Exploring the Utility of Distributed Representations in Unsupervised Learning of Morphology from Child-Directed Speech
Term Paper : Ling692C

Sree Harsha Ramesh
Department of Computer Science
University of Massachusetts Amherst
shramesh@cs.umass.edu

Abstract

The question of how a child acquires language from a mere exposure to natural language speech is a fascinating one that has received a great deal of attention over the years. In this project, I have tried to particularly address the aspect of explaining the production of English inflectional morphology by building a computational model that can learn dense vector representations of words from a corpus of child directed speech and give rise to interesting regularities between different morphologically related forms of words. We leverage this property to generate morphological representations for a few inflectional categories, and test their effectiveness in generation of inflected forms from base forms, and to also identify the inflectional category in a given inflected form. Remarkably, on comparing the morphological transformations learned by the vector space model vis-à-vis Linguistica (Lee and Goldsmith, 2016), we found that there are inflectional categories such as the past tense of irregular verbs, which were not explained by the minimum description length based models, but ours did.

1 Introduction

Vector-space word representations learned from large corpora have been shown to capture semantic and syntactic regularities in language. The regularities are observed as constant vector offsets between pairs of words sharing a particular relationship. The syntactic and morphological relationships inherent in the representations have been observed to include base/comparative/superlative forms of adjectives; singular/plural forms of common nouns; possessive/non-possessive forms of common nouns; and base, past and 3rd person present tense forms of verbs (Mikolov et al., 2013b). Thus, it begs the question of whether the distributed representations of words could actually provide cues for learning the internal structure of words and if it has any relationship to language acquisition in children. The latter question has been studied by (Albright and Hayes, 2003) and (Rumelhart and McClelland, 1985) in learning the past tenses of English verbs, albeit by only considering pairs of inflected forms of a verb while ignoring the context. So, in this paper I would be investigating if the said computational model trained on child-directed speech gives rise to pairs of words related morphologically.

2 Related Work

Unsupervised learning of morphology is of great interest to linguists and cognitive scientists, because it closely resembles the learning situation faced by humans acquiring their first language. A child acquiring English would not know at birth that the inflection *-ed* is a morph, and must learn it based on the linguistic input. (Lignos and Yang) provide an overview of the morphological learning problem in language acquisition, covering issues of data sparsity, productivity, and analogy. Most published work in computational morphology does not speak directly to the problem of human morphological acquisition, because the datasets used are mostly raw corpus text from adult language that is very much unlike child directed speech. Some recent work, however, does use child directed speech, e.g., (Frank et al., 2013) (who also make use of syntactic information, though they do batch learning). (Lee and Goldsmith, 2016) present preliminary results of incremental morphological learning using child

CHILDES Db	#Corp.	#Sentences	#Words
Eng-NA	48	2,078,904	8,916,872
Eng-UK	13	3,085,156	13,140,207
Total	**61**	**5,164,060**	**22,057,079**

Table 1: Statistics about the compilation of CHILDES English corpora used in this project.

directed speech, where they ran Linguistica 5 [1] to successfully induce morphological signatures.
There has not been a lot of work in exploring the use of word embeddings for modeling child language acquisition. However, (Grimm et al., 2015) were able to use prediction-based methods in order to model children's acquisition of syntactic categories, but they used a supervised neural network training where the word embeddings were labeled by syntactic category.

3 Approach

This section proposes a computational framework for using vector representations of words to generate inflected forms of words from their base forms, and also to identify the inflectional categories present in an inflected word.

3.1 Dataset

Child directed sentences were compiled from the CHILDES (MacWhinney, 2000) English-North American and English-UK databases by excluding the sentences spoken by children which were tagged as 'CHI' in the transcripts.
According to (Hart and Risley, 2003), in professional families, the average number of words spoken by adults to children, in a year, is estimated to be around 11.2 million, which translates to around 22 million words in two years. The number of words in the compiled corpus, as shown in Table 1, roughly corresponds to the experiences of a 2-year old.

