Canonicalizing Historical Text with CAB

Bryan Jurish
jurish@bbaw.de

Searching Linguistic Patterns in Large Text Corpora for Digital Humanities Research
ESU Digital Humanities 2016, Universität Leipzig
20th July, 2016

Overview

The Big Picture

- The Situation
- The Problem
- The Approach

Canonicalization Methods

- Type-wise Conflation
  - Lexicon, Transliteration, Phonetization, Rewrite Cascade
- Token-wise Disambiguation
  - Dynamic Hidden Markov Model
- Problems & Workarounds

Experiments

- Generative Canonicalization Methods
- Alignment-based Lexicon

— The Big Picture —

Deutsches Textarchiv (DTA)

CLARIN-D

Primary Goals

- digitize ∼ 1,300 print volumes, printed ∼ 1600-1900
  - first editions of respective works
  - detailed metadata, highly accurate transcriptions
- TEI-XML corpus encoding & storage
  - DTA base format (DTABf) dialect
- linguistic analysis (automated)
  - tokenization, normalization, PoS-tagging, lemmatization
- online search (DDC) http://www.ddc-concordance.org
  - lemma-based, PoS-sensitive, spelling-tolerant

In Numbers

(2016-07-14)

2,438	transcribed works
593,008	digitized pages
991,255,987	unicode characters
139,738,217	running words
447	additional works in DTAQ

Deutsches Textarchiv (DTA)

http://www.deutschestextarchiv.de

Anmelden (DTAQ) DWDS dlexDB CLARIN-D

D T A

Werke im Deutschen Textarchiv

Texte ▼ Projekt ▼ Dokumentation ▼ Impressum

suchen

in den Titeldaten · im Korpus · in der Dokumentation · Hilfe

Beispielanfragen: {Gott,Herr} && "das ewige Leben #2 geben" ehelichen with $p=VVINF "Auge um" && "Zahn um"

Deutsches Textarchiv

Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache

Das Deutsche Textarchiv stellt einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis 1900 bereit. Die Textauswahl erfolgte auf der Grundlage einer von Akademiemitgliedern erstellten und ausführlich kommentierten, umfangreichen Bibliographie. In Ergänzung wurden einschlägige Literaturgeschichten und (Fach-)Bibliographien ausgewertet. Aus der Gesamtliste der auf diesem Wege ermittelten Titel wurde von der DTA-Projektgruppe ein hinsichtlich der repräsentierten Textsorten und Disziplinen ausgewogenes Korpus zusammengestellt (weitere Informationen zur Textauswahl).

Um den historischen Sprachstand möglichst genau abzubilden, werden als Vorlage für die Digitalisierung in der Regel die Erstausgaben der Werke zugrunde gelegt. Das elektronische Volltextkorpus des DTA ist über das Internet frei zugänglich und dank seiner Aufbereitung durch (computer-)linguistische Methoden schreibweisentolerant über den gesamten jeweils verfügbaren Bestand durchsuchbar. Sämtliche Texte stehen zum Download zur Verfügung.
mehr ...

Neuigkeiten aus dem Projekt

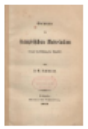
Das DTA auf der LREC 2016

Das DTA in Zahlen

2 438	Werke
593 008	digitalisierte Seiten
139 738 217	fortlaufende Wortformen
991 255 987	Zeichen (Unicode)
447	weitere Werke in DTAQ

Das DTA am 14. Juli 2016

Am 14. Juli 1789 fand der *Sturm auf die Bastille*, Symbol für die französische Revolution, statt. Die Bastille war ein Gefängnis in Paris und galt bei den Revolutionären als Sinnbild für das Ancient Regime. Ziel der bewaffneten Demonstrantenmenge waren außerem die dort gelagerten Munitionsvorräte. Bereits zwei Tage nach der Erstürmung begannen Abrissarbeiten, die sich über ein Jahr hinzogen. Noch heute ist der 14. Juli der Nationalfeiertag Frankreichs. Friedrich Christoph Dahlmann beschreibt die Vorgänge in seiner „Geschichte der französischen Revolution".



Dahlmann, Friedrich Christoph: Geschichte der französischen Revolution bis auf die Stiftung der Republik. Leipzig, 1845.

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, frölich, fröhlich, vrölich, fröhlig, frölig, ...
Herzenleid *"heart-sorrow"*	hertzenleid, herzenleit, hertzenleyd, hertzenleidt, herzenlaid, hertzenlaid, hertzenlaydt, ...

Conventional NLP Tools $\Longrightarrow$ Strict Orthography

- IR systems, PoS taggers, lemmatizers, morphological analyzers, ...
- Fixed lexicon keyed by (ortho)graphic 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)
  - ***spurious "noise"*** (poor model fit)
  - *. . . and more!*

The Approach: Canonicalization

a.k.a. (orthographic) 'standardization', 'normalization', 'modernization', . . .

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

In a Nutshell

- *Map* each word w to a unique canonical cognate $\widetilde{w}$
- *Defer* application analysis to canonical forms

Canonical Cognates

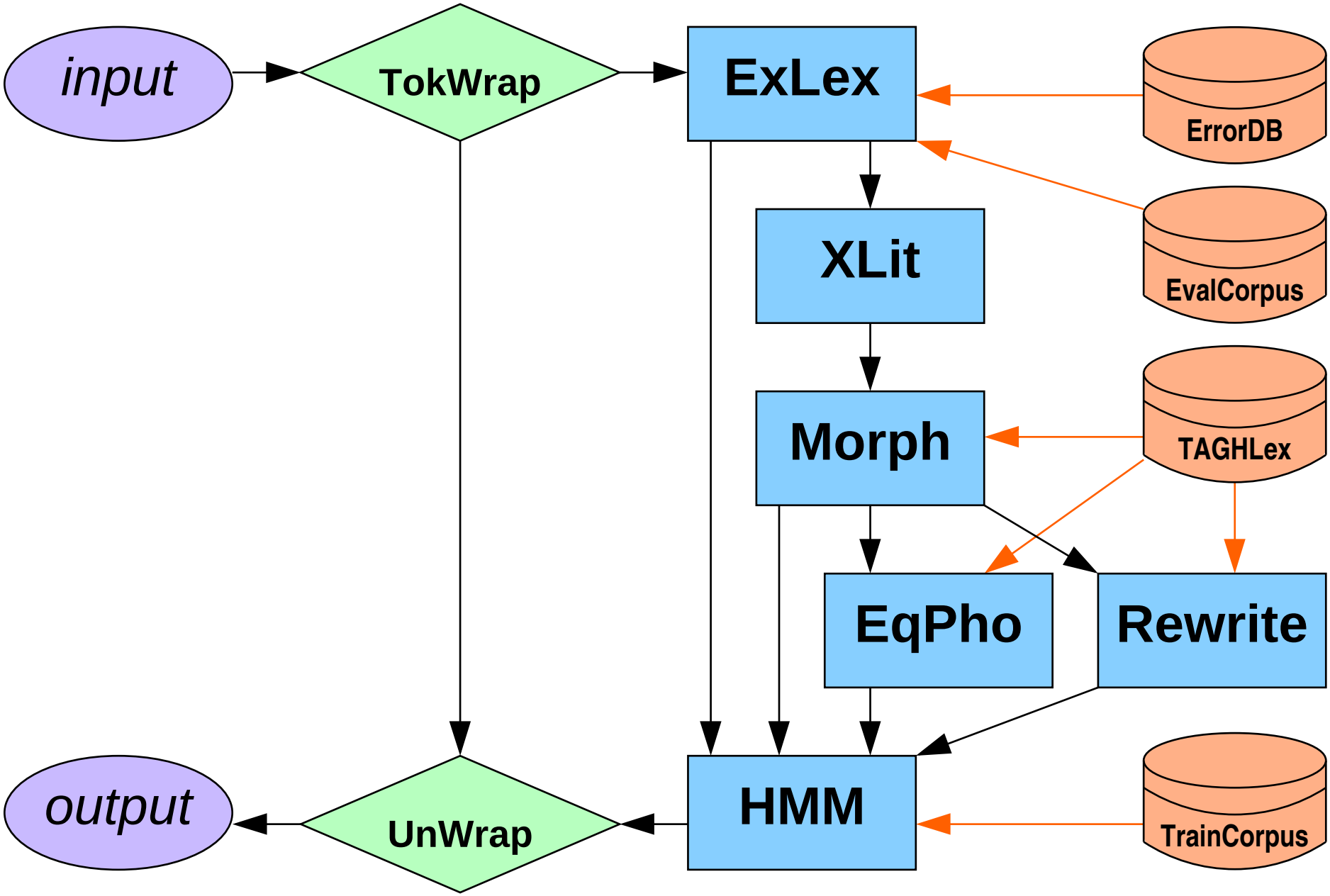
- Synchronically active "extant equivalent(s)" $\widetilde{w} \in \mathrm{Lex}$
- Preserve both root and relevant features of input

