

Cross-Lingual Query Suggestion Using Query Logs of Different Languages

Wei Gao^{1*}, Cheng Niu², Jian-Yun Nie³, Ming Zhou², Jian Hu²,
Kam-Fai Wong¹, Hsiao-Wuen Hon²

¹The Chinese University of Hong Kong

{wgao, kfwong}@se.cuhk.edu.hk

²Microsoft Research Asia

{chengniu, mingzhou, jianh, hon}@microsoft.com

³Université de Montréal

nie@iro.umontreal.ca

*This work was done when the author was visiting at Microsoft Research Asia

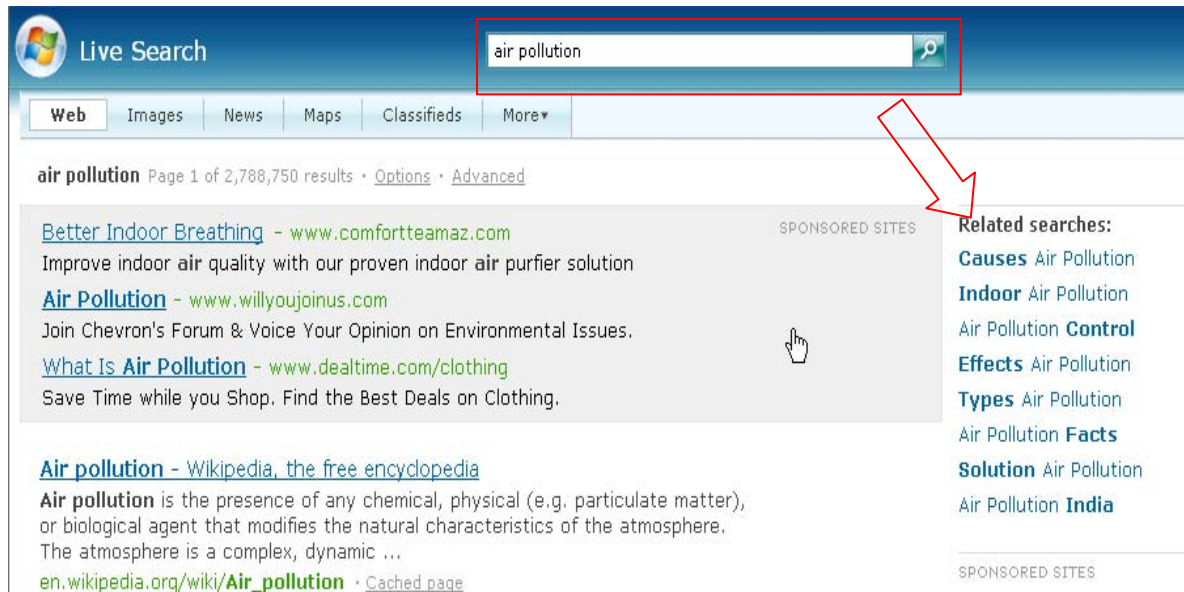


Outline

- **Introduction**
- Discriminative Model for Cross-Lingual Query Suggestion (CLQS)
- Mono-/Cross-Lingual Features
- CLIR with CLQS
- Performance Evaluation
- Conclusions

Query Suggestion

- Query Suggestion
 - A functionality that helps search engine users better specify their information needs with related queries having been frequently used by other users.
- Example – MSN Live Search



The screenshot shows the MSN Live Search interface. The search bar at the top contains the text "air pollution" and is highlighted with a red box. Below the search bar, there are tabs for "Web", "Images", "News", "Maps", "Classifieds", and "More". The search results are displayed on a page titled "air pollution" with 2,788,750 results. The results include sponsored sites and a "Related searches" section. A red arrow points from the search bar to the "Related searches" section.

air pollution Page 1 of 2,788,750 results • [Options](#) • [Advanced](#)

Better Indoor Breathing - www.comforteamaz.com
Improve indoor air quality with our proven indoor air purifier solution

Air Pollution - www.willyoujoinus.com
Join Chevron's Forum & Voice Your Opinion on Environmental Issues.

