Word Probability Re-Estimation Using Topic Modeling and Lexical Decision Data

Hronský, Ratislav; Keuleers, Emmanuel

Published in:
Proceedings of the Annual Meeting of the Cognitive Science Society

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Word Probability Re-Estimation Using Topic Modeling and Lexical Decision Data

Rastislav Hronsky (r.hronsky@tilburguniversity.edu)
Jheronimus Academy of Data Science
Sint Janssingel 92, 5211 DA 's-Hertogenbosch

Emmanuel Keuleers (e.a.keuleers@tilburguniversity.edu)
Department of Cognitive Science and Artificial Intelligence, Tilburg University
Warandelaan 2, 5037 AB Tilburg

Abstract

Two assumptions of psycholinguistic research are that text corpora can be used as a proxy of the language that people have been exposed to and that the reaction time with which people recognize words decreases with the probability (or frequency) of the words in a corpus. We propose a method that produces topic-specific word probabilities from a text corpus using latent Dirichlet allocation, then combines them to fit lexical decision reaction times and re-estimates word probabilities. We evaluated how well independent lexical decision reaction times could be predicted from re-estimated word probabilities compared to original probabilities, using independent lexical decision data. In an experiment designed to prove the concept, the re-estimated word frequency model explained up to 9.6% of additional variability in reaction times on group level and up to 2.9% on level of individual participants.

Keywords: psycholinguistics; lexical decision; language model; topic modeling; corpus; personalization

Introduction

Word probabilities are used in many areas, spanning from psycholinguistics to computational language modeling and artificial intelligence. While the main interest of a psychologist is to find models explaining the observed behavior well, the interest of an artificial intelligence researcher is to design machines that behave in a human-like manner. Both might benefit from the ability to estimate the word distribution of a personal language environment.

Getting a precise idea about what language a person is exposed to is difficult. Typically, researchers use very large text corpora as a proxy of a person's language environment and, although they take into account that some corpora better reflect the language environment than others (New, Brysbaert, Veronis, & Pallier, 2007; Brysbaert & New, 2009), this is still a one-size-fits-all approach. This is cause for concern, as it is not controversial that there are profound individual differences in familiarity with words. For instance, Mandera, Keuleers, and Brysbaert (2019) have shown that word prevalence (the proportion of a population that knows a word) can be markedly influenced by factors such as age, gender, location. For instance, in English, words such as *howitzer*, *thermistor*, *azimuth* are significantly more prevalent among men and words *peplum*, *tulle*, *chignon* among women, on average (Mandera et al., 2019).

In a recent effort to make corpora more reflective of the language environment, Johns, Jones, and Mewhort (2019) developed the *experiential optimization* method, which allows for the creation of customized corpora based on performance on a target task, such as the TOEFL synonym test (Landauer & Dumais, 1997), ratings of semantic similarity (Recchia & Jones, 2009), or lexical decision reaction time (Balota, Yap, Hutchison, et al., 2007). The core principle of experiential optimization is to assemble a customized corpus by using a hill-climbing algorithm that iteratively adds the best fitting section to the final corpus until the performance on the target task stops improving.

In the implementation of Johns et al. (2019), experiential optimization requires a corpus to be composed of sections that are labelled, for instance according to genre and author. This makes it difficult to apply the technique to corpora which do not have any such labeling. The method proposed in the current paper alleviates this concern by using topic modeling as a form of unsupervised labeling, thereby allowing the use of an arbitrary corpus. Instead of iteratively selecting the best fitting section of a corpus, like in experiential optimization, our method fits the topic model to the target task using least squares optimization. The fitted topic model can then be used to re-estimate word probabilities of all the words in the original corpus. The technique can be applied for data from particular groups or from individuals.

In this paper, we provide a proof-of-concept of the technique using lexical decision as the target task. In a lexical decision task (LDT), a person is shown a string of characters and asked, whether or not it is an actual word (Meyer & Schvaneveldt, 1971). In recent years, large amounts of lexical decision data have been collected for English (Balota, Yap, Cortese, et al., 2007; Keuleers, Lacey, Rastle, & Brysbaert, 2012) and other languages (Dutch Lexicon Project (Keuleers, Diependaele, & Brysbaert, 2010), French Lexicon project (Ferrand et al., 2010), developmental lexicon project (Schröter & Schröder, 2017), Chinese Lexicon project (Sze, Liow, & Yap, 2014)). As pointed out by Johns et al. (2019), most of the explained variance in lexical decision latency is accounted for by measures such as word frequency and contextual diversity (Adelman, Brown, & Quesada, 2006), which can be easily derived from text corpora (see Brysbaert, Mandera, and Keuleers (2018) for an overview).

