Query reformulation

model and patterns

Paolo Boldi
Francesco Bonchi
Carlos Castillo
Sebastiano Vigna

Università degli studi
di Milano, Italy

Yahoo! Research
Barcelona, Spain

Query reformulation

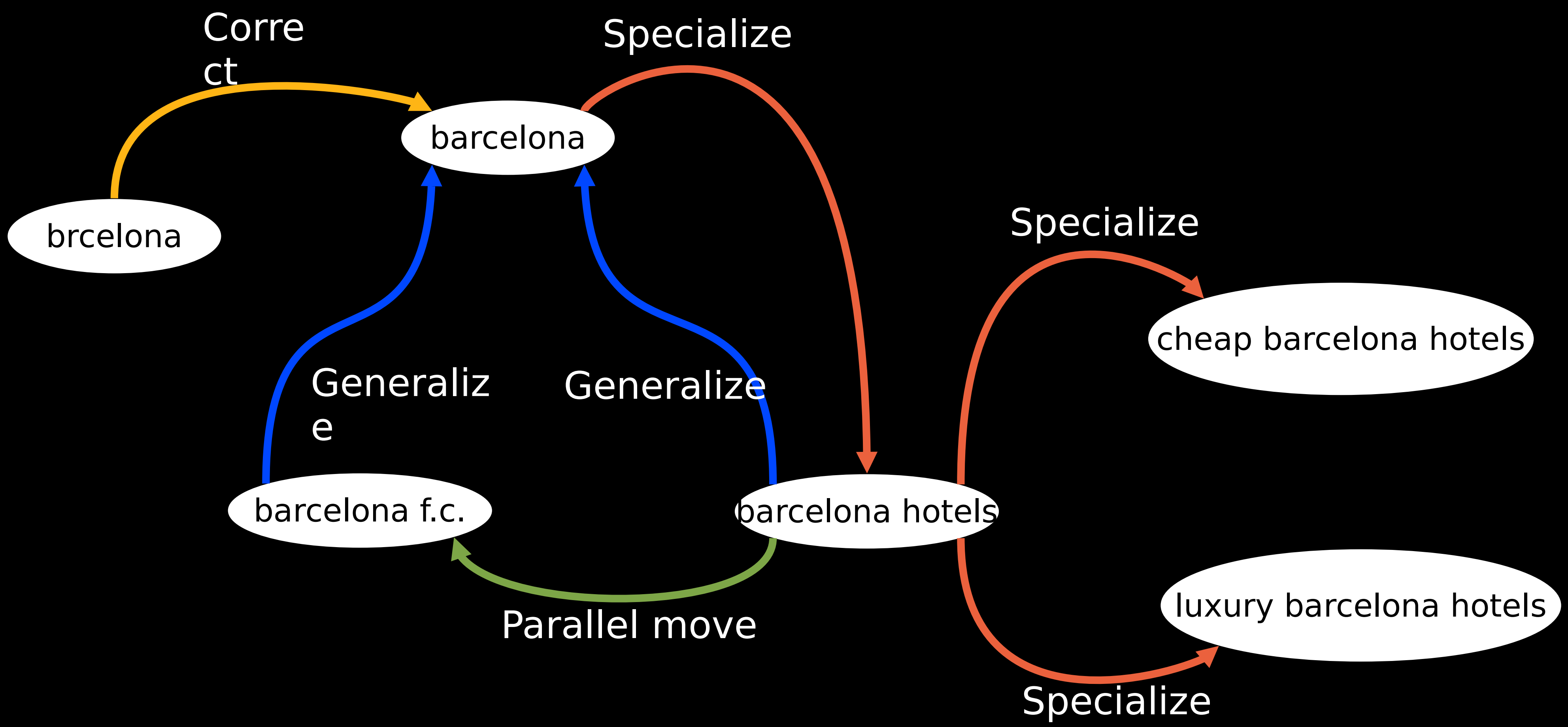
model and patterns:

from “dango” to “japanese cakes”

Paolo Boldi[M]
Francesco Bonchi[Y]
Carlos Castillo[Y]
Sebastiano Vigna[M]

[M] Università degli studi
di Milano, Italy

[Y] Yahoo! Research
Barcelona, Spain



Rieh, S. Y. and Xie, H: “Analysis of multiple query reformulations on the web”. IPM 32 (3) 2006.

Reformulation types

Error correction

startford cinema → stratford cinema

Generalization (“zoom out”)

barcelona hotels → barcelona

Specialization (“zoom in”)

barcelona soccer → barcelona camp nou

Rieh and Xie: “Analysis of multiple query reformulations”. IPM 2006.

Zoom-in, zoom-out, pan, names comes from Y!SAMA

Reformulation types

Rephrasing

- wikipedia english → english wikipedia
- robbs celebrities → robbs celebs

Parallel move

- barcelona → rome

Rieh and Xie: “Analysis of multiple query reformulations”. IPM 2006.

Why model reformulation types?