The CHILDES corpus, also comes with rich word-level morpheme analysis that was used as the ground truth in computing the accuracy of the model's morphological transformations.
We also annotated the verbs in the corpus as being regular vs irregular, by considering the list of irregular verbs from Wikipedia [2]. Of the 204 irregular verbs that are commonly used in standard modern English, 135 were represented in the corpus.
The rest of this section, describes how this dataset was used to compute word representations, and to subsequently learn morphological representations.

3.2 Word Representation

Models that learn representations of words as continuous vectors are built upon the Distributional Hypothesis (Harris, 1954):

'words that occur in similar context tend to have a similar meaning'

Distributional representations derived through context-predicting methods such as *word2vec* have been shown to model vector semantics better (Baroni et al., 2014) than context-counting methods such as Latent Semantic Analysis (Deerwester et al., 1990). Typically, the quality of the word vectors is measured by their ability to answer word analogy style questions of the kind : find a word that is similar to *"small"* in the same sense as *"biggest"* is similar to *"big"*. We can model this using vector algebra as, simply computing $\vec{X} = \overrightarrow{biggest} - \overrightarrow{big} + \overrightarrow{small}$. Then, we search in the vector space for the word closest to X measured by cosine distance as defined in Equation 1 below, and hope to get the vector for *"smallest"* as the answer.

$$\cos(\vec{x}, \vec{y}) = \frac{\tilde{\mathbf{x}} \cdot \tilde{\mathbf{y}}}{\|\tilde{\mathbf{x}}\|\|\tilde{\mathbf{y}}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

The *word2vec* model described in (Mikolov et al., 2013a) is a single-hidden-layer neural network trained to predict target words for neighboring words as shown in Figure 1. The training objective is similar to that of any language model, where we try to maximize the probability of seeing the target word given its context, as shown in Equation 2, where the maximum distance between the target word and a neighboring word is 2.

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t \mid w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2} \quad (2)$$

For learning the word vectors from the CHILDES corpus, we used the implementation of word2vec in the popular unsupervised semantic modeling software, *gensim* (Řehůřek and Sojka, 2010).

[1] https://linguistica-uchicago.github.io/lxa5/
[2] https://en.wikipedia.org/wiki/English_irregular_verbs

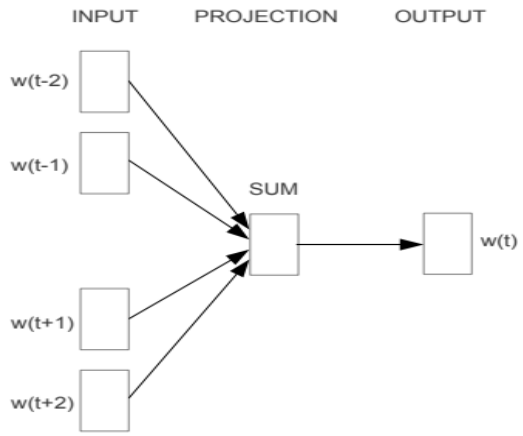

Figure 1: Word2Vec architecture: predicting the current word based on the context.

3.3 Morphological Representation

From a linguistic view, the lemma represents inflected word forms with the same meaning. Thus, the lemma of a word has a large influence on the kinds of context where the word appears. For example, the lemma *"eat"* will likely steer the context to food that is eaten or places where you can eat. By removing the vector influence of the lemma, the influence of the semantic context in the embedding will be reduced thereby emphasizing the morphological information.

As proposed in (Gieske, 2017), the morphology representation for w is obtained by subtracting the word embedding of its lemma w_{lemma} from the embedding of w:

$$\overrightarrow{w_{morph}} = \overrightarrow{w} - \overrightarrow{w_{lemma}} \tag{3}$$

Thus, the past tense morphological representation of the word *"eaten"* would be $\overrightarrow{eaten} - \overrightarrow{eat}$.

Such representations are created for the inflectional categories from the embeddings corresponding to each category, by using the morphological information of inflected forms. Some examples of inflection categories are shown in Table 2.