What Is Air Pollution - www.dealtime.com/clothing
Save Time while you Shop. Find the Best Deals on Clothing.

Air pollution - [Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/Air_pollution)
Air pollution is the presence of any chemical, physical (e.g. particulate matter), or biological agent that modifies the natural characteristics of the atmosphere. The atmosphere is a complex, dynamic ...
en.wikipedia.org/wiki/Air_pollution • [Cached page](#)

Related searches:
Causes Air Pollution
Indoor Air Pollution
Air Pollution Control
Effects Air Pollution
Types Air Pollution
Air Pollution Facts
Solution Air Pollution
Air Pollution India

SPONSORED SITES

More Example

- MSN and Google's keyword tool suggesting terms for pay-for-performance search market.

Keyword Tool

The Keyword Tool generates potential [keywords for your ad campaign](#) and reports their Google statistics, including search performance and seasonal trends. Start your search by entering your own keyword phrases or a specific URL. You can then add new keywords to the green box at the right. [Learn more](#)

Important note: Please note that we cannot guarantee that these keywords will improve your campaign performance. We also reserve the right to disapprove any new keywords you add. Keep in mind that you alone are responsible for the keywords you select and for making sure that your use of the keywords does not violate any applicable laws, including any applicable trademark laws. For more details, please review our [Terms and Conditions](#).

Results are tailored to **English, United States** [Edit](#)

Keyword Variations | **Site-Related Keywords**

Enter one keyword or phrase per line:

air pollution Use synonyms

Choose data to display: [?](#)

More specific keywords - sorted by relevance [?](#)

<u>Keywords</u>	<u>March Search Volume</u> ?	<u>Advertiser Competition</u> ?	Match Type: ?
air pollution	<div><div style="width: 50%;"></div></div>	<div><div style="width: 10%;"></div></div>	<input type="button" value="Add >"/>
air water pollution	<div><div style="width: 20%;"></div></div>	<div><div style="width: 0%;"></div></div>	<input type="button" value="Add >"/>
indoor air pollution	<div><div style="width: 30%;"></div></div>	<div><div style="width: 40%;"></div></div>	<input type="button" value="Add >"/>
air pollution effects	<div><div style="width: 40%;"></div></div>	<div><div style="width: 30%;"></div></div>	<input type="button" value="Add >"/>
ozone air pollution	<div><div style="width: 10%;"></div></div>	<div><div style="width: 0%;"></div></div>	<input type="button" value="Add >"/>
epa air pollution	<div><div style="width: 15%;"></div></div>	<div><div style="width: 20%;"></div></div>	<input type="button" value="Add >"/>



Query Suggestion and Query Expansion

- Query Suggestion vs. Query Expansion

	Query expansion	Query suggestion
Target	Terms and/or phrases	Full queries
Mechanism	Term/phrase extraction from documents	Query extraction from query logs

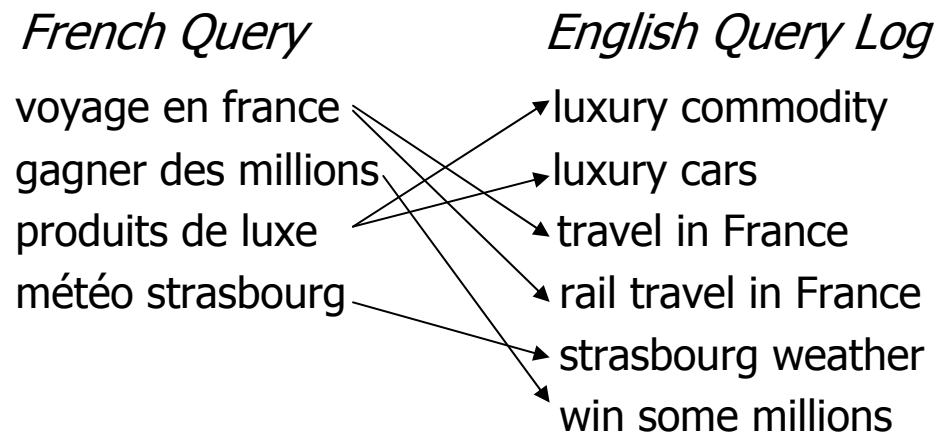


Cross-Lingual Query Suggestion (CLQS)

