Word Embeddings

What works, what doesn't, and how to tell the difference for applied research

Pedro L. Rodríguez Arthur Spirling

Vanderbilt University New York University

May 14, 2020

We spent thousands of dollars running hundreds of embedding models and performing thousands of human validation tasks to get these main takeaways so you don't have to

Pedro L. Rodríguez Arthur Spirling

Vanderbilt University New York University

May 14, 2020

Big Picture(s)

Big Picture(s)

Which one of these portraits is more realistic?

WELFARE

dependency

reform

Select the best candidate context word for the cue word provided by clicking on the respective checkbox below the word.

Decisions, Decisions

Decisions, Decisions

Explosion of interest in word embeddings.

Decisions, Decisions

Explosion of interest in word embeddings. These are real valued vectors that are used for two purposes:

Decisions, Decisions

Explosion of interest in word embeddings. These are real valued vectors that are used for two purposes:

1. feature representations for downstream NLP/ML tasks.

Decisions, Decisions

Explosion of interest in word embeddings. These are real valued vectors that are used for two purposes:

1. feature representations for downstream NLP/ML tasks.

2. tools for studying word usage and meaning—"semantics".

Decisions, Decisions

Explosion of interest in word embeddings. These are real valued vectors that are used for two purposes:

1. feature representations for downstream NLP/ML tasks.

2. tools for studying word usage and meaning—"semantics".

As with all such representational strategies (topic models, TF-IDF), there are (literally thousands) of modeling options available. . .

Decisions, Decisions

Explosion of interest in word embeddings. These are real valued vectors that are used for two purposes:

1. feature representations for downstream NLP/ML tasks.

2. tools for studying word usage and meaning—"semantics".

As with all such representational strategies (topic models, TF-IDF), there are (literally thousands) of modeling options available...

How should political scientists choose among them?

Why do we ask?

How should political scientists choose among them?

Some Background

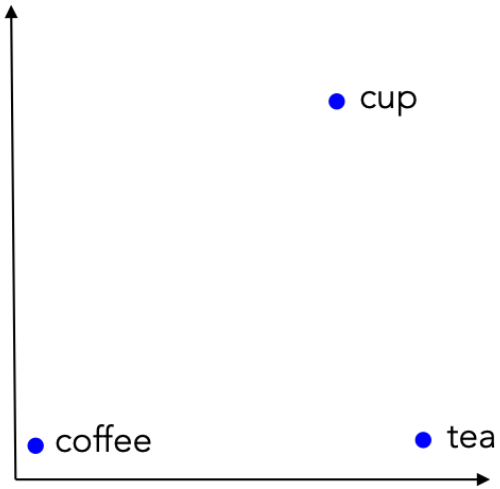
Firth quote, so you know we're legit

Firth quote, so you know we're legit

"You shall know a word by the company it keeps."

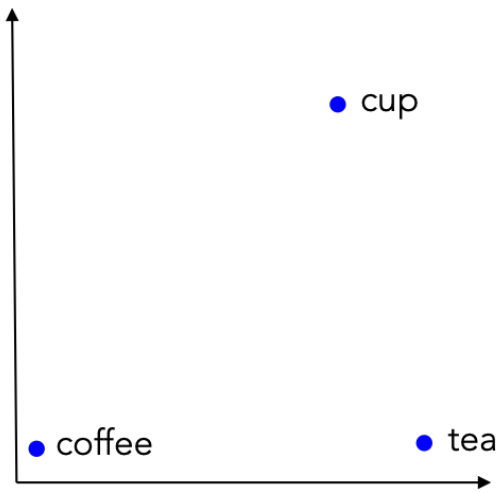
(Firth, 1957)

From Context to Embeddings



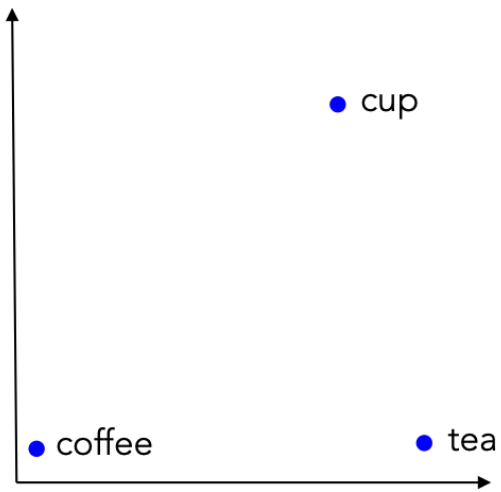
I had a cup of coffee this morning

From Context to Embeddings



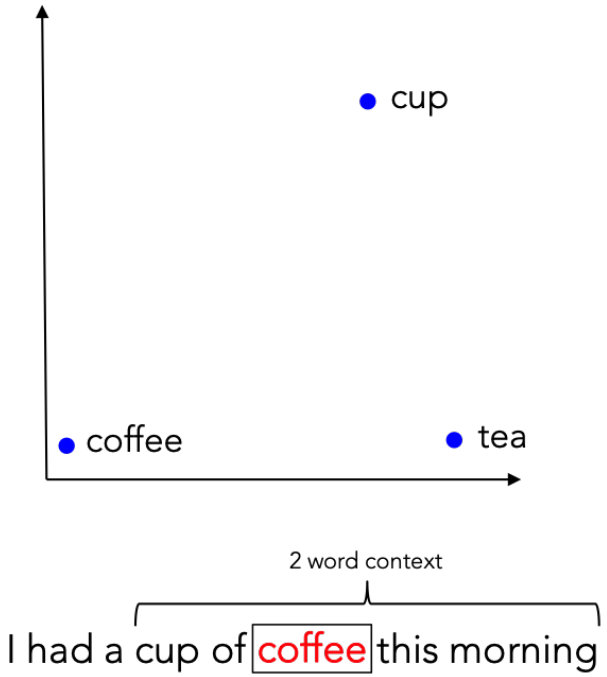
I had a cup of coffee this morning

From Context to Embeddings

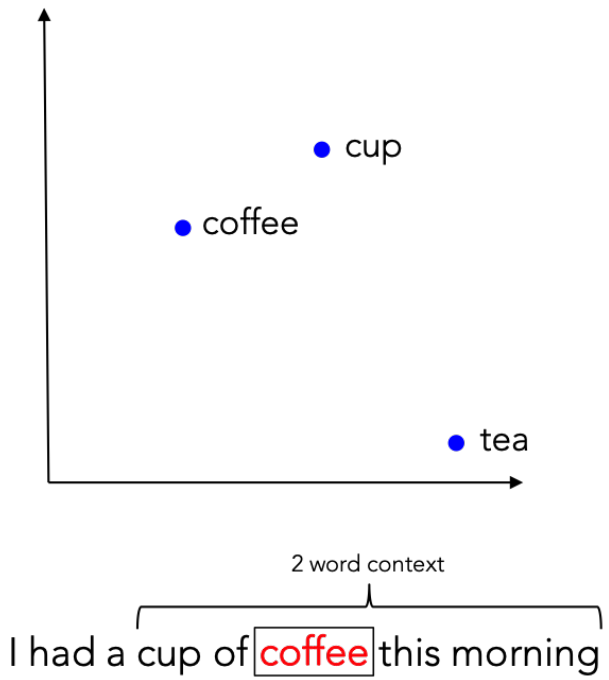


I had a cup of coffee this morning

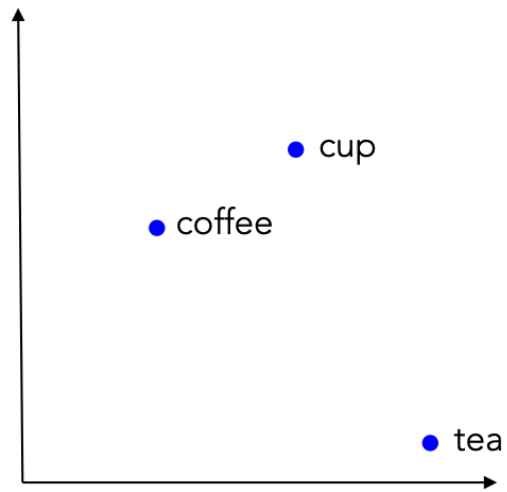
From Context to Embeddings



From Context to Embeddings

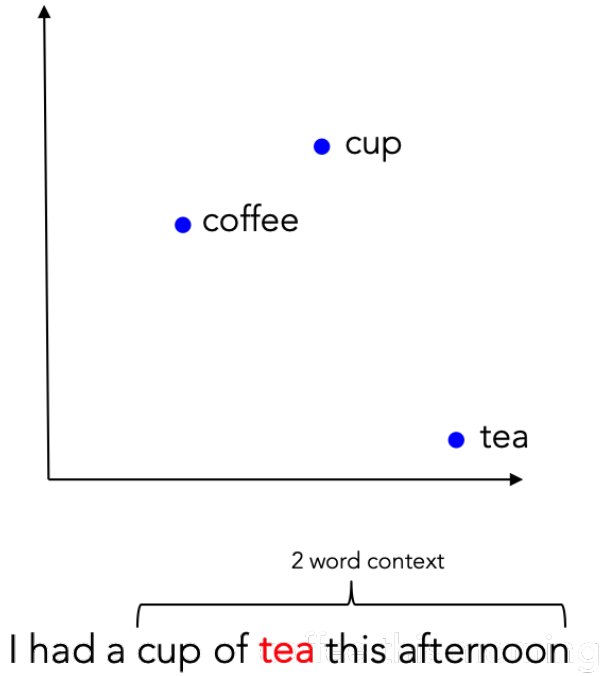


From Context to Embeddings

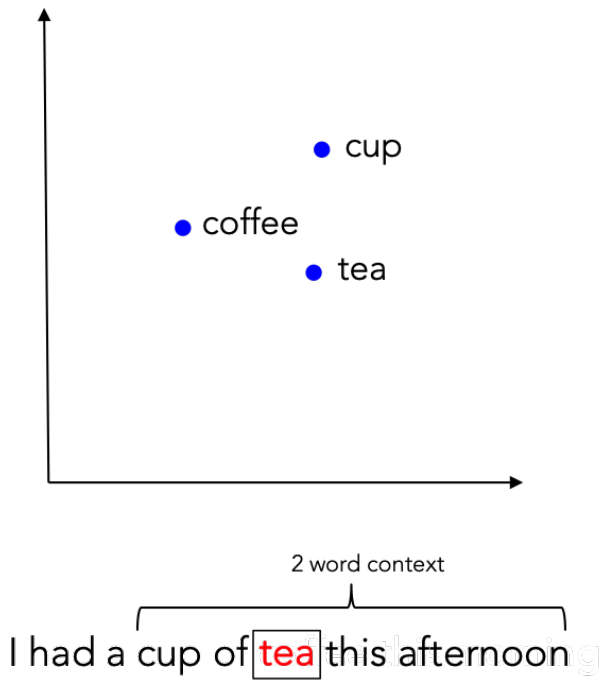


I had a cup of tea this afternoon

From Context to Embeddings



From Context to Embeddings



Some Foreground

I need to...

