Eötvös Loránd University
Faculty of Humanities
Thesis booklet

MÁRTON MAKRAI

SYMBOLIC AND DISTRIBUTED WORD REPRESENTATIONS

CHAPTERS ON LEXICAL RELATIONS AND CROSS-LINGUAL METHODS

DOI: 10.15476/ELTE.2023.350

Doctoral School of Linguistics
Head: Gábor Tolcsvai Nagy MHAS

Theoretical Linguistics PhD Programme
Head: Zoltán Bánréti CSc

Members of the Committee:
Miklós Törkenczy DSc, chair
Attila Novák PhD, opponent
Tibor Szécsényi PhD, opponent
Beáta Gyuris PhD
Veronika Vincze PhD

Supervisor:
András Kornai DSc

Budapest, December 2023

1 BACKGROUND AND OBJECTIVES

The thesis belongs to the quite interdisciplinary field of computational lexical semantics, on the border of computer science, linguistics, mathematics, psychology, and philosophy. The two main groups of meaning representations are that of symbolic and distributional ones, which are presented in chapters 2 and 3 of the dissertation, respectively. The main examples of the former are the various semantic networks, whose nodes represent concepts and the edges between them represent both fixed relations (belonging to the meaning of the words) and relations in the text to be processed, based on directly formalizing the explicit knowledge of experts. Of these, I consider the older ones (2.2) important for their basic ideas, and the newer ones (2.4) as potential or competing computational linguistic resources. The theoretical framework of this dissertation is provided by the `4lang` meaning representation theory, which, in addition to the paradigms already mentioned, also builds on cognitive semantics (2.3).

Distributional meaning representations (shallow or deep neural networks, word embeddings and contextualized word representations) can be trained without linguistic knowledge on unannotated text, they are more useful in practice than symbolic ones, but they are difficult to interpret. An important predecessor is word representations based on matrix decomposition, where the elements of the matrices to be decomposed represent associations between words (4.1.1) or co-occurrences (4.1.2–4.1.6). I describe the history of neural language modeling in a big-picture but relatively complete way (4.2 and 4.3) from around 2003 (when connectionist dreams started to come true, i.e. that human implicit language competence can be efficiently modeled by artificial neural networks) to attention-based deep language models. Very deep pre-trained language models, such as GPT, are only mentioned in the dissertation.

The `4lang` semantic framework was developed in an inter-institutional (SZTAKI, BME, Institute of Linguistics) project led by András Kornai. The system, which has been outlined in several

books and articles and partially implemented, can be most simply thought of as a graph with concepts as nodes and connected by three kinds of edges: in the case of two asymmetric roles (e.g. subject and object, possessor and possession, figure and ground), arrows labeled 1 and 2 are used (according to the two arguments of the head), and for a single role (e.g. IS-A, single-participant event, property), a 0-labeled arrow is used. The concepts corresponding to the nodes are rather abstract: they are language-independent, they have no part-of-speech (e.g. *use* and *usage* are the same node), and word-disambiguation is only done in the case of clear homonymy. This is the monosemy principle.

The methodology of the members of the research team applies the principle of stepwise lexical decomposition (Dik 1978) to digital dictionaries (digitized traditional dictionaries or *Wiktionary*). Each word in the entire vocabulary is represented by its definition, then the words in the definitions are replaced by their definitions, and so on. When no more substitutions can be made (because each word definition would bring back a previously eliminated word), we have the defining vocabulary. It was mostly my task to manually define the elements of such a defining vocabulary, following the principles of `4lang` and cognitive semantics. The resulting definitions served as input for many of the results of this dissertation.

The main chapters of the thesis deal with the representation of relations between words in symbolic and distributional models. Chapters 5 and 6 examine the relations between argument-bearing words and their arguments and modifiers. Chapter 5 does this in `4lang`, and chapter 6 in a distributional model. Chapter 7 investigates how various word relations – lexical relations proper, familiar from semantic nets or e.g. from the Conceptual Structures of Jackendoff (1983), analogical relations, and relations between words in different languages – appear in static word embeddings. The final chapter proposes a cross-lingual approach to word sense induction: it examines whether a word that appears ambiguous in a machine learning model is really ambiguous.

Section 3.3 applies the classical web search method *PageRank* to measure how important the symbols used in `4lang` definitions (i.e. lexical relations, deep cases – which play a major role in chapter 5 – and simple concepts like *place*) are in the process as words define each other.

The methodology of Chapter 5, on the so called deep cases of `4lang`, is perhaps the closest to that of computational lexicography: the task was to manually create the definitions of a basic vocabulary along cognitive semantic considerations. This was mostly don by me.

