Word embeddings

German Rigau i Claramunt

german.rigau@ehu.es

IXA group

Departamento de Lenguajes y Sistemas Informáticos

UPV/EHU

Word embeddings
Summary

- Distributional hypothesis:
  - “words that are used and occur in the same contexts tend to purport similar meanings” (Harris 1954)
  - “a word is characterized by the company it keeps” (Firth 1957)
- Distributional semantics:
  - Word Space models: vectors for representing words
    - Latent Semantic Analysis (LSA),
    - Singular Value Decomposition (SVD), random indexing,
    - Word embeddings
  - One word is a position in a n-dimensional *space*.
    - Similarity (cosine similarity, etc.)

Word embeddings
Tools

- word2vec (Google)
  - https://code.google.com/archive/p/word2vec/
- GloVe (Stanford)
  - https://nlp.stanford.edu/projects/glove/
- FastText (Facebook)
  - https://github.com/facebookresearch/fastText
  - https://embeddings.sketchengine.co.uk
- Language projections (UPV/EHU):
  - https://github.com/artetxem/vecmap

Word embeddings
Tools

- S-Space
  - https://github.com/fozziethebeat/S-Space/
- Semantic Vectors
  - https://github.com/semanticvectors/semanticvectors/
- Gensim
  - http://radimrehurek.com/gensim/index.html
- DISCO:
  - http://www.linguatools.de/disco/disco-builder.html
- Indra:
  - https://github.com/Lambda-3/Indra

Word embeddings

German Rigau i Claramunt
german.rigau@ehu.es
IXA group
Departamento de Lenguajes y Sistemas Informáticos
UPV/EHU