Bisher bei der Textanalyse…

Vektorisierung mit Wörtern als Features

- Reihenfolge nicht beachtet
- Zusammenhänge nicht betrachtet
- Semantik geht verloren

Vektorisierung mit N-Grammen als Features

- Reihenfolge beachtet
- Zusammenhänge in Form von Tupeln
- Abstraktion in Form von Semantik fehlt

Worte jeweils einzelne Entitäten

Kontext entscheidet über Semantik!

Distributional Hypothesis (Firth, 1957)

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

Beispiel: Was ist "tezgüino"? Was ähnelt "tezgüino"?

A bottle of _____ is on the table.
Everybody likes _____.
Don’t have _____ before you drive.
We make _____ out of corn.

EISENSTEIN, JACOB: *Natural Language Processing*. Georgia Tech, 2018. Ch. 14

Gesucht: Semantischer Zusammenhang von Worten

Dazu anderes Modell notwendig

- Worte können nicht einfach durchnummeriert werden
- Nummerierung beliebig
- Enthält zu wenig Informationen
- “Abstände” sind unbedeutend

Darstellung der Worte als “Wortvektoren”

- Niedrigdimensionaler Vektorraum (100 – 300 Dimensionen)
- Abstände und Ähnlichkeiten definiert
- “Vektorgleichungen”
- Einfacher als neuronales Netz

Konstruktionsversuche

Naiv: 1-Hot-Encoding

- Bei jedem Vektor immer nur ein 1 bei dem entsprechenden Wort setzen
- Kontext geht verloren
- Kein Ähnlichkeitsmaß

Stattdessen

- Neuronales Netz auf 1-Hot-Vektor trainieren
- Entspricht einem “verteilten” Wortvektor
- Trainingsprozess ist aufwändig

Trainingsziel und Gütefunktion

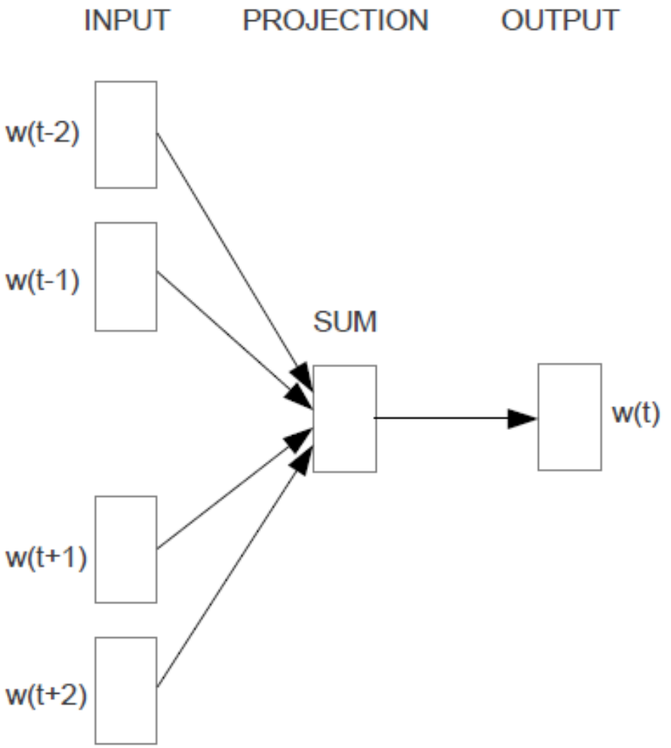
Ziel: Kontext zwischen den Worten herstellen

CBOW-Modell

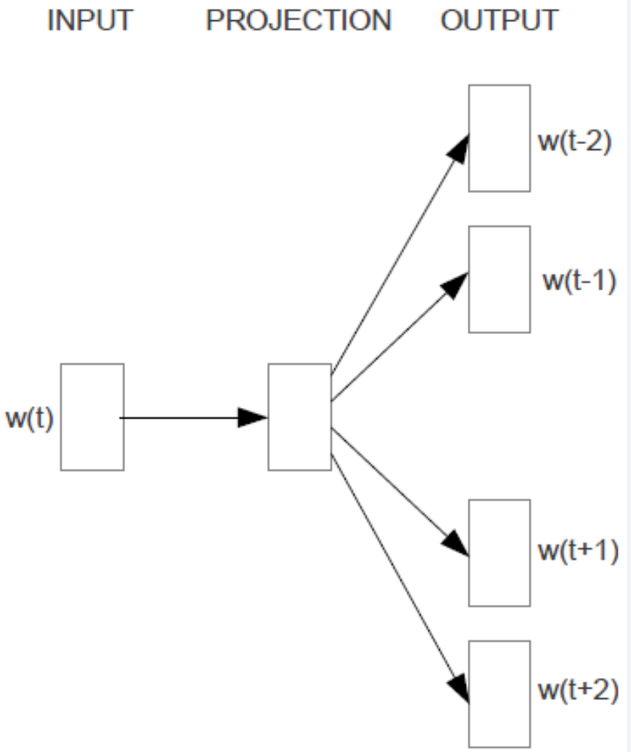
- Aus Kontext Wort bestimmen

Skipgram-Modell

- Aus Wort Kontext bestimmen
- Langsamer, genauer bei seltenen Worten



CBOW



Skip-gram

Eigenschaften der Wortvektoren

Ähnlichkeiten

- “Winkel” (Skalarprodukt) als Ähnlichkeitsmaß verwenden
- Semantische und syntaktische Ähnlichkeiten

Vektorgleichungen

- Beispiel: Differenzvektoren entsprechen dem Unterschied zwischen Mann und Frau