Chapter 6 – on associations between verbs, subjects and objects – is still motivated by computational lexicography, but it builds an relatively serious bulk of data science and a probabilistic apparatus, in so far as it generalizes normalized positive pointwise mutual information (PPMI) to the multivariate case, and demonstrates its usefulness in a tensor decomposition experiment.

The two methods in the section on hypernym extraction (7.1), namely formal concept analysis (FCA) and sparse (over-complete) word representations, belong to two rather different domains (mathematical linguistics vs. general machine learning/representation learning, respectively), but can be considered as discrete-combinatoric vs. statistical (relaxed) pairs of each other.

Section 7.3 on causality considers a conjecture derived from a two-dimensional visualization of static word vectors, and analyzes it in the original space, following the paradigm of representing analogy pairs as parallelograms – also called vector offset – to conclude that the meaning vector of a cause (e.g. *pain*) is composed of the vector of the cause (*hurt*) and that of a common causal element. This is a very abstract finding about the latent space of words.

Section 7.4 uses the popular methods of the 2013–2019 period, the aforementioned vector offset and linear word translation, for Hungarian and other medium-resource European languages. It was

a methodological novelty to extend both vector offset and linear translation from the word2vec model to the GloVe model.

In Section 7.5, I combined triangulation, a common method of word translation retrieval in early machine learning, with linear translation, which was also used in the previous section of this dissertation and in the next chapter.

The last chapter measures the precision of multi-sense word embeddings in detecting the ambiguity of words using linear translation.

3 THESES

1. I have proposed and successfully applied a method taken from web search, *PageRank*, to measure how important each node in the semantic network is in the recursive process as words define each other. In Makrai (2013), I demonstrated the method on the `4lang` version of that time. In the dissertation, I repeated the analysis with the definitions of Makrai (2014b).

2. I added deep cases to the handwritten, formulaic definitions of the basic vocabulary of `4lang`, thereby developing a deep case set (Makrai 2014b).

3. To model subject-verb-object co-occurrences, I generalized several weighted versions of PPMI to the higher-order (>2) case. By modeling third-order interactions of English subject-verb-object triples using tensor decomposition, I showed (Makrai 2022) that

    a) these weighted higher-order PPMI variants perform better than log frequency, vanilla PPMI and log Dice the, which were used as baselines.

    b) treating empty objects uniformly with filled ones benefits word representation.

c) latent dimensions obtained with non-negative decomposition (either canonical polyadic or Tucker) are semantically meaningful.

d) The difference between the embedding of each noun as subject and object is intuitively related to agenthood.

4. Lexical relations

a) With my co-author(s)[1], we successfully extracted hypernyms with sparse word representations. We won three sub-tasks of *SemEval-2018 Task 9* (Berend, Makrai, and Földiák 2018).

b) With my co-authors[2], we validated a word embedding obtained from the definition graph on the basis which possible subrelations of antonymy appear in it.

c) I have shown that in the classical word embedding Senna, the lines connecting different cause-effect pairs (e.g. *hurt–pain*) run near a common "causal center", i.e. the meaning of each effect (*pain*) is composed of the meaning of the corresponding cause (*hurt*) and a constant causal element (Makrai 2014a).

d) I created a benchmark for analogy relations in Hungarian, a language with rich morphology, and used it to test Hungarian static word embeddings, showing that morphological relations appear similar to those in languages with better resources (Makrai 2015).

5. Linear translation

a) I extended the linear translation method to European languages with medium resources and to the GloVe model (Makrai 2015).

1 The ratio of contributions is Berend:Makrai = 2:1. Péter Földiák contributed the FCA idea.

2 I sorted the antonym pairs into sub-relations. The statistical tests are equal contribution with Dávid Nemeskey.

b) I hybridized the old method for word translation retrieval, triangulation a.k.a. the pivot method, with linear word translation: the elements of a triangulated translation word pair list were given the figure of merit from linear translation. I have published the largest German-Hungarian dictionary (word pair list) of the time (Makrai 2016).

6. With my co-authors[3] we have proposed a method to evaluate multi-sense word embeddings (MSEs) as detectors of homonymy (Borbély et al. 2016). The method is taken from the context of embedding-based dictionary induction. I have shown which of the methods of this paradigm are useful when the source language model is multi-sense. For the evaluation itself, I have proposed two metrics: one measuring the quality of translation in general, and the other the precision of ambiguous word recognition. By comparing the two SOTA MSE models, I have shown that the expected trade-off holds between the two metrics: the more specific the vectors are, the easier it is to translate, but if they are too specific, the translations may coincide (Makrai and Lipp 2018).

