

12. Vorlesung

- Statistische Sprachmodelle für Information Retrieval
 - Allgemeiner Ansatz
 - Unigram Modell
 - Beziehung zum Vektorraummodell mit TF-IDF Gewichten
- Statistische Sprachmodelle zur Glättung
 - Idee von Glättung
 - Methoden
 - Jelinek-Mercer
 - Dirichlet
 - Absolute Discounting
 - Interpolation und Backoff
 - Vergleich der Methoden
- Die zwei Rollen von Glättung
 - Experimente
 - Einfluß der Anfragelänge und Ausführlichkeit
 - Zweistufiges Modell

Statistische Sprachmodelle für Information Retrieval

- Es gibt viele Modelle für Information Retrieval
 - Theoretische Modelle
 - Boolesches Modell
 - Fuzzy Set Modell
 - Probabilistische Modelle
 - Empirische Modelle
 - Vektorraummodell
 - TF-IDF Gewichte
- Statistische Sprachmodelle
 - Spracherkennung, Verarbeitung von natürlicher Sprache
 - Gute theoretische Grundlage
 - Interessante Beziehungen zu Empirischen Information Retrieval Modellen

Allgemeiner Ansatz

- Information Retrieval Problem
 - gegeben eine Anfrage $q=q_1q_2\dots q_n$ finde die relevanten Dokumente
- Ansatz basierend auf statistischen Sprachmodellen
 - bestimme Wahrscheinlichkeit, dass die Anfrage q durch ein probabilistisches Modell “generiert” wird, welches ein Dokument $d=d_1d_2\dots d_m$ beschreibt.
 - Ordne die Dokumente aus der Sammlung nach der Posterior Wahrscheinlichkeit $p(d|q)$ mit $p(d|q) \propto p(q|d)p(d)$
 - $p(d)$ ist die Prior Wahrscheinlichkeit, dass Dokument d ueberhaupt zu irgendeiner Anfrage relevant ist
 - $p(d)$ wird im weiteren als gleichverteilt angenommen und beeinflusst damit das Ranking nicht, d.h. nur die Likelihood $p(q|d)$ ist wichtig
 - spezielle Ansätze modellieren mit $p(d)$ nicht-textuale Eigenschaften von d , z.B. Dokumentlänge, Links in einer Webseite, Format und Stileigenschaften

Unigram Modell

- Wörter der Anfrage werden als unabhängig identisch verteilt angenommen, deshalb

$$p(q|d) = \prod_{i=1}^n p(q_i|d)$$

- Die Wahrscheinlichkeiten $p(q_i|d)$ sind multinomial verteilt, d.h. für jedes d gibt es einen Parametervektor

$$\vec{\nu}_d = (\nu_{d1}, \dots, \nu_{dn}) \text{ mit } \sum_i \nu_{di} = 1$$

der mit ML-Schätzer bestimmt ist $\nu_{dk} = \frac{c(w_k, d)}{\sum_{w' \in V} c(w', d)}$

- Die Likelihood für eine Anfrage q bezüglich eines Dokuments d ist dann $p(q|d) = p(q|\vec{\nu}_d)$

Beziehung zum Vektorraummodell mit TF-IDF Gewichten

- Oberflächlich scheint es einen fundamentalen Unterschied zwischen TF-IDF und stat. Sprachmodellen zu geben: stat. Sprachmodell haben kein IDF
- Unterscheidung der Verteilung für „gesehene“ Wörter und „nicht-gesehene“ Wörter

Glättung

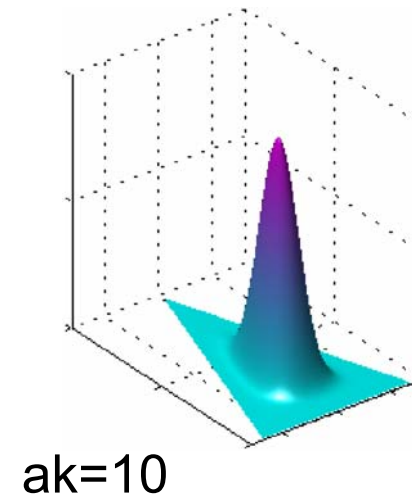
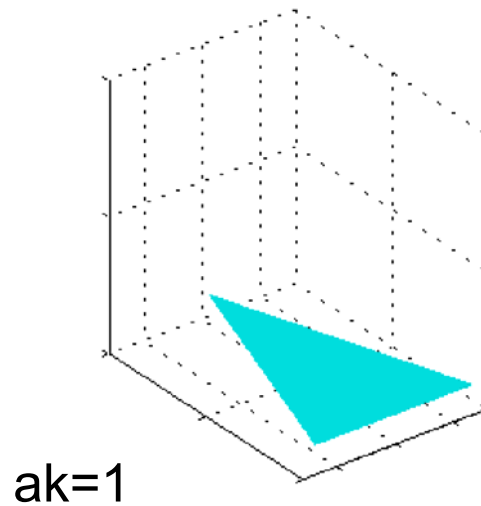
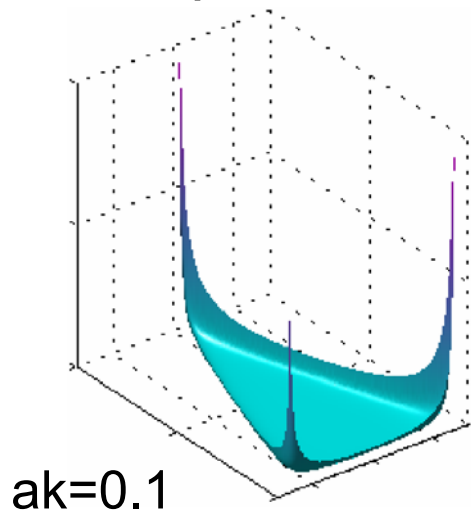
- Ungeglättetes Modell ist der Maximum-Likelihood Schätzer, d.h. normalisierte Anzahlen
- Problem
 - ML-Schätzer unterschätzt die Wahrscheinlichkeit von nicht-gesehenen Wörtern in einem Dokument
- Idee der Glättung
 - weise nicht-gesehenen Wörtern eine Wahrscheinlichkeit größer als Null zu
 - dies zieht notwendigerweise eine Abwertung der gesehenen Wörter nach sich
 - weise nicht-gesehenen Wörtern extra Wahrscheinlichkeit bezüglich eines Alternativmodelles zu

