Practical observations in applying neural word embeddings to machine translation

Michael Seeber
Skymind Labs and GalvanizeU
San Francisco, CA

Overview

- Hypothesis
- Pre-trained embeddings
- Bilingual dictionaries
- Translation methodology
- Experimental Results

Hypothesis

- Neural Word Embeddings
  - Vectors that represent Words

- Blueberry -> черника

- A translation matrix can translate words between two languages via their word vectors

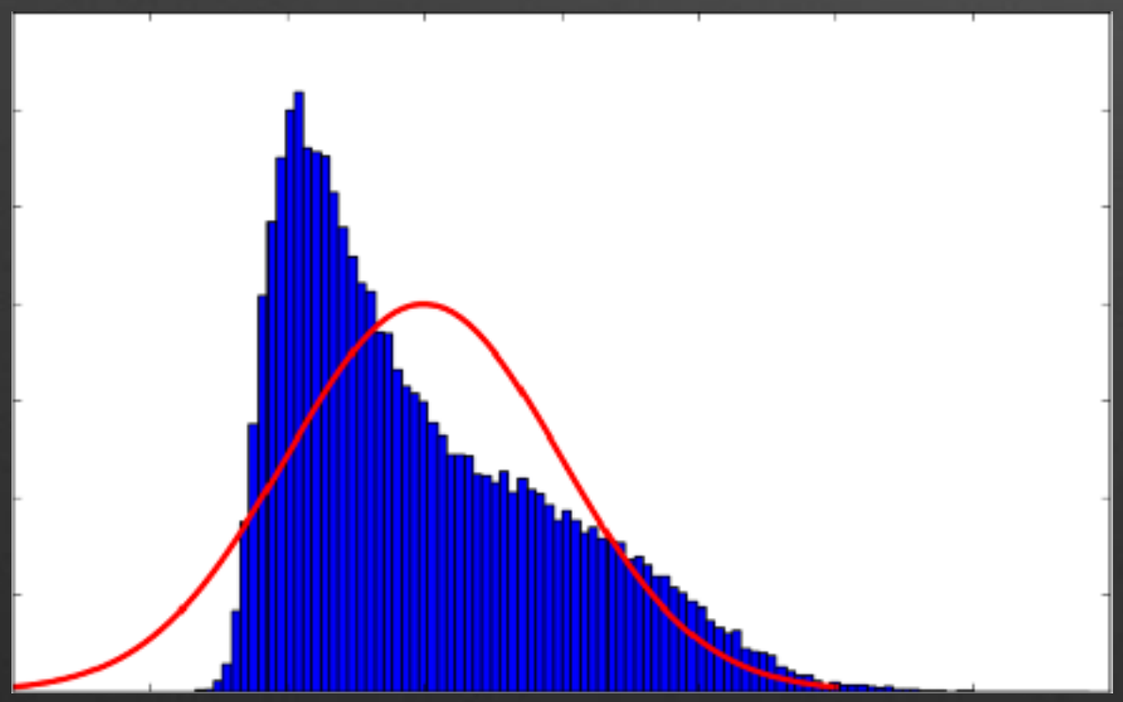
Pre-trained embeddings

Kyubyong

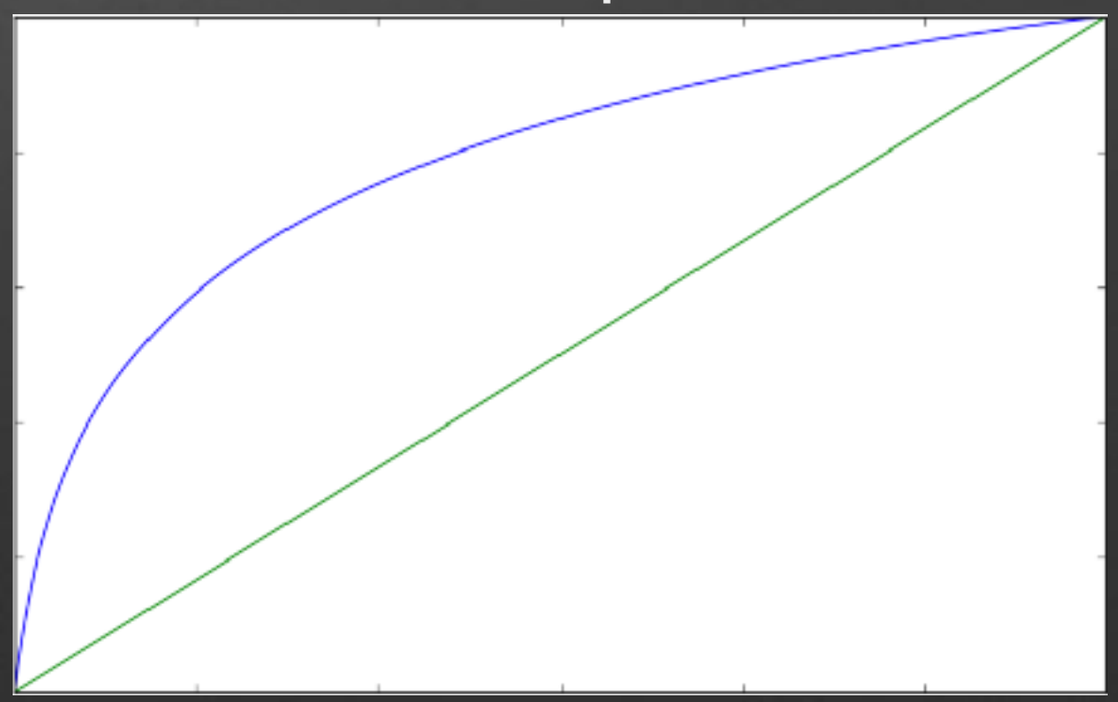
- 29 Languages
- Trained on Wikipedia
- Word2Vec Algorithm
- 50,000 Vocabulary Size (or lower)
- 300 Vector Size

