Exploiting Bilingual Word Embeddings to Establish Translational Equivalence

Tobias Eder

Übersetzung ohne Wörterbuch

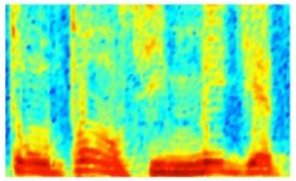
- Übersetzung auf bestimmter Domain
- Unbekannte Wörter im Text?
- Ohne Wörterbuch keine Übersetzung
- Domain-abhängige Übersetzung seltener Wörter

Übersicht

1. Motivation
2. Word Embeddings
3. Vektorraummodelle
   - Word2Vec
   - FastText
4. Lineare Abbildungen
5. Korpora und Experimentaufbau
6. Weitere Schritte / Regularisierung
7. Literaturangaben

Word Embeddings

AUDIO



Audio Spectrogram

DENSE

IMAGES

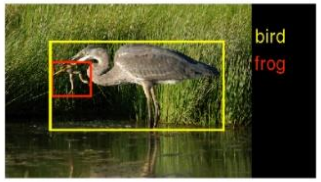


Image pixels

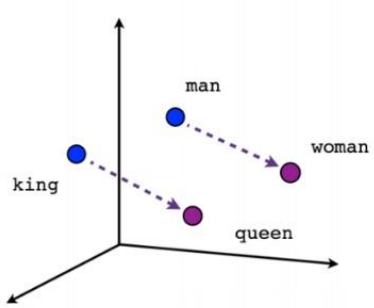
DENSE

TEXT

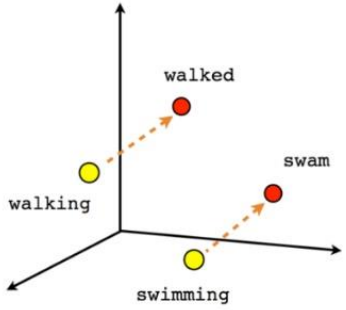
0	0	0	0.2	0	0.7	0	0	0	...	...