Jelinek-Mercer Methode

- Idee
 - interpoliere linear zwischen der Maximum Likelihood Schätzung für ein Dokument und dem Modell für die ganze Sammlung
 - Parameter Lambda kontrolliert den Einfluß der beiden Modelle

Bayesisches Glätten mittels Dirichlet Prior

- Unigram-Modell ist eine Multinomiale Verteilung mit einem Parameter-Vektor, der zu Eins summiert
- Dirichlet-Verteilung weist einem Vektor, der zu Eins summiert, eine Wahrscheinlichkeit zu
- Beispiele für eine Dirichlet Verteilung



Absolute Discounting

- Ähnlich zu Jelinek-Mercer, aber von $p_s(w|d)$ eine Konstante subtrahiert, statt mit $(1-\lambda)$ zu multiplizieren
-

Zusammenfassung

Method	$p_s(w d)$	α_d	Parameter
Jelinek-Mercer	$(1 - \lambda) p_{ml}(w d) + \lambda p(w \mathcal{C})$	λ	λ
Dirichlet	$\frac{c(w; d) + \mu p(w \mathcal{C})}{ d + \mu}$	$\frac{\mu}{ d + \mu}$	μ
Absolute discount	$\frac{\max(c(w; d) - \delta, 0)}{ d } + \frac{\delta d _u}{ d } p(w \mathcal{C})$	$\frac{\delta d _u}{ d }$	δ

- Alle Methoden lassen sich effizient implementieren
 - alle α_d lassen sich vorberechnen zur Indexzeit
 - das Gewicht eines Terms w aus der Anfrage q , der in einem Dokument d vorkommt, kann aus $p(w|\mathcal{C})$, $c(w,q)$ und $c(w,d)$ berechnet werden
 - Die Komplexität entspricht der des TF-IDF Modells

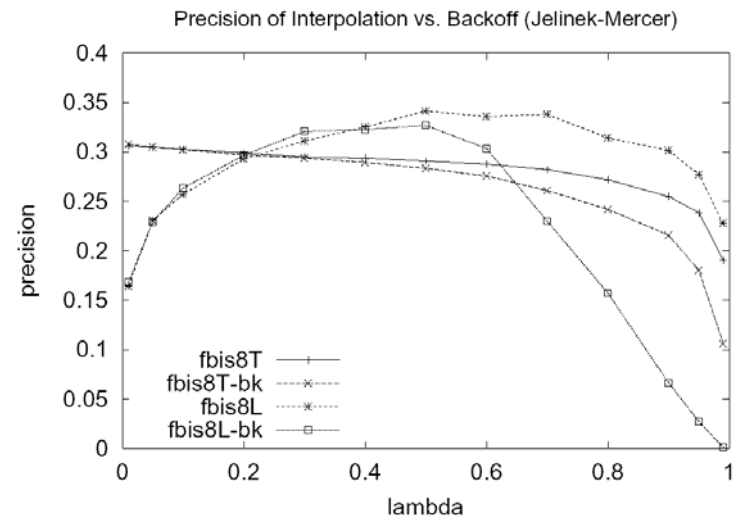
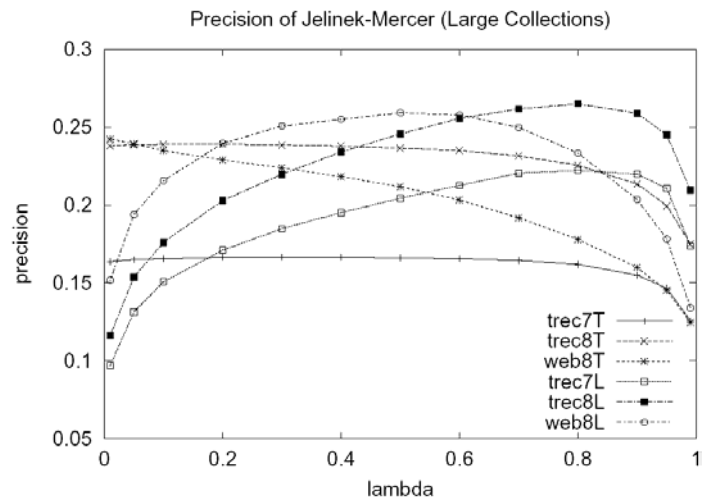
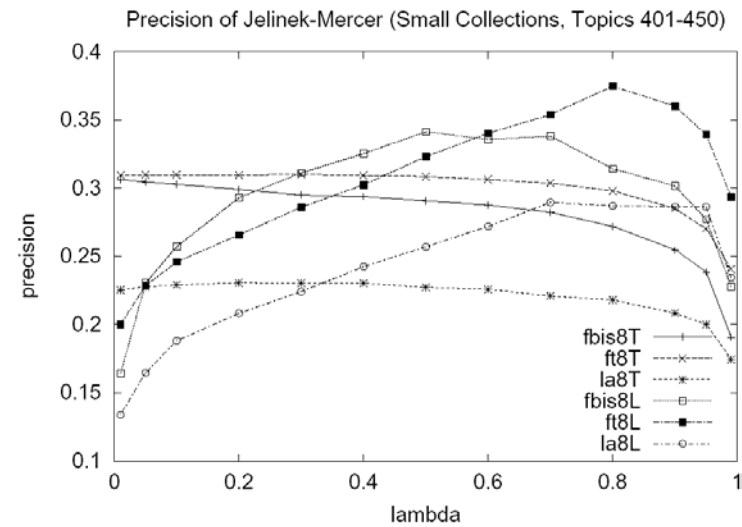
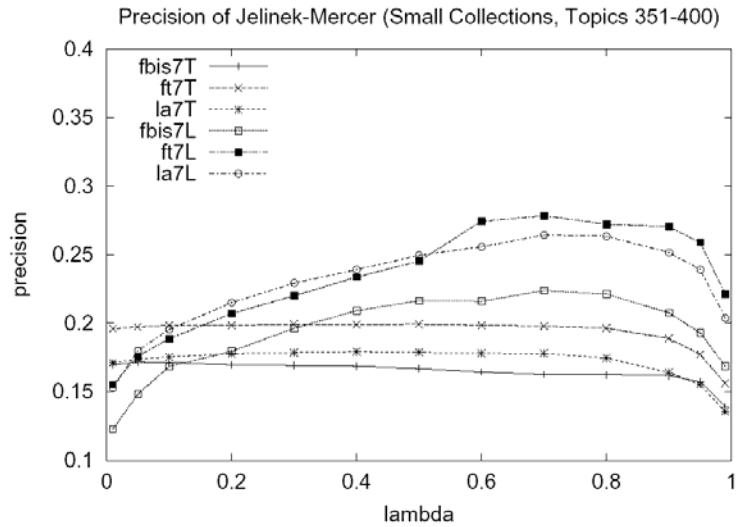
Interpolation und Backoff

