Naïve discrimination learning approach to polysemy

Dušica Filipović Đurđević (dusica.djurdjevic@f.bg.ac.rs)
Department of Psychology, Faculty of Philosophy, University of Belgrade
Laboratory for Experimental Psychology, Faculty of Philosophy, University of Belgrade
Laboratory for Experimental Psychology, Faculty of Philosophy, University of Novi Sad

Abstract

In this paper, the focus of the investigation is the power of the computational model which is based on the principles of discrimination learning to predict processing latencies of polysemous nouns. Discrimination learning has been shown as a powerful principle that can account for multiple language phenomena, such as various frequency effects, various morphological phenomena and so on. However, the application of the model in predicting semantic effects is at its very roots. Here, the model is applied to simulate processing of words with multiple related senses, by mapping the bigrams from the input to the output that is represented by an array of words that co-occur with the target word. The simulated reaction time was positively correlated with empirically observed processing latencies. Additionally, the regression model fitted to simulated reaction time revealed a significant effect of lemma frequency, word familiarity, number of senses, and redundancy of sense probability distribution. These effects mirrored the pattern of the effects observed with empirical reaction time, the only exception being the reversal of the direction of the number of senses effect. Taken together, the results add to the body of research that links polysemy to context variability and additionally demonstrate how such a link can be related to learning.

Keywords: distributional semantics; naive discrimination learning; polysemy.

Introduction

Discrimination learning, as defined by Rescorla (1988) is the process through which an organism learns the structure of the environment by discriminating the stimuli that serve as cues which are good predictors of the certain outcome from the stimuli that do not contribute to such prediction. Learning is thus seen as the dynamic process of the competition of cues, and knowledge, including language, as the system of cue-outcome association that is constantly updated (Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010).

In order to further apply this approach to language processing Baayen, Milin, Filipović Đurđević, Hendrix, & Marelli (2011) have proposed Naïve discriminative reader (NDR), a simple computational model based on the equations of Rescorla and Wagner (1972). The model includes no hidden layers and no feedback activation. The only process that the model subsumes is the mapping of the language input (e.g. bigrams) to the output (e.g. word meaning). The model has proven fruitful in accounting for multiple language phenomena, such as word frequency effect (Baayen, 2010), N-gram frequency effect (Baayen, Hendrix, & Ramscar, 2013), morphological family size effect (Baayen, et al., 2011), inflected paradigm typicality effect (Baayen, et al., 2011; Filipović Đurđević & Gatarić, 2018; Filipović Đurđević & Milin, 2018) and so on. However, in the previous demonstrations, the outcomes were treated as the pointers to the locations in multidimensional semantic space. Although semantics was not conceptually seen as a localized phenomenon, technically it was implemented in such a way.

In this paper, we aim to broaden the implementation of the model outcomes by applying the distributional semantics approach, i.e. by using co-occurrence vectors to specify the meaning of the word at the outcome level in the model. The distributional semantics models have been flourishing during the course of several decades to show that multiple phenomena related to processing of word meaning can be accounted for based on co-occurrence of words (Landauer and Dumais, 1997; Lund and Burgess, 1996; but lately also Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Chen, Corrado, & Dean, 2013). The meaning of a word is represented as the vector or a simple array of frequencies of the target word co-occurring with each of the context words (selected in advance). It has been demonstrated that such vectors do capture certain aspects of meaning, the most famous demonstration being the one of predicting human TOEFL synonymy choices based on the cosine distance between the vectors representing the words in question (Landauer & Dumais, 1997).