Since, word embeddings are learned in an unsupervised manner and are limited to a finite set of contexts observed in the training corpus, the morphological representations of inflectional categories taken from individual inflection pairs vary widely. Thus, a centroid representation, $z(\varphi)$ is computed for a given inflection category, φ using the methods of average centroid or prototype centroid described below(Gieske, 2017).

Inflection Category	Base Form	POS	Inflected Form
3S	walk	v	walks
PAST	walk	v	walked
IRR-PAST	swim	v	swam
PASTP	eat	v	eaten
PRESP	eat	v	eating
CP	big	adj	bigger
PL	toy	n	toys

Table 2: Inflection categories with tags from CHILDES : 3S - 3rd person singular; PAST - Simple Past; IRR-PAST - Irregular Past; PASTP - Past Participle; PRESP - Present Progressive; CP - Comparative; PL - Plural.

3.3.1 Average Centroid

To compute the average centroid $z(\varphi)$, for a given category φ we take the arithmetic average of all morphological representations w_{morph} of examples of an inflection category found in the vocabulary V_φ:

$$z(\varphi) = \frac{\sum_{w \epsilon V_\varphi} \overrightarrow{w_{morph}}}{|V_\varphi|} \tag{4}$$

For example, to compute the morphological representation of PAST, we take inflection pair examples such as $eat \rightarrow ate$, $walk \rightarrow walked$, $run \rightarrow ran$ and take an arithmetic average of morphological representations of all inflected forms of this category, as in Equation 4.

3.3.2 Prototype Centroid

Another representation for centroid is created by choosing the morphology representation which is most similar to the rest of the morphology representations of words corresponding to category φ:

$$z(\varphi) = argmax_{\overrightarrow{w_{morph}}} \sum_{w' \epsilon V_\varphi} cos(\overrightarrow{w_{morph}}, \overrightarrow{w'_{morph}}) \tag{5}$$

The similarity metric used is the cosine similarity defined in 1. For e.g., prototype centroid for PAST, would be the morphological representation corresponding to $walked$, if it were closest to all other past tense representations.

4 Results

This section discusses the results of two tasks - accuracy of correctly inflecting a base form and the

accuracy of identifying the inflectional categories in a given inflected form, and also provides a qualitative analysis of the results from Goldsmith's Linguistica.

4.1 Generation of Inflected Forms

Here, we evaluate the performance of the generation of the correct inflected form of a given stem and inflection category by providing the leave-one-out average and prototype centroid representations. The leave one out representations were computed, by taking the examples corresponding to the current stem being evaluated, out of the set of examples being used to compute the centroids. The inflected form is generated by finding the word whose vector is the most similar to the vector sum of the word vector of the stem, $\overrightarrow{stem}$ and the centroid of the given inflectional category, $z(\varphi)$. Since, we get a ranked list of most similar words, we have reported the accuracies by considering only the top ranked result as the answer, and also by considering the list of top n results as candidates for the answer. In Figure 2, the accuracies obtained by using the above mentioned methods are denoted *Accuracy* and *Accuracy@5* respectively.

There are few things to note from the results plotted in Figure 2. Generally, the accuracy increases by a huge margin, by looking at candidates beyond the top-ranked inflected form. This indicates that with further optimization of the centroid computation by using weighting schemes, we would be able to better capture the inflected forms of words. Also, almost always, the prototype centroid is better than the average centroid when $n = 1$, and but, this seems to change when we look at $n = 5$, indicating that the averaging technique needs more fine-tuning to work well when n=1, but it does seem to have a greater potential in being able to fetch the right candidates.

4.2 Identification of Inflectional Categories

In order to analyse the extent to which morphological information is contained in the word embeddings, we consider this task of classifying the inflection categories found in a given inflected word. By identifying the centroids closest to the morphology representation of the inflected verb w_{morph}, we find its corresponding inflectional categories.

