

# Towards Optimum Query Segmentation: In Doubt Without

Matthias Hagen

Martin Potthast

Anna Beyer

Benno Stein

Bauhaus-Universität Weimar  
matthias.hagen@uni-weimar.de

CIKM 2012  
Ka'anapali, Maui  
October 31, 2012

# All search engines ...



# All search engines ...



new york times square dance

# All search engines ... face the same problem



## Problem

What is the user's information need?

new york times square dance

Is it: new york times square dance ?



Image source: [<http://www.thepacktimes.com/n2/images/stories/large/2009/08/06/Bollywood1.jpg>]

"All the News  
That's Fit to Print"

# The New York Times

**Late Edition**  
Today, a shower, yielding to sun, high 64. Tonight, chilly, low 45. Tomorrow, sunny and cool, high 64. Yesterday's high 70, low 54. Weather map and details, Page 24.

VOL. CLVIII • No. 54,678

© 2009 The New York Times

NEW YORK, SUNDAY, MAY 17, 2009

\$5 beyond the greater New York metropolitan area. \$4.00



## Square Dance

**19 U.S. States Have Designated It As  
Their Official State Dance**

### From a Theory To a Consensus On Emissions

**Permits Gain Political  
Edge Over Taxation**

By JOHN M. BRODER

WASHINGTON — As Congress weighs imposing a mandatory limit on climate-altering gases — an outcome still far from certain — it is likely to turn to a system that sets a government ceiling on total emissions and allows polluting industries to buy and sell permits to meet it.

That approach, known as cap and trade, has been embraced by President Obama. Democratic leaders in Congress, mainstream environmental groups and a growing number of business interests, including energy-consuming industries like autos, steel and aluminum.

But not long ago, many of today's supporters dismissed the idea of tradable emissions permits as an industry-inspired Republican scheme to avoid the real costs of cutting air pollution. The right answer, they said, was strict government regulation, state-of-the-art technology and a federal tax on every ton of harmful emissions.

How did cap and trade, hatched as an academic theory in obscure economic journals half a century ago, become the policy of choice in the debate over how to slow the heating of the planet? And how did it come to eclipse the idea of simply skipping a tax on energy consumption that befores the public square or leaves the nation hostage to foreign oil

### CONSERVATIVES MAP STRATEGIES ON COURT FIGHT

**MEMOS OUTLINE ATTACKS**

**Hoping to Re-Energize  
G.O.P. by Opposing  
Obama's Choice**

By CHARLIE SAVAGE

WASHINGTON — If President Obama nominates Judge Diane B. Wood to the Supreme Court, conservatives plan to attack her as an "outspeaker" supporter of "abortion, including partial-birth abortion."

If he nominates Judge Sonia Sotomayor, they plan to accuse her of being "willing to expand constitutional rights beyond the text of the Constitution."

And if he nominates Kathleen M. Sullivan, a law professor at Stanford, they plan to denounce her as a "prominent supporter of homosexual marriage."

Preparing to oppose the confirmation of Mr. Obama's eventual choice to succeed Justice David H. Souter, who is retiring, conservative groups are working together to stockpile ammunition. Ten memorandums summarizing their research, obtained by The New York Times, provide a window onto how they hope to frame the coming debate.

The memorandums dissect possible nominees' records, noting statements the groups find objectionable on issues like abortion, campaign financing, the

# Segment your queries!

## The benefits

- Improved precision
- Potential disambiguation
- Reformulations on segment level

## The syntax

Quotes around segments: "new york" "times square" dance

# Segment your queries!

## The benefits

- Improved precision
- Potential disambiguation
- Reformulations on segment level

## The syntax

Quotes around segments: "new york" "times square" dance

## The “minor” issue . . .

Most web searchers are not even aware of the quotes option.



Automatic pre-retrieval query segmentation

## Automatic **pre-retrieval** query segmentation

Remark: Runtime is crucial!

# The computational problem as we see it

## Query Segmentation

- Given a keyword query
- Find the “best” segmentation

Remarks: We assume correct spelling!  
We do not change keywords!

