Learning Concept Abstractness Using Weak Supervision

Ella Rabinovich[▲][*]   **Benjamin Sznajder**[†]   **Artem Spector**[†]   **Ilya Shnayderman**[†]
Ranit Aharonov[†]   **David Konopnicki**[†]   **Noam Slonim**[†]

[†]IBM Research
▲ Dept. of Computer Science, University of Toronto
{benjams, artems, ilyashn, ranita, davidko, noams}@il.ibm.com
ella@cs.toronto.edu

Abstract

We introduce a weakly supervised approach for inferring the property of abstractness of words and expressions in the complete absence of labeled data. Exploiting only minimal linguistic clues and the contextual usage of a concept as manifested in textual data, we train sufficiently powerful classifiers, obtaining high correlation with human labels. The results imply the applicability of this approach to additional properties of concepts, additional languages, and resource-scarce scenarios.

1 Introduction

During the last decades, the influence of psycholinguistic properties of words on cognitive processes has become a major topic of scientific inquiry. Among the most studied psycholinguistic attributes are concreteness, familiarity, imagery, and average age of acquisition. *Abstractness* (the opposite of concreteness) quantifies the degree to which an expression denotes an entity that can be directly perceived by human senses.

Word abstractness ratings were first collected by Spreen and Schulz (1966) and Paivio et al. (1968), and made available in the MRC database (Coltheart, 1981) for 4,292 English words. Since its release, this database has stimulated research in a wide range of linguistic tasks, as well as artificial intelligence and cognitive studies. Despite their evident usefulness, resources providing abstractness ratings are relatively rare and of limited size. Here, we address the task of *automatically* inferring the abstractness rating of a concept by applying a *weakly supervised* approach that exploits minimal linguistic clues.

Studies on derivational morphological processes indicate that word meaning is often entailed by its morphology. As an example, word suffixation by *-ant* or *-ent* is used to denote a person, as

in *assistant*, while the suffix *-hood* yields nouns meaning "condition of being", as in *childhood*. A wide range of word-formation processes was described by Huddleston and Pullum (2002); in particular, the authors detail categories of suffixes that are used to derive words, broadly perceived as abstract, e.g., *-ism* as in *feminism*, or *-ness* as in *agreeableness*.

Concept abstractness indicators are also likely to be manifested in its contextual usage. Consider the two sentences below, each embedding abstract and concrete words – one describing *feminism* and the other *screwdriver* – respectively:

Second- and third-wave **feminism** *in China involved a reexamination of women's* <u>roles</u> *during the* <u>communist</u> <u>revolution</u> *and other* <u>reform</u> <u>movements</u>*, and new discussions about whether women's* <u>equality</u> *has been fully achieved.*

Many **screwdriver** <u>handles</u> *are not* <u>smooth</u> *and often not* <u>round</u>*, but have* <u>bumps</u> *or other irregularities to improve* <u>grip</u> *and to prevent the* <u>tool</u> *from* <u>rolling</u> *when on a flat* <u>surface</u>*.*

We hypothesize that the immediate neighborhood of a word as reflected in embedding sentences captures the signal of abstractness. In the examples above, several potential clues for the degree of word abstractness are underlined.

Correspondingly, we propose a method for inferring the degree of abstractness of concepts in the complete absence of labeled data, by exploiting (1) a minimal set of morphological word-formation clues; and (2) a text corpus for learning the context in which words tend to appear.

We demonstrate that this method allows us to infer the abstractness ratings of unigram, bigram and trigram Wikipedia concepts (titles) – the task that, to the best of our knowledge, was only addressed through manual labeling so far (Brysbaert et al., 2014). The main contribution of this work is, therefore, in the proposal and evaluation of a

---

*Work done while the author was at IBM Research.

weakly supervised methodology for inferring the abstractness rating of concepts, potentially applicable to additional languages. The suggested approach may also be applicable for predicting other word and concept properties, when those are manifested in both morphology and context. Finally, we release a dataset of 300K Wikipedia concepts automatically rated for their degree of abstractness, and additional 1500 unigram, bigram and trigram concepts annotated with both manual and predicted scores.[1]

