Using Unigram and Bigram Language Models for Monolingual and Cross-Language IR

Lixin Shi and Jian-Yun Nie

Dept. d'Informatique et de Recherche Opérationnelle
Université de Montréal

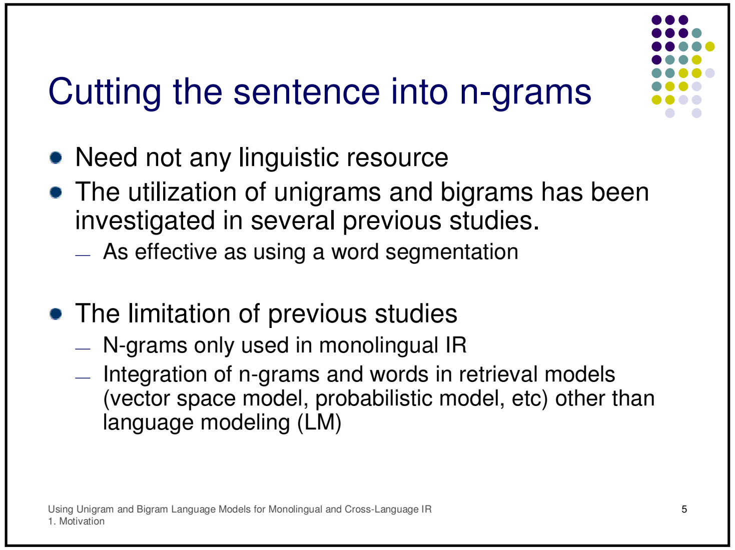
1. Motivation
2. Related Work
3. Using Different Indexing Units
4. Using Different Translation Units
5. Conclusion and Future Work

The difference between East-Asian and most European languages

- A common problem in East-Asian languages (Chinese, Japanese and Korean to some extent) is the lack of natural word boundaries.
- For information retrieval, we have to determine the index units first.
  - Using word segmentation
  - Cutting sentence into n-grams

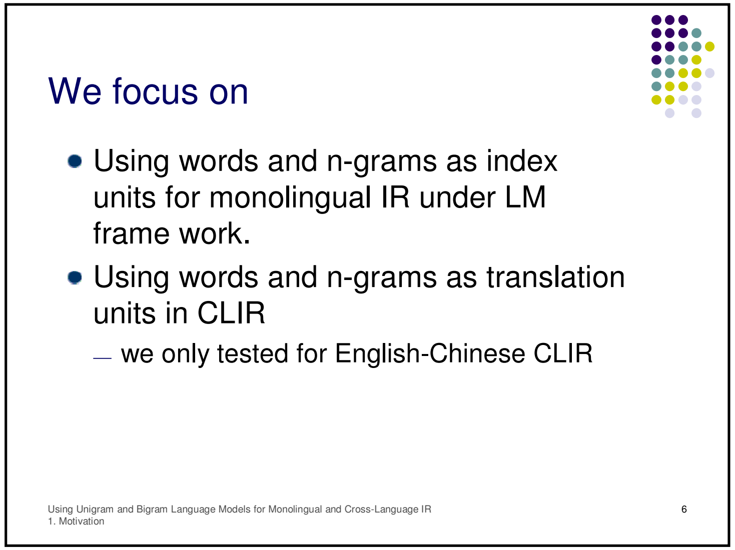
Word segmentation

- Based on rules, dictionaries and/or statistics
- Problems for information retrieval
  - Segmentation Ambiguity: The same string can be segmented into different words
    e.g. "发展中国家" ➔
    发展中(developing)/国家(country)
    发展(development)/中(middle)/国家(country)
    发展(development)/中国(China)/家(family)
  - If a document and a query are segmented into different words, there may be mismatch.
  - Two different words may have the same or related meaning, especially when they share come common characters.
    办公室(office) ↔ 办公楼(office building)