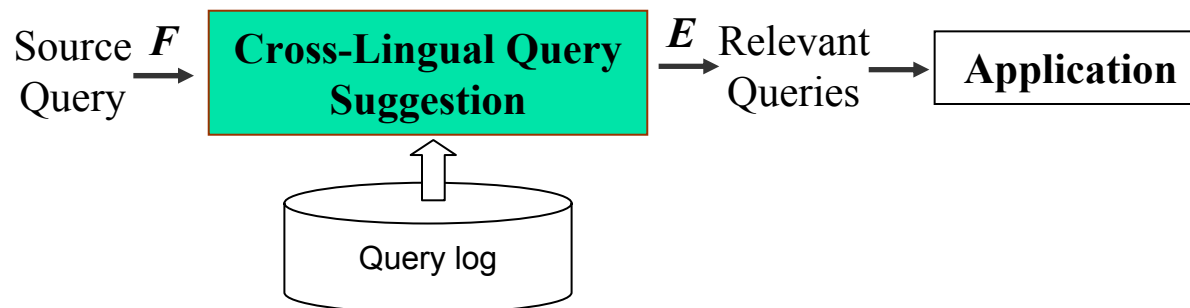
- Cross-Lingual Query Suggestion (CLQS): suggests related queries, but in a different language.
 - Example: (French) terrorisme international → (English) international terrorism, world terrorism, what is terrorism, terrorist attacks, terrorist groups, september 11, ...

CLQS by Query Log Mining

- A query in a source language is likely to have correspondents in the query log of the target language.



