Axiomatic Analysis of Cross-Language Information Retrieval

Razieh Rahimi[a], Azadeh Shakery[a], Irwin King[b]

[a]School of Electrical and Computer Engineering, University of Tehran

[b]Department of Computer Science and Engineering, The Chinese University of Hong Kong

Adoption of translation knowledge in CLIR models

A major challenge in Cross-Language Information Retrieval (CLIR) is adoption of translation knowledge in retrieval models, as it affects the term weighting which is known to highly impact the retrieval performance.

Axiomatic analysis

This analysis is based on formal constraints that any reasonable retrieval model should satisfy.

Our contribution

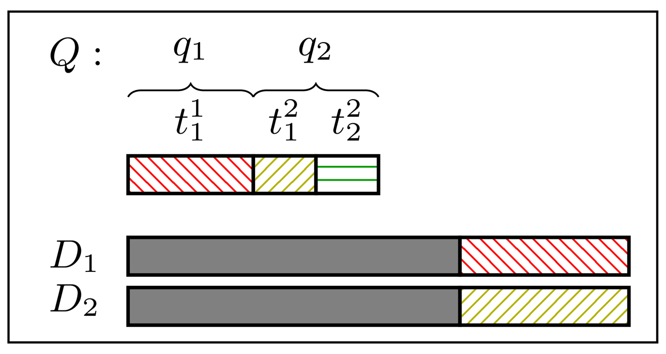
By adopting axiomatic analysis framework, we **formulate the impacts of translation knowledge on document ranking as constraints** that any cross-language retrieval model should satisfy.

CL-C1

The first constraint targets queries in which query terms have different numbers of translation alternatives in the target language, in particular when query terms are not ambiguous and translation alternatives are synonyms.

- $Q = \{q_1 q_2\}$: a two-term query.
- $p(t_1^1|q_1) = \alpha$,
  $p(t_1^2|q_2) = \beta$,
  $p(t_2^2|q_2) = \gamma$,
  such that $\beta + \gamma > \alpha$.
- q_2 is not ambiguous and its translations, t_1^2 and t_2^2, are synonyms or related words.
- $|D_1| = |D_2|$,
  $c(t_1^1, D_1) = c(t_1^2, D_2)$,
  Other translations of query terms do not occur in D_1 and D_2.
- $\mathrm{DV}(t_1^1) = \mathrm{DV}(t_1^2)$. (Term Discrimination Value)

$\Rightarrow$ $S(Q, D_2) > S(Q, D_1)$



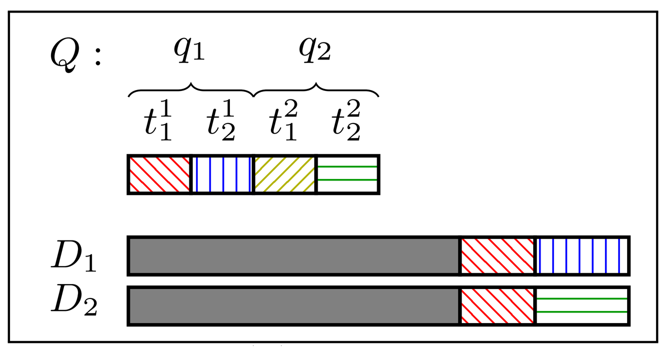
- D_1 and D_2 in the figure have equal occurrences of translations of query terms.
- Assume that these translations have the same discrimination value.
- According to translation probabilities, $p(t_1^1|q_1) > p(t_1^2|q_2)$, D_1 seems a better match to the query, because it contains t_1^1.
- However, considering that t_1^2 and t_2^2 are synonyms, we can say that $p(t_1^2|q_2) = \beta + \gamma$, which is greater than $p(t_1^1|q_1)$.
- In this case, weighting based on translation probabilities will artificially enhance query terms with fewer synonym translations, which this constraint intends to avoid.

CL-C2

The second CLIR constraint is about the coverage of translations of distinct query terms. Consider two documents that have the same total occurrences of translations of query terms and the same coverage of different translation alternatives of all query terms. The document that covers translations of more distinct original query terms should get a higher score.

- $Q = \{q_1 q_2\}$: a two-term query.
- $p(t_i^1|q_1) = p(t_j^2|q_2)$.
- $|D_1| = |D_2|$,
  $c(t_k^1, D_1) = c(t_k^1, D_2)$ where t_k^1 is a translation of q_1 and $k \neq i$,
  $c(t_i^1, D_1) = c(t_j^2, D_2)$,
  Other translations of query terms do not occur in D_1 and D_2.
- $\mathrm{DV}(t_i^1) = \mathrm{DV}(t_j^2)$.