Improved session segmentation

Improved recommendations

Improved session understanding in general

Research **agenda**

Automatically classify query reformulation types

Study **patterns** of query reformulation

C C S S G S ... S P S C S S ... *session DNA*

Annotate the query-flow graph

P. Boldi, F. Bonchi, C. Castillo, S. Vigna: “Query Reformulation Model and Patterns”. 2008.

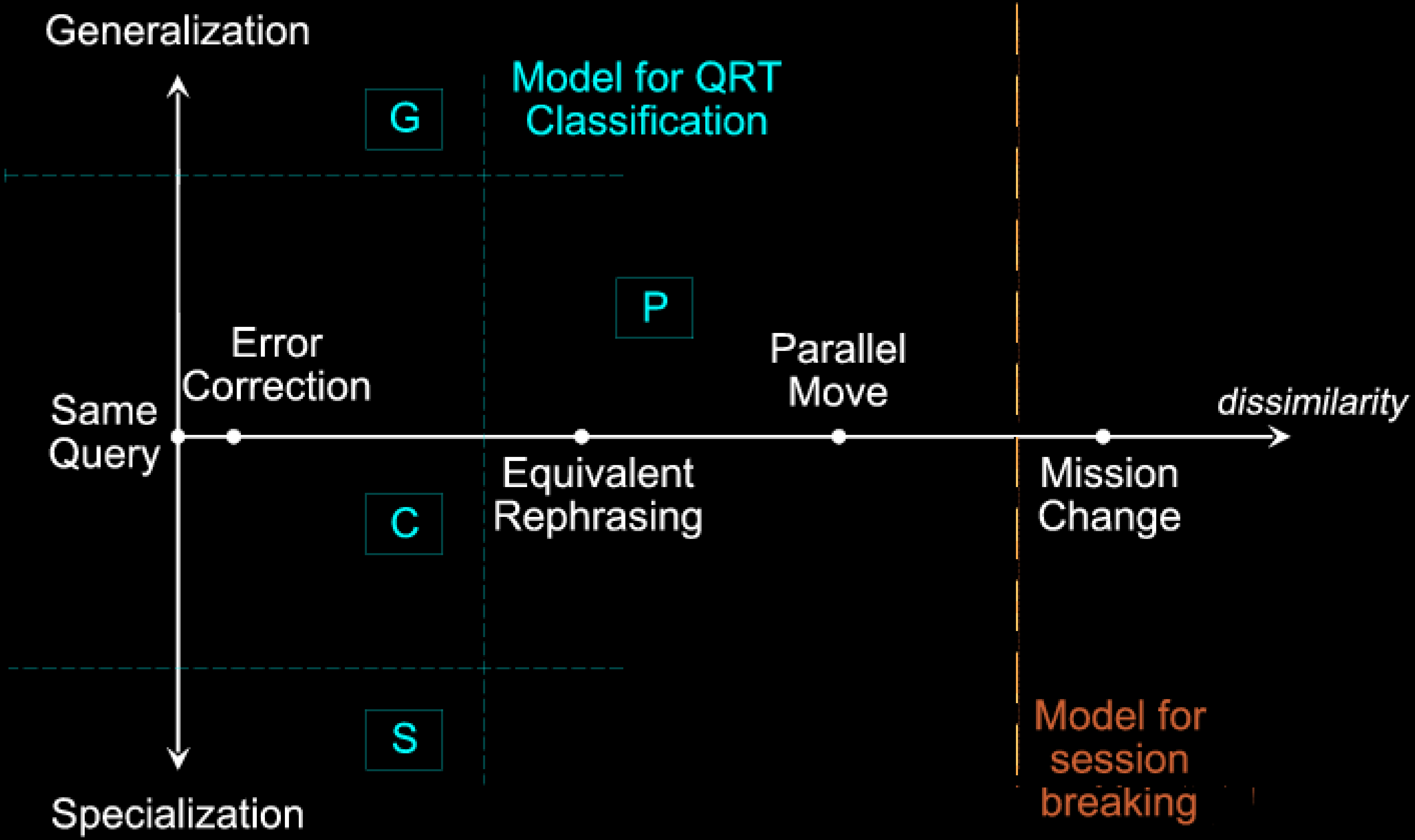
Research **agenda**

Automatically classify query reformulation types

Study **patterns** of query reformulation

C C S S G S … S P S C S S … *session DNA*

Annotate the query-flow graph



P. Boldi, F. Bonchi, C. Castillo, S. Vigna: “Query Reformulation Model and Patterns”. 2008.

Model for classification

Labeled examples

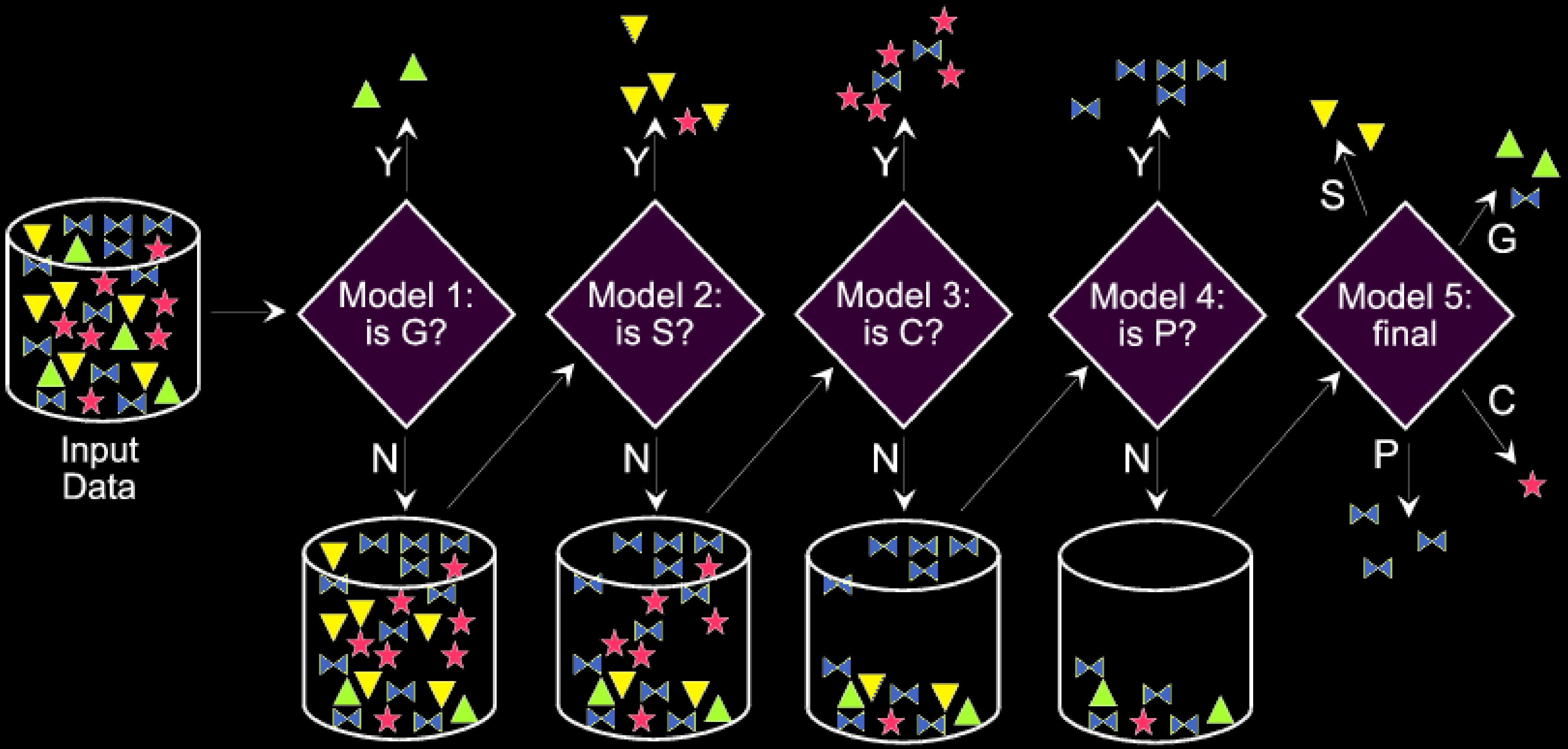
- 1,357 examples, 2/3 training 1/3 testing

Features

- Same as chains + edit distance + delta lengths + ...

Learning method

