Target Detection & Knowledge Learning for Domain-restricted Question Answering

Mengdi Zhang, Tao Huang, Yixin Cao, Lei Hou

Knowledge Engineering Group, Dept. of Computer Science and Technology, Tsinghua University

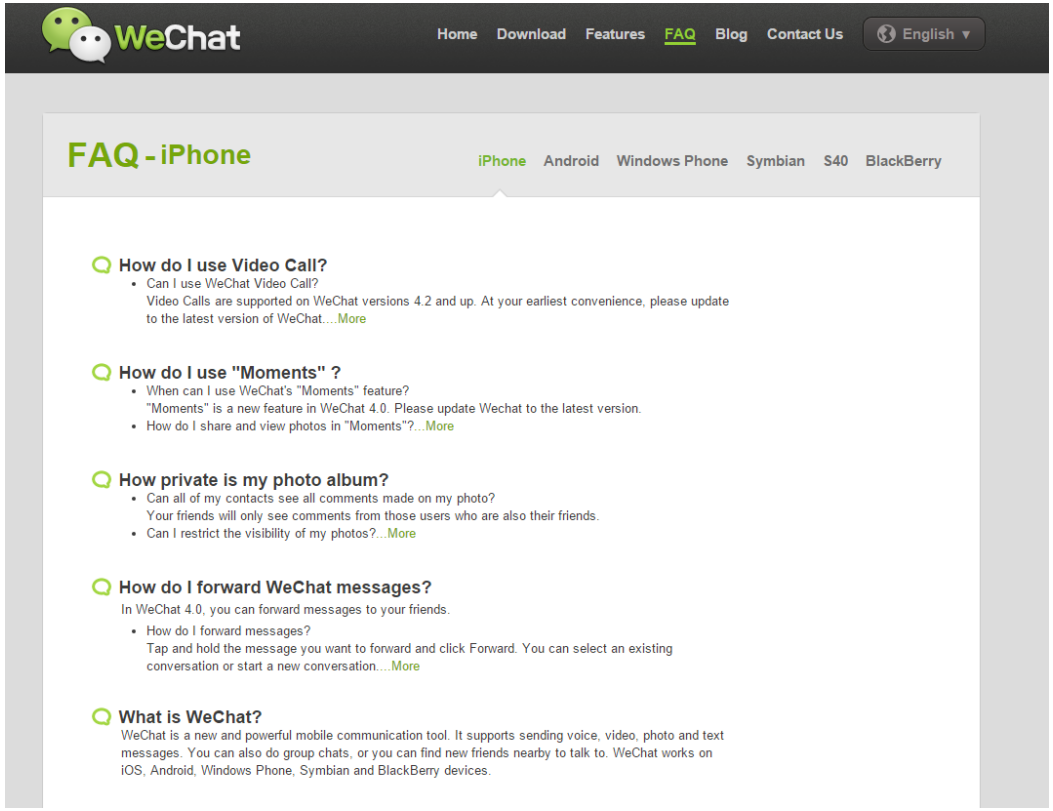
Outline

- Motivation & Challenges
- Approach
- Experiment
- Conclusion

Motivation

■Frequently Asked Question(FAQ)

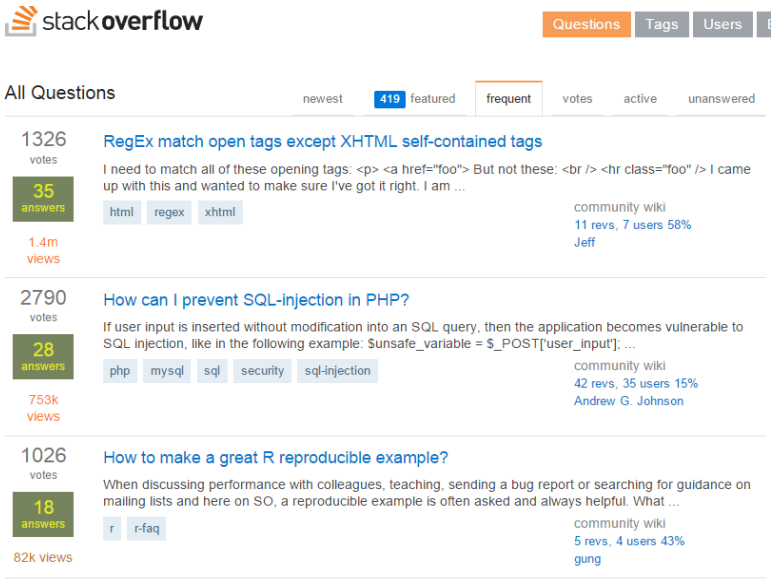
FAQs, are listed questions and answers, all supposed to be commonly asked in some context, and pertaining to a particular topic. (definition from Wikipedia)



