ICT-Crossn: The System of Cross-lingual Information Retrieval of ICT in NTCIR-7

Weihua Luo, Tian Xia, Ji Guo, Qun Liu

Multilingual Interaction Technology Laboratory (MITEL)
Institute of Computing Technology, Chinese Academy of Sciences

presented in NTCIR-7

2008-12-18

Outline

- Background
- The system architecture
- Query translation
- Document retrieval
- Experiments on the dry-run set
- Results on the formal run
- Conclusion

Outline

- Background
- The system architecture
- Query translation
- Document retrieval
- Experiments on the dry-run set
- Results on the formal run
- Conclusion

Background

- Tasks in NTCIR-7
  - Advanced Cross-lingual Information Access (ACLIA)
    - Information Retrieval for Question Answering (IR4QA)
      - A novel task in ACLIA task cluster in NTCIR-7
      - Motivation
        - CLIR has made a great progress
        - Find out which IR technique would help CCLQA
      - Subtasks
        - CS-CS
        - EN-CS √
        - CT-CT √
        - EN-CT
        - JA-JA
        - EN-JA

Difference between CLIR and IR4QA

INSTITUTE OF COMPUTING TECHNOLOGY

- Query

CLIR

```
<TOPIC>
<NUM>001</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>Time Warner, American Online (AOL), Merger, Impact</TITLE>
<DESC>Find reports about the impact of AOL/Time Warner merger.</DESC>
<NARR>
<BACK>Time Warner and American Online (AOL) announced a merger on January 10th, 2000. The market value was estimated at $US350 billion making it the biggest merger in the US.</BACK>
<REL>Comments on AOL/Time Warner merger's effects on Internet and entertainment media businesses are relevant. Descriptions of the development of the AOL/Time Warner merger are partially relevant. Information about the total amount and the transformation of ownership structure are irrelevant.</REL>
</NARR>
<CONC>Time Warner, American Online, AOL, Gerald Levin, merger, M&A, Merger and Acquisition, media, entertainment business</CONC>
</TOPIC>
```

IR4QA

```
<TOPIC ID="ACLIA1-CS-T41">
<QUESTION LANG="EN">
<![CDATA[ List the hazards of global warming.]]>
</QUESTION>
<QUESTION LANG="CS">
<![CDATA[ 列举全球气候变暖的危害。]]>
</QUESTION>
<NARRATIVE LANG="EN">
<![CDATA[ Users need to know the harm of global warming to human beings and the environment.]]>
</NARRATIVE>
<NARRATIVE LANG="CS">
<![CDATA[ 用户需要知道全球气候变暖对人类和环境有什么危害。]]>
</NARRATIVE>
</TOPIC>
```

Difference between CLIR and IR4QA

INSTITUTE OF COMPUTING TECHNOLOGY

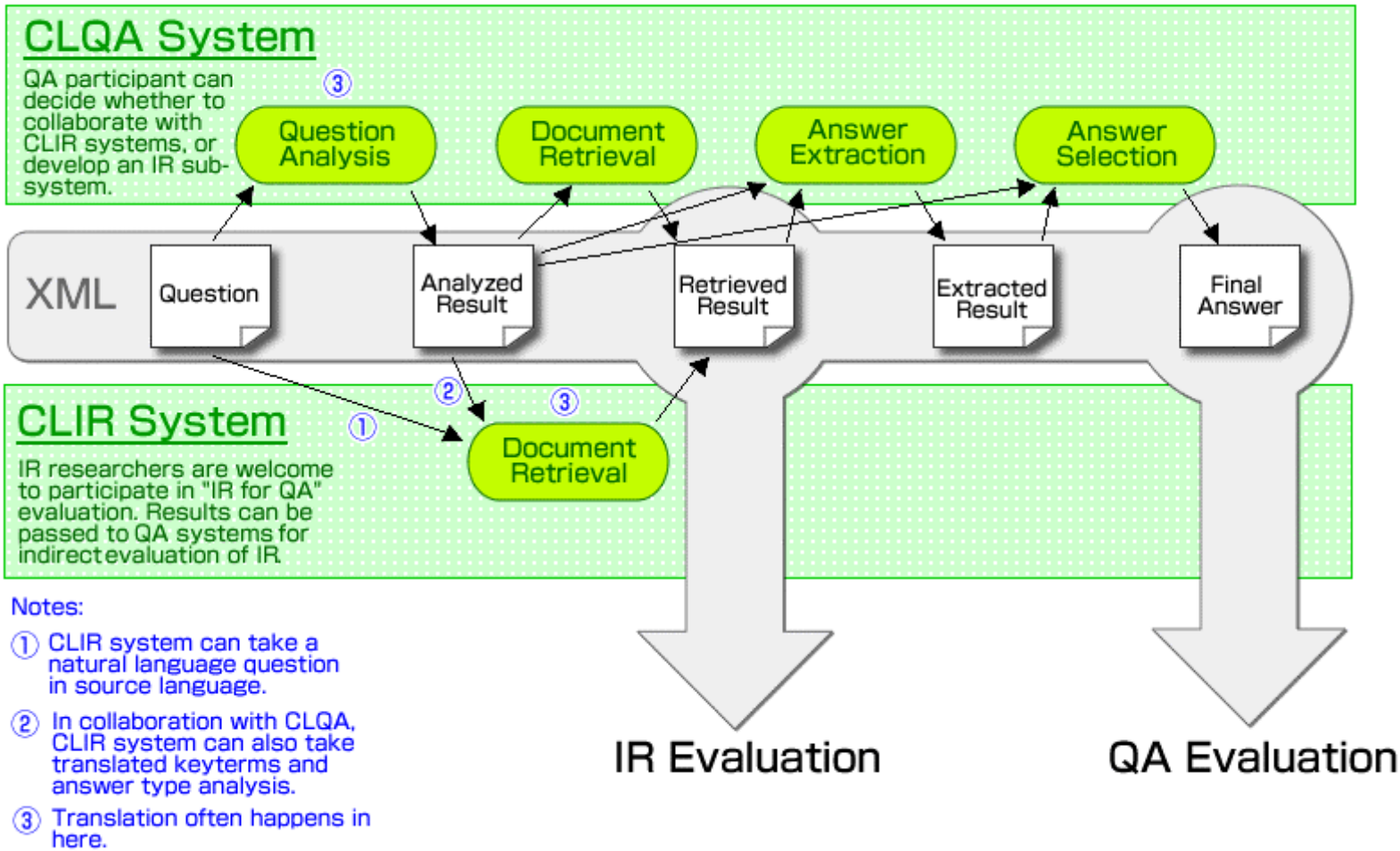
- Query
  - CLIR
    - Monolingual description of Information need (IN)
    - Different grained description of IN
      - TITLE → DESC → NARR
    - Accurate keywords
    - Detailed background and relevance judgement
  - IR4QA
    - Bilingual description of IN
    - One grained description of IN
      - QUESTION ≈ NARRATIVE
    - No keywords available
    - No background and relevance judgement in detail
    - Question type provided with topics

Difference between CLIR and IR4QA

INSTITUTE OF COMPUTING TECHNOLOGY

- Metrics of evaluation
  - CLIR
    - Traditional IR metrics: AP
  - IR4QA
    - IR evaluation
      - AP--primary
      - Q
      - nDCG
    - End-to-end evaluation with a QA system
      - F-score

Sample System Architecture and Data Flow



Notes:

① CLIR system can take a natural language question in source language.

② In collaboration with CLQA, CLIR system can also take translated keyterms and answer type analysis.

③ Translation often happens in here.

