NLP Resources for Dutch

Nov 24, 2022

Contents:

CHAPTER 1

Tools

List of NLP software for Dutch.

- **NLTK**
  - Good start for notebook(?): https://www.datacamp.com/community/tutorials/stemming-lemmatization-python
- SpaCy
- Pattern
- Flair

CHAPTER 2

Datasets

List of labeled datasets.

- Universal Dependencies
- 110k Dutch Book Reviews Dataset (110kDBRD)

CHAPTER 3

Models

List of pretrained word embeddings and other models.

3.1 Word embeddings

- fastText: fastText word vectors trained on Common Crawl and Wikipedia.
- Word2Vec: Word2Vec vectors trained by CLIPS on diffent Dutch corpora.
- Word2Vec: Word2Vec vectors trained by the Nordic Language Processing Laboratory (NLPL) on the CoNLL17 corpus.
- ConceptNet Numberbatch: multilingual word embeddings in the same semantic space. Built using an ensemble that combines data from ConceptNet, word2vec, GloVe, and OpenSubtitles 2016, using a variation on retrofitting.

3.2 ULMFiT

- Leiden fastai ULMFiT model: Trained on Wikipedia by the Text Mining and Retrieval research group at Leiden University.

3.3 BERT

- BERTje: Dutch pre-trained BERT model developed at the University of Groningen.
- RobBERT: Dutch BERT model using RoBERTa's pre-training, developed by KU Leuven. Trained on scraped data from the Dutch section of the OSCAR corpus. 4 smaller, distilled models are also available.
- BERT-NL: cased and uncased BERT model trained on the SoNaR corpus by the Text Mining and Retrieval research group at Leiden University.

- mBERT: Multilingual BERT model.

CHAPTER 4

Corpora

List of (largely) unlabeled text collections.

- SoNaR-500: over 500 million words, from different domains and genres, automatically tokenized, pos-tagged, lemmatised and named entities extracted.
- SoNaR-1: 1 million words, largely from SoNaR-500, with manually annotated named entities, corefences, spatial and temporal relations.
- SoNaR Nieuwe Media Corpus 1.0: tweets, chat and sms messages.

CHAPTER 5

Indices and tables

- genindex
- modindex
- search