I need to...

...automatically generate a dictionary from unlabeled corpora (Hamilton et al, 2016)

I need to...

...automatically generate a dictionary from unlabeled corpora (Hamilton et al, 2016)

...model ideology of parliamentarians (Rheult & Cochrane, 2019)

I need to...

...automatically generate a dictionary from unlabeled corpora (Hamilton et al, 2016)

...model ideology of parliamentarians (Rheult & Cochrane, 2019)

...improve performance of `readme` (Jerzak et al, 2018)

I need to...

...automatically generate a dictionary from unlabeled corpora (Hamilton et al, 2016)

...model ideology of parliamentarians (Rheult & Cochrane, 2019)

...improve performance of `readme` (Jerzak et al, 2018)

...understand how words have changed meaning over time (Rodman, 2019)

I need to...

...automatically generate a dictionary from unlabeled corpora (Hamilton et al, 2016)

...model ideology of parliamentarians (Rheult & Cochrane, 2019)

...improve performance of `readme` (Jerzak et al, 2018)

...understand how words have changed meaning over time (Rodman, 2019)

→ choose a model (Word2Vec, GloVe),

I need to...

...automatically generate a dictionary from unlabeled corpora (Hamilton et al, 2016)

...model ideology of parliamentarians (Rheult & Cochrane, 2019)

...improve performance of `readme` (Jerzak et al, 2018)

...understand how words have changed meaning over time (Rodman, 2019)

→ choose a model (Word2Vec, GloVe), an architecture within that model (CBOW, Skipgram),

I need to...

...automatically generate a dictionary from unlabeled corpora (Hamilton et al, 2016)

...model ideology of parliamentarians (Rheult & Cochrane, 2019)

...improve performance of `readme` (Jerzak et al, 2018)

...understand how words have changed meaning over time (Rodman, 2019)

→ choose a model (Word2Vec, GloVe), an architecture within that model (CBOW, Skipgram), parameters within that architecture (window size, embedding length)

I need to...

...automatically generate a dictionary from unlabeled corpora (Hamilton et al, 2016)

...model ideology of parliamentarians (Rheult & Cochrane, 2019)

...improve performance of `readme` (Jerzak et al, 2018)

...understand how words have changed meaning over time (Rodman, 2019)

→ choose a model (Word2Vec, GloVe), an architecture within that model (CBOW, Skipgram), parameters within that architecture (window size, embedding length) and a training set (pretrained, locally fit).

I need to...

choose a model (Word2Vec, GloVe), an architecture within that model (CBOW, Skipgram), parameters within that architecture (window size, embedding length) and a training set (pretrained, locally fit).

The Problem

The Problem

There are **no** generally accepted downstream tasks in political science:

The Problem

There are **no** generally accepted downstream tasks in political science: 'extrinsic' evaluation criteria make no sense (e.g. analogy banks, learner accuracy).

The Problem

There are **no** generally accepted downstream tasks in political science: 'extrinsic' evaluation criteria make no sense (e.g. analogy banks, learner accuracy).

So, embeddings are **only** useful to the extent they capture semantically meaningful information about politics:

The Problem

There are **no** generally accepted downstream tasks in political science: 'extrinsic' evaluation criteria make no sense (e.g. analogy banks, learner accuracy).

So, embeddings are **only** useful to the extent they capture semantically meaningful information about politics: focus on 'intrinsic' evaluation criteria.

The Problem

There are **no** generally accepted downstream tasks in political science: 'extrinsic' evaluation criteria make no sense (e.g. analogy banks, learner accuracy).

So, embeddings are **only** useful to the extent they capture semantically meaningful information about politics: focus on 'intrinsic' evaluation criteria.

But how can we evaluate this?

The Solution

The Solution

We propose a "Turing test":

The Solution

We propose a "Turing test": ask crowdworkers whether output from humans or machine (model) fits a cue better.

The Solution

We propose a "Turing test": ask crowdworkers whether output from humans or machine (model) fits a cue better.

We get remarkable, human-like performance from embeddings models in terms of meaning. ✓

Wait, there's more...

Wait, there's more...

Embeddings models have multiple parameters,

Wait, there's more...

Embeddings models have multiple parameters, especially window-size and embedding dimensions.

Wait, there's more...

Embeddings models have multiple parameters, especially window-size and embedding dimensions.

And should you use pretrained or locally fit?

Wait, there's more...

Embeddings models have multiple parameters, especially window-size and embedding dimensions.

And should you use pretrained or locally fit?

We use our technical critera on fit and stability and our Turing test to provide advice.

Wait, there's more...

Embeddings models have multiple parameters, especially window-size and embedding dimensions.

And should you use pretrained or locally fit?

We use our technical critera on fit and stability and our Turing test to provide advice.

Avoid small windows, few dimensions but otherwise results are robust to these parameter choices. ✓

Wait, there's more...

Embeddings models have multiple parameters, especially window-size and embedding dimensions.

And should you use pretrained or locally fit?

We use our technical critera on fit and stability and our Turing test to provide advice.

Avoid small windows, few dimensions but otherwise results are robust to these parameter choices. ✓

Pretrained embeddings work about as well as anything else. ✓

Implementing Choices

Implementing Choices

We train a series of (local) models.

Implementing Choices

We train a series of (local) models.

We focus on two hyperparameter choices (25 combinations):

Implementing Choices

We train a series of (local) models.

We focus on two hyperparameter choices (25 combinations):

window-size —1, 6, 12, 24 and 48

Implementing Choices

We train a series of (local) models.

We focus on two hyperparameter choices (25 combinations):

window-size —1, 6, 12, 24 and 48

embedding dimension —50, 100, 200, 300 and 450

Implementing Choices

We train a series of (local) models.

We focus on two hyperparameter choices (25 combinations):

window-size —1, 6, 12, 24 and 48

embedding dimension —50, 100, 200, 300 and 450

For each pair we estimate 10 sets of embeddings (250 in all).

Implementing Choices

We train a series of (local) models.

We focus on two hyperparameter choices (25 combinations):

window-size —1, 6, 12, 24 and 48

embedding dimension —50, 100, 200, 300 and 450

For each pair we estimate 10 sets of embeddings (250 in all).

Also compare to `GloVe` and `Word2Vec` (skip-gram) pretrained.

Evaluation Criteria

technical criteria: model loss and computation time.

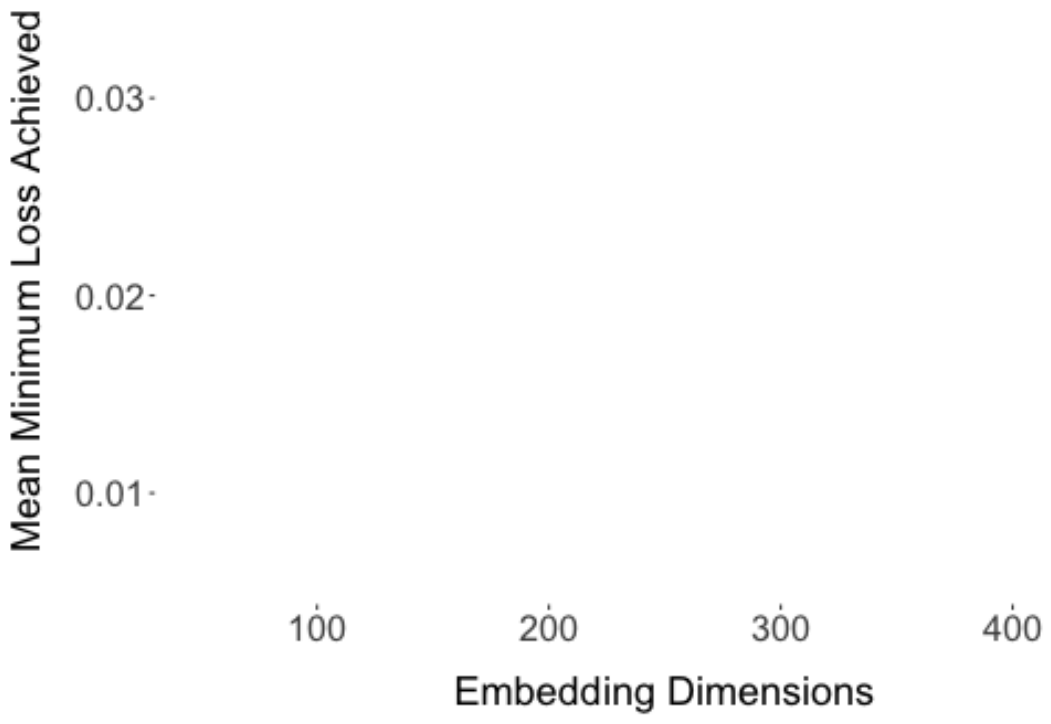
model variance (stability): within-model Pearson correlation of nearest neighbor rankings across multiple initializations.

query search ranking correlation: Pearson and rank correlations of cosine similarities.