Outline

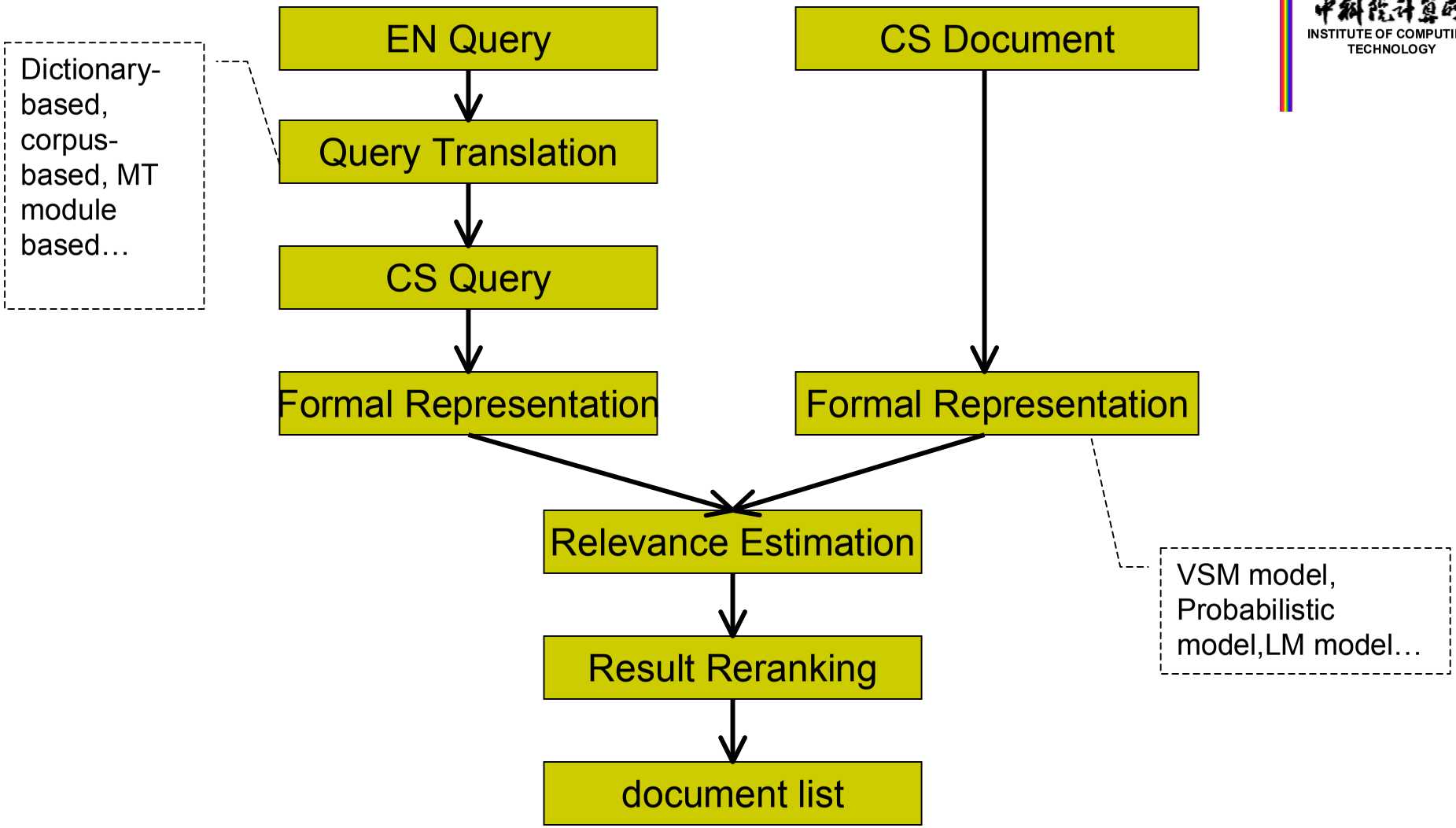
- Background
- The system architecture
- Query translation
- Document retrieval
- Experiments on the dry-run set
- Results on the formal run
- Conclusion

Our Implementation

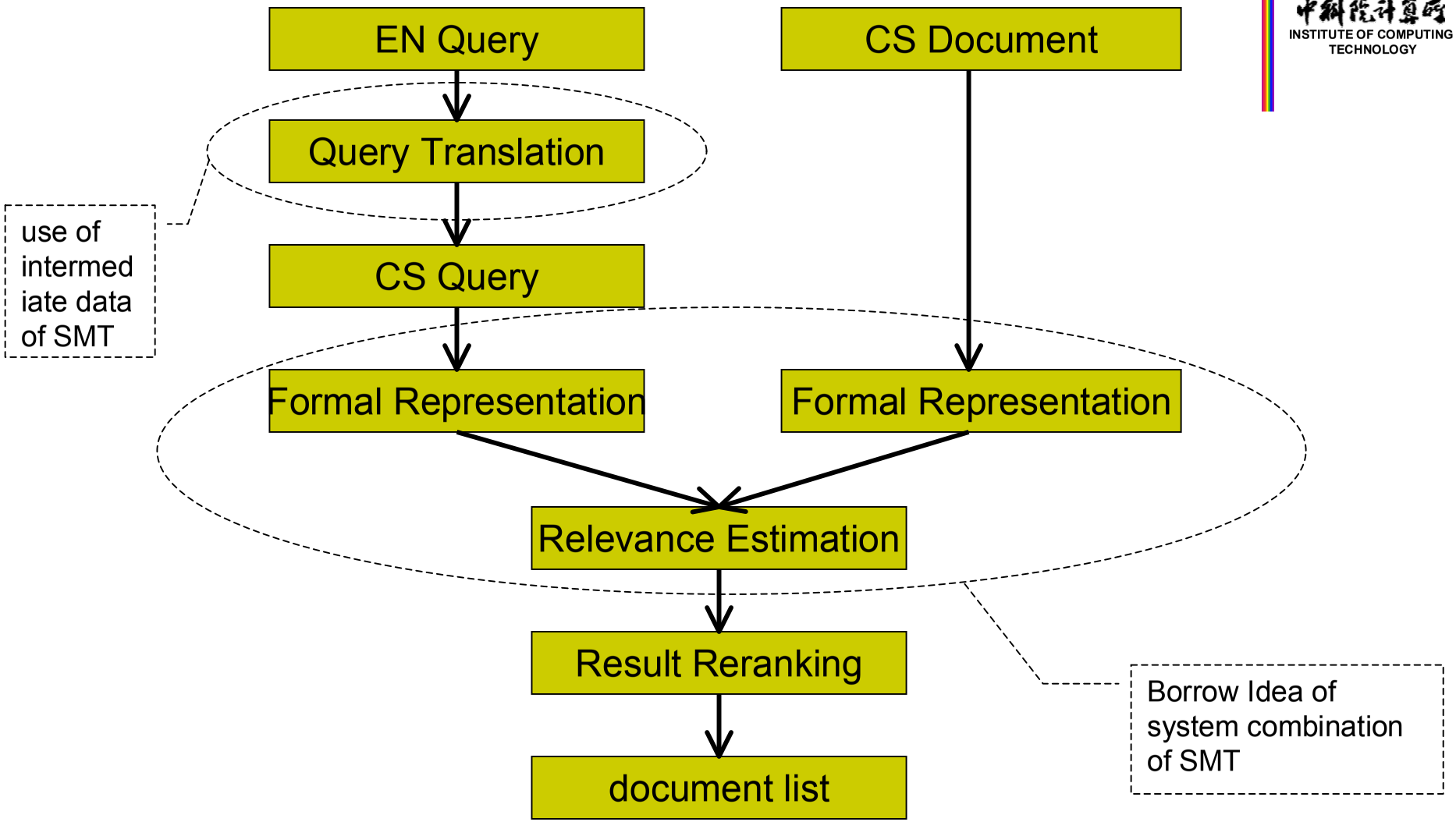
- Basic idea
  - A traditional CLIR framework still works
    - IR4QA is still an IR task
  - Major concerned issues

Modification should be made!