- Bisher vorgestellte Methoden sind Interpolationsmethoden
- Arbeitsweise
 - verringere Anzahl der gesehenen Wörter
 - angenommene zusätzliche Anzahlen werden auf gesehene und nicht-gesehene Wörter verteilt
- Möglicher Nachteil
 - allgemein häufiges Wort bekommt übermäßige viel zusätzlich angenommene Zähler, d.h. letztendlich zählt es mehr als es wirklich im Dokument vorkommt
- Backoff
 - vertraue ML-Schätzung bei hohen Anzahlen und verteile nur die Wahrscheinlichkeit von weniger häufigen Worten neu
 - zusätzliche Anzahl nur für nicht-gesehene Worte

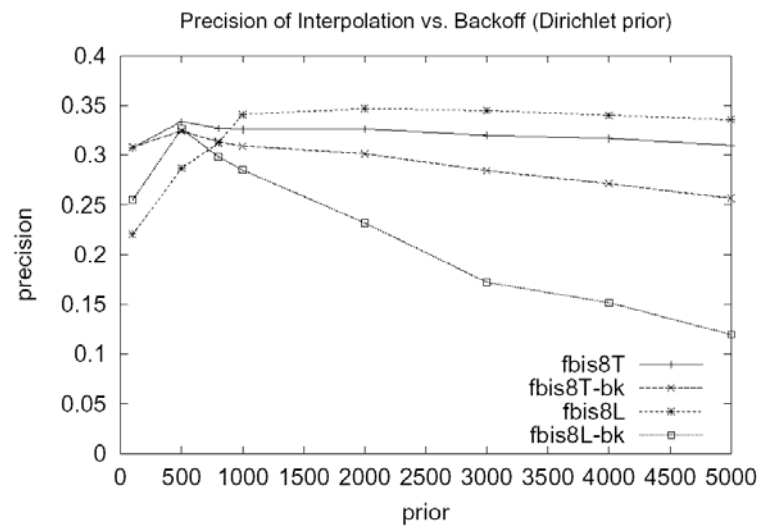
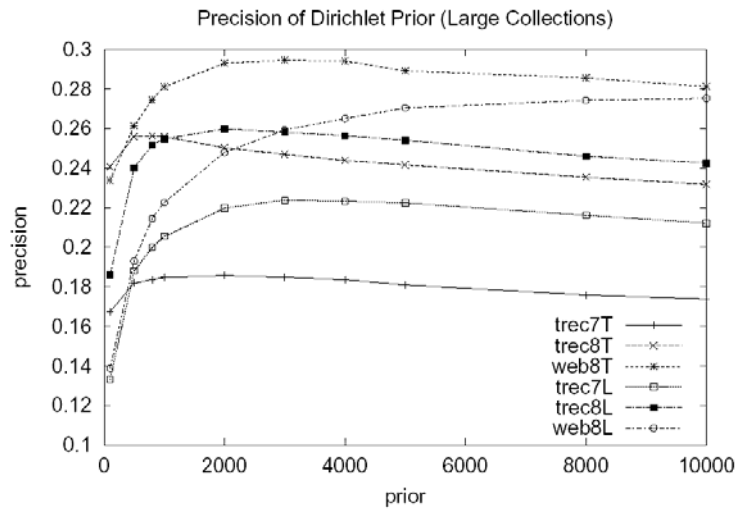
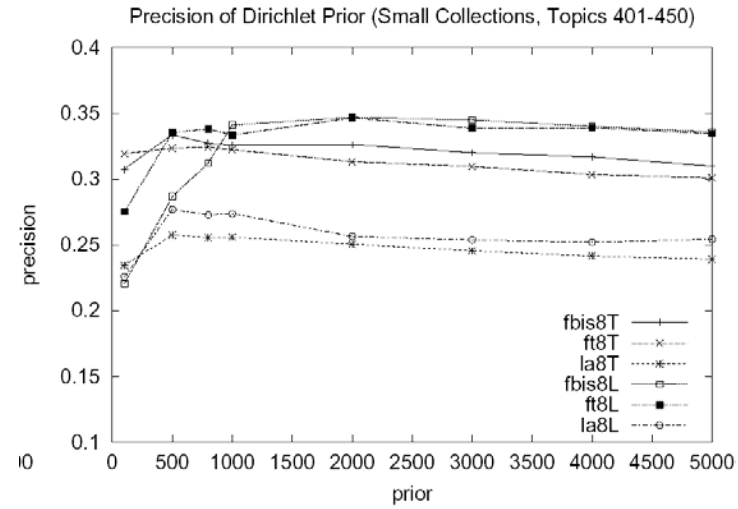
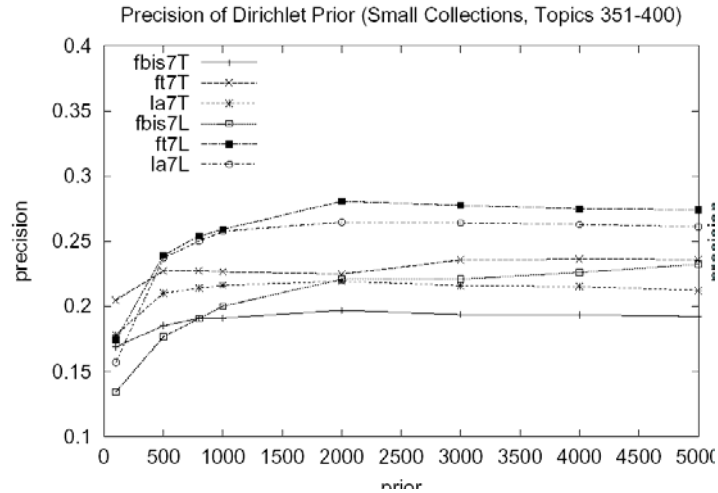
Vergleich der Methoden