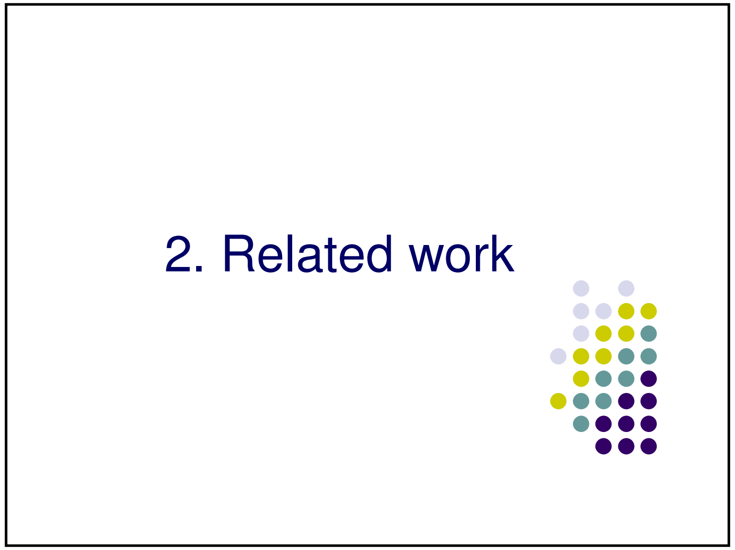
Cutting the sentence into n-grams

- Need not any linguistic resource
- The utilization of unigrams and bigrams has been investigated in several previous studies.
  - As effective as using a word segmentation
- The limitation of previous studies
  - N-grams only used in monolingual IR
  - Integration of n-grams and words in retrieval models (vector space model, probabilistic model, etc) other than language modeling (LM)

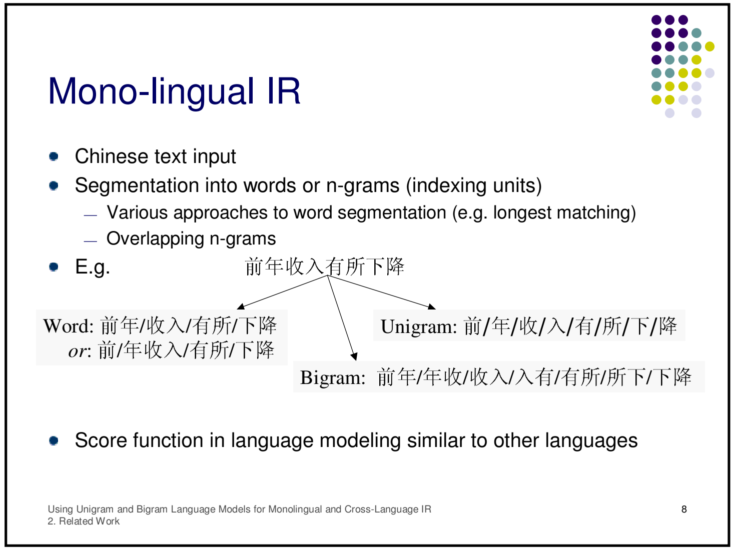


We focus on

- Using words and n-grams as index units for monolingual IR under LM frame work.
- Using words and n-grams as translation units in CLIR
  - we only tested for English-Chinese CLIR



2. Related work



Mono-lingual IR

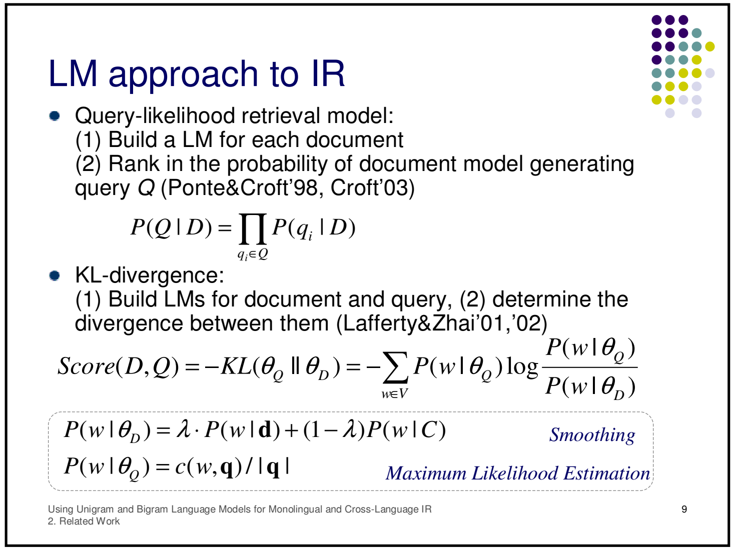
- Chinese text input
- Segmentation into words or n-grams (indexing units)
  - Various approaches to word segmentation (e.g. longest matching)
  - Overlapping n-grams
- E.g. 前年收入有所下降

