WebCLEF 2006 topic development

Krisztian Balog & Leif Azzopardi

Recall ...

- Topics at WebCLEF
  - 1940 in total
  - 320 manually created (see WebCLEF 2005 overview)
  - 1620 topics were automatically generated
    - 810 using a unigram model
    - 810 using a bigram model
- This talk: focus on automatically created topics

Background

- Azzopardi & De Rijke, Automatic Construction of Known-Item Finding Test Beds, SIGIR 2006
- Produce numerous known-item queries (query, item) at minimal cost

Algorithm

- Initialize empty query set $q = \{\}$
- Select doc d to be the known-item with probability $p(d)$
- Select query length k with prob $p(k)$
- Repeat k times:
  - Select a term t from doc model of d with probability $p(t|\theta_d)$
  - Add t to query q
- Record (d,q) to be known-item/query

Algorithm (2)

- Need to define *p*(*d*), *p*(*k*), and $p(t|\theta_d)$
  - Simulate the thought and behavior of using by using different distributions to characterize various types/styles of queries
  - E.g., $p(t|\theta_d)$ as a mixture of maximum likelihood estimate of term occurring in doc and a background model:

$$p(t|\theta_d) = (1 - \lambda)p(t|d) + \lambda p(t)$$

  - “As λ tends to 0, the user’s recollection of the doc improves”

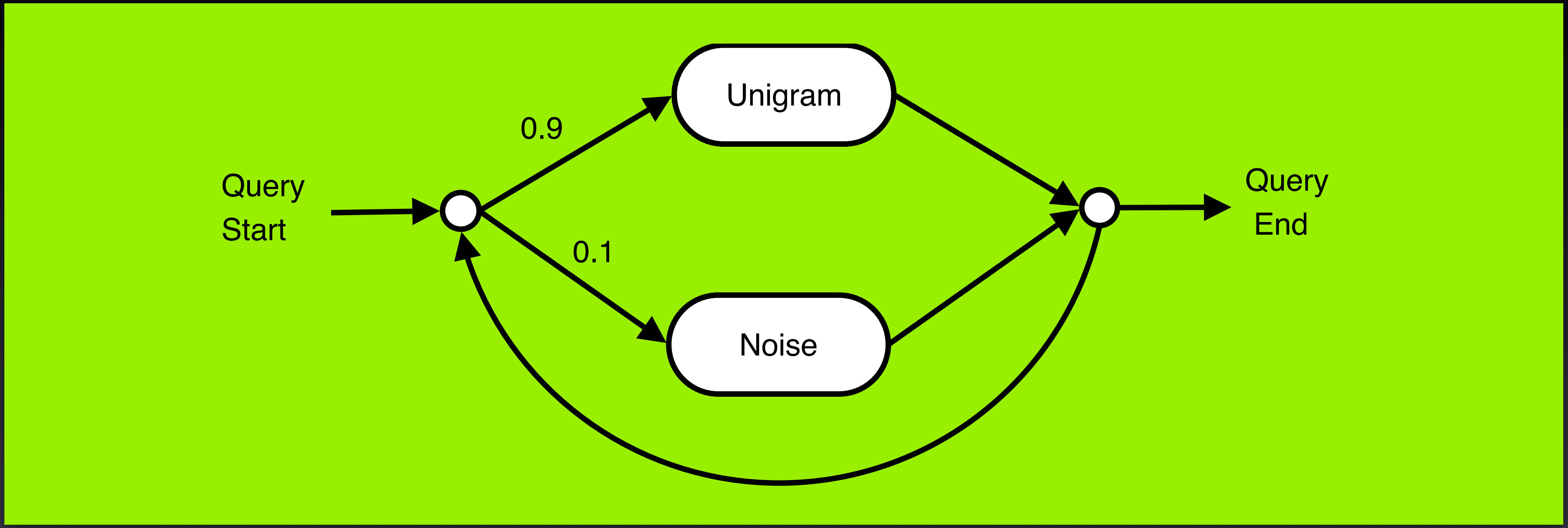
Algorithm (3)

