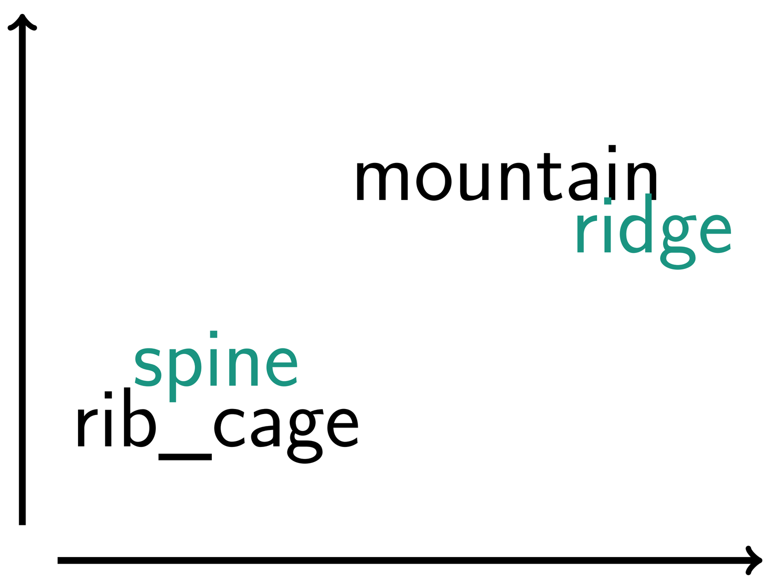
Resources compared	n	ρ
CED vs POS	42532	0.052
LDOCE vs POS	28549	**0.206**
`OSub` vs POS	48587	**0.141**
EKSZ vs POS	52158	0.080
NSZ vs POS	3532	0.046
`huang` vs POS	98405	0.026
AdaGram vs freq	399985	0.343
`huang` vs freq	94770	0.376
CED vs freq	36709	0.124
LDOCE vs freq	27859	0.317
`neela.s4` vs freq	94044	**0.649**
`neela.NP.30k` vs freq	94044	0.368
`neela.NP.6k` vs freq	94044	**0.635**
UMBC POS vs freq	136040	-0.054

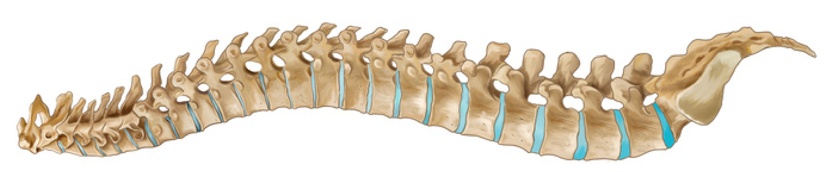
Table 2: Word sense distribution similarity with POS tag perplexity (top panel) and word frequency (bottom panel)

Conclusion of sense numbers

- exponential decay in lexicons not reflected by embeddings
- the real induction problem is 1-sense vs ambiguous

Resource	1	2	3	4	5	6+	Size	Mean	Std
CED	80,003	1,695	242	69	13	2	82,024	1.030	0.206
LDOCE	26,585	3,289	323	56	11	1	30,265	1.137	0.394
`OSub`	58,043	14,849	2,259	431	111	25	75,718	1.354	0.492
AdaGram	122,594	330,218	11,341	5,048	7,626	0	476,827	1.836	0.663
`huang`	94,070	0	0	0	0	6,162	100,232	1.553	2.161
`neela.30k`	69,156	0	30,000	0	0	0	99,156	1.605	0.919
`neela.NP.6k`	94,165	2,967	1,012	383	202	427	99,156	1.101	0.601
`neela.NP.30k`	71,833	20,175	4,844	1,031	439	834	99,156	1.411	0.924
`neela.s4`	574,405	0	0	4,000	0	0	578,405	1.021	0.249
EKSZ	66,849	628	57	11	1	0	121,578	1.012	0.119
NSZ (b)	5,225	122	13	3	0	0	5,594	1.029	0.191
`OSub`	159,843	9,169	229	3	0	0	169,244	1.144	0.199
AdaGram	135,052	76,096	15,353	5,448	6,513	0	238,462	1.626	0.910
`jiweil`	57,109	92,263	75,710	39,624	15,153	5,997	285,856	2.483	1.181
`neela.s2`	767,870	4,000	0	0	0	0	99,156	1.005	0.072
`neela.s4`	767,870	0	0	4,000	0	0	99,156	1.016	0.215