- Verschiedene Testdaten für ad hoc Retrieval
- Vorverarbeitung
 - Porter Stemming
 - keine Stopwörter entfernen
- Anfragen
 - nur Titel
 - ausführliche Version: Titel + Beschreibung + Erzählung

Jelinek-Mercer

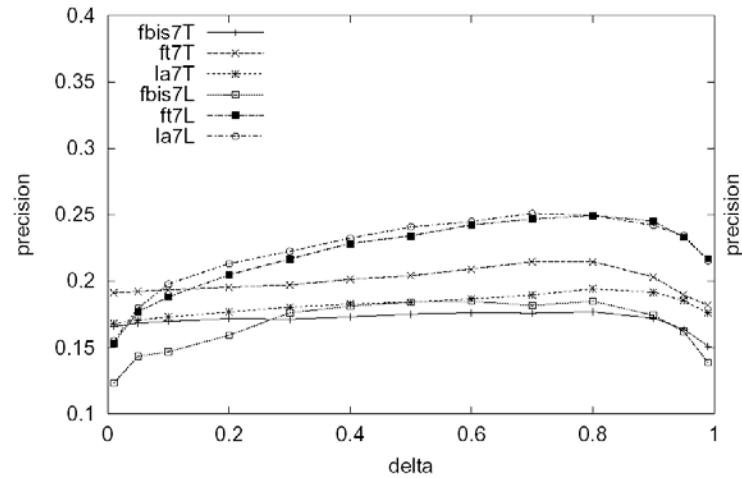


Dirichlet Priors

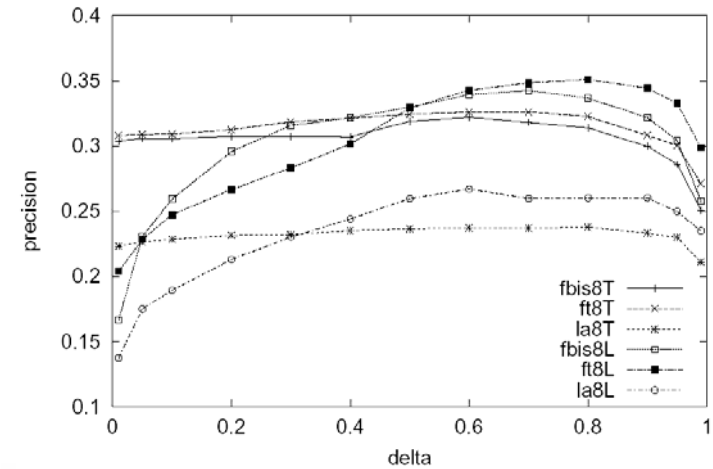


Absolute Discounting

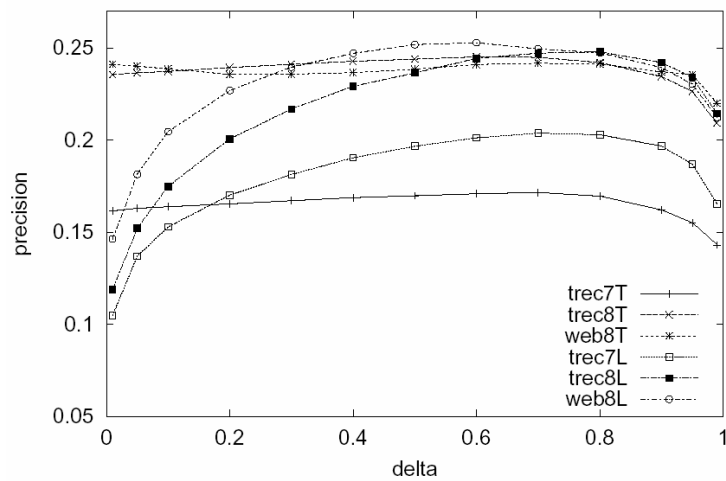
Precision of Absolute Discounting (Small Collections, Topics 351-400)



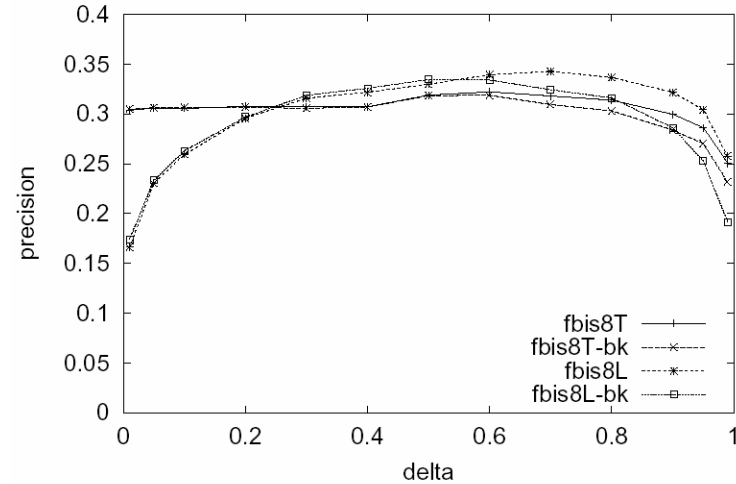
) Precision of Absolute Discounting (Small Collections, Topics 401-450)



Precision of Absolute Discounting (Large Collections)



Precision of Interpolation vs. Backoff (Absolute Discounting)

