Neue Verfahren und Evaluierungsmaße für Anfragesegmentierungen

Anna Beyer

Bauhaus-Universität Weimar
anna.beyer@uni-weimar.de

Bachelorverteidigung
3. Februar 2012

Gliederung

Motivation

1 Milliarde Suchanfragen täglich

Quelle: http://www.google.com/insidesearch/underthehood.html

Motivation


```
tokio hotel bill
```

Was sucht der Nutzer?

Motivation

`tokio  hotel bill`

Quelle: http://ahoy.tk-jk.net/MoreImages4/WillyHotelBill.jpg

Motivation

`tokio hotel bill`

`tokio hotel bill`

Quelle: http://ahoy.tk-jk.net/MoreImages4/WillyHotelBill.jpg

Quelle: http://www.popartuk.com/g/l/lg840631.jpg

Anfragesegmentierung

Anfragesegmentierung

Definition

Einteilung einer Suchanfrage in Sinneinheiten

Syntax

Kennzeichnung der Segmente durch Anführungszeichen

z. B. `"tokio hotel" "bill"`

Vorteil

Verbesserte Retrieval Performance

[Li et al., SIGIR 2011]

Anfragesegmentierung

Situation

Segmentierung bei weniger als 1,12 % der Suchanfragen

[White und Morris, SIGIR 2007]

Anfragesegmentierung

Situation

Segmentierung bei weniger als 1,12 % der Suchanfragen

[White und Morris, SIGIR 2007]

Lösung

Automatische Segmentierung vor der Suche

Segmentierungsverfahren

Segmentierungsverfahren

Problemdefinition

- Gegeben — Anfrage mit n Worten
- Gesucht — "Beste" Segmentierung aus 2^{n-1} möglichen Segmentierungen

Beispiel

Anfrage	`new york times`
Segmentierungen	`new | york | times`
	`new | york times`
	`new york | times`
	`new york times`

Verwandte Arbeiten

PMI-basiert	[Risvik et al., WWW 2003] [Jones et al., WWW 2006] [Huang et al., WWW 2010]
Überwachtes Lernen	[Bergsma and Wang, EMNLP-CoNLL 2007] [Bendersky et al., SIGIR 2009]
Häufigkeitsbasiert	[Zhang et al., ACL-IJCNLP 2009] [Hagen et al., SIGIR 2010] [Mishra et al., WWW 2011] [Li et al., SIGIR 2011]
Wikipedia-basiert	[Tan and Peng, WWW 2008] [Hagen et al., WWW 2011]
Retrieval feedback	[Brenes et al., CERI 2010] [Bendersky et al., CIKM 2010] [Bendersky et al., ACL 2011]

Wikipedia-basiertes Verfahren [Hagen et al., WWW 2011]

Idee

Berechnung einer Punktzahl für jede mögliche Segmentierung

Beispiel

`new york times`

Wikipedia-basiertes Verfahren [Hagen et al., WWW 2011]

Idee

Berechnung einer Punktzahl für jede mögliche Segmentierung

Beispiel

Mögliche Segmente

Segment s
new york
york times
new york times

Wikipedia-basiertes Verfahren [Hagen et al., WWW 2011]

Idee

Berechnung einer Punktzahl für jede mögliche Segmentierung

Beispiel

Häufige Phrasen = bessere Segmente

Segment s	$freq(s)$
`new york`	165,36
`york times`	17,60
`new york times`	17,55

Wikipedia-basiertes Verfahren [Hagen et al., WWW 2011]

Idee

Berechnung einer Punktzahl für jede mögliche Segmentierung

Beispiel

Wikipedia-Titel = hochwertige Segmente

Segment s	$freq(s)$	Wiki
`new york`	165,36	✓
`york times`	17,60	-
`new york times`	17,55	✓

Wikipedia-basiertes Verfahren [Hagen et al., WWW 2011]

Idee

Berechnung einer Punktzahl für jede mögliche Segmentierung

Beispiel

Wikipedia-Titel = hochwertige Segmente

Segment s	$freq(s)$	Wiki	$freq'(s)$
`new york`	165,36	✓	165,36
`york times`	17,60	-	17,60
`new york times`	17,55	✓	165,36

Wikipedia-basiertes Verfahren [Hagen et al., WWW 2011]