Our solution is based on two assumptions:

1. Participants' reaction times in a lexical decision task partly reflect the word probabilities in their language environment.

2. A corpus consisting of a large set of semantically diverse

documents allows for identification of a smaller set of semantically coherent components for each of which there is a distinct set of word probabilities, for instance using topic modeling.

Using the two assumptions, the proposed method uses topic modeling to find a topic space and estimates probabilities for every word in the topic vocabulary given lexical decision reaction times. Since lexical decisions are made for single words, the method is constrained to unigram representations.

By conducting two experiments, we addressed the following research questions:

1. How much additional variability in lexical decision reaction times can be explained by re-estimating word probabilities for groups and individuals?
2. How does limiting the amount of words and corresponding reaction times shown to the model affect the performance?

We recorded up to 9.59% of additionally explained variability in aggregate reaction times by the re-estimated word probabilities at 80 topics, which was the highest value we experimented with. When fitted on individual level reaction times, on average, the re-estimated word probabilities additionally explained up to 2.9% of variability compared to the original corpus probabilities. As the proportion of words shown to the model decreased, so did the R^2 on reaction times. This effect was moderated by the number of topics.

Method

For the sake of the present study, we formally define the following terms:

- A *vocabulary* is a set of words denoted by $V = \{w_1, w_2, ..., w_N\}$. In the context of the present work, a *word* is the elementary discrete unit of a unigram language model.
- *Topics* represent the hidden semantic structure of a corpus denoted by $Z = \{z_1, z_2, ..., z_M\}$ with M being the number of topics. A topic z_m predicates a word distribution over a vocabulary V_{TM} represented by a vector β_m with elements equivalent to $\beta_m[i] = p(w_i|z_m)$. The columns vectors β_m comprise a matrix β of shape $|V_{TM}|, M$.
- *Reaction times* of an individual j to words V_{RT} are represented by a vector R_j containing elements indexed by $i \in \{1, 2, ..., |V_{RT}|\}$. The value of $R_j[i]$ is equivalent to $f(RT)$ (see Formula 2). In case the trial was inaccurately decided (participant did not recognize the word) the value is undefined.
- Let $g = \{1, 2, ..., J\}$ be a group of J participants and $Y_g = \left[R_1, R_2, ..., R_J\right]$ be a matrix with column vectors being the participants' reaction times R_j. *Aggregate reaction times* of the group g are equivalent to a vector $\bar{Y_g}$ with elements equal to means of the corresponding rows of the matrix Y_g while ignoring the undefined elements.

Model

The proposed technique consists of two main steps:

1. Estimation of the matrix β (only necessary once).
2. Finding a linear combination of topics x that best fits the reaction times (per participant/group).

The first step corresponds to obtaining a transformation from topic space to word distribution, $\mathbb{R}^M \rightarrow \mathbb{R}^{|V_{TM}|}$. For this purpose, we adopted a well established generative topic modeling method, the *latent Dirichlet allocation* (LDA; Blei, Ng, and Jordan (2003)). The vanilla LDA models a corpus as a collection of documents, each of which is assumed to be generated by a unique combination of hidden variables (topics). The topics are identified based on words that tend to occur in similar context which makes the estimation of the matrix β fully unsupervised. The number of topics M is a hyper-parameter.

The second step corresponds to finding the most likely linear combination of topics given the reaction times. We operationalized it by solving the following linear least squares optimization problem:

$$\arg\min_x \|Ax - y\|_2 \quad (1)$$

The matrix A is a rescaled representation of topics β defined as $A = \left[g(\beta_1), ..., g(\beta_M)\right]$ where $g(\beta_m) = log_{10}(\beta_m / min(\beta_m))$. Rescaling is needed because of the Zipfian distribution of word probabilities; the logarithm of the probability is a more suitable unit for a linear model. Depending on whether the topic mixture should model aggregate or individual level reaction times, the dependent variable y is selected to be $\bar{Y_g}$, or R_j.