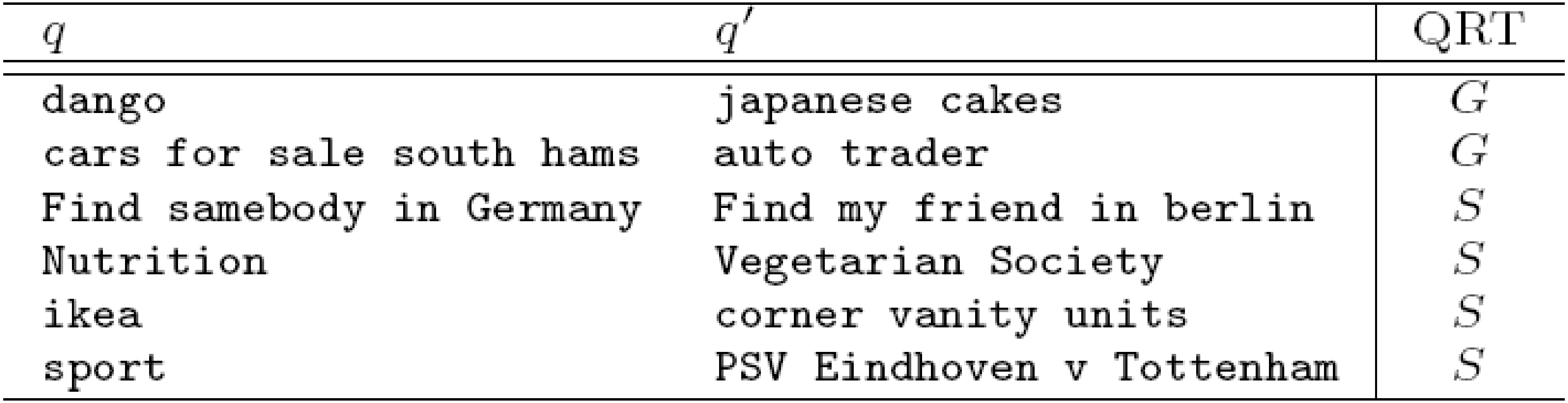
- Find easy cases first, solve hard cases later


Input
Data
Model 1:
is G?
Y
N
Model 2:
is S?
Y
N
Model 3:
is C?
Y
N
Model 4:
is P?
Y
N
Model 5:
final
S
G
C
P

Rule 1 of model 1: $is\_G?$	**Rule 1 of model 2:** $is\_S?$
if $terms.cosine > 0.47$ **and** $deltaLenRel \leq -0.37$ **then** $is\_G? = Y$	**if** $ngrams.cosine > 0.42$ **and** $terms.deltaLen > 1$ **then** $is\_S? = Y$

Rule 1 of model 3: $is\_C?$	**Rule 1 of model 4:** $is\_P?$
if $avgSessPosition \leq 1.91$ **and** $levenshtein \leq 3$ **then** $is\_C? = Y$	**if** $avgRelPosition > 0.65$ **and** $terms.jaccard \leq 0.25$ **and** $deltaLen \leq 5$ **and** $terms.deltaLen > 0$ **then** $is\_P? = Y$

Example classifier output



q	q'	QRT
dango	japanese cakes	G
cars for sale south hams	auto trader	G
Find samebody in Germany	Find my friend in berlin	S
Nutrition	Vegetarian Society	S
ikea	corner vanity units	S
sport	PSV Eindhoven v Tottenham	S