Conflation Relation $\sim_r$

- *Binary relation* on strings (words) in $\mathcal{A}^*$ $(\sim_r \subseteq \mathcal{A}^* \times \mathcal{A}^*)$
- Prototypically a true *equivalence relation* *(reflexive, transitive, symmetric)*
  - $w \sim_r v$ holds iff w and v are conflated by relation $\sim_r$
  - equivalence class $[w]_r = \{v \in \mathcal{A}^* \mid w \sim_r v\}$

— Canonicalization Methods —

DTA::CAB System Architecture



Input & Tokenization

CLARIN-D

garbage in ⇝ garbage out

Graphical Errors

(cf. Haaf, Wiegand & Geyken, 2013)

Print	wit ↦	*wett	≠ mit	*"with"*
Transcription	vft ↦	*fest	≠ und (vn̄)	*"and"*
Whitespace	daßſie ↦	*Dash|sie	≠ daß sie	*"that she"*

Tokenization Errors

(cf. Jurish & Würzner, 2013)

Abbrev	Durchl[.] ↦	*torkle	≠ Durchl.	*"highness"*
	Superint[.] ↦	*Super|rind	≠ Superint.	*"superintendent"*
	vnterthän[.] ↦	*Unter|teen	≠ untertän.	*"subserviently"*
Space	ists ↦	*Ists$_{\texttt{NN.gen}}$	≠ ist es	*"is it"*
	ers ↦	*Airs	≠ er es	*"he it"*
Newline	tau[\n]sent ↦	*Tau sehnt	≠ tausend	*"thousand"*

‘Exception’ Lexicon (exlex)

$\mathrm{exlex} : \mathcal{A}^* \to \mathcal{A}^* : w \mapsto \widetilde{w}$

Sketch

- Deterministic type-wise mapping
- Overrides all other canonicalizers

Sources

(Jurish, Drotschmann & Ast, 2013)

- Alignment-based corpus http://kaskade.dwds.de/demo/dtaec
  - 126 volumes (1780-1901), 5.6M tokens, 212k types
- Online error database http://kaskade.dwds.de/demo/caberr
  - includes basic parallel inflection paradigms

Weaknesses

(cf. Kempken et al., 2006; Gotscharek et al., 2009b)

- can’t handle any *ambiguity*
- can’t handle *productive morphological processes*
- alignment-based bootstrapping *(Jurish & Ast, 2015)*

  $\rightsquigarrow$ bogus identity alignments $(w \mapsto w)$

exlex: Bogus Identities

- Assumed: modern edition $\Longrightarrow$ strict orthography
- Implicitly accepted **identity pairs** $(w \mapsto w)$
  - *ca.* 59% types, 87% tokens identical modulo transliteration
- Not always justified by the editions used ***(oops)***

Letter Case	bruder $\mapsto$	*bruder	$\neq$ Bruder	*"brother"*
	trost $\mapsto$	*trost	$\neq$ Trost	*"comfort"*
Extinct Forms	ward $\mapsto$	*ward	$\neq$ wurde	*"was"*
	däuchte $\mapsto$	*däuchte	$\neq$ dünkte	*"seems"*
Prosodic Foot	andre $\mapsto$	*andre	$\neq$ andere	*"other"*
	eignen $\mapsto$	*eignen	$\neq$ eigenen	*"own"*
Dialect	kömmt $\mapsto$	*kömmt	$\neq$ kommt	*"comes"*
	nich $\mapsto$	*nich	$\neq$ nicht	*"not"*
Apostrophes	in's $\mapsto$	*in's	$\neq$ ins	*"into the"*
	s'ist $\mapsto$	*s'ist	$\neq$ es ist	*"it is"*

Deterministic Transliteration (xlit)

CLARIN-D

Sketch

$$w \sim_{\text{xlit}} v :\Leftrightarrow \text{xlit}^*(w) = \text{xlit}^*(v)$$

(Jurish 2008, 2010b,c)

- **Idea:** account for *extinct characters*
- **Implementation:** $\mathcal{O}(1)$ character lookup table
- mostly useful as *preprocessor* for subsequent methods

Successes

Long-'s'	Abſtand	↦ Abstand	*"distance"*
Superscript-'e'	nötig	↦ nötig	*"necessary"*
Diacritics	Hochzît	↦ Hochzeit	*"wedding"*

Failures

Diacritics	wôl	↦ *wol	≠ wohl	*"well"*
Extant Characters	Thür	↦ *Thür	≠ Tür	*"door"*
	vn̄	↦ *vn (⇝ Vene)	≠ und	*"and"*

TAGH Morphology (morph)

Sketch

(Geyken & Hanneforth, 2006)

- Model (modern) morphological processes as WFST
- Provides *weighted target language* for subsequent methods
  - analysis cost ≈ derivational complexity
- "Modern-wins" filtering

$$w \mapsto w :\Leftarrow w \in \pi_1(M_{\text{morph}})$$

Overgeneration: 'modern' form *shouldn't* always win

Andre	↦	*André$_{\text{NE}}$	≠ Andere	*"other"*
from	↦	*from$_{\text{FM.en}}$	≠ fromm	*"pious"*
hiebei	↦	*Hieb|ei	≠ hierbei	*"hereby"*
Zeugnuß	↦	*Zeug|nuß	≠ Zeugnis	*"report"*
keyserlich	↦	*Keys$_{\text{NE}}$|erl$_{\text{NE}}$|ich	≠ kaiserlich	*"imperial"*
Proceſs	↦	*Process	≠ Prozeß	*"process"*

- **Workaround:** *safety heuristics*, e.g. *$_{\text{FM}}$, *$_{\text{NE}}$, *-Ei, *-Nuß, ...

Phonetic Equivalence (eqpho)

CLARIN-D

Sketch

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

- **Idea:** conflate words by *phonetic form* *(Jurish, 2008, 2010b)*
- **Implementation:** text-to-speech rule-set *(Möhler et al., 2001)*
  - modified & compiled as FST M_{pho}
  - online $k-$best equivalence cascade search *(Jurish, 2010a)*

$$C_{\text{eqpho}} := M_{\text{pho}} \circ M_{\text{pho}}^{-1} \circ \pi_1(M_{\text{morph}})$$

Problems

Kopfe	↦	*Kopfweh	≠ Kopf	*"head"*
wes	↦	*Wehs	≠ wessen	*"whose"*
gewegen	↦	*Geh|wegen	≠ gewogen	*"weighted"*
maiestat	↦	*Mais|tat	≠ Majestät	*"majesty"*
Moyſe	↦	*Mäuse	≠ Mose	*"Moses"*
Troglodyt	↦	*Troglo|duett	≠ Troglodyt	*"troglodyte"*

- **Workarounds:** target language pruning, cascade lookup cutoffs

Rewrite Cascade (rw)

$$\begin{aligned} \mathrm{best}_{\mathrm{rw}}(w) &:= \arg\min_{v \in \mathcal{A}^*} [\![M_{\mathrm{rw}} \circ \pi_1(M_{\mathrm{morph}})]\!](w, v) \\ w \sim_{\mathrm{rw}} v &:\Leftrightarrow \mathrm{best}_{\mathrm{rw}}(w) = \mathrm{best}_{\mathrm{rw}}(v) \end{aligned}$$

