Natural Language Processing

SoSe 2017

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Tokenization and Segmentation

Dr. Mariana Neves

May 15th, 2017

Tokenization

- Separation of words in a sentence

„Latest figures from the US government show the trade deficit with China reached an all-time high of $365.7bn (£250.1bn) last year. By February this year it had already reached $57bn.“

„Latest figures from the US government show the trade deficit with China reached an all time high of $ 365.7 bn (£ 250.1 bn) last year . By February this year it had already reached $ 57 bn .“

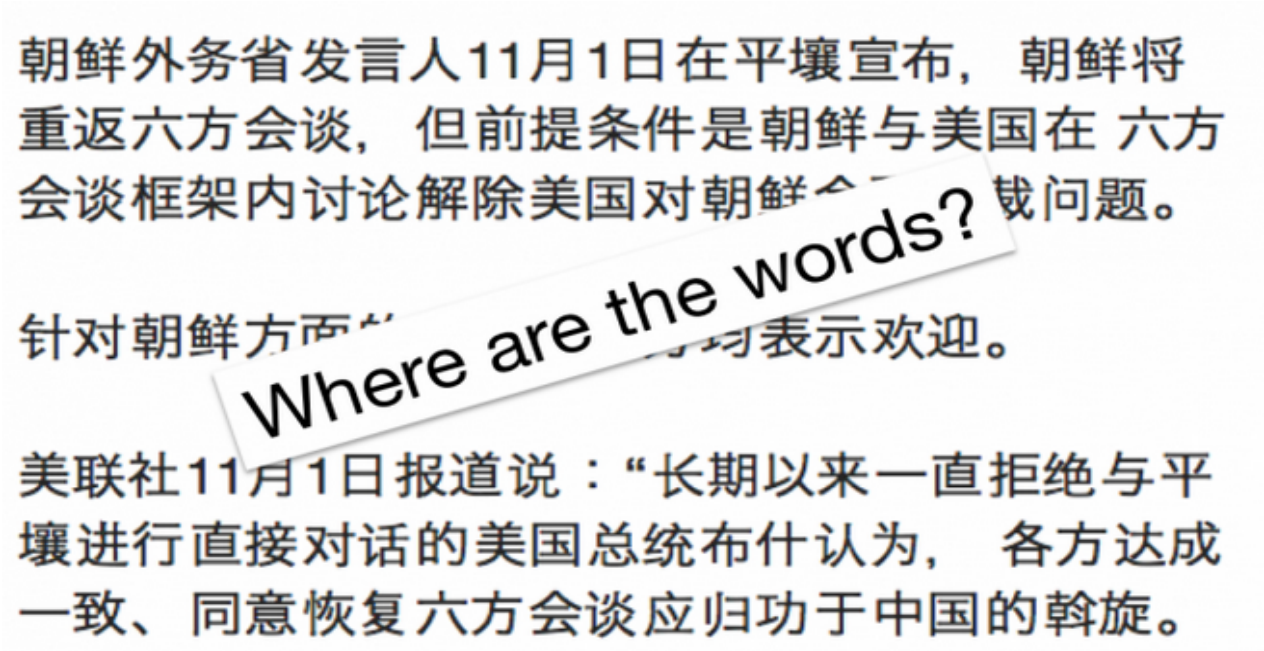
(http://www.bbc.com/news/election-us-2016-36185012)

Tokenization

- Issues related to tokenization:
  - Separators: punctuations
    - Exceptions: „m.p.h“, „Ph.D“
  - Expansions: „we're“ = „we are“
  - Multi-words expressions: „New York“, „doghouse“

Segmentation = Tokenization

- Word segmentation: separation of the morphemes but also tokenization for languages without 'space' character



(http://www.basistech.com/better-multilingual-search/)

Segmentation?

Sentence separation (splitting)

- Also usually based on punctuations (.?!)
  - Exceptions: „Mr.“, „4.5“

Approaches for tokenization

- Based on regular expressions

- Based on rules or machine learning
  - Binary classifers that decides whether a certain punctuation is part of a word or not

Approaches for segmentation

- Maximum matching approach
  - Based on a dictionary
  - Longest sequence of letters that forms a word

- Palmer (2000):

thetabledownthere

thetabledownthere

thetabledownthere

thetabledownthere

Neural networks for segmentation

- Chinese word segmentation formalized as a character-based sequence labeling task where only contextual information within fixed sized local windows and simple interactions between adjacent tags can be captured.