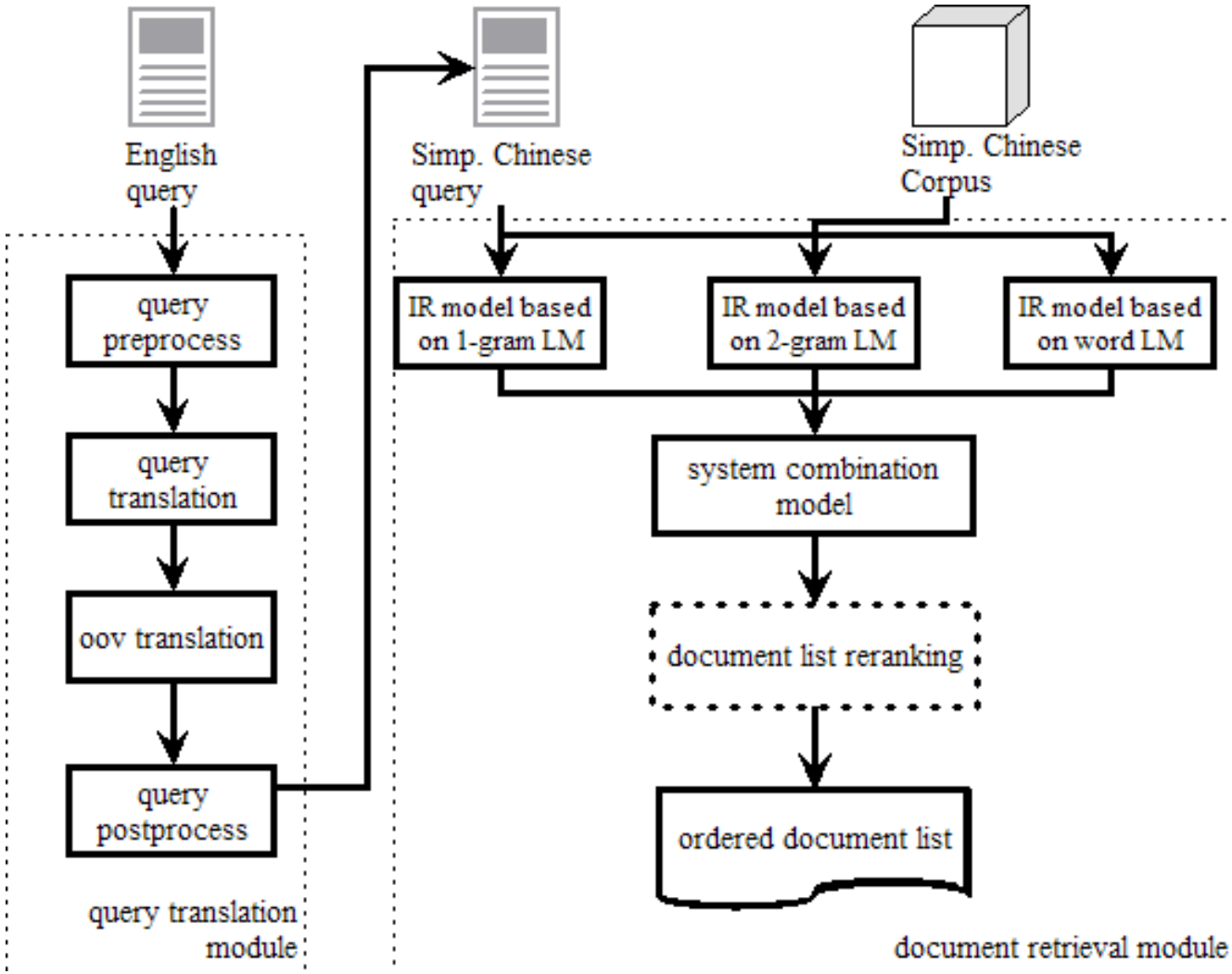
Traditional CLIR Flowchart



Our Improvement for IR4QA



ICT-Crossn


English query
Simp. Chinese query
Simp. Chinese Corpus
query preprocess
query translation
oov translation
query postprocess
query translation module
IR model based on 1-gram LM
IR model based on 2-gram LM
IR model based on word LM
system combination model
document list reranking
ordered document list
document retrieval module

Outline

- Background
- The system architecture
- Query translation
- Document retrieval
- Experiments on the dry-run set
- Results on the formal run
- Conclusion

Query Translation

- Basic ideas
  - Phrase translation
    - Full text translation is unfit for short questions
    - Phrases resolve some ambiguities
    - Phrase based SMT creates huge phrase tables
  - OOV words translation
    - Many OOV words provide key information of questions
    - Any dictionary has a limited coverage
    - Corresponding translation may be found with search engines

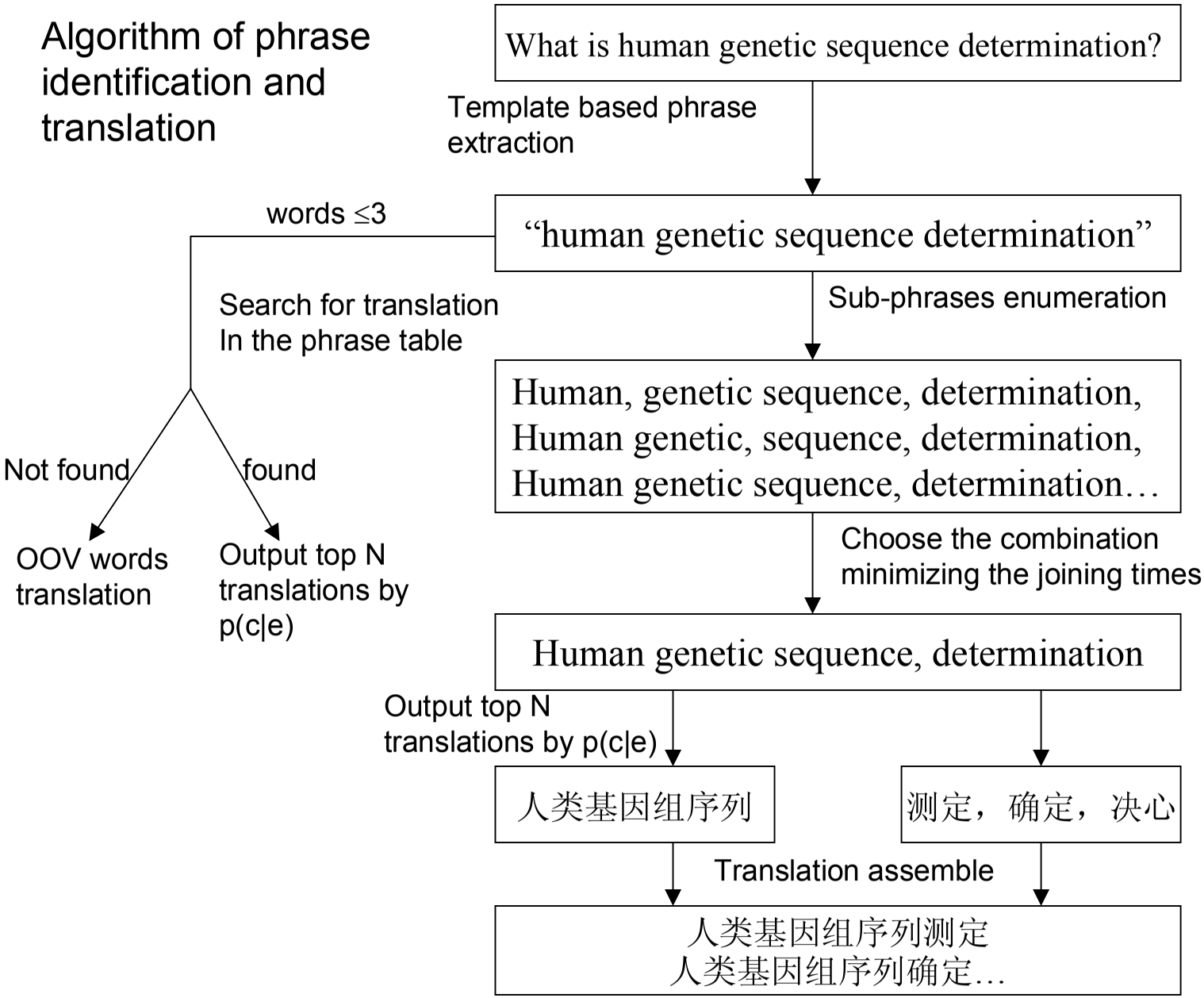
Phrase Translation

- Phrase table
  - Trained on 5M pairs of EN-CS sentence
  - Word alignment with Giza++
  - Phrase extraction and probabilities estimation with a tool in Mencius (our phrase-based SMT decoder)(F.J. Och, Hermann Ney, 2008)
  - Size up to 17GB
  - The phrase table is filtered by the test set