Sketch

- **Idea:** map words to *nearest* extant type *(Jurish, 2010b,c)*
  - generalized string edit distance *(Damerau, 1964; Levenshtein, 1966)*
  - computable even for infinite lexica *(Mohri, 2002; Jurish, 2010a)*
- **Implementation:** online $(k = 1)$−best cascade search
  - editor WFST M_{rw} compiled from ca. 300 SPE-style rules
  - weight interpolation constants λ_{rw} and λ_{morph}

Problems

ehligen	↦	*Elchen	≠ ehelichen	*"marital"*
gewüntschten	↦	*Gewinde|sekten	≠ gewünschten	*"desired"*
Predigampt	↦	*Birdie|gambit	≠ Predigamt	*"ministry"*
Verdamnuß	↦	*Dominus	≠ Verdammnis	*"damnation"*
Weſſier	↦	*Wessi	≠ Wesir	*"vizier"*

- **Workarounds:** punitive target weights, lookup cutoffs

Token-wise Disambiguation (hmm)

CLARIN-D

Idea

(Mays et al., 1991; Brill & Moore, 2000; Jurish, 2010c)

- Allow high-recall overgeneration at type level
- Recover precision using token-level context

Implementation: Dynamic Hidden Markov Model (HMM)

- **States** are word-conflator pairs $\mathcal{Q} = (\mathcal{W} \cup \{\mathbf{u}\}) \times R$
- **Observations** are input strings $\mathcal{O}_S = \bigcup_{i=1}^{n_S} \{w_i\} \subset \mathcal{A}^*$
- **Transitions** *(static)* $A\big(\langle \widetilde{w}_i, r_i \rangle_{i=1}^m\big) \approx p\big(\widetilde{w}_m \big| \widetilde{w}_1^{m-1}\big)$
- **Lexicon** *(dynamic)*: Maxwell-Boltzmann distribution

$$B\big(\langle \widetilde{w}, r \rangle, w\big) \approx \frac{b^{\beta d_r(w, \widetilde{w})}}{\sum_{r' \in R} \sum_{\widetilde{w}' \in \downarrow[w]_{r'}} b^{\beta d_{r'}(w, \widetilde{w}')}}$$

  - b, β are global model parameters ($b \geq 1, \beta \leq 0$)
  - $d_r(w, \widetilde{w})$ depends on conflator r
- **Lookup** (`moot`): Viterbi Algorithm *(Viterbi, 1967)*

HMM Example

Input	Dete	ſammlete	Steyne	im	Rockermel

HMM Example

Input	Dete	ſammlete	Steyne	im	Rockermel
xlit	*Dete*	*sammlete*	*Steyne*	*im*	*Rockermel*

HMM Example

Input	Dete	ſammlete	Steyne	im	Rockermel
xlit	*Dete*	*sammlete*	*Steyne*	*im*	*Rockermel*
pho	$\emptyset$	$\emptyset$	$\{Steine\}$	$\left\{ \begin{array}{c} im, \\ ihm \end{array} \right\}$	$\{Rockärmel\}$

HMM Example

Input	Dete	sammlete	Steyne	im	Rockermel
xlit	*Dete*	*sammlete*	*Steyne*	*im*	*Rockermel*
pho	$\emptyset$	$\emptyset$	{*Steine*}	{*im*, *ihm*}	{*Rockärmel*}
rw	*Tete*$\langle 1 \rangle$	*sammelte*$\langle 5 \rangle$	*Steine*$\langle 1 \rangle$	*im*$\langle 0 \rangle$	*Rockermehl*$\langle 10 \rangle$

HMM Example

Input	Dete	sammlete	Steyne	im	Rockermel
xlit	*Dete*	*sammlete*	*Steyne*	*im*	*Rockermel*
pho	$\emptyset$	$\emptyset$	{*Steine*}	{*im*, *ihm*}	{*Rockärmel*}
rw	*Tete*$\langle 1 \rangle$	*sammelte*$\langle 5 \rangle$	*Steine*$\langle 1 \rangle$	*im*$\langle 0 \rangle$	*Rockermehl*$\langle 10 \rangle$
hmm	*Dete*	*sammelte*	*Steine*	*im*	*Rockärmel*

HMM Example

Input	Dete	sammlete	Steyne	im	Rockermel
xlit	*Dete*	*sammlete*	*Steyne*	*im*	*Rockermel*
pho	$\emptyset$	$\emptyset$	{*Steine*}	{*im*, *ihm*}	{*Rockärmel*}
rw	*Tete*$\langle 1 \rangle$	*sammelte*$\langle 5 \rangle$	*Steine*$\langle 1 \rangle$	*im*$\langle 0 \rangle$	*Rockermehl*$\langle 10 \rangle$
hmm	*Dete*	*sammelte*	*Steine*	*im*	*Rockärmel*
Output	**Dete**	**sammelte**	**Steine**	**im**	**Rockärmel**

HMM Problems

Sparse and/or Inappropriate Training Data

- many "normal" words treated as *unknown* ($\mathbf{u}$)
- poor handling of historical *syntax*
  - *In deym dienst bestendig bleyben / die trubsall vnns nicht abtreiben*
    M. Luther et al.: Eyn Enchiridion oder Handbuchlein, 1524
  - *Vnd mich deucht / das er den Hyacinthum auff einer ſeite lieb hatte*
    G. Rollenhagen: Vier Bücher wunderbarlicher . . . himmel, 1603
  - *Daher newlichs einer geſagt: die tortur iſt allmächtig*
    F. von Spee: Gewissens-Buch, Von Processen Gegen die Hexen, 1647
  - *kaum ſieht er hinein, ſo erblickt er ein Medaillon*
    Goethe: Wilhelm Meisters Lehrjahre, Bd. 4, 1796

Edelgebohrnen	↦	*Tele|gebahren	≠ edelgeborenen	*"noble-born"*
potz [blitz]	↦	*Potts	≠ potz$_{\text{ITJ}}$	*"gee [whiz]"*
Secretärsstelle	↦	*Secretärsstelle	≠ Sekretärsstelle	*"secretary's job"*
unwiſend	↦	*anwesend	≠ unwissend	*"unknowing"*
Phaſmate	↦	*Faß|matte	≠ Phasmate	*"phantoms"*

- **Workarounds:** language guessing, PoS heuristics, exlex, . . .

“Real” Problems: Target Lexicon

CLARIN-D

Extinct or Missing Target Lexemes

elektroskopische	↦	*Elektro|ska|piefke	*“electroscopic”*
Sakramentierer	↦	*Sakrament|irrer	*“sacramentists”*
glorwürdigen	↦	*Chlor|würdigen$_{\mathtt{VV}}$	*“glory-worthy”*

Extinct or Missing Morphological Processes

aller-	aller|größte	↦	*grüßest	*“greatest of all”*
hier-	hier|nächst	↦	*hin|nagst	*“here next”*
ob-	ob|angeregter	↦	*Oboen|gerechter	*“aforementioned”*
-e	Kopf|e	↦	*Kopf|weh	*“head”*
-ig	standhaft|ig	↦	*stoned|hastig	*“steadfast”*
-lich	auflös|lich	↦	*auslöschlich	*“soluble”*

Lexical Shift / Codification

bälder	↦	*Bälde$_{\mathtt{NN}}$	*“sooner”*
neulicher	↦	*neulich$_{\mathtt{ADV}}$	*“recent”*

More “Real” Problems: Interference

Proper Names

NE ↦ *Lex	Carlowitz	↦	*Gorilla|witz$_{NN}$	≠ Carlowitz
	Philipson	↦	*File|bison$_{NN}$	≠ Philipson
	Moyſe	↦	*Mäuse$_{NN}$	≠ Moses
NE +flect	Jeſ|um	↦	*Jass|ohm	≠ Jesus
	Mathilde|n	↦	*modelten	≠ Mathilde
	George|ns	↦	*Chargen	≠ Georg

Non-Lexical Material

from	↦	*from$_{FM.en}$	≠ fromm	*“pious”*
medicinæ	↦	*Medizin|ah	≠ medicinae	*“medical” (FM.lat)*
mon Dieu	↦	*Mohn Dia	≠ mon Dieu	*“my God” (FM.fr)*
Zn	↦	*Zen	≠ Zn	*“zinc” (chem.)*

