Are Web Search Queries Evolving into a Language of their own?

Rishiraj Saha Roy and Niloy Ganguly (IIT Kharagpur)
Monojit Choudhury (Microsoft Research India)

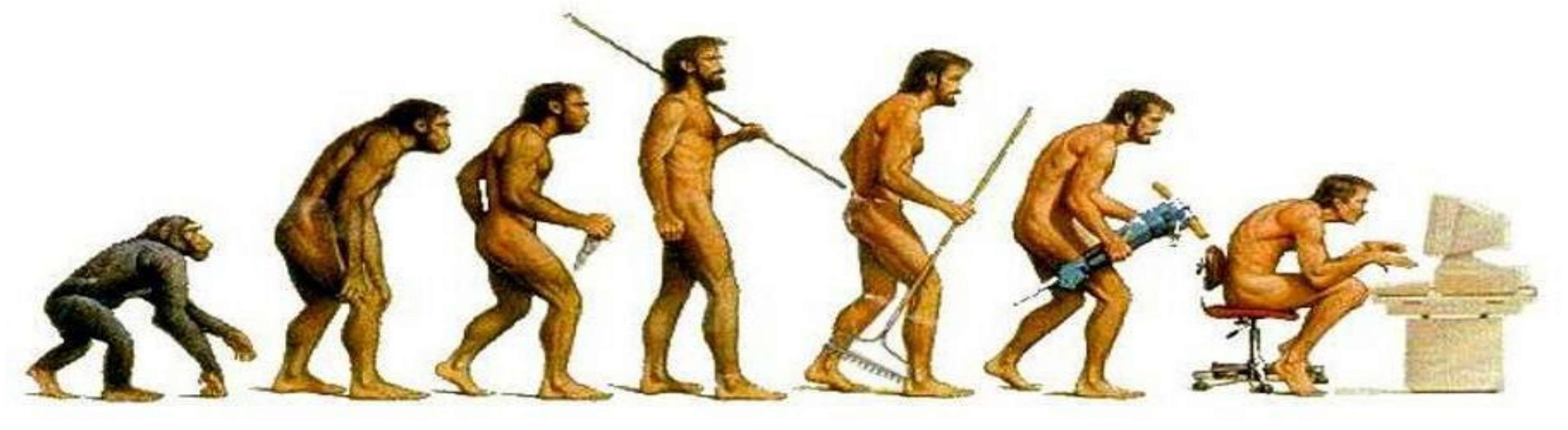


Image source: http://shishikwena.blogspot.in/

Introduction

Web users communicate their information need to a search engine through queries. Queries have a structure far simpler than NL, but more complex than the commonly assumed bag-of-words model. In fact, Web search queries define a new and fast evolving language of its own, whose dynamics is governed by the behavior of the search engine towards the user and that of the user towards the engine.