Word: 前年/收入/有所/下降
or: 前/年收入/有所/下降

Unigram: 前/年/收/入/有/所/下/降

Bigram: 前年/年收/收入/入有/有所/所下/下降

- Score function in language modeling similar to other languages



LM approach to IR

- Query-likelihood retrieval model:
  (1) Build a LM for each document
  (2) Rank in the probability of document model generating query *Q* (Ponte&Croft'98, Croft'03)

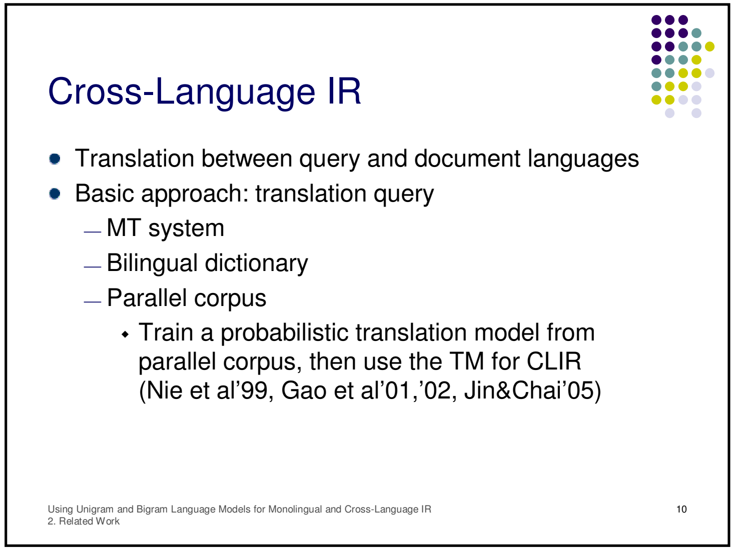
$$P(Q \mid D) = \prod_{q_i \in Q} P(q_i \mid D)$$

- KL-divergence:
  (1) Build LMs for document and query, (2) determine the divergence between them (Lafferty&Zhai'01,'02)

$$Score(D,Q) = -KL(\theta_Q \parallel \theta_D) = -\sum_{w \in V} P(w \mid \theta_Q) \log \frac{P(w \mid \theta_Q)}{P(w \mid \theta_D)}$$

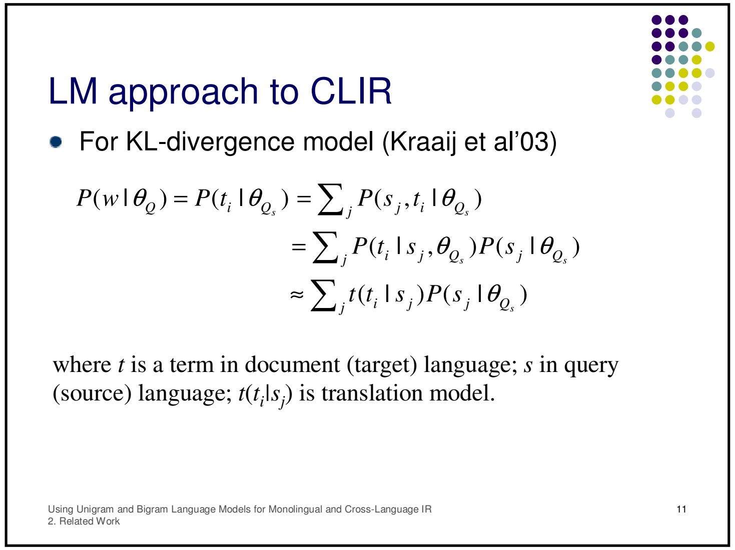
$$P(w \mid \theta_D) = \lambda \cdot P(w \mid \mathbf{d}) + (1-\lambda) P(w \mid C) \quad \textit{Smoothing}$$