The focus of the current research is on the polysemous words, i.e. words that refer to multiple related senses (e.g. *paper* as scientific paper, or *paper* as material). Processing advantage of polysemous words compared to unambiguous words was observed in numerous studies (Beretta, Fiorentino, & Poeppel, 2005; Klepousnitou, 2012; Pylkkänen, Llinas, & Murphy, 2006; Rodd, Gaskell, & Marslen-Wilson, 2002; for a comprehensive review see Eddington & Tokowicz, 2015). Additionally, it was shown that an increase in a number of senses was followed by a decrease in processing latencies. Research in Serbian revealed that, in addition to a number of senses, the balance of sense probabilities also affected processing: the more balanced relative frequencies of individual senses were, the faster the processing was (Filipović Đurđević, 2007). The balance of sense probabilities was operationalized as information theory measure of redundancy (Cover & Thomas, 1991; Shannon, 1948), higher redundancy being related to more balanced sense probabilities. Evidence from several lines of research suggested that the observed processing advantage of polysemous words was related to variability in contexts in which polysemous words appeared (Adelman, Brown, Quesada, 2006; Filipović Đurđević & Kostić, 2009).

In this research, this hypothesis was tested by applying the discrimination learning approach (Baayen et al., 2011) to simulate the effects of polysemy in the Serbian language.

Method

Polysemous words were selected from Filipović Đurđević and Kostić (2017) database of polysemous nouns. The selection was based on the frequency of occurrence in Ebart media database (http://www.arhiv.rs), which was used for building semantic vectors. Only the words that occurred more than 500 times in the database were selected, which led to the selection of 130 polysemous words. For these words, processing latencies were taken from visual lexical decision task experiment of Filipović Đurđević (2007). Data on (log) lemma frequency were taken from Kostić (1999), whereas data on word length in letters, word familiarity, number of senses, and redundancy were taken from Filipović Đurđević and Kostić (2017).

The simulation

The simulation which is based on naïve discriminative learning was conducted in **R** statistical software (R Core Team, 2017), by using **ndl** package (Arppe et al., 2015), following the procedure of Baayen et al. (2011).

Specification of model input

The input consisted of bigrams that constituted each word form in the nominative singular, as presented in the experiment (e.g. for word *linija* the input bigrams would be: *#l, li, in, ni, ij, ja, a#*).

Specification of model output

The output consisted of the lemma associated with the given word form and of the semantic vectors of that word form. The vectors were built separately for individual occurrences of all inflected forms of the target words by moving a 7-point window through the raw text of Ebart media database (http://www.arhiv.rs) of approximately 65 million words (Filipović Đurđević & Kostić, 2009; Schütze, 1998). These vectors were simply arrays of context words that co-occurred with the target polysemous word (three positions to the left, or three positions to the right).

Following the simplest approach described in Baayen et al., (2011), the simulated reaction time was estimated as the inverse summed activation for the word form which was presented in the experiment in which the empirical reaction time was collected (Filipović Đurđević, 2007).

Results

The simulated reaction time was positively correlated with reaction time observed in Filipović Đurđević (2007): r = .334, t (128) = 3.880, p = .0002 (95% CI: r = .17 – r = .48; Figure 1).

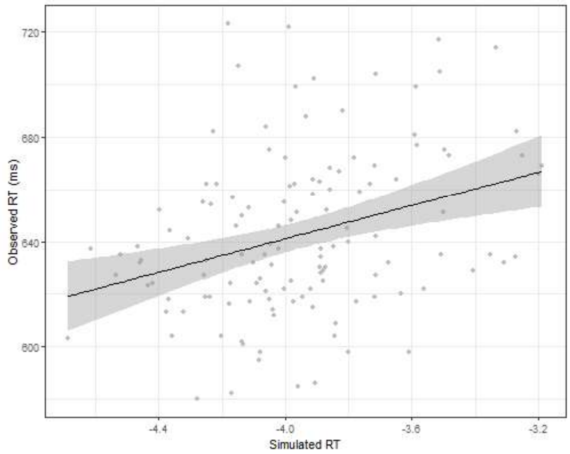

Figure 1: The relation between the simulated and the observed processing latencies.

Additionally, linear regression model fitted to simulated reaction time revealed a significant effect of lemma frequency, word familiarity, number of senses, and redundancy of sense probability distribution (Table 1). All of the observed effects were comparable to those in the model fitted to behavioural data. Importantly, an increase in redundancy was related to an increase in both behavioural RT and simulated RT. However, although there was a significant effect of the number of senses on simulated RT, the direction of this effect was reversed as compared to that observed with fitting behavioural RT.

Table 1: Coefficients from linear regression fitted to simulated reaction times for 130 polysemous nouns.