***Are Web Search Queries* a self-organizing system of language?**

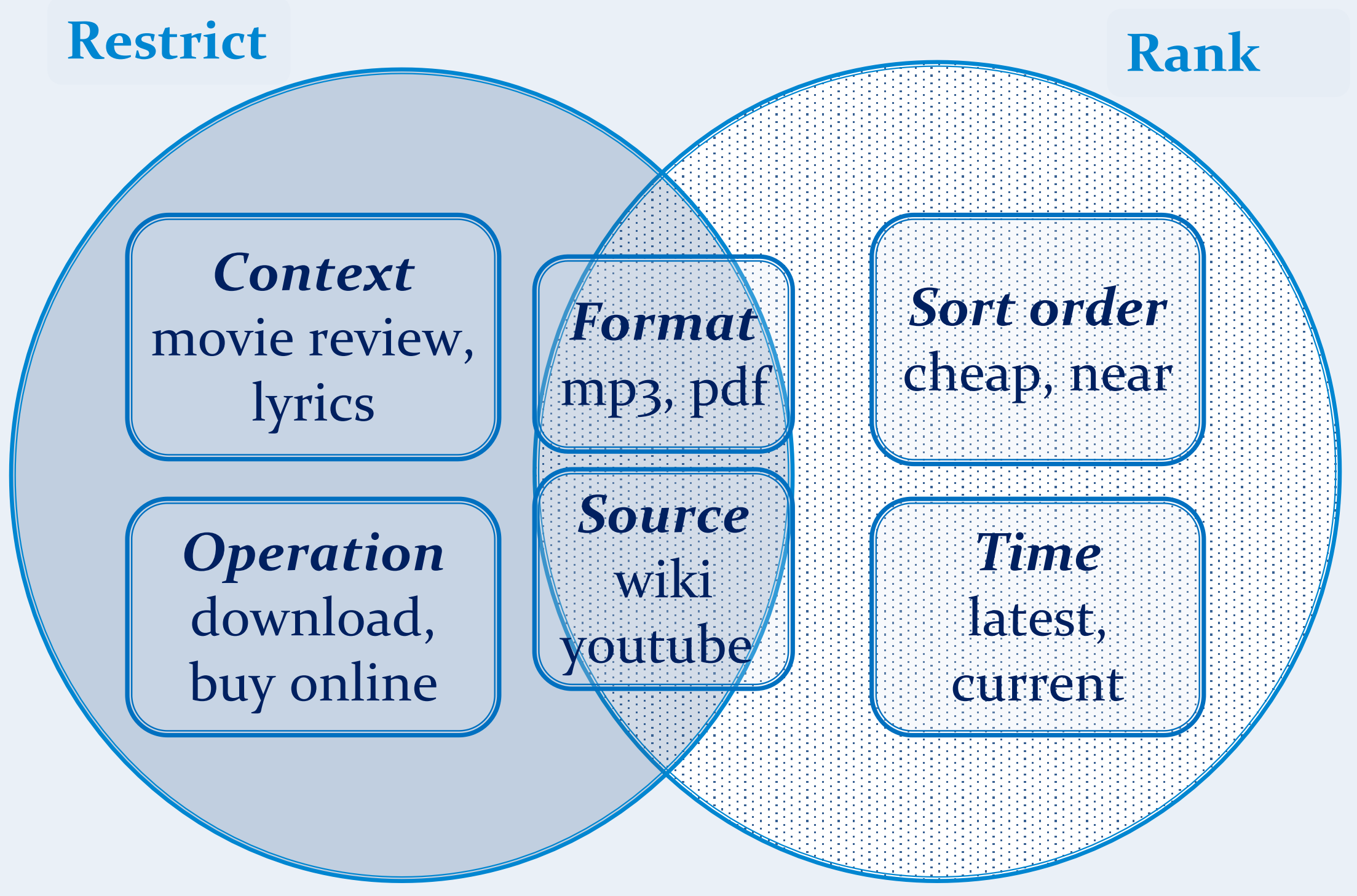
Identifying Structural Units

ORIGINAL LOG	SEGMENTED LOG
larry the lawnmower tv show	*larry the lawnmower | tv show*
our lady of lourdes seven hills	*our lady of lourdes | seven hills*
grand theft auto 3 ps2 cheats	*grand theft auto 3 | ps2 cheats*
lyrics for my name is	*lyrics for | my name is*
never miss a beat mp3 download	*never miss a beat | mp3 download*
room to let perth wa	*room to let | perth wa*
as time goes by sheet music	*as time goes by | sheet music*
villeroy and boch kitchen sinks	*villeroy and boch | kitchen sinks*
we are the people song lyrics	*we are the people | song lyrics*
worlds best chocolate chip cookies	*worlds best | chocolate chip cookies*
another way to die piano sheet	*another way to die | piano sheet*
vanilla ice cream in french	*vanilla ice cream | in french*
piano chords suddenly i see	*piano chords | suddenly i see*
paris by night sydney 2009	*paris by night | sydney 2009*
kiss the rain piano sheet	*kiss the rain | piano sheet*
lyrics for when im gone	*lyrics for | when im gone*
play free solitaire card game	*play free | solitaire card game*
eb games forest hill chase	*eb games | forest hill chase*