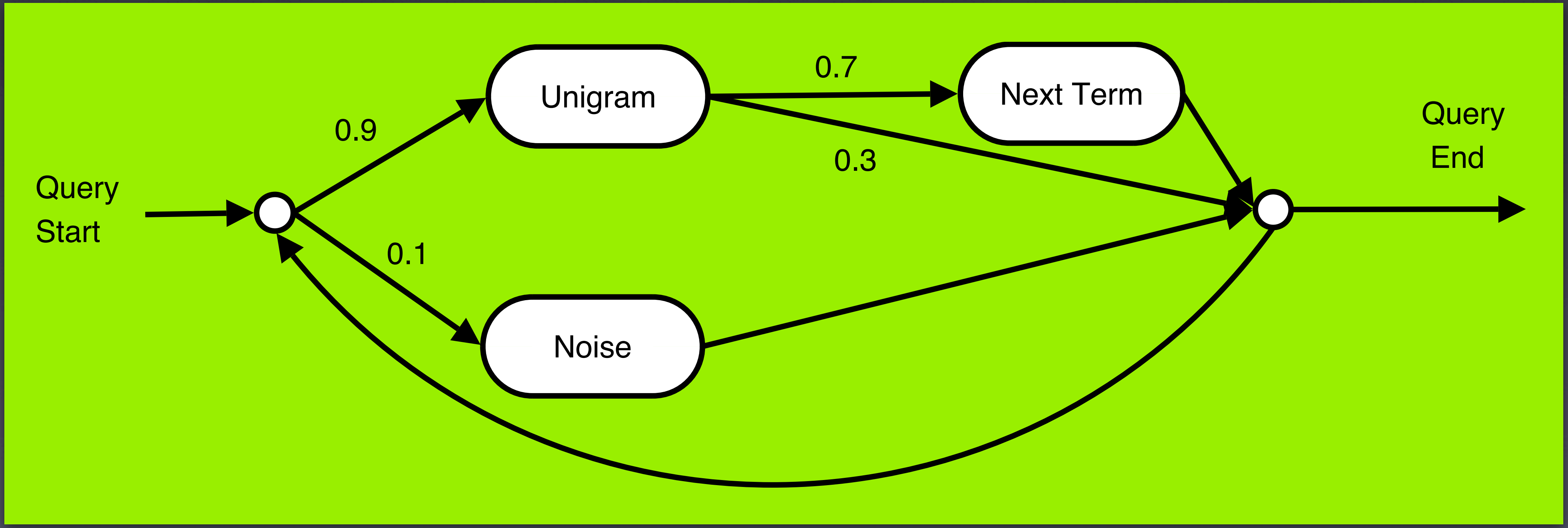
- Generate different types of queries by using different information to estimate the probability of a term being recalled by the user, $p(t|d)$
  - *popular*: most popular/common terms
  - *discriminative*: according to inverse doc freq
  - *uniform*: indiscriminate recollection
- Uniform produced queries most similar to real ones (TREC-ent, email search)

WebCLEF 2006 settings

- Uniform sampling
- Include query noise and phrase extraction
    - *auto-uni*
    - *auto-bi*

Query generation



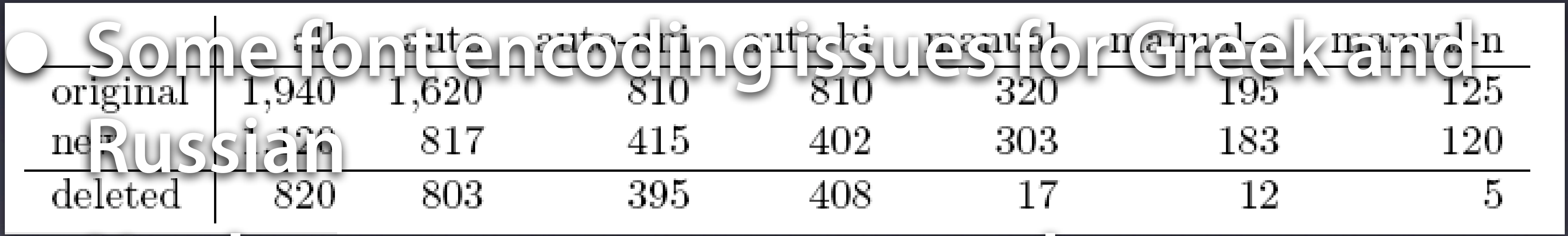


More details

- Indexing and sampling performed using the Lemur language modeling toolkit
  - Query length *k* selected using a Poisson distribution with mean 3
  - Restrictions on sampled query terms
    - Size at least 3, no numeric characters
  - Document prior *p*(*d*) uniform
- 27 primary domains
  - 30 topics auto-uni, 30 topics auto-bi

