Inferring Psycholinguistic Properties of Words

Gustavo Henrique Paetzold and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{ghpaetzold1,l.specia}@sheffield.ac.uk

Abstract

We introduce a bootstrapping algorithm for regression that exploits word embedding models. We use it to infer four psycholinguistic properties of words: Familiarity, Age of Acquisition, Concreteness and Imagery and further populate the MRC Psycholinguistic Database with these properties. The approach achieves 0.88 correlation with human-produced values and the inferred psycholinguistic features lead to state-of-the-art results when used in a Lexical Simplification task.

1 Introduction

Throughout the last three decades, much has been found on how the psycholinguistic properties of words influence cognitive processes in the human brain when a subject is presented with either written or spoken forms. A word's Age of Acquisition is an example. The findings in (Carroll and White, 1973) reveal that objects whose names are learned earlier in life can be named faster in later stages of life. Zevin and Seidenberg (2002) show that words learned in early ages are orthographically or phonologically very distinct from those learned in adult life.

Other examples of psycholinguistic properties, such as Familiarity and Concreteness, influence one's proficiency in word recognition and text comprehension. The experiments in (Connine et al., 1990; Morrel-Samuels and Krauss, 1992) show that words with high Familiarity yield lower reaction times in both visual and auditory lexical decision, and require less hand gesticulation in order to be described. Begg and Paivio (1969) found that humans are less sensitive to changes in wording made to sentences with high Concreteness words.

When quantified, these aspects can be used as features for various Natural Language Processing (NLP) tasks. The Lexical Simplification approach in (Jauhar and Specia, 2012) is an example. By combining various collocational features and psycholinguistic measures extracted from the MRC Psycholinguistic Database (Coltheart, 1981), they trained a ranker (Joachims, 2002) that reached first place in the English Lexical Simplification task at SemEval 2012. Semantic Classification tasks have also benefited from the use of such features: by combining Concreteness with other features, (Hill and Korhonen, 2014) reached the state-of-the-art performance in Semantic Composition (*denotative/connotative*) and Semantic Modification (*intersective/subsective*) prediction.

Despite the evident usefulness of psycholinguistic properties of words, resources describing such properties are rare. The most extensively developed resource for English is the MRC Psycholinguistic Database (Section 2). However, it is far from complete, most likely due to the inherent cost of manually entering such properties. In this paper we propose a method to automatically infer these missing properties. We train regressors by performing bootstrapping (Yarowsky, 1995) over the existing features in the MRC database, exploiting word embedding models and other linguistic resources for that (Section 3). This approach outperform various strong baselines (Section 4) and the resulting properties lead to significant improvements when used in Lexical Simplification models (Section 5).

Proceedings of NAACL-HLT 2016, pages 435–440,
San Diego, California, June 12-17, 2016.

2 The MRC Psycholinguistic Database

Introduced by Coltheart (1981), the MRC (Machine Readable Dictionary) Psycholinguistic Database is a digital compilation of lexical, morphological and psycholinguistic properties for 150,837 words. The 27 psycholinguistic properties in the resource range from simple frequency measures (Rudell, 1993) to elaborate measures estimated by humans, such as Age of Acquisition and Imagery (Gilhooly and Logie, 1980). However, despite various efforts to populate the MRC Database, these properties are only available for small subsets of the 150,837 words.

We focus on four manually estimated psycholinguistic properties in the MRC Database:

- **Familiarity**: The frequency with which a word is seen, heard or used daily. Available for 9,392 words.
- **Age of Acquisition**: The age at which a word is believed to be learned. Available for 3,503 words.
- **Concreteness**: How "palpable" the object the word refers to is. Available for $8,228$ words.
- **Imagery**: The intensity with which a word arouses images. Available for 9,240 words.

All four properties are real values, determined based on different quantifiable metrics. We focus on these properties since they have been proven useful and are some of the most scarce in the MRC Database. As we discussed in Section 1, these properties have been successfully used in various approaches for Lexical Simplification and Semantic Classification, and yet are available for no more than 6% of the words in the MRC Database.

3 Bootstrapping with Word Embeddings

In order to automatically estimate missing psycholinguistic properties in the MRC Database, we resort to bootstrapping. We base our approach on that by (Yarowsky, 1995), a bootstrapping algorithm which aims to learn a classifier over a reduced set of annotated training instances (or "seeds"). It does so by performing the following five steps:

1. Initialise training set S with the seeds available.
2. Train a classifier over S.
3. Predict values for a set of unlabelled instances U.
4. Add to S all instances from U for which the prediction confidence c is equal or greater than ζ.
5. If at least one instance was added to S, go to step 2, otherwise, return the resulting classifier.