$$P(w \mid \theta_Q) = c(w, \mathbf{q}) / |\mathbf{q}| \quad \textit{Maximum Likelihood Estimation}$$



Cross-Language IR

- Translation between query and document languages
- Basic approach: translation query
  - — MT system
  - — Bilingual dictionary
  - — Parallel corpus
    - Train a probabilistic translation model from parallel corpus, then use the TM for CLIR (Nie et al'99, Gao et al'01,'02, Jin&Chai'05)

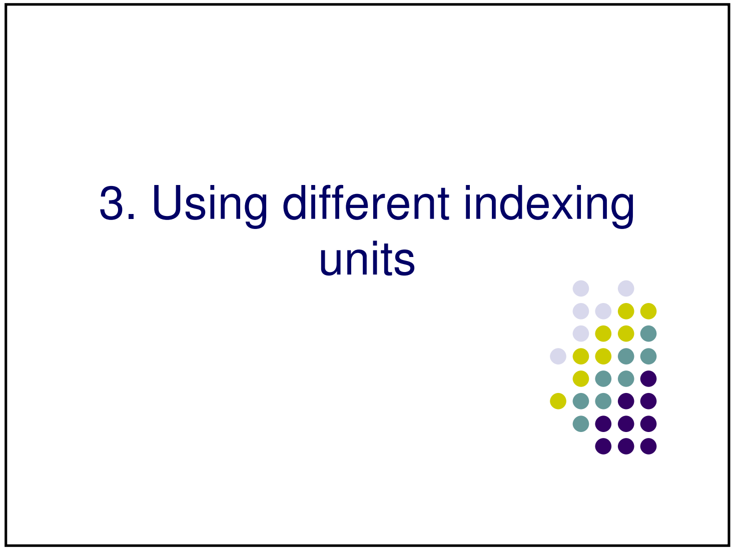


LM approach to CLIR

- For KL-divergence model (Kraaij et al'03)

$$\begin{aligned} P(w \mid \theta_Q) = P(t_i \mid \theta_{Q_s}) &= \textstyle\sum_j P(s_j, t_i \mid \theta_{Q_s}) \\ &= \textstyle\sum_j P(t_i \mid s_j, \theta_{Q_s}) P(s_j \mid \theta_{Q_s}) \\ &\approx \textstyle\sum_j t(t_i \mid s_j) P(s_j \mid \theta_{Q_s}) \end{aligned}$$

where *t* is a term in document (target) language; *s* in query (source) language; $t(t_i|s_j)$ is translation model.



3. Using different indexing units

Different indexing units

- Single index
  - Unigram (single character)
  - Bigram
  - Word

“国企研发投资”
U: 国/企/研/发/投/资
B: 国企/企研/研发/发投/投资
W: 国企/研发/投资

$$Score(D,Q) = -KL(\theta_Q \parallel \theta_D)$$

- Problems with single index
  - Words can be segmented in different ways
  - Closely related words cannot match

Combining different indexes

- Combine words with characters or bigrams and characters
  - Merging indexes
    - WU: Word & Unigram
    - BU: Bigram & Unigram