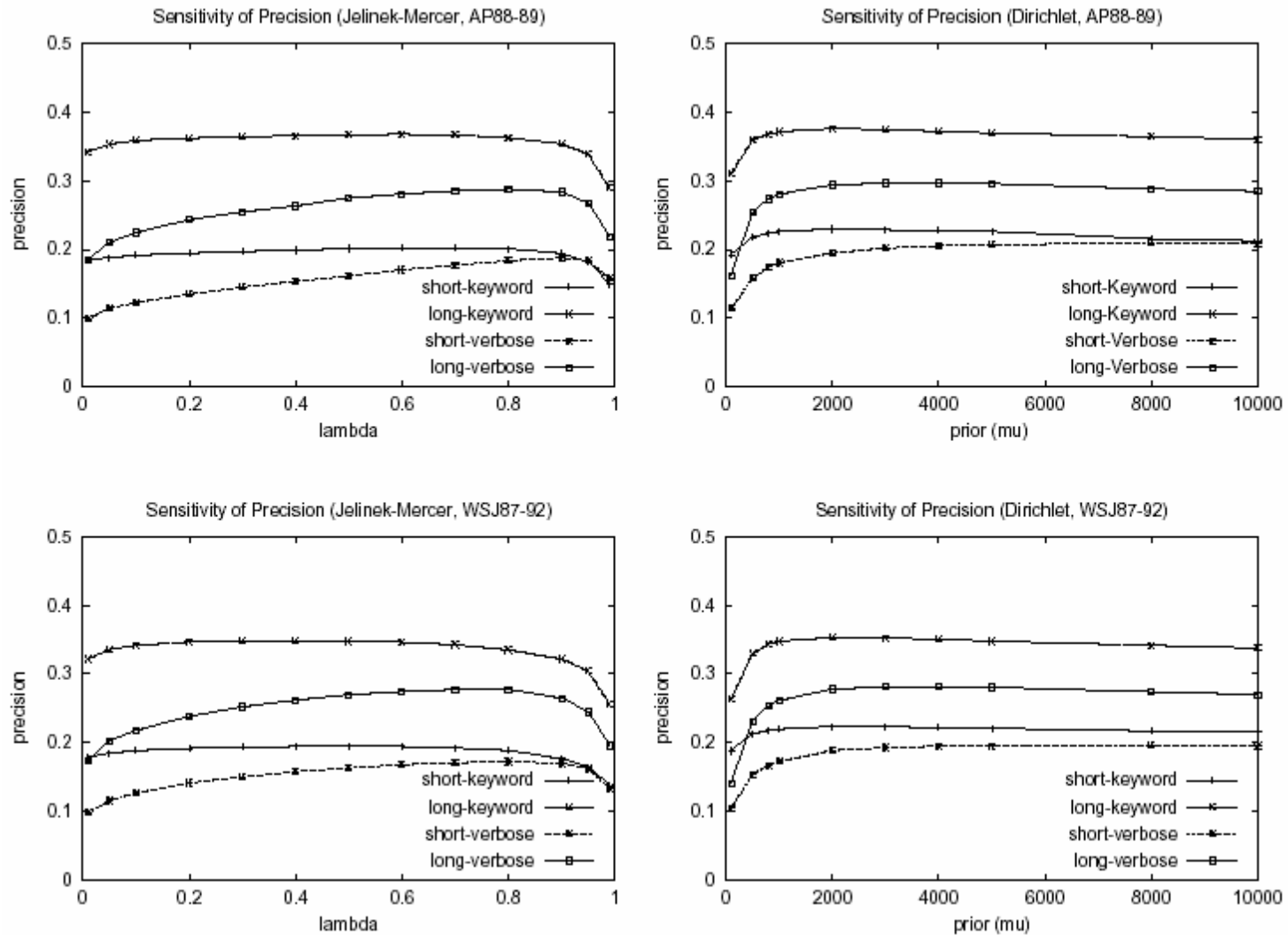


Vergleich bezüglich der Anfragen

Collection	Method	Parameter	Avg. Prec.	Prec@10	Prec@20
fbis7T	JM	$\lambda = 0.05$	0.172	0.284	0.220
	Dir	$\mu = 2,000$	0.197	0.282	0.238
	Dis	$\delta = 0.8$	0.177	0.284	0.233
ft7T	JM	$\lambda = 0.5$	0.199	0.263	0.195
	Dir	$\mu = 4,000$	0.236	0.283	0.213
	Dis	$\delta = 0.8$	0.215	0.271	0.196
la7T	JM	$\lambda = 0.4$	0.179	0.238	0.205
	Dir	$\mu = 2,000$	0.220*	0.294*	0.233
	Dis	$\delta = 0.8$	0.194	0.268	0.216
fbis8T	JM	$\lambda = 0.01$	0.306	0.344	0.282
	Dir	$\mu = 500$	0.334	0.367	0.292
	Dis	$\delta = 0.5$	0.319	0.363	0.288
ft8T	JM	$\lambda = 0.3$	0.310	0.359	0.283
	Dir	$\mu = 800$	0.324	0.367	0.297
	Dis	$\delta = 0.7$	0.326	0.367	0.296
la8T	JM	$\lambda = 0.2$	0.231	0.264	0.211
	Dir	$\mu = 500$	0.258	0.271	0.216
	Dis	$\delta = 0.8$	0.238	0.282	0.224
trec7T	JM	$\lambda = 0.3$	0.167	0.366	0.315
	Dir	$\mu = 2,000$	0.186*	0.412	0.342
	Dis	$\delta = 0.7$	0.172	0.382	0.333
trec8T	JM	$\lambda = 0.2$	0.239	0.438	0.378
	Dir	$\mu = 800$	0.256	0.448	0.398
	Dis	$\delta = 0.6$	0.245	0.466	0.406
web8T	JM	$\lambda = 0.01$	0.243	0.348	0.293
	Dir	$\mu = 3,000$	0.294*	0.448*	0.374*
	Dis	$\delta = 0.7$	0.242	0.370	0.323
Average	JM	—	0.227	0.323	0.265
	Dir	—	0.256*	0.352*	0.289*
	Dis	—	0.236	0.339	0.279

Collection	Method	Parameter	Avg. Prec.	Prec@10	Prec@20
fbis7L	JM	$\lambda = 0.7$	0.224	0.339	0.279
	Dir	$\mu = 5,000$	0.232	0.313	0.249
	Dis	$\delta = 0.6$	0.185	0.321	0.259
ft7L	JM	$\lambda = 0.7$	0.279	0.331	0.244
	Dir	$\mu = 2,000$	0.281	0.329	0.248
	Dis	$\delta = 0.8$	0.249	0.317	0.236
la7L	JM	$\lambda = 0.7$	0.264	0.350	0.286
	Dir	$\mu = 2,000$	0.265	0.354	0.285
	Dis	$\delta = 0.7$	0.251	0.340	0.279
fbis8L	JM	$\lambda = 0.5$	0.341	0.349	0.283
	Dir	$\mu = 2,000$	0.347	0.349	0.290
	Dis	$\delta = 0.7$	0.343	0.356	0.274
ft8L	JM	$\lambda = 0.8$	0.375	0.427	0.320
	Dir	$\mu = 2,000$	0.347	0.380	0.297
	Dis	$\delta = 0.8$	0.351	0.398	0.309
la8L	JM	$\lambda = 0.7$	0.290*	0.296	0.238
	Dir	$\mu = 500$	0.277	0.282	0.231
	Dis	$\delta = 0.6$	0.267	0.287	0.222
trec7L	JM	$\lambda = 0.8$	0.222	0.476	0.401
	Dir	$\mu = 3,000$	0.224	0.456	0.383
	Dis	$\delta = 0.7$	0.204	0.460	0.396
trec8L	JM	$\lambda = 0.8$	0.265	0.504	0.434
	Dir	$\mu = 2,000$	0.260	0.484	0.4
	Dis	$\delta = 0.8$	0.248	0.518	0.428
web8L	JM	$\lambda = 0.5$	0.259	0.422	0.348
	Dir	$\mu = 10,000$	0.275	0.410	0.343
	Dis	$\delta = 0.6$	0.253	0.414	0.333
Average	JM	—	0.280	0.388	0.315*
	Dir	—	0.279	0.373	0.303
	Dis	—	0.261	0.379	0.304

