Corpus based Identification of Text Segments

Thomas Ebert
Betreuer: MSc. Martin Schmitt

Übersicht

1. Motivation
2. Ziel der Arbeit
3. Vorgehen
4. Probleme
5. Evaluierung
6. Erkenntnisse und offene Fragen

Motivation

Motivation

- Textsegment $\rightarrow$ bedeutungstragende Einheit
- Morphem
- Wort
- Phrase
- Satz
- Topic (Thema eines Abschnitts)

Motivation

- Textaufbereitung für NLP-Aufgabe meist wortbasierend (Tokenisierung)
- Wort ist nicht eindeutig definiert aber intuitiv
- Tokenisierung ist sehr fehleranfällig, lokale Anpassungen nötig
- Ist das intuitive Konzept "Wort"die beste Art für einen Computer einen Text zu segmentieren?

Ziel der Arbeit

Ziel der Arbeit

- Entwicklung eines Algorithmus', der einen eingegebenen Satz oder Text in seine 'besten' Segmente (Buchstaben N-Gramme) zerlegt.
- Ist der nicht-symbolische Ansatz besser als der wortbasierte Ansatz?
- Welche Chancen und Risiken bietet der nicht-symbolische Ansatz?

Vorgehen

Vorgehen

- Extrahieren von N-Grammen der Länge 1 bis 10 aus dem Wikipedia Korpus (Englisch)
- Korpus enthält unannotierte Rohtexte
- Erste 10.000 Texte (22.650.880 Zeichen) des Korpus werden zum extrahieren verwendet

Vorgehen

- Frequenzliste für die N-Gramme wird erstellt
- N-Gramme werden mit einem Gütemaß bewertet
- Gütemaß = n*log(freq)
- n = N-Gramm-Länge
- freq = absolute Häufigkeit des N-Gramms

Vorgehen

- Zum Testen wird ein Satz eingegeben
- Der Satz wird in die N-Gramme mit den höchsten Gütemaßen zerlegt.

Probleme

Probleme

- Mit der Größe der Eingabe, steigt die Laufzeit exponentiell
- Lösung: heuristischer Ansatz
- Größe des Fensters (Window) festlegen
- Berechnung der höchsten Güte ist nicht mehr garantiert, aber Segmentierung ist ggf. noch besser als bei symbolischem Ansatz

Evaluierung

Evaluierung

- Evaluierung von Text Segmenten ist schwierig
- Häufig Uneinigkeit über die Granularität von Segmenten
- Je nach Anwendung können Fehler relevant oder irrelevant sein
  z.B. bei IR kann die Korrektheit von Segmentgrenzen vernachlässigt werden,
  bei "news boundary detection"nicht.
- Auswirkung auf die Endanwendung (z.B. IR, Sentiment Analysis) wird als Maß verwendet.

Evaluierung

- Verwendung von word2vec um Buchstaben N-Gramm embeddings zu erhalten
- Sentiment Analyse auf Satzebene zur Evaluation
- Verwendung von Movie Review Data
- Vergleich mit Word embeddings

Evaluierung

- Mögliches Modell: (Cho et al., 2014)
  Sigmoid auf dem letzten Zustand eines LSTM-Encoders
  LSTM (Long-Short-Term Memory)
- Die Sigmoidfunktion wird auf die Summe der gewichteten Eingabewerte angewendet um ein Ergebnis zu erhalten.

Erkenntnisse und offene Fragen

Erkenntnisse und offene Fragen

- Auch Buchstaben NGramme weisen ein Zipfsche Verteilung auf
- Häufigste N-Gramme größer 3 enthalten Funktionswörter
- Häufigste N-Gramme größer 8 enthalten Inhaltswörter

Erkenntnisse und offene Fragen

- Noch keine Ergebnisse für die Evaluierung vorhanden
- Andere Möglichkeit um N-Gramme zu extrahieren?
- Ist das Ergebnis der Evaluierung schon aussagekräftig?

References I

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio.
Learning phrase representations using rnn encoder-decoder for statistical machine translation.
arXiv preprint arXiv:1406.1078, 2014.

F. Y. Choi.
Advances in domain independent linear text segmentation.
In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33. Association for Computational Linguistics, 2000.

J. C. Reynar.
Topic segmentation: Algorithms and applications.
IRCS Technical Reports Series, page 66, 1998.

References II

H. Schuetze.
Nonsymbolic text representation.
arXiv preprint arXiv:1610.00479.