One critical difference between this approach and ours is that our task requires regression algorithms instead of classifiers. In classification, the prediction confidence c is often calculated as the maximum signed distance between an instance and the estimated hyperplanes. There is, however, no analogous confidence estimation technique for regression problems. We address this problem by using word embedding models.

Embedding models have been proved effective in capturing linguistic regularities of words (Mikolov et al., 2013b). In order to exploit these regularities, we assume that the quality of a regressor's prediction on an instance is directly proportional to how similar the instance is to the ones in the labelled set. Since the input for the regressors are words, we compute the similarity between a test word and the words in the labelled dataset as the maximum cosine similarity between the test word's vector and the vectors in the labelled set.

Let M be an embeddings model trained over vocabulary V, S a set of training seeds, ζ a minimum confidence threshold, $sim(w, S, M)$ the maximum cosine similarity between word w and S with respect to model M, R a regression model, and $R(w)$ its prediction for word w. Our bootstrapping algorithm is depicted in Algorithm 1.

Algorithm 1: Regression Bootstrapping

```
input: M, V, S, ζ;
output: R;
repeat
    Train R over S;
    for w ∈ V − S do
        if sim(w, S, M) ≥ ζ then
            Add ⟨w, R(w)⟩ to S;
        end
    end
until ‖S‖ converges ;
```

We found that 64,895 out of the 150,837 words in the MRC database were not present in either WordNet or our word embedding models. Since our bootstrappers use features extracted from both these resources, we were only able to predict the Familiarity, Age of Acquisition, Concreteness and Imagery values of the remaining 85,942 words in MRC.

4 Evaluation

Since we were not able to find previous work for this task, in these experiments, we compare the performance of our bootstrapping strategy to various baselines. For training, we use the Ridge regression algorithm (Tikhonov, 1963). As features, our regressor uses the word's raw embedding values, along with the following 15 lexical features:

- Word's length and number of syllables, as determined by the Morph Adorner module of LEXenstein (Paetzold and Specia, 2015).
- Word's frequency in the Brown (Francis and Kucera, 1979), SUBTLEX (Brysbaert and New, 2009), SubIMDB (Paetzold and Specia, 2016), Wikipedia and Simple Wikipedia (Kauchak, 2013) corpora.
- Number of senses, synonyms, hypernyms and hyponyms for word in WordNet (Fellbaum, 1998).
- Minimum, maximum and average distance between the word's senses in WordNet and the thesaurus' root sense.
- Number of images found for word in the Getty Images database[1].

We train our embedding models using word2vec (Mikolov et al., 2013a) over a corpus of 7 billion words composed by the SubIMDB corpus, UMBC webbase[2], News Crawl[3], SUBTLEX (Brysbaert and New, 2009), Wikipedia and Simple Wikipedia (Kauchak, 2013). We use 5-fold cross-validation to optimise parameters: ζ, embeddings model architecture (CBOW or Skip-Gram), and word vector size (from 300 to 2,500 in intervals of 200). We include four strong baseline systems in the comparison:

- **Max. Similarity**: Test word is assigned the property value of the closest word in the training set, i.e. the word with the highest cosine similarity according to the word embeddings model.
- **Avg. Similarity**: Test word is assigned the average property value of the n closest words in the training set, i.e. the words with the highest cosine similarity according to the word embeddings model. The value of n is decided through 5-fold cross validation.
- **Simple SVM**: Test word is assigned the property value as predicted by an SVM regressor (Smola and Vapnik, 1997) with a polynomial kernel trained with the 15 aforementioned lexical features.
- **Simple Ridge**: Test word is assigned the property value as predicted by a Ridge regressor trained with the 15 aforementioned lexical features.
- **Super Ridge**: Identical to Simple Ridge, the only difference being that it also includes the words embeddings in the feature set. We note that this baseline uses the exact same features and regression algorithm as our bootstrapped regressors.

The parameters of all baseline systems are optimised following the same method as with our approach. We also measure the correlation between each of the aforementioned lexical features and the psycholinguistic properties. For each psycholinguistic property, we create a training and a test set by splitting the labelled instances available in the MRC Database in two equally sized portions. All training instances are used as seeds in our approach. As evaluation metrics, we use Spearman's (ρ) and Pearson's (r) correlation. Pearson's correlation is the most important indicator of performance: an effective regressor would predict values that change linearly with a given psycholinguistic property.