92% accuracy in the 4-classes problem

Research **agenda**

Automatically classify query reformulation types

Study **patterns** of query reformulation

C C S S G S ... S P S C S S ... *session DNA*

Annotate the query-flow graph

P. Boldi, F. Bonchi, C. Castillo, S. Vigna: “Query Reformulation Model and Patterns”. 2008.

Datasets

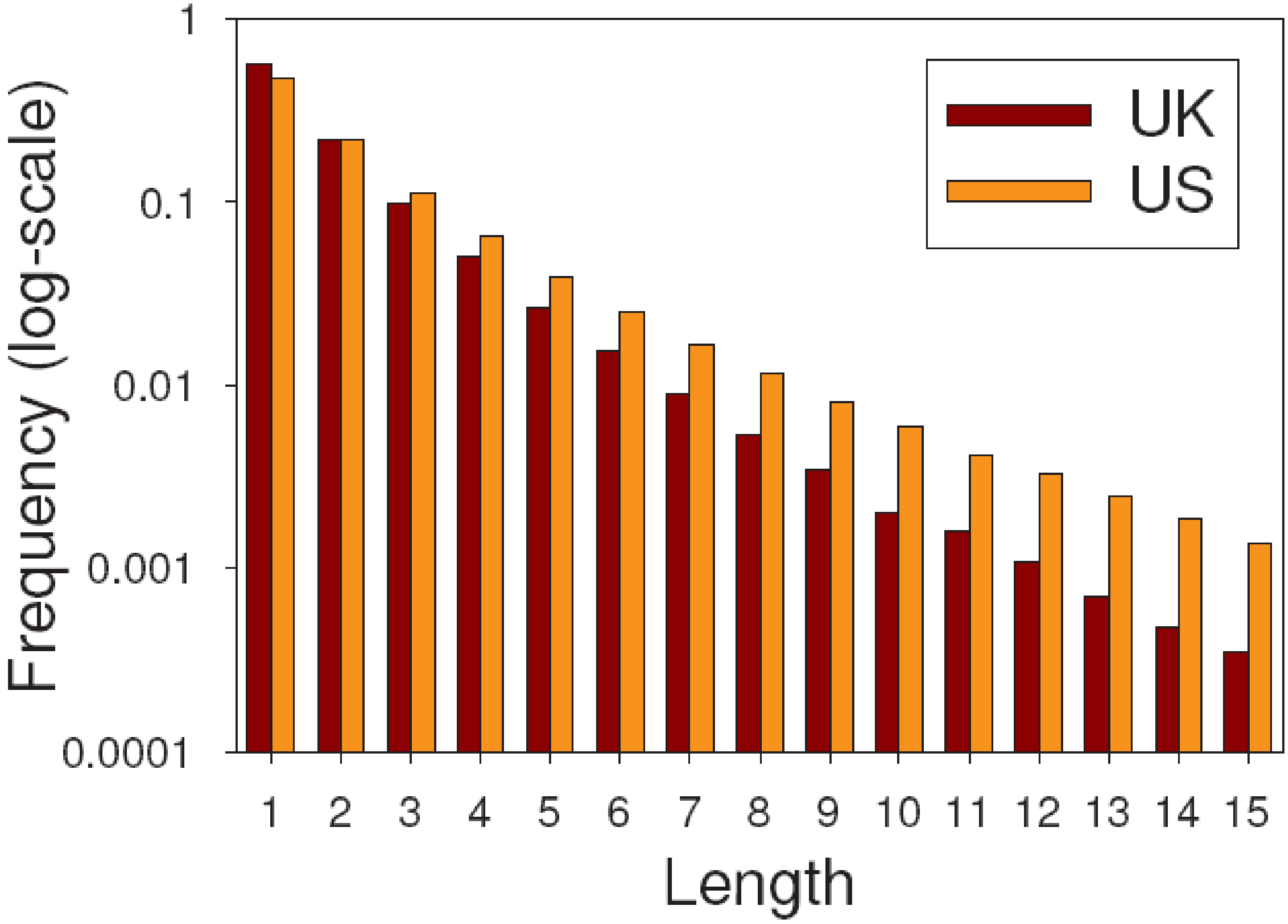