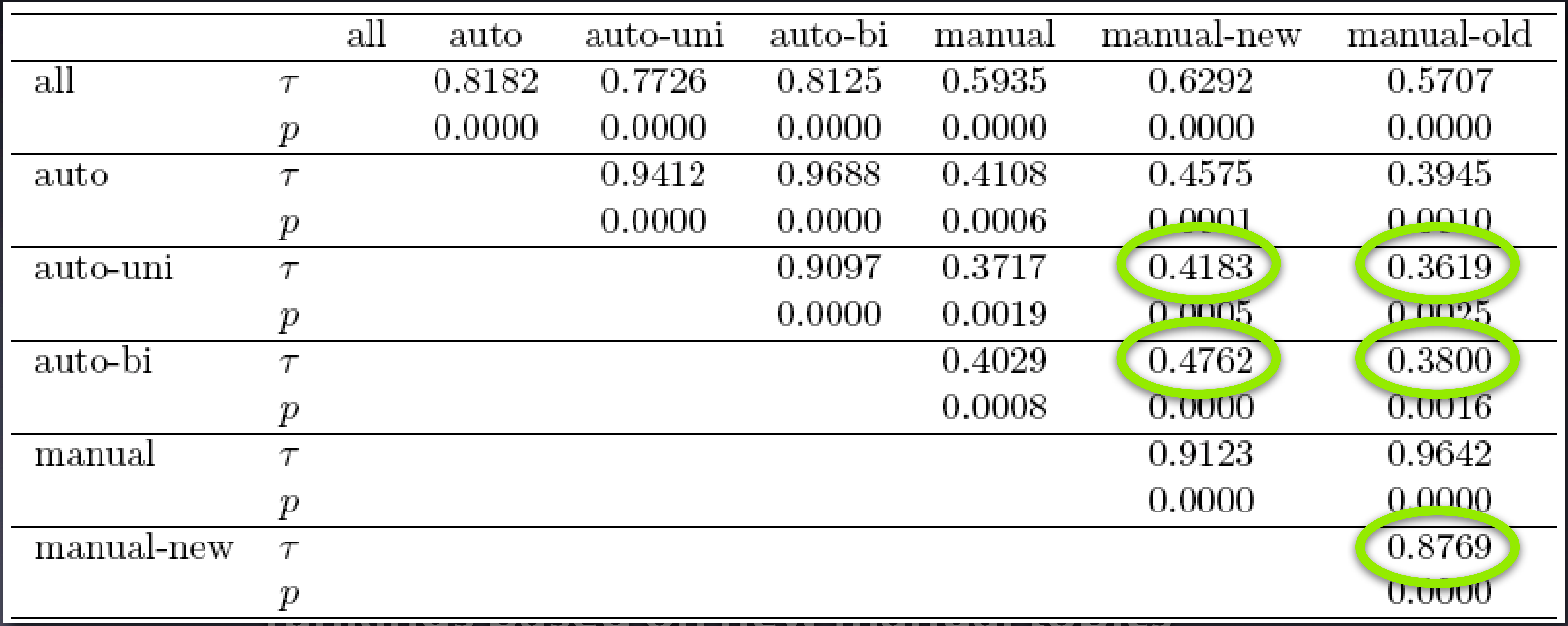
Issues

- Performance on automatic topics frequently poor
  - Mixed language (e.g., Portuguese plus English from navigation panel)
- Some font encoding issues for Greek and Russian
- No document structure used



	all	auto	auto-[illegible]	auto-[illegible]	manual	manual-[illegible]	manual-n
original	1,940	1,620	810	810	320	195	125
new	1,120	817	415	402	303	183	120
deleted	820	803	395	408	17	12	5

Rankings of runs



		all	auto	auto-uni	auto-bi	manual	manual-new	manual-old
all	τ		0.8182	0.7726	0.8125	0.5935	0.6292	0.5707
	p		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
auto	τ			0.9412	0.9688	0.4108	0.4575	0.3945
	p			0.0000	0.0000	0.0006	0.0001	0.0010
auto-uni	τ				0.9097	0.3717	0.4183	0.3619
	p				0.0000	0.0019	0.0005	0.0025
auto-bi	τ					0.4029	0.4762	0.3800
	p					0.0008	0.0000	0.0016
manual	τ						0.9123	0.9642
	p						0.0000	0.0000
manual-new	τ							0.8769
	p							0.0000

rankings based on new manual topics

Next steps

- Further experimentation with term dependencies, document structure, … in the generation process
- Generator will be made available