Metaphor recognition via concreteness/abstractness

Anne-Kathrin Bugert

Figurative Language Resolution
Institut für Computerlinguistik
Ruprecht-Karls-Universität Heidelberg

May 29, 2019

Table of contents

Turney et al. 2011

Motivation

Motivation

Lakoff and Johnson 1980

metaphor is a method for transferring knowledge from a concrete domain to an abstract domain

Motivation

Lakoff and Johnson 1980

metaphor is a method for transferring knowledge from a concrete domain to an abstract domain

$\rightarrow$ Hypothesis: degree of abstractness in a word's context is correlated with the likelihood that the word is used metaphorically

Example

- L: He *shot down* my plane.
  - $\rightarrow$ C_1: He *fired at* my plane.
  - $\nrightarrow$ A_1: He *refuted* my plane.
- M: He *shot down* my argument.
  - $\nrightarrow$ C_2: He *fired at* my argument.
  - $\rightarrow$ A_2: He *refuted* my argument.

Abstractness and Concreteness

Abstractness and Concreteness

- concrete words refer to things, events, and properties that we can perceive directly with our senses (trees, walking, red)
- abstract words refer to ideas and concepts that are distant from immediate perception (economics, calculating, disputable)

Abstractness Score

$$A(word) = \sum_{aword \in Awords} sim(word, aword) - \sum_{cword \in Cwords} sim(word, cword)$$

- abstractness of a given word: sum of similarity with twenty abstract paradigm words minus sum of similarity with twenty concrete paradigm words
- linear normalization to map the calculated abstractness value to range from 0 (highly concrete) to 1 (highly abstract)

Semantic Similarity

- corpus: $5x10^{10}$ words (280 GB of plain text) from university websites
- vocabulary: terms (words and phrases) of the WordNet lexicon with a frequency of 100 or more in the corpus (114,501 terms)

Semantic Similarity

- corpus: $5x10^{10}$ words (280 GB of plain text) from university websites
- vocabulary: terms (words and phrases) of the WordNet lexicon with a frequency of 100 or more in the corpus (114,501 terms)

- search up to 10,000 phrases per term (phrase: the given term plus four words to the left and four words to the right)

$\rightarrow$ word-context frequency matrix F with 114,501 rows and 139,246 columns
  - rows: terms in WordNet
  - columns: unigrams in WordNet with a frequency of 100 or more in the corpus
    - unigram represented by two columns, one marked left and one marked right

Semantic Similarity

- new matrix X with PPMI
- smoothed with a truncated Singular Value Decomposition (SVD)
- $X = U_k \Sigma_k V_k^t$

Semantic Similarity

- new matrix X with PPMI
- smoothed with a truncated Singular Value Decomposition (SVD)
- $X = U_k \Sigma_k V_k^t$
  - parameter k controlls the number of latent factors
  - parameter p adjust the weights of the factors

  $\rightarrow$ latent meaning

  $\rightarrow$ noise reduction

  $\rightarrow$ sparsity reduction
- terms represented by matrix $U_k \Sigma_k^p$ which has 114,501 rows (one for each term) and k columns (one for each latent contextual factor)
- semantic similarity of two terms is given by the cosine of the two corresponding rows in $U_k \Sigma_k^p$

Paradigm Words – Data

MRC Psycholinguistic Database Machine Usable Dictionary

- includes 4,295 words rated with degrees of abstractness by humans
- ratings range from 158 (highly abstract) to 670 (highly concrete)
- half of the words to train and other half to validate the algorithm

Paradigm Words – Data

MRC Psycholinguistic Database Machine Usable Dictionary

- includes 4,295 words rated with degrees of abstractness by humans
- ratings range from 158 (highly abstract) to 670 (highly concrete)
- half of the words to train and other half to validate the algorithm

Abstract Words	Rating	Concrete Words	Rating
as	158	ape	654
of	180	grasshopper	660
apt	183	tomato	662
however	186	milk	670

