A Probabilistic Concept Annotation for IT Service Desk Tickets

Ea-Ee Jan[*], *Kuan-Yu Chen*[*†], *and Tsuyoshi Ide*[*]

[*]IBM Service Delivery and Risk Analytics & [†]National Taiwan University

Motivations

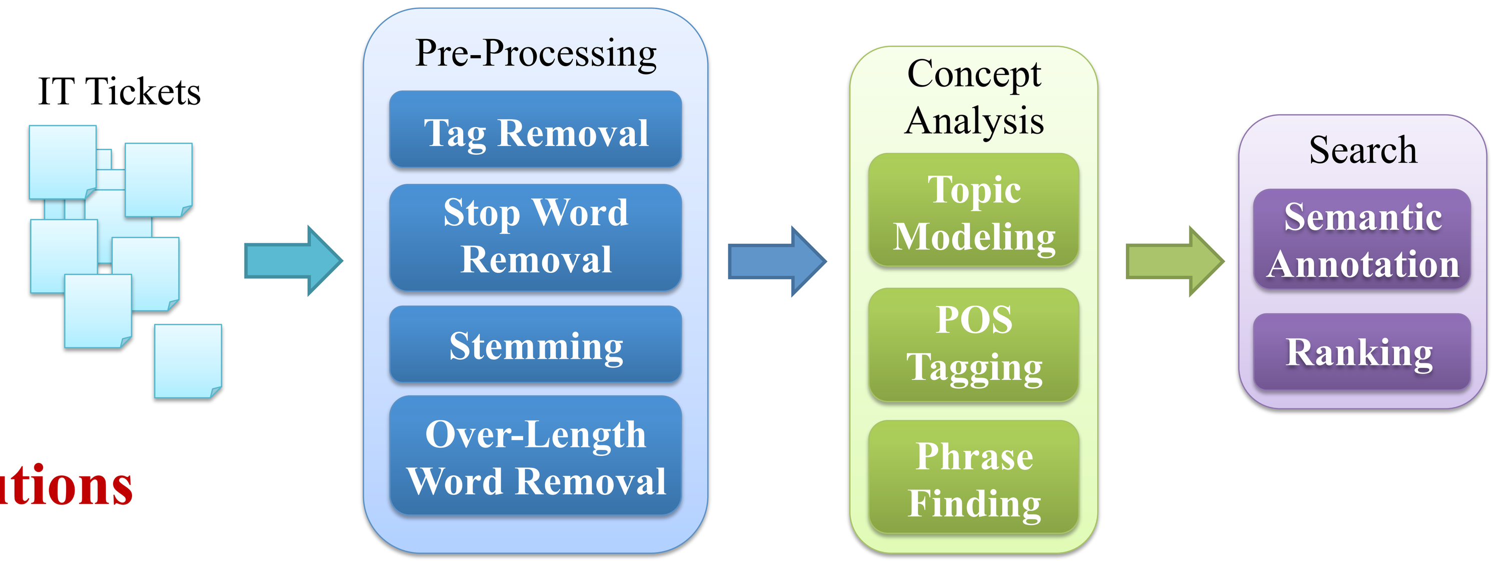
➢IT Service desk is a tens million dollars business for an enterprise

➢Millions of IT service desk tickets are created yearly to address business users' IT related problems

- password reset
- firewall not working
- how to setup mail box

➢It is critical to know what key IT problems have been dealt with

- what are the **pain points**
- what are the pain points **distributions**



The Proposed Framework

➢**Text Normalization Pre-processing**

☐ Text pre-processor for handling noisy text from IT service desk tickets

- Xml tag, stop words removal, stemming, punctuations and abbreviation normalization
- Word length feature is used to remove email, http link and other functionless words

➢**Concept Analysis**

☐ Topic models for anatomizing normalized tickets

- Topic modeling can be used to dissect word usage cues in each document
- Assuming a latent topic conveys ideas which are common to a subset of the input data
- Beyond bag of words

☐ Represent topics by **readable descriptions (phrases)** instead of word distributions given by the topic model to better visualize topics

- Phrases are composed by *n*-gram tokens, filtering by predefined Part of Speech patterns
- The most suitable phrase to represent a given topic is determined by

$$P(Phrase_i|T_k)=\frac{P(T_k\mid Phrase_i)P(Phrase_i)}{P(T_k)}\approx \underbrace{P(T_k\mid Phrase_i)}_{\text{The relevance degree between the pair of topic and phrase}}\underbrace{P(Phrase_i)}_{\text{The naturalness of the phrase in a language}}$$

➢**Search**

☐ IT ticket search needs to address both the **precision** and **confidence** score of each document assigned to a topic. It is different from traditional IR approach,