Yahoo! UK search engine

- 3.4M chains containing 6.6M queries

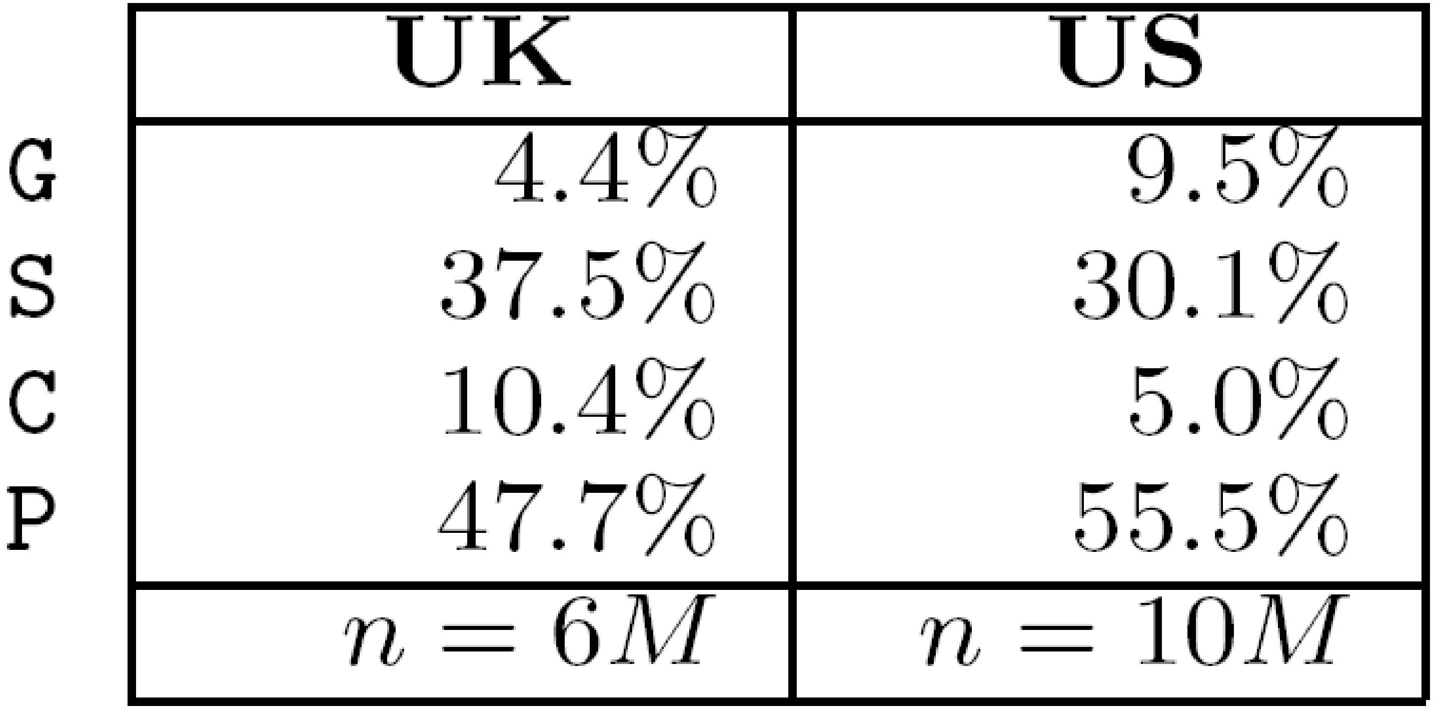
Yahoo! US search engine

- 4.0M chains containing 10.5M queries

Distribution of chain length

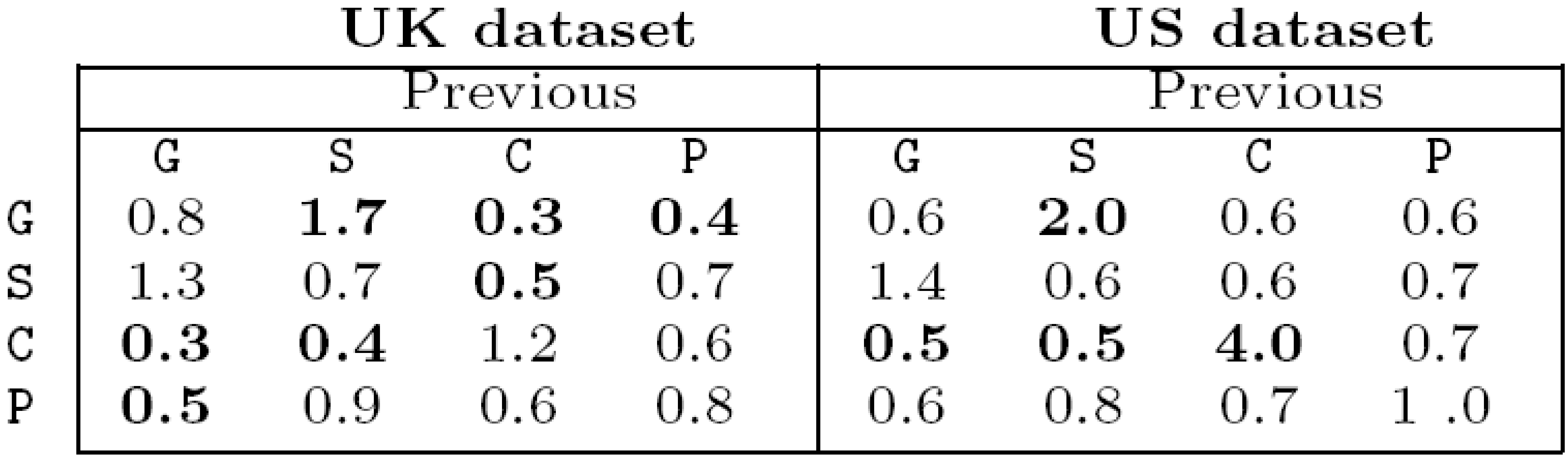


Distribution of reformulation types



	UK	**US**
G	4.4%	9.5%
S	37.5%	30.1%
C	10.4%	5.0%
P	47.7%	55.5%
	$n = 6M$	$n = 10M$