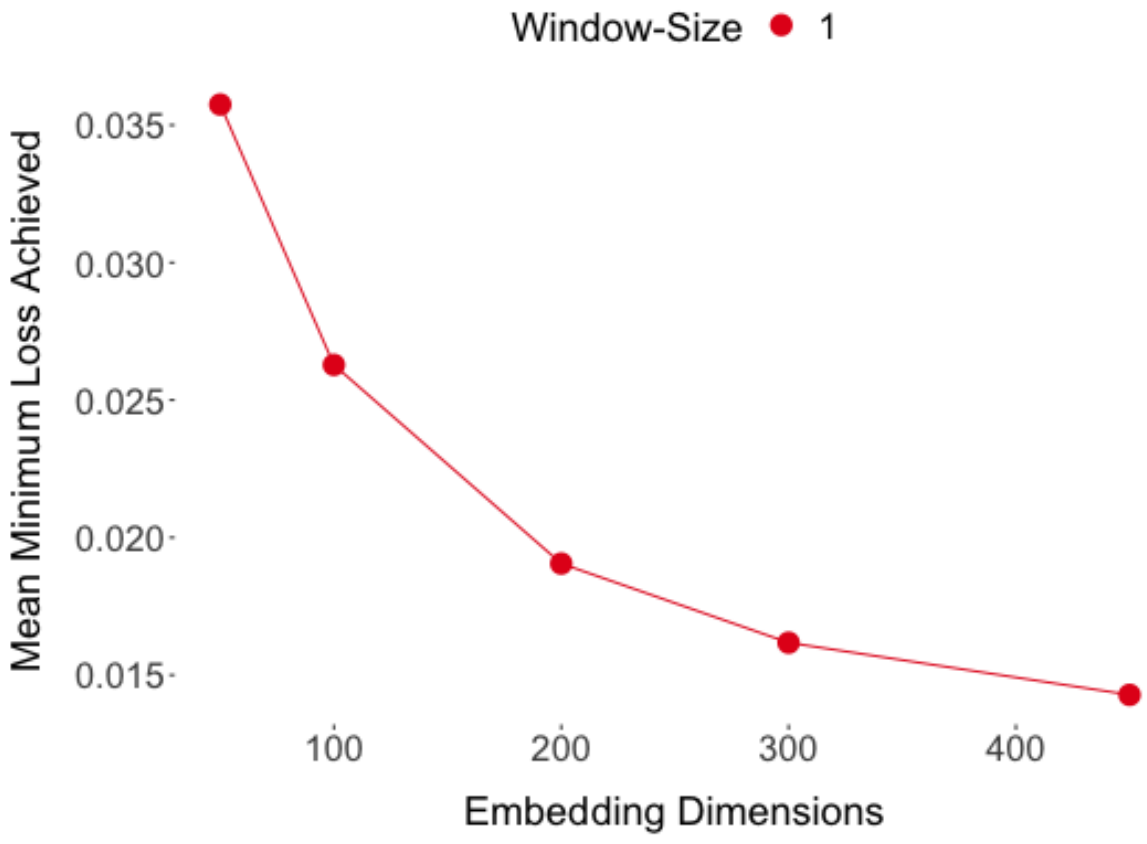
human preference: a "Turing test" assessment and rank deviations from human generated lists.