Einfluß der Anfrage Länge und Ausführlichkeit



Überlegungen

- Beobachtung
 - Retrieval Performanz ist für Schlüsselwörter weniger sensitiv als für ausführliche Anfragen
 - Suoptimales Glätten schadet ausführlichen Anfragen wesentlich mehr als Schlüsselwortanfragen
- Schlußfolgerung
 - Glätten ist verantwortlich, für das „Erklären“ von allgemeinen Worten in einer Anfrage
- Die beiden Rollen des Glättens
 - Beheben des Problems kleiner Stichproben, d.h. das „Erklären“ von nicht-gesehenen Worten, abhängig vom Dokument
 - das „Erklären“ von allgemeinen Worten in einer Anfrage, abhängig von der Anfrage
- Modelle
 - Dirichlet Priors sind gut für das „Erklären“ von nicht-gesehenen Worten, weil Dokument-spezifische ML-Korrektur
 - Jelinek-Mercer sind gut für das „Erklären“ von allgemeinen Worten, das Dokument-unabhängige ML-Korrektur

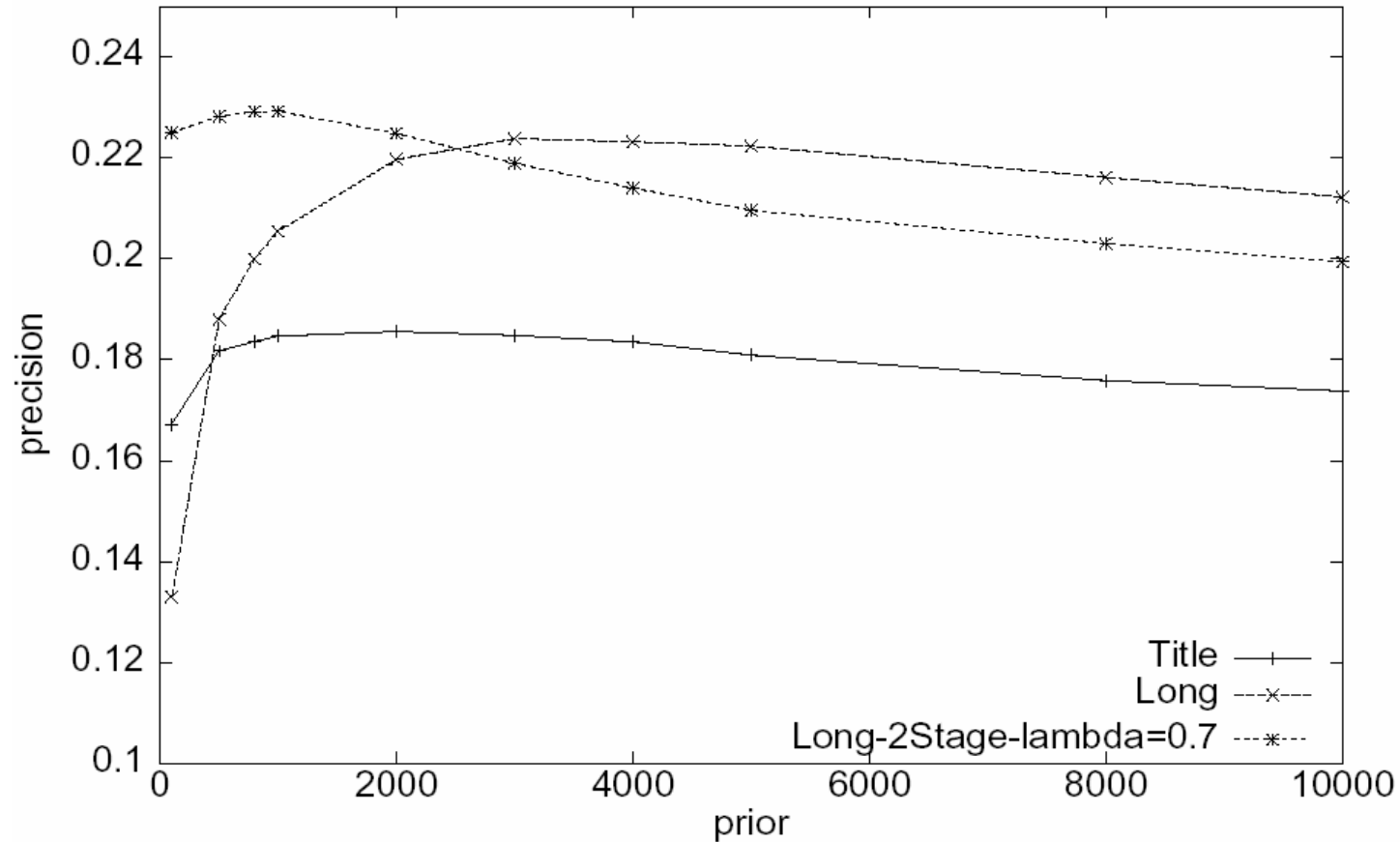
Zweistufiges Modell

- Unterscheide Glättung der Anfrage und Glättung der Dokumente
- Modell
 - glätte zuerst ein Dokument-Modell mittels Dirichlet-Priors
 - glätte in einem zweiten Schritte diese Schätzung mittels Jelinek-Mercer bezüglich eines Modells für Anfragen

