TODAY'S MENU

CWS: EVALUATION AND LEXICON-BASED SYSTEMS

- Some *historical* aspects
- Evaluation and Bakeoffs
- DIY maximum forward matching segmentation

"WORDHOOD", THE BIG PICTURE

IN LINGUISTICS

- an old but still **open debate**
  - 字／詞 distinction attributed to (章士釗, 1907)
- about **how** and **what** to define
  - For Modern Chinese (Kratochvil, 1967) (Duanmu, 98) (Packard, 2000) (N'Guyen, 2006)
  - But not only ! (Haspelmath, 2011)
- even about specific items
  - 以前／以後 (Liu, 2013)

IN LANGUAGE TECHNOLOGIES AND NLP

- (Wu, 2003)
  However, **we do not have to wait for linguists to reach a consensus before we do segmentation in NLP**. In computer applications, we are more concerned with "segmentation units" than "words". While words are supposed to be well-defined, unambiguous and static linguistic entities, segmentation units are not.
- but this appently practical approach sweeps many issues under the carpet

SHORT CHRONOLOGY

- no resources
  - only unsupervised methods (Sproat & Shih, 1990)
- electronic dictionnaries
  - lexicon-based is possible
- Proper evaluation was needed
  - segmented corpora made available
  - (questionable?) guidelines
- Machine Learning was possible
  - (at some point the linguists disappeared)
- What is next ?
  - (back to unsupervised methods ? imho, better for linguistics)

SEGMENTATION BAKEOFFS

various available corpora for training and evaluation, six bakeoff where held between 2003 and 2012

- The first two demonstrated the supremacy of Machine Learning
- The third tried to focus on Named Entities and OOVs
- The fourth added tagging
- The fifth explored domain adaptation (without much success of semi-supervised learning)
- the sixth focused on micro-blogging

EVALUATION METRICS

from Information Retrieval

- Boundary-based F-score
- Word-based F-score

(See printed document)

LONGUEST-MATCH SEGMENTATION

is one of the simpler yet quite efficient lexicon-based algorithms (given a good lexicon)

- used as topline and baseline in the Bakeoffs

```
1) start at the begining of the corpus
2) use the lexicon to find the longuest word that can be matched at current p
osition
3) consider it to be the next token
4) move to the end of the next token in the corpus
5) go back to 2) until you reach the end of the corpus
```

(manual) demo with Unitex

YOUR TURN !

- Write a script to run the evaluation
- Write a longuest-match segmentation program, try it with
  - 萌典 lexicon
  - training data
  - testing data
- comment on the scores

if time allows

- compare with zpar
- try to modify your script to get better results (should be done on dev set)