## Example

Given the query	<code>new york times square dance</code>
Solutions could be	<code>"new york" "times square" dance</code> <code>"new york times" "square dance"</code>
But not (word order!)	<code>"new york" "dance times square"</code> (a Latin dance studio in NYC)

# Standing on the shoulders of ...

Mutual information	[Risvik et al., WWW 2003] [Jones et al., WWW 2006] [Huang et al., WWW 2010]
Supervised learning	[Bergsma and Wang, EMNLP-CoNLL 2007] [Bendersky et al., SIGIR 2009]
Unsupervised learning	[Tan and Peng, WWW 2008] [Zhang et al., ACL-IJCNLP 2009]
Retrieval feedback	[Brenes et al., CERI 2010] [Bendersky et al., CIKM 2010] [Bendersky et al., ACL 2011]
Query log	[Mishra et al., WWW 2011] [Li et al., SIGIR 2011] [Roy et al., SIGIR 2012]
Web frequencies	[Hagen et al., SIGIR 2010] [Hagen et al., WWW 2011]

# Same same but different

## Typical algorithmic scheme: one strategy for all queries

- Segment every possible phrase
- More frequent phrases = better segments
- Boost Wikipedia titles

## Major difference

Source and processing of frequencies

# Same same but different

## Typical algorithmic scheme: one strategy for all queries

- Segment every possible phrase
- More frequent phrases = better segments
- Boost Wikipedia titles

## Major difference

Source and processing of frequencies

## Question

Is it really reasonable to treat every query in the same way?

# How would humans do it?

## The study

- 54 000 queries (3–10 keywords) from “filtered” AOL log
  - Sampling follows frequency and length distribution
  - 50% noun phrase queries
- 10 annotators per query via Mechanical Turk
  - 1800 workers in total (300 queries per worker)

## Key finding

- Segmentation behavior depends on query type
  - Noun phrase queries: most keywords in segments
  - Other queries: most keywords not in segments

# How would humans do it?

## The study

- 54 000 queries (3–10 keywords) from “filtered” AOL log
  - Sampling follows frequency and length distribution
  - 50% noun phrase queries
- 10 annotators per query via Mechanical Turk
  - 1800 workers in total (300 queries per worker)

## Key finding

- Segmentation behavior depends on query type
  - Noun phrase queries: most keywords in segments
  - Other queries: most keywords not in segments



## Hybrid scheme: different strategies for different query types

- Noun phrase queries
  - Segment every possible phrase
  - Use state of the art (e.g., [Hagen et al., WWW 2011])
- Other queries
  - Segment only low risk phrases
  - In doubt, leave phrases without quotes

## Hybrid scheme: different strategies for different query types

- Noun phrase queries
  - Segment every possible phrase
  - Use state of the art (e.g., [Hagen et al., WWW 2011])
- Other queries
  - Segment only low risk phrases
  - In doubt, leave phrases without quotes
  - **New method required ...**

# KISS – Keep it simple and stupid!



Image source: [[http://1.bp.blogspot.com/\\_UjOZKrzYyS4k/THRrh8KPvVI/AAAAAAAAAoc/Bx3HjnlRytc/s400/lipstick-mirror.jpg](http://1.bp.blogspot.com/_UjOZKrzYyS4k/THRrh8KPvVI/AAAAAAAAAoc/Bx3HjnlRytc/s400/lipstick-mirror.jpg)]

# Wikipedia titles are high quality phrases



Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact Wikipedia

Toolbox  
Print/export

Languages  
العربية  
বাংলা  
Català  
Česky

Article **Discussion**

Read

## Times Square


From Wikipedia, the free encyclopedia  
(Redirected from [Times square](#))

*For the subway station, see [Times Square - 42nd Street \(New York City Subway\)](#). For other uses, see [Times Square \(disambiguation\)](#).*

**Times Square** is a major commercial intersection in the borough of Manhattan in New York City, at the junction of Broadway and Seventh Avenue and stretching from West 42nd to West 47th Streets. The extended Times Square area, also called the Theatre District, consists of the blocks between Sixth and Eighth Avenues from east to west, and West 40th and West 53rd Streets from south to north, making up the western part of the commercial area of Midtown Manhattan.



# (Most) Wikipedia titles are high quality phrases



WIKIPEDIA  
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia

▼ Interaction

- Help
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia

► Toolbox

► Print/export

▼ Languages

- Dansk
- Español
- Français
- Slovenščina
- Svenska

Article

[Discussion](#)

Read

[View source](#)

[View history](#)

## Toilet paper orientation

From Wikipedia, the free encyclopedia

There are two choices of **toilet paper orientation** when using a **toilet roll holder** with a horizontal **axle parallel** to the wall:



The over orientation



The under orientation

## Hybrid strategy mix to mimic human behavior

Noun phrase queries:	Use [Hagen et al., WWW 2011]
Other queries:	Segment only Wikipedia titles

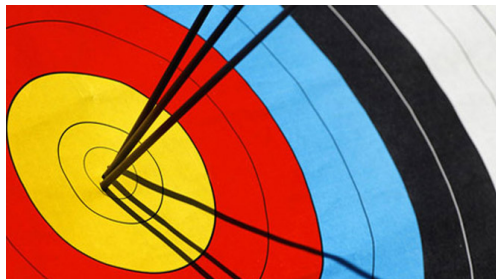
## Hybrid strategy mix to mimic human behavior

Noun phrase queries:	Use [Hagen et al., WWW 2011]
Other queries:	Segment only Wikipedia titles

## Question

But does this yield the “best” segmentation?

