Comparing Canonicalizations of Historical German Text

Bryan Jurish

jurish@bbaw.de

Project *“Deutsches Textarchiv”*
Berlin-Brandenburg Academy of Sciences
Berlin, Germany

SIGMORPHON 2010
Uppsala, Sweden
15 July, 2010

Overview

The Big Picture

- The Situation
- The Problem
- The Proposal

Canonicalization Methods

- Phonetic Identity
- Levenshtein Edit Distance
- Heuristic Rewrite Transducer

Evaluation

- Test Corpus
- Evaluation Measures
- Results

The Big Picture

The Situation

Historical Text $\not\ni$ Orthographic Conventions

- also applies to OCR text, E-Mail SMS, Tweets, . . .
- High variance of graphemic forms

fröhlich *“joyful”*	frölich, fröhlich, vrœlich, frœlich, froͤlich, froͤhlich, vrölich, fröhlig, frölig, . . .
Herzenleid *“heart-sorrow”*	hertzenleid, herzenleit, hertzenleyd, hertzen-laidt, hertzenlaydt, herzenleyd, . . .

Conventional NLP Tools $\Rightarrow$ Strict Orthography

- Document indexers, PoS taggers, stemmers, morphological analyzers, parsers, . . .
- Fixed lexicon keyed by orthographic form
- Extant lexemes only

The Problem

	Conventional	**Tools**
$\oplus$	**Historical**	**Corpus**
=		***Soup***

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

The Proposal

In a Nutshell

þe	Olde	Wydgett	Shoppe
↓	↓	↓	↓
the	*old*	*widget*	*shop*

- *Conflate* each word w with its *canonical cognates* $\widetilde{w}$
- *Defer* application analysis to canonical forms

$$\text{analyses}_R(w) := \bigcup_{\widetilde{w} \in \text{Lex} \cap [w]_R} \text{analyses}(\widetilde{w})$$

Canonical Cognates

- Synchronically active "extant equivalents" $\widetilde{w} \in \text{Lex}$
- Preserve both root and relevant features of input

Conflation Relation

- *Binary relation* $\sim_R$ on strings (words) in $\mathcal{A}^*$
- Prototypically a true *equivalence relation*

Canonicalization Methods

Phonetic Conflation: Sketch

Idea

(Jurish, 2008)

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

$$w \sim_{\mathrm{Pho}} v :\Leftrightarrow \mathrm{pho}(w) = \mathrm{pho}(v)$$

Phonetization: Letter-to-Sound (LTS) Conversion

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

$$M_{\sim_{\mathrm{Pho}}} = M_{\mathrm{Pho}} \circ M_{\mathrm{Pho}}^{-1} \circ \mathrm{Id}(\mathrm{Lex})$$

Phonetic Conflation: Problems

Insufficient **(too permissive)**

- Phonetic Identity $\not\Rightarrow$ Lexical Equivalence
- ***Precision Errors*** (conflated but not equivalent)
- Not too dangerous (yet)

usz–Uhus	vil–fiel	in–ihn
“out”–“owls”	*“much”–“fell”*	*“in”–“him”*

Unnecessary **(too strict)**

- Phonetic Identity $\not\Leftarrow$ Lexical Equivalence
- ***Recall Errors*** (equivalent but not conflated)
- This is the **more severe** of the two problems!

guot–gut	tiuvel–Teufel	umb–um
“good”	*“devil”*	*“around”*

Levenshtein Conflation: Sketch

Idea

- Relax strict identity criterion (improve recall)
- Map each input word to "nearest" extant type
  - string edit distance *(Levenshtein, 1966)*
  - computable even for infinite lexica *(Mohri, 2002)*

Gory Details

$$\mathrm{best}_{\mathrm{Lev}}(w) \quad := \quad \arg\min_{v \in \mathrm{Lex}} [\![M_{\mathrm{Lev}}]\!](w, v)$$

$$w \sim_{\mathrm{Lev}} v \quad :\Leftrightarrow \quad \mathrm{best}_{\mathrm{Lev}}(w) = \mathrm{best}_{\mathrm{Lev}}(v)$$

- Synchronic lexicon $\mathrm{Lex} \subseteq \mathcal{A}^*$
  - TAGH input language *(Geyken & Hanneforth, 2006)*
- Edit Distance WFST M_{Lev}
- Best-first search using `gfsmxl` C library

Levenshtein Conflation: Problems

Search Space too Large

- Backtracking & heap maintainence are $\mathcal{O}(|\mathcal{A}| \cdot |w|)$
- *circa* 150 times slower than phonetic conflation

Metric Granularity too Coarse

- No context-sensitivity $c(th \to t) = c(uhu \to uu) = 1$
- No target-sensitivity $c(\ddot{u} \to i) = c(\ddot{u} \to x) = 1$
- Examples for $d_{\text{Lev}} = 1$

w	$\text{best}_{\text{Lev}}(w)$		$\widetilde{w}$	
aug	aus	*"out"*	auge	*"eye"*
faszt	fast	*"almost"*	fasst	*"grabs"*
ouch	buch	*"book"*	auch	*"also"*
ram	rat	*"advice"*	rahm	*"cream"*
vol	volk	*"people"*	voll	*"full"*