Ambiguity / Divergence

wit ↦ mit, wie, weit, wir, wit
widder ↦ wider, wieder, weder, Widder

— Experiments —

Experiment 1: Generative Canonicalizers

CLARIN-D

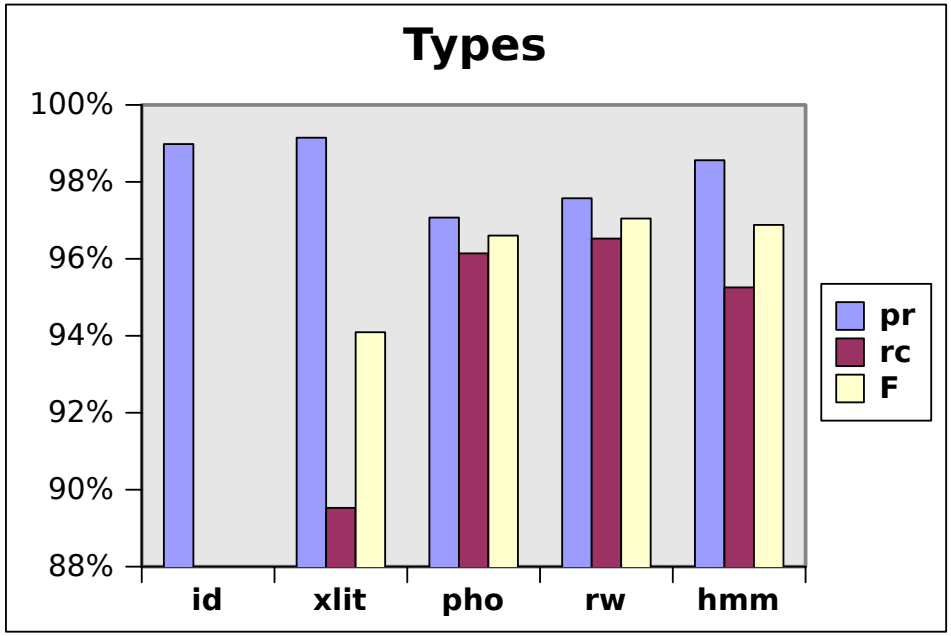
DTA Evaluation Corpus *(Jurish, Drotschmann & Ast, 2013)*

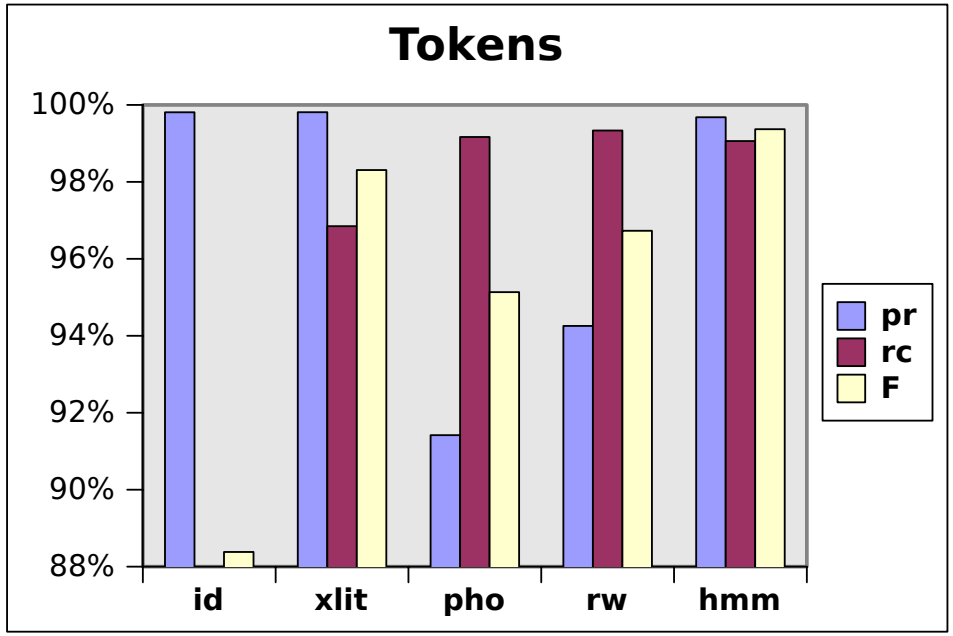
- Gold standard subset of *DTA* Phase 1
  - 13 volumes *(1780 – 1880)*
  - 114,542 tokens; 10,070 types *(alphabetic only)*
- *Canonical cognate* assigned to each token
  - automatic alignment with conventional edition
  - 3-pass manual review
    - Type-wise *(conservative)*
    - Token-wise *(unverified tokens only)*
    - "Suspicious" pairs *(heuristic selection)*

Evaluation Measures

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

Experiment 1: Results





	% Types			% Tokens		
	pr	rc	F	pr	rc	F
id	99.0	59.2	74.1	99.8	79.3	88.4
xlit	***99.1***	89.5	94.1	***99.8***	96.8	98.3
pho	97.1	96.1	96.6	91.4	99.2	95.1
rw	97.6	***96.5***	***97.0***	94.3	***99.3***	96.7
hmm	98.6	95.3	96.9	99.7	99.1	***99.4***

Experiment 2: Alignment-based Lexicon

CLARIN-D

'Evaluation' Corpus $\rightsquigarrow$ Ground-Truth Relevance

$$\mathrm{relevant}(w, \widetilde{w}) := \{(v, \widetilde{v}) : \widetilde{v} = \widetilde{w}\}$$

- Most thoroughly annotated corpus subset
- 13 volumes $\sim$ 320k tokens $\sim$ 28k types (words only)

'Training' Corpus $\rightsquigarrow$ Canonicalization Lexicon (lex)

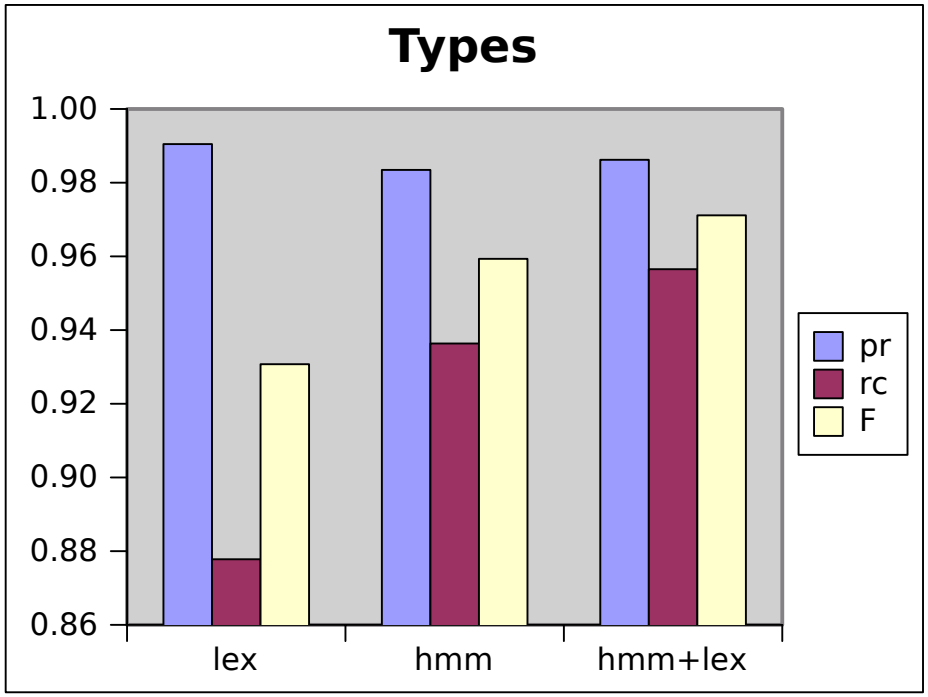
$$\mathrm{lex}(w) = \begin{cases} \arg\max_{\widetilde{w}} f(w, \widetilde{w}) & \text{if } f(w) > 0 \\ w & \text{otherwise} \end{cases}$$