“国企研发投资”
WU: 国企/研发/投资/国/企/研/发/投/资
BU: 国企/企研/研发/发投/投资/国/企/研/发/投/资

  - Multiple indexes
    - B+U: Interpolate Bigram and Unigram

$$\begin{tikzcd}
 & Q_U \arrow[r] & -KL(Q_U,D_U) & D_U \arrow[l] & \\
Q \arrow[ur] \arrow[dr] & & & & D \arrow[ul] \arrow[dl] \\
 & Q_B \arrow[r] & -KL(Q_U,D_U) & D_B \arrow[l] &
\end{tikzcd}$$

$$Score(D,Q) = \sum_i \alpha_i Score_i(D,Q)$$

Experiment Setting

	NTCIR3/4		NTCIR5/6	
	Collections	#doc (KB)	Collections	#doc(KB)
Cn	CIRB011 CIRB020	381	CIRB040r	901
Jp	Mainichi98/99 Yomiuri98+99	594	Mainichi00/01r Yomiuri00+01	858
Kr	Chosunilbo98/99 Hankookilbo	254	Chosunilbo00/01 Hankookilbo00/01	220

	NTCIR3	NTCIR4	NTCIR5	NTCIR6
Numbers of topics	50	60	50	50

Using different index units for C/J/K monolingual IR on NTCIR4/5

Run	Means Average Precision (MAP)											
	U		B		W		BU		WU		0.3B+0.7U	
	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax
C-C-T-N4	.1929	.2370	.1670	.2065	.1679	.2131	.1928	.2363	.1817	.2269	**.1979**	**.2455**
C-C-T-N5	**.3302**	.3589	.2713	.3300	.2676	.3315	.2974	.3554	.3017	.3537	.3300	**.3766**
J-J-T-N4	.2377	.2899	.2768	.3670	–	–	.2807	**.3722**	–	–	**.2873**	.3664
J-J-T-N5	.2376	.2730	.2471	.3273	–	–	.2705	.3458	–	–	**.2900**	**.3495**
K-K-T-N4	.2004	.2147	.3873	.4195	–	–	**.4084**	**.4396**	–	–	.3608	.3889
K-K-T-N5	.2603	.2777	.3699	.3996	–	–	**.3865**	**.4178**	–	–	.3800	.4001

- Surprisingly, U is better than B and W for Chinese
- Interpolating unigram and bigram (B+U) has the best performance for Chinese and Japanese.
- However, BU and B are the best for Korean.

Analysis of monolingual IR results

- NTCIR 5 Topic 18
  - 烟 草 商 诉 讼 赔 偿 (Tobacco business, accusation, compensation)
  - **Word:** 烟草商(Tobacco business) 诉讼(accusation) 赔偿(compensation)
  - Unigram (0.7659) > Word(0.1625)
  - The relevant document *udn_xxx_20000716_0463237* includes 烟草,公司,业者, 香烟 ,烟商, but cannot match “烟草商”.
    It’s ranked 4th with unigram index, but 62nd with word index.
- NTCIR 5 Topic 24
  - 经 济 舱 综 合 症 候 群 航 班 (Economy class, syndrome, flight)
  - **Word:** 经济(economy) 综合症(syndrome) 候(wait) 航班(flight)
  - Ubigram(.7607)>Word(0.0002)
  - “..综合症候..” is segmented into “../综合症/候/..”
    It cannot match “症候” (syndrome).
  - The irrelevant document *udn_xxx_20011227_1251132* is retrieved only due to 综合症.
- The combination of unigrams with words or bigrams help solve these problems