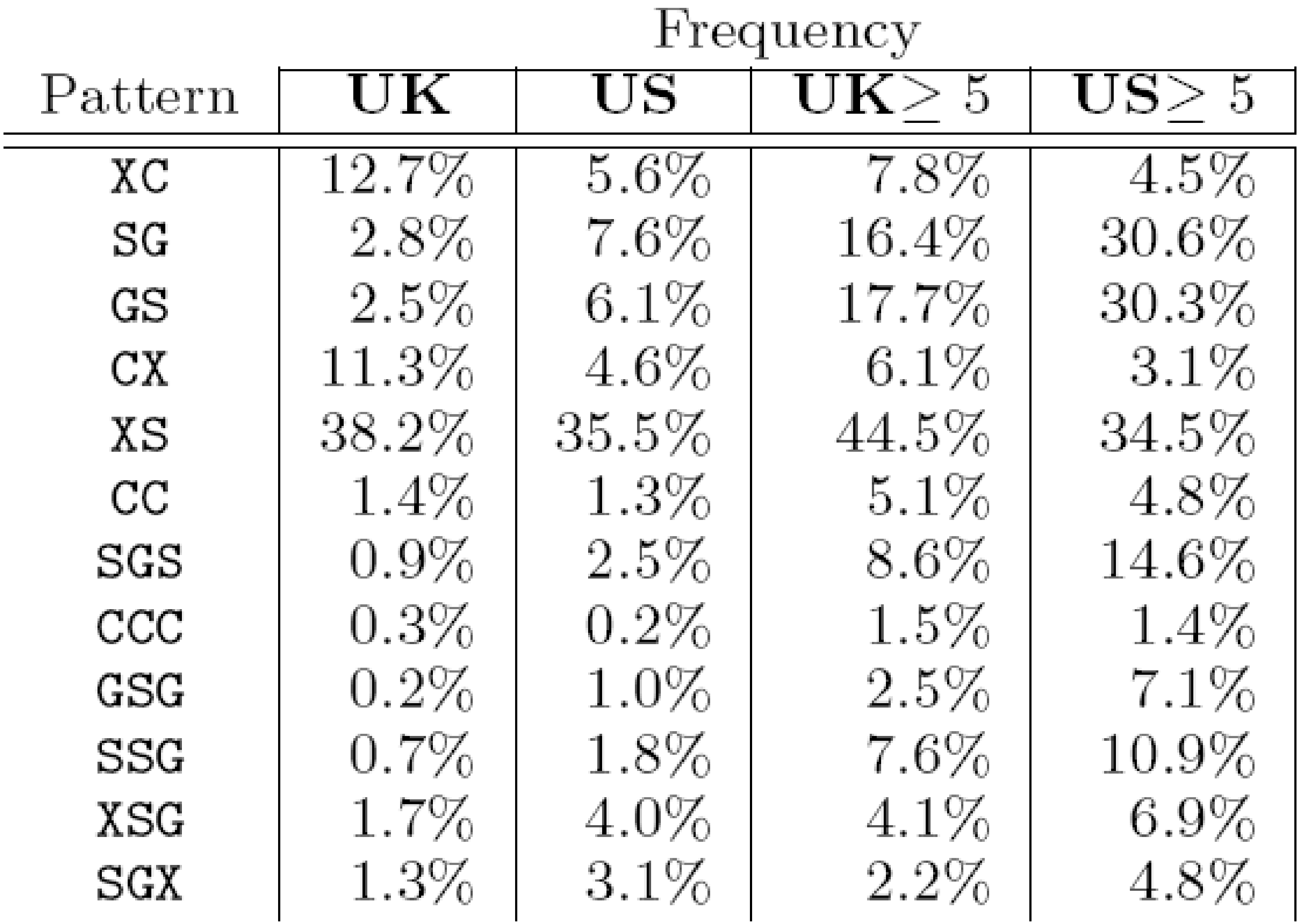
Conditional probability wrt prior
P(x|previous=y) / P(x)



	UK dataset				**US dataset**			
	Previous				Previous			
	G	S	C	P	G	S	C	P
G	0.8	**1.7**	**0.3**	**0.4**	0.6	**2.0**	0.6	0.6
S	1.3	0.7	**0.5**	0.7	1.4	0.6	0.6	0.7
C	**0.3**	**0.4**	1.2	0.6	**0.5**	**0.5**	**4.0**	0.7
P	**0.5**	0.9	0.6	0.8	0.6	0.8	0.7	1 .0