Sample of CS-EN Bilingual Phrase Table

! 三 井 住友	! Sumitomo Mitsui	0:0 1:2 2:2 3:1 3:2	1 0.018866 1 0.120792 2.718 1
! 三 井 住友 金融	! Sumitomo Mitsui Financial	0:0 1:2 2:2 3:1 3:2 4:3	1 0.0129453 1 0.0073817 2.718 1
! 三 井 住友 金融 集团	! Sumitomo Mitsui Financial Group	0:0 1:2 2:2 3:1 3:2 4:3 5:4	1 0.00932347 1 0.00318336 2.718 1
" , 以 对付	" to deal with the	0:0 2:2 3:2 3:3	1 0.000132428 0.5 0.000141711 2.718 1
" , 以 对付	" to deal with	0:0 2:2 3:2 3:3	1 0.000132428 0.5 0.000511601 2.718 1
" , 以 协调	to coordinate its own	2:0 3:1 3:2 3:3	0.333333 8.20519e-06 1 1.46173e-06 2.718 1
" , 以便 使	" so that the	0:0 2:1 2:2 3:1	0.333333 6.48644e-05 0.5 0.000702031 2.718 1
" , 以便 使	" so that	0:0 2:1 2:2 3:1	0.333333 6.48644e-05 0.5 0.00253444 2.718 1
广播 电视	Broadcast	0:0 1:0	0.444444 0.105186 0.166667 0.024001 2.718 1
广播 电视	broadcast and tv	0:0 1:2	1 0.106304 0.0833333 0.000460165 2.718 1
广播 电视	broadcasting	0:0 1:0	0.142857 0.0856858 0.166667 0.0846071 2.718 1
广播 电视台	TV station	0:0 1:0 1:1	0.0754717 0.00252064 0.5 0.0257354 2.718 1
广播 电视台	feeding radio and television	0:0 0:1 1:3	1 0.0685612 0.5 2.82344e-05 2.718 1
国际 发展 局	International Development Bureau	0:0 1:1 2:2	1 0.280224 1 0.00156186 2.718 1
国际 发展 局 援助	International Development Bureau Assistance	0:0 1:1 2:2 3:3	1 0.0941678 1 2.42523e-05 2.718 1
国际 发展 局 援助 局	International Development Bureau Assistance Department	0:0 1:1 2:2 3:3 4:4	1 0.00106589 1 1.82349e-07 2.718 1
国际 发展署 代表团	Delegation Visits	0:0 1:0 1:1 2:0	0.117647 0.000173761 1 0.091007 2.718 1
国际 法	international law	0:0 1:1	0.2 0.19695 0.75 0.0735244 2.718 1
国际 法	the international law	0:1 1:2	0.5 0.19695 0.25 0.020366 2.718 1
国际 法庭	International Court of Justice ,	0:0 1:1 1:3	1 0.111924 0.0238096 8.55327e-06 2.718 1
国际 法庭	International Court of Justice	0:0 1:1 1:3	0.6 0.111924 0.0714286 5.25604e-05 2.718 1
国际 反 恐	international anti terrorism	0:0 1:1 2:2	1 0.0935168 0.166667 0.0128866 2.718 1
国际 反 恐	international counter terrorism	0:0 1:1 2:2	1 0.0077033 0.166667 0.000422512 2.718 1
国际 反 恐	international terrorism	0:0 1:1 2:1	0.0384615 0.0118484 0.25 0.069292 2.718 1
外事	of the Foreign Affairs	0:2 0:3	0.75 0.0421837 0.0769231 0.00190038 2.718 1
外事	the Foreign Affairs	0:1 0:2	0.818182 0.0421837 0.115385 0.0142804 2.718 1
外事	the foreign affairs	0:1 0:2	0.214286 0.00529871 0.0384615 0.0015063 2.718 1
外事 人才	personnel of foreign affairs	0:2 0:3 1:0	1 0.000712337 1 9.76949e-05 2.718 1
外事 委员会	Affairs Committee	0:0 1:1	0.4 0.042495 0.0833333 0.0592912 2.718 1
外事 委员会	Foreign Affairs Committee ,	0:0 0:1 1:2	0.8 0.0268476 0.0833333 0.00196545 2.718 1
外事 委员会	Foreign Affairs Committee of the	0:0 0:1 1:2	0.666667 0.0268476 0.0185185 0.000445208 2.718 1
外事 委员会	Foreign Affairs Committee of	0:0 0:1 1:2	0.777778 0.0268476 0.0324074 0.00160727 2.718 1
外事 委员会	Foreign Affairs Committee	0:0 0:1 1:2	0.763285 0.0268476 0.365741 0.0120779 2.718 1

Algorithm of phrase identification and translation



Phrase translation

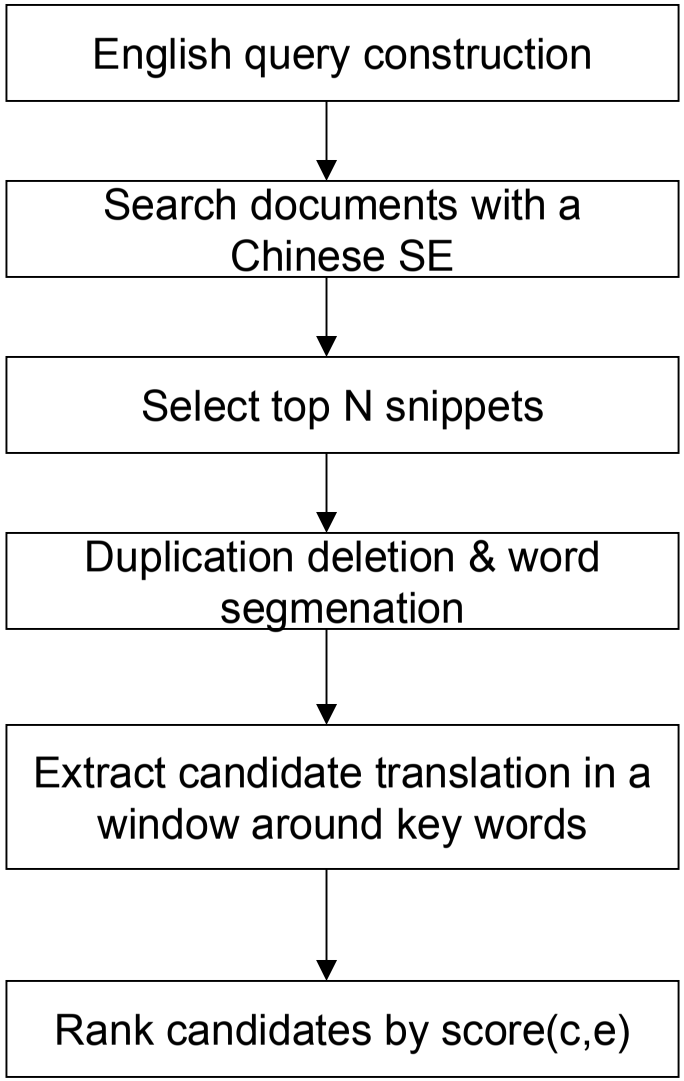
- Additional comments
  - The particle phrase/word is removed
  - The phrase occurring in many questions is removed
  - A word sequence with the first letter of its each word capitalized is viewed as a whole phrase
    - Balkan Syndrome
    - Edmund Ho Hau Wah
    - Parade of The Century
    - …

OOV words translation

- OOV words
  - Non-exists in the phrase table
  - Retrieve candidate translation with search engine (SE) (Ying Zhang, Phil Vines, 2004)
  - Rank candidates with a linear function

$$score(c,e) = \sum_i \omega_i f_i(c,e)$$

    - Parenthesis
    - Similarity of transliteration
    - co-occurrence probability
  - Weights are tuned in a greedy way


English query construction
Search documents with a Chinese SE
Select top N snippets
Duplication deletion & word segmenation
Extract candidate translation in a window around key words
Rank candidates by score(c,e)

Outline

- Background
- The system architecture
- Query translation
- Document retrieval
- Experiments on the dry-run set
- Results on the formal run
- Conclusion

Document Retrieval

- Basic ideas
  - Different models have different performance on different task
  - System combination in proper way improves over a single model (L.X.Shi, J.Y.Nie, 2007)
    - Various models are complementary
  - Lemur toolkit