The results of CJK monolingual IR on NTCIR6

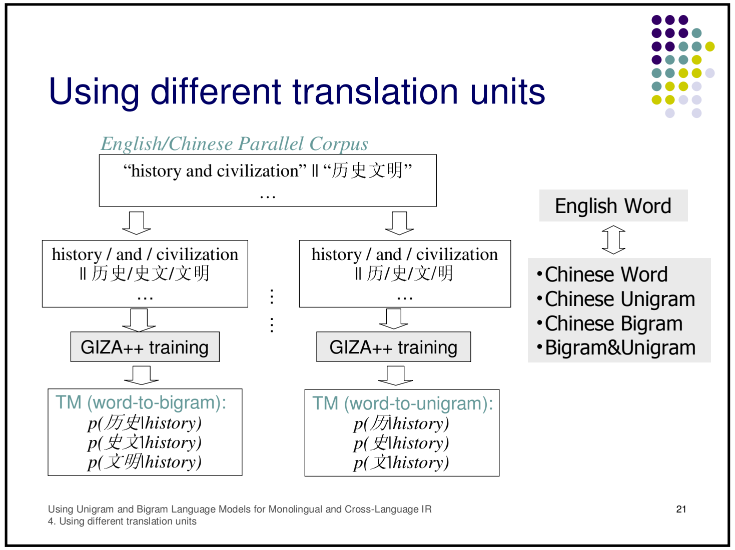
Run-id	RALI without pseudo feedback		RALI with pseudo feedback		Average MAP of all NTCIR6 runs	
	Rigid	Relax	Rigid	Relax	Rigid	Relax
C-C-T	.2139	.3022	.2330	.3303	.2269	.3141
C-C-D	.1671	.2376	.2031	.2907	.2354	.3294
J-J-T	.2426	.3171	.2576	.3343	.2707	.3427
J-J-D	.1877	.2485	.2292	.3052	.2480	.3214
K-K-T	.3332	.3939	.3460	.4130	.3833	.4644
K-K-D	.2623	.2970	.3287	.3945	.3892	.4678

- Our submission: Chinese&Japanese: U+B; Korean K-K-T:BU, K-K-D:U
- Our results are lower than average MAPs of NTCIR6:
  - We only aimed to compare index units using the basic IR technique
  - After apply a simple pseudo relevance feedback the results become more comparable to average MAPs.
- Globally, combining n-grams is a reasonable alternative to word segmentation
- (This is not new.)

4. Using different translation units

Existing approaches

- Translating English words to Chinese words
- Possibly cutting Chinese words into n-grams
- Then monolingual retrieval in Chinese
- Problem:
  - Coverage of Chinese words in the linguistic resources (dictionary, parallel corpus)
  - Variation of spelling in Chinese
  - Possible solution: also translating into n-grams ?



Using different translation units

English/Chinese Parallel Corpus

"history and civilization" ǁ "历史文明"
…

history / and / civilization
ǁ 历史/史文/文明
…

GIZA++ training

TM (word-to-bigram):
p(历史|history)
p(史文|history)
p(文明|history)

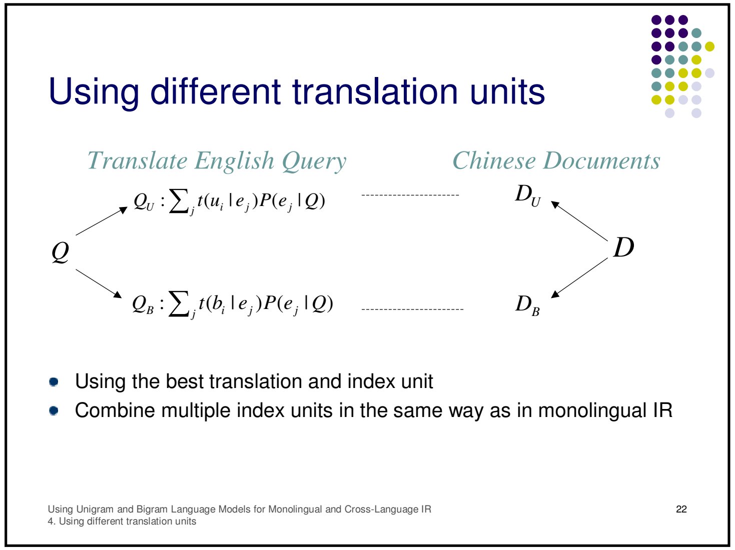
history / and / civilization
ǁ 历/史/文/明
…

GIZA++ training

TM (word-to-unigram):
p(历|history)
p(史|history)
p(文|history)

