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 tag set 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.

![Diagram of query segmentation process]

Unsupervised POS induction is constructed by including all POS tagged patterns in the lexicon $\mathcal{L}_o$. For each POS n-gram or POS pattern $P$, we count the number of times $P$ appears in $\mathcal{L}_o$, denoted by $\text{count}(P, i)$. We define a score for $P$ as follows:

$$\text{score}(P, i + 1) = \text{score}(P, i) \times \text{count}(P, i)$$

Here, iteration $i \geq 0$ and $\alpha$ is a tuning parameter. We define $\text{score}(P, 0) = 1$.

The WAS for every unique n-gram $w$ is augmented as:

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

POS tagging is aiding common segmentation. We also construct the lexicon $\mathcal{L}_{i+1}$ as follows:

$$\text{POS}(\mathcal{L}_{i+1}) = \text{index } j \text{ of the underlying POS pattern of } w. \text{ score}(w, 0) = \text{WAS}(w) \text{ of } \mathcal{L}_{i+1}$$

Iterate till convergence ($\mathcal{L}_{i+1} = \mathcal{L}_i$)

Example Segments:

- Tagset: Segmented query
  - Orig: picture | in | picture | lctv (0.61)
  - PTB/UTS: picture in | picture | lctv (0.79)
  - Bie-S: picture in | picture | lctv (0.75)

- All tags: samsung | omnia | free | games (0.69)
- Orig: richard burns rally | pc | cheats (0.67)
- All tags: richard burns | rally | pc cheats (0.75)

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