Table: Examples of abstract and concrete words from the MRC Dictionary

Paradigm Words

- empty set of paradigm words
- add one word at time, alternating between adding a word to the concrete paradigm words and the abstract paradigm words
- add the paradigm word that resulted in the highest Pearson correlation with the ratings of the training words
- stop after forty paradigm words (to prevent overfitting)
  - Pearson correlation training set: 0.8600
  - Pearson correlation testing set: 0.8064

Paradigm Words – Validation

- binary classification task from testing data
- median of ratings of the 2,147 words
- words with an abstractness above the median assigned to class 1, words below the median to class 0
- algorithm to guess the rating of each word in the test set, calculated median guess, likewise assigned to classes 0 and 1
- guesses were 84.65% accurate

Paradigm Words

Concrete Paradigm Words			Abstract Paradigm Words		
Order	Word	Correlation	Order	Word	Correlation
1	donut	0.4447	2	sense	0.6165
3	antlers	0.6582	4	indulgent	0.6973
5	aquarium	0.7150	6	bedevil	0.7383
7	nursemaid	0.7476	8	improbable	0.7590
9	pyrethrum	0.7658	10	purvey	0.7762
11	swallowwort	0.7815	12	pigheadedness	0.7884
13	strongbox	0.7920	14	ranging	0.7973
15	sixth-former	0.8009	16	quietus	0.8067
17	restharrow	0.8089	18	regularisation	0.8123
19	recorder	0.8148	20	creditably	0.8188

Table: Half of the forty paradigm words and the Pearson correlation on the training set.

Abstractness Ratings

- assign abstractness ratings to every term in the matrix
- 114,501 ratings would have a Pearson correlation of 0.81 with human ratings and an accuracy of 85% on binary (abstract or concrete) classification

Experiments

Experiments

- abstractness ratings to generate features for supervised machine learning
- learning algorithm: logistic regression as implemented in Weka
  - parameter settings:
    - $R = 0.2$ (for robust ridge regression)
    - $M = -1$ (for unlimited iterations)

First experiment: Adjectives

- 100 adjective-noun phrases labeled denotative (literal) or connotative (metaphorical or nonliteral) by five annotators, according to the sense of the adjective
  - deep snow → denotative
  - deep appreciation → connotative
- use abstractness rating of the noun (context) to predict whether the adjective (the target) was used in a metaphorical or literal sense
- algorithm predict labels with average accuracy of 79%

First experiment: Adjectives

- five adjectives: dark, deep, hard, sweet, warm
- for each: twenty word pairs in which the first word is the adjective and the second is a noun
  - Corpus of Contemporary American English (COCA)
  - find nouns that follow each adjective in the corpus and sort adjective-noun pairs by frequency
  - minimum PMI of 3 between adjective and noun

Adjective-Noun Pairs	Noun Abstractness
dark glasses	0.26826
dark chocolate	0.28211
dark energy	0.66297
dark mood	0.61858

Table: Some examples of adjective-noun pairs and the abstractness rating of the noun

First experiment: Adjectives

- five annotators: judge whether the use of the adjective is a denotation or a connotation
- "Denotation is the most direct or specific meaning of a word or expression while connotation is the meaning suggested by the word that goes beyond its literal meaning."
- Interjudge reliability: Cronbach's Alpha equal to 0.95

First experiment: Adjectives

- logistic regression with ten-fold cross-validation to predict each judge's denotative and connotative labels
- feature: abstractness rating of the noun
- algorithm predicts labels with average accuracy of 79%

Judge	Accuracy	Majority
1	0.730	0.590
2	0.810	0.570
3	0.840	0.560
4	0.790	0.510
5	0.780	0.520
Average	0.790	0.550

Table: The accuracy of logistic regression at predicting the labels of each judge

→ supports hypothesis that the abstractness of the context is predictive of whether an adjective is used in a literal or metaphorical sense

Second Experiment: Known Verbs

- TroFi (Trope Finder) Example Base of literal and nonliteral usage
- 50 verbs in 3 737 labeled sentences
- in each sentence target verb is labeled L (literal) or N (nonliteral)
- nonliteral includes metaphorical as a special case
    - Other types of nonliteral usage include idiomatic and metonymical, most of the nonliteral cases in TroFi are metaphorical

Example

L: An Energy Department spokesman says the sulfur dioxide might be simultaneously recoverable through the use of powdered limestone, which tends to *absorb* the sulfur.

