The Power of Naïve Query Segmentation

Matthias Hagen Martin Potthast Benno Stein Christoph Bräutigam

The Problem

Web queries are keyword based. `san jose yellow pages`

Queries with quotated segments perform better. `"san jose" "yellow pages"`

Users are often not aware of the quotation option.

↓

Automatic query segmentation

Our Naïve Approach

Get Google n-gram counts for segments

segment s	count(s)
`san jose`	14 495 804
`san jose yellow`	8 822
`san jose yellow pages`	8 739
`jose yellow`	8 831
`jose yellow pages`	8 745
`yellow pages`	41 380 676

→

Score every segmentation S as follows:

$$\text{score}(S) = \sum_{\substack{s \in S,\\ |s| > 1}} |s|^{|s|} \cdot \text{count}(s)$$

segmentation S	score(S)
`"san" "jose" "yellow pages"`	165 522 704
`"san" "jose yellow" "pages"`	35 324
`"san" "jose yellow pages"`	236 115
`"san jose" "yellow" "pages"`	57 983 216
`"san jose" "yellow pages"`	223 505 920
`"san jose yellow" "pages"`	238 194
`"san jose yellow pages"`	8 948 736

←

Choose segmentation with highest score

`"san jose" "yellow pages"`

Evaluation

Annotator	Accuracy Measure	MI	[1]	[2]	[3]	Naïve
A	query	0.274	**0.638**	0.526		0.536
	break	0.693	**0.863**	0.810		0.807
	seg prec	0.469		0.657	0.652	**0.665**
	seg rec	0.534		0.657	0.699	**0.708**
	seg F	0.499		0.657	0.675	**0.686**
B	query	0.244		**0.494**		0.380
	break	0.634		**0.802**		0.752
	seg prec	0.408		0.623	**0.632**	0.519
	seg rec	0.472		0.640	**0.659**	0.626
	seg F	0.438		0.631	**0.645**	0.568
C	query	0.264		**0.494**		0.454
	break	0.666		**0.796**		0.772
	seg prec	0.451		**0.634**	0.614	0.581
	seg rec	0.519		0.642	0.649	**0.653**
	seg F	0.483		**0.638**	0.631	0.615
Agree	query	0.343	**0.717**	0.671		0.627
	break	0.728	**0.892**	0.871		0.851
	seg prec	0.510		0.767	**0.772**	0.718
	seg rec	0.550		0.782	**0.826**	0.778
	seg F	0.530		0.774	**0.746**	0.746

Query corpus (Bergsma and Wang [1]):
- 500 queries from AOL query log
- 3 human Annotators → A, B, C
- 220 queries identically segmented → Agree

Accuracy measures:
- query — ratio of queries that exactly match annotator's segmentation
- break — ratio of decisions between two words that match annotator's decisions
- seg prec — precision of derived segments
- seg rec — recall of derived segments
- seg F — F-Measure

→

Results:
- Naïve always in a 0.1-range compared to best approach
- Naïve segments 3000 queries per second with 12GB RAM

[1] S. Bergsma and Q.I. Wang. Learning noun phrase query segmentation. In *EMNLP-CoNNL 2007*, pages 819-826.
[2] B. Tan and F. Peng. Unsupervised query segmentation using generative language models and Wikipedia. In *WWW 2008*, pages 347-356.
[3] C. Zhang, N. Sun, X. Hu, T. Huang, and T.-S. Chua. Query segmentation based on eigenspace similarity. In *ACL-IJCNLP 2009*, pages 185-188.