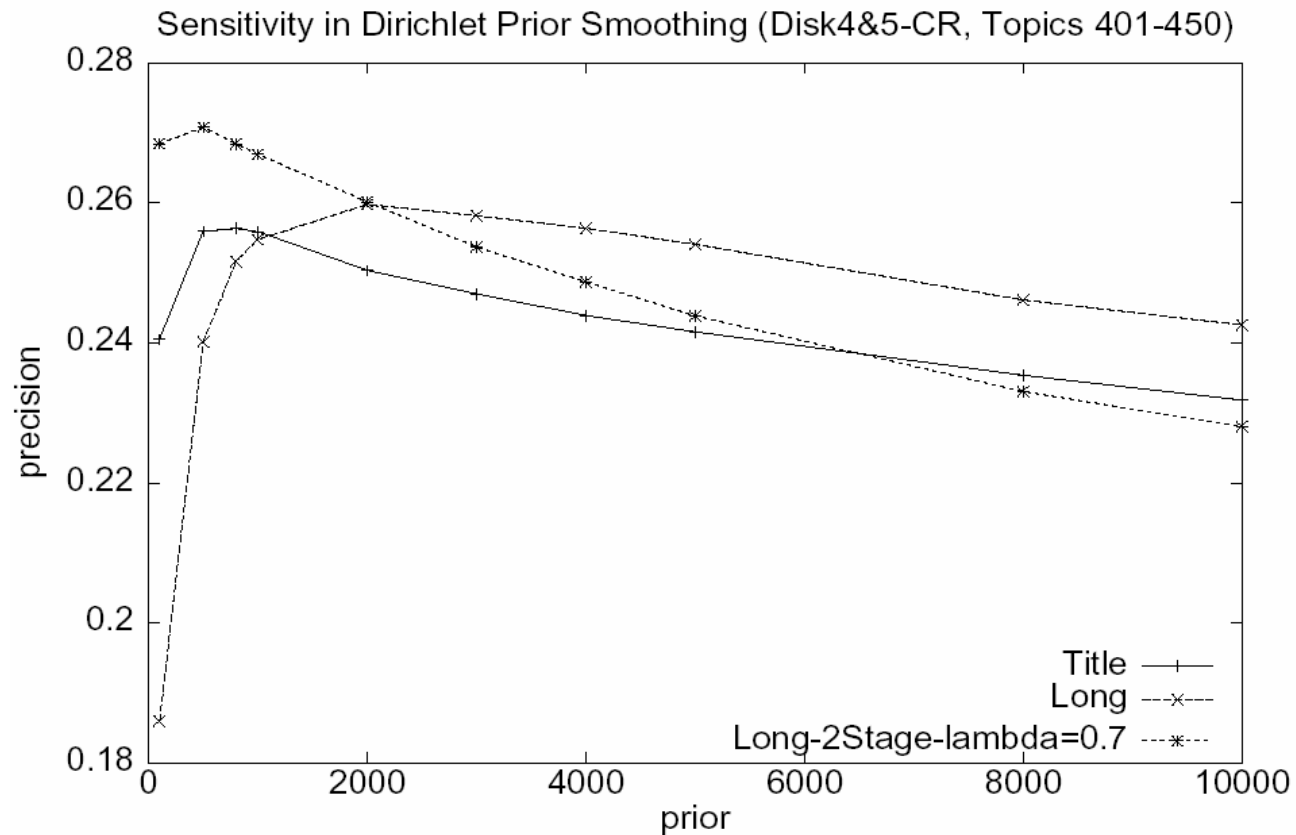
$$p_{\lambda, \mu}(w | d) = (1 - \lambda) \frac{c(w; d) + \mu p(w | \mathcal{C})}{|d| + \mu} + \lambda p(w | \mathcal{U})$$

Experimente

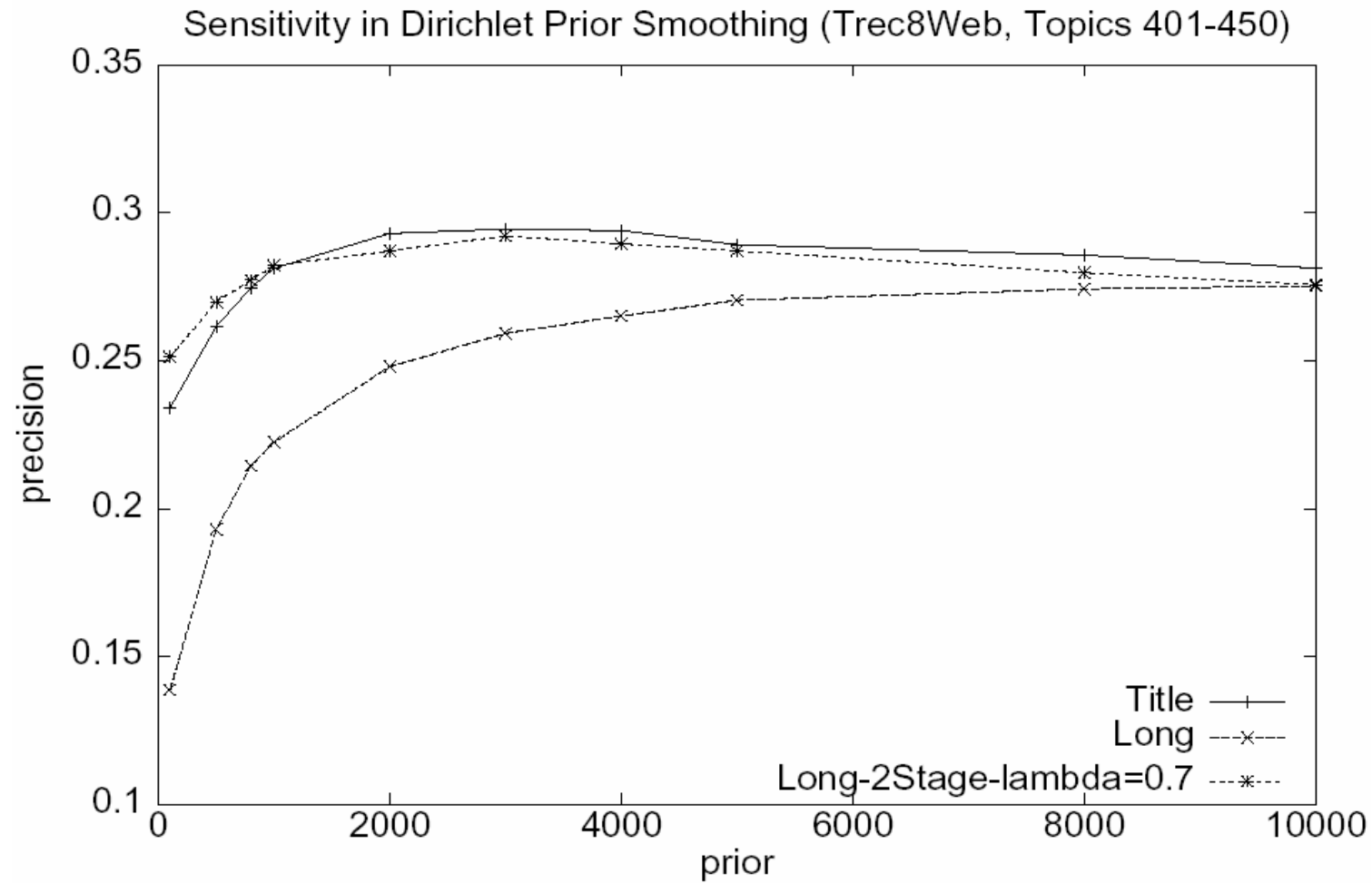
Sensitivity in Dirichlet Prior Smoothing (Disk4&5-CR, topics 351-400)