2 Related work

A large body of research addressed the relations of word abstractness and cognitive processes (Connell and Lynott, 2012; Gianico-Relyea and Altarriba, 2012; Oliveira et al., 2013; Nishiyama, 2013; Paivio, 2013; Barber et al., 2013). Computational investigation of word abstractness and concreteness has been a prolific field of recent research, laying out an empirical foundation for the theoretically motivated hypotheses on the characteristics of these properties. A ranker trained on psycholinguistic features extracted from the MRC database (in combination with other features) reached first place in the English Lexical Simplification task at SemEval 2012 (Jauhar and Specia, 2012). Hill and Korhonen (2014) achieved state-of-the-art performance in Semantic Composition and Semantic Modification prediction by including concreteness in the set of features used by the model.

Along the years, several works extended the seed MRC dataset by employing various *supervised* machine learning techniques, further utilizing the extended dataset for tasks of lexical simplification (Paetzold and Specia, 2016b,a), cross-lingual metaphor detection (Tsvetkov et al., 2013), literal and metaphorical sense identification (Turney et al., 2011), as well as readability assessment of Brazilian Portuguese (dos Santos et al., 2017). Feng et al. (2011) exploited word attributes from WordNet, properties extracted from the CELEX database, and Latent Semantic Analysis over a large text corpus for building a linear regression model predicting abstractness rate; the model accounted for 64% variance of human annotations.

A comprehensive survey of psycholinguistic and memory research on word concreteness is presented in Brysbaert et al. (2014) (BWK), who conducted a large-scale manual annotation of concreteness ratings for over 40K concepts, further used by Rothe et al. (2016) to infer concreteness ratings for the whole Google News lexicon. To the best of our knowledge, our work is the first attempt to automatically infer the property of concept abstractness in the complete absence of labeled data.

3 Predicting concept abstractness

3.1 Abstractness indicators

Nominalization is a word-formation process that involves the formation of nouns from bases of other classes by means of affixation. As an example, a derivational suffix can be added to an adjective (capable+*ity* for *capability*) or a verb (react+*tion* for *reaction*) to create a noun. Various word-formation processes often enrich words with meaning associated with certain semantic grouping. Huddleston and Pullum (2002) detail nominalization processes that serve to form nouns denoting a "state" or "condition of being", which in turn are broadly associated with abstractness. As such, the suffixes *-ety, -ity* and *-ness* carry over the general meaning of "quality or state of being" and the suffix *-ism* is used to form nouns denoting a range of doctrines, beliefs and movements (Huddleston and Pullum, 2002). Additional suffixes that tend to form English nouns with high degree of abstractness include *-ance, -ence, -ation, -ution, -dom, -hood, -ship* and *-y*.

3.2 Dataset

We used the English Wikipedia[2] article titles as a proxy for retrieving frequently used single- and multi-word expressions, thereby associating over 5M Wikipedia titles with concepts.

Training data We chose two abstractness signals, manifested by the suffixes *-ism* and *-ness*, representing different types of abstract meanings. We extracted 1,040 potentially abstract unigram Wikipedia titles suffixed by either of the two (the positive class). The – admittedly noisy – concrete (negative) class was generated by randomly selecting the same number of unigram concepts from the complementary set of titles.In both cases, we set a threshold[3] on the frequency of a concept in the

corpus, and filtered out non-alphabetic unigrams and unigrams containing special characters. We assessed the quality of the positive and negative weakly-labeled training unigrams by manual annotation of their level of abstractness, obtaining abstractness prior of 93% in the set of presumably abstract concepts, and concreteness prior of 81% for the opposite class.

Given this set of weakly-labeled positive and negative concepts, we randomly selected a set of Wikipedia sentences that include any of these concepts (equally split by positive and negative unigrams), to be used in the training phase, while limiting sentence length to the range of 10 to 70 tokens. This step resulted in about 400K train sentences in each class, 800K in total. The final preprocessing phase involved masking a sentence concept with a generic token, aiming to prevent the classifier from training on the concept itself, and instead training on its contextual usage.

