Evaluating word vectors

VL Embeddings

Uni Heidelberg

SS 2019

How to evaluate word vectors?

- **Intrinsic** vs. **extrinsic** evaluation
- Intrinsic
  - evaluation on a dataset created for a specific task
    e.g.: word similarity (semantic, syntactic), word analogy, ...
  - easy to compare your model to other models
  - fast to compute
  - useful for understanding which parameters matter

How to evaluate word vectors?

- **Intrinsic** vs. **extrinsic** evaluation

- Intrinsic
  - evaluation on a dataset created for a specific task
    e.g.: word similarity (semantic, syntactic), word analogy, ...
  - easy to compare your model to other models
  - fast to compute
  - useful for understanding which parameters matter
  - **not clear how meaningful for real-world tasks**

How to evaluate word vectors?

- **Intrinsic** vs. **extrinsic** evaluation

- Intrinsic
  - evaluation on a dataset created for a specific task
    e.g.: word similarity (semantic, syntactic), word analogy, ...
  - easy to compare your model to other models
  - fast to compute
  - useful for understanding which parameters matter
  - **not clear how meaningful for real-world tasks**

- Extrinsic
  - evalation on real-world task → more meaningful
  - might take a long time
  - harder to compare to other models/systems
    (harder to isolate the effect of the embeddings)

  → keep system fixed, plug in different embedding types

Intrinsic word vector evaluation

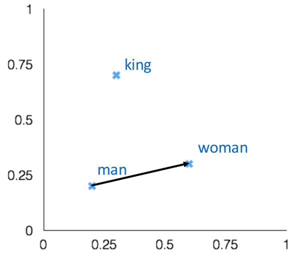
Word vector analogies

A is to B what C is to ?
e.g. man is to women what king is to ?

$$d = argmax_i \frac{(x_b - x_a + x_c)^T x_i}{||x_b - x_a + x_c||}$$

Evaluate word vectors by how well they capture intuitive semantic and syntactic analogies:

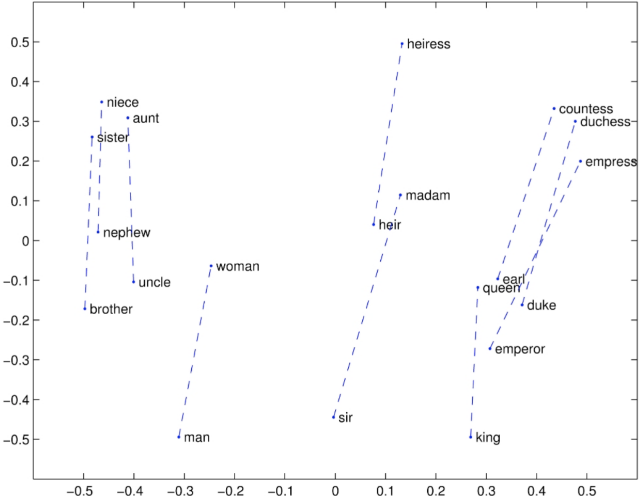
- substract man from woman and add king
- find vector with highest cosine similarity to A - B + C



From R. Socher's slides for CS224d (2016) https://cs224d.stanford.edu/lectures/CS224d-Lecture3.pdf