The results are illustrated in Table 1. While the similarity-based approaches tend to perform well for Concreteness and Imagery, typical regressors capture Familiarity and Age of Acquisition more effectively. Our approach, on the other hand, is consistently superior for all psycholinguistic properties, with both Spearman's and Pearson's correlation

[1] http://developers.gettyimages.com/

[2] http://ebiquity.umbc.edu/resource/html/id/351

[3] http://www.statmt.org/wmt11/translation-task.html

System	Familiarity		Age of Acquisition		Concreteness		Imagery	
	ρ	r	ρ	r	ρ	r	ρ	r
Word Length	-0.238	-0.171	0.501	0.497	-0.170	-0.195	-0.190	-0.193
Syllables	-0.168	-0.114	0.464	0.458	-0.207	-0.238	-0.218	-0.224
Freq: SubIMDB	0.798	**0.725**	**-0.679**	**-0.699**	0.048	0.003	0.208	0.170
Freq: SUBTLEX	**0.827**	0.462	-0.646	-0.251	0.028	0.137	0.187	0.265
Freq: SimpleWiki	0.725	0.488	-0.453	-0.306	0.015	0.145	0.119	0.247
Freq: Wikipedia	0.694	0.283	-0.349	-0.112	-0.076	0.081	0.027	0.134
Freq: Brown	0.706	0.608	-0.380	-0.395	-0.155	-0.214	-0.054	-0.107
Sense Count	0.471	0.363	-0.429	-0.391	0.020	-0.017	0.119	0.059
Synonym Count	0.411	0.336	-0.381	-0.357	-0.036	-0.047	0.070	0.035
Hypernym Count	0.307	0.295	-0.411	-0.387	0.167	0.088	0.268	0.160
Hyponym Count	0.379	0.245	-0.324	-0.196	0.120	0.002	0.196	0.023
Min. Sense Depth	-0.347	-0.072	0.366	0.055	0.151	-0.185	0.127	-0.224
Max. Sense Depth	-0.021	-0.008	-0.197	-0.196	**0.447**	**0.455**	**0.415**	**0.414**
Avg. Sense Depth	-0.295	-0.256	0.215	0.183	0.400	0.428	0.345	0.347
Img. Search Count	0.544	0.145	-0.325	-0.033	-0.037	-0.073	0.117	-0.059
Max. Similarity	0.406	0.402	0.445	0.443	0.742	0.743	0.618	0.605
Avg. Similarity	0.528	0.527	0.536	0.535	0.826	0.819	0.733	0.707
Simple SVM	0.835	0.815	0.778	0.770	0.548	0.477	0.555	0.528
Simple Ridge	0.832	0.815	0.785	0.778	0.603	0.591	0.620	0.613
Super Ridge	**0.847**	**0.833**	**0.827**	**0.820**	**0.859**	**0.852**	**0.813**	**0.800**
Bootstrapping	**0.863**	**0.846**	**0.871**	**0.862**	**0.876**	**0.869**	**0.835**	**0.823**

Table 1: Regression correlation scores. In bold are the highest scores within a group (features, baselines, proposed approach), and underlined the highest scores overall.

scores varying between 0.82 and 0.88. The difference in performance between the Super Ridge baseline and our approach confirm that our bootstrapping algorithm can in fact improve on the performance of a regressor.

The parameters used by our bootstrappers, which are reported below, highlight the importance of parameter optimization in out bootstrapping strategy: its performance peaked with very different configurations for most psycholinguistic properties:

- **Familiarity**: 300 word vector dimensions with a Skip-Gram model, and $\zeta = 0.9$.
- **Age of Acquisition**: 700 word vector dimensions with a CBOW model, and $\zeta = 0.7$.
- **Concreteness**: 1,100 word vector dimensions with a Skip-Gram model, and $\zeta = 0.7$.
- **Imagery**: 1,100 word vector dimensions with a Skip-Gram model, and $\zeta = 0.7$.

Interestingly, frequency in the SubIMDB corpus[4], composed of over 7 million sentences extracted from subtitles of "family" movies and series, has good linear correlation with Familiarity and Age of Acquisition, much higher than any other feature. For Concreteness and Imagery, on the other hand, the results suggest something different: the further a word is from the root of a thesaurus, the most likely it is to refer to a physical object or entity.