English Word

- Chinese Word
- Chinese Unigram
- Chinese Bigram
- Bigram&Unigram



Using different translation units

Translate English Query

Chinese Documents

Q

$Q_U : \sum_j t(u_i \mid e_j) P(e_j \mid Q)$

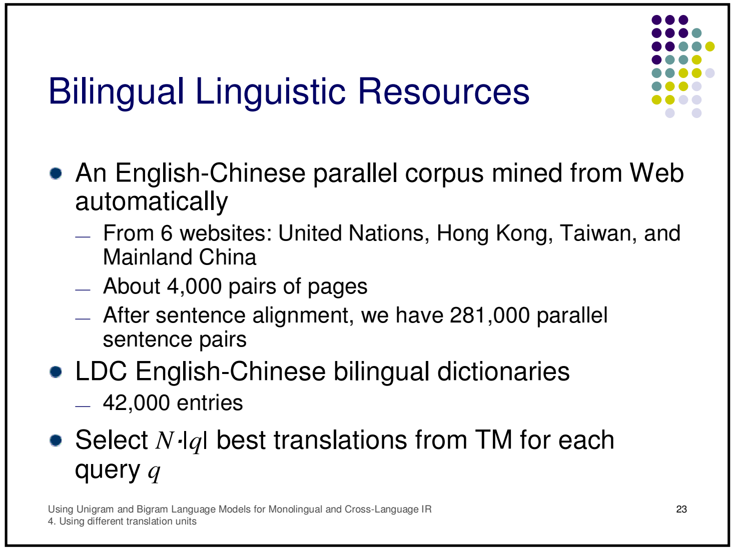
$Q_B : \sum_j t(b_i \mid e_j) P(e_j \mid Q)$

D_U

D_B

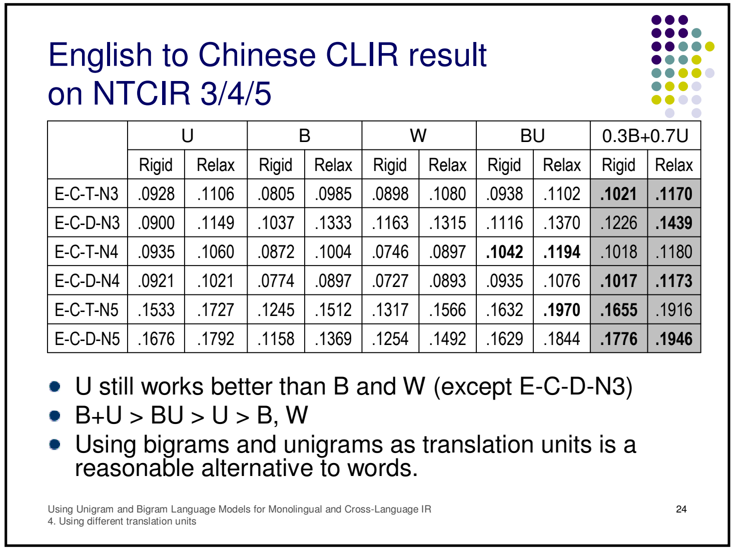
D

- Using the best translation and index unit
- Combine multiple index units in the same way as in monolingual IR



Bilingual Linguistic Resources

- An English-Chinese parallel corpus mined from Web automatically
  - From 6 websites: United Nations, Hong Kong, Taiwan, and Mainland China
  - About 4,000 pairs of pages
  - After sentence alignment, we have 281,000 parallel sentence pairs
- LDC English-Chinese bilingual dictionaries
  - 42,000 entries
- Select $N \cdot |q|$ best translations from TM for each query q



