

Evaluating Word Embeddings Using a Representative Suite of Practical Tasks

Neha Nayak, Gabor Angeli, Christopher Manning
Stanford University

August 12, 2016



WHAT DO WE WANT FROM AN EVALUATION?

- ▶ Predict performance in downstream tasks
- ▶ Word similarity doesn't give us this
- ▶ Solution : evaluate on downstream tasks

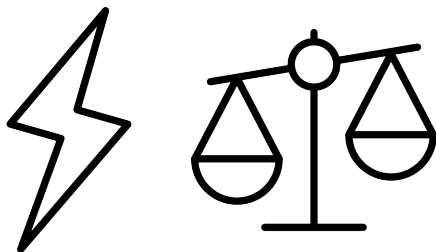
WHY DOES WORD SIMILARITY ENDURE AS AN EVALUATION METHOD?

WHY DOES WORD SIMILARITY ENDURE AS AN EVALUATION METHOD?



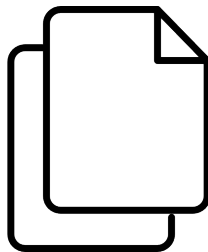
- ▶ Fast

WHY DOES WORD SIMILARITY ENDURE AS AN EVALUATION METHOD?



- ▶ Fast
- ▶ Unbiased

WHY DOES WORD SIMILARITY ENDURE AS AN EVALUATION METHOD?



- ▶ Fast
- ▶ Unbiased
- ▶ Replicable

PROPOSED EVALUATION



- ▶ A fast, unbiased, and replicable *extrinsic evaluation*
- ▶ Simple neural models
- ▶ For a suite of downstream tasks
- ▶ Chosen to test both syntactic and semantic properties

TASKS



	Syntactic	Semantic
Word level	Part of Speech tagging	Named Entity Recognition
Phrase level	Chunking	Phrase-level NLI
Sentence level		Sentiment, Question classification

TASKS

	Syntactic	Semantic
Word level	Part of Speech tagging	Named Entity Recognition
Phrase level	Chunking	Phrase-level NLI
Sentence level		Sentiment, Question classification

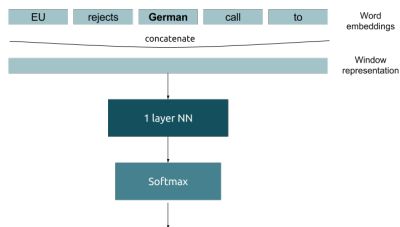
- ▶  Fast: limited tasks, small datasets
- ▶  Unbiased: standard datasets

MODELS

- ▶  Fast: simple models
- ▶  Unbiased: few hyperparameters

MODELS

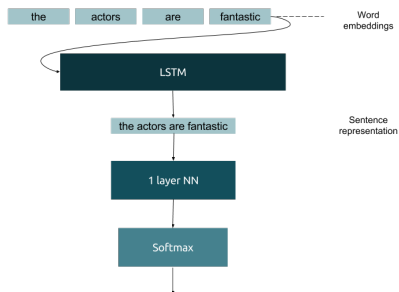
- ▶ Window model: POS Tagging, NER, Chunking



- ▶ ⚡ Fast: simple models
- ▶ ⚖️ Unbiased: few hyperparameters

MODELS

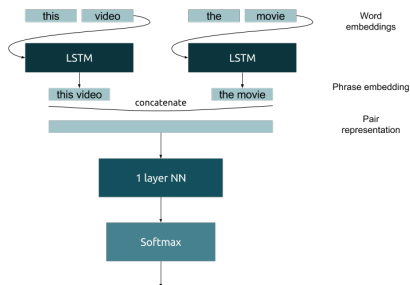
- ▶ Sentence model: Sentiment classification, Question classification