Generalizations appear after specializations
Corrections follow more corrections

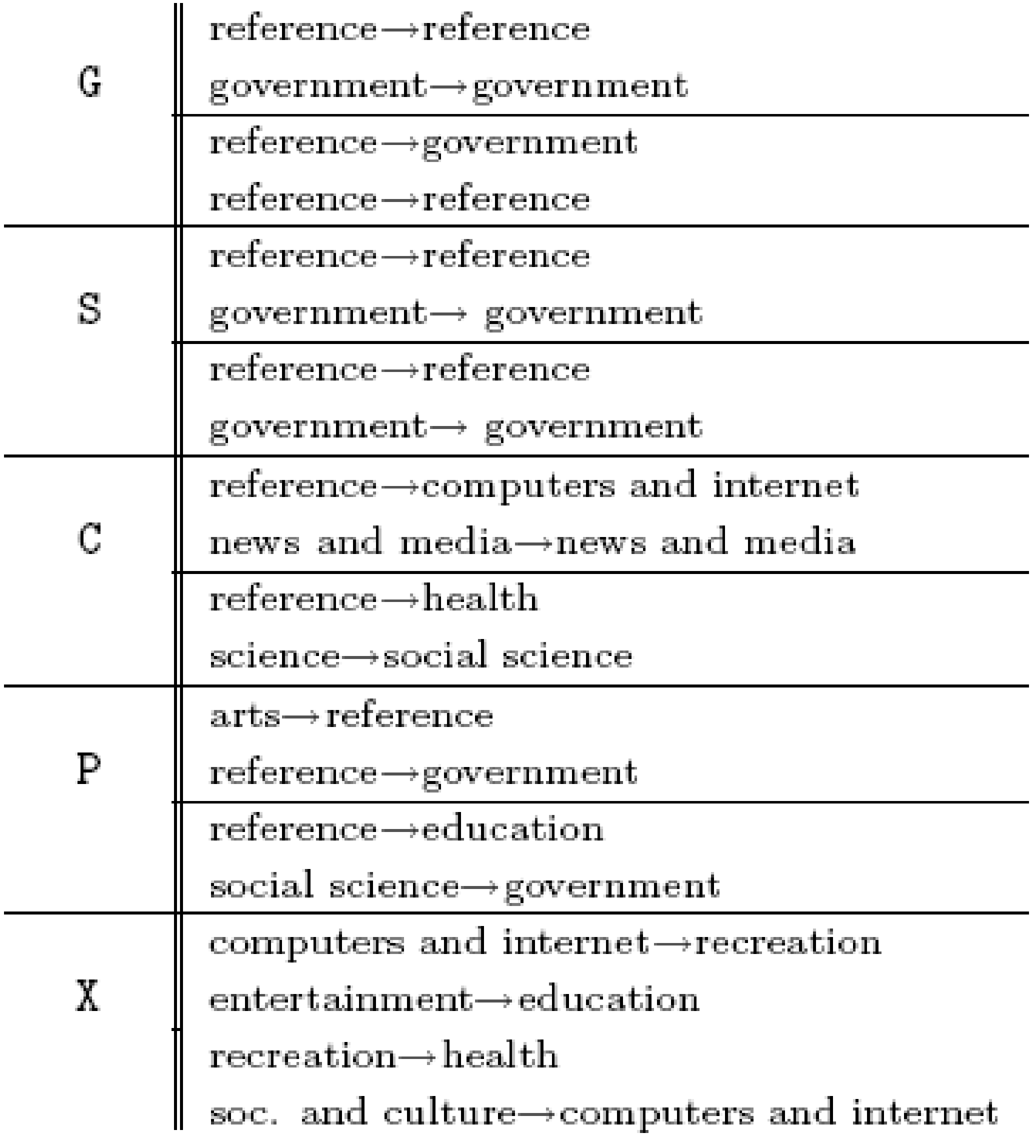
Salient patterns



	Frequency			
Pattern	**UK**	**US**	**UK**≥ 5	**US**≥ 5
XC	12.7%	5.6%	7.8%	4.5%
SG	2.8%	7.6%	16.4%	30.6%
GS	2.5%	6.1%	17.7%	30.3%
CX	11.3%	4.6%	6.1%	3.1%
XS	38.2%	35.5%	44.5%	34.5%
CC	1.4%	1.3%	5.1%	4.8%
SGS	0.9%	2.5%	8.6%	14.6%
CCC	0.3%	0.2%	1.5%	1.4%
GSG	0.2%	1.0%	2.5%	7.1%
SSG	0.7%	1.8%	7.6%	10.9%
XSG	1.7%	4.0%	4.1%	6.9%
SGX	1.3%	3.1%	2.2%	4.8%

Specialization/Generalization pairs
Corrections beginning or ending a chain

Topical patterns



G	reference→reference
	government→government
	reference→government
	reference→reference
S	reference→reference
	government→ government
	reference→reference
	government→ government
C	reference→computers and internet
	news and media→news and media
	reference→health
	science→social science
P	arts→reference
	reference→government
	reference→education
	social science→government
X	computers and internet→recreation
	entertainment→education
	recreation→health
	soc. and culture→computers and internet

Research **agenda**

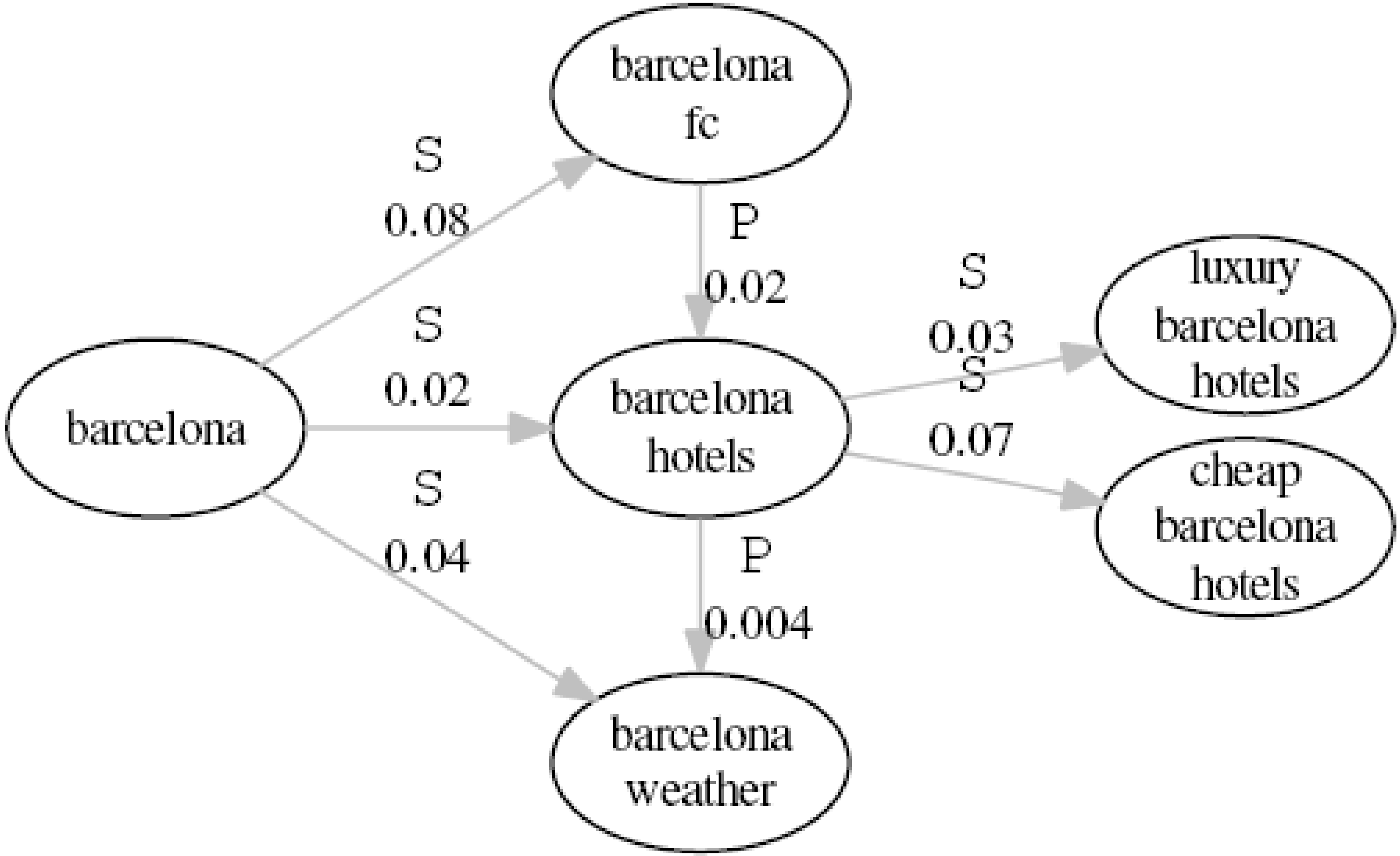
Automatically classify query reformulation types