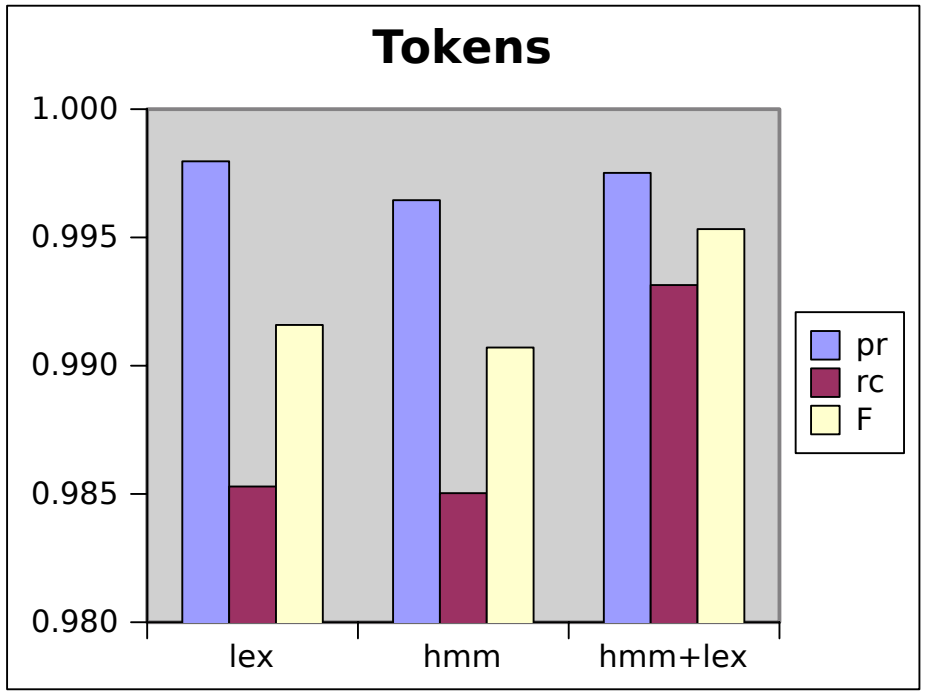
- Strictly disjoint from test corpus (by author)
- 101 volumes $\sim$ 3.5M tokens $\sim$ 158k types (words only)

Evaluation

- Simulated information retrieval $(\mathrm{pr}, \mathrm{rc}, \mathrm{F})$ *(van Rijsbergen, 1979)*
- Tested methods: id, lex, hmm, $\mathrm{hmm{+}lex}$

Experiment 2: Results





	% Types			% Tokens		
	pr	rc	F	pr	rc	F
id	***99.1***	55.7	71.3	***99.8***	78.5	87.9
lex	99.0	87.8	93.1	***99.8***	98.5	99.2
hmm	98.3	93.6	95.9	99.6	98.5	99.1
hmm+lex	98.6	***95.7***	***97.1***	***99.8***	***99.3***	***99.5***

Concluding Remarks

CLARIN-D

Historical Text and Conventional Tools

don't play together nicely "out of the box"

Canonicalization Methods

- static lexicon — *$\sim$ effective but sparse*
- transliteration — *$\sim$ quick and dirty*
- phonetic equivalence — *$\sim$ elegant but coarse*
- rewrite cascade — *$\sim$ flexible but costly*
- HMM disambiguator — *$\sim$ precision recovery*

Advanced Topics

- parameter optimization — *$\geq$27 free parameters*
- rewrite editor — *induction from training pairs*
- HMM smoothing — *PoS, semantics, morphs, . . .*

þe Olde Lasst Slyde

("The End")

===== Gory Details =====

Gory Details

- Related work
- LTS FST construction
- Dictionary lemma instantiation
- Online k-best lookup algorithm
- HMM disambiguator
- Evaluation Measures

Online Demos

- http://kaskade.dwds.de/demo/cab — canonicalization server
- http://kaskade.dwds.de/demo/caberr — error database
- http://kaskade.dwds.de/demo/dtaec — evaluation corpus editor
- http://kaskade.dwds.de/dta-ecp — evaluation corpus pruning
- http://kaskade.dwds.de/dstar/dta — corpus search
- http://www.deutschestextarchiv.de — top-level DTA web site

Related Work

Spelling Correction

CLARIN-D

Damerau (1964); Levenshtein (1966)

- string edit distance ("DL distance")

Kernighan et al. (1990); Church & Gale (1991)

- explicit error models
- local k-gram context disambiguation

Mays, Damerau & Mercer (1991)

- word trigram model
- full sentential context

Brill & Moore (2000)

- generic k-local string-to-string error operations

Diachronic Conflation: English

Hartlib Papers, Sheffield *(16th–18th C.)*
(Rogers & Willett, 1991; Robertson & Willett, 1992; 1993)

- phonetic digest codes
  - SOUNDEX *(Russel, 1918)*
  - PHONIX *(Gadd, 1988; 1990)*
- approximate (inverse) matching
  - string edit distance *(Levenshtein, 1966)*
  - letter n-gram distance *(Pollock & Zamora, 1984)*
- improved recall ($\mathrm{rc}_{\mathrm{typ}} \approx 96\%$) for $k_{\mathrm{Lev}^{-1}} \in \{10, 20\}$

UCREL/UCLan VARD Project, Lancaster *(16th–19th C.)*
(Rayson, Archer & Smith, 2005; Baron & Rayson, 2008; 2011)

- manual or trained letter-replacement heuristics
  - unweighted k-local rules *(Rayson & Baron, p.c.)*
- improved accuracy *vs.* conventional spell-checkers
  - $\approx$ 50% token-wise error reduction ($\mathrm{rw} \approx 46\%$)

Diachronic Conflation: German

RSNSR Project, Duisburg-Essen ***14th–19th C.***

"Regelbasierte Suche in Textdatenbanken mit nichtstandardisierter Rechtschreibung"
Kempken, 2005; Pilz et al., 2006, 2008; Ernst-Gerlach & Fuhr, 2006, 2007, 2010; ...

- inverse canonicalization for information retrieval
- trained edit distance *(Cendrowska, 1987; Ristad & Yianilos, 1998)*
- improved type-wise precision & recall *vs.* DL distance
  - $\Delta(\mathrm{pr}_{\mathrm{typ}}) \approx 39\% \quad ; \quad \Delta(\mathrm{rc}_{\mathrm{typ}}) \approx 31\%$ $(\mathrm{rw} \approx 56\%; 45\%)$

IMPACT Project (OCR), CIS München ***16th–19th C.***

(Gotscharek, Neumann, Reffle, Ringlstetter & Schulz, 2009a; b; c)

- "hypothetical dictionary" for canonicalization
  - unweighted k-local rules *(Ringlstetter, p.c.)*
  - 19th C. tok: $\mathrm{pr} = 97.7\%; \mathrm{rc} = 95.0\%$ $(\mathrm{hmm} = 99.7\%; 99.1\%)$
- "channel error profiles" for precision recovery

LTS FST Construction

LTS FST: Method

CLARIN-D

Step 1: Rule Match Detection

- Aho-Corasick Pattern Matchers (ACPMs)
- Distinguish left *vs.* right of current I/O position
- Intermediate alphabets: matched rule subsets
- Composition of (reversed) partial ACPMs models LTS configurations

Step 2: Output Filter

- Composed with match-detection FST
- Selects applicable rules
- Simulates position increment

Aho-Corasick Pattern Matching

CLARIN-D

Given

- A set $P = \{p_1, \ldots, p_{n_P}\} \subseteq \Sigma^*$ of **patterns**

Task

- **Identify all occurrences** of any pattern $p \in P$ in an input string $w \in \Sigma^*$

Aho-Corasick Algorithm

(Aho & Corasick, 1975)

- Constructs a **pattern-matching FST** $\mathrm{AC}(P) : \mathcal{A}^* \to \wp(P)^*$

- **Linear** runtime complexity: $\mathcal{O} = \mathcal{O}(|w|)$
  - More like $\mathcal{O}(|w| + \sum_{i=1}^{n_P} |p_i| + \sum_{j=1}^{|w|} |P \sim_j w|)$

. . . but who's counting?