- This is due to high penalty of human cost incurred by search errors

$$S(D_j,T_k)=\frac{P(D_j\mid T_k)}{\sum_{k'=1}^{K}P(D_j\mid T_{k'})}\approx\frac{P(D_j\mid Phrase_k)}{\sum_{k'=1}^{K}P(D_j\mid Phrase_{k'})}$$

$$P(D_j\mid Phrase_k)=\prod_{l=1}^{|D_j|}P(w_l\mid Phrase_k)=\prod_{l=1}^{|D_j|}\Big[\underbrace{\alpha\cdot P_{\mathrm{U}}(w_l\mid Phrase_k)}_{\text{Literal Term Matching}}+\underbrace{\beta\cdot P_{\mathrm{T}}(w_l\mid Phrase_k)}_{\text{Concept Matching}}+\underbrace{\gamma\cdot P_{\mathrm{BG}}(w_l)}_{\text{Avoid Zero Probability}}\Big]$$

Literal Term Matching
password reset vs. reset password

Concept Matching
buy notebook vs. purchase NB

Avoid Zero Probability

Experimental Results

➢Dataset: the applications portals (AP) and the mailbox problems (MB)

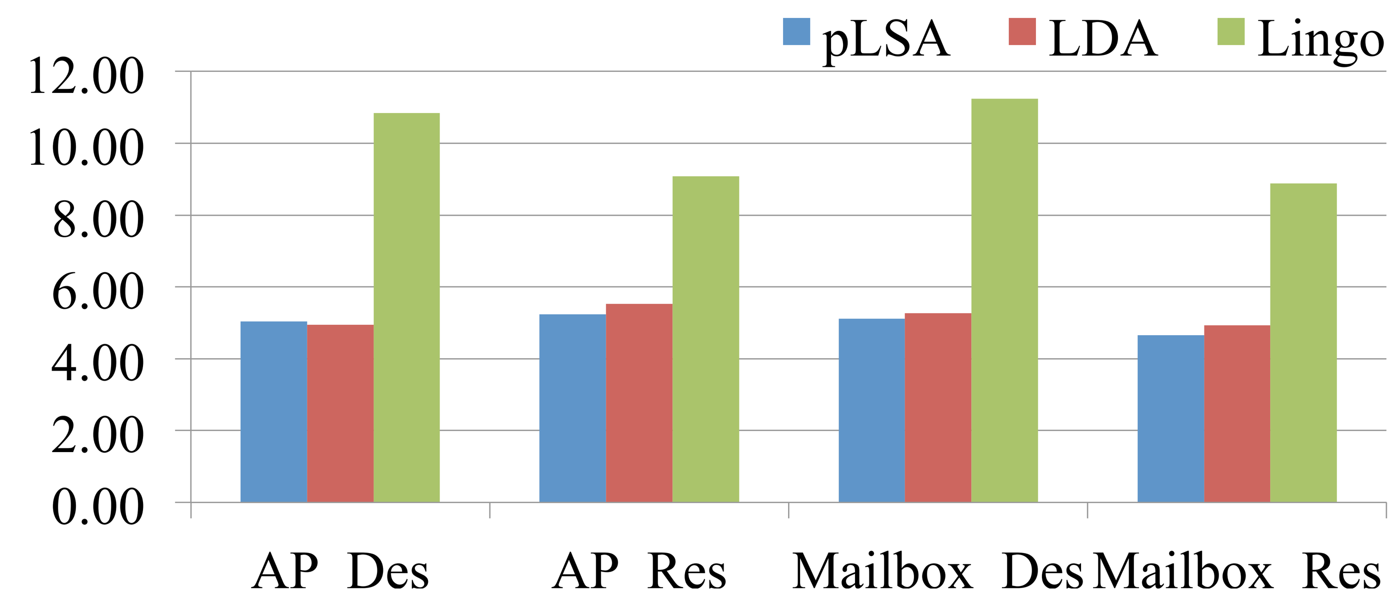
- AP related to many applications, cover boarder spectrum
- MB problems are more specific

➢We proposed to judge the work by employing Dunn index (DI) and Davies-Bouldin index (DBI) due to lack of hand labeled references for assessments.

- Larger DI/Smaller DBI indicate better cluster integrities

➢The DBI index indicates that the proposed framework consistently outperform Lingo

$$\mathrm{DBI}=\frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\frac{\delta_i+\delta_j}{\left\|\mu_i-\mu_j\right\|_2},$$



$$\mathrm{DI}=\frac{\min\limits_{1\le i,j\le K,\, i\neq j}\left\|\mu_i-\mu_j\right\|_2}{\max\limits_{1\le k\le K,\, s_m^k,s_n^k\in C_k}\left\|s_m^k-s_n^k\right\|_2},$$

$\left\|\mu_i-\mu_j\right\|_2$ is pariwise centroid distance,

δ_i is average distance of all elmenets in a cluster

$\left\|s_m^k-s_n^k\right\|_2$ is pairwise element distance in a cluster

➢The DI shows that the proposed framework outperform Lingo in both AP and MB datasets