The final vector of *re-estimated word probabilities* is obtained by (1) multiplying the matrix A by the vector x, the best fitting topic combination, and (2) transforming it with exponentiation and normalization such that the entries sum to 1.

The vocabulary of the vector y containing the lexical decision information, V_{RT}, may only be a subset of the vocabulary of the topic model (and hence the matrix β), V_{TM}. While the final re-estimation is done for all words in V_{TM}, the least squares optimization is performed on the subset of rows of β that correspond to words in $V_{RT} \cap V_{TM}$ (see Figure 1 for an illustration) [1].

We used the following implementations for the aforementioned techniques: `gensim.models.LdaMulticore` for LDA (Řehůřek & Sojka, 2010), and for least squares `sklearn.linear_model.LinearRegression` (Pedregosa et al., 2011).

Evaluation

In psycholinguistics, the most common method to quantify the fit of a corpus to lexical decision data is to compute the

[1] This is useful because lexical decision data are typically not available for every word in a corpus.

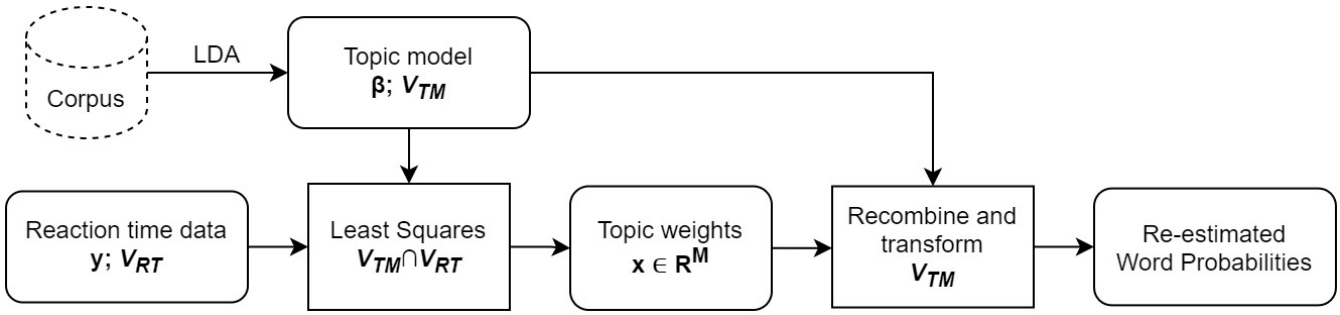

Figure 1: An overview of the method. Where appropriate, we denote the corresponding data structure and vocabulary.

R^2, proportion of explained variance, between the aggregated, standardized reaction times and log-scaled word frequencies. Excluding the inaccurately decided trials from the computation is a common practice. We adopted the practice in the present study, while working with log-scaled probabilities instead of frequencies.

The baseline model performance is the R^2 achieved by the log-scaled word probabilities derived from the corpus processed as described in subsection Corpus.

The predictive capabilities of the model can be assessed by working with two disjoint vocabularies, V_{train}, V_{test}, and only using the reaction times for words in V_{train} to find the best topic combination and evaluate the performance on reaction times for words in V_{test}.

Word Recognition Data

The British Lexicon Project (BLP; Keuleers et al. (2012)) contains lexical decisions for 78 individuals, who were students and employees of Royal Holloway, University of London. The participants were divided in one group of 38 and one group of 40, with each group assigned to a different set of word stimuli. Each participant responded to all of the word stimuli for their group, resulting in reaction times for 14,365 words per participant. On average, participant's accuracy level was more than 80%, as a sufficient level of accuracy was one of the requirements for not being excluded from the BLP.

The raw reaction times, measured in milliseconds, were standardized assuming a log-normal distribution. Because the log-normal distribution requires an absolute zero point, we shifted the reaction times toward zero. Additionally, we reflected the variable about the y-axis by multiplying it by -1, such that higher values could be interpreted as faster reactions (see Figure 2). In order to eliminate variance caused by individual differences and situational context, the procedure was performed per participant and block (participants responded in blocks of 500 trials). The final transformation f, depicted in Figure 2, can be summarized as follows:

$$f(RT) = -zscore(log(RT - min(RT))) \quad (2)$$

Corpus

The English section of the OpenSubtitles [2] dataset (Lison & Tiedemann, 2016), release 2018, is an extensive collection of translated movie subtitles. We decided to use this corpus throughout the experiments for topic modeling purposes for the following reasons: it is open and accessible which makes it a suitable resource for reproducible scientific experiments, and subtitle-based corpora are simply a suitable linguistic resource for explaining variance in lexical decision performance (Brysbaert & New, 2009).