For example, 3 shows the cosine similarity scores between the morphological representation of the inflected verbs - *sang* and *walked* computed using 3 and the average centroid representation of the inflection categories. We can see that *sang* has been correctly identified as being an irregular past tense form of a verb, and *walked* has been identified to have the regular past tense form, and also the past participle form of a verb. There are also some negative similarity scores for the egregious pairs such as PL - *walked*, 3S - *sang*, and CP - *walked*. These findings indicate that the vector representations of words have morphological properties embedded in them.

Inflection Category	**Similarity Score *sang***	**Similarity Score *walked***
PL	0.0497672	-0.0224439
PAST	0.133487	**0.314452**
IRR-PAST	**0.142298**	0.29334
PASTP	0.0207341	0.271016
CP	0.116213	-0.0490698
3S	-0.0105123	0.0226166

Table 3: Inflection identification: similarity scores between different inflection categories and the morp. representation of inflected verbs - *sang*, and *walked*. Expected categories are the most similar.

Corpus level quantitative results for inflection category identification can be seen in Figure 3. By looking at the confusion matrix shown in Figure 4, we can see that the confusion is mostly between the classes - PAST and PASTP which could be explained by words like the regular verbs walked,talked,laughes that share the same inflection for both of these tenses. But there are some weird confusions like between PL and Irregular - Past, which I can't think of anything to attribute to.

4.3 Comparison with Linguistica 5

Linguistica 5 is the latest version of Goldsmiths Linguistica (Goldsmith, 2001, 2006), for unsupervised learning of morphology from a given corpus, using the minimum description length (MDL) algorithm. On running algorithm on the same corpus used by the vector space model, there were a total of 222 morphological signatures identified, with the top few signatures listed in the Table 4. It was interesting to observe from the full list of signatures and their associated stems, that there were quite a few regular inflected forms such as the plu-