Linear translation

- Mikolov et al. (2013b)

$$W : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2} \qquad z \approx Wx$$

- learning the mapping: supervised by a seed dictionary

$$\min_W \sum_i ||Wx_i - z_i||^2$$

- 5 K train + 1 K test
- generate or score translations
- different metrics, Euclidean distance in training and cosine similarity in collection of translations.

. . . from MSEs

- idea: from MSE to traditional (single)
- MSEs contain global vectors for each word form
- mapping trained in traditional fashion with global vectors as sources
- seed: `OSub`
- generation of translations
  - baseline: traditional (with global as source)
  - sources is MSE, target remains single-sense
  - evaluation
- target embedding from Mikolov et al. (2013a)

	global	sense
AdaGram 800 a.05 m100	21.7%	25.1%
AdaGram 800 a.01 m100	10.8%	21.0%
jiweil	32.2%	8.3%

Table 3: Precision of Hungarian to English translation

Glue code for the experiments

- `https://github.com/hlt-bme-hu/multiwsi`

References

S. Bartunov, D. Kondrashkin, A. Osokin, and D. Vetrov. Breaking sticks and ambiguities with adaptive skip-gram. *ArXiv preprint*, 2015.

B. K. Boguraev and E. J. Briscoe. *Computational Lexicography for Natural Language Processing*. Longman, 1989.

P. Halácsy, A. Kornai, L. Németh, A. Rung, I. Szakadát, and V. Trón. Creating open language resources for Hungarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 203–210. ELRA, 2004.

L. Han, A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese. Umbc_ebiquity-core: Semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 44–52, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

E. Huang, R. Socher, C. Manning, and A. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 873–882, Jeju Island, Korea, 2012. Association for Computational Linguistics.

N. Ittzés, editor. *A magyar nyelv nagyszótára III-IV*. Akadémiai Kiadó, 2011.

J. Li and D. Jurafsky. Do multi-sense embeddings improve natural language understanding? In *EMNLP*, 2015.

M. Miháltz. *Semantic resources and their applications in Hungarian natural language processing*. PhD thesis, Pázmány Péter Catholic University, 2010. URL `https://itk.ppke.hu/uploads/articles/163/file/Mihaltz_diss.pdf`.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *Proceedings of the ICLR 2013*, 2013a.

T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. Xiv preprint arXiv:1309.4168, 2013b.

A. Neelakantan, J. Shankar, A. Passos, and A. McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *EMNLP*, 2014.

F. Pusztai, editor. *Magyar értelmező kéziszótár*. Akadémiai Kiadó, 2003.

G. Recski, G. Borbély, and A. Bolevácz. Building definition graphs using monolingual dictionaries of Hungarian. In A. Tanács, V. Varga, and V. Vincze, editors, *XI. Magyar Számitógépes Nyelvészeti Konferencia [11th Hungarian Conference on Computational Linguistics]*, 2016.

J. Reisinger and R. J. Mooney. Multi-prototype vector-space models of word meaning. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics, 2010.

J. M. Sinclair. *Looking up: an account of the COBUILD project in lexical computing*. Collins ELT, 1987.

F. Tian, H. Dai, J. Bian, B. Gao, R. Zhang, E. Chen, , and T. Y. Liu. A probabilistic model for learning multi-prototype word embeddings. In *COLING*, pages 151–160, 2014.

J. Tiedemann. Parallel data, tools and interfaces in OPUS. In N. Calzolari, editor, *LREC*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.