N: He said that MMWEC will have to *absorb* only $4 million in additional annual costs now paid by the Vermont utilities.

Second Experiment: Known Verbs

- duplicate the setup of Birke and Sarkar 2006

$\rightarrow$ learn separate model for each individual verb

- average f-score of 63.9%, comparable to 64.9% by Birke and Sarkar 2006

Second Experiment: Known Verbs

- same subset as Birke and Sarkar 2006: 25 verbs in 1,965 sentences, manually labeled
- create a vector with five features for each sentence:
  1. the average abstractness ratings of all nouns, excluding proper nouns
  2. the average abstractness ratings of all proper nouns
  3. the average abstractness ratings of all verbs, excluding the target verb
  4. the average abstractness ratings of all adjectives
  5. the average abstractness ratings of all adverbs
- set the average to a default value of 0.5 when there were no words for a given part of speech

Example

L: An Energy Department spokesman says the sulfur dioxide might be simultaneously recoverable through the use of powdered limestone, which tends to *absorb* the sulfur.

L: $< 0.3873, 0.5397, 0.6375, 0.2641, 0.5835 >$

N: He said that MMWEC will have to *absorb* only $4 million in additional annual costs now paid by the Vermont utilities.

N: $< 0.6120, 0.3726, 0.6699, 0.5612, 0.5000 >$

Second Experiment: Known Verbs

- weight of each context word may depend on the part of speech of the context
- logistic regression algorithm determines the appropriate weighting, based on the training data

Second Experiment: Known Verbs

- weight of each context word may depend on the part of speech of the context
- logistic regression algorithm determines the appropriate weighting, based on the training data

- separate model learned for each individual verb
- ten-fold cross-validation for each verb to learn and test logistic regression models

Birke and Sarkar 2006 scorings

- *Literal recall = correct literals in literal cluster / total correct literals*
  - 100% if there are no literals
- *Literal precision = correct literals in literal cluster / size of literal cluster*
  - 100% if there are no nonliterals in the literal cluster and 0% otherwise
- *f-score = (2 · precision · recall) / (precision + recall)*
- nonliteral precision and recall are defined similarly
- average precision is the average of literal and nonliteral precision; similarly for average recall
- overall performance: f-score of average precision and average recall
- Turney et al. 2011 modified f-score (0/0=0): precision of a class is 0% if the algorithm never guesses that class

Second Experiment: Known Verbs

Algorithm	Accuracy	F-score (0/0=0)	F-score (0/0=1)
Concrete-Abstract	0.734	0.631	0.639
Birke-Sarkar	*NA*	*NA*	0.649
Majority Class	**0.697**	**0.408**	0.629
Probability Matching	**0.605**	**0.500**	**0.500**

Table: The performance with known verbs.

- statistical significance (paired t-test): bold font when the performance is significantly below the performance of Concrete-Abstract

Third Experiment: Unknown Verbs

- TroFi Example Base
  - "new" verbs for training (appear in 1,772 sentences)
  - "old" verbs for testing (appear in 1,965 sentences)
- all training sentences used together to learn a single logistic regression model

Algorithm	Accuracy	F-score (0/0=0)	F-score (0/0=1)
Concrete-Abstract	0.686	0.673	0.681
Birke-Sakar	*NA*	*NA*	0.649
Majority Class	**0.697**	**0.408**	**0.629**
Probability Matching	**0.605**	**0.500**	**0.500**

Table: The performance with unknown verbs.

logistic regression coefficients

	Feature	Coefficient
1	AvgNounAbs	11.4117
2	AvgProbAbs	0.7250
3	AvgVerbAbs	-0.5528
4	AvgAdjAbs	1.1478
5	AvgAdvAbs	-0.2013
6	Intercept	-5.9436

Table: The logistic regression coefficients for class N.

