Decision Lists (Yarowsky, 1994)

Susanne Fertmann
22. November 2010

Ein Vortrag basierend auf:
DECISION LISTS FOR LEXICAL AMBIGUITY RESOLUTION: Application to Accent Restoration in Spanish and French, David Yarowsky

Problem

- In einem akzentfreien Text sollen Akzente gesetzt werden (Großbuchstaben in Überschriften)

> appeler l'autre *cote* de l'atlantique → côté
> vivre sur notre *cote* ouest toujours verte → côte

- cote → côté, côte, cote, coté
- Klassifizierung → Taggen
- Ambiguität: semantisch unterschiedliche Wörter, selbe Repräsentation

- Bsp. im Deutschen: Großschreibung
  - LEBEN (Verb vs. Nomen) → Leben/leben
  - LIEBE (Adjektiv vs. Verb vs. Nomen) → Liebe/liebe

Ziel

- Klassifikator um Akzente richtig zu setzen = Decision List

LogL	Evidence	Classification
8,28	PREPOSITION que *terminara*	terminara
7,24	de que *terminara*	terminara
7,14	para que *terminara*	terminara
6,87	y *terminara*	terminará
6,64	WEEKDAY (within ±k words)	terminará
5,82	NOUN que *terminara*	terminará
5,45	domingo (within ±k words)	terminará

Überblick

- Problem
- Algorithmus zum Erstellen der Decision Lists
- Evaluierung
- Ausblick

Häufige Ambiguitäten

- Spanisch:
  - -o/-ó (completo/completó)
    - Präsens vs. Präteritum, oft auch noch Nomen
  - -ara/ará (terminara/terminará)
    - Subjunktiv vs. Futur
    - 3.Pers. Sg. bei allen Verben auf –ar (schwerer mit Part-of-speech-Tagging)
  - secretaria/secretaría, sabana/sábana
    - Rein semantisch
- Französisch:
  - -e/-é: Präsens vs. Vergangenheit
  - traité/traite, marche/marché, côte/côté
    - Viele rein semantische

Vorteile der Akzentrestauration

- Unterschiedliche Typen von Ambiguität: semantisch und syntaktisch
- Große Datenmengen
  - durch Löschen von Akzenten
  - leichte Evaluierung
  - nicht subjektiv und keine Handarbeit
- Kommerzielle Anwendungen (Grammatik- und Rechtschreibprüfung, automatische Akzentsetzung beim Tippen)

Der Algorithmus: Überblick

- Ambiguitäten finden
- Kontexte sammeln
- Kollokationen untersuchen
- Sortieren in Decision Lists
  - Optional: Pruning & Interpolation
  - Decision Lists für allgemeine Klassen
- Anwendung

Schritt 1: Ambiguitäten finden

De-accented form	Accent Pattern	%	Number
cout	coût	100%	330
coute	coûté	53%	107
	coûte	47%	96
cote	côté	69%	2645
	côte	28%	1040
	cote	3%	99
	coté	<1%	15

- Grundlage: Korpus mit Akzenten
- Zeitungen: 20Mio (frz) bzw. 49Mio (span) Wörter

Schritt 2:
Kontexte sammeln

- ±k Wortfenster → Trainingsset:

Pattern	Context
côté	du laisser de *cote* faute de temps
côté	appeler l'autre *cote* de l'atlantique
côté	passe de notre *cote* de la traniere
côte	vivre sur notre *cote* ouest toujours verte
côte	creer sur la *cote* du labrador des
côte	ravaillaient *cote* a cote, ils

Schritt 3:
Kollokationen untersuchen

- Kollokation hier im weiteren Sinne
  - Wörter in der Umgebung
- Bestimmte Kollokationen erfordern bestimmte Akzent-Pattern
- Auswahl der nützlichsten Kollokationen

Schritt 3:
Kollokationen untersuchen

Position	Collocation	côte	côté
-1 w	du *cote*	0	536
	la *cote*	766	1
	un *cote*	0	216
	notre *cote*	10	70
+1 w	*cote* ouest	288	1
	cote du	55	156
+1w,+2w	*cote* du gouvernement	0	62
-2w,-1w	cote a *cote*	23	0
±k w	poisson (in ±k words)	20	0
±k w	ports (in ±k words)	22	0
±k w	opposition (in ±k words)	0	39

Schritt 3:
Kollokationen untersuchen

- sprachspezifisches Wissen nicht notwendig, aber nützlich
- Französisch: kein weiteres Sprachwissen
  - Basis: nur Verteilung im Originaltext
  - Schnell, günstig
- Spanisch: zusätzliche Informationen (ermöglichen Generalisierungen)
  - Morphologische Analyse
  - Definieren von Wortklassen (weekday, month)
  - Wortarten (grobe Analyse: Nomen, Adjektiv, Artikel, etc.)
    - PoS-Tagger waren damals teuer
    - Informationen aus Wörterbüchern
    - Annäherung durch Suffixe (-aba, -amos)