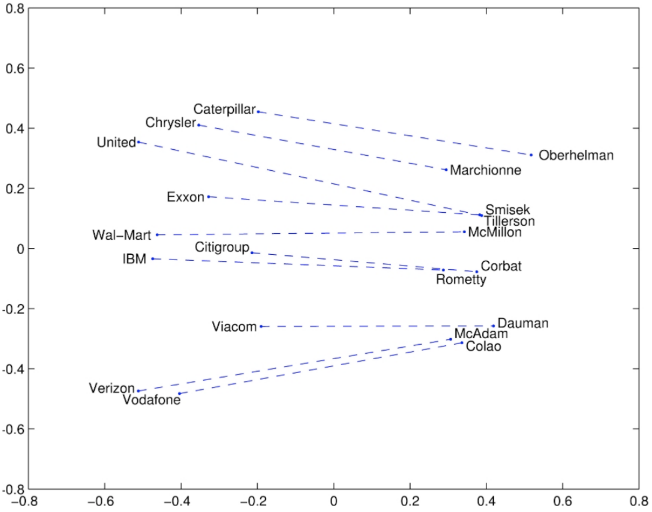
Intrinsic word vector evaluation

Word analogies (GloVe) – Examples



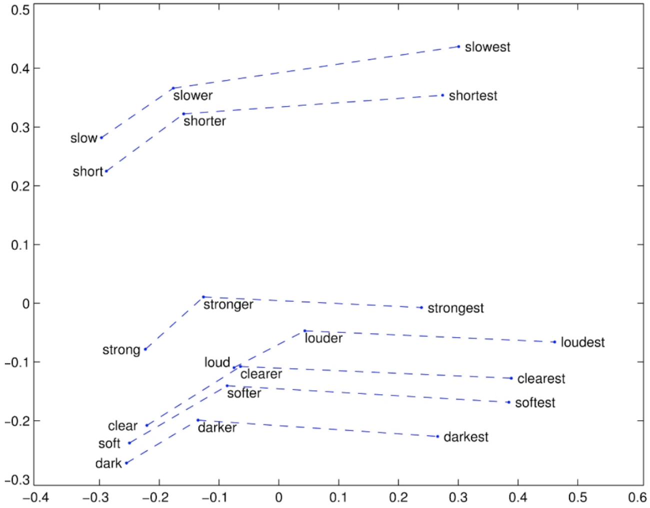
Intrinsic word vector evaluation

Word analogies (GloVe) – Examples



Intrinsic word vector evaluation

Word analogies (GloVe) – Examples



Datasets for intrinsic word vector evaluation

Word vector analogies: Syntactic and semantic examples from
`http://download.tensorflow.org/data/questions-words.txt`
(Mikolov et al. 2013)

city-in-state
Chicago Illinois Houston Texas
Chicago Illinois Philadelphia Pennsylvania
Chicago Illinois Dallas Texas
Chicago Illinois Detroit Michigan
Chicago Illinois Boston Massachusetts
...

From R. Socher's slides for CS224d (2016) https://cs224d.stanford.edu/lectures/CS224d-Lecture3.pdf

Datasets for intrinsic word vector evaluation

Word vector analogies: Syntactic and semantic examples from
`http://download.tensorflow.org/data/questions-words.txt`
(Mikolov et al. 2013)

capital-world
Abuja Nigeria Accra Ghana
Abuja Nigeria Algiers Algeria
Abuja Nigeria Ankara Turkey
Abuja Nigeria Apia Samoa
Abuja Nigeria Asmara Eritrea
...

From R. Socher's slides for CS224d (2016) https://cs224d.stanford.edu/lectures/CS224d-Lecture3.pdf

Datasets for intrinsic word vector evaluation

Word vector analogies: Syntactic and semantic examples from
`http://download.tensorflow.org/data/questions-words.txt`
(Mikolov et al. 2013)

gram4-superlative
bad worst big biggest
bad worst cold coldest
bad worst cool coolest
bad worst fast fastest
bad worst good best
...

From R. Socher's slides for CS224d (2016) `https://cs224d.stanford.edu/lectures/CS224d-Lecture3.pdf`

Impact of dimension size on analogy task

Compare different word embedding models and hyperparameters for analogy task

- Do more dimensions help?
- How important is corpus size?
- How important is the domain/genre of your corpus?
- Which model is better for capturing syntax/semantics?

Impact of dimension size on analogy task

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

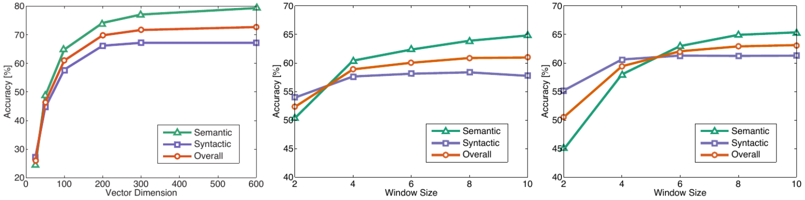
Percentage accuracy on analogy dataset.