Idee

Berechnung einer Punktzahl für jede mögliche Segmentierung

Beispiel

Normalisierung mit Segmentlänge

Segment s	$freq(s)$	Wiki	$freq'(s)$	$weight(s)$
`new york`	165,36	✓	165,36	330,72
`york times`	17,60	-	17,60	35,20
`new york times`	17,55	✓	165,36	496,08

Wikipedia-basiertes Verfahren [Hagen et al., WWW 2011]

Idee

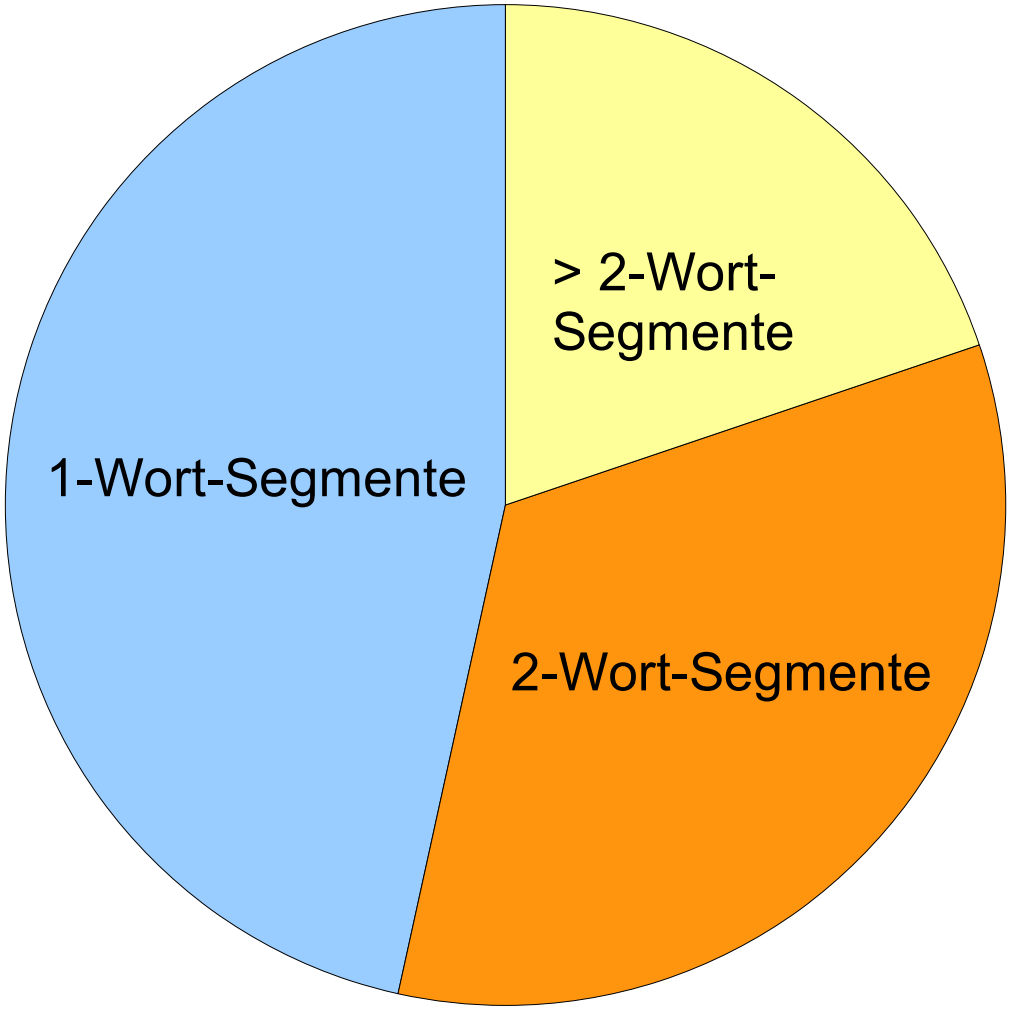
Berechnung einer Punktzahl für jede mögliche Segmentierung

Beispiel

$$\text{Punktzahl} = \text{Summe Gewichte}$$

Rang	Segmentierung S	$score(S)$
1	`new york times`	496,08
2	`new york` | `times`	330,72
3	`new` | `york times`	35,20
4	`new` | `york` | `times`	0,00