Study **patterns** of query reformulation
C C S S G S ... S P S C S S ... *session DNA*

Annotate the query-flow graph

P. Boldi, F. Bonchi, C. Castillo, S. Vigna: “Query Reformulation Model and Patterns”. 2008.

Example annotated sub-graph



Interesting properties

Let G, S, P, C represent the corresponding slice

of the query-flow graph

Correlated pairs:

G and S^T, S and G^T (tend to be anti-symmetric)

C and C^T, P and P^T (tend to be symmetric)

Entropy measures

Transition-type entropy

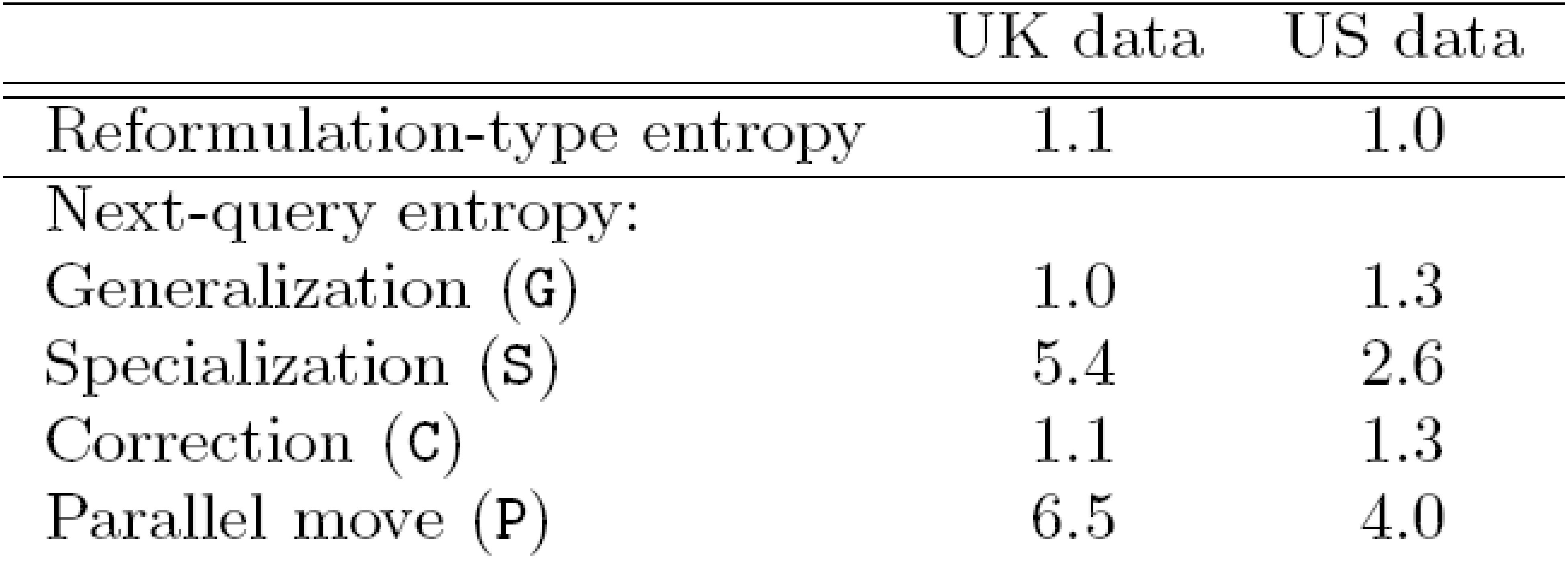
- Maximum 2 bits (4 transition types)

Next-query entropy

- Maximum $\log_2$(|Queries|-1)

Note: US data was large, dropped count=1

Average entropy (freq > 100)



	UK data	US data
Reformulation-type entropy	1.1	1.0
Next-query entropy:		
Generalization (G)	1.0	1.3
Specialization (S)	5.4	2.6
Correction (C)	1.1	1.3
Parallel move (P)	6.5	4.0

Specializatio:	$2^{5.4} = 42$	$2^{2.6} = 6$
Parallel move	$2^{6.5} = 91$	$2^{4.0} = 16$

Conclusions

High accuracy in 4-classes: 92%

Specializations and Generalizations alternate

Corrections are common at the beginning and at the end of a chain

Large entropy in specializations/parallel moves

Follow-up work: query recommendation

Q&A