Prior to training models and running experiments, the dataset was processed to better fit the needs of the approach. First, we excluded the movies released in 2018 (only 5 movies) and before 2000, in order to limit the text to relatively recent language as well as make it computationally easier to process. From the resulting set of 97,388 movies we picked the first subtitle file per movie, since the original dataset contained multiple alternative translations for some movies. Next, all of the utterances were lemmatized with *spaCy* (Honnibal, Montani, Van Landeghem, & Boyd, 2020) and filtered to only include words from the BLP. The resulting corpus contained 373 million tokens and 25,626 distinct words that constitute the final topic model vocabulary V_{TM}. The discrepancy between this vocabulary size and number of words in the BLP (28,730) was mainly due to the fact that the lemmatizer returned nouns in singular form and verbs in present tense; therefore, some stimuli from the BLP in past tense or plural form were not matched by lemmas in the corpus.

Experiments

We performed several word probability re-estimation experiments for individual level data, group level data, situations with limited training sets, and different topic models for three values of the hyper-parameter value $M \in \{5, 20, 80\}$.

Experiment 1: Full vocabulary

The goal of the initial experiment was to establish the performance on lexical decision data using the full set of available words. This purpose was conveniently served by the two

[2] http://www.opensubtitles.org/

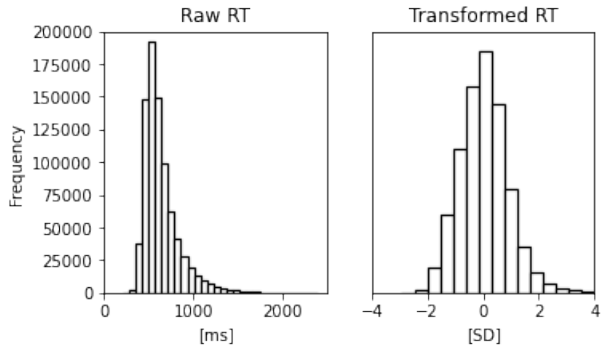

Figure 2: Histogram of all reaction times from the BLP before (left) and after (right) applying the transformation f.

Table 1: Results of the first experiment in terms of differences in R^2 between fitted models and baseline models (BL R^2) across different amounts of topics used. The upper part shows results for aggregate reaction times of all the combinations of training and test data GR1 and GR2. The average increase in R^2 per individual when fitted on individual and aggregate data is listed in the bottom part.

		BL R^2	$M = 5$	$M = 20$	$M = 80$
Groups					
Train	Test				
GR1	GR1	.2688	+.0131	+.0732	+.0959
GR2	GR1	.2688	+.0130	+.0718	+.0876
GR1	GR2	.2521	+.0146	+.0720	+.0788
GR2	GR2	.2521	+.0146	+.0733	+.0863
Individuals					
Basis					
Agg.	Mean	.0657	+.0033	+.0163	+.0197
	SD	.0271	.0018	.0061	.0072
Ind.	Mean	.0657	+.0038	+.0185	+.0273
	SD	.0271	.0019	.0062	.0069

participant groups originating from the design of the BLP which provide aggregate data for two disjoint vocabularies. The two groups, GR1 and GR2, were sized 38 and 40 participants. We estimated the topic coefficients for both groups and then tested for every combination of them. Next, we applied the method to every individual and computed two values of R^2, one based on aggregate data of the group that the participant belongs to, and one based on the participant's single responses. For both we report the average R^2 per participant and the standard deviation.

Results The results of the inter-group evaluation are listed in the upper half of the Table 1. The difference in baseline performance between the two test sets was about 1.6% in terms of R^2; the corresponding difference to the rescaled corpora was consistently positive with an increasing value for higher number of topics used. The steps from $M = 5$ to $M = 20$ and then further to $M = 80$ respectively accounted for $\sim 5.9\%$ and $\sim 1.5\%$ of additional variance explained, on average. When evaluating on the same set of data as used for training, there was a marginal increase in R^2 compared to using the complementary set. This effect was symmetrical across the two groups and got progressively stronger with higher values of M, whilst achieving the largest difference of $\sim 0.8\%$ for the test setup on the first group and $M = 80$.