Evaluation data A randomly selected set of 1500 Wikpedia concepts (with the minimum of 500 occurrences per concept), split equally between unigrams, bigrams and trigrams, and distinct from the training set, was used for testing prediction. We henceforth refer to this set of concepts as the *evaluation set*. Each of these concepts was manually annotated for abstractness on the 1–7 scale by seven in-house labelers, using an adaptation of the guidelines by Spreen and Schulz (1966) to the multi-word scenario:

Words or phrases may refer to persons, places and things that can be seen, heard, felt, smelled or tasted or to more abstract concepts that cannot be experienced by our senses. The purpose of this task is to rate a list of concepts with respect to "concreteness" in terms of sense-experience. Any expression that refers to objects, materials or persons should receive a high concreteness rating; any expression that refers to an abstract concept that cannot be experienced by the senses should receive a low concreteness rating. Concrete concepts typically have physical or concrete existence, while abstract do not. Think of the concepts "onion" and "nationalism" – "onion" can be experienced by our senses and therefore should be rated as concrete (1); "nationalism" cannot be experienced by the senses as such and therefore should be rated as abstract (7).

Word polysemy is a common challenge in tasks related to lexical semantics. As such, our percep-

tion of the concreteness rate of the concept *bank* may vary depending on whether a financial institution or a river bank is concerned. While we could not avoid this issue altogether (since working with pre-trained word representations that do not carry disambiguation information), we ensured that all in-house labelers annotated the same word sense by providing them with Wikipedia definition of the most frequent sense of a concept.

The final abstractness score was computed as the average over individual annotations. The average pairwise weighted Kappa agreement[4] on the entire set of 1500 concepts was 0.65.

3.3 Classification models

We hypothesize that words that share similar degree of abstractness tend to share certain similarities in their contextual usage; that, in contrast to concepts that exhibit opposite abstractness rate. Indeed, a statistical significance test applied to the (weak) positive and negative training data (Section 3.2) reveals markers such as {*parish, movement, century, spiritual, life, doctrine, nature, regime*} sharing excessive frequency in sentences containing abstract concepts. The very essence of this phenomenon is captured by distributed word representations (Mikolov et al., 2013; Pennington et al., 2014), a.k.a. word embeddings, learned based on the contextual usage of words. We therefore trained three classifiers, each exploiting different language properties, as described below.

Naive Bayes (NB) Using solely word counts in textual data, we used a simple probabilistic Naive Bayes classifier, with a bag-of-words feature set extracted from the 800K sentences containing positive and negative training concepts. Given a sentence containing a test concept, its degree of abstractness was defined as the posterior probability assigned by the classifier. Aiming at robust classification, we retrieved 500 sentences containing each test concept from the corpus. Consequently, the final abstractness score of a concept was calculated by averaging the predictions assigned by the classifier to individual sentences.

Nearest neighbor We used the nearest neighbors algorithm, specifically, its *radius-based* version (NN-RAD), using the pre-trained GloVe embeddings (Pennington et al., 2014). This classifier

---

[4]We used the implementation in `http://scikit-learn.org`, with "quadratic" scheme.

estimates the degree of concept abstractness given only its distributional representation.

The abstractness score of a test concept was computed by the ratio of its abstract neighbors to the total number of concepts within the predefined radius, where the entire set of neighbors is limited to the concepts in the weakly-labeled training set. The proximity threshold (radius) was set to 0.25, w.r.t. the cosine similarity between two embedding vectors.[5] Multi-word concepts were subject to more careful processing, where the classifier computed a multi-word concept representation as an average of representations of its individual words, and further estimated the abstractness score of the obtained embedding. In case that one of a concept constituents was not found in embeddings, we excluded the concept from computation.

RNN Aiming at exploiting both embeddings and textual data, we utilized a bidirectional recurrent neural network (RNN) with one layer of forward and backward LSTM cells. Each cell has width of 128, and is wrapped by a dropout wrapper with keep probability 0.85. An attention layer was created in order to weigh words according to their proximity to the train/test concept. The output of the LSTM cells is passed to the attention layer which reduces it to the size of 100. The output of the attention layer is passed to a fully connected layer which produces the final prediction of the abstractness level of a concept. GloVe embeddings with 300 dimensions were used as word representations. Given a set of sentences containing a test concept, its final abstractness score was computed by applying the averaging procedure described for the Naive Bayes classifier.

4 Results

We demonstrate that trained models discover linguistic patterns associated with abstract meaning (beyond those known at training), and furthermore yield abstractness scores that correlate significantly with human annotations.