- CLQS based on mining query logs of difference languages



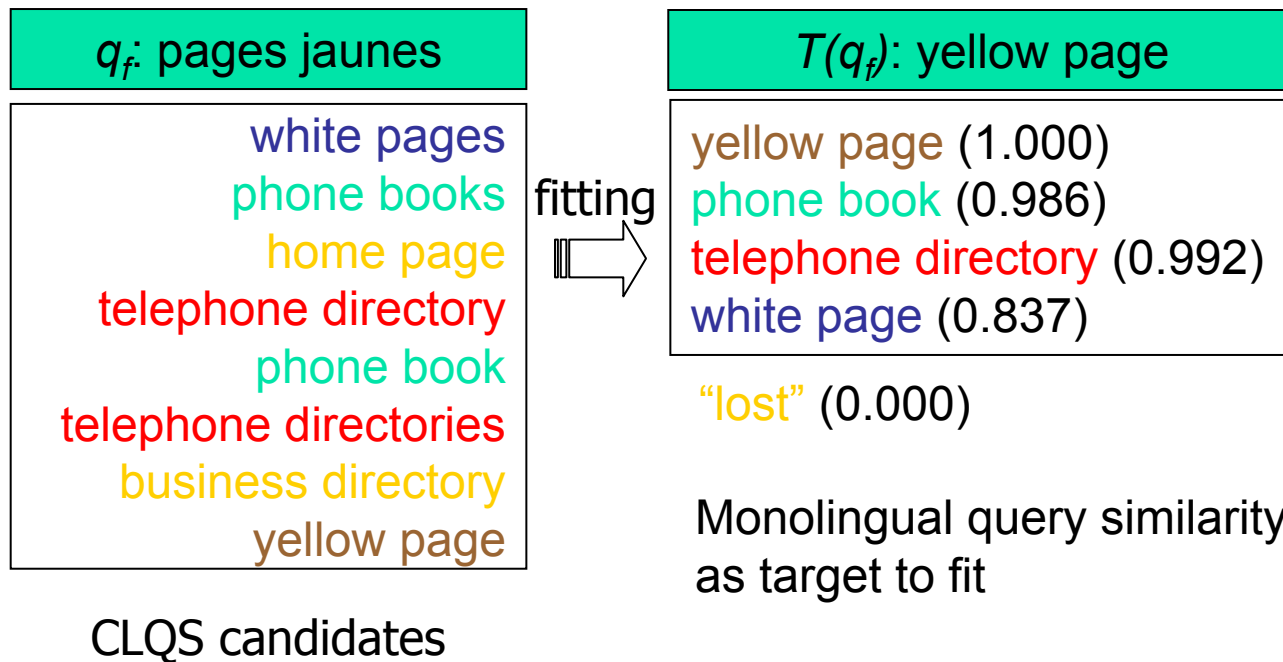


Outline

- Introduction
- Discriminative Model for Cross-Lingual Query Suggestion (CLQS)
- Mono-/Cross-Lingual Features
- CLIR with CLQS
- Performance Evaluation
- Conclusions

Principled Approach to Cross-lingual Similarity Estimation

- Central Task: Define and estimate cross-lingual query similarity
- Principled approach to similarity estimation: the cross-lingual similarity is equal to the monolingual similarity between the target query and the source query's translation.





SVM Regression for CLQS

- Regression model for learning the cross-lingual query similarity function.

$$sim_{CL}(q_f, q_e) = sim_{ML}(T_{q_f}, q_e)$$

- Advantages:

1. Only need a list of manually created query-translation pairs.
2. Used to fit any proper monolingual query similarity measure.
3. A principled way to integrate multiple features.

- SVM Regression:

$$sim_{CL}(q_f, q_e) = \bar{w} \cdot \bar{\phi}(f(q_f, q_e))$$



Outline

- Introduction
- Discriminative Model for Cross-Lingual Query Suggestion (CLQS)
- **Mono-/Cross-Lingual Features**
- CLIR with CLQS
- Performance Evaluation
- Conclusions



Monolingual Query Similarity

- Monolingual query similarity
 - Combining both query content-based similarity and click-based similarity, estimated from query log (Wen et al., *ACM TOIS*, 2002).

$$sim_{ML}(p, q) = \lambda * sim_{content}(p, q) + (1 - \lambda) * sim_{click-through}(p, q)$$

$$sim_{content}(p, q) = \frac{KN(p, q)}{Max(kn(p), kn(q))}$$

KN(x,y): # of query words in common;
kn(x): # of query words in x

$$sim_{click-through}(p, q) = \frac{RD(p, q)}{Max(rd(p), rd(q))}$$

RD(x,y): # of common URLs;
rd(x): # of clicked URLs of x



Cross-Lingual Features

- Queries q_f and q_e are bilingually similar if
 - they are translatable by a bilingual dictionary.
 - they are statistically associated in word-aligned parallel data.
 - their query words co-occur frequently on eb pages.
 - q_e is monolingually similar with queries generated as above.

Cross-Lingual Features (1): Dictionary-based Translation with Disambiguation

- Dictionary-based translation disambiguation using word co-occurrence statistics

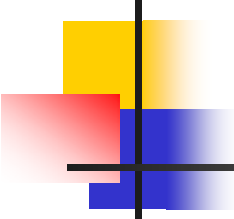
$$q_f = \{w_{f1}, w_{f2}, \dots, w_{fn}\} \quad T(w_{fi}) = \{t_{i1}, t_{i2}, \dots, t_{im}\}$$

$$MI(t_{ij}, t_{kl}) = P(t_{ij}, t_{kl}) \log \frac{P(t_{ij}, t_{kl})}{P(t_{ij})P(t_{kl})}$$

$$S_{dict}(q_f, T_{q_f}) = \sum_{i,k, i \neq k} MI(t_{ij}, t_{kl})$$

q_f :	la	seconde	guerre	mondiale
	an	<u>second</u>	<u>war</u>	<u>world</u>
	<u>the</u>	<u>II</u>	wartime	worldwide
	a		quarrel	war
	it		warfare	

- Use top-4 translations to retrieve target queries, and assign corresponding translation scores.