The bottom part of the Table 1 lists the results for average proportion of variance explained per participant. We can observe patterns similar to the previous scenario: the difference in R^2 was consistently positive for both models, it progressed with higher values of M, and the first increase of M improved R^2 more significantly than the second one (on average, 1.4% and 0.61% respectively). Additionally, the rate at which the performance rose with higher values of M was higher for the model trained on the individual's reaction times as opposed to aggregate group reaction times.

Discussion The fact that the re-estimated word probabilities consistently explained more variance than the original corpus, even when the training group was different than testing group, indicates the validity of the method. The fact that the performance difference between probabilities estimated on the testing data versus training data was more pronounced at higher values of M suggests that increasing the degree of granularity in the topics leads to a more fine-grained re-estimation. This effect was especially evident when both trained and evaluated on individual level data where, unlike in other scenarios, the performance did not appear to start converging even at the highest value of M.

Experiment 2: Limited Vocabulary

The second experiment was designed to assess the prediction robustness of the technique by progressively decreasing the percentage of words used for training, thus making the least squares problem less overspecified. The complementary set of unseen words was used for testing. The experiment was adapted for both individual and group level reaction times. In case of individuals, we computed the mean performance per participant and the standard deviation. We report results on the first group only, since there was no significant difference between the groups. The training set proportions were 50%, 25%, and 12.5%; the values for number of topics remained unchanged. In order to eliminate effects of chance under various train-test splits, we ran every experiment setting multiple times with a differently shuffled vocabulary and estimated the average R^2 (100 times for groups and 10 times for individuals). Lastly, we computed the upper bound performance (UB) achieved by training and testing on the same, full vocabulary.

Results The results for aggregate reaction times are shown in the upper part of the Table 2. Whilst the baseline performance was stable across different test set sizes, the average increase in R^2 progressed with more topics used in the model. There was an exception to this in the case of train set propor-

Table 2: The results of the experiment 2 reported as differences to the baseline performance BL R^2 for the various re-estimated probability distributions. The top and bottom part show data for group and individual level prediction respectively. The UB rows represent the upper bound. The column N lists the average number of words used in the training vocabulary.

Ratio	N	BL R^2	$M = 5$	$M = 20$	$M = 80$
Group 1					
UB	12,807	.2688	+.0131	+.0732	+.0959
50%	6,359	.2688	+.0126	+.0704	+.0845
25%	3,180	.2685	+.0121	+.0684	+.0766
12.5%	1,589	.2690	+.0111	+.0640	+.0610
Individuals of group 1					
UB	Mean	.0691	+.0039	+.0194	+.0292
	SD	.0276	.0022	.0065	.0072
50%	Mean	.0691	+.0027	+.0139	+.0088
	SD	.0276	.0029	.0075	.0084
25%	Mean	.0694	+.0021	+.0107	-.0010
	SD	.0274	.0027	.0072	.0077
12.5%	Mean	.0693	+.0005	+.0052	-.0149
	SD	.0273	.0033	.0070	.0077

tion of 12.5% where at $M = 80$ the increase was .3% lower than at $M = 20$. Similar to the first experiment, the step in number of topics from 5 to 20 resulted in larger increase in R^2 ($\sim 5.6\%$ on average) than from 20 to 80 ($\sim .7\%$ on average). The increase was the closest to the upper bound at train set proportion of 50% and it decreased toward the lowest proportion value.

The results for average fit per participant are listed in the bottom half of Table 2. In this case, we can observe a slightly different pattern for increasing values of M; in terms of M, the performance increased from 5 to 20 by $\sim .8\%$ on average, but from 20 to 80 it decreased by $\sim 1.2\%$ for all train set proportions, on average. Furthermore, at $M = 80$ and proportions 25% and 12.5%, there was a decrease in performance w.r.t. the baseline; the best performing models were achieved at $M = 20$.

Discussion The second experiment demonstrated that the word probabilities can be re-estimated well even based on a very limited set of words in the training set, especially in case of aggregate reaction time data. In this setup it also seems to be important to choose M cautiously, because increasing the value only helped up to a certain point. When using individual level reaction times, modeling by 80 topics was counterproductive in all situations but mostly with smaller train set sizes. This effect aligns with the common issue of overfitting where high number of model parameters combined with small number of training examples hinders the capability to generalize beyond training data and regularization needs to take place.

