# Linguistic Annotation of Computer-Mediated Communication (not only) an Explorative Analysis

Kay-Michael Würzner
Lothar Lemnitzer
Alexander Geyken
Bryan Jurish

University of Potsdam

Berlin-Brandenburgische Akademie of Sciences and Humanities

# The Big Picture

- DERIK (Beißwenger et al. 2012)
  - Reference corpus for CMC
  - Representative and balanced
  - Joint project of University of Dortmund and BBAW

**Acquisition** Collecting data (in progress)

**Mark-Up** Standard format for different sources (i.e. `TEI`, Lemnitzer and Ermakova)

**Annotation** Adding linguistic information

# What we understand as *linguistic annotation*:

**Tokenizing** detection of word and sentence boundaries

**Lemmatization** identification of a word's base form and possible part of speech (PoS)

**PoS-Tagging** computation of a word's most probable PoS in context

→ Requirement for search engine indexing

# Tokenizing – Task

- Segmentation of continuous text into words (or *tokens*) and sentences
- Classification of certain classes of tokens
  - Abbreviations
  - Numbers
  - Special characters
  - Foreign alphabets
- Normalization of hyphenations

# Tokenizing – Challenges

- Reduced or inconsistent punctuation

  Hallo haben sie fragen zum BAföG

- Special vocabulary

  Wenn Ihre Antwort komplett ist, tippen Sie bitte am Ende ein *E*

- Sequences of special characters

  ??????????? was das

- Addressing of chat partners

  asset1, . . . ; keller > > . . . ; baloo: . . . ; @günni . . . etc.

# Tokenizing – Approaches

- Rule-based
  - Set of regular expressions defining tokens
  - Construction of deterministic finite-state automaton
  - Usually done by a *scanner generator* (i.e. `flex`, `re2c`)

```
[A-z]+          {return WORD;}
[0-9]+          {return INT;}
"." / [ \n\t]   {return SB;}
```

  - very fast but highly error-prone

# Tokenizing – Approaches II

- Statistical
  - Detection of likely token or sentence boundaries using context information
  - Kiss and Strunk (2006)
    - Unsupervised; trained on raw text
    - Initial, minimal rule-based step
    - Disambiguates · using an indepency assumption:

$$\boxed{\mathrm{P}(\cdot|w) = p = \mathrm{P}(\cdot|\neg w)}$$

    - Additional features as word length and the number of internal periods to refine decision
    - Used in `OpenNLP`

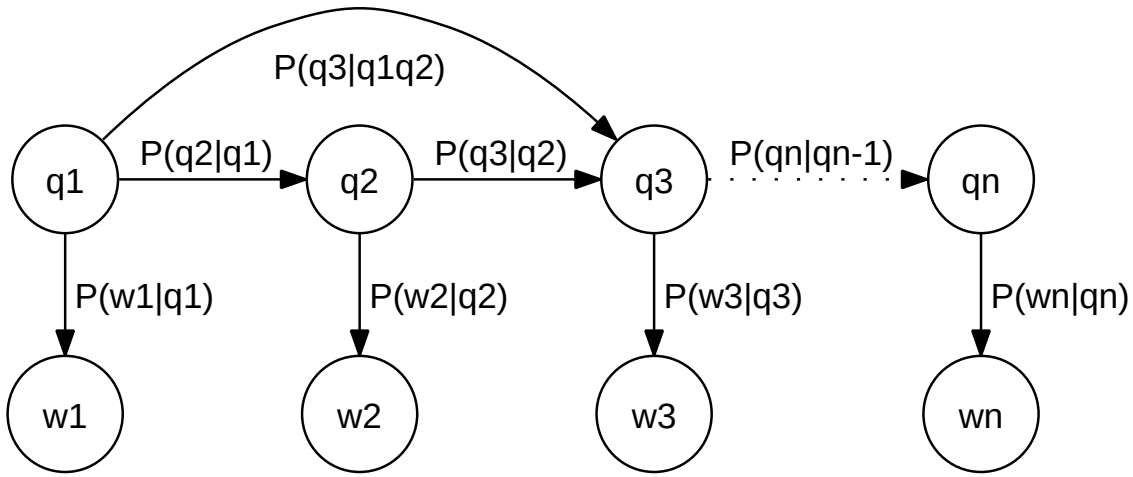
# Tokenizing – Approaches III

- Statistical
  - Laboreiro et al. (2010)
    - Especially dedicated to micro blogging
    - Supervised; trained on pre-tokenized text
    - Initial, minimal rule-based step
    - Training of a support vector machine classifier
    - Very promising results but **no end-of-sentence detection**

# Tokenizing – Our Approach

- Tokenizing as PoS tagging
- `moot` (Jurish 2003)
  - 2nd order Hidden Markov Model
  - *Observations* words; *states* tags
  - Viterbi optimization ...
- Minimal rule set
- Aggregation of (pre-)tokens
- Small tagset