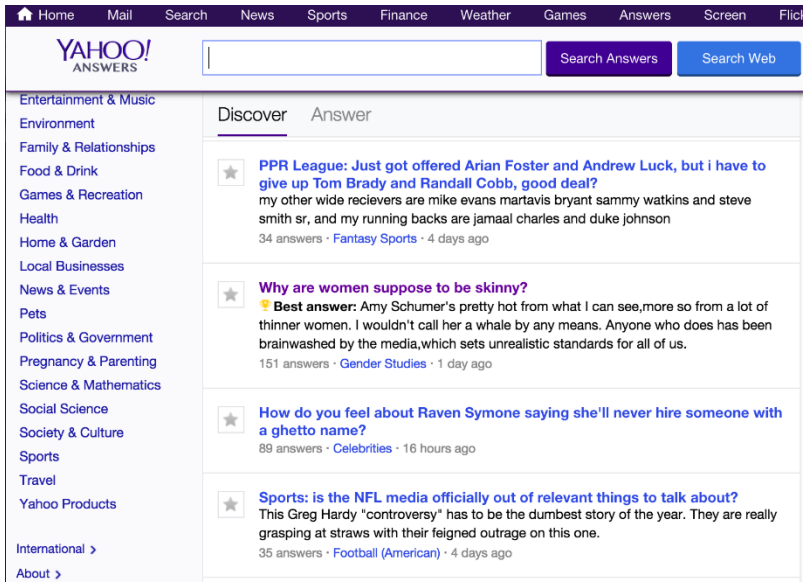
Company FAQ page

Motivation

■Question Answering Community



Stack Overflaw



Yahoo! Answers

Motivation

(StackExchange)

Choose a Site

Data updated Sep 20 at 6:09

Stack Overflow Q&A for programmers	10m questions	17m answers	42m comments	42k tags	visit site →
Mathematics Q&A for people studying math at any level and professionals in related fields	489k questions	713k answers	2.2m comments	1.2k tags	visit site →
Super User Q&A for computer enthusiasts and power users	282k questions	436k answers	971k comments	5.1k tags	visit site →
Server Fault Q&A for system administrators and IT professionals	205k questions	361k answers	679k comments	3.4k tags	visit site →
Ask Ubuntu Q&A for Ubuntu users and developers	202k questions	267k answers	631k comments	2.9k tags	visit site →
TeX - LaTeX Q&A for users of TeX, LaTeX, ConTeXt, and related typesetting system	96k questions	129k answers	512k comments	1.3k tags	visit site →
Meta Stack Exchange Q&A about the Stack Exchange Network	74k questions	115k answers	580k comments	1.3k tags	visit site →
Unix and Linux Q&A for users of Linux, FreeBSD and other Un*x-like operating systems.	73k questions	115k answers	302k comments	2.1k tags	visit site →
Stack Overflow на русском Q&A for программистов	66k questions	94k answers	259k comments	3.1k tags	visit site →
Statistical Analysis Q&A for statisticians, data analysts, data miners and data visualization experts	64k questions	66k answers	252k comments	1.2k tags	visit site →

Motivation

- Data gets bigger. (10M+)
- Question gets more domain specific. (Products, Coding, etc)

We need an automatic domain-restricted FAQ answering framework!

Motivation

- The Domain Restricted FAQ Answering Task
  - Definition 1: **QA Pair**
    - Pair of <Question, Answer>, provided by expert, listed in FAQ set .
  - Definition 2: **User's query**
    - The question asked by user
  - Problem 1: **FAQ Retrieval**
    - Given a user query, return the best matched answer from the FAQ set.

