Canonicalization Techniques for Historical German Text

Bryan Jurish

jurish@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstrasse 22/23 · 10117 Berlin · Germany

November 19, 2009

Overview

The Big Picture

- **The Situation**: unconventional text corpora
- **The Problem**: conventional tools $\rightsquigarrow$ low coverage
- **The Proposal**: conflation & canonical form(s)

Three Canonicalization Techniques

- Sketches
- Examples
- Evaluation

Future Directions

- “Wouldn’t it be nice . . . ”
- Next Steps
- Selected Software

The Big Picture

The Situation: Corpora

"Unconventional" Text Corpora

- ***Historical text***
- Spoken language transcriptions
- OCR output
- Non-standard dialects *(e.g. E-Mail, SMS, Tweets, ...)*

Lexical "Conventions"

- Extinct or dialect-specific lexemes
- Require manual attention

Orthographic Conventions

- Extinct, dialect-specific, or spurious lexical variants
- Can be handled automatically (to some extent)

The Situation: Text Technologies

Conventional Text Technologies

- Document indexers
- Part-of-speech taggers
- Word stemmers
- Morphological analyzers

Common Characteristics

- Fixed lexicon accessed *via* orthographic form
- Extant lexemes only

Desideratum

- Apply existing tools to “unconventional” corpora

. . . but . . .

The Problem

	Conventional	Tools
+	Unconventional	Corpus
		Soup

- Corpus variants missing from application lexicon
- Low coverage (many unknown types)
- Poor recall (much relevant data not retrieved)
- Degraded accuracy (poor model fit)
- . . . *and more!*

The Proposal

Conflation & Canonical Form(s)

- Collect variant forms into equivalence classes
- Represent classes by (extant) canonical elements

Analysis by Disjunction

- Analyze "extinct" form w by disjunction over extant members of its equivalence class $[w]$:

$$\text{analyses}(w) := \bigcup_{v \in [w]} \text{analyses}(v)$$

- Expect improved recall, some loss of precision

. . . A Case in Point

Base Corpus

- Verse quotations from the *e-DWB1* *(Bartz et al., 2004)*
- 6,581,501 tokens of 322,271 graphemic types
- Indexed with the TAXI corpus indexing system

Preprocessing & Filtering

- UTF-8 $\longrightarrow$ ISO-8859-1 (*e.g.* $\text{œ} \mapsto oe$, $\overset{e}{o} \mapsto \ddot{o}$, $\hat{o} \mapsto o$, . . .)
- ***removed*** non-alphabetic & foreign material
- 5,491,982 tokens of 318,383 graphemic types

Conventional Analysis

- TAGH morphology FST *(Geyken & Hanneforth, 2006)*

Canonicalization Techniques

Phonetic Canonicalization: Sketch

Idea

(Jurish, 2008)

- Map each word w to a unique phonetic form $\mathrm{pho}(w)$
- Conflate words with identical phonetic forms

$$[w]_{\mathrm{pho}} := \{v : \mathrm{pho}(v) = \mathrm{pho}(w)\}$$

Phonetization: Letter-to-Sound (LTS) Conversion

- Well-known in text-to-speech (TTS) research
- `ims_german` LTS rule-set *(Möhler et al., 2001)*
  - for `festival` TTS system *(Black & Taylor, 1997)*
  - slightly modified for historical input
  - compiled as a finite-state transducer (FST)

Phonetic Canonicalization: Problems

Insufficient (too permissive)

- Phonetic Identity $\not\Rightarrow$ Lexical Equivalence
- ***Precision Errors*** (conflated but not equivalent)
  - *(hân–Hahn), (niht–Niet), (vil–fiel), (usz–Uhus), . . .*
- Not too dangerous (yet)

Unnecessary (too strict)

- Phonetic Identity $\not\Leftarrow$ Lexical Equivalence
- ***Recall Errors*** (equivalent but not conflated)
  - *(guot–gut), (pflag–pflegte), (tiuvel–Teufel), (umb–um), . . .*
- This is the **more severe** of the two problems!