5 Psycholinguistic Features for LS

Here we assess the effectiveness of our bootstrappers in the task of Lexical Simplification (LS). As shown in (Jauhar and Specia, 2012), psycholinguistic features can help supervised ranking algorithms capture word simplicity. Using the parameters described in Section 4, we train bootstrappers for these two properties using all instances in the MRC Database as seeds. We then train three rankers with (W) and without (W/O) psycholinguistic features:

[4] http://ghpaetzold.github.io/subimdb

- **Horn** (Horn et al., 2014): Uses an SVM ranker trained on various n-gram probability features.
- **Glavas** (Glavaš and Štajner, 2015): Ranks candidates using various collocational and semantic metrics, and then re-ranks them according to their average rankings.
- **Paetzold** (Paetzold and Specia, 2015): Ranks words according to their distance to a decision boundary learned from a classification setup inferred from ranking examples. Uses n-gram frequencies as features.

We use data from the English Lexical Simplification task of SemEval 2012 to assess systems' performance. The goal of the task is to rank words in different contexts according to their simplicity. The training and test sets contain 300 and 1,710 instances, respectively. The official metric from the task – TRank (Specia et al., 2012) – is used to measure systems' performance. As discussed in (Paetzold, 2015), this metric best represents LS performance in practice. The results in Table 2 show that the addition of our features lead to performance increases with all rankers. Performing F-tests over the rankings estimated for the simplest candidate in each instance, we have found these differences to be statistically significant ($p < 0.05$). Using our features, the **Paetzold** ranker reaches the best published results for the dataset, significantly superior to the best system in SemEval (Jauhar and Specia, 2012).

	TRank	
Ranker	W/O	W
Best SemEval	-	0.602
Horn	0.625	**0.635**
Glavas	0.623	**0.636**
Paetzold	0.653	**0.657**

Table 2: Results on SemEval 2012 LS task dataset

6 Conclusions

Overall, the proposed bootstrapping strategy for regression has led to very positive results, despite its simplicity. It is therefore a cheap and reliable alternative to manually producing psycholinguistic properties of words. Word embedding models have proven to be very useful in bootstrapping, both as surrogates for confidence predictors and as regression features. Our findings also indicate the usefulness of individual features and resources: word frequencies in the SubIMDB corpus have a much stronger correlation with Familiarity and Age of Acquisition than previously used corpora, while the depth of a word's in a thesaurus hierarchy correlates well with both its Concreteness and Imagery.

In future work we plan to employ our bootstrapping solution in other regression problems, and to further explore potential uses of automatically learned psycholinguistic features.

The updated version of the MRC resource can be downloaded from http://ghpaetzold.github.io/data/BootstrappedMRC.zip.

References

Ian Begg and Allan Paivio. 1969. Concreteness and imagery in sentence meaning. *Journal of Verbal Learning and Verbal Behavior*, 8(6):821–827.

Marc Brysbaert and Boris New. 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*, 41:977–990.

John B Carroll and Margaret N White. 1973. Word frequency and age of acquisition as determiners of picture-naming latency. *The Quarterly Journal of Experimental Psychology*, 25(1):85–95.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505.

Cynthia M Connine, John Mullennix, Eve Shernoff, and Jennifer Yelen. 1990. Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(6):1084.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Brown University*.

Kenneth J Gilhooly and Robert H Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4):395–427.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd ACL*.

Felix Hill and Anna Korhonen. 2014. Concreteness and subjectivity as dimensions of lexical meaning. In *Proceedings of ACL*, pages 725–731.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd ACL*, pages 458–463.

S. Jauhar and L. Specia. 2012. Uow-shef: Simplex–lexical simplicity ranking based on contextual and psycholinguistic features. In *Proceedings of the 1st SemEval*, pages 477–481.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM*, pages 133–142.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st ACL*, pages 1537–1546.

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*, pages 746–751.

Palmer Morrel-Samuels and Robert M Krauss. 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3):615.

Gustavo Henrique Paetzold and Lucia Specia. 2015. Lexenstein: A framework for lexical simplification. In *Proceedings of The 53rd ACL*.

Gustavo Henrique Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of The 30th AAAI*.

Gustavo Henrique Paetzold. 2015. Reliable lexical simplification for non-native speakers. In *Proceedings of the 2015 NAACL Student Research Workshop*.

Allan P. Rudell. 1993. Frequency of word usage and perceived word difficulty: Ratings of Kuera and Francis words. *Behavior Research Methods*.

Alex Smola and Vladimir Vapnik. 1997. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 1st SemEval*, pages 347–355.

Andrey Tikhonov. 1963. Solution of incorrectly formulated problems and the regularization method. In *Soviet Math. Dokl.*, volume 5, pages 1035–1038.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics.

Jason D Zevin and Mark S Seidenberg. 2002. Age of acquisition effects in word reading and other tasks. *Journal of Memory and language*, 47(1):1–29.