	Estimate	SE	t	Pr(>|t|)
Intercept	-4.19	.08	-50.96	<.001
Word length in letters	-.00	.03	-.05	.960
(log) Lemma frequency	-.09	.03	-3.05	.003
Word familiarity	-.06	.03	-2.05	.042
Number of senses	.06	.03	2.16	.033
Redundancy	1.49	.49	3.04	.003

Discussion

The discrimination learning approach to language processing has been successfully combined with distributional semantics to account for processing of Serbian polysemous nouns. The processing latencies which were simulated in this way were related to standard psycholinguistic variables, as well as to descriptions of polysemy: number of senses and balance of sense probabilities (redundancy). However, while the latter affected simulated latencies in the expected direction, the direction of the former was reversed, thus opening the space for future investigation. In spite of this, it can be concluded that the pioneering attempt to bring together the two approaches seems to be the fruitful ground to further understand the processing of lexical semantics in the light of simple learning principles.

Acknowledgements

This research is supported by the Ministry of Education, Science and Technological Development, Republic of Serbia (179033, 179006).

References

Adelman, J., Brown, G., Quesada, J. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17* (9), 814.

Arppe, A., Hendrix, P., Milin, P., Baayen, R.H., Sering, T., & Shaoul, C. (2015). *ndl: Naive Discriminative Learning. R package version 0.2.17*. http://CRAN.R-project.org/package=ndl

Azuma, T. (1996). Familiarity and relatedness of word meanings: Ratings for 110 homographs. *Behavior Research Methods, Instruments, and Computers*, 28(1), 109–124.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*, 436-461.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, 56, 329–347.

Baayen, R. H., Milin, P., Filipovic Đurđević, D., Hendrix, P. & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review 118*, 438-482.

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: an MEG study. *Cognitive Brain Research, 24*(1), 57–65.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley & Sons. https://doi.org/10.1002/0471200611

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: the current state of the literature. *Psychonomic Bulletin & Review, 22*(1), 13–37.

Filipović Đurđević, D. (2007). *Polysemy effect in processing of Serbian nouns*. PhD thesis, University of Belgrade.

Filipović Đurđević, D. & Gatarić, I. (2018). Simultaneous effects of inflectional paradigms and classes in processing of Serbian verbs. *Psihologija, 51*(3), 259–288. https://doi.org/10.2298/PSI170811015F

Filipović Đurđević, D., Đurđević, Đ., & Kostić, A. (2009). Vector based semantic analysis reveals absence of competition among related senses. *Psihologija, 42*(1), 95-106.

Filipović Đurđević, D., & Kostić, A. (2017). Number, Relative Frequency, Entropy, Redundancy, Familiarity, and Concreteness of Word Senses: Ratings for 150 Serbian Polysemous Nouns. In S. Halupka-Rešetar and S. Martínez-Ferreiro (Eds.) *Studies in Language and Mind 2* (pp.13-77). RS, Novi Sad: Filozofski fakultet u Novom Sadu. http://digitalna.ff.uns.ac.rs/sadrzaj/2017/978-86-6065-446-7

Filipović Đurđević, D., & Milin, P., (2018). Information and Learning in Processing Adjective Inflection, *Cortex*, https://doi.org/10.1016/j.cortex.2018.07.020

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language, 123*(1), 11–21.

Kostić Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika.* Beograd: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203–208.

Medijska dokumentacija Ebart. <http://www.arhiv.rs/>

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781* [cs]. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Pylkkänen, L., Llinas, R. & Murphy, G. (2006). Representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience. 18*(1), 1–13.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science, 34*, 909 –957.

Rescorla, R. A. (1988). Pavlovian conditioning: It's Not What You Think It Is. *The American psychologist. 43*. 151-60.

Rescorla, R.A. & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical Conditioning II, Appleton-Century-Crofts*, pp. 64–99.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Languag*e, *4*6, 245–266.

Shannon, C. E. (1948). *A mathematical theory of communication*. Bell System Technical Journal, XXVII: 379-423.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics, 24* (1), 97–123