(i)vLBL: Mnih et al. (2013); SG/CBOW: Mikolov et al. (2013);

HPCA: Hellinger PCA (Lebret and Collobert 2014); SVD-S: $\sqrt{M}$; SVD-L: $log(1+M)$)

Impact of context window size on analogy task

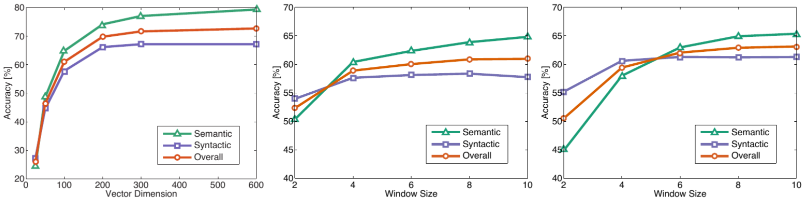
- Evaluate window size for symmetric vs. asymmetric contexts



(a) Symmetric context (b) Symmetric context (c) Asymmetric context

Impact of context window size on analogy task

- Evaluate window size for symmetric vs. asymmetric contexts

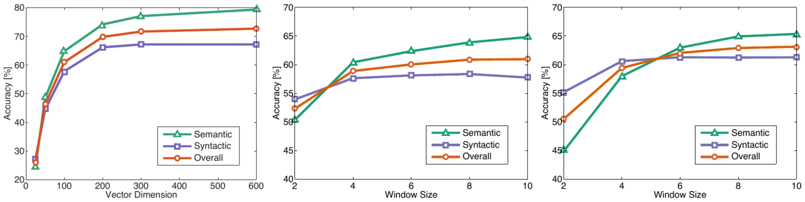


(a) Symmetric context (b) Symmetric context (c) Asymmetric context

- Asymmetric contexts: left context only
- Best dimension size: ≈ 300
- Best window size: 8
- But results might be different for downstream tasks (and also for other languages)

Impact of context window size on analogy task

- Evaluate window size for symmetric vs. asymmetric contexts



(a) Symmetric context (b) Symmetric context (c) Asymmetric context

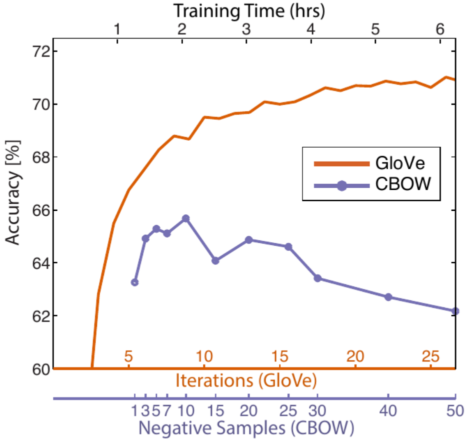
- Asymmetric contexts: left context only
- Best dimension size: $\approx$ 300
- Best window size: 8
- But results might be different for downstream tasks (and also for other languages)

Parameter choice: trade-off between accuracy and efficiency

From R. Socher's slides for CS224d (2016) https://cs224d.stanford.edu/lectures/CS224d-Lecture3.pdf