French

Distribution of L2 Norms



PCA Cumulative Explained Variance

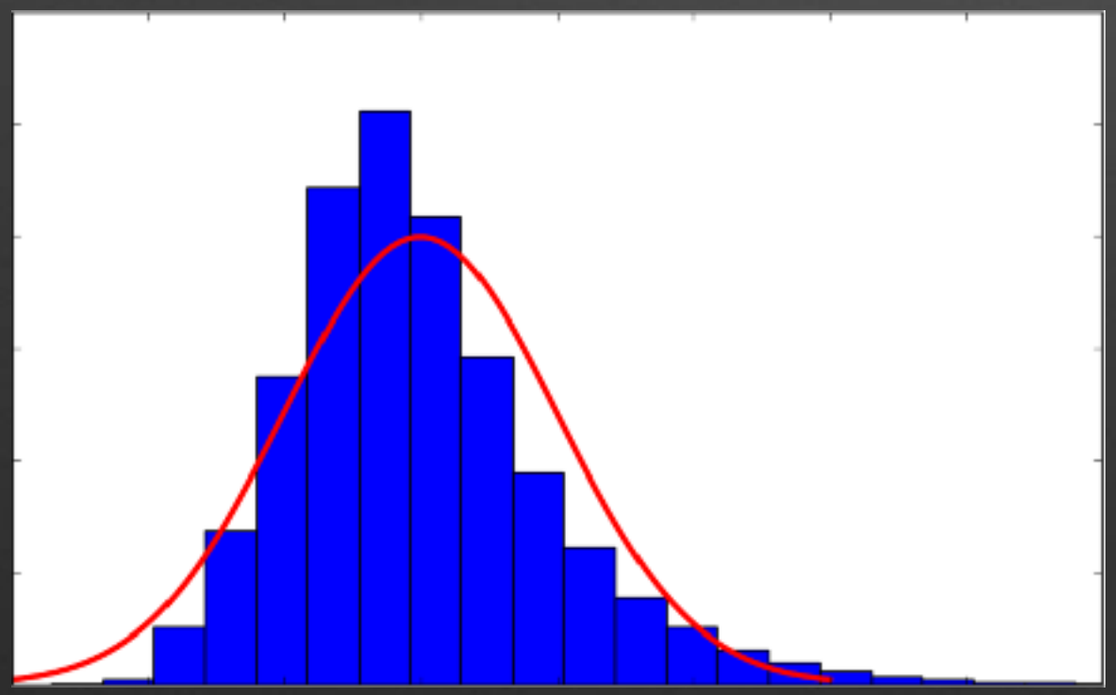


Polyglot

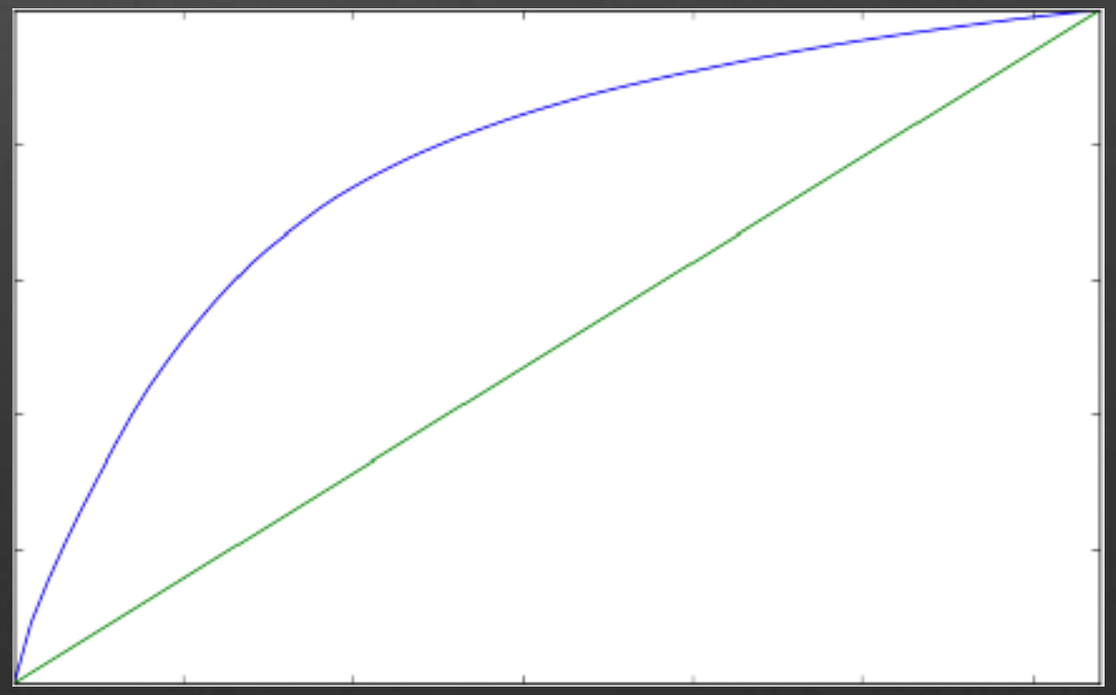
- 136 Languages
- Trained on Wikipedia
- 100,000 Vocabulary Size (or lower)
- 64 Vector Size