# Different meanings of “best”



Accuracy



TREC-style



# How to measure accuracy of segmentations?

## The standard corpus

[Bergsma and Wang, EMNLP-CoNLL 2007]

- 500 queries (only noun phrases!)
- 3 annotators segmented all queries
- Accuracy against individual annotator or best fit

# How to measure accuracy of segmentations?

## The standard corpus

[Bergsma and Wang, EMNLP-CoNLL 2007]

- 500 queries (only noun phrases!)
- 3 annotators segmented all queries
- Accuracy against individual annotator or best fit

## Example

Reference	"new york times"	"square dance"	(2 segments)
Computed	"new york times"	square dance	(3 segments)

# How to measure accuracy of segmentations?

## The standard corpus

[Bergsma and Wang, EMNLP-CoNLL 2007]

- 500 queries (only noun phrases!)
- 3 annotators segmented all queries
- Accuracy against individual annotator or best fit

## Example

Reference	"new york times"	"square dance"	(2 segments)
Computed	"new york times"	square dance	(3 segments)

- Query accuracy: 0 (computed  $\neq$  reference)
- Precision: 0.33 (1 out of 3 computed segments correct)
- Recall: 0.5 (1 out of 2 reference segments found)
- Break accuracy: 0.75 (3 out of 4 potential segment breaks correct)

# How to measure accuracy of segmentations?

## The standard corpus

[Bergsma and Wang, EMNLP-CoNLL 2007]

- 500 queries (only noun phrases!)
- 3 annotators segmented all queries
- Accuracy against individual annotator or best fit

## Example

Reference	"new york times"	"square dance"	(2 segments)
Computed	"new york times"	square dance	(3 segments)

- Query accuracy: 0 (computed  $\neq$  reference)
- Precision: 0.33 (1 out of 3 computed segments correct)
- Recall: 0.5 (1 out of 2 reference segments found)
- Break accuracy: 0.75 (3 out of 4 potential segment breaks correct)

# How to measure accuracy of segmentations?

## The standard corpus

[Bergsma and Wang, EMNLP-CoNLL 2007]

- 500 queries (only noun phrases!)
- 3 annotators segmented all queries
- Accuracy against individual annotator or best fit

## Example

Reference	"new york times"	"square dance"	(2 segments)
Computed	"new york times"	square dance	(3 segments)

- Query accuracy: 0 (computed  $\neq$  reference)
- Precision: 0.33 (1 out of 3 computed segments correct)
- Recall: 0.5 (1 out of 2 reference segments found)
- Break accuracy: 0.75 (3 out of 4 potential segment breaks correct)

# How to measure accuracy of segmentations?

## The standard corpus

[Bergsma and Wang, EMNLP-CoNLL 2007]

- 500 queries (only noun phrases!)
- 3 annotators segmented all queries
- Accuracy against individual annotator or best fit

## Example

Reference	"new york times" "square dance"	(2 segments)
Computed	"new york times"   square dance	(3 segments)

- Query accuracy: 0 (computed  $\neq$  reference)
- Precision: 0.33 (1 out of 3 computed segments correct)
- Recall: 0.5 (1 out of 2 reference segments found)
- Break accuracy: 0.75 (3 out of 4 potential segment breaks correct)

## Our large-scale corpus

- 54 000 queries
- 10 annotators per query
- Each annotator segmented only small fraction

# Rethinking reference selection for larger corpora

## Our large-scale corpus

- 54 000 queries
- 10 annotators per query
- Each annotator segmented only small fraction

## Example

In the corpus

8 votes      "new york" "times square" dance

2 votes      "new york times" "square dance"

Computed      "new york times" "square dance"



# Rethinking reference selection for larger corpora

## Our large-scale corpus

- 54 000 queries
- 10 annotators per query
- Each annotator segmented only small fraction

## Example

In the corpus

8 votes      "new york" "times square" dance

2 votes      "new york times" "square dance"