Word, context, or document vectors

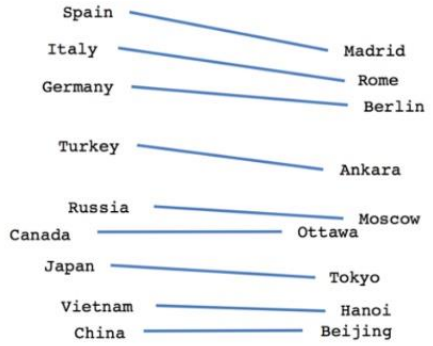
SPARSE



Male-Female

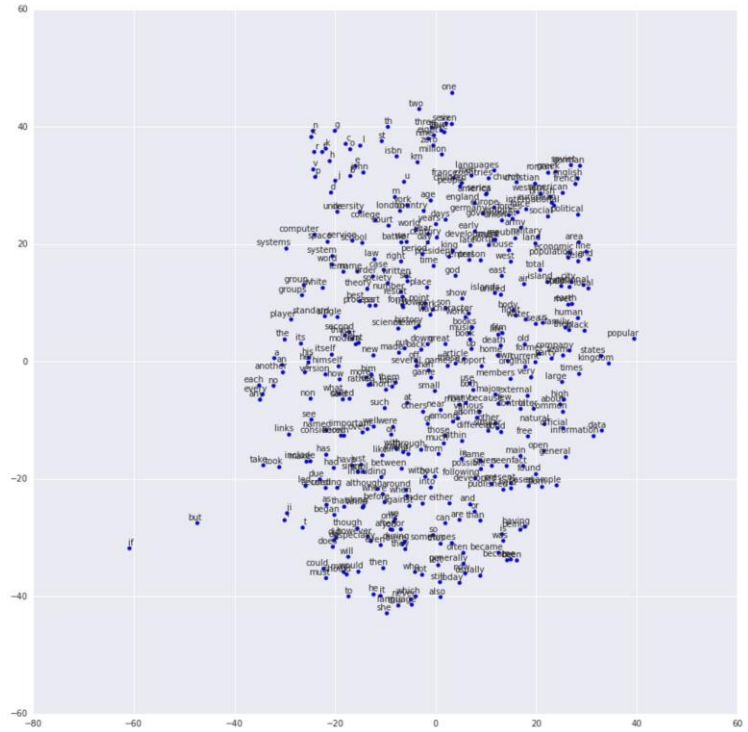


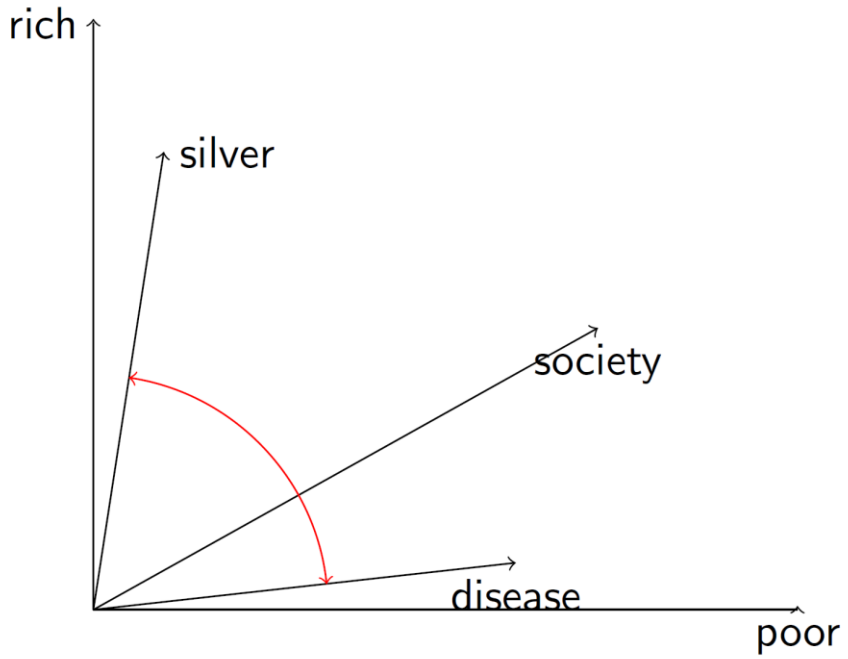
Verb tense



Country-Capital

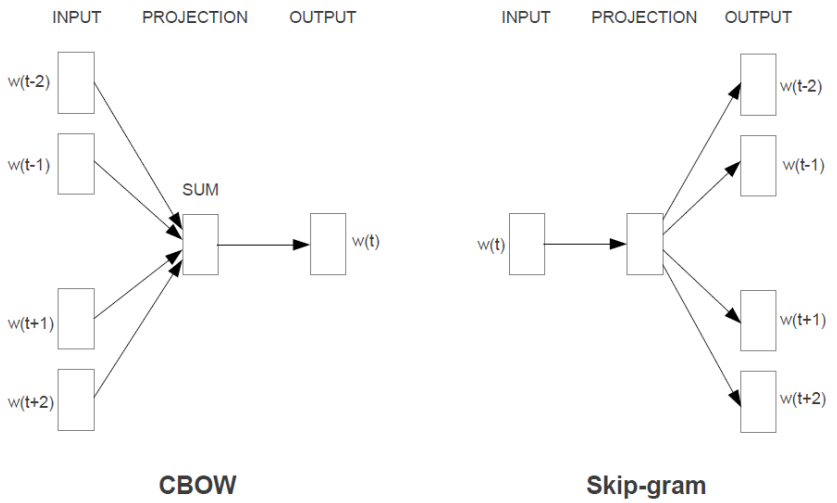
Word Embeddings





Word2Vec

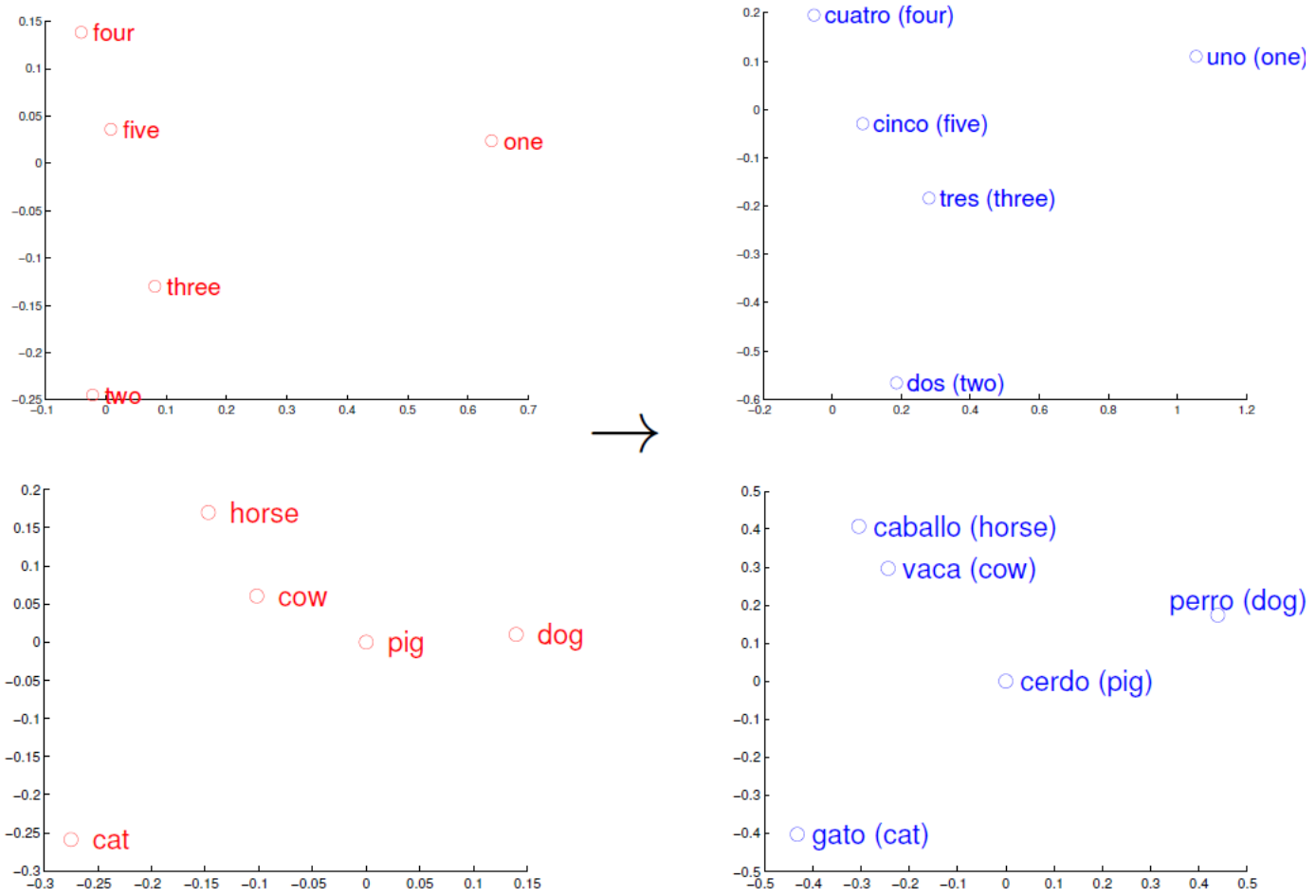
- Google (2013)
- Word-Embedding Toolkit
- CBOW und Skipgram Modelle