Experimente



Experimente



Bestimmen der Parameter

- Parameter der zweistufigen Modells
 - bestimme μ der Dirichlet mittels Leaf-One-Out Cross Validerung
 - bestimme λ der Jelinek-Mercer Methode mittels EM

Experimente

Collection	Query	Method	Avg. Prec.	(Median)	InitPr	Pr@10	Pr@20
AP88-89	SK	Best JM	0.203	(0.194)	0.573	0.310	0.283
		Best Dir	0.230	(0.224)	0.623	0.356	0.332
		Two-Stage	0.222*		0.611	0.358	0.317
	LK	Best JM	0.368	(0.362)	0.767	0.509	0.469
		Best Dir	0.376	(0.368)	0.755	0.506	0.475
		Two-Stage	0.374		0.754	0.505	0.480
	SV	Best JM	0.188	(0.158)	0.569	0.309	0.272
		Best Dir	0.209	(0.195)	0.609	0.338	0.304
		Two-Stage	0.204		0.598	0.339	0.305
	LV	Best JM	0.288	(0.263)	0.711	0.430	0.391
		Best Dir	0.298	(0.285)	0.704	0.453	0.403
		Two-Stage	0.292		0.689	0.444	0.400
WSJ87-92	SK	Best JM	0.194	(0.188)	0.629	0.364	0.330
		Best Dir	0.223	(0.218)	0.660	0.412	0.376
		Two-Stage	0.218*		0.662	0.409	0.366
	LK	Best JM	0.348	(0.341)	0.814	0.575	0.524
		Best Dir	0.353	(0.343)	0.834	0.562	0.507
		Two-Stage	0.358		0.850*	0.572	0.523
	SV	Best JM	0.172	(0.158)	0.615	0.346	0.314
		Best Dir	0.196	(0.188)	0.638	0.389	0.333
		Two-Stage	0.199		0.660	0.391	0.344
	LV	Best JM	0.277	(0.252)	0.768	0.481	0.452
		Best Dir	0.282	(0.270)	0.750	0.480	0.442
		Two-Stage	0.288*		0.762	0.497	0.449
ZF1-2	SK	Best JM	0.179	(0.170)	0.455	0.220	0.193
		Best Dir	0.215	(0.210)	0.514	0.265	0.226
		Two-Stage	0.200		0.490	0.256	0.227
	LK	Best JM	0.306	(0.290)	0.675	0.345	0.300
		Best Dir	0.326	(0.316)	0.681	0.376	0.322
		Two-Stage	0.322		0.696	0.368	0.322
	SV	Best JM	0.156	(0.139)	0.450	0.208	0.174
		Best Dir	0.185	(0.170)	0.456	0.225	0.185
		Two-Stage	0.181		0.487	0.246*	0.203
	LV	Best JM	0.267	(0.242)	0.593	0.300	0.258
		Best Dir	0.279	(0.273)	0.606	0.329	0.272
		Two-Stage	0.279*		0.618	0.334	0.278

Experimente

Collection	Query	Method	Avg. Prec.	(Median)	InitPr	Pr@10	Pr@20
Disk4&5 (-CR)	Trec7-SK	Best JM	0.167	(0.165)	0.632	0.366	0.315
		Best Dir	0.186	(0.182)	0.688	0.412	0.342
		Two-Stage	0.182		0.673	0.420	0.357
	Trec7-SV	Best JM	0.173	(0.138)	0.646	0.392	0.342
		Best Dir	0.182	(0.168)	0.656	0.416	0.340
		Two-Stage	0.181		0.655	0.416	0.348
	Trec7-LV	Best JM	0.222	(0.195)	0.723	0.476	0.401
		Best Dir	0.224	(0.212)	0.763	0.456	0.383
		Two-Stage	0.230		0.760	0.466	0.412
	Trec8-SK	Best JM	0.239	(0.237)	0.621	0.438	0.378
		Best Dir	0.256	(0.244)	0.717	0.448	0.398
		Two-Stage	0.257		0.719	0.448	0.405
	Trec8-SV	Best JM	0.231	(0.192)	0.687	0.416	0.357
		Best Dir	0.228	(0.222)	0.670	0.400	0.337
		Two-Stage	0.231		0.719	0.440	0.354
	Trec8-LV	Best JM	0.265	(0.234)	0.789	0.504	0.434
		Best Dir	0.260	(0.252)	0.753	0.484	0.400
		Two-Stage	0.268		0.787	0.478	0.400
Web	Trec8-SK	Best JM	0.243	(0.212)	0.607	0.348	0.293
		Best Dir	0.294	(0.281)	0.756	0.448	0.374
		Two-Stage	0.278*		0.730	0.426	0.358
	Trec8-SV	Best JM	0.203	(0.191)	0.611	0.340	0.284
		Best Dir	0.267	(0.249)	0.699	0.408	0.325
		Two-Stage	0.253		0.680	0.398	0.330
	Trec8-LV	Best JM	0.259	(0.243)	0.790	0.422	0.348
		Best Dir	0.275	(0.248)	0.752	0.410	0.343
		Two-Stage	0.284		0.781	0.442	0.362