Document Retrieval

- Index models? Retrieval models?
  - Index models
  - Various models with minor difference is easy to combine
  - Index models have less parameters to be tuned
  - A linear function and the objective function

$$score(d,q) = \sum \alpha_j score_j(d,q)$$

$$\delta(\bigcup_{i=1}^{n} Model_i) = MAP(\bigcup_{i=1}^{n} Model_i)$$

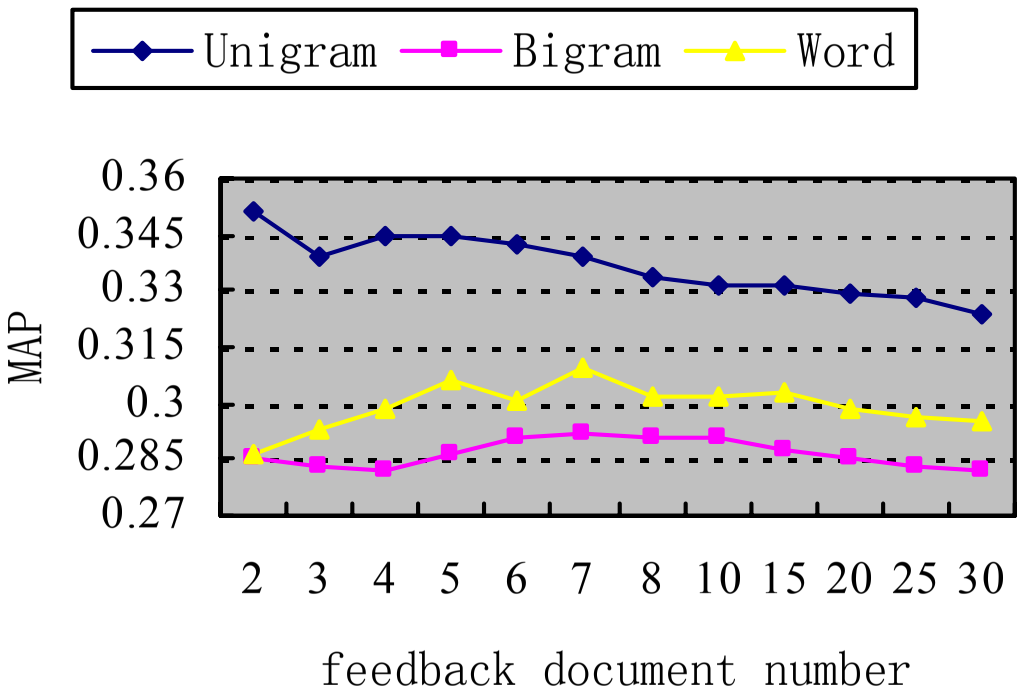
IR model selection

The results of different IR models and index units for Chinese monolingual IR on the NTCIR-5 test set

IR model	Mean Average Precision								
	T run			D run			N run		
	Unigram	Bigram	Word	Unigram	Bigram	Word	Unigram	Bigram	Word
tf_idf	0.2497	0.2664	0.2515	0.2164	0.2465	0.2052	0.2716	0.3032	0.2729
fb_tf_idf	0.2724	**0.3162**	0.2950	0.2569	0.3088	**0.2608**	0.2844	0.3362	0.2837
okapi	0.2609	0.2646	0.2582	0.2065	0.2393	0.1769	0.2835	0.3172	0.2737
fb_okapi	0.2931	0.3005	**0.3277**	0.1946	**0.3130**	0.2345	0.1678	**0.3390**	0.2773
kl_jm	0.2369	0.2366	0.2499	0.2218	0.2151	0.1851	0.2916	0.2970	0.2685
kl_dir	0.3031	0.2615	0.2706	0.2588	0.2270	0.2013	0.3154	0.3019	0.2790
kl_abs	0.3045	0.2519	0.2578	0.2302	0.2217	0.1823	0.3183	0.3013	0.2624
mixfb_kl_dir	**0.3443**	0.2860	0.3057	**0.2933**	0.2983	0.2513	**0.3343**	0.3378	**0.2992**

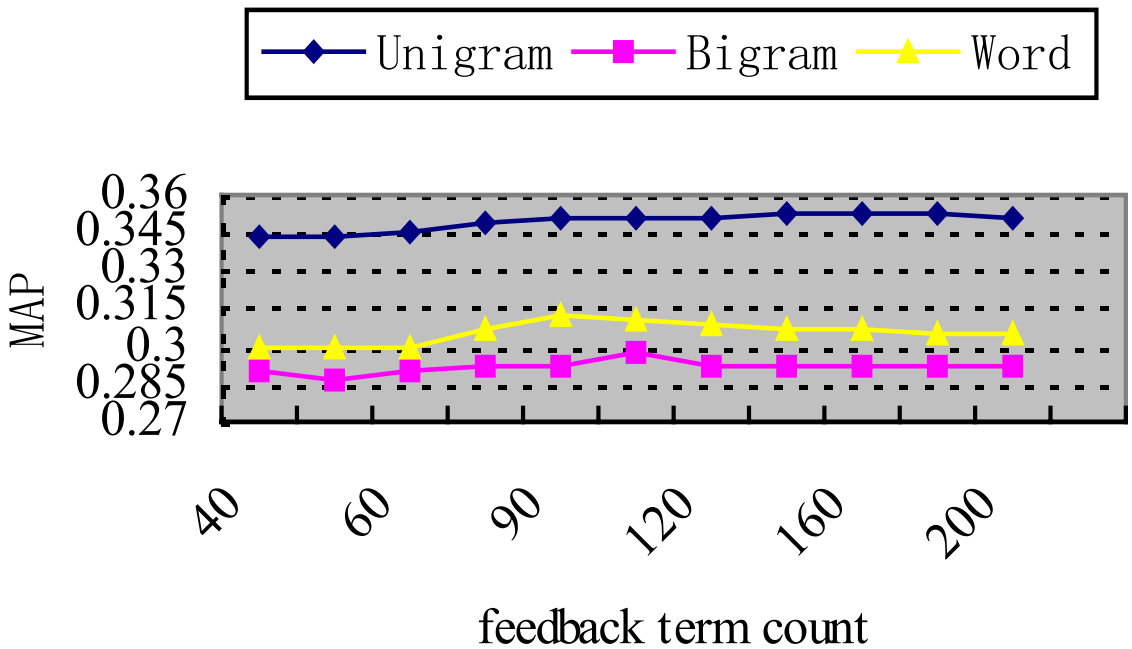
Parameter optimaztion

The results of different index units and feedback document number



Parameter optimaztion

The results of different index units and feedback term count



Index Unit Combination

- Retrieval model
  - language model, Dirrichlet smoothing, with feedback
- Score normalization
  - divide by the score of top 1 doc→[0,1]
- Index model combination
  - Each index model finishes retrieval and feedback respectively
  - the system combines their final score

Outline

- Background
- The system architecture
- Query translation
- Document retrieval
- Experiments on the dry-run set
- Results on the formal run
- Conclusion

Development set

- Construction
  - Motivation
    - Evaluation metric of IR4QA ≠ Evaluation metric of CLIR
    - Parameters tuned on previous test sets may well be unfit for IR4QA
  - Process
    - Download topics and answers in the dry run set
    - Convert according to specified XML format
    - Mark documents containing answers as relevant