Schritt 3:
Kollokationen untersuchen

Collocation	terminara	terminará
PREPOSITION que *terminara*	32	0
de que *terminara*	15	0
para que *terminara*	14	0
NOUN que *terminara*	0	13
carrera que *terminara*	0	3
reunion que terminara	0	2
que *terminara*	42	37
WEEKDAY (within ±k words)	0	23
domingo (within ±k words)	0	10

Schritt 4:
Sortieren in Decision Lists

- Log-likelihood

$$Abs(Log(\frac{Pr(Accent\_Pattern_1|Collocation_i)}{Pr(Accent\_Pattern_2|Collocation_i)}))$$

Collocation	terminara	terminará
PREPOSITION que *terminara* de que *terminara*	32 15	0 0
que *terminara*	42	37

- Abs(log($\frac{P(terminara|que\_t)}{P(terminará|que\_t)}$)) = Abs(log($\frac{42}{37}$)) = 0,055
- Kollokationen, die am eindeutigsten zu *einem* Pattern gehören, haben einen höheren LL
  → **sortiert** Kollokationen **nach Informativität**
- Smoothing nötig

Schritt 4:
Sortieren in Decision Lists

LogL	Evidence	Classification
8,28	PREPOSITION que *terminara*	terminara
7,24	de que *terminara*	terminara
7,14	para que *terminara*	terminara
6,87	y *terminara*	terminará
6,64	WEEKDAY (within ±k words)	terminará
5,82	NOUN que *terminara*	terminará
5,45	domingo (within ±k words)	terminará

- Klassifikator für neue Beispiele
- Die oberste passende Regel wird gewählt
- Klassifikator jetzt eigentlich fertig

Schritt 5:
Optional: Pruning & Interpolation

Wahrscheinlichkeiten

Global	vs.	**Residual**
(= vom gesamten Trainingset) Besser geschätzt, weniger relevant		(= vom restlichen Set, wenn keine Regel passt) Relevanter, schlecht geschätzt

- Kombinieren? In welchem Verhältnis?
  → **Interpolation**
- Globale Wahrscheinlichkeiten reichen für dieses Problem aus, für alles andere lohnt der Aufwand nicht

Schritt 5:
Optional: Pruning & Interpolation

- Pruning: „redundancy by subsumption“

LogL	Evidence	Classification
8,28	PREPOSITION que *terminara*	terminara
~~7,24~~	~~de que *terminara*~~	~~terminara~~
~~7,14~~	~~para que *terminara*~~	~~terminara~~
6,87	y *terminara*	terminará
6,64	WEEKDAY (within ±k words)	terminará
5,82	NOUN que *terminara*	terminará
~~5,45~~	~~domingo (within ±k words)~~	~~terminará~~

- Ist ein Bigramm schon desambiguierend, werden Trigramme erst gar nicht generiert

Schritt 5:
Optional: Pruning & Interpolation

- Pruning:
  - Regeln, die zu mehr falschen als richtigen Entscheidungen führen, werden entfernt (Overfitting)
  - Bei Platz-Problemen: nicht aufgerufene Regeln werden gelöscht
  - wichtige Informationen können verloren gehen
  - 3% Verlust in Performanz vs. 90% Reduzierung von Platz
  - Hier nicht angewendet

Schritt 6:
Decision Lists für allgemeine Klassen

- Ähnliche Ambiguitäten → generelle Decision List
- -ara/-ará bei spanischen Verben auf –ar (Subjunktiv vs. Futur)
- Erfasst auch Verben, die in den Trainingsdaten nicht angemessen abgedeckt werden
- Jedes Wort auf -ara ähnliche Repräsentation im Trainingsset, damit tatsächliche Eigenschaften der Klasse und nicht des häufigsten Wortes
- Test, ob ein Wort in diese Klasse passt, sonst wortspezifische Decision List
- Im spanischen: -o/-ó, -e/é, -ara/-ará, -aran/-arán

Der Algorithmus: Überblick

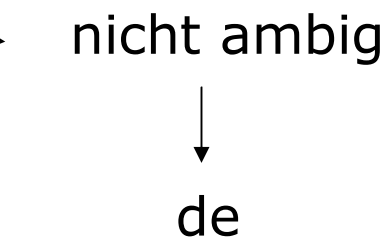
- Ambiguitäten finden
- Kontexte sammeln
- Kollokationen untersuchen
- Sortieren in Decision Lists
  - Optional: Pruning & Interpolation
  - Decision Lists für allgemeine Klassen
- Anwendung