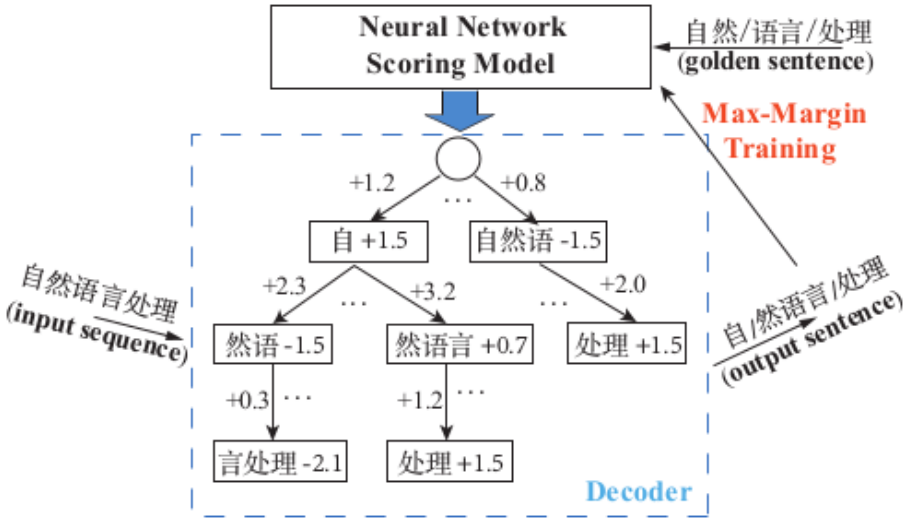
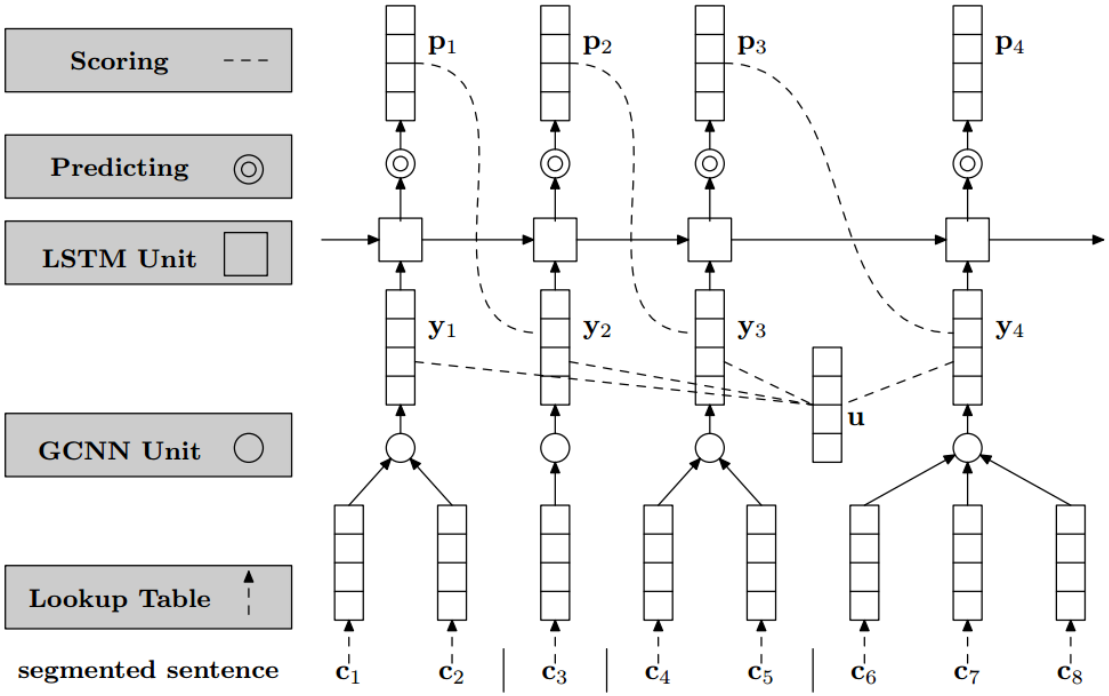


Figure 1: Our framework.

(https://arxiv.org/abs/1606.04300)

Neural networks for segmentation



(https://arxiv.org/abs/1606.04300)

Tools for tokenization

- Spacy: https://spacy.io/
- OpenNLP: https://opennlp.apache.org/
- Stanford CoreNLP: https://stanfordnlp.github.io/CoreNLP/
- deeplearning4j tokenizer: https://deeplearning4j.org/tokenization
- Neural tokenizer: https://github.com/Kyubyong/neural_tokenizer

Exercise

- Project: choose a tokenizer and try it in your document collection.
  - Manually check a sample of the results.

Further reading

- NLP book: Chapter 3