Technical Criteria: Model Loss

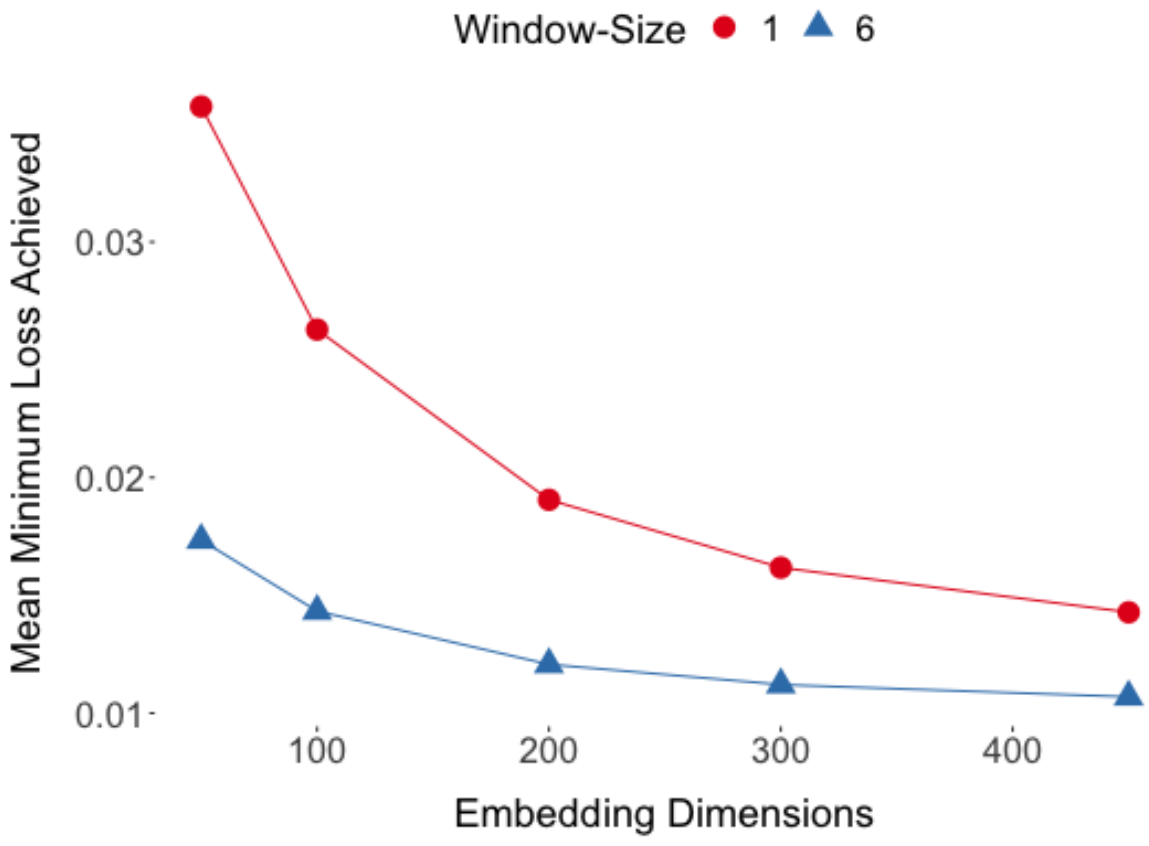
Technical Criteria: Model Loss



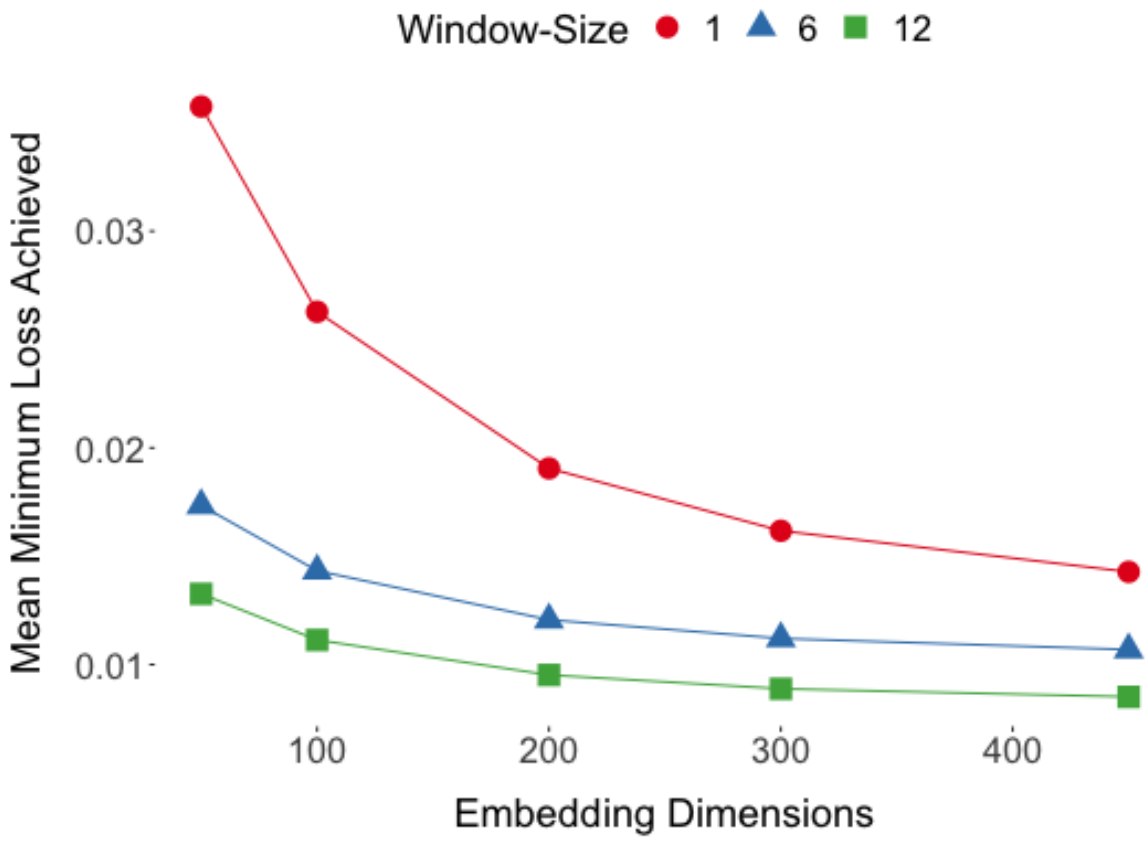
Technical Criteria: Model Loss



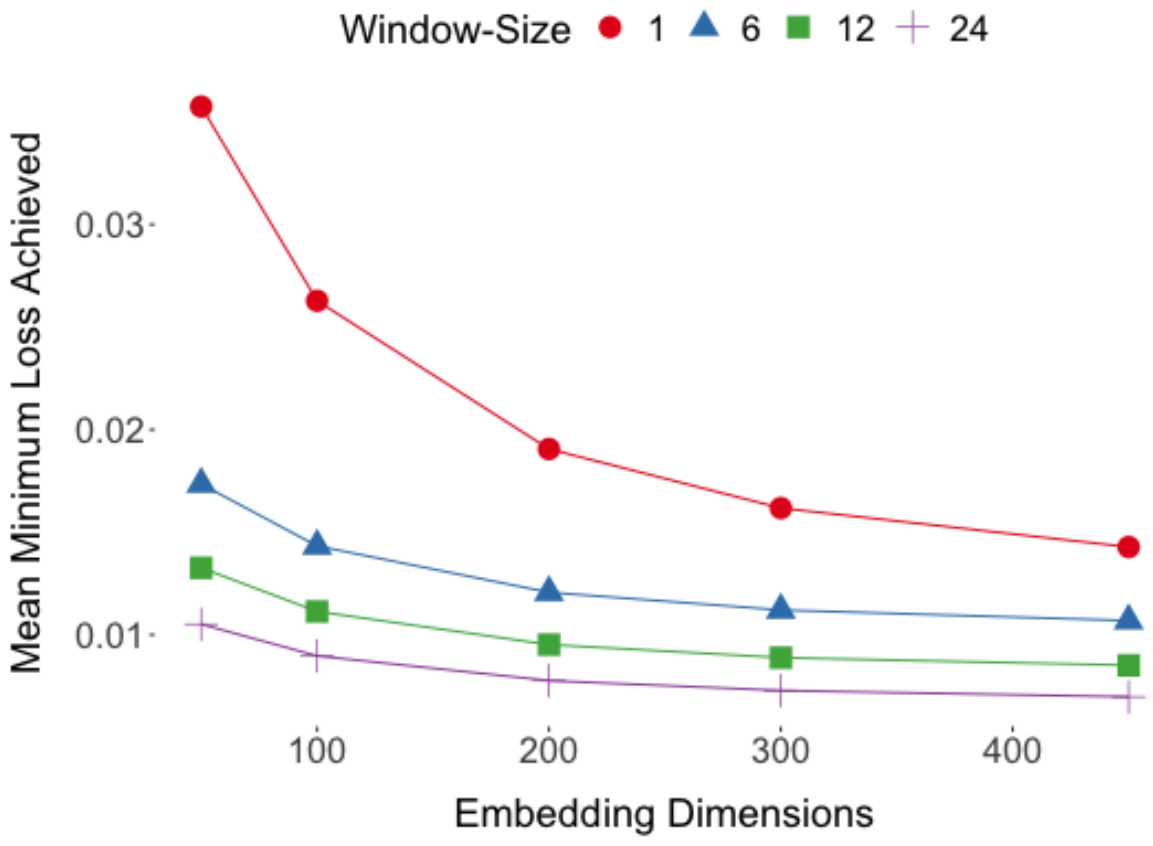
Technical Criteria: Model Loss



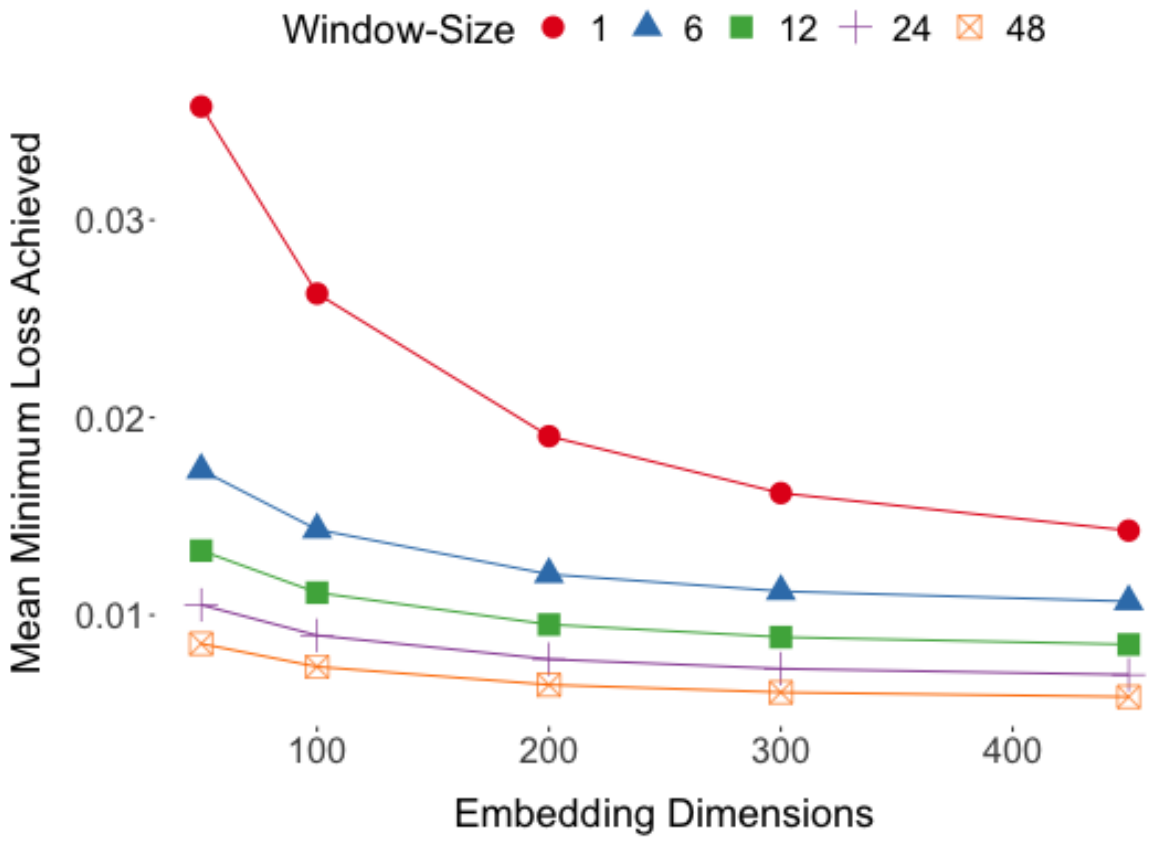
Technical Criteria: Model Loss



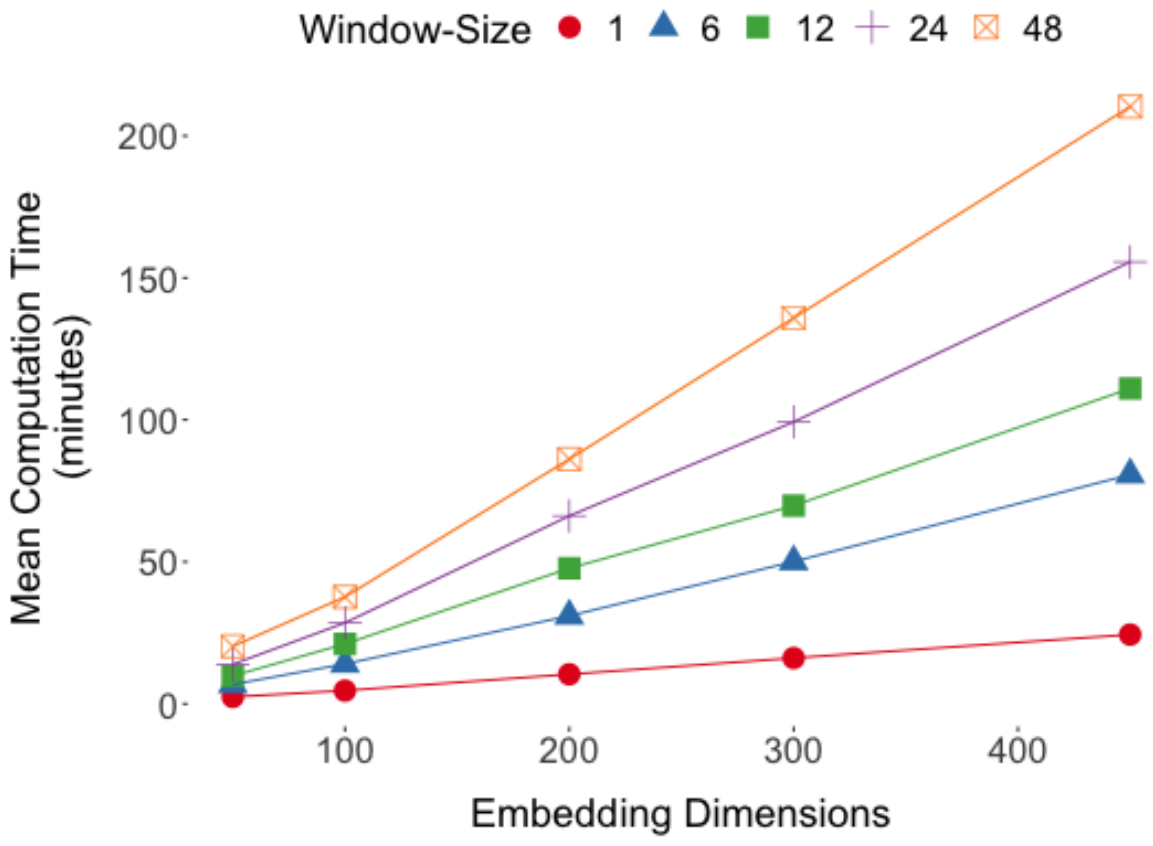
Technical Criteria: Model Loss



Technical Criteria: Model Loss



Technical Criteria: Computation Time



Evaluation Criteria

technical criteria: model loss and computation time.

model variance (stability): within-model Pearson correlation of nearest neighbor rankings across multiple initializations.

query search correlations: Pearson and rank correlations of cosine similarities.

human preference: a "Turing test" assessment and rank deviations from human generated lists.

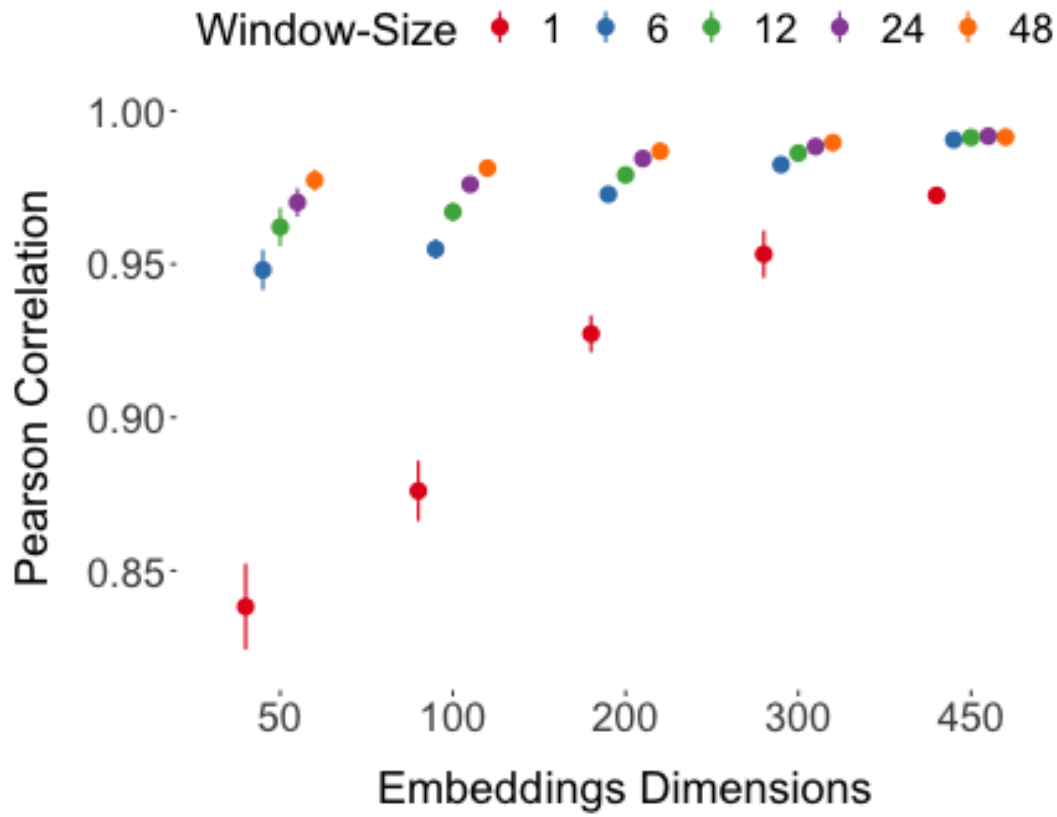
Evaluation Criteria

technical criteria: model loss and computation time.

model variance (stability): within-model Pearson correlation of nearest neighbor rankings across multiple initializations.

query search correlations: Pearson and rank correlations of cosine similarities.

human preference: a "Turing test" assessment and rank deviations from human generated lists.