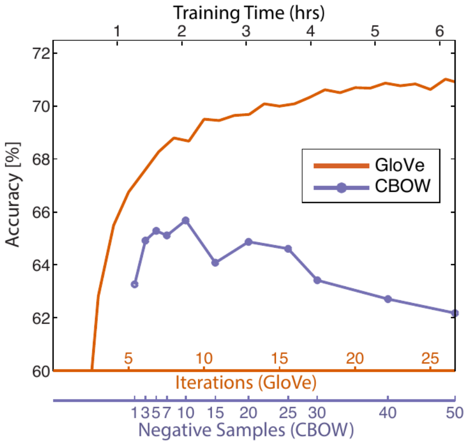
Training time for different embeddings

- Direct comparison: CBOW and GloVe



Training time for different embeddings

- Direct comparison: CBOW and GloVe

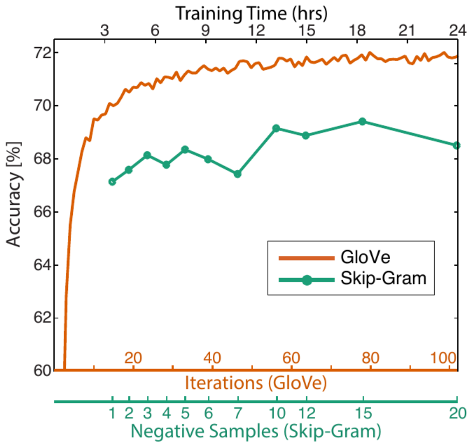


- But: CBOW trained for only 1 iteration – fair comparison?

From R. Socher's slides for CS224d (2016) https://cs224d.stanford.edu/lectures/CS224d-Lecture3.pdf

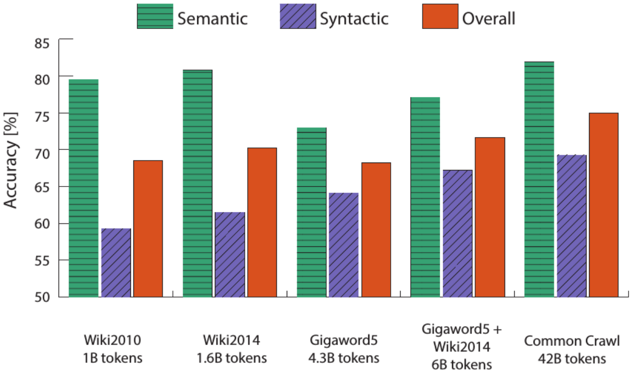
Training time for different embeddings

- Direct comparison: Skip-Gram and GloVe



Impact of data size and domain on GloVe

- More data is better
- Wikipedia better than news (for analogy dataset)



Datasets for word similarity evaluation

- Word similarity: Correlation between cosine similarity (or other distance measure) and human judgments
- **WordSim353** (word similarity and relatedness) (http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/)

Word 1	Word 2	Human (mean)
tiger	cat	7.35
tiger	tiger	10.00
book	paper	7.46
computer	internet	7.58
plane	car	5.77
professor	doctor	6.62
stock	phone	1.62
stock	CD	1.31
stock	jaguar	0.92

Intrinsic evaluation based on word similarity

History

- Rubenstein and Goodenough (1965):
  - first word similarity task with 65 word pairs and judgments by human raters
- Goal: test distributional hypothesis (Harris, 1954)
  - R&G found positive correlation between contextual similarity and human-annotated similarity of word pairs

Datasets for word similarity evaluation

- WS353 (Mikolov et al. 2013): similar and related words
- RG (Rubenstein and Goodenough, 1965): 65 word pairs assessed by semantic similarity with a scale from 0 to 4
- MC (Miller and Charles, 1991): subset of RG containing 10 pairs with high similarity, 10 with middle similarity and 10 with low similarity
- SCWS (Huang et al., 2012) $\Rightarrow$ similarity ratings for different word senses
- RW (Luong et al., 2013) $\Rightarrow$ 2,034 pairs of rare words assessed by semantic similarity with a scale from 0 to 10