Experiments on the dry-run set

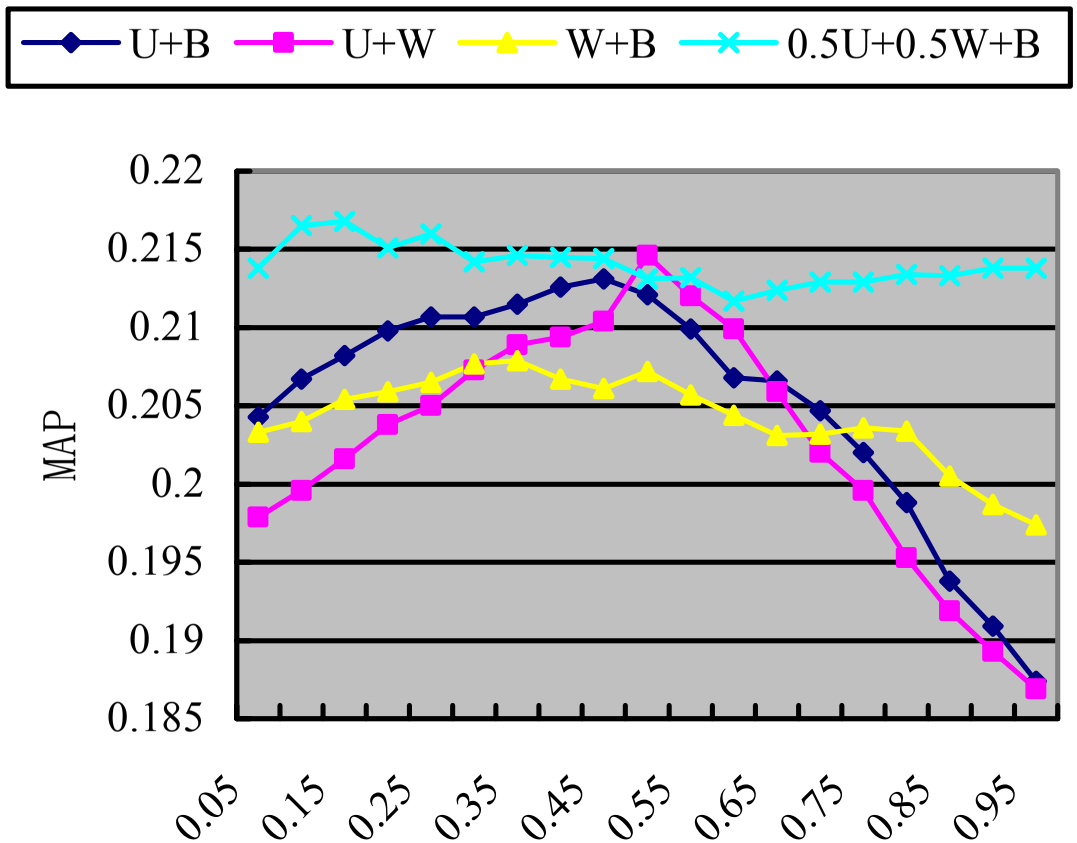
- Development set

Task	Question Type	Question number
EN-CS	Event	14
	Relationship	10
	Biography	21
	Definition	33
CT-CT	Event	8
	Relationship	10
	Biography	24
	Definition	29

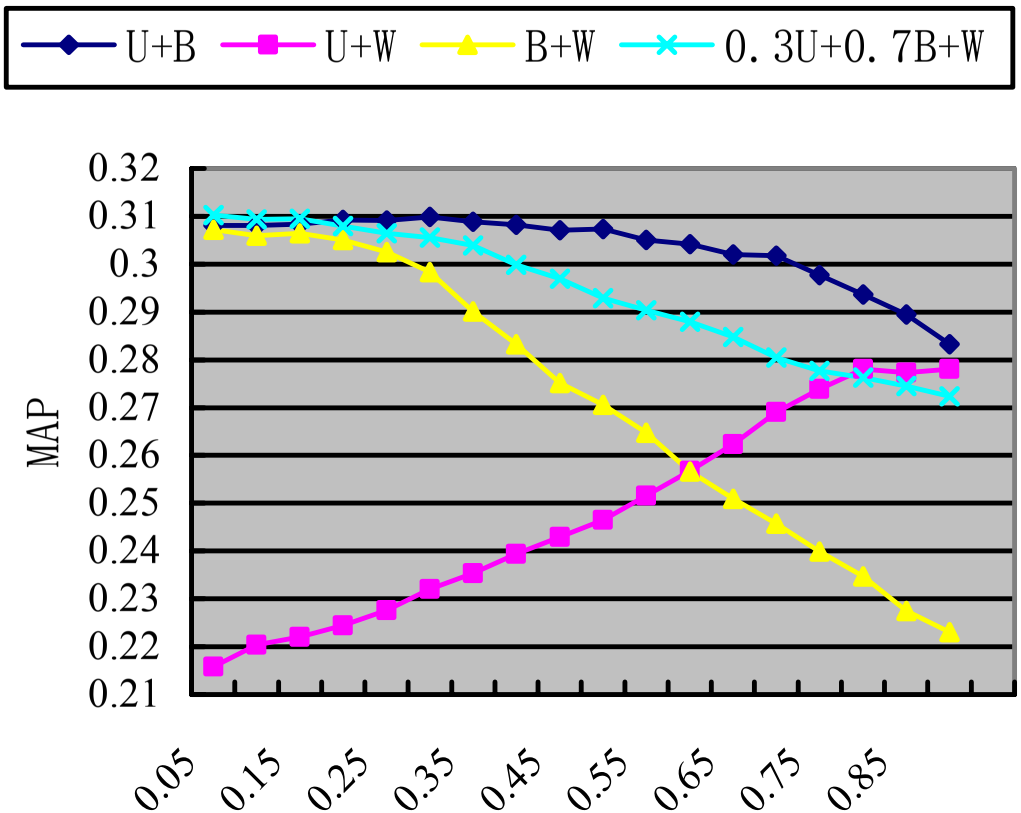
Sample topics in development data set

Topics	Relevant documents
What is Mayday?	za2001_038799 za2001_041818 za2001_040008 za1999_028092 za2001_042120 za2001_003573 za2000_014785 za2001_033102
List events related to Russian submarine Kursk explosion.	XIN20010312.0245 XIN20000815.0073 XIN20011009.0072 za2001_000149
What is the International Criminal Court?	XIN19980728.0164 XIN19980617.0100 za2000_061113
What is the relationship between Shintaro Ishihara and the literary circle?	za1999_010451 za1999_007192 za1999_012952 za1999_007192 XIN19990425.0127
印尼暴動與峇里島觀光有何關連？ (What is the relationship between riots in Indonesia and tourism in Bali?)	udn_xxx_19980911_0015 udn_xxx_19980517_0264 udn_xxx_19980521_0269 udn_xxx_19990512_0009 udn_xxx_19981012_0046
禽流感是在說什麼呢? (What is Avian influenza?)	udn_xxx_20010926_1110898 mhn_xxx_20010901_1073944 mhn_xxx_20010926_1110315 edn_xxx_20011203_1214182
誰是連戰？ (Who is Lien Chan?)	udn_xxx_20010521_0913048 udn_xxx_19990322_0237 udn_xxx_19990902_0408 udn_xxx_20000309_0049710 udn_xxx_19990507_0272
神舟號太空船 (List events about Shenzhou spacecrafts.)	udn_xxx_19991123_0131 udn_xxx_19991130_0249 udn_xxx_19991124_0194 udn_xxx_20000108_0255269 udn_xxx_19991122_0061 udn_xxx_19991123_0131 udn_xxx_20010111_0719983 udn_xxx_20010118_0730638 udn_xxx_19991122_0147 udn_xxx_20001106_0618907

The Results of Index Models Combination in EN-CS Task



The Results of Index Models Combination in CT-CT Task



Outline

- Background
- The system architecture
- Query translation
- Document retrieval
- Experiments on the dry-run set
- Results on the formal run
- Conclusion