ACPM Construction Sketch

Input

- **Prefix Tree Acceptor** (trie) for P

Output

- **FST** $\mathrm{AC}(P) : \mathcal{A}^* \to \wp(P)^*$ such that

$$[\mathrm{AC}(P)](w) = \mathrm{cat}_{i=1}^{|w|} \left\{ p \in P \;\middle|\; p = w_{i-|p|+1}^{i} \right\}$$

Method

- Transition function $\quad \texttt{goto} : Q \times \mathcal{A} \to Q$
- Failure function $\quad \texttt{fail} : Q \to Q$
- Output function $\quad \texttt{out} : Q \to \wp(P)$

Together, `goto`, `fail`, and `out` define a conventional arc-dependent input-deterministic transition function

$$\delta : Q \times \mathcal{A} \to Q \times \wp(P)$$

ACPM Limitations & Workarounds

CLARIN-D

Delayed Output

- ACPM output occurs at pattern terminus
- **Problem** for LTS rule lookahead / -behind
- **Solution**: construct 2 independent ACPMs

Information Loss

- ACPM outputs only pattern sets, not input symbols
- **Problem** for dual-ACPM construction (need both)
- **Solution**: preserve input in lookbehind ACPM

Fixed Increment

- ACPM outputs a pattern set for each input symbol
- **Problem** for LTS rule-dependent increment
- **Solution**: filter out unneeded outputs

LTS FST: Match Detection

CLARIN-D

Left-Context Matching (Lookbehind)

$$
\begin{aligned}
M_L &\cong \mathrm{AC}\big(\{\alpha \mid (\alpha[\beta]\gamma \to \pi) \in R\}\big) \\
&: \mathcal{A}_w^* \to (\mathcal{A}_w \times \wp(R))^* \\
&: w \mapsto \mathrm{cat}_{i=1}^{|w|} \left\langle w_i, \left\{ (\alpha[\beta]\gamma \to \pi) \in R \;\middle|\; w_{i-|\alpha|}^{i-1} = \alpha \right\} \right\rangle
\end{aligned}
$$

Target & Right-Context Matching (Lookahead)

$$
\begin{aligned}
M_R &\cong \mathrm{rev}\Big(\mathrm{AC}\big(\{\mathrm{rev}(\beta\gamma) \mid (\alpha[\beta]\gamma \to \pi) \in R\}\big)\Big) \\
&: (\mathcal{A}_w \times \wp(R))^* \to \wp(R)^* \\
&: \langle w_i, S_i \rangle_{i=1}^{|w|} \mapsto \\
&\qquad \mathrm{cat}_{i=1}^{|w|} \left(S_i \cap \left\{ (\alpha[\beta]\gamma \to \pi) \in R \;\middle|\; w_i^{i+|\beta\gamma|-1} = \beta\gamma \right\} \right)
\end{aligned}
$$

Match Detection FST (Current Match Subset)

$$
M_{LR} = M_L \circ M_R \;:\; \mathcal{A}_w{}^* \to \wp(R)^* \;:\; w \mapsto \mathrm{cat}_{i=1}^{|w|} R \sim_i w
$$

LTS FST: Output Filter

Notation

- Let $\mathcal{A}_{LR} \subseteq \wp(R)$ be the output alphabet of M_{LR}
- For each $S \in \mathcal{A}_{LR}$, let $(\alpha_S[\beta_S]\gamma_S \rightarrow \pi_S) = \min_{\prec} S$

Output Filter Construction

- Define output filter $M_O : {\mathcal{A}_{LR}}^* \rightarrow {\mathcal{A}_p}^*$ as:

$$M_O = \left(\bigcup_{S \in \mathcal{A}_{LR}} \left[\underbrace{(S : \pi_S)}_{\text{apply}} \underbrace{(\mathcal{A}_{LR} \times \{\varepsilon\})^{|\beta_S|-1}}_{\text{increment}} \right] \right)^*$$

LTS Transducer

- Then, the desired LTS FST can be defined as:

$$M_{\text{pho}} = (M_{LR} \circ M_O) = (M_L \circ M_R \circ M_O) : \mathcal{A}_w^* \rightarrow \mathcal{A}_p^*$$

Lemma Instantiation

Lemma Instantiation: Sketch

CLARIN-D

Idea

- Exploit dictionary-corpus structure
- Assume each quote $q \in \mathcal{Q}$ contains at least one instance $i \in \mathcal{W}$ of the associated dictionary lemma $\ell \in \mathcal{L}$

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: Method

Empirical Probability Estimates

$$\begin{aligned}
\mathrm{P}(\ell, i) &= \frac{\sum_{w_i \in \mathrm{pho}^{-1}(i)} \sum_{w_\ell \in \mathrm{pho}^{-1}(\ell)} \mathrm{f}(W{=}w_i, L{=}w_\ell)}{|Corpus|} \\
\mathrm{P}(\ell) &= \sum_i \mathrm{P}(\ell, i) \\
\mathrm{P}(i) &= \sum_\ell \mathrm{P}(\ell, i)
\end{aligned}$$

Pointwise Mutual Information

$$\begin{aligned}
\mathrm{I}(\ell, i) &= \log_2 \frac{\mathrm{P}(\ell,i)}{\mathrm{P}(\ell)\mathrm{P}(i)} \\
\tilde{\mathrm{I}}(i|\ell) &= \frac{\mathrm{I}(\ell,i) - \min \mathrm{I}(\ell,\mathcal{W})}{\max \mathrm{I}(\ell,\mathcal{W}) - \min \mathrm{I}(\ell,\mathcal{W})} \\
\tilde{\mathrm{I}}(\ell|i) &= \frac{\mathrm{I}(\ell,i) - \min \mathrm{I}(\mathcal{L},i)}{\max \mathrm{I}(\mathcal{L},i) - \min \mathrm{I}(\mathcal{L},i)}
\end{aligned}$$

Edit-Distance Threshhold

$$d_{\max}(\ell, i) = \min\{|\ell|, |i|\} - 1$$

Lemma Instantiation: Method (cont.)

CLARIN-D

Phonetic Similarity

$$\mathrm{sim}(\ell, i) \quad = \quad \begin{cases} \frac{d_{\max}(\ell,i) - d_{\mathrm{Lev}}(\ell,i)}{d_{\max}(\ell,i)} & \text{if } d_{\mathrm{Lev}}(\ell, i) \leq d_{\max}(\ell, i) \\ 0 & \text{otherwise} \end{cases}$$

Instantiation Likelihood Function L

$$\mathrm{L}(i, \ell) \quad = \quad \frac{\mathrm{sim}(\ell,i) \times (\tilde{\mathrm{I}}(\ell|i) + \tilde{\mathrm{I}}(i|\ell))}{2}$$

Instantiation Heuristic

$$\begin{aligned} \mathcal{I}(\cdot) \quad &: \quad \mathcal{Q} \to \mathcal{W} \\ &: \quad q \mapsto \arg\max_{w \in q} \mathrm{L}(\mathrm{pho}(w), \mathrm{pho}(\mathrm{lemma}(q))) \\ w \sim_{\mathrm{li}} v \quad &:\Leftrightarrow \quad (w \sim_{\mathrm{pho}} v) \text{ or} \\ &\quad \Big(\mathrm{lemma}\big(\mathcal{I}^{-1}(w)\big) \cap \mathrm{lemma}\big(\mathcal{I}^{-1}(v)\big) \neq \emptyset\Big) \end{aligned}$$