# Tokenizing – Our Approach II




$$\tau(w_{i\ldots n}) = \arg\max_{q_{1\ldots n}\in T^n} P(q_{1\ldots n}|w_{1\ldots n})$$

$$P(q_{1\ldots n}|w_{1\ldots n}) = \prod_{i=1}^{n} P(w_i|q_i)P(q_i|q_{i-2}q_{i-1})$$

# Tokenizing – Our Approach III

- Classification of pre-tokens wrt.
  - Class $C = \{alpha, alpha\text{-}stopword, numeric, eos, dot, comma, quote, other\}$
  - Spelling $S = \{upper, lower, caps, *\}$
  - Length $L = \{1, \leq 3, \leq 5, long\}$
  - Abbreviation $A = \{known, unknown\}$
- Classification of token state wrt.
  - Beginning of token $\texttt{bot} \in \mathbb{B}$
  - Beginning of sentence $\texttt{bos} \in \mathbb{B}$

# Tokenizing – Our Approach IV

- Text $T \in K \times S \times L \times A$ (minus „impossible“ combinations)

  $\rightarrow$ *Aggregation* of tokens

- Tag $\in T \times \mathbb{B} \times \mathbb{B}$ (minus „impossible“ combinations)

  $\rightarrow$ Tagset!

# Tokenizing – Example

| | |
|---|---|
| Das | cls:alpha−sw_cas:upper_abbr:uk_len:3_bos:1_bot:1 |
| Unternehmen | cls:alpha_cas:upper_abbr:uk_len:long_bos:0_bot:1 |
| verkauft | cls:alpha_cas:lower_abbr:uk_len:long_bos:0_bot:1 |
| er | cls:alpha−sw_cas:lower_abbr:uk_len:3_bos:0_bot:1 |
| 1984 | cls:num_cas:∗_abbr:uk_len:5_bos:0_bot:1 |
| fuer | cls:alpha−sw_cas:lower_abbr:uk_len:3_bos:0_bot:1 |
| 2,5 | cls:num_cas:∗_abbr:uk_len:3_bos:0_bot:1 |
| Milliarden | cls:alpha_cas:upper_abbr:uk_len:long_bos:0_bot:1 |
| . | cls:dot_cas:∗_abbr:uk_len:1_bos:0_bot:1 |

# Tokenizing – Evaluation

- Dortmund chat corpus (Storrer and Beißwenger)
  - Base corpus containing $140{,}000$ posts ($\approx 10^6$ tokens)
  - Freely available release corpus ($\approx 500{,}000$ tokens)
  - Chats from different contexts: university, media, „Plauder-Chats“
- Manually tokenized subset (ca. $100{,}000$ tokens)
- Includes different chat scenarios ($1:1$, $1:n$, $m:n$)

# Tokenizing – Evaluation II

| | manual | | automatic | |
|---|---|---|---|---|
| + Items | 105,279 | (100.0 %) | 103,290 | (100.0 %) |
| – Match | 100,337 | (**95.3** %) | 100,337 | (97.1 %) |
| – NoMatch | 4,942 | (4.7 %) | 2,953 | (2.9 %) |
| + Items | 39,747 | (100.0 %) | 39,158 | (100.0 %) |
| – Match | 38,515 | (**96.9** %) | 38,515 | (98.4 %) |
| – NoMatch | 1,232 | (3.1 %) | 643 | (1.6 %) |

- Still losing 34 % of the unmarked sentence boundaries
- Training on Tiger: 92 % correct for CMC and 99 % for Tiger

# Morphological Analysis – Task

- Categorisation of words regarding their *possible* PoS
- Mapping of words to their base form (i.e. *lemma*)
- Due to word formation processes, infinite number of possible words

# Morphological Analysis – Challenges

- Writing errors

  dadruch

- Missing whitespace

  3,9Millionen

- Dialect-specific words

  zwüsched, zwüschäd, zwüschet

- Complex bounded compounds

  26/26-Lösung, 2-Fach-BA

# Morphological Analysis – Approach

- Finite-state morphology
  - Large lexicon (word, category, inflection class)
  - Productive rules to mimic word formation processes
  - Implemented with (weighted) finite-state transducers
  - Weights as complexity estimate of the word formation

$$\mathtt{NN} \cdot \mathtt{NN} \rightarrow \mathtt{NN}\ \langle 10 \rangle$$

  - Beesley and Karttunen (2003), Schmid et al. (2004), Geyken and Hanneforth (2006)

# Morphological Analysis – Example

[1]> apply "Telekommunikation"

Telekommunikation[NN SemClass=abstr Gender=fem Number=sg Case=*]

tele/K+Kommunikation[NN SemClass=abstr Gender=fem Number=sg Case=*] <5>

Tele/N#Kommunikation[NN SemClass=abstr Gender=fem Number=sg Case=*] <10>

...

Telekom/PN#Muni/N#Kat/N#Ion[NN SemClass=k_g_dingnat Gender=neut Number=sg Case=nom_acc_dat] <30>

...

