Kilian Evang
University of Groningen

Valerio Basile
University of Groningen

Johan Bos
University of Groningen

A Text Denormalization Algorithm Producing Training Data for Text Segmentation

Problem

raw text

```
una comunita' latina.
```

normalized text

```
una comunità latina .
```

?

only tokenized text

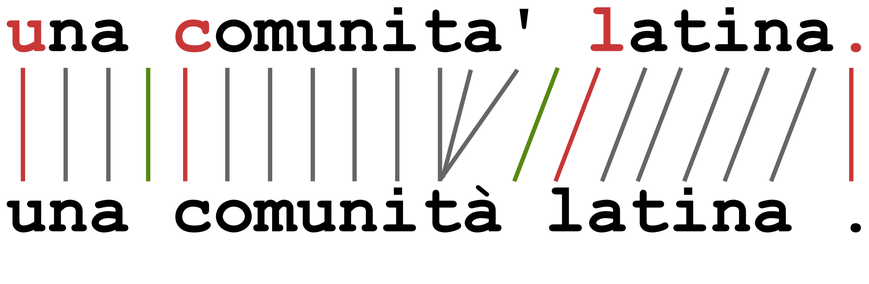
```
una comunita' latina .
```

Solution

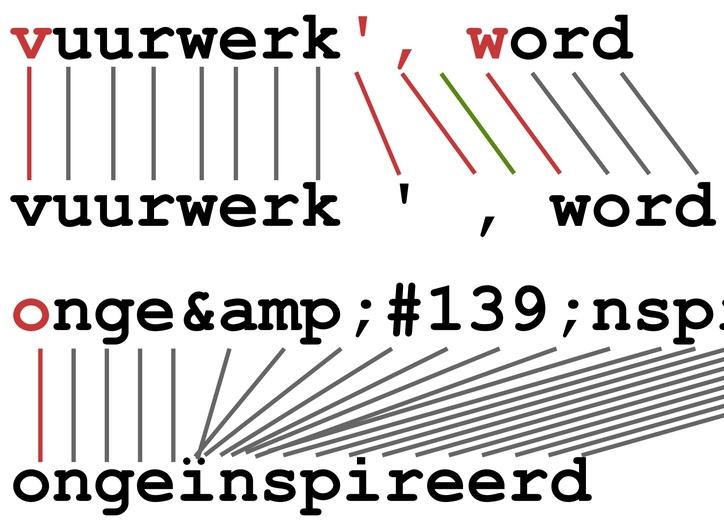
1. align Levenshtein matches

2. co-align heuristically

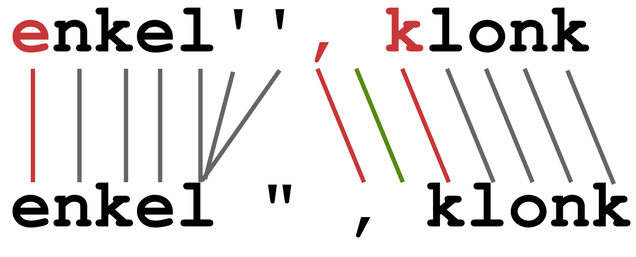
3. transfer boundaries

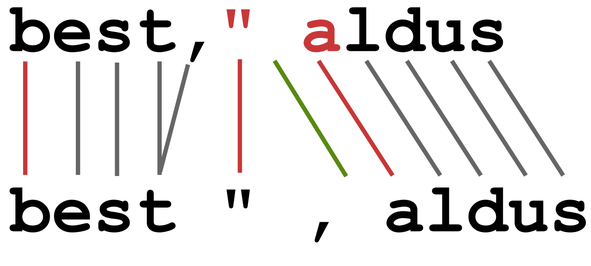


What works



What doesn’t work yet





Conclusion

- Algorithm transfers token boundaries form normalized to raw text
- Levenshtein algorithm provides principled starting point
- Some normalizations require additional heuristics
- Application: train statistical text segmentation tools (Evang et al. 2013)

https://github.com/texttheater/iobify