Anwendung von Decision Lists

EL HORARIO DE VERANO TERMINARA ESTE DOMINGO 25 DE OCTUBRE

aplicación de esta medida generó un ahorro estimado de mil 300 millones de kilowatts-hora en el país. […]Este domingo 25 de octubre, a las 02:00 de la mañana, termina el Horario de Verano, por lo que se invita a la población a atrasar una hora sus relojes la noche del sábado 24 de octubre.
La Comisión Nacional para el Uso Eficiente de la Energía (CONUEE) informó que la aplicación de esta medida generó un ahorro estimado de mil 300 millones de kilowatts-hora en el país en este año, lo cual equivale aproximadamente al consumo total anual de los estados de Campeche, Colima o Nayarit en el año 2008.
En los siete meses de implementación, el horario de verano evitó emitir a la atmósfera alrededor de 1.6 millones de toneladas equivalentes de bióxido de carbono, el principal gas de efecto invernadero, y se dejaron de quemar, cerca de 3.1 millones de barriles equivalentes de petróleo para la generación de electricidad.

http://www.informador.com.mx/mexico/2009/147934/6/horario-de-verano-terminara-el-domingo-25-de-octubre.htm



Anwendung von Decision Lists

EL HORARIO DE VERANO TERMINARA ESTE DOMINGO 25 DE OCTUBRE

aplicación de esta medida generó un ahorro estimado de mil 300 millones de kilowatts-hora en el país. […]Este domingo 25 de octubre, a las 02:00 de la mañana, termina el Horario de Verano, por lo que se invita a la población a atrasar una hora sus relojes la noche del sábado 24 de octubre.
La Comisión Nacional para el Uso Eficiente de la Energía (CONUEE) informó que la aplicación de esta medida generó un ahorro estimado de mil 300 millones de kilowatts-hora en el país en este año, lo cual equivale aproximadamente al consumo total anual de los estados de Campeche, Colima o Nayarit en el año 2008.
En los siete meses de implementación, el horario de verano evitó emitir a la atmósfera alrededor de 1.6 millones de toneladas equivalentes de bióxido de carbono, el principal gas de efecto invernadero, y se dejaron de quemar, cerca de 3.1 millones de barriles equivalentes de petróleo para la generación de electricidad.

http://www.informador.com.mx/mexico/2009/147934/6/horario-de-verano-terminara-el-domingo-25-de-octubre.htm

ambig

Log L	Evidence	Classification
8,28	PREPOSITION que *terminara*	terminara
6,87	y *terminara*	terminará
6,64	WEEKDAY (within ±k words)	terminará
5,82	NOUN que *terminara*	terminará

terminará

Anwendung von Decision Lists

- Jedes Wort aus einem Text wird in einer Tabelle nachgeschlagen
- Nicht-ambige Wörter: das richtige Pattern wird ausgegeben
- Ambige Wörter: Weiterleiten zur Decision List (des Wortes oder der Klasse)
- Die am weitesten oben stehende passende Regel wird gesucht (Kontext)
- Das wahrscheinlichste Pattern wird ausgegeben
- Wenn keine Regel passt: Default-Eintrag

Anwendung von Decision Lists

- Kombinieren aller passenden Regeln vs. Benutzten einer einzelnen
  - Selten unterschiedliche Entscheidung
  - Wörter haben in einer bestimmten Kollokation nur *eine* Bedeutung (Yarowsky, 1993)
  - Nur eine Regel: effizienter, schneller, leichteres Generalisieren

Evaluierung

- Objektiv und automatisch möglich, da Originalkorpus mit Akzenten
- Problem: Fehler im Originalkorpus i/í, ca. 4% der Akzente fehlen

Performanz

Ambige Wörter (18% der Token):

Type	Frequency	Agreement	Baseline
-o/-ó	81%	98%	86%
-ara/-ará, -aran/-arán	4%	92%	84%
Other	15%	98%	94%
Total		98%	93%

Performanz

Pattern 1	Pattern 2	Agreement	Basel.	N
terminara	terminará	82,9%	59%	218
llegara	llegará	78,4%	64%	860
anuncio	anunció	98,4%	57%	9459
regalo	regaló	90,7%	56%	280

Pattern 1	Pattern 2	Agreement	Basel.	N
décidé	décide	96,5%	64%	3667
côté	côte	98,1%	69%	3865

Zusammenfassung

- Decision Lists = geordnete Liste mit Regeln zum Klassifizieren
- Einfacher, effizienter Algorithmus
- Anwendung: Akzentrestauration

- Auch anwendbar bei anderen Desambiguierungs-Problemen?
  - Großschreibung

Ausblick

- Pony: Tier vs. Haarschnitt
  - Der Pony → Haarschnitt
  - Das Pony → Tier
  - Zucht (±k Wörter) → Tier
  - Pferd (±k Wörter) → Tier
  - Frisur (±k Wörter) → Haarschnitt
  - lang (±k Wörter) → Haarschnitt

  - Homonymie
  - Funktioniert das auch bei komplizierteren Ambiguitäten?