# Morphological Analysis – Evaluation

- No gold standard for standard or non-standard text available
- Alternatively, evaluation of coverage of TAGH (Geyken and Hanneforth 2006)

| | Token | Types |
|---|---|---|
| *known* | 84,599 | 10,346 |
| *unknown* | 9,372 | 5,459 |
| Coverage % | 88.92 | 47.24 |

- Recognition rate of TAGH for modern newspaper text: 99 %

# Morphological Analysis – Improvements

- Error tolerant morphological analysis: Jurish (2011)
  - Developed for historical texts
  - Uses phonological similarity and edit distance

| | Token | Types |
|---|---|---|
| *known* | 84,599 | 10,346 |
| *unknown* | 9,372 | 5,459 |
| *fixed* | 7,183 | 3,984 |
| Coverage % | 97.41 | 85.74 |

- Historical model applied to test corpus without adjustments

# Morphological Analysis – Top 10 Improvements

| Type | Frequency | Type | Frequency |
|---|---|---|---|
| *hi* | 98 | *gibts* | 33 |
| *nich* | 51 | *kinder* | 32 |
| *nix* | 46 | *musik* | 31 |
| *leute* | 42 | *z.B.* | 29 |
| *adelheid* | 40 | *tach* | 27 |

# Part-of-Speech Tagging – Task and Approaches

- Disambiguation of possible PoS of a word in context
- (partly) rule-based (Brill 1995) vs. fully statistical (Brants 2000)
- Supervised (trained on manually tagged text)
- Training materials available in the form of tree banks (leaves and pre-terminals)

# Part-of-Speech Tagging – Challenges

- Mimicry of spoken language
- Ellipses
- Single word sentences

Geeehts?
Ich!!!
Ein sehr starkes Schlafmittel...

# Part-of-Speech Tagging – Approach

- `moot` (Jurish, 2003)
- Trained on `Tiger`
- 2nd order *Hidden Markov* Modell: *observations* words, *states* PoS tags
- Viterbi maximization over PoS tag sequences for a sentence
- Optional restriction to categories assigned by the morphology

# Part-of-Speech Tagging – Example (Jurish 2003)

| | | | | | | |
|---|---|---|---|---|---|---|
| **Input:** | Linda | wird | die | Mannschaft | verstärken | . |
| **Morphological Analysis:** | $\left\{\begin{array}{l}\texttt{NE}.first,\\ \texttt{NE}.last\end{array}\right\}$ | $\left\{\begin{array}{l}\texttt{VAFIN}.3rd.sg.pres,\\ \texttt{VVFIN}.3rd.sg.pres,\end{array}\right\}$ | $\left\{\begin{array}{l}\texttt{ART}.sg.nom.fem,\\ \vdots\\ \texttt{PDS}.nom.sg.fem,\\ \vdots\\ \texttt{PRELS}.acc.pl\end{array}\right\}$ | $\left\{\begin{array}{l}\texttt{NN}.masc.sg.nom,\\ \vdots\\ \texttt{NN}.fem.sg.*\end{array}\right\}$ | $\left\{\begin{array}{l}\texttt{VVFIN}.1st.pl.pres,\\ \vdots\\ \texttt{VVINF}\end{array}\right\}$ | $\{\$.\}$ |
| **Tag Extraction:** | $\{\texttt{NE}\}$ | $\left\{\begin{array}{l}\texttt{VVFIN},\\ \texttt{VAFIN}\end{array}\right\}$ | $\left\{\begin{array}{l}\texttt{ART},\\ \texttt{PDS},\\ \texttt{PRELS}\end{array}\right\}$ | $\{\texttt{NN}\}$ | $\left\{\begin{array}{l}\texttt{VVFIN},\\ \texttt{VVINF}\end{array}\right\}$ | $\{\$.\}$ |
| **Disambiguation:** | NE | VAFIN | ART | NN | VVINF | $. |

# Part-of-Speech Tagging – Evaluation

- ≈ 10 % of the test corpus manually annotated
- Not (yet) enough to train a model but to evaluate existing tools
- `moot` vs. `TreeTagger`

| | `moot` | $\texttt{moot}_{tagh}$ | $\texttt{moot}_{fixed}$ | `TreeTagger` |
|---|---|---|---|---|
| *match* | 7,424 | 6,999 | 7,948 | 7,673 |
| *nomatch* | 2,908 | 3,330 | 2,386 | 2,661 |
| Correctness % | 71.9 | 67.8 | 76.9 | 74.3 |

- `tree-tagger -cap-heuristics -sgml -eos-tag "<eos>"`

# Summary

- Promising results for tokenizing
- Morphological analysis yet to be evaluated
- PoS tagging disastrous: better models and pre-processing needed

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

# Many thanks for your attention

Big shout to Gabriella Pein, Maria Ermakova, Michael Beißwenger, Julian Heister und Frank Wiegand