Improving Unsupervised Query Segmentation using Parts-of-Speech Sequence Information

Rishiraj Saha Roy, Yogarshi Vyas, Niloy Ganguly (IIT Kharagpur)
Monojit Choudhury (Microsoft Research)

Unsegmented Query Log

Lexicon L_o

Other n-gram

POS tagged Query Log

POS tagged Lexicon L_i

POS tagged Other *n*-gram

Scored POS pattern list

Augmented Lexicon

Other n-gram

Segmented query log

IR-based Evaluation

Word Association Score (WAS) calculation (Mishra et al. WWW'2011)

POS Tagging
*The*_DT *rolling*_VBG *stones*_NNS

- **Penn Tree bank tagset** (36 tags) Stanford log-linear POS tagger [Toutanova et al. 2003]
- **Universal POS tagset** (12 tags) [Petrov et al. 2012] Stanford log-linear POS tagger
- **Unsupervised POS induction** (405 tags) [Biemann, 2006]

POS Transfer

N-gram Computation

- For each POS *n*-gram (or POS pattern) P_j (say, DT-VBG-NNS), we count the number of times P_j appears in L_i, denoted by $count(P_j,i)$
- We define a score for P_j as follows:

$$score(P_j, i+1) = score(P_j, i)\ln(e + \alpha e^{-i/\ln(1+count(Pj,i))})$$

- Here, iteration $i \geq 0$ and α is a tuning parameter. We define $score(P_j, 0) = 1$.

Score augmentation

- The WAS for every unique n-gram $\mathbf{w}$ is augmented as:

$$score(\mathbf{w}, i+1) = score(\mathbf{w}, i) \times score(P_{POS(\mathbf{w})}, i)$$

- $POS(\mathbf{w})$ = index j of the underlying POS pattern of $\mathbf{w}$. $score(\mathbf{w}, 0) = WAS(\mathbf{w})$
- L_{i+1} is constructed by including all $\mathbf{w}$ for which $score(\mathbf{w}, i) \geq \delta$ (user defined threshold).

Iterate till convergence ($L_{i+1} \equiv L_i$)

Segment queries using the final lexicon (Mishra et al. WWW'2011)

IR Performance of the segmented queries (Mishra et al. WWW'2011)

Metric	Orig	PTB	UTS	Bie-S
nDCG@5	0.743	0.751*	0.751*	0.751*
nDCG@10	0.747	0.753	0.752	0.752
MAP	0.901	0.905	0.905	0.905
MRR	0.587	0.601	0.598	0.602

Query segmentation is the process of breaking down Web search queries into their constituent structural units, thereby aiding the retrieval process. This study connects two orthogonal approaches to segmentation or chunking of text fragments – those that rely on purely statistical word association measures and those that try to incorporate linguistic information, used commonly for Natural Language chunking.

Our initial experiments show that POS tagging does improve query segmentation. Although we do not observe any performance benefits of using a specific POS tagset or tagging approach over others, we do observe that tagset and taggers designed for POS tagging English text help improve the segmentation of a complementary set of queries than the ones which are benefitted by unsupervised POS induction. Thus, appropriately combining these two approaches is expected to lead to better segmentation.

Example Clusters from Bie-S

Cluster 1: bake, casserole, dessert, fry, meatballs,
Cluster 2: athletics, baseball, cycling, football,
Cluster 3: army, citizenship, customs, defence,
Cluster 4: battlefield, diablo, godfather, hitman,

Example Segmentations

Tagset	Segmented query
Orig	picture | in | picture | lcd tv (0.61)
PTB/UTS	picture in picture | lcd tv (0.79)
Bie-S	picture in | picture | lcd tv (0.75)
Orig	samsung i900|omnia|free |games (0.69)
All tags	samsung i900 |omnia |free games (0.81)
Orig	richard burns rally | pc | cheats (0.67)
All tags	richard burns | rally | pc cheats (0.75)

Number of Winning and losing queries

Tagset	Gain	Same	Loss
PTB	67 (+0.048)	150	33 (-0.060)
UTS	57 (+0.055)	162	32 (-0.068)
Bie-S	67 (+0.042)	140	43 (-0.050)

Contact: *rishiraj@cse.iitkgp.ernet.in, monojitc@microsoft.com*