



NLP Lab 3

Wrap up & Word Embeddings

Agenda

- CV: Reminder
- Sentiment Analyse besprechen
- Text Generation besprechen
- Word Embeddings

CV: Reminder

- Es haben noch nicht alle die Lösung der letzten CV-Aufgabe abgegeben
- ⇒ Notebooks bitte an Stanislav schicken!
stanislav.frolov@inovex.de

Sentiment Analyse

- kurze Vorstellung der Datenaufbereitung
- Evaluationsergebnisse

Text Generation

- kurze Vorstellung aktueller Stand
- Fortsetzung heute (ggf. parallel zu Embeddings)

Word Embeddings

Beispiel

Bellend rennt ... der Katze hinterher.

Word Embeddings

- Bisher: Repräsentation von Wörtern als one-hot-encoded Vektoren
- Word Embeddings:
 - reellwertige Vektoren mit N Dimensionen
 - Komponente im Vektor bildet Eigenschaft des Wortes ab
- wichtige Publikation:
 - Efficient Estimation of Word Representations in Vector Space – Mikolov et al. 2013

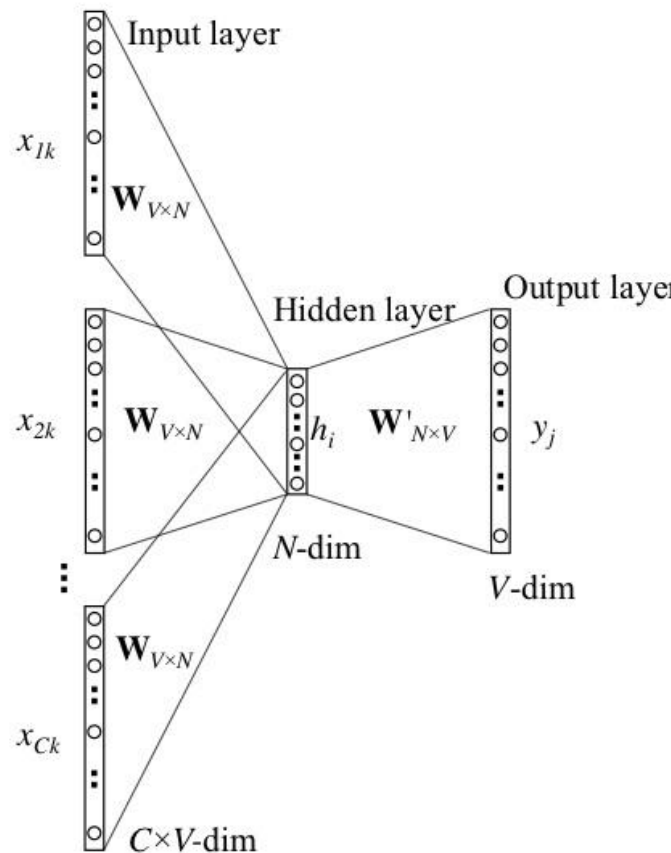
Word Embeddings

- Wie kann man Wörter auf Vektoren abbilden?
- Trick:
 - Fake Task
 - (einfaches) neuronales Netz, das Wörter aus ihrem Kontext vorhersagt
 - Vektoren werden aus den Gewichtsmatrizen des neuronalen Netz extrahiert
 - Ergebnisse interessieren eigentlich nicht

Word Embeddings

continuous bag-of-words Model

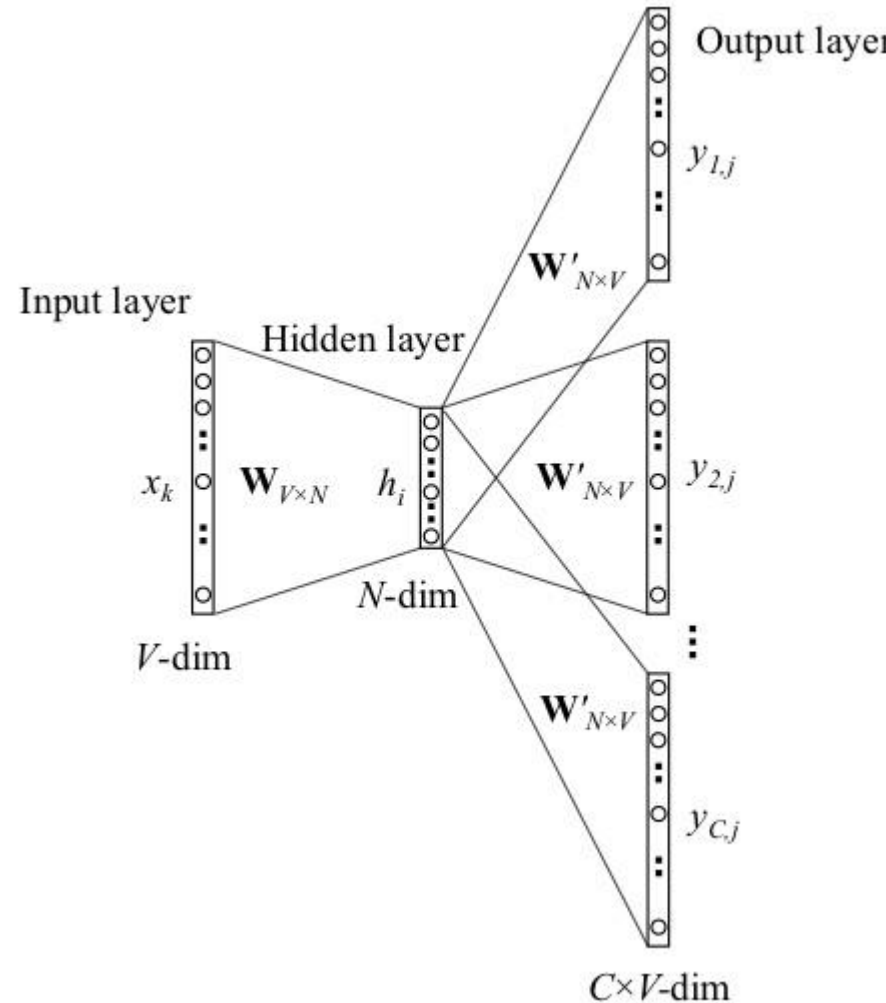
- Input layer:
 - Kontextwörter als Vektoren
- Hidden layer:
 - N Dimensionen
 - Durchschnitt über Inputvektoren
- Output layer:
 - Wort als Vektor
- Loss: Softmax



Word Embeddings

Skip Gram Model

- Input
 - Wort als Vektor
- Hidden layer:
 - N Dimensionen
- Output:
 - $C \times V$ dimensionaler Vektor
- Loss: Softmax



Word Embeddings

Implementierungen

- GloVe
 - <https://nlp.stanford.edu/projects/glove> (Stanford)
- FastText
 - <https://fasttext.cc> (Facebook)
- Word2Vec
 - <https://code.google.com/archive/p/word2vec> (Google)
- Gensim
 - <https://radimrehurek.com/gensim/models/word2vec.html>

<https://github.com/hskaailabnlp/embeddings>