Not an easy task

Challenges

Question Understanding

- Casual Forms:
  - Short
    - 网银密码忘了
  - Informal
    - 朋友在国外，我可以打电话到银行给他汇钱吗？(from user)
    - 电话银行对外转账是否支持外币？(from expert)
- Diverse Expressions:
  - Vocabulary Gap between user and expert
    - 花钱吗 vs 收费
  - Expression variation among users
    - 花钱吗，要钱吗，有多贵
- Special Domain Restriction:
  - Domain Knowledge: unavailable and expensive
    - **95566**电话银行
    - **U**盾密码

Challenges

1. How to understand the intension of user?
   - 朋友在国外，我可以打电话到银行给他汇钱吗？
   - 手机银行需要花钱吗

2. How to conquer the domain knowledge?
   - 95566电话银行
   - U盾密码

Challenges

- Lexical Form:
  - Short : ambiguous
  - Informal : hard to parse
- Content: Vocabulary Gap between User and Expert
  - 花钱 vs 收费
  - Expression Variation among Users
    - 花钱，要钱，有多贵

Parsing

- Special Domain Restriction:
  - Domain Knowledge: unavailable and expensive
    - 95566电话银行
    - U盾密码

Knowledge Base

Challenges

- Parser works badly in short & informal text.
- KB is too expensive to construct.

- Prefer a cheap but effective solution.
- Let's jump into the data.

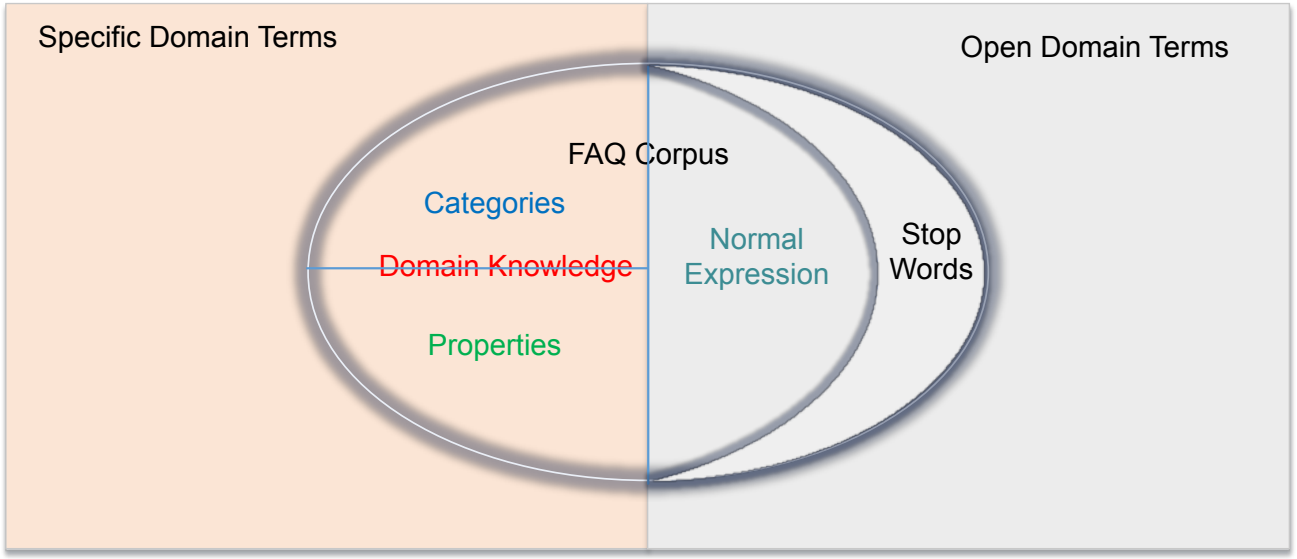
Observation

- The FAQ set:
  - 通过个人网上银行的跨行快汇功能可否向农业银行汇款？
  - 通过个人网上银行的跨行快汇功能可否向平安银行汇款？
  - 个人网上银行中“购买基金”在什么位置？
  - 如何通过个人网上银行查询工银e支付的转账明细？

- One step closer:
  - 通过个人网上银行的跨行快汇功能可否向农业银行汇款？
  - 通过个人网上银行的跨行快汇功能可否向平安银行汇款？
  - 个人网上银行中“购买基金”在什么位置？
  - 如何通过个人网上银行查询工银e支付的转账明细？

Observation

Vocabulary Space of Specific Domain(e.g. Bank)



Example: 如何通过个人网上银行查询工银e支付的转账明细?