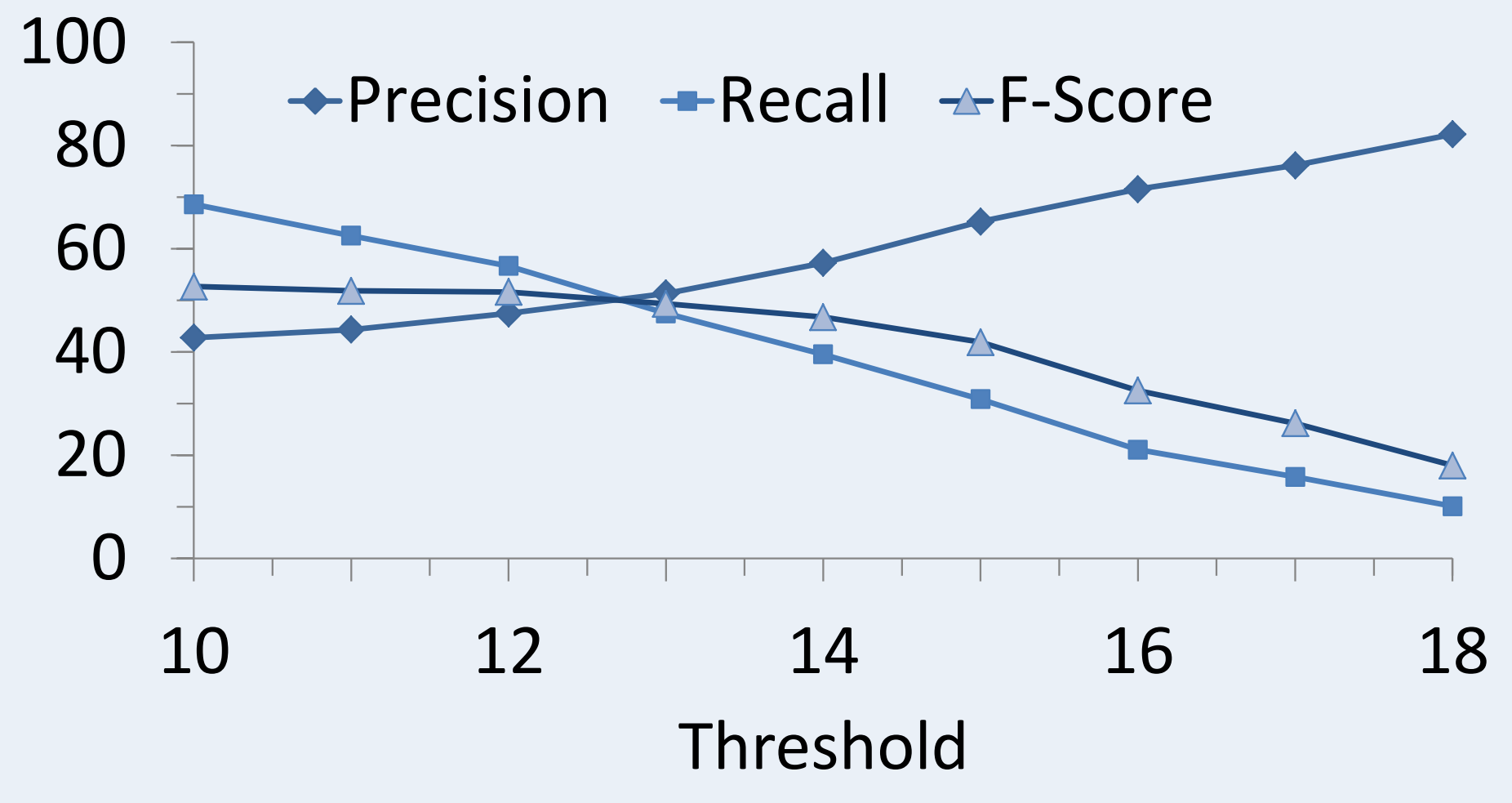
- $\boldsymbol{N}$: Number of queries in log containing n-gram $\boldsymbol{M}$
- $\boldsymbol{k}$: Queries containing words of $\boldsymbol{M}$ in any order
- $\boldsymbol{E}$: Expected count of $\boldsymbol{M}$ under a bag-of-words *NULL*
- If $\mathbf{P}[\boldsymbol{E} \geq \boldsymbol{N}] < \boldsymbol{exp}(-\frac{2(\boldsymbol{N}-\boldsymbol{E})^2}{\boldsymbol{k}})$, bounded by a threshold, holds, then $\boldsymbol{M}$ is declared a *query segment*
- Use PMI-based scheme with Wikipedia titles for named entities
- Use dynamic programming to search over all possible segmentations

Discovering Syntactic Categories

- Queries contain two types of segments – content and intent
- Analogous to content and function words of NL
- Study distributional statistics of query segments – frequency, co-occurrence counts and entropy, clustering coefficient
- Annotate segments manually
- Co-occurrence entropy found to be the best discriminator
- Intent units can be classified



A taxonomy of intent segments in Web search queries.