English

Distribution of L2 Norms



PCA Cumulative Explained Variance

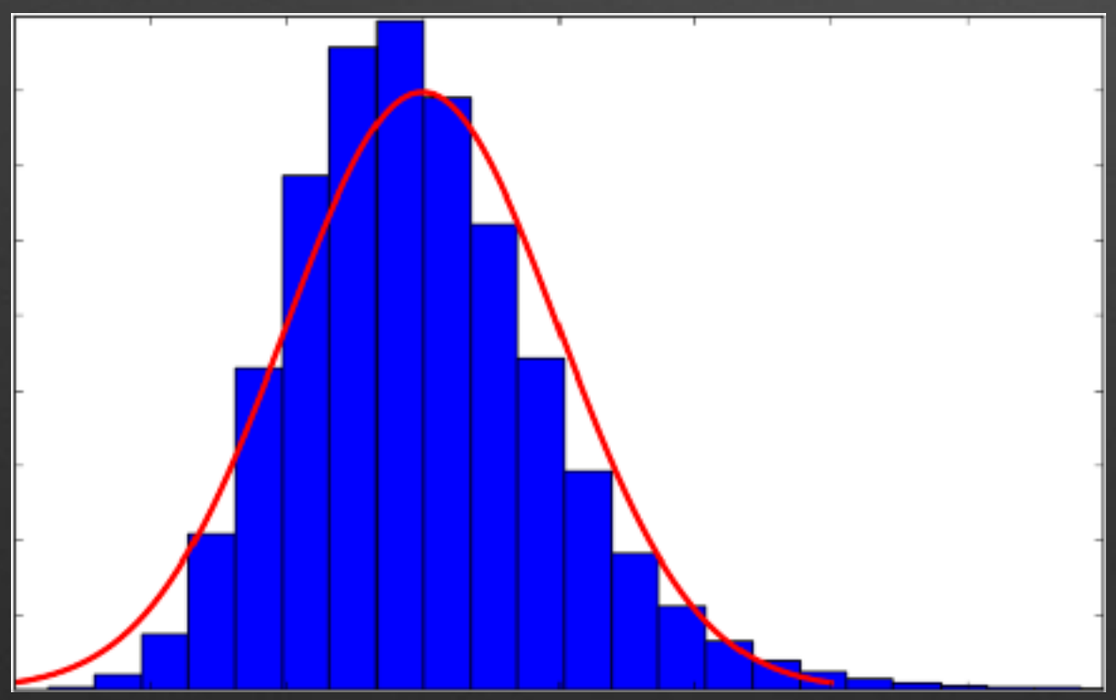


Fasttext

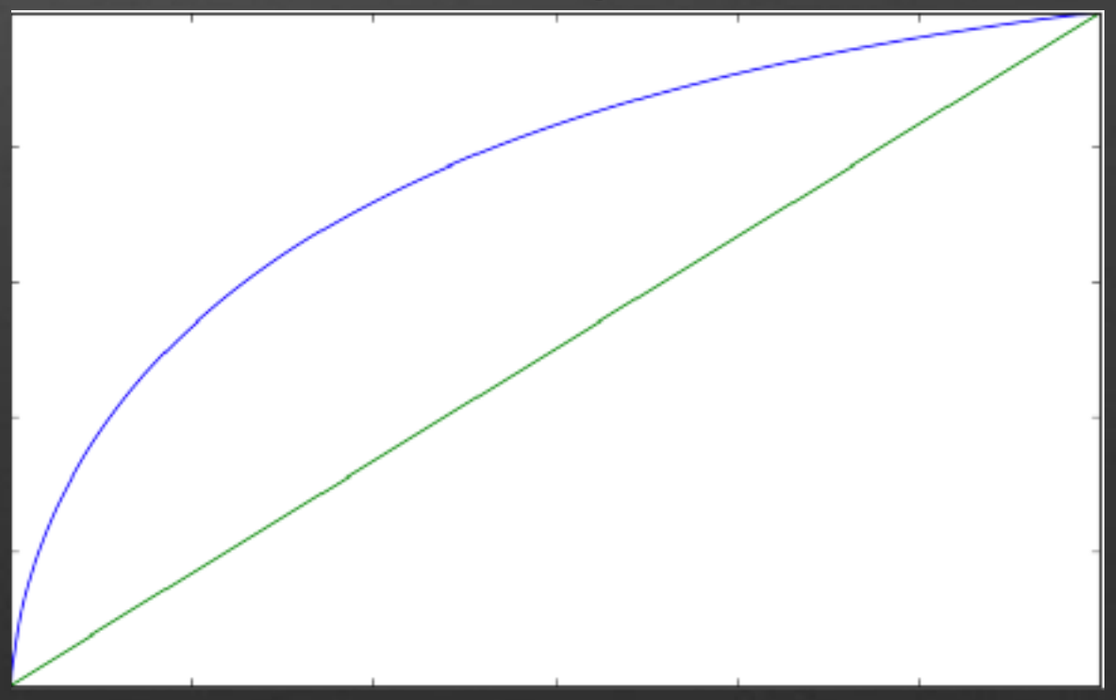
- 90 Languages
- Trained on Wikipedia
- Fasttext Algorithm
- ~1,000,000 Vocabulary Size
- 300 Vector Size