In addition to the theses listed above, still in the topic on the dissertation, I published some work on word ambiguity (Makrai 2007) and on spreading activation in the semantic network of `4lang` (Nemeskey et al. 2013). Kornai and Makrai (2013) and Kornai et al. (2015) introduced `4lang` in Hungarian and English respectively. More generally, the research team's understanding of hybrid data-driven computational linguistics was summarized in Ács et al. (2018). Makrai and Sass (2018) discussed the background of the power law distribution in a different context.

3 The cross-lingual part of the first article was an equal contribution with Gábor Borbély; András Kornai advised.

THE CANDIDATE'S PUBLICATIONS

Ács, J., G. Borbély, M. Makrai, D. Nemeskey, G. Recski, and A. Kornai. 2018. "Hibrid nyelvtechnológiák." *Magyar Tudomány* 6. (Cited on page 7).

Berend, Gábor, Márton Makrai, and Péter Földiák. 2018. "300-sparsans at SemEval-2018 Task 9: Hypernymy as interaction of sparse attributes." In *Proceedings of The 12th International Workshop on Semantic Evaluation,* 928–934. New Orleans, Louisiana: Association for Computational Linguistics, June. https://doi.org/10.18653/v1/S18-1152. (Cited on page 6).

Borbély, Gábor, Márton Makrai, Dávid Márk Nemeskey, and András Kornai. 2016. "Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation." In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP,* 83–89. Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2515. (Cited on page 7).

Kornai, András, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. 2015. "Competence in lexical semantics." In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics,* 165–175. Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.18653/v1/S15-1019. (Cited on page 7).

Kornai, András, and Márton Makrai. 2013. "A 4lang fogalmi szótár." In *IX. Magyar Számitógépes Nyelvészeti Konferencia,* edited by Attila Tanács and Veronika Vincze, 62–70. (Cited on page 7).

Makrai, Márton. 2007. *Többértelműségek magyar mondatok számítógépes elemzésében – a* meg *szó szófajának vizsgálata gyakoriságokkal.* https://hlt.bme.hu/en/publ/makrai07. Témalaborlolgozat. (Cited on page 7).

———. 2013. "Fogalmak fontossága a definíciós gráf vizsgálatával [Importance of concepts based on the analysis of the definition graph." In *VII. Alkalmazott Nyelvészeti Doktoranduszkonferencia,* edited by Tamás Váradi. MTA Nyelvtudományi Intézet Budapest. ISBN: 978-963-9074-59-0. http://www.nytud.hu/alknyelvdok13/proceedings13/ANyD7-Makrai-Marton.pdf. (Cited on page 5).

———. 2014a. "Causality in vectors space language models." In *Spring Wind,* 6:192–200. Association of Hungarian PhD / DLA Students (DOSZ). (Cited on page 6).

———. 2014b. "Deep cases in the `4lang` concept lexicon." In *X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014),* edited by Attila Tanács, Viktor Varga, and Veronika Vincze, 50–57 (in Hungarian), 387 (English abstract). ISBN: 978-963-306-246-3. (Cited on page 5).

———. 2015. "Comparison of distributed language models on medium-resourced languages." In *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015),* edited by Attila Tanács, Viktor Varga, and Veronika Vincze, 22–33. Szegedi Tudományegyetem Informatikai Tanszékcsoport. ISBN: 978-963-306-359-0. (Cited on page 6).

———. 2016. "Filtering Wiktionary triangles by linear mapping between distributed models." In *LREC.* (Cited on page 7).

———. 2022. "Three-order normalized PMI and other lessons in tensor analysis of verbal selectional preferences." In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia,* edited by Gábor Berend, Gábor Gosztolya, and Veronika Vincze, 105–120. Szegedi Tudományegyetem TTIK, Informatikai Intézet. ISBN: 978-963-306-848-9. (Cited on page 5).

Makrai, Márton, and Veronika Lipp. 2018. “Do multi-sense word embeddings learn more senses?” In *K + K = 120 Workshop Dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays.* (Cited on page 7).

Makrai, Márton, and Bálint Sass. 2018. “A szöveg mint skálafüggetlen hálózat.” In *XIV. Magyar Számítógépes Nyelvészeti Konferencia.* (Cited on page 7).

Nemeskey, Dávid, Gábor Recski, Márton Makrai, Attila Zséder, and András Kornai. 2013. “Spreading activation in language understanding.” In *Proceedings of the 9th International Conference on Computer Science and Information Technologies (CSIT 2013),* 140–143. Yerevan, Armenia: Springer. https://hlt.bme.hu/media/pdf/nemeskey_2013.pdf. (Cited on page 7).