- 1 to 5 are the five features
- 6 is the constant term in the regression equation
- abstractness of nouns (excluding proper nouns) has largest weight in predicting whether the target is in class N

Conclusion

Conclusion

- algorithm for the degree of abstractness of a word
  - corpus?
  - paradigm words?

Conclusion

- algorithm for the degree of abstractness of a word
  - corpus?
  - paradigm words?

- abstractness of the context is predictive of whether an adjective is used in a literal or metaphorical sense
  - only for concrete target words?

Questions?

Köper and Schulte im Walde 2017

Contribution

Contribution

- compare supervised techniques to learn and extend abstractness ratings for huge vocabularies
- learn and investigate norms for multi-word units by propagating abstractness to verb-noun pairs
- show that multisense abstractness ratings are potentially useful for metaphor detection
- *publish automatically created abstractness norms for 3 million English words and multi-words as well as automatically created sense-specific abstractness ratings*

Comparison of Approaches & Ressources

Comparison of Approaches

- Approaches:
  - Turney et al. 2011: requires vector representation and annotated training samples of words
  - distributional vectors implicitly encode attributes such as abstractness

  $\rightarrow$ directly feed the vector representation of a word into a classifier
    - linear regression (L-Reg)
    - regression forest (Reg-F)
    - a fully connected feed forward neural network with up to two hidden layers (NN)

Comparison of Approaches

- Vector representations:
  - compare vectors between 50 and 300 dimensions
  - Glove vectors (Pennington et al. 2014)
    - trained on 6billion tokens of Wikipedia plus Gigaword (V=400K)
  - word2vec cbow model (Mikolov et al. 2013)
    - trained on a Google internal news corpus with 100billion tokens (V=3million)

Comparison of Approaches

- ratings from Brysbaert et al. 2014 for training and testing
  - 20% test (7990) and 80% training (31 964), 1 000 ratings from training data for hyper parameter tuning
- evaluation: comparing new created ratings against test (gold) ratings using Spearman's rank-order correlation

Comparison of Approaches

	T&L 03	L-Reg.	Reg-F.	NN
Glove50	.76	.76	.78	**.79**
Glove100	.80	.79	.79	**.85**
Glove200	.78	.78	.76	**.84**
Glove300	.76	.78	.74	**.85**
W2V300	.83	.84	.79	**.90**

Table: Spearman's ρ for the test ratings. Comparing representations and regression methods.

Comparison of Ressources

- abstractness ratings for the entire vocabulary of W2V300 dataset
- compare the correlation with other existing norms of abstractness
  - MRC Psycholinguistic Database
  - ratings from Brysbaert et al. 2014
  - automatically created ratings from Turney et al. 2011
- map ratings to an interval ranging from very abstract (0) to very concrete (10)
- common subset contains 3 665 ratings

Comparison of Ressources

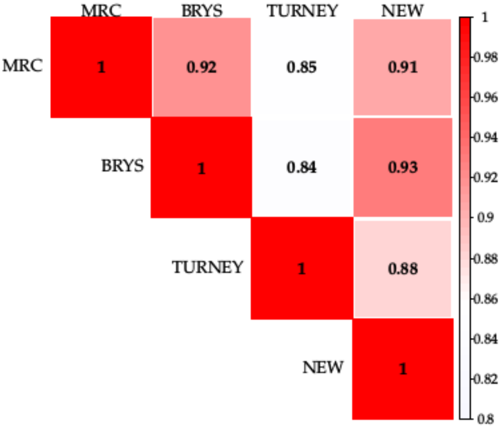

Figure: Pairwise Spearman's ρ on commonly covered subset. Red = high correlation

Abstractness for Phrases

Abstractness for Phrases

- dataset: collection from Mohammad et al. 2016, who annotated different senses of WordNet verbs for metaphoricity
- same subset of verb-direct object and verb-subject relations as used in Shutova et al. 2016
- web corpus ENCOW14
  - remove words and phrases that appeare less than 50 times in the corpus
  - selection covers 535 pairs, 238 metaphorical and 297 literal