Window-Size
1
6
12
24
48
1.00
0.95
0.90
0.85
Pearson Correlation
50
100
200
300
450
Embeddings Dimensions

Evaluation Criteria

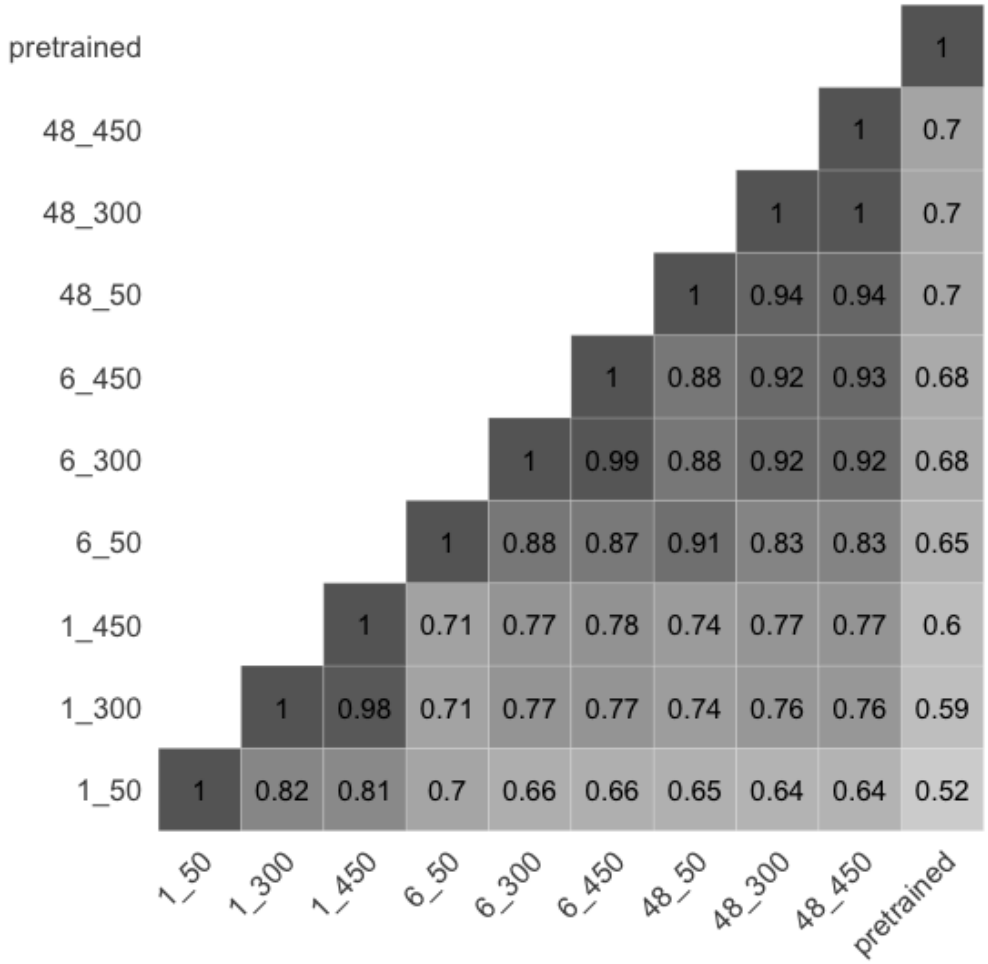
technical criteria: model loss and computation time.

model variance (stability): within-model Pearson correlation of nearest neighbor rankings across multiple initializations.

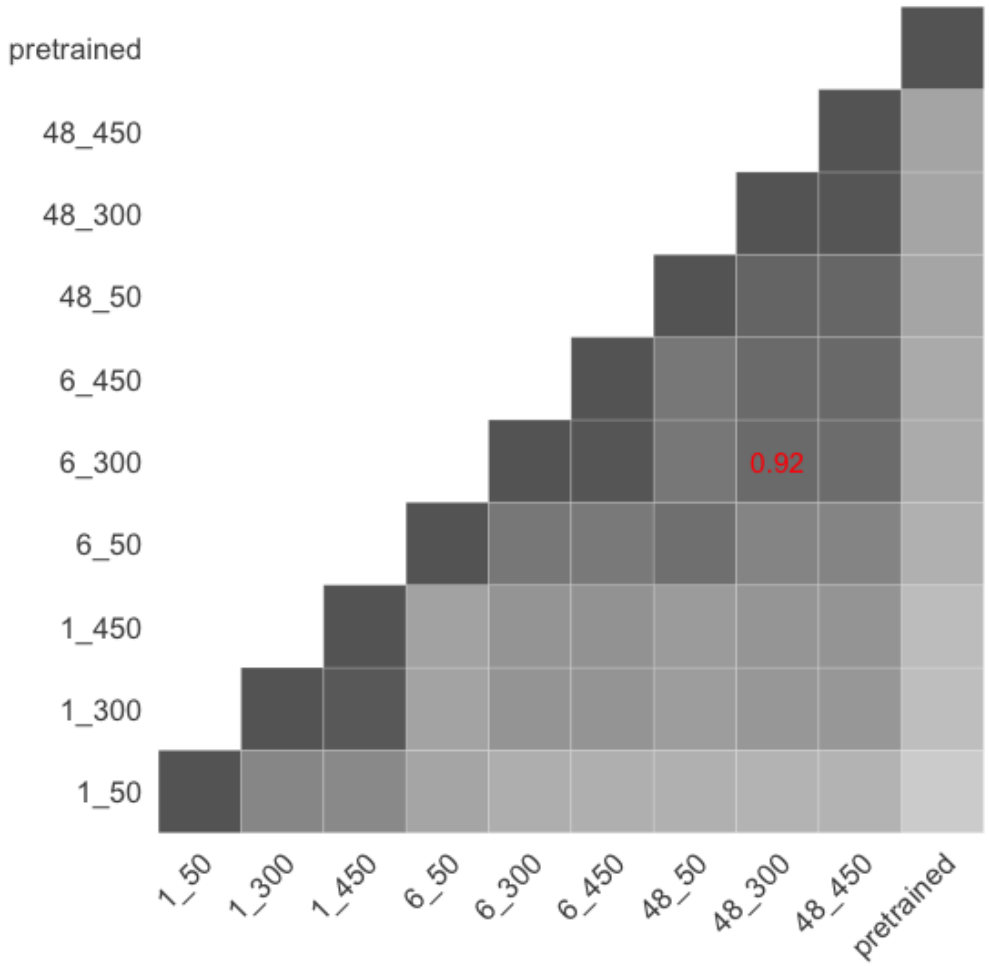
query search correlations: Pearson and rank correlations of cosine similarities.

human preference: a "Turing test" assessment and rank deviations from human generated lists.

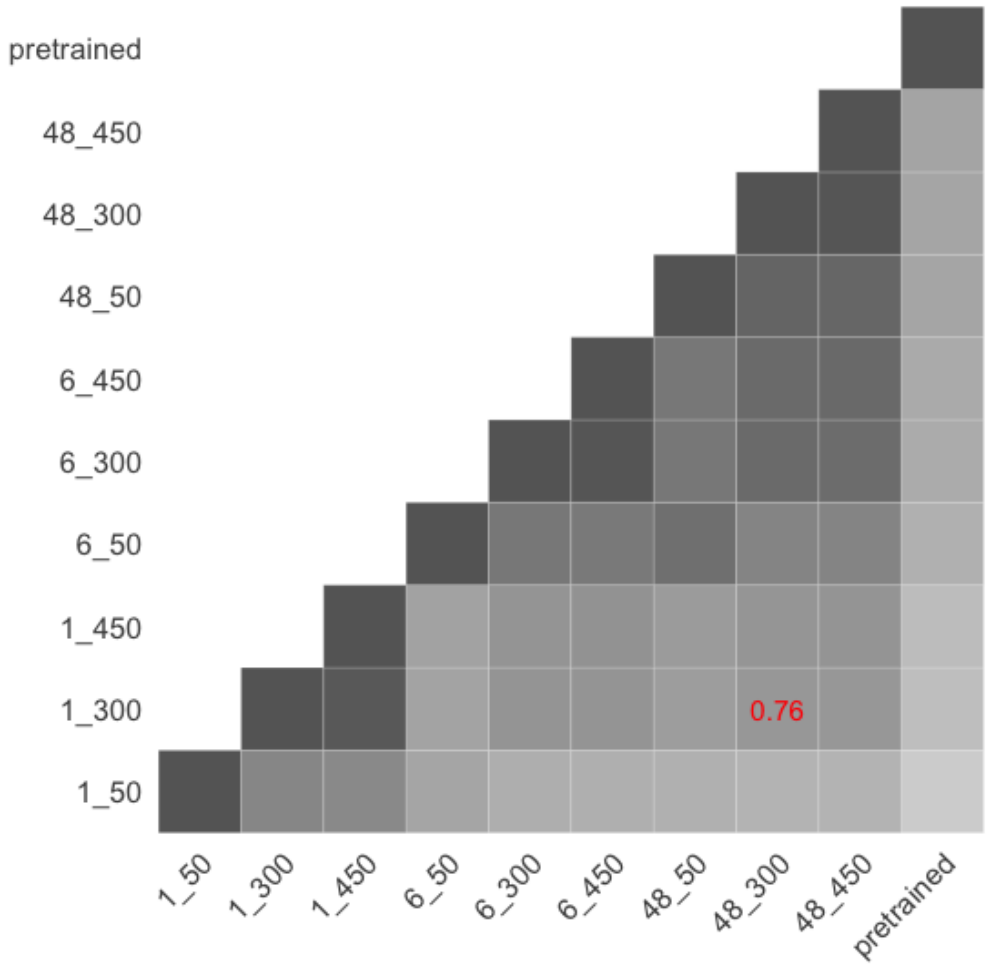
Query Search Correlations (curated queries)