Cross-Lingual Features (2): Translation Score based on Parallel Corpora

- Parallel corpus as an assistant bilingual resource.
 - Word alignment optimization: GIZA++ (Och and Ney, 2003)
 - Given q_f , retrieve queries from the log containing the aligned words of q_f .
 - Associate the candidate queries with bi-directional similarity score based on IBM model 1 (Brown et al., 1993) translation probability:

$$S_{IBM1}(q_f, q_e) = \sqrt{P_{IBM1}(q_f | q_e) \cdot P_{IBM1}(q_e | q_f)}$$

$$P_{IBM1}(t | s) = \frac{1}{(|s| + 1)^{|t|}} \prod_{j=1}^{|t|} \sum_{i=0}^{|s|} p(t_j | s_i) = \frac{1}{(|s| + 1)^{|t|}} \prod_{j=1}^{|t|} \sum_{i=0}^{|s|} \frac{C(t_j, s_i) + \delta}{C(s_i) + \delta N}$$



Cross-Lingual Features (3): Online Mining for Related Queries

- If a target-language query often co-occurs with the source-language query in many web pages, they are likely to be semantically related.

[La Vie En Rose - Julien and Sophie of **Jeux D'enfants**](#)

The TFL approved fanlisting for the relationship between Julien and Sophie from the movie

Jeux D'enfants (Love Me If You Dare)

[julienophie.boorns.net/](#) - 3k - [Cached](#) - [Similar pages](#)

[Musées de France - Les **jeux d'enfants** de Bruegel](#)

Les **jeux d'enfants** de Bruegel. « Back to list. Les **jeux d'enfants** de Bruegel. Réf. Prix. JA104792, € 14.50, Add to cart. Prices shown include VAT. ...

[www.museesdefrance.com/produits/details/JA104792](#) - 10k - [Cached](#) - [Similar pages](#)

[**Jeux D'Enfants** - Love Me If You Dare Photos - Yahoo! Movies UK](#)

Photo from the film **Jeux D'Enfants** - Love Me If You Dare.

[uk.movies.yahoo.com/j/Jeux-DEnfants-Love-Me-If-You-Dare/photos-46957-46958.html](#) - 7k -

[Cached](#) - [Similar pages](#)

[**Jeux D'Enfants** - Love Me If You Dare Movie - Yahoo! Movies UK](#)

Jeux D'Enfants - Love Me If You Dare movie information including cast & crew details, photos and movie trailers from Yahoo! Movies UK.

[uk.movies.yahoo.com/j/Jeux-DEnfants-Love-Me-If-You-Dare/index-46958.html](#) - 15k -

[Cached](#) - [Similar pages](#)

[Video: **Jeux d'enfants** - Dans le coeur d'une fée](#)

View "**Jeux d'enfants**" at YouTube. P'tite fée' s rating: ★. Tags: jeux,; d'enfants. This also appears in: QdJ: Si j'étais actrice. ...

[laptitefee.vox.com/library/video/6a00d10a7ad4938bfa00cd971b98f84cd5.html](#) - 33k -

[Cached](#) - [Similar pages](#)

[Love Me If You Dare \(**Jeux d'enfants**\) - Movie Reviews, Photos ...](#)

Love Me If You Dare (**Jeux d'enfants**), Movie Ratings & Reviews, Cast & Crew, Clips & Videos, Posters & Gallery, Layouts & Lists, Fan Club & Showtimes.

[www.flixster.com/movie/love-me-if-you-dare-jeux-denfants](#) - 69k - [Cached](#) - [Similar pages](#)