More datasets for word similarity evaluation

Name	Description
SimVerb-3500	3,500 pairs of verbs assessed by semantic similarity (that means that pairs that are related but not similar have a fairly low rating) with a scale from 0 to 4.
MEN (Marco, Elia and Nam)	3,000 pairs assessed by semantic relatedness with a discrete scale from 0 to 50.
RW (Rare Word)	2,034 pairs of words with low occurrences (rare words) assessed by semantic similarity with a scale from 0 to 10.
SimLex-999	999 pairs assessed with a strong respect to semantic similarity with a scale from 0 to 10.
SemEval-2017	500 pairs assessed by semantic similarity with a scale from 0 to 4 prepared for the SemEval-2017 Task 2. Contains words and collocations (climate change).
MTurk-771 WordSim-353	771 pairs assessed by semantic relatedness with a scale from 0 to 5. 353 pairs assessed by semantic similarity with a scale from 0 to 10.
MTurk-287	287 pairs assessed by semantic relatedness with a scale from 0 to 5.
WordSim-353-REL	252 pairs, a subset of WordSim-353 containing no pairs of similar concepts.
WordSim-353-SIM	203 pairs, a subset of WordSim-353 containing similar or unassociated (to mark all pairs that receive a low rating as unassociated) pairs.
Verb-143	143 pairs of verbs assessed by semantic similarity with a scale from 0 to 4.
YP-130 (Yang and Powers)	130 pairs of verbs assessed by semantic similarity with a scale from 0 to 4.
RG-65 (Rubenstein and Goodenough)	65 pairs assessed by semantic similarity with a scale from 0 to 4.
MC-30 (Miller and Charles)	30 pairs, a subset of RG-65 which contains 10 pairs with high similarity, 10 with middle similarity and 10 with low similarity.

https://github.com/vecto-ai/word-benchmarks

Evaluation of different embeddings on word similarity task

- Spearman rank correlation with human judgments

Model	Size	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	72.7	75.1	56.5	37.0
CBOW†	6B	57.2	65.6	68.2	57.0	32.5
SG†	6B	62.8	65.2	69.7	58.1	37.2
GloVe	6B	65.8	72.7	77.8	53.9	38.1
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	**75.9**	**83.6**	**82.9**	**59.6**	**47.8**
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

All vectors with dimension=300, CBOW* contains phrase vectors

Problems for intrinsic evaluation

Faruqui, Tsvetkov, Rastogi and Dyer (2016): Problems with Evaluation of Word Embeddings Using Word Similarity Tasks

- Word similarity as a proxy for word vector evaluation

  $\Rightarrow$ correlate the distance between vectors and human judgments of *semantic similarity*

- Advantages
  - fast and computationally efficient
- But: is it reliable?

Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

  Kaffee – Tee

Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

  Auto – Zug

Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

  Baum – Blume

Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

  Tasse – Kaffee

Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

  Tasse – Kaffee

- *Similarity* often confused with *relatedness*

  ⇒ *cup* and *coffee* are rated more similar than *car* and *train* in WordSim353

  - similar problems with other datasets,
    e.g. MEN (Bruni et al., 2012)

Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

  Tasse – Kaffee

- *Similarity* often confused with *relatedness*

  ⇒ *cup* and *coffee* are rated more similar than *car* and *train* in WordSim353

  - similar problems with other datasets, e.g. MEN (Bruni et al., 2012)

⇒ Word vectors that capture this difference get punished

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:

Dackel – Fernseher

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:

Dackel – Fernseher
Dackel – Karotte

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:

Dackel – Fernseher
Dackel – Karotte
Dackel – Siamkatze

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:

Dackel – Fernseher
Dackel – Karotte
Dackel – Siamkatze

Dackel – Pudel

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:

Dackel – Fernseher
Dackel – Karotte
Dackel – Siamkatze

Dackel – Pudel
Dackel – Terrier

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:

Dackel – Fernseher	Dackel – Pudel
Dackel – Karotte	Dackel – Terrier
Dackel – Siamkatze	Dackel – Siamkatze

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:

Dackel – Fernseher	Dackel – Pudel
Dackel – Karotte	Dackel – Terrier
Dackel – Siamkatze	Dackel – Siamkatze

Human judgments can vary, depending on context

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity dependent on word sense
- How similar are:

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity dependent on word sense
- How similar are:

Maus – Katze

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity dependent on word sense
- How similar are:

Maus – Katze
Maus – Keyboard

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity dependent on word sense
- How similar are:

  Maus – Katze
  Maus – Keyboard
  Katze – Keyboard

Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity dependent on word sense
- How similar are:

Maus – Katze
Maus – Keyboard
Katze – Keyboard

Only one vector per word but more than one word sense
⇒ Session on Multisense word embeddings (July 9)

Intrinsic evaluation based on word similarity

No standardised splits – overfitting

- Good practice for ML
  - Split data into train, dev, test set
  - Select best model on dev, evaluate on test $\rightarrow$ avoid overfitting!
- For word similarity tasks
  - no standard splits, vectors are optimised on the test sets $\rightarrow$ overfitting
- Datasets are often quite small
  - further splits might make results unreliable

Overfitting

Possible Solutions

- Use **one** dataset for tuning, evaluate on **all other** datasets (Faruqui and Dyer 2014)
- Use **all** available datasets for tuning (Lu et al. 2015)
  1. choose hyperparameters with **best average** performance across **all** tasks
  2. choose hyperparameters that beat the baseline vectors on **most** tasks
- Makes sure that model generalises well across different tasks

Intrinsic evaluation based on word similarity

Statistical significance

- Significance testing important especially for non-convex objectives whith multiple locally optimal solutions
- Rastogi et al. (2015) observed that improvements obtained by models on a small word similarity dataset were insignificant
- Compute statistical significance for word similarity evaluation (see Faruqui et al. 2016)

Intrinsic evaluation based on word similarity

Low correlation with extrinsic tasks

- Chiu, Korhonen & Pyysalo (2016):
  Intrinsic evaluation of word vectors fails to predict extrinsic performance
    - possible reason: failure to distinguish similarity from relatedness

Intrinsic evaluation based on word similarity

Low correlation with extrinsic tasks

- Chiu, Korhonen & Pyysalo (2016):
  Intrinsic evaluation of word vectors fails to predict extrinsic performance
  - possible reason: failure to distinguish similarity from relatedness

- Artetxe, Labaka, Lopez-Gazpio and Agirre (2018):
  Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation
  - intrinsic evaluation not a good predictor for performance in downstream applications

References

- Mikolov, Yih and Zweig: (2013): Linguistic regularities in continuous space word representations. NAACL 2013.
- Faruqui, Tsvetkov, Rastogi and Dyer (2016): Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. The 1st Workshop on Evaluating Vector Space Representations for NLP, Berlin, Germany.
- Artetxe, Labaka, Lopez-Gazpio and Agirre (2018): Uncovering Divergent Linguistic Information in Word Embeddings with Lessons for Intrinsic and Extrinsic Evaluation. CoNLL 2018. Brussels, Belgium.
- Rubenstein and Goodenough (1965): Contextual correlates of synonymy. Communications of the ACM 8(10):627–633.
- Harris, Z. (1954). Distributional structure. Word, 10(23): 146-162.
- Multimodal Distributional Semantics E. Bruni, N. K. Tran and M. Baroni. Journal of Artificial Intelligence Research 49: 1-47.
- Collobert, Weston Bottou, Karlen, Kavukcuoglu and Kuksa (2011): Natural Language Processing (almost) from Scratch. Journal of Machine Learning Research 12 (2011) 2461-2505.
- Lu, Wang, Bansal, Gimpel and Livescu (2015): Deep multilingual correlation for improved word embeddings. NAACL 2015.
- Rastogi, Van Durme and Arora (2015): Multiview LSA: Representation learning via generalized CCA. NAACL 2015.
- Chiu, Korhonen and Pyysalo (2016): Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance. ACL 2016.

- Data and Code
  - Code for Artetxe etal. (2018): `https://github.com/artetxem/uncovec`
  - The MEN dataset `https://staff.fnwi.uva.nl/e.bruni/MEN`
  - Datasets for word vector evaluation `https://github.com/vecto-ai/word-benchmarks`