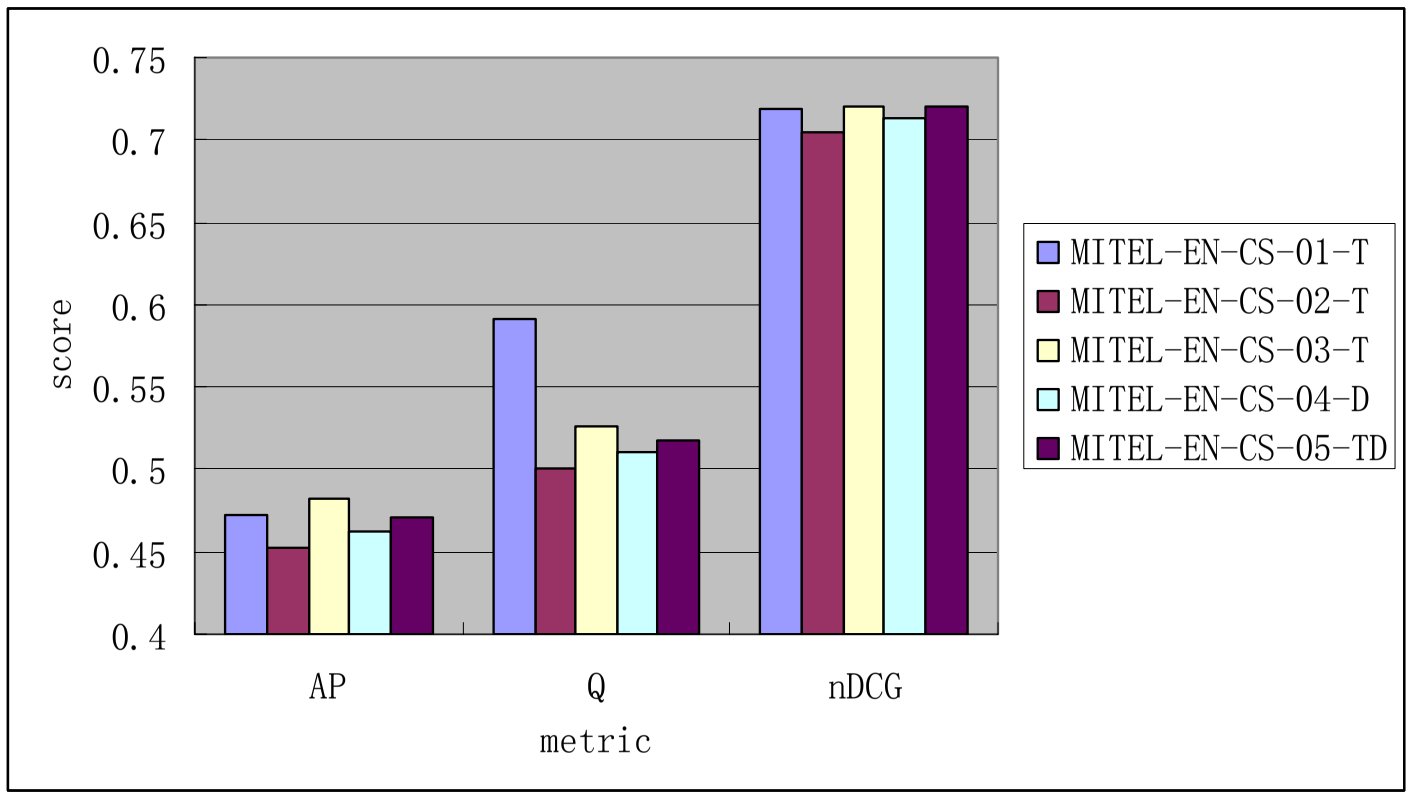
Results on the formal run

- Our runs

Run ID	Description
MITEL-EN-CS-01-T	0.5U+0.5W+0.15B, question part is used
MITEL-EN-CS-02-T	0.5U+0.5W, question part is used
MITEL-EN-CS-03-T	0.45U+0.55B, question part is used
MITEL-EN-CS-04-D	0.5U+0.5W+0.15B, narrative part is used
MITEL-EN-CS-05-TD	0.5U+0.5W+0.15B, both question and narrative part are used
MITEL-CT-CT-01-T	0.3U+0.7B, question part is used
MITEL-CT-CT-02-T	0.3U+0.7B+0.05W, question part is used
MITEL-CT-CT-03-D	0.3U+0.7B, narrative part is used
MITEL-CT-CT-04-T	Bigram, question part is used

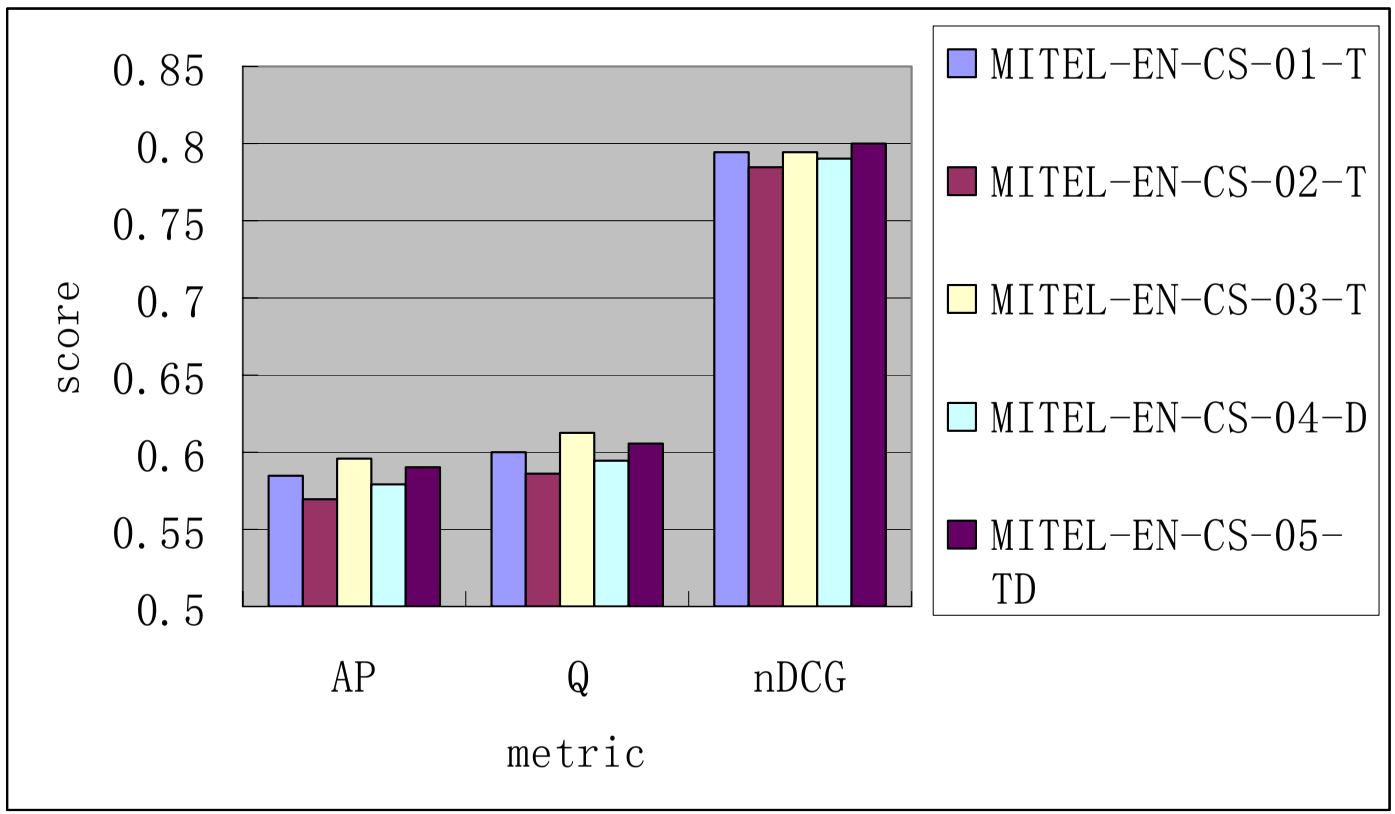
Performance in CLIR

The performance based on pseudo-qrels in the NTCIR-7 formal result (EN-CS)