Segment-Analyse



550 000 Segmentierungen

WT-Baseline

Idee

Nur Wikipedia-Titel in einer Anfrage segmentieren

Beispiel

Anfrage `pictures of new york in the 1920s`

WT-Baseline

Idee

Nur Wikipedia-Titel in einer Anfrage segmentieren

Beispiel

Anfrage `pictures of new york in the 1920s`

WT-Baseline

Idee

Nur Wikipedia-Titel in einer Anfrage segmentieren

Beispiel

Anfrage	`pictures of new york in the 1920s`
Segmentierung	`pictures | of | new york | in | the | 1920s`

Performance-Vergleich

	Wiki-basiert	WT-Baseline
Nominalanfragen		
Nicht-Nominalanfragen		

Trainingsset ca. 5 000 Anfragen

Performance-Vergleich

	Wiki-basiert	WT-Baseline
Nominalanfragen	**0,580**	0,497
Nicht-Nominalanfragen		

Trainingsset ca. 5 000 Anfragen

Performance-Vergleich

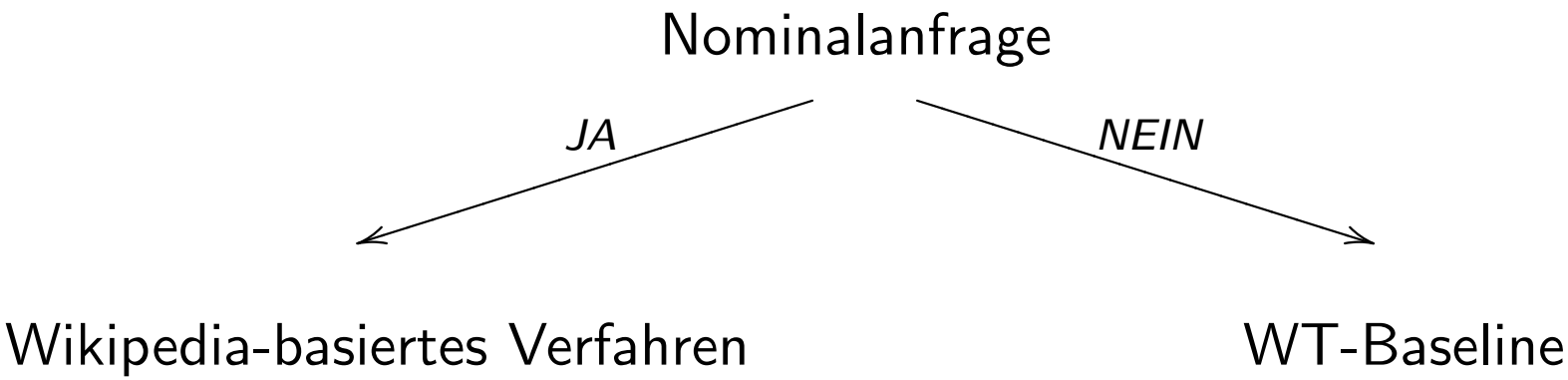
	Wiki-basiert	WT-Baseline
Nominalanfragen	**0,580**	0,497
Nicht-Nominalanfragen	0,268	**0,492**

Trainingsset ca. 5 000 Anfragen

Hybrides Verfahren

	Wiki-basiert	WT-Baseline
Nominalanfragen	**0,580**	0,497
Nicht-Nominalanfragen	0,268	**0,492**

Trainingsset ca. 5 000 Anfragen



Hybrides Verfahren

Wiki-basiert	WT-Baseline	Hybrid
0,412	0,502	**0,535**

Testset ca. 50 000 Anfragen

Evaluierung

Evaluierung

Grundlage

Manuelle Segmentierungen in Evaluierungskorpora

Beispiel

Stimmen	Segmentierung
2	`tokio hotel | bill`
1	`tokio | hotel bill`

Evaluierungsmaße

Bisher

Referenz = "Ähnlichste" Segmentierung im Korpus

Beispiel

Vorhersage `tokio | hotel bill`

Stimmen	Segmentierung
2	`tokio hotel | bill`
1	`tokio | hotel bill`

Evaluierungsmaße

Bisher

Referenz = "Ähnlichste" Segmentierung im Korpus

Beispiel

Vorhersage `tokio | hotel bill`

Referenz $\Longrightarrow$

Stimmen	Segmentierung
2	`tokio hotel	bill`
1	`tokio	hotel bill`

Evaluierungsmaße

Bisher

Referenz = "Ähnlichste" Segmentierung im Korpus

Beispiel

Vorhersage `tokio | hotel bill`

	Stimmen	Segmentierung
	9	`tokio hotel | bill`
Referenz $\Longrightarrow$	1	`tokio | hotel bill`