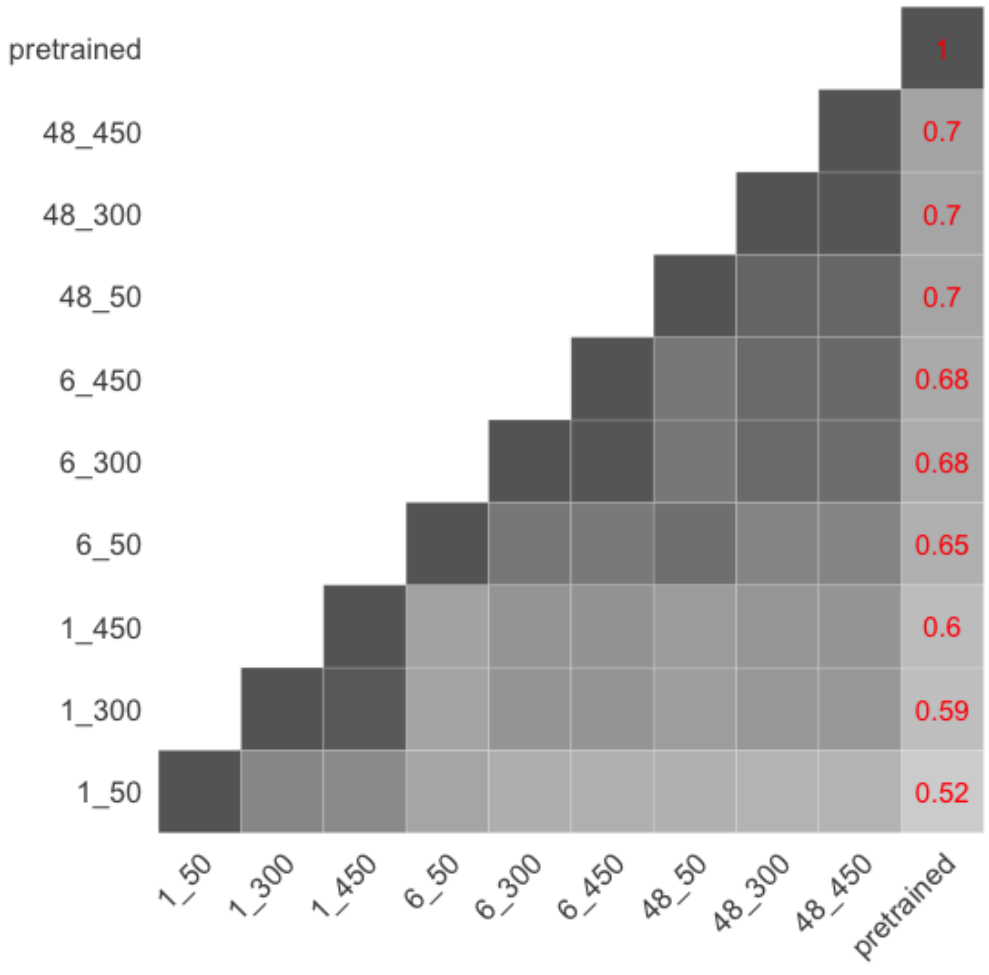
Query Search Correlations (curated queries)



Query Search Correlations (curated queries)



Query Search Correlations (curated queries)



Evaluation Criteria

technical criteria: model loss and computation time.

model variance (stability): within-model Pearson correlation of nearest neighbor rankings across multiple initializations.

query search correlations: Pearson and rank correlations of cosine similarities.

human preference: a "Turing test" assessment and rank deviations from human generated lists.

RShiny App: Definitions

Context Words

A famous maxim in the study of linguistics states that:

You shall know a word by the company it keeps. (Firth, 1957)

This task is designed to help us understand the nature of the ''company'' that words ''keep'': that is, their CONTEXT.

Specifically, for a CUE WORD, its CONTEXT WORDS include words that:

- Tend to occur in the vicinity of the CUE WORD. That is, they are words that appear close to the CUE WORD in written or spoken language.

AND/OR

- Tend to occur in similar situations to the CUE WORD in spoken and written language. That is, they are words that regularly appear with other words that are closely related to the CUE WORD.

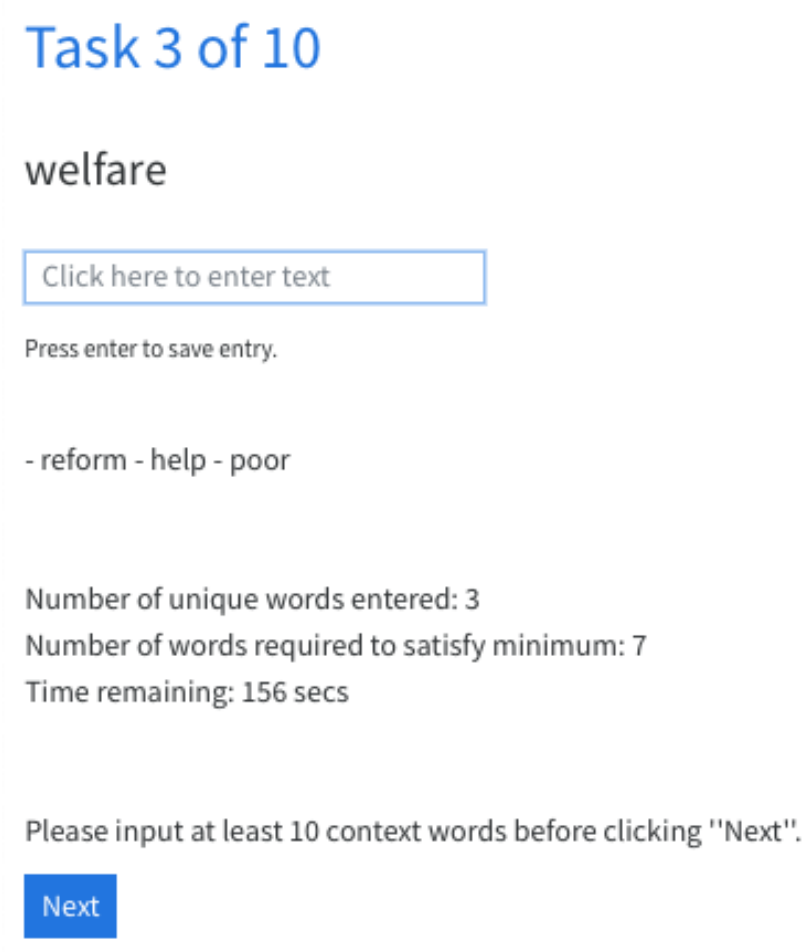
For example, CONTEXT WORDS for the cue word COFFEE include:

1. *cup* (tends to occur in the vicinity of COFFEE).
2. *tea* (tends to occur in similar situations to COFFEE, for example when discussing drinks).