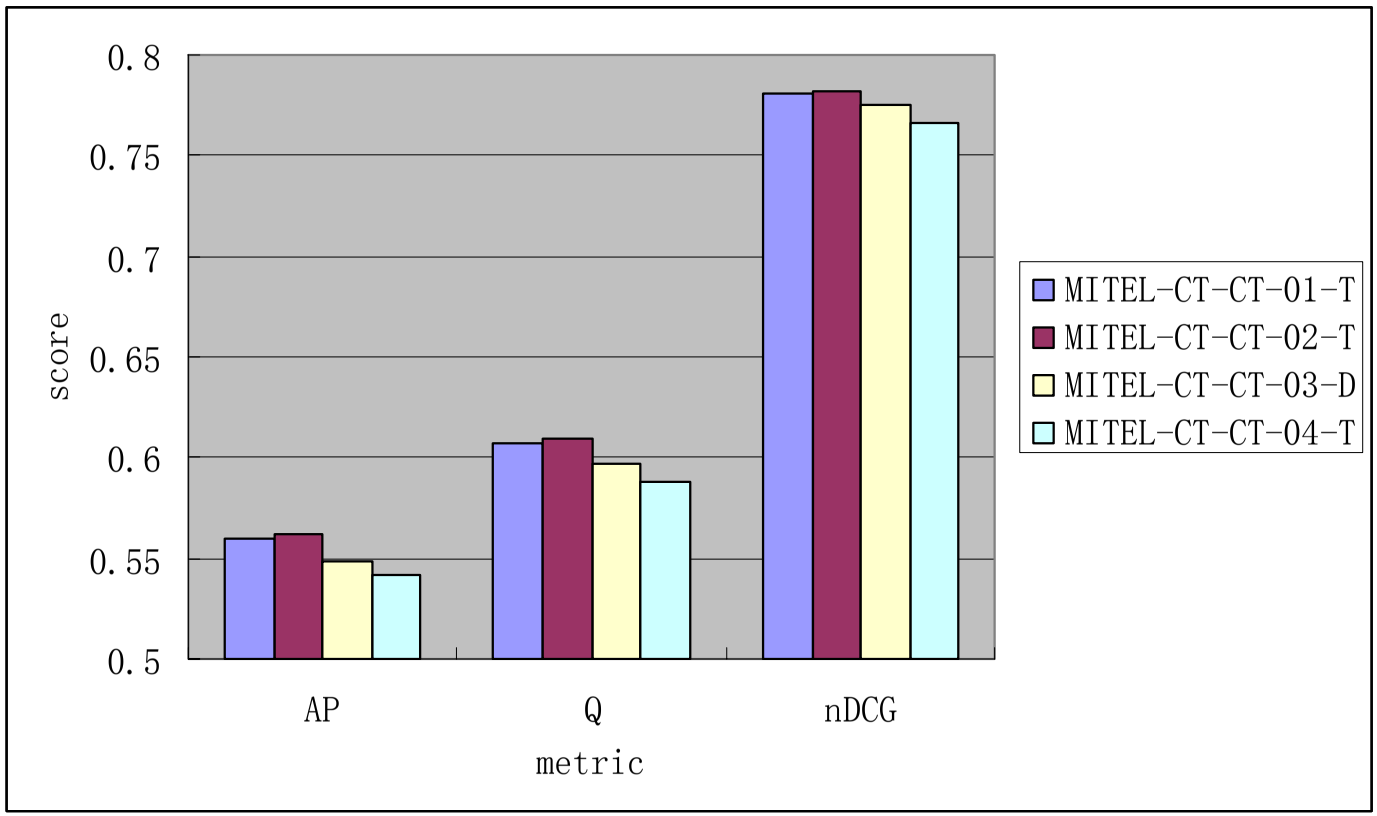
Performance in CLIR

The performance based on real qrels in the NTCIR-7 formal result (EN-CS)



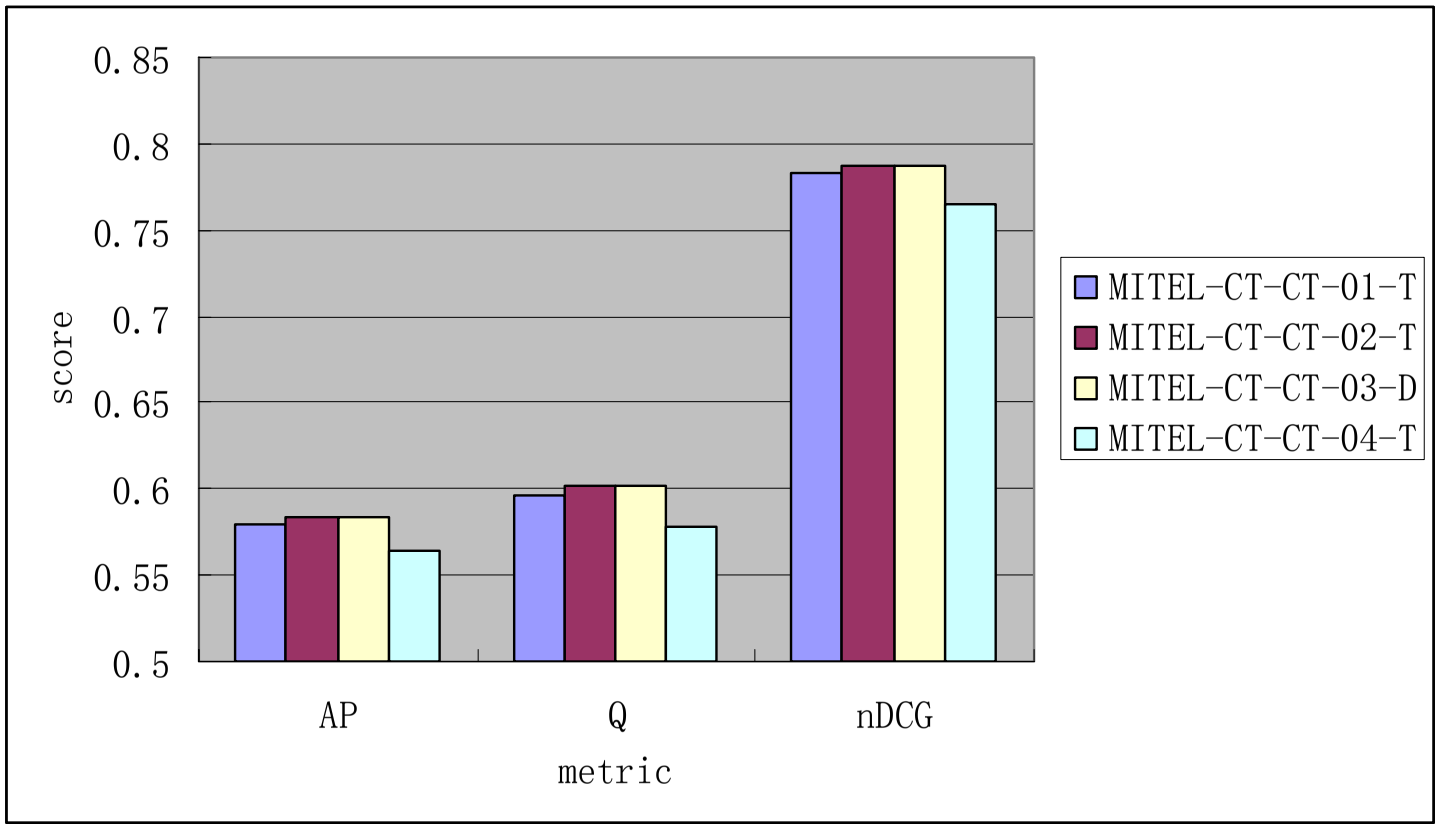
Performance in CLIR

The performance based on pseudo-qrels in the NTCIR-7 formal result (CT-CT)



Performance in CLIR

The performance based on real qrels in the NTCIR-7 formal result (CT-CT)



Performance in CCLQA with a QA system

INSTITUTE OF COMPUTING TECHNOLOGY

- Performance in CCLQA with a QA system

IR4QA \ CCLQA	ATR/NiCT	CMUJAV	KECIR
MITEL	0.2750	0.1982 (0.2168)	0.1829

EN-CS Collaboration Task (F3 scores based on automatic evaluation)

Analysis

- Overall recall with top 1000 over 90% (dry run)
- Precision relatively low
- Model combination helps recall more
- Evaluation metric of IR4QA differs from the one of CLIR
  - “Who is Hu Jintao?”
  - Bibliography of Hu is more appropriate than a meeting he attends
  - Model combination doesn’t change this!

Outline

- Background
- The system architecture
- Query translation
- Document retrieval
- Experiments on the dry-run set
- Results on the formal run
- Conclusion

Conclusion

- IR4QA is a novel but challenging task
- ICT-Crossn achieves a good performance
  - A phrase table of SMT + OOV words translation based on SE
  - Index model combination
  - The construction of Development set based on the dry run set

Conclusion

- Problems
  - identification and translation of phrases
    - “natural gas hydrates”
  - reranking
    - A reranking method for CLIR of I2R doesn’t work in IR4QA(L.P.Yang, D.H.Ji, 2005)
- Future work
  - Phrases translation
  - High recall & low precision mean…
    - Reranking should work
    - A better objection function for IR4QA should be explored

Information retrieval for question answering:
a challenging task but a long way to go

INSTITUTE OF COMPUTING TECHNOLOGY

Thank you!

Many thanks to my co-authors Tian Xia, Ji Guo and Qun Liu

Many thanks to Dr. Zhongjun He for his SMT system -- Mencius