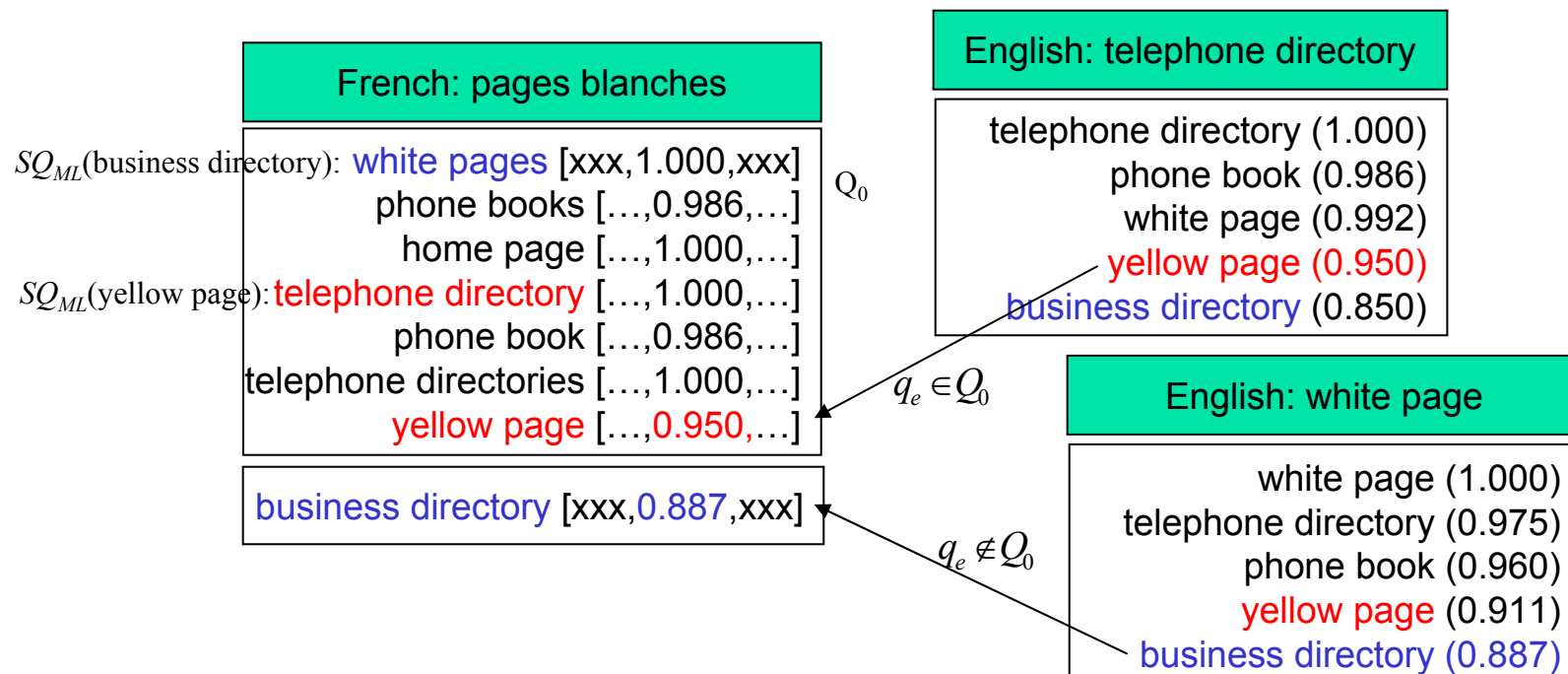
Cross-Lingual Features (3): Online Mining for Related Queries (cont')

- Format bilingual search queries:
 - la seconde guerre mondiale: (la seconde guerre mondiale) *AND* (the *OR* la *OR* a *OR* it) *AND* (second *OR* II) *AND* (war *OR* wartime *OR* quarrel *OR* warfare) *AND* (world *OR* worldwide *OR* war)
- Co-Occurrence Double Checking (CODC): two objects a and b are considered to have association if b can be found by using a as query, and vice versa (Chen et al., ACL, 2006).

$$S_{CODC}(q_f, q_e) = \begin{cases} 0, & \text{if } freq(q_e @ q_f) \cdot freq(q_f @ q_e) = 0 \\ e^{\log \left[\frac{freq(q_e @ q_f) \cdot freq(q_f @ q_e)}{freq(q_f) \cdot freq(q_e)} \right]^\alpha}, & \text{otherwise} \end{cases}$$

Cross-Lingual Features (4): Monolingual Query Suggestion

- Further improve the recall of CLQS for a given set of target candidate queries Q_0 expanded by using monolingual query suggestion.