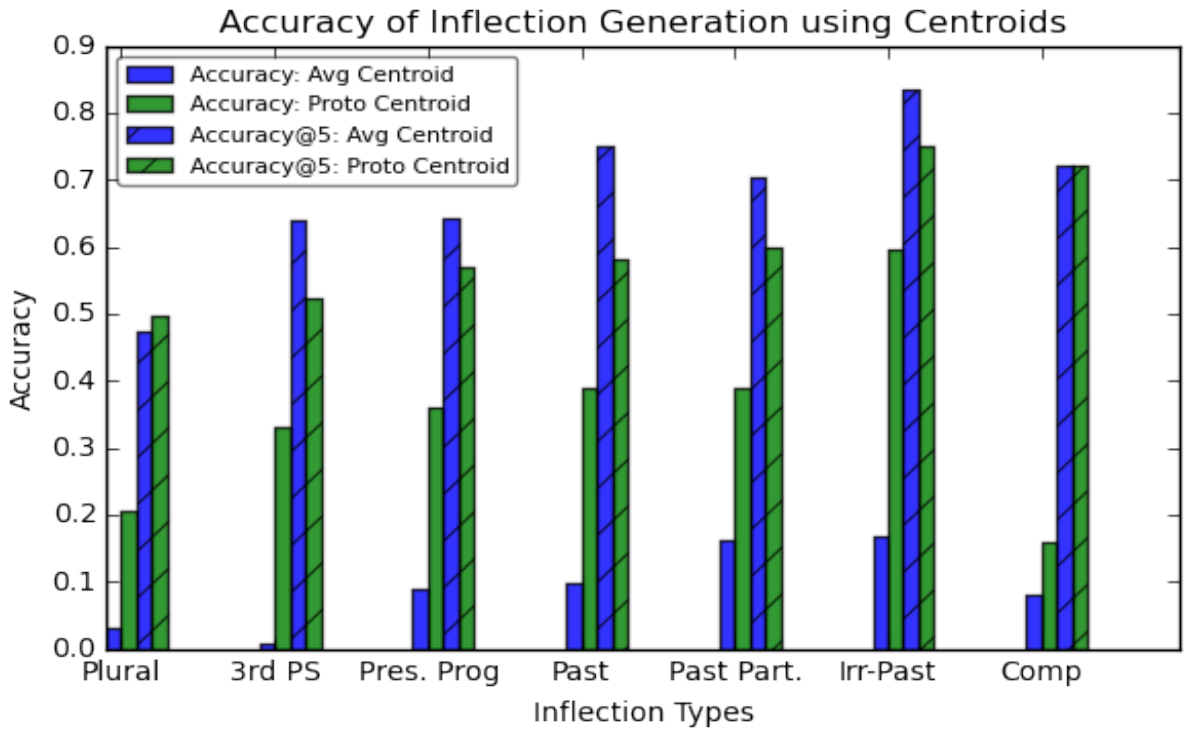

Figure 2: Accuracy of inflection generation from base forms, using average and prototype centroids, evaluated at $n = 1$, and $n = 5$. n is the number of top ranked candidates considered.

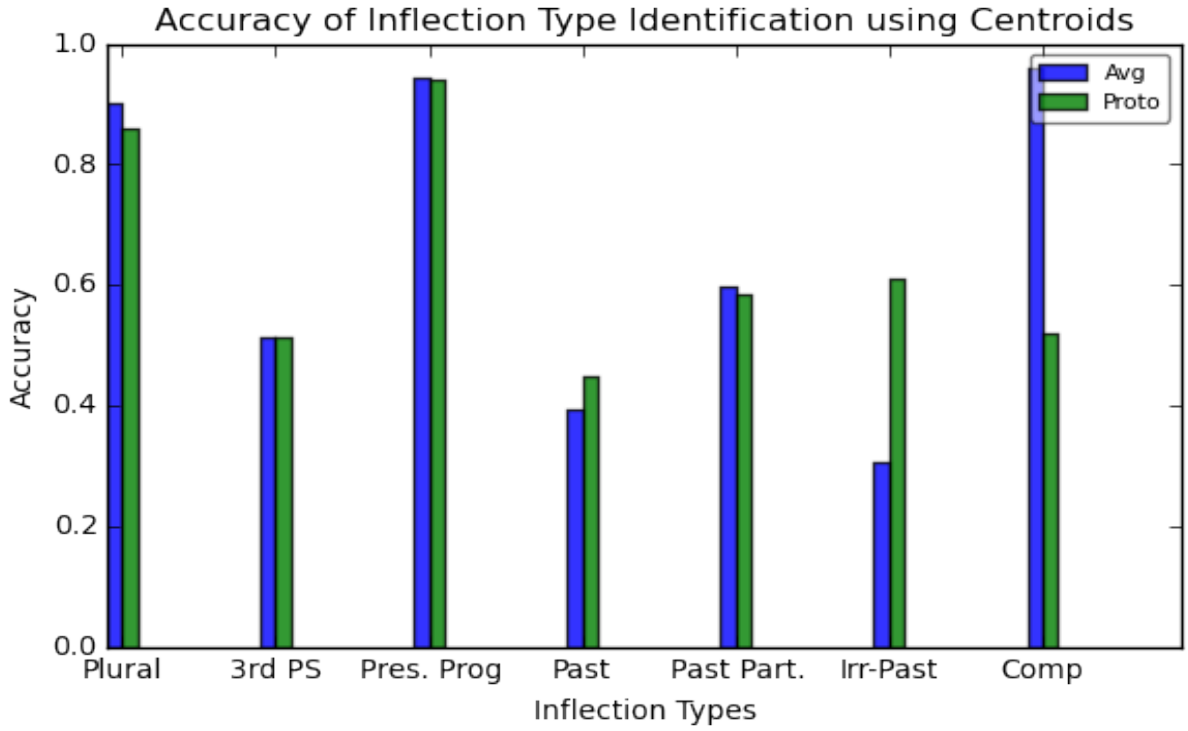

Figure 3: Acuracy of inflection category identification measured as a percentage of inflected forms classified correctly.