fastText

- Facebook Research (2016)
- Word-representation learning
  - Mit Subword-Information (Buchstaben n-Gramme)
- Word-vectors für OOV Wörter
- Textklassifikation mit linearem Modell

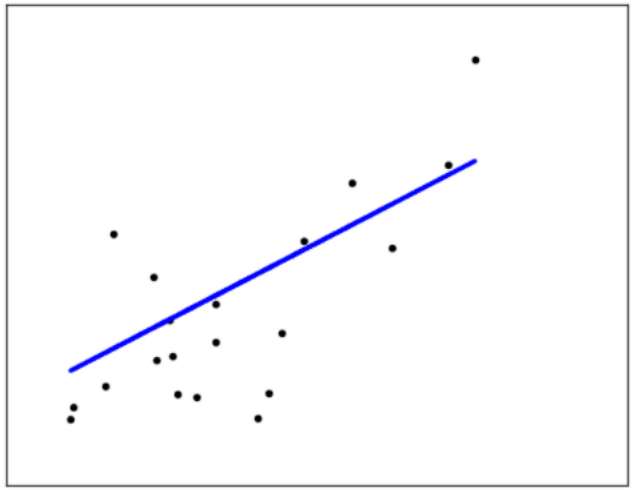
Lineare Abbildungen



Lineare Abbildungen

Lineare Regression:

$$\min_w ||Xw - y||_2^2$$



Ridge Regression (L2-Regularisierung):

$$\min_w ||Xw - y||_2^2 + \alpha||w||_2^2$$

Korpora und Experimentaufbau

- Vier unterschiedliche parallele Korpora:
  - General (ca 110M Tokens)
  - Medical Big (ca 50M Tokens)
  - EMEA (ca 4M Tokens)
  - TED Talks (ca 2M Tokens)
- Unterschiedliche Embeddings (CBOW, Skipgram)
- Übersetzung Englisch – Deutsch
- Kleiner paralleler Korpus (ca. 5000 Wörter)

Korpora und Experimentaufbau

- Auswahl an Worten aus Korpus (ca 1000 hochfrequente)
- Abbildung mit Regressions-Modell
- Domänespezifische Testsets
- Unterschiedliche Performance der Modelle

Weitere Schritte

- Niedrigfrequente Wörter?
- Bessere Abbildungen?
- Andere Regularisierungsmethoden?
- Evaluation auf OOV-Wörtern in fastText

Literaturangaben

- Bojanowski, Piotr; Grave, Edouard; Joulin, Armand; Mikolov, Tomas: “Enriching Word Vectors with Subword Information”. In: arXiv:1607.04606. 2016.
- Ishiwatari, Shonosuke; Kaji, Nobuhiro; Yoshinaga, Naoki; Toyoda, Masahi; Kitsuregawa, Masaru: “Accurate Cross-lingual Projectio between Count-based Word Vectors by Exploiting Translatable Context Pairs”. In: Proceedings of the 19th Conference on Computational Language Learning. 2015.
- Mikolov, Tomas; Le, Quoc V; Sutskever, Ilya: “Exploiting Similarities among Languages for Machine Translation”. In: arXiv:1309.4168. 2013
- Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey: “Efficient Estimation of Word Representation in Vector Space”. In: Proceedings of Workshop at ICLR. 2013.
- Mikolov, Tomas; Yih, Wen-tau; Zweig, Geoffrey: “Linguistic Regularities in Continuous Space Word Representations”. In: Proceedings of NAACL-HLT. 2013.