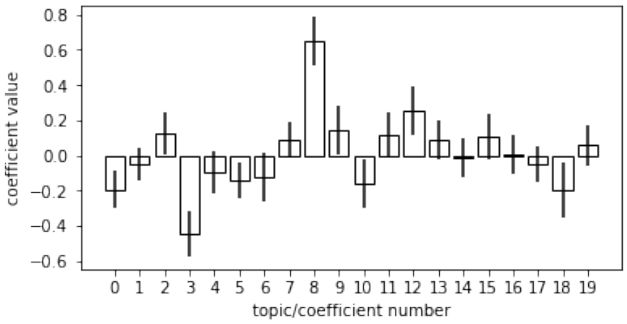

Figure 3: Distribution of topics coefficients at $M = 20$ aggregated over all 78 participants. The bars correspond to the average coefficient value; the whiskers represent the standard deviation.

Further Analysis

To shed more light on the nature of the results, we inspected the resulting topic distributions and re-estimated word probabilities. Firstly, we noticed that when the topic models were fitted to lexical decision data of individual participants, the average topic coefficients for all participants were significantly non-zero, as shown in Figure 3. This finding suggests that the discrepancy between the original corpus as a reference to lexical exposure of the participants is more pronounced than differences between individual participants. Additionally, we noticed that the mean of all coefficients was consistently positive and also consistently included some negative components. This implies that the technique allows for the identification of topics that are particularly bad reflections of an individual or a group's language environment. In contrast to experiential optimization, which only allows for a customized corpus to be built by adding words, the current technique can be thought of as also allowing subtraction of material.

In Figure 4 we can see that the re-estimated word probabilities (left) are less scattered at the lower end, suggesting better prediction for words which tend to be reacted to quickly despite their low frequency. There is also a larger concentration of points at values of roughly -4.6, because there is a subset of words which are not clearly associated to any topic; therefore, no combination of topics can change their position on the x-axis of the scatter plot. These words, which have a very low frequency, probably lack the structural variety to be topic-specific, and therefore cannot be re-estimated.

General Discussion

The word probability re-estimation technique results in a substantial increase in explained variance for aggregated lexical decision data, relative to baseline corpus word frequencies. Fitting a topic combination using the lexical decision data from one group of participants in the BLP resulted in re-estimated word probabilities that explained nearly 9% more variance in lexical decision latencies for the other

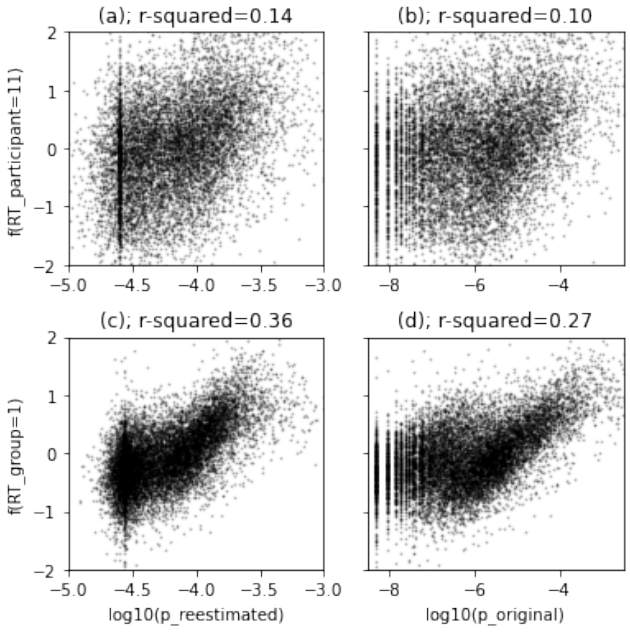

Figure 4: Scatter plots depicting the relationship between word probabilities (log-scale) and reaction times. The top panels show reaction times for a single participant, while the bottom panels shows aggregate reaction times. The left part corresponds to the re-estimated log-probabilities and the right part correspond to the original log-probabilities. The number of topics was $M = 80$.

group of participants, compared to the original word probabilities. Hence, this technique may be of particular value to researchers in psycholinguistics, where many experiments require frequency effects to be tightly controlled.