Evaluierungsmaße

Bisher

Referenz = "Ähnlichste" Segmentierung im Korpus

Beispiel

Vorhersage `tokio | hotel bill`

Referenz $\Longrightarrow$

Stimmen	Segmentierung	
9	`tokio hotel	bill`	$\Longrightarrow$ Faire Referenz
1	`tokio	hotel bill`	

Evaluierungsmaße

Neu

Erzeugung der Referenz unter Berücksichtigung aller Segmentierungen

Beispiel

Stimmen	Segmentierung
4	`tokio hotel | bill`
3	`tokio | hotel | bill`
2	`tokio hotel bill`
1	`tokio | hotel bill`

Evaluierungsmaße

Neu

Erzeugung der Referenz unter Berücksichtigung aller Segmentierungen

Beispiel

Stimmen	Segmentierung
4	tokio hotel | bill
3	tokio | hotel | bill
2	tokio hotel bill
1	tokio | hotel bill

NB	B	NB	B
tokio	hotel		bill
6	4	3	7

Evaluierungsmaße

Neu

Erzeugung der Referenz unter Berücksichtigung aller Segmentierungen

Beispiel

Stimmen	Segmentierung
4	tokio hotel | bill
3	tokio | hotel | bill
2	tokio hotel bill
1	tokio | hotel bill

NB	B	NB	B
tokio	hotel		bill
6	4	3	7

⇓

tokio hotel | bill

Genauigkeit

Beispiel

Referenz	`tokio  hotel` | `bill`	(2 Segmente)
Vorhersage	`tokio` | `hotel` | `bill`	(3 Segmente)

Genauigkeit

Beispiel

Referenz	`tokio hotel | bill`	(2 Segmente)
Vorhersage	`tokio | hotel | bill`	(3 Segmente)

query 0,00 Vorhersage $\neq$ Referenz

Genauigkeit

Beispiel

Referenz	`tokio  hotel` | `bill`	(2 Segmente)
Vorhersage	`tokio` | `hotel` | `bill`	(3 Segmente)

query	0,00	Vorhersage $\neq$ Referenz
segPrec	0,33	1 von 3 der vorhergesagten Segmente korrekt

Genauigkeit

Beispiel

Referenz	`tokio hotel` | `bill`	(2 Segmente)
Vorhersage	`tokio` | `hotel` | `bill`	(3 Segmente)

query	0,00	Vorhersage $\neq$ Referenz
segPrec	0,33	1 von 3 der vorhergesagten Segmente korrekt
segRec	0,50	1 von 2 der referenzierten Segmente korrekt

Genauigkeit

Beispiel

Referenz	`tokio hotel` | `bill`	(2 Segmente)
Vorhersage	`tokio` | `hotel` | `bill`	(3 Segmente)

query	0,00	Vorhersage $\neq$ Referenz
segPrec	0,33	1 von 3 der vorhergesagten Segmente korrekt
segRec	0,50	1 von 2 der referenzierten Segmente korrekt
break	0,50	1 von 2 Break-Entscheidungen korrekt

Genauigkeit

	Wiki-basiert	WT-Baseline	Hybrid
query	0,412	0,502	**0,535**
segPrec	0,606	0,673	**0,693**
segRec	0,565	**0,708**	0,703
break	0,717	0,778	**0,785**

Testset ca. 50 000 Anfragen

Laufzeit

	Wiki-basiert	WT-Baseline	Hybrid
Anfragen/s	3 358	3 797	2 918

Testset ca. 50 000 Anfragen

Laufzeit

	Wiki-basiert	WT-Baseline	Hybrid
Anfragen/s	3 358	3 797	2 918

Testset ca. 50 000 Anfragen

$$1\,\text{Milliarde Anfragen/Tag} = 12\,000\,\text{Anfragen/s}$$

Zusammenfassung

Zusammenfassung

Ergebnisse

- WT-Baseline
- Hybrides Verfahren
- Fairere Evaluierungsmaße

Zusammenfassung

Ergebnisse

- WT-Baseline
- Hybrides Verfahren
- Fairere Evaluierungsmaße

Ausblick

- Optimierung
- Retrieval Performance

Zusammenfassung

Ergebnisse

- WT-Baseline
- Hybrides Verfahren
- Fairere Evaluierungsmaße

Ausblick

- Optimierung
- Retrieval Performance

Vielen Dank