Problem Definition

Definition 2 ***Target-word.*** *We define the target-word* w^t *as a word which can stand for the main meanings of the question, i.e., user's intention. There are usually more than one target-word in a question. We represent a question* $Q_i = \{w_1, w_2, ..w_m\}$ *in a ranked list of target-words* $Q_i^t = \{w_1^t, w_2^t, .., w_k^t\}$*, where* $k \leq m$*.*

Example:

通过个人网上银行的跨行快汇功能可否向农业银行汇款？

Problem Definition

Definition 3 ***Domain Knowledge.*** *We define a domain knowledge structure embedded in the questions: service category and its properties, as* $< C, P_1, P_2, ..P_c >$. *A question can be categorized by this two-layer labels. Domain knowledge* $K = \{< C_i, P_{i1}, P_{i2}, ..P_{ic} > | i = 1, 2, ...d\}$ *is defined for FAQ corpus* S, *where* d *means the number of services included in* S.

Example:

Table 1. A snippet domain knowledge of banking

Category	Properties
phone-bank	query open-account register close-account
电话银行	查询 开户 注册 销户
text-bank	password binding query
短信银行	密码 绑定 查询
e-pay	query close-account register remittance
e支付	查询 注销 注册 汇款
noble-metal	sale price buying specification
贵金属	销售 零售价 购买 规格
fund	investment custody subscribing purchase
基金	定投 托管 认购 申购

Problem Definition

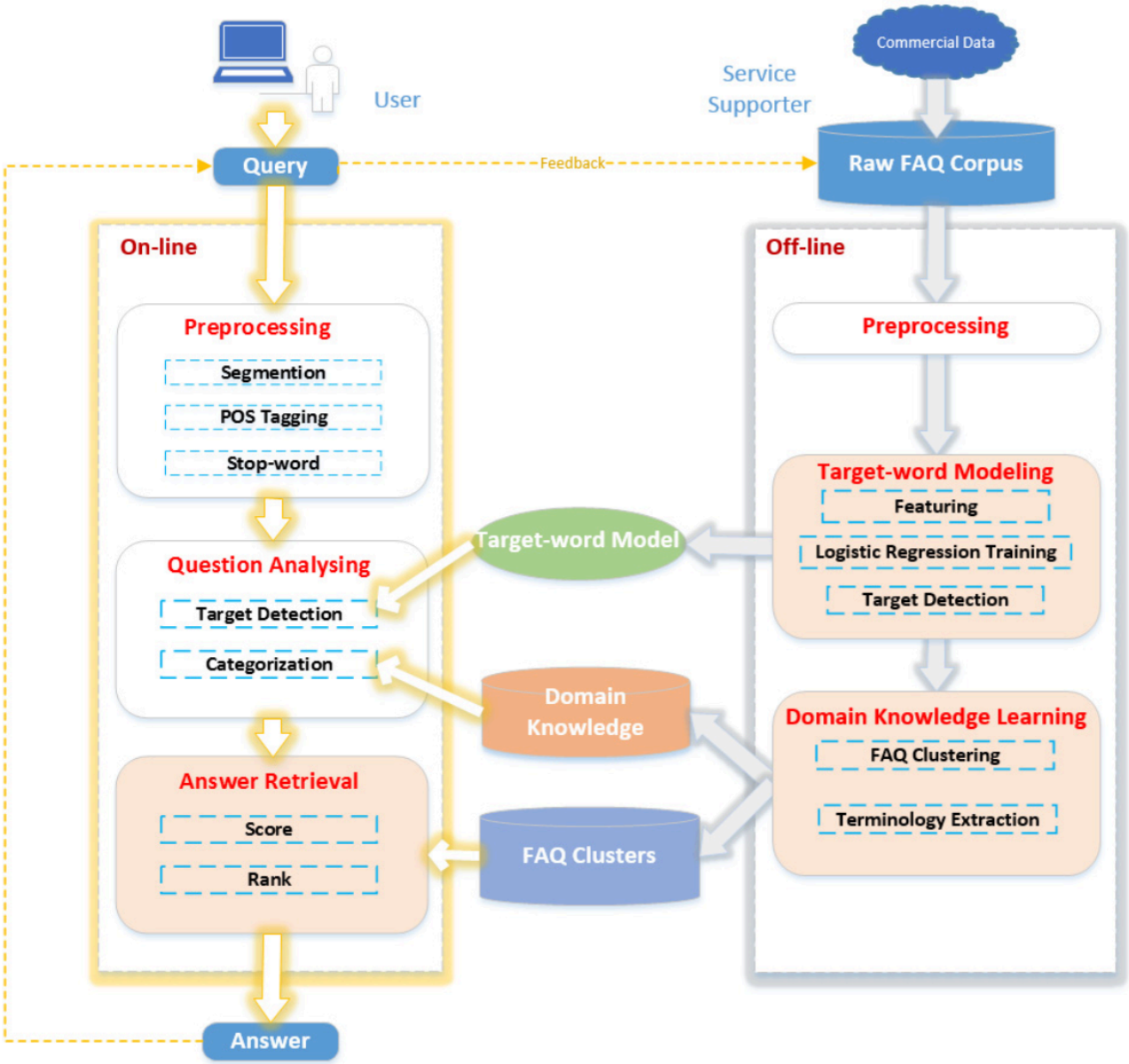
Problem 1 ***Data-driven FAQ Answering.*** *Given a FAQ corpus* $S = \{< Q_i, A_i > | i = 1, 2.., n\}$ *and a user's query* q*, the goal is to firstly detect the target-word and learn the domain knowledge* K *from* S*, and finally find a list of QA Pair* $p \in S$ *for* q*, ranked by a function* $Score(q, p)$ *which measures the similarity between* p *and* q *based on the target-words and domain knowledge obtained previously.*