German

Distribution of L2 Norms



PCA Cumulative Explained Variance

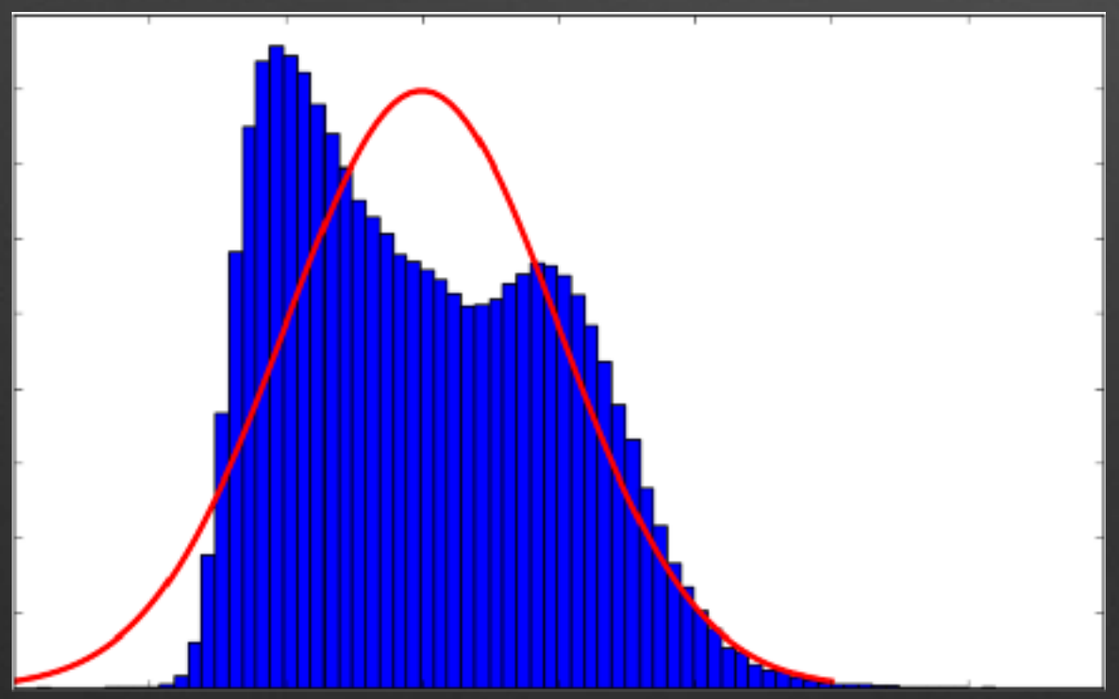


Zeroshot

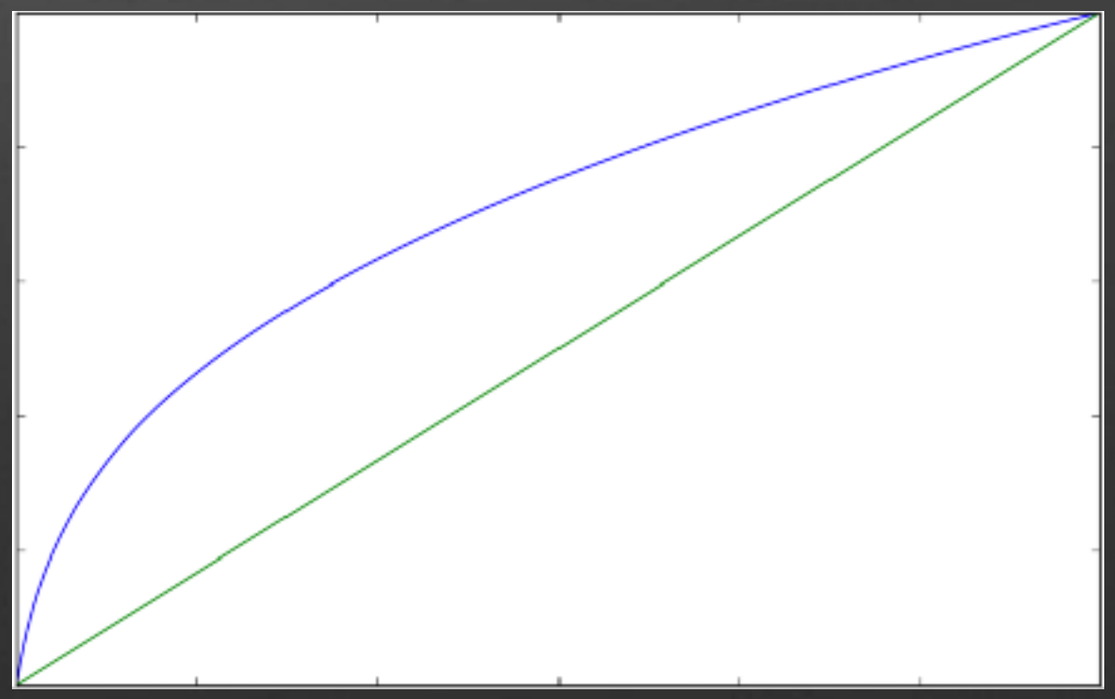
- 2 Languages
- Trained on British National Corpus and WaCky
- Word2Vec Algorithm
- 200,000 Vocabulary Size
- 300 Vector Size

Italian

Distribution of L2 Norms



PCA Cumulative Explained Variance



Bilingual dictionaries

- Google Translate
  - English -> German
  - German -> English
  - English -> Russian
  - Russian -> English

- Georgiana Dinu (Zeroshot)
  - English -> Italian

Translation methodology

- Find a translation matrix, *T*, by minimizing the mean squared error:

$$min_T \sum_{i=1}^{n_{Train}} \|v_{1,i}T - v_{2,i}\|^2$$

EXPERIMENTAL DESIGN

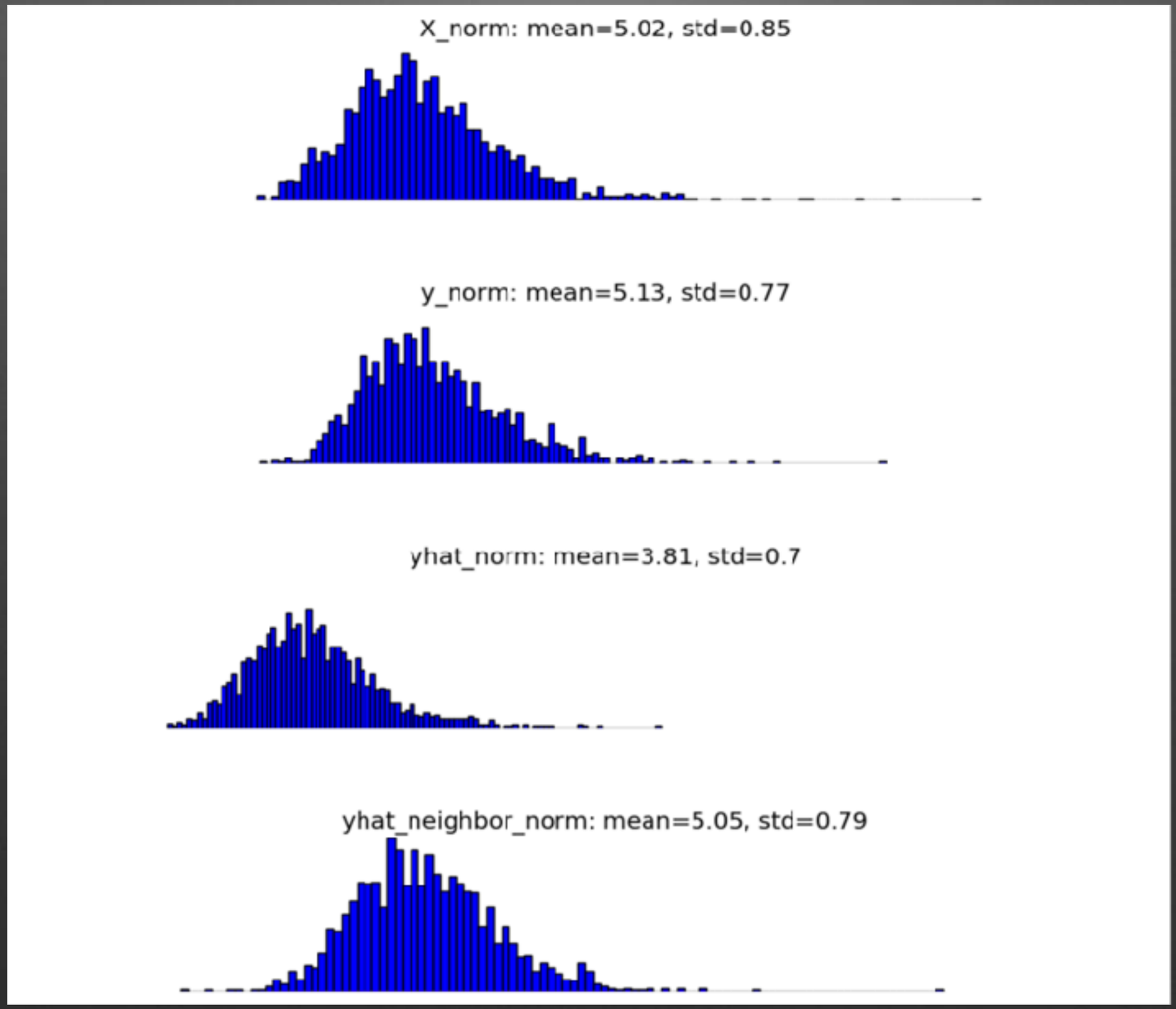
	Embedding	Translation	Train/Test Size	Train/Test Sampling	Max Neighbors
Experiment Label					
A	Fasttext	Enlish -> Russian	5,000 / 1,500	Random sample	944,211
B	Fasttext	English -> Russian	5,000 / 1,500	Random sample most frequent words	944,211
C	Fasttext	English -> German	5,000 / 1,500	Random sample	1,137,616
D	Fasttext	English -> German	5,000 / 1,500	Random sample most frequent words	1,137,616
E	Zeroshot	English -> Italian	5,000 / 1,869	Stratified sample word-frequency buckets	200,000

Experimental results

Accuracy

- A: 3.9%
- B: 46.4%
- C: 21.9%
- D: 63.6%
- E: 27.9%

Vector norms



Isotropy

- A: 0.32
- B: 0.38
- C: 0.36
- D: 0.43
- E: 0.47

Conclusion