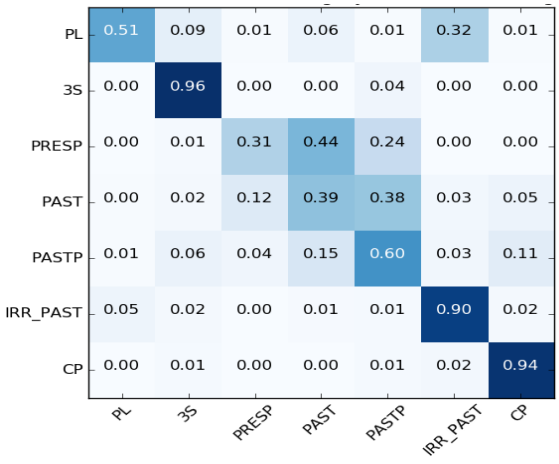

Figure 4: Confusion matrix for inflection category identification using Average Centroids. On the Y-axis are the true categories, and on the X-axis are the predicted categories.

Morphological Signature	Example Stem	#Stems
NULL/s	pennie	1435
's/NULL	merchant	1028
NULL/g	forgettin	435
's/NULL/s)	pheasant	387
NULL/d	startle	154
d/r	discovere	150
NULL/ly	partial	121

Table 4: Morphological signatures and example stems identified by Linguistica 5.

rals - *chairs, carrots, trucks* and past - *collected, lived, laughed* were unaccounted for. Remarkably, none of the irregular verbs identified in the corpus, were actually accounted for in the signatures as well. This does make sense because MDL works by computing count based probabilities of string affixes and edit distances, and the irregular forms do not fit the pattern, because by definition, they have significantly different affixes compared to other regular verbs.

5 Conclusion

In this project, we looked at the effectiveness of using distributed representations of words learned from child directed corpus, for producing inflections of words, and also for identifying the inflection category present in a given inflection form, through the use of centroid representations. We also found that the model is particularly good at capturing irregular past tenses, which unsupervised morphological models like Linguistica completely fail to account for.

However, a major limitation of the model discussed in this study, is that the lemma is assumed to be known beforehand for computing morphological representations devoid of other semantic influence. Also, the computation of centroids is currently supervised as it relies on morphologically annotated base-inflected pairs. It would be interesting to study if there are a few canonical inflection pairs in a given category that have an exemplar effect on learning the morphological representation of the whole category.

The effectiveness of using word2vec embeddings in NLP applications, has led to a surge in unsupervised approaches to learn semantically motivated word representations, foremost among which are Glove (Pennington et al., 2014) and Fasttext (Bojanowski et al., 2016). So, it would be interesting to repeat this study using these embeddings. Also of considerable import would be to look at different language settings. I'm excited about the future direction of this work!

6 Acknowledgements

I would like to thank Gaja and Carolyn for providing valuable suggestions during their office hours, which helped immensely in conducting this research.

References

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition* 90(2):119–161.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*. pages 238–247.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6):391.

Stella Frank, Frank Keller, and Sharon Goldwater. 2013. Exploring the utility of joint morphological and syntactic learning from child-directed speech. In *EMNLP*. pages 30–41.

SA Gieske. 2017. Inflecting verbs with word embeddings: A systematic investigation of morphological information captured by german verb embeddings .

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics* 27(2):153–198.

John Goldsmith. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12(4):353–371.

Robert Grimm, Giovanni Cassani, Walter Daelemans, and Steven Gillis. 2015. Towards a model of prediction-based syntactic category acquisition: First steps with word embeddings. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*. pages 28–32.

Zellig S Harris. 1954. Distributional structure. *Word* 10(2-3):146–162.

Betty Hart and Todd Risley. 2003. The early catastrophe. *American Educator* 27(4):6–9.

Jackson Lee and John Goldsmith. 2016. Linguistica 5: Unsupervised learning of linguistic structure. In *HLT-NAACL Demos*. pages 22–26.

Constantine Lignos and Charles Yang. ???? Morphology and language acquisition. *Cambridge Handbook of Morphology. Cambridge University Press, To appear* .

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *hlt-Naacl*. volume 13, pages 746–751.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50. http://is.muni.cz/publication/884893/en.

David E Rumelhart and James L McClelland. 1985. On learning the past tenses of english verbs. Technical report, California Univ San Diego La Jolla Inst For Cognitive Science.