Online k-Best Lookup

Modified Dijkstra (1959) Algorithm

```
function DIJKSTRA-KBEST-STRINGS(w⃗, C, k, c_max, x_max)
    S := {⟨q_0, ε⟩}                                          /* Initialize */
    d[·] := {⟨q_0, ε⟩ ↦ 1̄}                                   /* Sparse partial minimum-cost map */
    d_Π,k[·] := ∅                                            /* Track k-best final configurations */
    while S ≠ ∅                                              /* Main loop */
        ⟨q, s⟩ := arg min_{⟨q,s⟩∈S, <_K} d[q, s]             /* Best-first search */
        S := S \ {⟨q, s⟩}
        if x_max = 0 then break                              /* Too many extractions */
        x_max := x_max − 1
        if d[q, s] >_K c_max then break                      /* Cost upper-bound */
        if q ∈ F then                                        /* Finality check */
            d_Π,k[q, s] := d[q, s]
            if |d_Π,k| = k then break                        /* Greedy termination */
        foreach e ∈ ARCS(w⃗, C, q)                            /* Outgoing arcs (online expansion) */
            d′ := d[q, s] ⊗ c[e]                             /* Accumulate */
            s′ := s ⌢ o[e]                                   /* Append */
            if d′ <_K COST(d[·], ⟨n[e], s′⟩) then            /* Relax */
                d[n[e], s′] := d′
                S := S ∪ {⟨n[e], s′⟩}
    return d_Π,k[·]                                          /* Final output */
```

k-Best Lookup: Subroutines

```
function ARCS(w⃗, C, q)                                   /* Online cascade composition */
    return EXPAND-ARCS(C, q, 1, w⃗[q[1]]) ∪ EXPAND-ARCS(C, q, 1, ε)

function EXPAND-ARCS(C, q, i, a)                         /* Recursive guts for ARCS() */
    local A := ∅
    foreach e ∈ E_i ∪ { (q[i], q[i], ε, ε, 1̄) } : p[e] = q[i] & i[e] = a
        if i = |C| then
            A := A ∪ {e}
        else
            foreach e′ ∈ EXPAND-ARCS(C, q, i + 1, o[e])
                A := A ∪ { (⟨p[e], p[e′]⟩, ⟨n[e], n[e′]⟩, i[e], o[e′], c[e] ⊗ c[e′]) }
    return A

function COST(d[·], q)                                   /* Minimum cost tracking for sparse partial map */
    if d[q] defined then return d[q]
    return 0̄
```

Complexity

$$
\begin{aligned}
\mathcal{O}(\textsc{Dijkstra}) &= \mathcal{O}\big(|E| + |V| \log |V|\big) \\
\mathcal{O}(\textsc{Kbest}[-x_{\max}]) &= \mathcal{O}\Big(|E_{\Pi,k}| + \mathcal{O}(\textsc{arcs})\, |V_{\Pi,k}| \log |V_{\Pi,k}|\Big) \\
\mathcal{O}(\textsc{Kbest}[+x_{\max}]) &= \mathcal{O}\big(x_{\max} \times \mathcal{O}(\textsc{arcs}) \log x_{\max}\big) \\
\mathcal{O}(\textsc{arcs}) &= \mathcal{O}\big(\textstyle\prod_{i=1}^{n_C} \deg(q \in Q_i)\big)
\end{aligned}
$$

where:

$$
\begin{aligned}
V_{\Pi,k} &= \Big\{ \langle q, s\rangle \in Q \times \mathcal{B}^* \;\Big|\; d[q,s] \leq_{\mathcal{K}} \max\big(\mathrm{rng}(d_{\Pi,k})\big) \Big\} \\
E_{\Pi,k} &= \Big\{ \big\langle \langle q, s\rangle, e\big\rangle \in V_{\Pi,k} \times E \;\Big|\; p[e] = q \Big\}
\end{aligned}
$$

- Fibonacci Heap *(Fredman & Tarjan, 1987)*
- Linear sorted arc expansion

Degenerate Cycles

A cycle $\pi \in E^*$ is ***"degenerate"*** if $\exists \pi' \in E^*, \forall n \in \mathbb{N}$:

- $\pi'\pi \quad \in \quad \Pi(q_0, Q)$ /* *accessible* */
- $p[\pi] \quad = \quad n[\pi]$ /* *cyclic* */
- $i[\pi] \quad = \quad \varepsilon$ /* *empty input* */
- $o[\pi] \quad \neq \quad \varepsilon$ /* *non-empty output* */
- $c\left[\pi^n\right] \quad \leq_{\mathcal{K}} \quad c_{\max}$ /* *cost-free* */

Consequence: Infinite Loop

- Accessibility $\rightsquigarrow$ *enqueue-ability*
- Empty input $\rightsquigarrow$ *word-length independence*
- Non-empty output $\rightsquigarrow$ *new queue items*
- Cost freedom $\rightsquigarrow$ *iterated* RELAX, EXTRACT

$\rightsquigarrow$ ***KABOOM!***

HMM Disambiguator

Basic Model

Common Variables

- $\mathcal{W} \subset \widetilde{\mathcal{A}}^*$ — known extant words
- $\mathbf{u} \notin \mathcal{W}$ — designated "unknown word" symbol
- $S = \langle w_1, \ldots, w_{n_S} \rangle$ — current input sentence
- $R = \{r_1, \ldots, r_{n_R}\}$ — (opaque) type-wise conflators
- $m \in \mathbb{N}$ — model order (m-grams)

Disambiguator $D = \langle \mathcal{Q}, \mathcal{O}_S, \Pi, A, B_S \rangle$

- $\mathcal{Q} = (\mathcal{W} \cup \{\mathbf{u}\}) \times R$
- $\mathcal{O}_S = \bigcup_{i=1}^{n_S} \{w_i\}$
- $\Pi : \mathcal{Q} \to [0,1] : q \mapsto p(Q_1 = q)$
- $A : \mathcal{Q}^m \to [0,1] : q_1^m \mapsto p(Q_i = q_m | Q_{i-m+1}^{i-1} = q_1^{m-1})$
- $B_S : \mathcal{Q} \times \mathcal{O}_S \to [0,1] : \langle q, o \rangle \mapsto p(O = o | Q = q)$

General Properties

Probability Computation

- Sentence probability
  - $p(S = w_1^{n_S}) = \sum_{q_1^{n_S} \in \mathcal{Q}^{n_S}} p(S = w_1^{n_S}, Q = q_1^{n_S})$
- Path probability
  - $p(S = w_1^{n_S}, Q = q_1^{n_S}) = \prod_{i=1}^{n_S} p(q_i | q_{i-m+1}^{i-1}) p(w_i | q_i)$

Markov Assumptions

- m-local state dependencies
  - $p(q_i | q_1^{i-1}, w_1^{i-1}) = p(q_i | q_{i-m+1}^{i-1})$
- Context-independent observations (surface forms)
  - $p(w_i | q_1^i, w_1^{i-1}) = p(w_i | q_i)$

Transition Probabilities

Assumptions

- $p(Q = \langle \widetilde{w}_q, r_q \rangle) = p(W = \widetilde{w}_q) p(R = r_q)$
  : independence of extant form from conflator
- $p(R = r) = \frac{1}{n_R}$ *: conflators are uniformly distributed*

Transition Probability Estimation

- $\Pi(q) :\approx p\left(W_1 = \widetilde{w}_q\right) / n_R$
- $A(q_1^m) :\approx p\left(W_i = \widetilde{w}_{q_m} \middle| W_{i-m+1}^{i-1} = \widetilde{w}_{q_1}^{q_{m-1}}\right) / n_R$
- $\equiv$ word m-gram model for contemporary forms

Smoothing

- Unknown words: $\mathrm{f}(\mathbf{u}) :\approx \frac{1}{2}$ *(Lidstone, 1920)*
- Linear interpolation of m-grams *(Jelinek & Mercer, 1980; 1985)*

Lexical Probabilities

Maxwell-Boltzmann Approximation

$$B\big(\langle \widetilde{w}, r\rangle, w\big) :\approx \frac{b^{\beta d_r(w,\widetilde{w})}}{\sum_{r'\in R}\sum_{\widetilde{w}'\in\downarrow[w]_{r'}} b^{\beta d_{r'}(w,\widetilde{w}')}}$$

Parameters

$$\begin{aligned}
b &= 2 \\
\beta &= -1 \\
R &= \{\mathrm{xlit}, \mathrm{pho}, \mathrm{rw}\} \\
\downarrow[w]_r &= [w]_r \cap \mathcal{W} \\
d_{\mathrm{xlit}}(w,\widetilde{w}) &= 2/|w| && \text{if } \widetilde{w} = \mathrm{xlit}^*(w) \\
d_{\mathrm{pho}}(w,\widetilde{w}) &= 1/|w| && \text{if } \widetilde{w} \in \downarrow[w]_{\mathrm{pho}} \\
d_{\mathrm{rw}}(w,\widetilde{w}) &= [\![M_{\mathrm{rw}} \circ A_{\mathrm{Lex}}]\!](w,\widetilde{w})/|w| && \text{if } \widetilde{w} = \mathrm{best}_{\mathrm{rw}}(w)
\end{aligned}$$

