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?
Is it: New York Times Square Dance?!

Square Dance

19 U.S. States Have Designated It As Their Official State Dance
Segment your queries!

**The benefits**
- Improved precision
- Potential disambiguation
- Reformulations on segment level

**The syntax**
Quotes around segments: "new york" "times square" dance

Hagen, Potthast, Beyer, Stein
Towards Optimum Query Segmentation: In Doubt Without
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.
The way out . . .

Automatic pre-retrieval query segmentation
The way out . . .

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

<table>
<thead>
<tr>
<th>Given the query</th>
<th>new york times square dance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solutions could be</td>
<td>&quot;new york&quot; &quot;times square&quot; dance</td>
</tr>
<tr>
<td></td>
<td>&quot;new york times&quot; &quot;square dance&quot;</td>
</tr>
<tr>
<td>But not (word order!)</td>
<td>&quot;new york&quot; &quot;dance times square&quot;</td>
</tr>
<tr>
<td></td>
<td>(a Latin dance studio in NYC)</td>
</tr>
<tr>
<td>Method</td>
<td>References</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Mutual information</td>
<td>[Risvik et al., WWW 2003]</td>
</tr>
<tr>
<td></td>
<td>[Jones et al., WWW 2006]</td>
</tr>
<tr>
<td></td>
<td>[Huang et al., WWW 2010]</td>
</tr>
<tr>
<td>Supervised learning</td>
<td>[Bergsma and Wang, EMNLP-CoNLL 2007]</td>
</tr>
<tr>
<td></td>
<td>[Bendersky et al., SIGIR 2009]</td>
</tr>
<tr>
<td>Unsupervised learning</td>
<td>[Tan and Peng, WWW 2008]</td>
</tr>
<tr>
<td></td>
<td>[Zhang et al., ACL-IJCNLP 2009]</td>
</tr>
<tr>
<td>Retrieval feedback</td>
<td>[Brenes et al., CERI 2010]</td>
</tr>
<tr>
<td></td>
<td>[Bendersky et al., CIKM 2010]</td>
</tr>
<tr>
<td></td>
<td>[Bendersky et al., ACL 2011]</td>
</tr>
<tr>
<td>Query log</td>
<td>[Mishra et al., WWW 2011]</td>
</tr>
<tr>
<td></td>
<td>[Li et al., SIGIR 2011]</td>
</tr>
<tr>
<td></td>
<td>[Roy et al., SIGIR 2012]</td>
</tr>
<tr>
<td>Web frequencies</td>
<td>[Hagen et al., SIGIR 2010]</td>
</tr>
<tr>
<td></td>
<td>[Hagen et al., WWW 2011]</td>
</tr>
</tbody>
</table>
Typical algorithmic scheme: one strategy for all queries

- Segment every possible phrase
- More frequent phrases $\Rightarrow$ 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
### 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
Algorithmic exploitation of key finding

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
Algorithmic exploitation of key finding

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!
Wikipedia titles are high quality phrases

Towards Optimum Query Segmentation: In Doubt Without
(Most) Wikipedia titles are high quality phrases

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
<table>
<thead>
<tr>
<th>Hybrid strategy mix to mimic human behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Noun phrase queries:</strong> Use [Hagen et al., WWW 2011]</td>
</tr>
<tr>
<td><strong>Other queries:</strong> Segment only Wikipedia titles</td>
</tr>
</tbody>
</table>
In doubt without

<table>
<thead>
<tr>
<th>Hybrid strategy mix to mimic human behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun phrase queries: Use [Hagen et al., WWW 2011]</td>
</tr>
<tr>
<td>Other queries: Segment only Wikipedia titles</td>
</tr>
</tbody>
</table>

**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

Example

Reference

"new york times" "square dance"

(2 segments)

Computed

"new|york|times" square dance

(3 segments)

Query accuracy: 0 (computed ≠ 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

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
</tr>
<tr>
<td>Computed</td>
</tr>
</tbody>
</table>
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 ≠ 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 ≠ 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

<table>
<thead>
<tr>
<th>Reference</th>
<th>&quot;new york times&quot; &quot;square dance&quot; (2 segments)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computed</td>
<td>&quot;new york times&quot; square dance (3 segments)</td>
</tr>
</tbody>
</table>

- 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

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

[Bergsma and Wang, EMNLP-CoNLL 2007]

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)
Rethinking reference selection for larger corpora

Our large-scale corpus

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

Example:

- In the corpus:
  - 8 votes: "new york" "times square" dance
  - 2 votes: "new york times" "square dance"

Computed reference: "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 a small fraction

**Example**

<table>
<thead>
<tr>
<th>In the corpus</th>
<th>Votes</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>&quot;new york&quot; &quot;times square&quot; dance</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>&quot;new york times&quot; &quot;square dance&quot;</td>
</tr>
</tbody>
</table>

**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 a 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 a small fraction

**Example**

In the corpus

<table>
<thead>
<tr>
<th>Votes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>&quot;new york&quot; &quot;times square&quot; dance</td>
</tr>
<tr>
<td>2</td>
<td>&quot;new york times&quot; &quot;square dance&quot;</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>[Hagen et al., WWW 2011]</th>
<th>PMI</th>
<th>Wikipedia titles</th>
<th>hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.481</td>
<td>0.520</td>
<td>0.638</td>
<td><strong>0.644</strong></td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>[Hagen et al., WWW 2011]</th>
<th>PMI</th>
<th>Wikipedia titles</th>
<th>hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.481</td>
<td>0.520</td>
<td>0.638</td>
<td>0.644</td>
</tr>
</tbody>
</table>

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*
Performance à la TREC

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

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Bing</th>
<th>Indri</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun phrases</td>
<td>[WWW 2011]</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>other queries</td>
<td>Wikipedia titles</td>
<td>Wikipedia titles</td>
<td>[WWW 2011]</td>
</tr>
</tbody>
</table>
### Observations

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

---

<table>
<thead>
<tr>
<th></th>
<th>Bing</th>
<th>Indri</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hybrid-B</td>
<td>OPT</td>
</tr>
<tr>
<td>all queries</td>
<td>0.148</td>
<td>0.170</td>
</tr>
<tr>
<td>noun phrase</td>
<td>0.138</td>
<td>0.156</td>
</tr>
<tr>
<td>other</td>
<td>0.162</td>
<td>0.189</td>
</tr>
</tbody>
</table>

---

Still not the top reached
What about efficiency?

Runtime

Memory footprint
Memory and throughput

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

(Ref: 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!
What we have done

Main results

- Human segmentation behavior
  \(\leftrightarrow\) *noun phrase vs. other queries*

- Hybrid segmentation scheme
  \(\leftrightarrow\) *in doubt without*

- Accuracy and TREC evaluation
  \(\leftrightarrow\) *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
  \(\rightarrow\) *noun phrase vs. other queries*
- Hybrid segmentation scheme
  \(\rightarrow\) *in doubt without*
- Accuracy and TREC evaluation
  \(\rightarrow\) *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
  \[\text{noun phrase vs. other queries}\]
- Hybrid segmentation scheme
  \[\text{in doubt without}\]
- Accuracy and TREC evaluation
  \[\text{accuracy \neq retrieval gain}\]

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

Thank you 🌻