Main Tasks

- Task 1. **Target-word Detection**
  - Logistic Regression Classification
- Task 2. **Domain Knowledge Learning**
  - FAQ Clustering
  - Terminology Extraction
- Task 3. **Answer Retrieval**
  - Query Classification
  - Target-word Based BM25 Ranking

Framework

Semi-automated Domain-restricted FAQ Answering Framework (SDFA)


User
Service Supporter
Commercial Data
Query
Feedback
Raw FAQ Corpus
On-line
Off-line
Preprocessing
Segmention
POS Tagging
Stop-word
Preprocessing
Target-word Modeling
Featuring
Logistic Regression Training
Target Detection
Target-word Model
Question Analysing
Target Detection
Categorization
Domain Knowledge
Domain Knowledge Learning
FAQ Clustering
Terminology Extraction
Answer Retrieval
Score
Rank
FAQ Clusters
Answer

Main Tasks

- Task 1. **Target-word Detection**
  - Logistic Regression Classification
- Task 2. **Domain Knowledge Learning**
  - FAQ Clustering
  - Terminology Extraction
- Task 3. **Answer Retrieval**
  - Query Classification
  - Target-word Based BM25 Ranking

Target-word Detection

- Supervised Classification
  - Logistic Regression
  - Word Features

Ask three student to label the Target-word training set.

Lexical Features	Semantic Features
1. BOW 2. Position 3. Length 4. Term Frequency in Corpus	1. POS tag of Wi-1 2. POS tag of Wi 3. POS tag of Wi+1

Main Tasks

- Task 1. **Target-word Detection**
  - Logistic Regression Classification
- Task 2. **Domain Knowledge Learning**
  - FAQ Clustering
  - Terminology Extraction
- Task 3. **Answer Retrieval**
  - Query Classification
  - Target-word Based BM25 Ranking

Domain Knowledge Learning

- Step 1: Cluster the FAQ set
  - partition the corpus into categories
    - DBSCAN
- Step 2: Extract the terminology
  - distill the two-layer terminology structure of a service
    - Top-1 target-word of the cluster => **category**
    - Top-N target-words of the cluster => **properties**
    - Manually check again, guided by the target-words

Main Tasks

- Task 1. **Target-word Detection**
  - Logistic Regression Classification
- Task 2. **Domain Knowledge Learning**
  - FAQ Clustering
  - Terminology Extraction
- Task 3. **Answer Retrieval**
  - Query Classification
  - Target-word Based BM25 Ranking

Answer Retrieval

- Step1: Categorize user's query based on domain knowledge
  - the relevant documents should have the same categories and properties
- Step2: Find the related QA Pair by the target-word based BM25 algorithm.
  - overlap target-word will be rewarded
  - different target-word will be punished

$$P(rel|q,p) \propto \sum_{q,tf} \lambda_i * w_i(tf)$$

Experiment

- Data
  - Two Chinese Bank FAQ Corpus

Table 2. Statistics on Datasets.