$\Rightarrow$ $S(Q, D_2) > S(Q, D_1)$



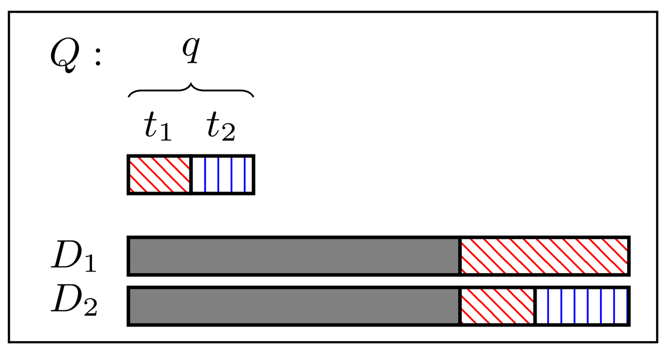
- t_1^1 and t_2^1 occur in document D1 with the same total number as the occurrences of t_1^1 and t_2^2 in document D2.
- But, D_1 covers only the translations of one query term q_1 while D_2 covers the translations of both query terms q_1 and q_2.
- Assume t_2^1 and t_2^2 have the same discrimination value.
- D_2 should get a higher score since it covers translations of more distinct original query terms.

CL-C3

The third constraint is about the coverage of different translation alternatives of a query term.

- $Q = \{q\}$: a query with only one term.
- $p(t_1|q) = p(t_2|q)$.
- $|D_1| = |D_2|$,
  $c(t_1, D_1) = c(t_1, D_2) + c(t_2, D_2)$,
  $c(t_2, D_1) = 0$,
  $c(t_1, D_2) > 0$,
  $c(t_2, D_2) > 0$,
  Other translations of q do not occur in D_1 and D_2.
- $\mathrm{DV}(t_1) = \mathrm{DV}(t_2)$.

$\Rightarrow$ $S(Q, D_2) > S(Q, D_1)$



- D_1 and D_2 have the same total occurrences of t_1 and t_2.
- D_2 covers two distinct translations of query term q, while D_1 covers only one translation of q.
- Assuming that t_1 and t_2 have the same discrimination value, D_2 should get a higher score w.r.t. query Q.

Constraint analysis on CLIR models

Summary of constraint analysis results for two CLIR models

	Corpus-based CLIR models	
	Probabilistic Structured Queries (PSQ)	LM-based query translation approach
	$\mathrm{TF}(q_i, D) = \sum_{w_t \in V_t} p(w_t|q_i)\mathrm{TF}(w_t, D)$, $\mathrm{DF}(q_i) = \sum_{w_t \in V_t} p(w_t|q_i)\mathrm{DF}(w_t)$.	$\mathrm{S}(Q, D) = \sum_{w_t \in V_t} p(w_t|\theta'_Q)\log p(w_t|\theta_D)$, $p(w_t|\theta'_Q) = \sum_{w_s \in V_s} p(w_t|w_s)p(w_s|\theta_Q)$.
CL-C1	✗	✗
CL-C2	✓	✗
CL-C3	✗	✓

Experiments

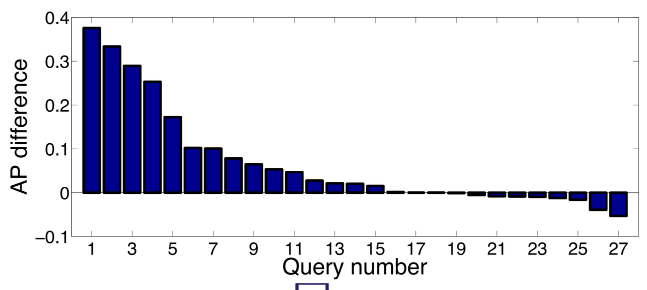
- Manually select queries in which the query terms have different numbers of synonymous translations in a translation model, trained on a parallel corpus.
- Study the effectiveness of using all translation alternatives for each query term, where:

"Syn" strategy: Each translation is weighted according to the translation model

"All" strategy: Considering the most probable translation with probability 1 and considering all other translations as instances of the most probable translation

Performance of CLIR models on selected queries.

Method	RUN	MAP (% All)	P@10	R@1000
LM-Based	All	0.2292	0.4704	0.6489
	Syn	**0.2959*** (29.1%)	**0.5333**	**0.7226**
PSQ	All	0.2513	0.4148	0.6814
	Syn	**0.2815** (12.0%)	**0.4444**	**0.7069**



Importance of satisfying CL-C1