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

Lixin Shi & Jian-Yun Nie
RALI, Dept. d'Informatique et de Recherche Opérationnelle
Université de Montréal
{shilixin, nie}@iro.umontreal.ca

Motivation

- No natural word boundaries in Chinese and Japanese: need to determine the index unit first.
  - Using word segmentation
  - Cutting sentence into n-grams
- Both types of indexes have been used in monolingual IR
- Question: How do they compare as translation units in CLIR?

Problems of word segmentation in information retrieval

- Segmentation Ambiguity:
  "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)
- Different words may have the same or related meaning, especially when they share common characters.
  办公室(office) ↔ 办公楼(office building)

Using different index units

- W (Word): Sentences are segmented into words.
- U (Unigram): Sentences are cut into single characters.
- B (Bigram): Sentences are cut into overlapping bigrams of characters.
- WU (Word and Unigram): Sentences are segmented into both words and single characters.
- BU (Bigram and Unigram): Sentences are cut into both overlapping character bigrams and single characters.
- B+U: Interpolating Bigram(B) and Unigram(U)

$$Score_{B+U}(D,Q) = \lambda Score_B(D,Q) + (1-\lambda) Score_U(D,Q)$$

LM approach for monolingual IR and CLIR

KL-divergence between query language model and document language model

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

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

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

LM For CLIR:

$$P(w|\theta_Q) = P(t_i|\theta_{Q_s}) = \sum_j P(s_j, t_i|\theta_{Q_s})$$
$$= \sum_j P(t_i|s_j, \theta_{Q_s}) P(s_j|\theta_{Q_s}) \approx \sum_j t(t_i|s_j) P(s_j|\theta_{Q_s})$$

	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

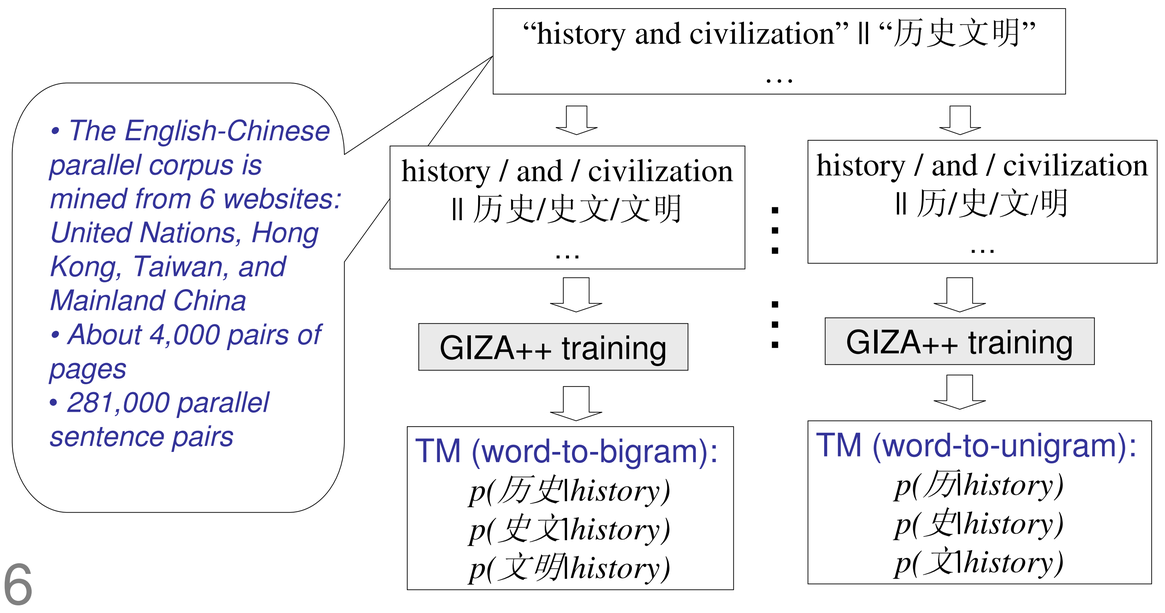
- Interpolating B and U is the best for Chinese and Japanese. But, BU and B are better for Korean

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

N-gram approach is comparable to the average results of NTCIR6

Using different translation units

Translate English Words to Chinese Words(W), Unigrams(U), Bigrams(B), or Bigrams&Unigrams(BU).



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**

- Using bigrams and unigrams as translation units seems a reasonable alternative to words.
- Combinations of bigrams and unigrams usually produce higher effectiveness.

Conclusion

- N-grams produce results comparable to the average results of NTCIR6 in Chinese, Japanese and Korean.
- For Chinese
  - N-grams are generally as effective as words for monolingual IR.
  - For Cross-language IR, n-grams approaches can be even better than dictionary-based word translation
- N-grams can be interesting alternative indexing and translation units to words.
- Worth further investigations.