Lemma Instantiation: Sketch

Idea

(Jurish, 2008)

- Exploit dictionary-corpus structure
- Assume each quote contains an instance of the associated dictionary lemma

String Edit Distance

(Levenshtein, 1966; Baroni et al., 2002)

- Relax strict identity criterion

Pointwise Mutual Information

(McGill, 1955; Church & Hanks, 1990)

- Filter out “random” phonetic similarities

Restrict Comparisons

- Compare only lemma-instance pairs
- Over ***10 thousand times faster*** (*vs.* all word pairs)

Lemma Instantiation: Problems

Restricted Domain

- Only applicable to dictionary-structured corpora
- MI technique requires (Quote $\rightarrow$ Lemma) mapping

Sparse Data $\rightsquigarrow$ Overspecification

- Extensional solution only valid for input corpus
- No inflectional paradigms
  - *post-hoc* workaround for nouns, verbs *(Seeker, 2008)*

Levenshtein Metric Granularity: too Coarse

- No context-sensitivity $c(th \rightarrow t) = c(h \rightarrow \varepsilon) = 1$
- No target-sensitivity $c(\ddot{u} \rightarrow i) = c(\ddot{u} \rightarrow x) = 1$

Rewrite Cascade: Sketch

Idea: Generalized Edit Distance via WFSTs

- Replace coarse Levenshtein metric
- Allow *a priori* linguistic "hints"
  - *e.g.* attenuate edit costs for vowel shifts (*ü/i, iu/eu, . . .*), voicing alternations (*k/g, b/p, . . .*), *etc.*
- Can be fine-tuned to handle dialect-specific quirks
- Can be used for arbitrary input (not just dictionaries)

Combinatorial Headaches

$$|Q_{(A \circ B)}| = O(|Q_A| \times |Q_B|)$$

- Offline cascade compilation; online composition

Approach

- On-the-fly computation of cascaded n-best lookup
- Variant of A*/Dijkstra Algorithm *(Dijkstra, 1959; Hart et al., 1968)*

Examples: TAGH Coverage Errors

da sah ich sitzen siben frawen
radweisz umb einen külen brunnen.

vil manige sêle er zuhte
dem tiuvel ûʒ sînem rachen.

*war

ir keinr nam war
wa ieder lag am rangen.

genuoge wurden verbrant,
versteinet und mit swerte erslagen

≠ *versteint*

Examples: Phonetic Canonicalization

sieben

da sah ich sitzen siben frawen
radweisz umb einen külen brunnen.

kühlen

*viel, *fiel* *Seele*

vil manige sêle er zuhte
dem tiuvel ûz sînem rachen.

aus *seinem*

ihr *nahm* **war*, *wahr*

ir keinr nam war
wa ieder lag am rangen.

**Ider*

verbrannt

genuoge wurden verbrant,
versteinet und mit swerte erslagen

≠ *versteint*

Examples: Lemma Instantiation

sieben

da sah ich sitzen siben frawen
radweisz umb einen külen brunnen.

radweise *kühlen*

*viel, *fiel* *Seele* *zuckte*

vil manige sêle er zuhte
dem tiuvel ûz sînem rachen.

Teufel *aus* *seinem*

ihr *nahm* **war, wahr*

ir keinr nam war
wa ieder lag am rangen.

**Ider, jeder*

verbrannt

genuoge wurden verbrant,
versteinet und mit swerte erslagen

≠ versteint *Schwert*

Examples: Rewrite Cascade

sieben Frauen
da sah ich sitzen siben frawen
radweisz umb einen külen brunnen.
radweise, *radweiß um kühlen

viel, *fiel manche Seele zuckte
vil manige sêle er zuhte
dem tiuvel ûz sînem rachen.
Teufel aus seinem

ihr keiner nahm *war, wahr
ir keinr nam war
wa ieder lag am rangen.
wo *Ider, jeder

genug verbrannt
genuoge wurden verbrant,
versteinet und mit swerte erslagen
≠ versteint Schwert erschlagen

Examples: Mapping Errors

sieben *Frauen*

da sah ich sitzen siben frawen

radweisz umb einen külen brunnen.

*radweise, *radweiß* *um* *kühlen*

*viel, *fiel* *manche* *Seele* *zuckte*

vil manige sêle er zuhte

dem tiuvel ûz sînem rachen.

