A General-Purpose Machine Learning Method for Tokenization and Sentence Boundary Detection

Valerio Basile Johan Bos Kilian Evang

University of Groningen
{v.basile,johan.bos,k.evang}@rug.nl

Computational Linguistics in the Netherlands 2013

Tokenization: a solved problem?

- Problem: tokenizers are often rule-based: hard to maintain, hard to adapt to new domains, new languages
- Problem: word segmentation and sentence segmentation often seen as separate tasks, but they inform each other
- Problem: most tokenization methods provide no alignment between raw and tokenized text (Dridan and Oepen, 2012)

Research Questions

- Can we use machine learning to avoid hand-crafting rules?
- Can we use the same method across domains and languages?
- Can we combine word and sentence boundary detection into one task?

Method: IOB Tagging

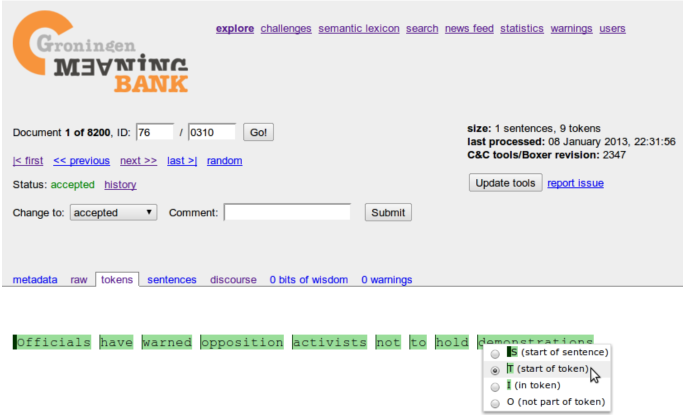
- widely used in sequence labeling tasks such as shallow parsing, named-entity recognition
- we propose to use it for word and sentence boundary detection
- label each character in a text with one of four tags:
  - **I:** inside a token
  - **O:** outside a token
  - B: two types
    - **T:** beginning of a token
    - **S:** beginning of the first token of a sentence

IOB Tagging: Example

```
It didn't matter if the faces were male,
SIOTIITIIOTIIIIIOTIOTIIOTIIIIOTIIIOTIIITO
female or those of children. Eighty-
TIIIIIOTIOTIIIIOTIOTIIIIIIITOSIIIIIIO
three percent of people in the 30-to-34
IIIIIOTIIIIIIOTIOTIIIIIOTIOTIIOTIIIIIIIO
year old age range gave correct responses.
TIIIOTIIOTIIOTIIIIOTIIIOTIIIIIIOTIIIIIIIIT
```

- Note: discontinuous tokens are possible (Eighty-three)

Acquiring Labeled Data: correcting a Rule-Based Tokenizer



Method: Training a Classifier

- We use Conditional Random Fields (CRF)
- State of the art in sequence labeling tasks
- Implementation: Wapiti (http://wapiti.limsi.fr)

Features Used for Learning

- current Unicode character
- label on previous character
- different kinds of contexts:
  - either Unicode characters in the context
  - or Unicode categories of these characters
- Unicode categories less in number (31), but also less informative than characters
- context windows sizes: 0, 1, 2, 3, 4 to the right and left of current character

Experiments

- Three datasets (different languages, different domains):
  - Newswire English
  - Newswire Dutch
  - Biomedical English

Creating the Datasets

- (Newswire) English: **Groningen Meaning Bank** (manually checked part)
  - 458 documents, 2,886 sentences, 64,443 tokens
  - already exists in IOB format
- Newswire Dutch: **Twente News Corpus** (subcorpus: two days from January 2000)
  - 13,389 documents, 49,537 sentences, 860,637 tokens
  - inferred alignment between raw and tokenized text
- Biomedical English: **Biocreative1**
  - 7,500 sentences, 195,998 tokens (sentences are isolated, only word boundaries)
  - inferred alignment between raw and tokenized text

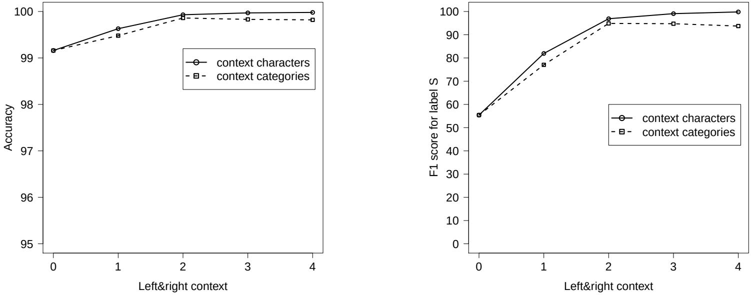
Baseline Experiment

- Newswire English without context features
- Confusion matrix:

		predicted label I	T	O	S
gold label	I	21,163	45	0	0
	T	26	5,316	0	53
	O	0	0	5,226	0
	S	4	**141**	0	123

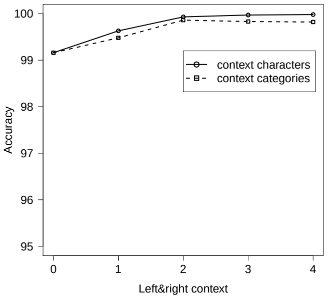
- Main difficulty: distinguishing between T and S

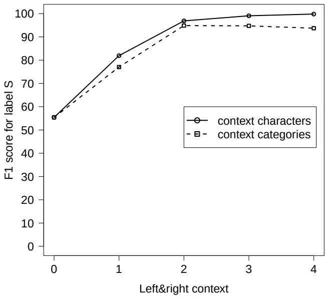
How Much Context Is Needed?