4.1 Revealing abstractness markers

We automatically scored 100K unigram Wikipedia concepts for abstractness with all classifiers and extracted the set of suffixes that share excessive frequency in the top-k abstract concepts using the statistical proportion test. More specifically,

we applied the test to the exhaustive list of all three-character English suffixes (e.g., *-aaa, -aab*), counting their occurrences in the subset of concepts with the highest abstractness scores[6] (the population under test) and in the remainder (the background). Our hypothesis was that suffixes associated with abstract meaning in the literature will be over-represented in the population of concepts ranked as abstract by the classifiers. The top-10 suffixes, scored by their statistical significance p-value[7] were {*-ism, -ity, -ion, -sis, -ics, -ess, -phy, -nce, -ogy, -ing*} – suffixes broadly associated with abstractness in the literature (where all suffixes but two are distinct from the training data). The underlying concept examples included {*illegalism, modernity, antireligion, henosis, politics, lawlessness, ecosophy, conscience, ideology, enabling*} – words broadly perceived as abstract.

4.2 Abstractness rating

Table 1 presents a few examples of abstract and concrete concepts, as identified by manual annotation, along with their abstractness score as predicted by the RNN classifier (Section 3.3).

abstract		concrete	
concept	score	concept	score
marxism	0.972	plywood	0.000
islamophobia	0.969	Wiltshire	0.000
affirmative action	0.844	moonlight	0.058
absolute monarchy	0.842	convoy	0.112
sincerity	0.836	gadget	0.120

Table 1: Examples of concepts found as abstract/concrete (above/below the average score of 0.5) via manual annotation, along with their score as predicted by RNN.

Table 2 presents the Pearson correlation between the abstractness scores as assigned by the classifiers and the manual annotations over the evaluation set. We also present the correlation of scores produced by our classifiers to the set of Wikipedia concepts from the manually annotated MRC database (MRC-seed, Section 1), and to the set of 5883 noun concepts[8] from manually annotated BWK dataset (Brysbaert et al., 2014).

Evidently, the best results are obtained by the RNN classifier, yielding up to 0.740 correlation

---

[5]The radius was tuned on the set of 500 unigrams.

[6]We used the set of 18% highest ranked concepts – the fraction of abstract concepts in a sample population, as estimated by manual labeling.

[7]In all cases the obtained p-value was practically zero.

[8]Only concepts that can be mapped to a corresponding Wikipedia page were considered.

test set	Naive Bayes	NN-RAD	RNN
BWK	**0.657**	0.622	0.634
MRC-seed	**0.674**	0.576	0.669
1-grams	0.679	0.638	**0.740**
2-grams	0.565	0.515	**0.666**
3-grams	0.412	0.467	**0.490**

Table 2: Correlation of abstractness scores assigned by the classifiers to manual annotations.

with human annotations. Notably, the simple Naive Bayes, utilizing only textual data, yields results of reasonable quality; the broad implication of this outcome lies in the potential applicability of this approach to resource-scarce scenarios where high quality word embeddings are not available. Interestingly, while using Google word2vec embeddings (instead of Glove) yielded similar results, utilizing fastText pre-trained representations (Joulin et al., 2016) obtained more accurate ranking, e.g., the NN-RAD classifier yielded correlation of 0.688 for the BWK dataset, compared to 0.622 obtained using Glove (Table 2). We attribute this improvement to the fact that fastText embeddings better capture morphological word properties and cover more extensive vocabulary.

The relatively low correlation obtained with trigram concepts can be explained by the inherent complexity introduced by the multi-word scenario, challenging still further the subjective human perception of abstractness. While inter-labeler agreement for unigrams and bigrams was 0.72 and 0.66, respectively, it only reached 0.54 for trigrams, supporting the aforementioned hypothesis.