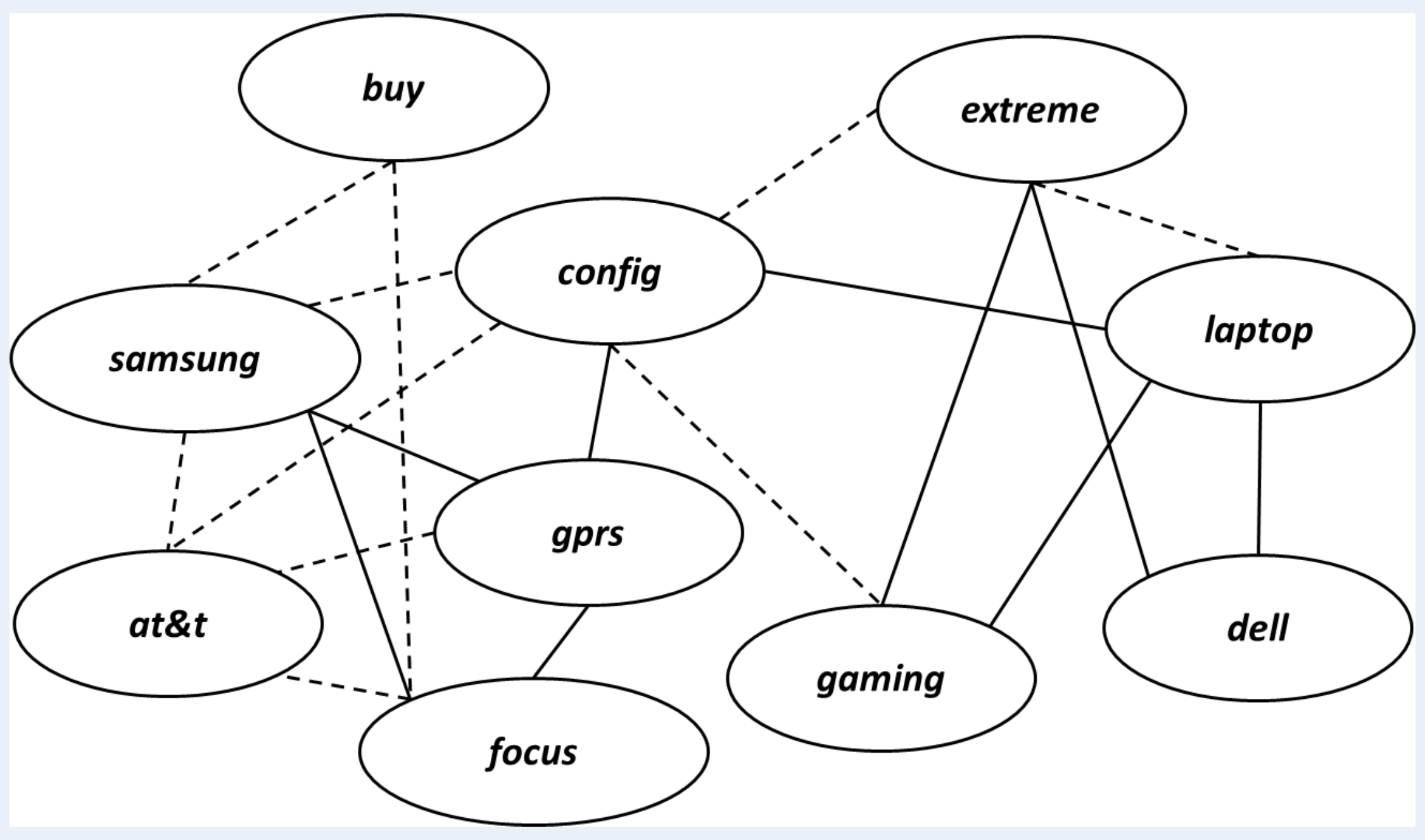


Intent segment detection against human annotations (reverse for content)

Applications

- Sponsored search
- Query suggestions
- Intent diversification

Applying Complex Network Modeling



A word co-occurrence network (WCN) built from a toy query log.

Natural language	Queries
Kernel and periphery	Kernel and periphery
Content and function units both in kernel and periphery	Content and intent units both in kernel and periphery
Kernel – 1000 units	Kernel – 500 units
Periphery – 84000 units	Periphery – 1200000 units
Small world effect	No small world effect
Sentences formed by units from kernel and periphery, or only kernel	Queries mostly formed by units from kernel and periphery, or only periphery
Intra-kernel edges dominate	Kernel-periphery edges dominate
Kernel more tightly coupled	Kernel less tightly coupled

Publications

1. Rishiraj Saha Roy, Niloy Ganguly, Monojit Choudhury and Srivatsan Laxman, "An IR-based Evaluation Framework for Web Search Query Segmentation", in SIGIR '12, pp. 881 – 890.
2. Rishiraj Saha Roy, Monojit Choudhury and Kalika Bali, "Are Web Search Queries an Evolving Protolanguage?", in Evolang IX, pp. 304 – 311. [BEST RESEARCH POSTER AWARD]
3. Rishiraj Saha Roy, Niloy Ganguly, Monojit Choudhury and Naveen Kumar Singh, "Complex Network Analysis Reveals Kernel-Periphery Structure in Web Search Queries", in QRU '11, SIGIR Workshop, pp. 5 – 8.
4. Nikita Mishra, Rishiraj Saha Roy, Niloy Ganguly, Srivatsan Laxman and Monojit Choudhury, "Unsupervised Query Segmentation Using only Query Logs", in WWW '11 Posters, pp. 91 – 92.

This work has applications in query understanding. We should also use this well-preserved data for studying language evolution.

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