	#QA Pairs	#Extended Questions	#Test Set	#Target-word Train Set
Bank1	48,495	127,026	4,336	2,272
Bank2	2,399	42,404	5,536	500

Experiment

- Data
  - Two Chinese Bank FAQ Corpus

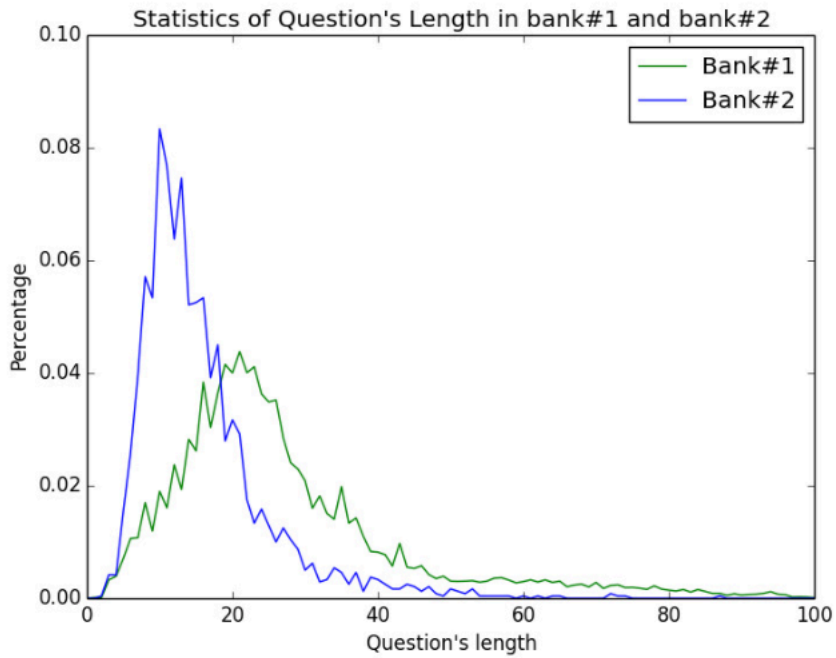


Fig. 2. The Distribution of Question's Length.

Experiment

- Evaluation Measures
  - Precision @1
  - Precision @5
  - Mean Reciprocal Rank(MRR)

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

Experiment

- Experiments Design
  - Baseline:
    - Cosine
    - BM25
  - Target-word Model Testing:
    - BM25 vs BM25t
  - Domain Knowledge Validation:
    - BM25t vs BM25t+Class
  - Whole Framework:
    - BM25t+Class+Punish

Experiment

- Result

Table 3. Overall results

Method	Bank1			Bank2		
	Precision@1	*Precision@5*	*MRR*	*Precision@1*	*Precision@5*	*MRR*
Cosine	41.3%	64.5%	55.7%	45.4%	68.1%	57.1%
BM25	61.1%	79.4%	68.2%	62.8%	84.3%	70.3%
$BM25^t$	63.6%	81.7%	70.0%	64.2%	87.0%	73.9%
$BM25^t$ +Class	63.5%	81.3%	69.8%	64.1%	86.7%	73.6%
$BM25^t$+Class+Punish	66.6%	84.1%	73.9%	65.3%	88.2%	74.6%

Conclusion

- A semi-automatic FAQ answering framework: SDFA
  - Score the target-word to detect user's intention
  - Cluster the FAQ corpus to learn a light-weight domain -knowledge structure
  - Rank QA pairs by target-word based BM25
- Data-driven fashion
  - Sometime, the data itself carries abundant knowledge, good enough for the end-to-end application.

Thank you. Question?