Individual lexical decision data are far more noisy than the aggregate data. Based on an experiment in which participants were presented with 500 identical trials on multiple occasions, Diependaele, Brysbaert, and Neri (2012) point out that the same lexical decision was made on only around 83% of the trials and, in terms of reaction times, the signal from repeated trials only explained about 8% of the variability of the initial signal. This upper bound on the amount of variance that can be explained in individual lexical decision data is reflected in our results, where explained variance for individual-level data was in the range of $6\% - 10\%$. Still, our results showed that, on average, word probability re-estimation increased explained variance in individual lexical decision times by nearly 3%, approaching the upper bound suggested by Diependaele et al. (2012).

One limitation of the present study is the lack of an extensive topic model evaluation. In the design of the experiments conducted in this study, the effect of mere dimensionality of the least squares optimization is confounded with the effect of the actual topic model quality resulting from the choice of M. This way, the results may create an impression that simply increasing the value of M leads to better results; however, this is meaningful only up to a certain point. The semantic diversity of corpora is affected by factors such as text source and size and should be rigorously estimated. The recommendations for finding the right value of M include measuring the topic model perplexity on a held-out test set, or various types of topic coherence (Röder, Both, & Hinneburg, 2015). In future studies, evaluating the topic model quality and the hyper-parameter setting needs more attention. Additionally, when working with larger values of M and/or limited sets of reaction times, it might be necessary to regularize the least squares problem because of overfitting.

In contrast to Johns et al. (2019), one limitation of the word probability re-estimation method is that it cannot be used for tasks that require local context for words, such as building lexical representations, an area where experiential optimization is extensively presented. Our technique does not preserve the structure of the corpus or the order of words. Therefore, it is restricted to unigram, bag-of-words language models. In future work, the presented technique may be extended to higher-fidelity language modeling by employing a topic modeling method capable of retaining the sequential structure of text.

Lastly, it is important to acknowledge that, while the present study uses lexical decision data for both fitting and validation, it would be interesting to see how well do re-estimated word probabilities transfer to language tasks beyond the training task, such as reading text measured by eye tracking.

Conclusion

Motivated by methodological innovation in computational psycholinguistics as well as potential application in artificial intelligence systems involving language, we demonstrated that behavioral data, such as reaction times from lexical decision experiments, can be used to re-estimate word probabilities so that they provide a better proxy to language environment, both at the group level and at the individual level. The present work improves on experiential optimization by being unsupervised, but, unlike experiential optimization, can only be applied to tasks requiring isolated word probabilities.

Acknowledgements

This research was funded by ITK as part of the SmartDATA (10028312) project as part of the KPN Responsible AI lab.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological science*, *17*(9), 814–823.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ... Treiman, R. (2007). *The english lexicon project* (Vol. 39) (No. 3). Springer New York LLC.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... Treiman, R. (2007). The english

lexicon project. *Behavior research methods*, *39*(3), 445–459.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. , *3*, 993–1022.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018, feb). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45–50.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Diependaele, K., Brysbaert, M., & Neri, P. (2012). How noisy is lexical decision? *Frontiers in Psychology*, *3*(SEP).

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... Pallier, C. (2010). The french lexicon project: Lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior research methods*, *42*(2), 488–496.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python.* Zenodo.

Johns, B. T., Jones, M. N., & Mewhort, D. J. (2019, feb). Using experiential optimization to build lexical representations. *Psychonomic Bulletin and Review*, *26*(1), 103–126.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 dutch mono-and disyllabic words and nonwords. *Frontiers in psychology*, *1*, 174.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012, mar). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304.

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.

Lison, P., & Tiedemann, J. (2016, may). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In N. C. C. Chair) et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016).* Paris, France: European Language Resources Association (ELRA).

Mandera, P., Keuleers, E., & Brysbaert, M. (2019). Recognition times for 62 thousand english words: Data from the english crowdsourcing project. *Behavior research methods*, 1–20.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, *90*(2), 227.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677. doi: 10.1017/S014271640707035X

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, *12*, 2825–2830.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, *41*(3), 647–656.

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth acm international conference on web search and data mining* (pp. 399–408).

Schröter, P., & Schröder, S. (2017). The developmental lexicon project: A behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods*, *49*(6), 2183–2203.

Sze, W. P., Liow, S. J. R., & Yap, M. J. (2014). The chinese lexicon project: A repository of lexical decision behavioral responses for 2,500 chinese characters. *Behavior Research Methods*, *46*(1), 263–273.