Abstractness for Phrases

- vector representations for a verb-noun phrase using word2vec and the same hyper-parameters used for the W2V300 embeddings together with the best learning method (NN)
- abstractness ratings for all three constituents: verb, noun and the entire phrase

- rating score and the Area Under Curve (AUC) metric
- also results based on cosine similarity and feature combinations

Abstractness for Phrases

Feat.	Name	Type	AUC
-	Random	baseline	.50
1	V-NN	cosine	.75
2	V-Phrase	cosine	.70
3	NN-Phrase	cosine	.68
4	V	rating	.53
5	NN	rating	**.78**
6	Phrase	rating	.71
Comb	1+2+3	cosine	.75
Comb	4+5+6	rating	.74
Comb	all(1-6)	mixed	.80
Comb	1+5+6	best	**.84**

Table: AUC Score single features and combinations. Classifying literal and metaphorical phrases based on Mohammad et al. 2016 dataset.

Sense-specific Abstractness Ratings

Sense-specific Abstractness Ratings

- automatically learned multi-sense abstractness ratings
- different vector representation per word sense
- Pelevina et al. 2016 performs sense learning after single senses have been learned

Sense-specific Abstractness Ratings

- apply multi-sense learning technique to W2V300 with default settings
- propagate abstractness to every newly created sense representation
- disambiguate the word sense by comparing the sense-specific vector representation to all context words

Sense-specific Abstractness Ratings

- VU Amsterdam Metaphor Corpus
  - 23 113 verb tokens in running text, annotated as literally or metaphorically
- TroFi metaphor dataset
  - 50 verbs and 3 737 labeled sentences
- ten-fold cross-validation over the entire data
- For the VUA aditionally results using the same training/test split as in Beigman Klebanov et al. 2016

Sense-specific Abstractness Ratings

- five feature dimensions (Turney et al. 2011) plus dimensions for subject and object:
  1. Rating of the verbs subject
  2. Rating of the verbs object
  3. Average rating of all nouns (excluding proper names)
  4. Average rating of all proper names
  5. Average rating of all verbs, excluding the target verb
  6. Average rating of all adjectives
  7. Average rating of all adverbs
- balanced Logistic Regression classifier (Beigman Klebanov et al. 2016)

Sense-specific Abstractness Ratings

Feat.	TroFi(10F)	VUA(10F)	VUA(Test)
1S	.72	.42	.44
MS	**.74**	**.44***	**.46**
1S(+L)	.74	**.61**	**.62**
MS(+L)	**.75**	**.61**	**.62**

Table: F-score (Metaphor). Classifying literal and metaphorical verbs based on the VUA and TroFi dataset. MS = multisense, 1S= single sense.

- lemma of the target verb (+L) to describe performance with respect to the state of the art (Beigman Klebanov et al. 2016)
- difference in performance of single and multi-sense ratings is statistically significant on the full VUA dataset, using the χ^2 test and $*$ for $p < 0.05$

Conclusion

Conclusion

- compare methods to propagate abstractness norms
- norms for multi-words phrases
- sense specific norms useful for metaphor detection

Questions?

Literature I

Beata Beigman Klebanov et al. “Semantic classifications for detection of verb metaphors”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2016, pp. 101–106.

Julia Birke and Anoop Sarkar. “A clustering approach for nearly unsupervised recognition of nonliteral language”. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. “Concreteness ratings for 40 thousand generally known English word lemmas”. In: *Behavior research methods* 46.3 (2014), pp. 904–911.

Literature II

Maximilian Köper and Sabine Schulte im Walde. "Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses". In: *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*. 2017, pp. 24–30.

George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 1980.

Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. "Metaphor as a medium for emotion: An empirical study". In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 2016, pp. 23–33.

Maria Pelevina et al. "Making Sense of Word Embeddings". In: *ACL 2016* (2016), p. 174.

Literature III

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

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. "Black holes and white rabbits: Metaphor identification with visual features". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016, pp. 160–170.

Peter D Turney et al. "Literal and metaphorical sense identification through concrete and abstract context". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, pp. 680–690.