4.3 Varying the size of a test set

How many sentences containing a test concept suffice for a reliable prediction? We address this question by limiting the number of (randomly chosen) sentences used for rating. While the correlation obtained by RNN with 500 sentences containing a test concept reached 0.740 (Table 2), as little as 10, and even 5 sentences yielded correlation of 0.706 and 0.675, respectively, implying the efficiency and effectiveness of the presented approach in the availability of only little data. The plot in Figure 1 presents the correlation of the RNN and NB classifiers to label as function of number of (randomly sampled) sentences used for evaluation. Each such experiment (e.g., using 1, 5, 10 sentences) was averaged over 50 runs; the average correlation to label, as well as standard devi-

ation, are plotted on the chart. The constant correlation yield by the (text-independent) NN-RAD algorithm is illustrated by the vertical line.
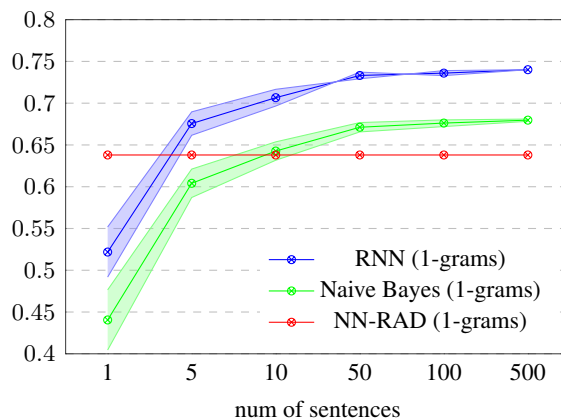

Figure 1: Average correlation (and standard deviation) to manual annotation as function of number of sentences used for evaluation.

4.4 Comparison to supervised models

Tsvetkov et al. (2013) used supervised learning algorithm to propagate abstractness scores to words using pre-trained word representations. Utilizing vector elements as features, they trained a supervised classifier, and predicted the degree of abstractness for unseen words. Abstractness rankings from the MRC database were used as a training set, and the classifier predictions were binarized into abstract-concrete boolean indicators using predefined thresholds. The authors obtained 94% accuracy when tested on held-out data.

5 Conclusions

We presented a weakly supervised approach for inferring the degree of concept abstractness. Our results demonstrate that a minimal morphological signal and a textual corpus are sufficient to train classifiers that yield relatively accurate predictions, that in turn can be used to unravel additional linguistic patterns indicative of the same property. Our future plans include exploring the value of the proposed methodology with other languages and additional properties.

Acknowledgments

References

Horacio A Barber, Leun J Otten, Stavroula-Thaleia Kousta, and Gabriella Vigliocco. 2013. Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and language*, 125(1):47–53.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

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

Louise Connell and Dermot Lynott. 2012. Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, 125(3):452–465.

Shi Feng, Zhiqiang Cai, Scott A Crossley, and Danielle S McNamara. 2011. Simulating human ratings on word concreteness. In *Proceedings of the FLAIRS Conference*.

Jennifer L Gianico-Relyea and Jeanette Altarriba. 2012. Word concreteness as a moderator of the tip-of-the-tongue effect. *The Psychological Record*, 62(4):763–776.

Felix Hill and Anna Korhonen. 2014. Concreteness and subjectivity as dimensions of lexical meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 725–731.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The cambridge grammar of English*. Cambridge University Press.

Sujay Kumar Jauhar and Lucia Specia. 2012. UOW-SHEF: Simplex–lexical simplicity ranking based on contextual and psycholinguistic features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 477–481. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in neural information processing systems*, pages 3111–3119.

Ryoji Nishiyama. 2013. Dissociative contributions of semantic and lexical-phonological information to immediate recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2):642–648.

J Oliveira, MV Perea, V Ladera, and P Gamito. 2013. The roles of word concreteness and cognitive load on interhemispheric processes of recognition. *Laterality: Asymmetries of Body, Brain and Cognition*, 18(2):203–215.

Gustavo Paetzold and Lucia Specia. 2016a. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679.

Gustavo Paetzold and Lucia Specia. 2016b. Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 435–440.

A Paivio. 2013. Dual coding theory, word abstractness, and emotion: a critical review of kousta et al.(2011). *Journal of experimental psychology.*, 142(1):282.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1.

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*, pages 1532–1543.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Leandro B. dos Santos, Magali Sanches Duran, Nathan Siegle Hartmann, Arnaldo Candido, Gustavo Henrique Paetzold, and Sandra Maria Aluisio. 2017. A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In *International Conference on Text, Speech, and Dialogue*, pages 281–289. Springer.

Otfried Spreen and Rudolph W Schulz. 1966. Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior*, 5(5):459–468.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690. Association for Computational Linguistics.