Click ''Next'' to continue

Next

RShiny App-1: Context Word Generation



RShiny App-2: "Turing" Context Word Evaluation

RShiny App-2: "Turing" Context Word Evaluation

RShiny App-2: "Turing" Context Word Evaluation

RShiny App-2: "Turing" Context Word Evaluation

Machine: 1 - Human: 0

Candidate: Local 6-300

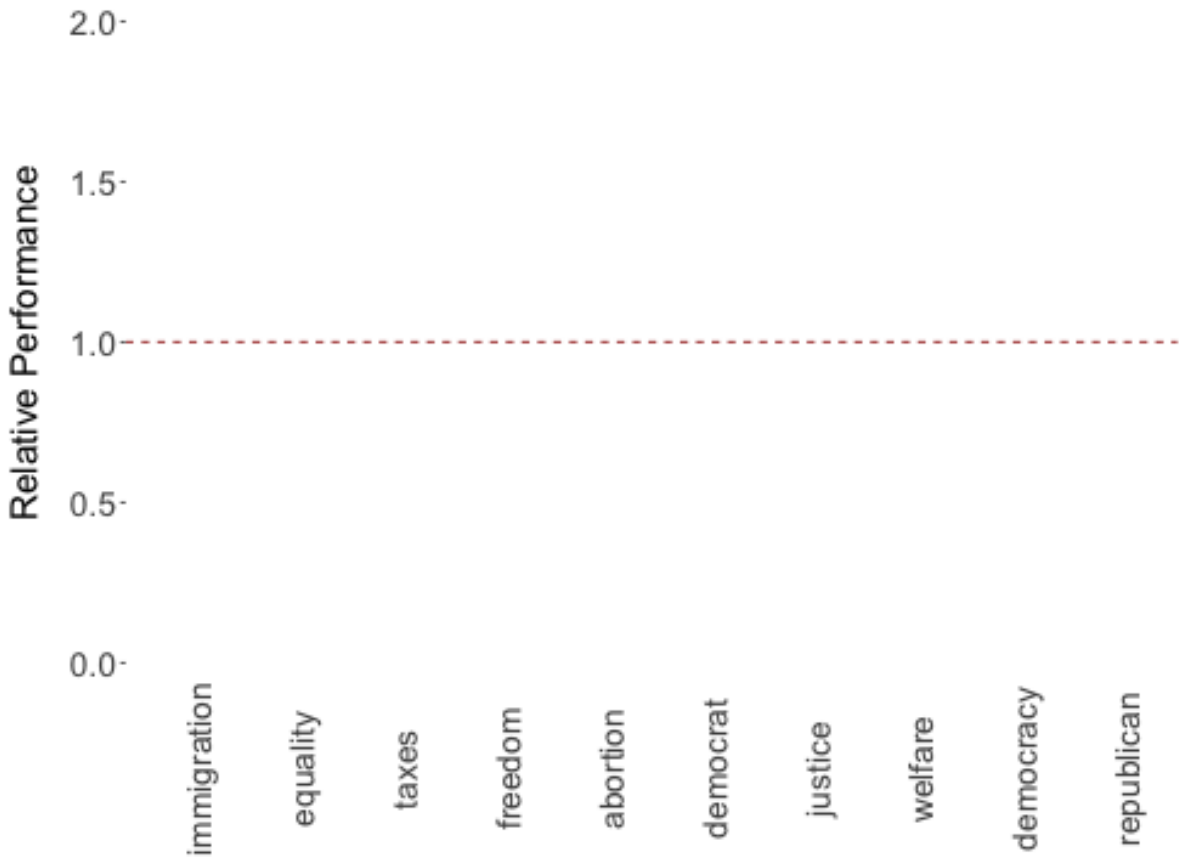
Baseline: Human

Candidate: Local 6-300

Baseline: Human

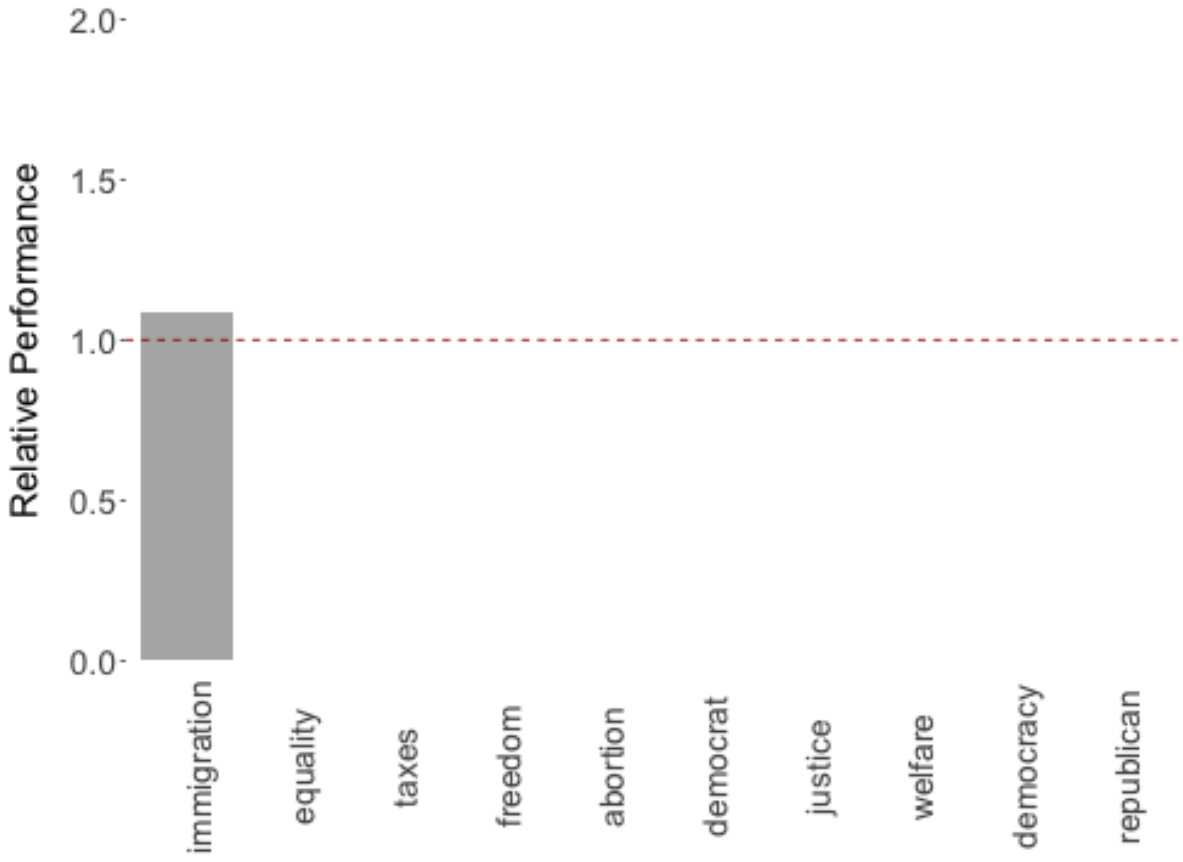
Candidate: Local 6-300

Baseline: Human



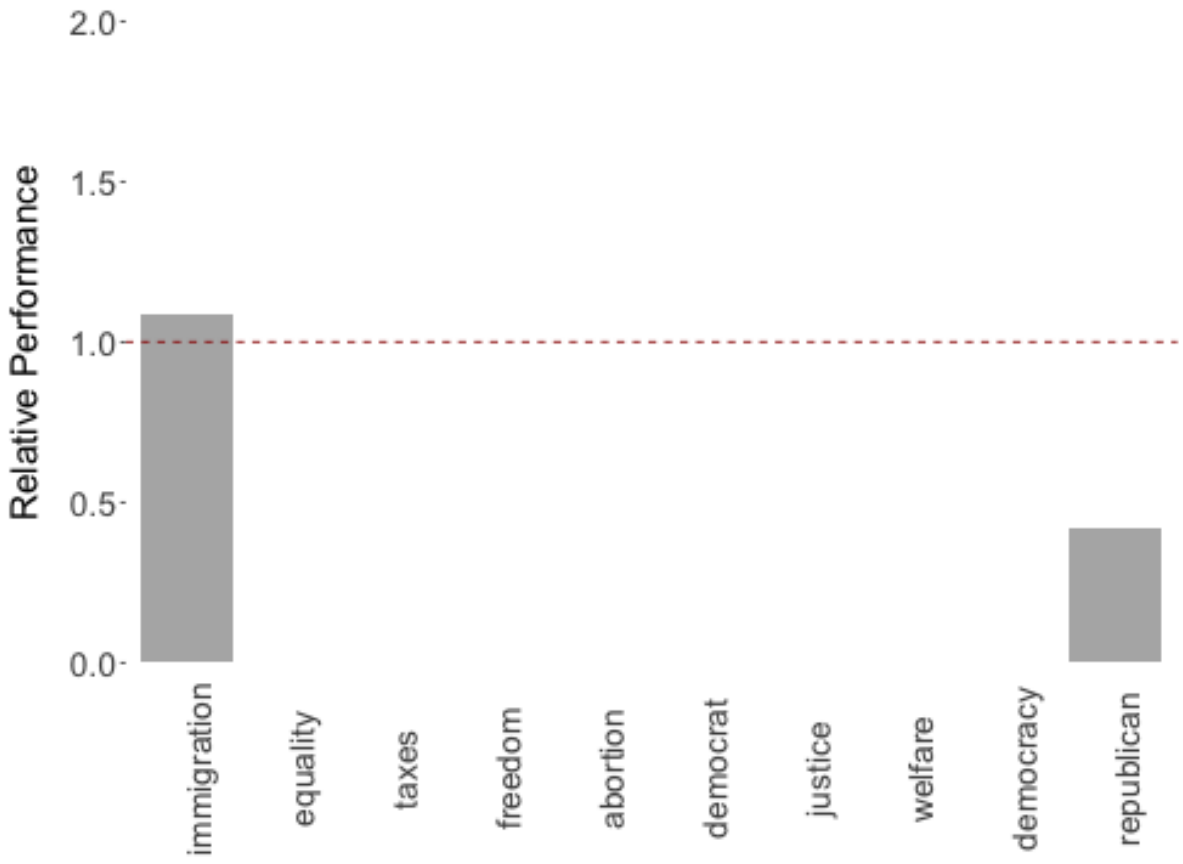
Candidate: Local 6-300

Baseline: Human



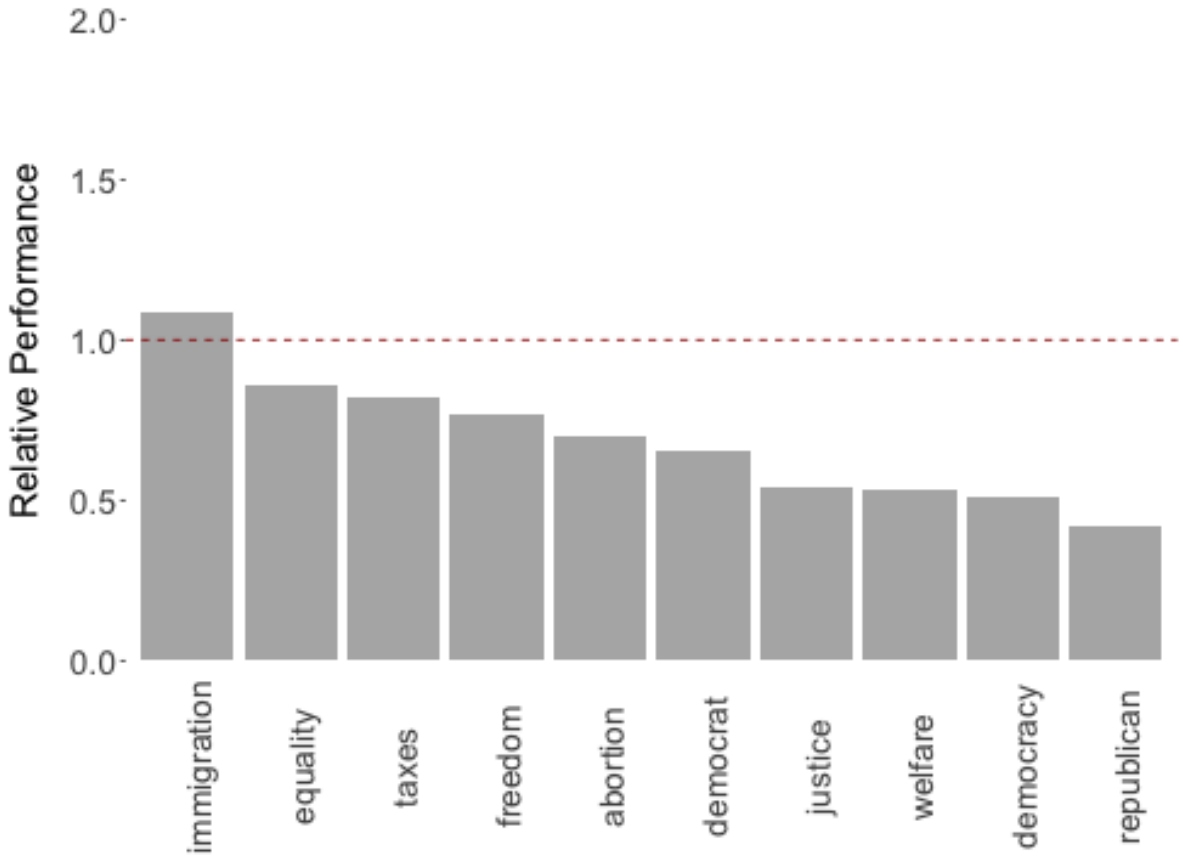
Candidate: Local 6-300

Baseline: Human



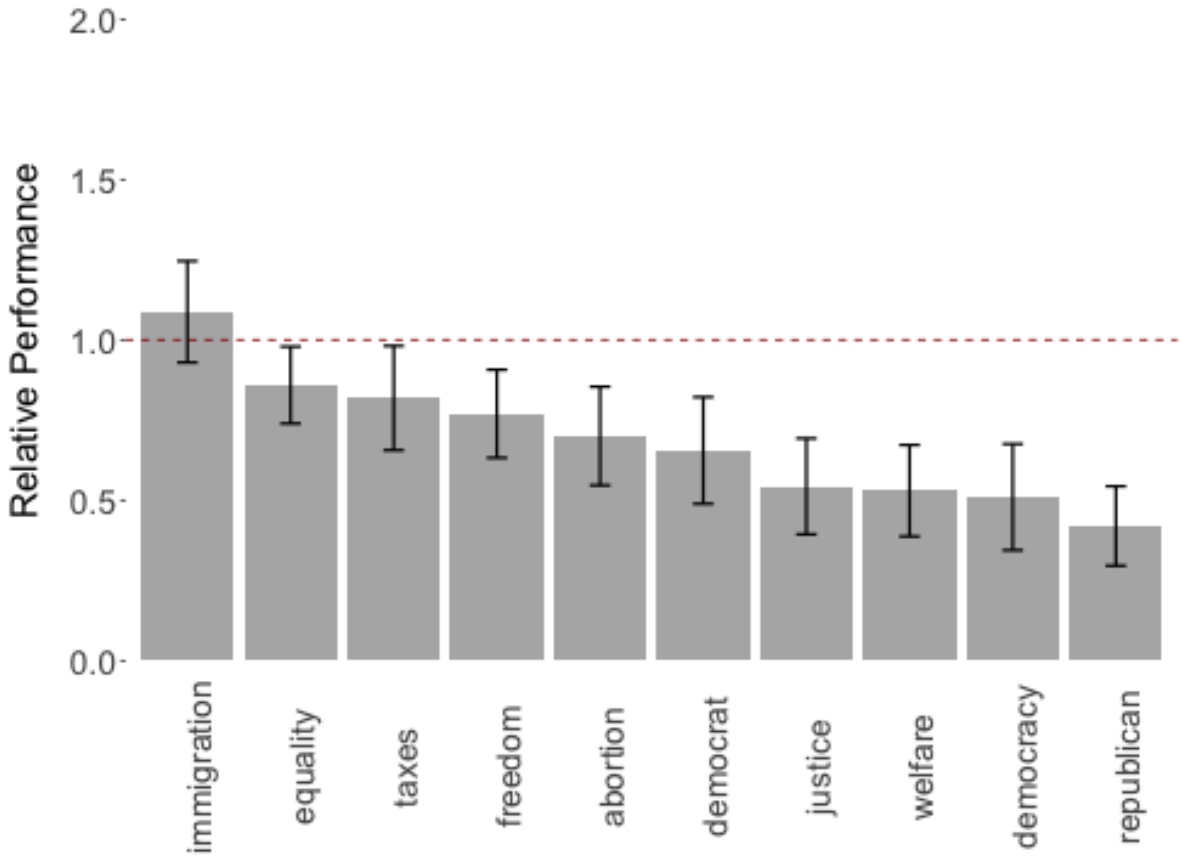
Candidate: Local 6-300

Baseline: Human



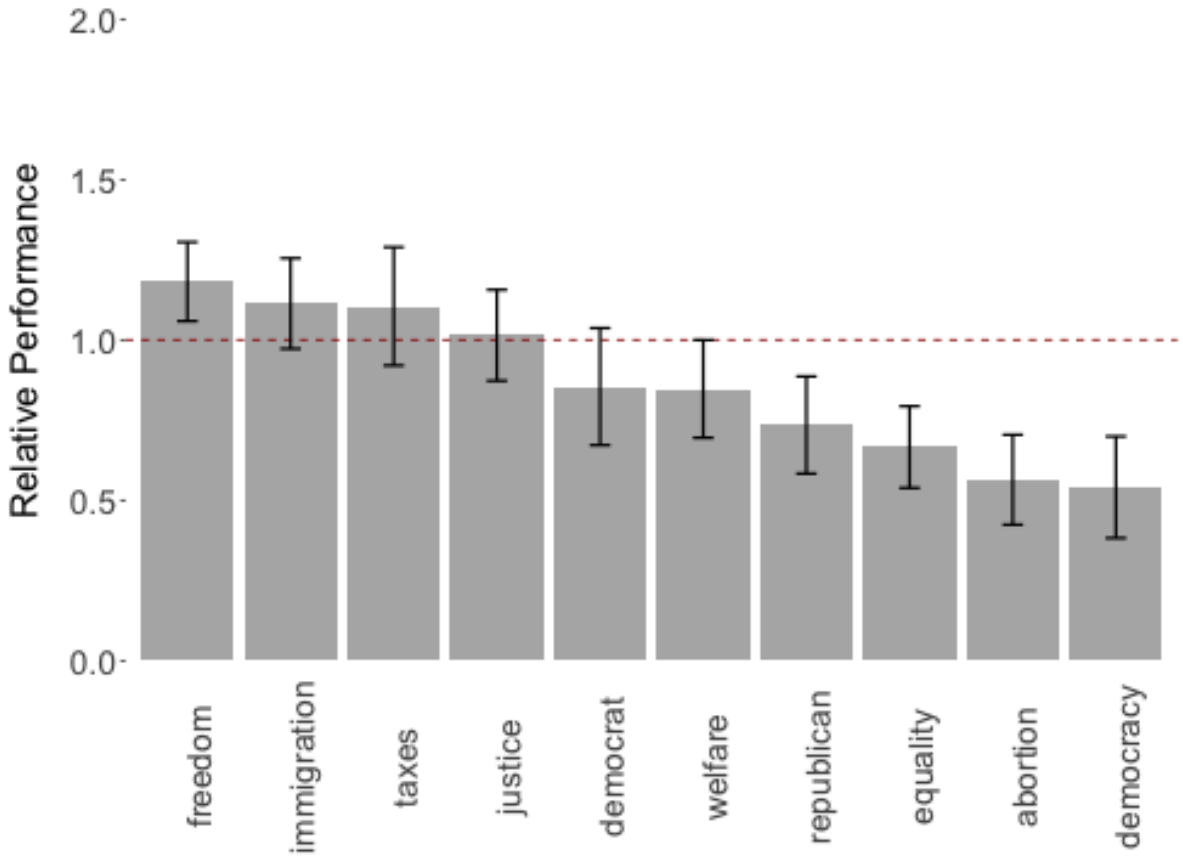
Candidate: Local 6-300

Baseline: Human



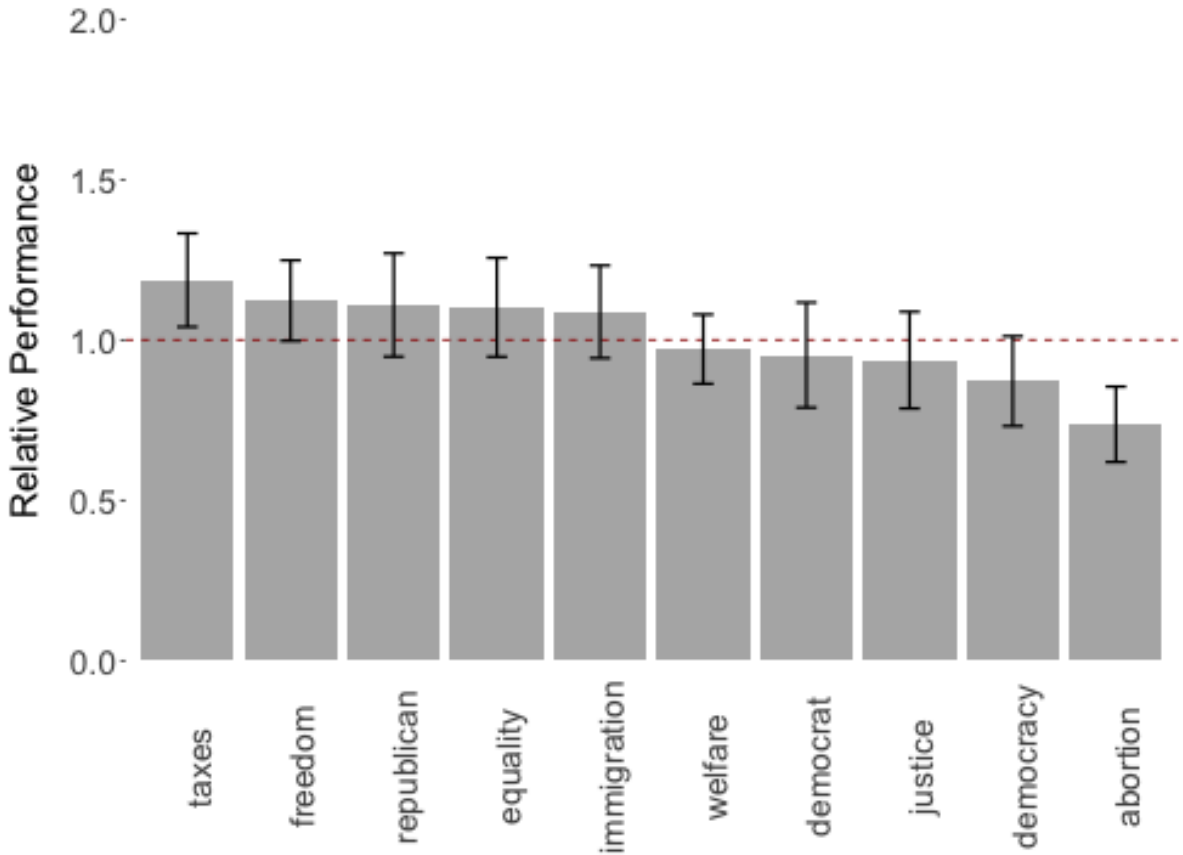
Candidate: GloVe pretrained

Baseline: Human



Candidate: `GloVe` pretrained

Baseline: `W2V` pretrained



Wrapping Up

Wrapping Up

We get remarkable, human-like performance from embeddings models in terms of meaning. ✓

Wrapping Up