English to Chinese CLIR result on NTCIR 3/4/5

	U		B		W		BU		0.3B+0.7U	
	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax
E-C-T-N3	.0928	.1106	.0805	.0985	.0898	.1080	.0938	.1102	**.1021**	**.1170**
E-C-D-N3	.0900	.1149	.1037	.1333	.1163	.1315	.1116	.1370	.1226	**.1439**
E-C-T-N4	.0935	.1060	.0872	.1004	.0746	.0897	**.1042**	**.1194**	.1018	.1180
E-C-D-N4	.0921	.1021	.0774	.0897	.0727	.0893	.0935	.1076	**.1017**	**.1173**
E-C-T-N5	.1533	.1727	.1245	.1512	.1317	.1566	.1632	**.1970**	**.1655**	.1916
E-C-D-N5	.1676	.1792	.1158	.1369	.1254	.1492	.1629	.1844	**.1776**	**.1946**

- U still works better than B and W (except E-C-D-N3)
- B+U > BU > U > B, W
- Using bigrams and unigrams as translation units is a reasonable alternative to words.

Analysis of CLIR result

- NTCIR5 Topic 18: Tobacco business, accusation, compensation (烟草商，訴訟，賠償)
- MAP(BU)=0.1164 > MAP(W)=0.0044
  - Query translated by Bigram&Unigram TM:

偿 0.2601	**烟 0.2531**	**补偿 0.2127**	**补 0.2018**
业 0.1788	烟酒 0.1254	**商 0.1121**	**偿贸 0.1042**
指 0.0930	及 0.0926	**控 0.0795**	**企 0.0641**
企业 0.0639	**告 0.0638**	经 0.0602	**赔偿 0.0553**
草 0.0547	的指 0.0545	**赔 0.0537**	**指控 0.0497**
烟草 0.0484	务 0.0408	…	

  - Query translated by Word TM

补偿贸易 0.3523	烟酒 0.3453	**补偿 0.3349**	**企业 0.1923**
赔偿 0.1772	**指控 0.1558**	**烟草 0.1260**	公卖 0.1018
商务 0.0944	**经营 0.0877**	创业 0.0801	生意 0.0797
商 0.0778	用品 0.0728	**指责 0.0618**	业务 0.0547
至于 0.0540	商业 0.0536	台商 0.0476	报告 0.0462
事业 0.0456	组织 0.0415	…	

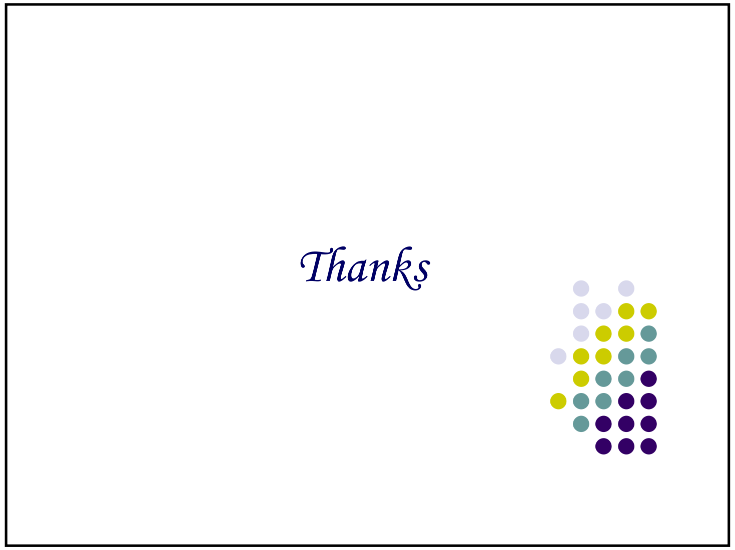
5. Conclusion and future work

Conclusion

- Our experimental results show that n-grams are generally as effective as words for monolingual and Cross-language IR in Chinese. For Japanese and Korean, n-grams approaches are comparable to the average results of NTCIR6.
- We tested creating different types of index separately, then grouping them during the retrieval process. We found that this approach is slightly more effective for Chinese and Japanese.
- Overall, n-grams can be interesting alternative indexing and translation units to word.

Future work

- We noticed that a type of index unit has variable effectiveness for different queries.
- Not reasonable to assign the same weight to a type of index for all queries
- Future work:
  - Make the weight dependent on query words.
  - Better parameter tuning methods



Thanks