- ▶ ⚡ Fast: simple models
- ▶ ⚖️ Unbiased: few hyperparameters

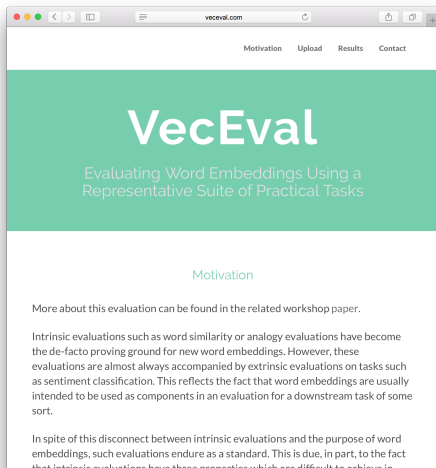
MODELS

- ▶ Phrase pair model: Phrase-level NLI

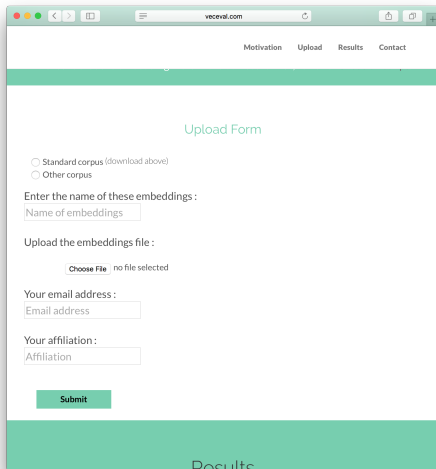


- ▶ ⚡ Fast: simple models
- ▶ ⚖️ Unbiased: few hyperparameters

A STANDARD, REPLICABLE EVALUATION



A STANDARD, REPLICABLE EVALUATION



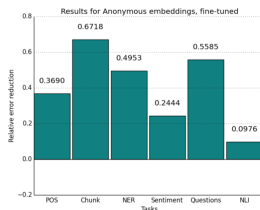
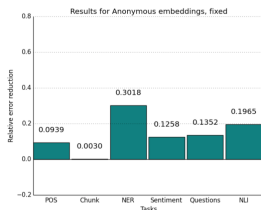
The screenshot shows a web browser window with the URL 'veceval.com'. The page has a navigation menu with links for 'Motivation', 'Upload', 'Results', and 'Contact'. The main content area is titled 'Upload Form' and contains the following elements:

- Two radio buttons for selecting the corpus: 'Standard corpus (download above)' and 'Other corpus'.
- A text input field labeled 'Enter the name of these embeddings:' with the placeholder text 'Name of embeddings'.
- A file upload section labeled 'Upload the embeddings file:' with a 'Choose File' button and the text 'no file selected'.
- An email input field labeled 'Your email address:' with the placeholder text 'Email address'.
- An affiliation input field labeled 'Your affiliation:' with the placeholder text 'Affiliation'.
- A green 'Submit' button.

At the bottom of the page, the word 'Results' is partially visible.

A STANDARD, REPLICABLE EVALUATION

Results for Anonymous embeddings

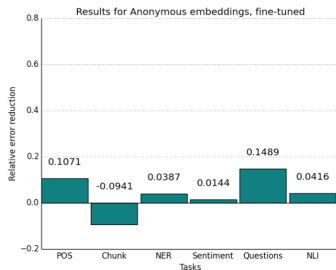
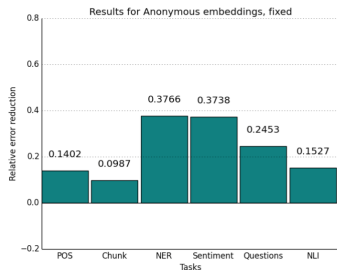


	POS (Acc.)	Chunk (Acc.)	NER (F1)	Sentimen (Acc.)	Questions (Acc.)	NLI (Acc.)
Fixed (Baseline)	86.30	82.30	94.40	65.10	72.50	45.00
Fine-tuned (Baseline)	82.80	77.60	94.20	69.20	84.00	46.40
	97.88	97.60	97.90	83.90	91.70	50.10
	96.90	92.80	96.10	76.70	84.40	43.90

Table 1: Raw results on downstream tasks

	WordSim	Analogy (Sem.)	Analogy (Syn.)
Anonymous vectors	0.640	52.0	63.0
Baseline	0.560	38.0	48.0

A STANDARD, REPLICABLE EVALUATION



- ▶ Both fixed and fine-tuned settings
- ▶ Relative error reduction

THANKS!

www.veceval.com