Computed      "new york times" "square dance"

- Traditional query accuracy:    1      (best fit reference)
- Weighted by achieved votes:    0.25    (times 2/8 votes ratio)
- Absolute majority reference:    0      (no match)

# Rethinking reference selection for larger corpora

## Our large-scale corpus

- 54 000 queries
- 10 annotators per query
- Each annotator segmented only small fraction

## Example

In the corpus

8 votes "new york" "times square" dance

2 votes "new york times" "square dance"

Computed "new york times" "square dance"

- Traditional query accuracy: 1 (best fit reference)
- Weighted by achieved votes: 0.25 (times 2/8 votes ratio)
- Absolute majority reference: 0 (no match)

# Rethinking reference selection for larger corpora

## Our large-scale corpus

- 54 000 queries
- 10 annotators per query
- Each annotator segmented only small fraction

## Example

In the corpus

8 votes      "new york" "times square" dance

2 votes      "new york times" "square dance"

Computed      "new york times" "square dance"

- Traditional query accuracy:    1      (best fit reference)
- Weighted by achieved votes:    0.25    (times 2/8 votes ratio)
- Absolute majority reference:    0      (no match)

# Accuracy evaluation results

## Query accuracy on our corpus (weighted + majority reference)

[Hagen et al., WWW 2011]	PMI	Wikipedia titles	hybrid
0.481	0.520	0.638	<b>0.644</b>

## Observations

- Previous state of the art worse than PMI baseline
- Wikipedia titles form a very strong new baseline

# Accuracy evaluation results

## Query accuracy on our corpus (weighted + majority reference)

[Hagen et al., WWW 2011]	PMI	Wikipedia titles	hybrid
0.481	0.520	0.638	<b>0.644</b>

## Observations

- Previous state of the art worse than PMI baseline
- Wikipedia titles form a very strong new baseline

## Question

What does accuracy tell about the retrieval impact?

## The framework

- Document set: ClueWeb09
- Queries: Topics from Web and Million query tracks
- Search engines: Bing and Indri

Remark: 355 queries, 60% noun phrases

## The framework

- Document set: ClueWeb09
- Queries: Topics from Web and Million query tracks
- Search engines: Bing and Indri

Remark: 355 queries, 60% noun phrases

## Hybrid variants

	Accuracy	Bing	Indri
noun phrases	[WWW 2011]	None	None
other queries	Wikipedia titles	Wikipedia titles	[WWW 2011]

# Still not the top reached

## nDCG@10

	Bing		Indri	
	hybrid-B	OPT	hybrid-I	OPT
all queries	0.148	<b>0.170</b>	0.228	<b>0.308</b>
noun phrase	0.138	<b>0.156</b>	0.232	<b>0.311</b>
other	0.162	<b>0.189</b>	0.222	<b>0.304</b>

## Observations

- Noun phrase queries often best without segmentation
- Tailored hybrid variants outperform other segmentations
- But OPT (always best nDCG segmentation) still ahead of hybrid



# What about efficiency?



Runtime



Memory footprint

## System and implementation details

- Standard quad-core PC running Ubuntu 12.04
- Hash table for normalized frequencies
- Needs 59MB, 2 GB, or 12 GB of RAM [Brants et al., EMNLP-CoNLL 2007]

## Throughput

3 000–4 000 queries per second

Remark: A load of 1 billion queries per day means 12 000 queries per second.

Almost the end: The take-home messages!

## Main results

- Human segmentation behavior  
↪ *noun phrase vs. other queries*
- Hybrid segmentation scheme  
↪ *in doubt without*
- Accuracy and TREC evaluation  
↪ *accuracy  $\neq$  retrieval gain*

## Future work

- Optimum segmentation
- Larger retrieval study
- Diverse query types
- When to segment?

# What we have (not) done

## Main results

- Human segmentation behavior  
↪ *noun phrase vs. other queries*
- Hybrid segmentation scheme  
↪ *in doubt without*
- Accuracy and TREC evaluation  
↪ *accuracy  $\neq$  retrieval gain*

## Future work

- Optimum segmentation
- Larger retrieval study
- Diverse query types
- When to segment?

# What we have (not) done

## Main results

- Human segmentation behavior  
↪ *noun phrase vs. other queries*
- Hybrid segmentation scheme  
↪ *in doubt without*
- Accuracy and TREC evaluation  
↪ *accuracy  $\neq$  retrieval gain*

## Future work

- Optimum segmentation
- Larger retrieval study
- Diverse query types
- When to segment?

**Thank you**  