Runtime Disambiguation

Viterbi Algorithm

(Viterbi, 1967)

$$\hat{q}_1^{n_S} = \textsc{Viterbi}(S, D) = \arg \max_{q_1^{n_S} \in \mathcal{Q}^{n_S}} p(q_1^{n_S}, S|D)$$

Disambiguator Output

$$\hat{S} = \langle \hat{w}_1, \ldots, \hat{w}_{n_S} \rangle$$

where:

$$\hat{w}_i := \begin{cases} \text{witness}\big(\downarrow[w]_{r_{\hat{q}_i}}\big) & \text{if } \widetilde{w}_{\hat{q}_i} = \mathbf{u} \\ \widetilde{w}_{\hat{q}_i} & \text{otherwise} \end{cases}$$

Evaluation Measures

Evaluation Measures: Basic

Common Definitions

- canonicalizers $C = \{c_1, \ldots, c_{n_C}\}$
- test corpus $G = \langle g_1, \ldots, g_{n_G} \rangle$
  - $g_i = \langle w_i, \widetilde{w}_i, [w_i]_{c_1}, \ldots, [w_i]_{c_{n_C}} \rangle \in \mathcal{A}^* \times \mathcal{A}^* \times \wp(\mathcal{A}^*)^{n_C}$
- queries $Q = \bigcup_{i=1}^{n_G} \{\widetilde{w}_i\}$

Relevance & Retrieval

for each canonicalizer $c \in C$ and query $q \in Q$, define:

- by token

$$\begin{aligned} \mathrm{relevant}_{\mathrm{tok}}(q) &= \{i \in \mathbb{N} \mid q = \widetilde{w}_i\} \\ \mathrm{retrieved}_{\mathrm{tok},c}(q) &= \{i \in \mathbb{N} \mid q \in [w_i]_c\} \end{aligned}$$

- by type

$$\begin{aligned} \mathrm{relevant}_{\mathrm{typ}}(q) &= \bigcup_{i \in \mathrm{relevant}_{\mathrm{tok}}(q)} \{w_i\} \\ \mathrm{retrieved}_{\mathrm{typ},c}(q) &= \bigcup_{i \in \mathrm{retrieved}_{\mathrm{tok},c}(q)} \{w_i\} \end{aligned}$$

Precision & Recall

For each evaluation mode $m \in \{\mathrm{tok}, \mathrm{typ}\}$, define:

Precision

$$\mathrm{pr}_{m,c} = \frac{\left|\bigcup_{q \in Q} \mathrm{retrieved}_{m,c}(q) \cap \mathrm{relevant}_m(q)\right|}{\left|\bigcup_{q \in Q} \mathrm{retrieved}_{m,c}(q)\right|}$$

- Likelihood of ***relevance***, given ***retrieval***

Recall

$$\mathrm{rc}_{m,c} = \frac{\left|\bigcup_{q \in Q} \mathrm{retrieved}_{m,c}(q) \cap \mathrm{relevant}_m(q)\right|}{\left|\bigcup_{q \in Q} \mathrm{relevant}_m(q)\right|}$$

- Likelihood of ***retrieval***, given ***relevance***

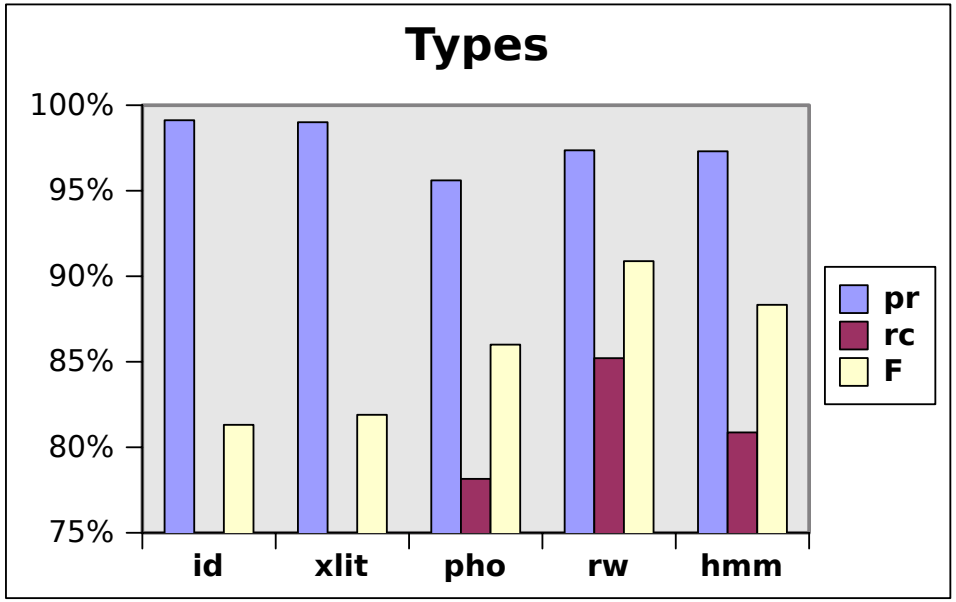
Precision-Recall Composite F

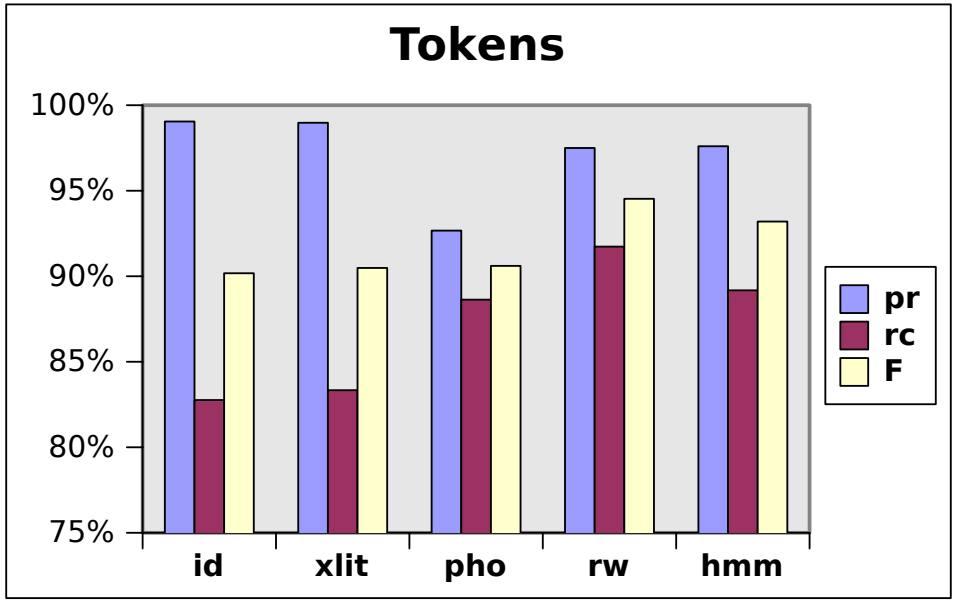
$$\mathrm{F}_{m,c} = \frac{2 \times \mathrm{pr}_{m,c} \times \mathrm{rc}_{m,c}}{\mathrm{pr}_{m,c} + \mathrm{rc}_{m,c}}$$

- Unweighted harmonic average

(van Rijsbergen, 1979: "E")

Evaluation: DWB-1 Verse





	% Types			% Tokens		
	pr	rc	F	pr	rc	F
id	***99.1***	68.9	81.3	***99.0***	82.8	90.2
xlit	99.0	69.8	81.9	99.0	83.3	90.5
pho	95.6	78.2	86.0	92.7	88.6	90.6
rw	97.4	***85.2***	***90.9***	97.5	***91.7***	***94.5***
hmm	97.3	80.9	88.3	97.6	89.2	93.2