Rewrite Cascade: Sketch

Idea: Generalized Edit Distance via WFSTs

- Replace coarse Levenshtein metric
- Reduce search space
- Attenuate edit costs for *e.g.*
  - elision $mp \to m / \_\_\# \langle 1 \rangle, \quad n \to en / \_\_\# \ \langle 5 \rangle$
  - vowel shift $o \to a \ / \_\_u \ \langle 1 \rangle, \quad o \to a \ / \_\_ \ \langle 9 \rangle$
  - (un)voicing $p \to b \ / \_\_ \ \langle 5 \rangle, \quad b \to p \ / \_\_ \ \langle 8 \rangle$
  - corpus quirks $sz \to ß \ / \_\_ \ \langle 1 \rangle, \quad f \to s \ / \_\_ \ \langle 10 \rangle$

Implementation

- Heuristic *"rewrite" transducer* M_{rw} replaces M_{Lev}

$$w \sim_{\text{rw}} v :\Leftrightarrow \text{best}_{\text{rw}}(w) = \text{best}_{\text{rw}}(v)$$

- 306 manually constructed SPE-style two-level rules
- *circa* 40 times faster than Levenshtein conflation

Rewrite Cascade: Problems

Resource-Intensive

- Heuristic rule-set must be manually developed
  - requires “expert” knowledge
  - time-consuming task

Language-Specific

- No immediate generalization to other languages

Computationally Expensive

- *circa* 4 times slower than Pho
- . . . still a big improvement over Lev

Evaluation

Evaluation: Basics

Gold Standard Test Corpus G

- Historical German verse from *e-DWB1* *(Bartz et al., 2004)*
- 11,242 tokens; 4157 types
- Canonical cognate manually assigned to each token

Evaluation Measures

- Simulated information retrieval task
- Type- and token-wise precision (pr), recall (rc), and F

Evaluation: Results

R	Type-wise %			Token-wise %		
	pr	rc	F	pr_f	rc_f	F_f
Id	***99.9***	70.8	82.9	***99.1***	83.7	90.7
Pho	96.7	80.1	87.6	92.7	89.6	91.1
Lev	96.6	78.9	86.9	97.2	87.8	92.2
rw	98.5	88.4	***93.2***	98.2	93.4	***95.8***
Pho|Lev	94.1	84.3	88.9	91.3	91.6	91.5
Pho|rw	96.1	***89.8***	92.8	92.5	***94.5***	93.5

Evaluation: Results: Id

R	**Type-wise %** pr	rc	F	**Token-wise %** pr_f	rc_f	F_f
Id	***99.9***	70.8	82.9	***99.1***	83.7	90.7
Pho	96.7	80.1	87.6	92.7	89.6	91.1
Lev	96.6	78.9	86.9	97.2	87.8	92.2
rw	98.5	88.4	***93.2***	98.2	93.4	***95.8***
Pho|Lev	94.1	84.3	88.9	91.3	91.6	91.5
Pho|rw	96.1	***89.8***	92.8	92.5	***94.5***	93.5

Id: **naïve string identity**

- Most precise, but worst recall
- Especially poor recall for low-frequency types
- Historical text really ***is*** tricky!

Evaluation: Results: Pho

R	Type-wise % pr	rc	F	Token-wise % pr_f	rc_f	F_f
Id	***99.9***	70.8	82.9	***99.1***	83.7	90.7
Pho	96.7	80.1	87.6	92.7	89.6	91.1
Lev	96.6	78.9	86.9	97.2	87.8	92.2
rw	98.5	88.4	***93.2***	98.2	93.4	***95.8***
Pho|Lev	94.1	84.3	88.9	91.3	91.6	91.5
Pho|rw	96.1	***89.8***	92.8	92.5	***94.5***	93.5

Pho: Phonetic conflation

- Poor token-wise precision
- Small number of errors for high-frequency types
  - *in–ihn* (“in”–“him”)
  - *wider–wieder* (“against”–”again”)

Evaluation: Results: Lev

R	Type-wise % pr	rc	F	Token-wise % pr_f	rc_f	F_f
Id	***99.9***	70.8	82.9	***99.1***	83.7	90.7
Pho	96.7	80.1	87.6	92.7	89.6	91.1
Lev	96.6	78.9	86.9	97.2	87.8	92.2
rw	98.5	88.4	***93.2***	98.2	93.4	***95.8***
Pho|Lev	94.1	84.3	88.9	91.3	91.6	91.5
Pho|rw	96.1	***89.8***	92.8	92.5	***94.5***	93.5

Lev: **Levenshtein conflation**

- ***No*** recall improvement *vs.* Pho
  - too many spurious conflations
  - union Pho|Lev does somewhat better

Evaluation: Results: rw

R	Type-wise % pr	rc	F	Token-wise % pr_f	rc_f	F_f
Id	***99.9***	70.8	82.9	***99.1***	83.7	90.7
Pho	96.7	80.1	87.6	92.7	89.6	91.1
Lev	96.6	78.9	86.9	97.2	87.8	92.2
rw	98.5	88.4	***93.2***	98.2	93.4	***95.8***
Pho|Lev	94.1	84.3	88.9	91.3	91.6	91.5
Pho|rw	96.1	***89.8***	92.8	92.5	***94.5***	93.5

rw: **Heuristic rewrite transducer**

- Best method overall
  - *circa* 60% fewer recall errors *vs.* string identity
- Recall further improved by including Pho

Conclusion

Summary

- Historical text corpora and conventional tools
  won't play together nicely
- Best canonicalization by heuristic rewrite FST
  - implementing linguistic intuitions helps!
- Phonetic, Levenshtein methods more accessible
  - improved by exception lexica, cost upper bounds

Next Steps

- Larger corpus *(under construction)*
- Precision recovery for overgeneration *(alpha)*
- Language-independent (pseudo-)metrics

þe Olde Laſſt Slyde

(“The End”)

Thank you for listening!

http://www.deutschestextarchiv.de