We get remarkable, human-like performance from embeddings models in terms of meaning. ✓

Avoid small windows, few dimensions but otherwise results are robust to these parameter choices. ✓

Wrapping Up

We get remarkable, human-like performance from embeddings models in terms of meaning. ✓

Avoid small windows, few dimensions but otherwise results are robust to these parameter choices. ✓

Pretrained, 'default' embeddings work about as well as anything else. ✓

Wrapping Up

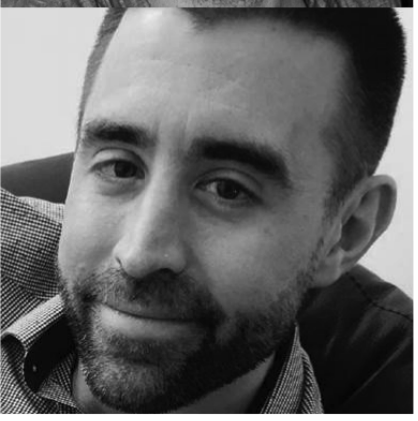
We get remarkable, human-like performance from embeddings models in terms of meaning. ✓

Avoid small windows, few dimensions but otherwise results are robust to these parameter choices. ✓

Pretrained, 'default' embeddings work about as well as anything else. ✓

`GloVe` appears to be more robust than `Word2vec`, but both are equally liked by humans. ✓

Thank you!



R software package to apply our proposed metrics and framework: coming soon.

GitHub:
`http://github.com/ArthurSpirling/EmbeddingsPaper`

Paper:
`../Paper/Embeddings_SpirlingRodriguez.pdf`

FAQ:
`../Project_FAQ/faq.md`