Teufel *aus* *seinem*

ihr *keiner* *nahm* **war, wahr*

ir keinr nam war

wa ieder lag am rangen.

wo **Ider, jeder*

genug *verbrannt*

genuoge wurden verbrant,

versteinet und mit swerte erslagen

≠ versteint *Schwert* *erschlagen*

Evaluation: Coverage

Method	Types	Tokens
+TAGH	42.4 %	83.7 %
+TAGH / pho	54.6 %	91.5 %
+TAGH / li	66.7 %	94.4 %
+TAGH / rw	80.4 %	97.3 %
(any)	**86.1 %**	**98.4 %**
Error Reduction	**75.9 %**	**90.2 %**

Evaluation: Gold Standard

Gold Standard Corpus

- Verse quotes from 1 volume (`gr01`) of the *e-DWB1*
- 11,818 tokens of 4,382 graphemic types

Manual Annotations (by source type)

- Source type “class”
- Extant target type(s)
- “Expert review” option for problematic types

Class	Description	Types	Tokens
LEX	“normal” lexical word	4,089	11,502
NE	proper name	118	127
FM	“foreign” material	65	69
GONE	lexical, extinct root	28	31
XY, ...	*(other material)*	82	89

Evaluation *vs.* Gold Standard

- Quantitative evaluation, restricted to LEX words
- Pairwise measures for equivalence relations

(Hatzivassiloglou & McKeown, 1993; Schulte im Walde, 2003)

- Precision $pr = P(\text{Relevant}|\text{Retrieved}) = tp/(tp + fp)$
- Recall $rc = P(\text{Retrieved}|\text{Relevant}) = tp/(tp + fn)$
- Harmonic Average $F = \left(\frac{1}{2}pr^{-1} + \frac{1}{2}rc^{-1}\right)^{-1} = 2 \times pr \times rc/(pr + rc)$

	% Type Pairs			% Token Pairs		
Method	**pr**	**rc**	**F**	**pr**	**rc**	**F**
tagh	***99.5***	45.7	62.7	***100.0***	90.2	94.8
tagh, pho	98.1	66.8	79.5	99.6	93.7	96.6
tagh, li	86.7	62.5	72.7	99.8	93.0	96.3
tagh, rw	96.1	73.4	83.3	99.8	93.6	96.6
tagh, pho, li	87.2	70.5	78.0	99.5	93.8	96.6
tagh, pho, rw	94.1	77.3	***84.9***	98.8	***96.3***	97.5
tagh, li, rw	84.0	76.2	79.9	99.6	95.7	***97.6***
tagh, pho, li, rw	83.5	***78.6***	81.0	98.7	***96.3***	97.5

Future Directions

“Wouldn’t it be nice . . . ”

Evaluation by Date of Origin

- Requires (Quote → Date) mapping
- Work in progress

Exception Lexicon

- For high-frequency types & common errors
- Currently defined only for gold-standard source types

Inflectional Paradigms

- Extend lemma-instantiation heuristics
- Untested, only defined for nouns & verbs

Dictionary Prose Evidence

- Heuristic exist, but are quite unreliable

Next Steps

Part-of-Speech Tagging

- `moot` tagger/disambiguator suite *(Jurish, 2003)*
- **To Do:** Adapt to accomodate non-canonical input

Context-Sensitive Canonicalization

- HMM disambiguation of canonical surface form
- **To Do:** Dynamic `moot` lexica, parameter estimation

Lemma Selection

- Conflate & evaluate by lemma (not by surface form)
- **To Do:** univocal (Token → Lemma) mapping

Dialect-Dependent Specialization

- Select rewrite rules based on source “dialect”
- **To Do:** Training data, dialect guesser, & more . . .

Selected Software

Lingua::LTS Rule Compiler

- Used to construct LTS and Rewrite-Editor (W)FSTs

Taxi::Mysql Document Indexing System

- Taxi/Grimm http://services.dwds.de:8765/
- Taxi/Grimm/WordMapper http://services.dwds.de:8764/
- Taxi/DTA http://services.dwds.de:9876/

DTA::CAB “Cascaded Analysis Broker”

- Online robust analysis daemon
- XML-RPC web-service
- HTML demo http://www.deutschestextarchiv.de/cab/

The End

Thank you for listening!