- results shown for GMB (trained on 80%, tested on 10% development set)
- performance almost constant after left&right window size 2

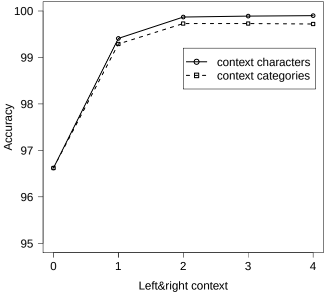
Characters or Categories?

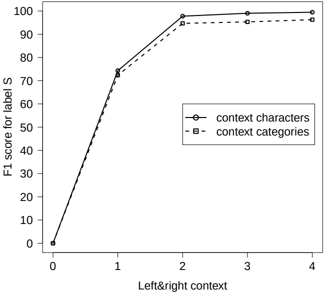




- character features perform well, categories overfit

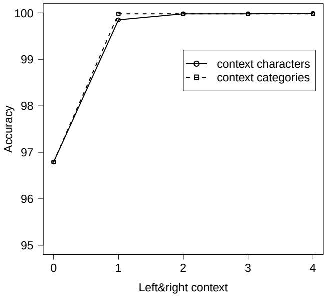
Applying the Method to Dutch

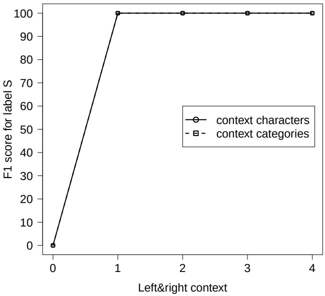




- results shown for TwNC (trained on 80%, tested on 10% development set)

Applying the Method to Biomedical English





- results shown for Biocreative1 (trained on 80%, tested on 10% development set)
- in this corpus: sentences isolated, sentence boundary detection trivial

What Kinds of Erros Does More Context Fix?

- examples from English newswire, 2-window vs. 4-window character models

```
context:  er Iran to the U.N. Security Council, whi
gold:     IIOTIIIOTIOTIIOTIIIOTIIIIIIIOTIIIIIIITOTII
2-window: IIOTIIIOTIOTIIOTIIIOSIIIIIIIOTIIIIIIITOTII
4-window: IIOTIIIOTIOTIIOTIIIOTIIIIIIIOTIIIIIIITOTII

context:  by Sunni voters. Shi'ite leaders have not
gold:     TIOTIIIIOTIIIIIITOSIIIIIIOTIIIIIIOTIIIOTII
2-window: TIOTIIIIOTIIIIIITOSIITIIIOTIIIIIIOTIIIOTII
4-window: TIOTIIIIOTIIIIIITOSIIIIIIOTIIIIIIOTIIIOTII
```

Examples of Errors Still Made by the Best Model

- examples from English newswire, 4-window character model
- probable causes: too simple features, not enough training data

context: ive arms race it can**n**ot win. Taiwan split
gold: IIIOTIIIOTIIIOTIOTII**T**IIOTIITOSIIIIIOTIIIII
4-window: IIIOTIIIOTIIIOTIOTII**I**IIOTIITOSIIIIIOTIIIII

context: ally paved with gold**?** **M**oses Bittok probab
gold: IIIIOTIIIIOTIIIOTIII**T**O**S**IIIIOTIIIIIOTIIIII
4-window: IIIIOTIIIIOTIIIOTIII**I**O**T**IIIIOTIIIIIOTIIIII

Is It Fast Enough?

- Tested on 4-core, 2.67 GHz desktop machine
- Training: around 1’30” for best model on 40,000 Dutch sentences
- Labeling: around 3,000 sentences/second

Future Work

- Compare with existing rule-based tokenizers
- Compare with existing sentence-boundary detectors
- Can we build universal models (trained on mixed-language, mixed-domain corpora)?
- Experiment with more complex features
- Software release

Conclusions

- Word and sentence segmentation can be recast as a combined tagging task
- Supervised learning: shift of labor from writing rules to correcting labels
- Learning this task with CRF achieves high speed and accuracy
- Our tagging method does not lose the connection between original text and tokens
- Possible drawback of tagging method: no changes to original text possible, e.g. normalization of punctuation etc.

References I

Dridan, R. and Oepen, S. (2012). Tokenization: Returning to a long solved problem — a survey, contrastive experiment, recommendations, and toolkit —. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–382, Jeju Island, Korea. Association for Computational Linguistics.