Recap: SVM Regression for CLQS

- Regression model for learning the cross-lingual query similarity function.

$$sim_{CL}(q_f, q_e) = sim_{ML}(T_{q_f}, q_e)$$

- Advantages:

1. Only need a list of manually created query-translation pairs.
2. Used to fit any proper monolingual query similarity measure.
3. A principled way to integrate multiple features.

- SVM Regression:

$$sim_{CL}(q_f, q_e) = \bar{w} \cdot \bar{\phi}(f(q_f, q_e))$$



Outline

- Introduction
- Discriminative Model for Cross-Lingual Query Suggestion (CLQS)
- Mono-/Cross-Lingual Features
- **CLIR with CLQS**
- Performance Evaluation
- Conclusions



CLIR with CLQS

- Besides as a standalone system, CLQS can be leveraged to supported CLIR
 - Given q_f , compute CLQS $\{q_e\}$
 - For each q_e , perform monolingual IR based on BM25 model (Robertson et al., 1995)
 - Documents are merged and re-ranked by the sum of BM25 scores



Outline

- Introduction
- Discriminative Model for Cross-Lingual Query Suggestion (CLQS)
- Mono-/Cross-Lingual Features
- CLIR with CLQS
- Performance Evaluation
- Conclusions



Performance Evaluation

- Data Resources

- CLIR: F-to-E
- 1-month MSN English query log, 7.29M queries, freq>5
- 4,171 manually created F-E query-translation pairs, 70% as training set, 20% as test set, 10% as development set
- Bilingual dictionary: 120,000 F-E entries
- Europarl F-E parallel corpus (<http://people.csail.mit.edu/koehn/publications/europarl>)
- CLIR Benchmark collection: TREC6 CLIR French-English dataset
 - Document set: AP88-90 newswire, 750MB
 - Query set: 25 F-E queries pairs (CL#01-25), avg. length=3.3

Objective CLQS Performance

- Benchmark CLQS by comparing with Monolingual Query Suggestion (MLQS)

- Mean-square-error (MSE) of SVM Regression

$$MSE = \frac{1}{l} \sum_i \left[sim_{CL}(q_{fi}, q_{ei}) - sim_{ML}(T_{q_{fi}}, q_{ei}) \right]^2$$

- Classification precision (P) and recall (R)

$$P = \frac{|S_{CLQS} \cap S_{MLQS}|}{|S_{CLQS}|} \quad R = \frac{|S_{CLQS} \cap S_{MLQS}|}{|S_{MLQS}|}$$

- CLQS performance with different feature settings

Features	Regression	Classification	
	MSE	Precision	Recall
DD	0.274	0.732	0.098
DD+PC	0.224	0.713	0.125
DD+PC+Web	0.115	0.808	0.192
DD+PC+Web+MLQS	0.174	0.796	0.421

DD: dictionary only;

DD+PC: dictionary and parallel corpora;

DD+PC+Web: dictionary, parallel corpora, and web mining;

DD+PC+Web+MLQS: dictionary, parallel corpora, web mining, and monolingual query suggestion



Subjective CLQS Performance

- Human subjective test on CLQS relevancy
 - 200 French queries from French log not in training examples
 - 1,727 English queries are produced by the model, avg. 8.7 suggestions per query.
 - 1,407 are recognized as relevant. Accuracy=80.9%
 - An example (CL14): “terrorisme international” (international terrorism)

International terrorism (0.991); what is terrorism (0.943); counter terrorism (0.920); terrorist (0.911); terrorist attack (0.898); international terrorist (0.853); world terrorism (0.845); global terrorism (0.833); transnational terrorism (0.821); human rights (0.811); terrorist groups (0.777); patterns of global terrorism (0.762); september 11 (0.734)



CLIR Performance using CLQS

- For comparisons, we run 4 experiments using
 - CLQS-based CLIR (all features)
 - MT-based query translation (MT): a commercial F-to-E MT system, i.e. Google's translation tool (http://www.google.com/language_tools)
 - Dictionary-based query translation (DT): implementation of translation disambiguation based on co-occurrence statistics (Ballestors and Croft, 1998)
 - Post-translation expansion (Ballestor and Croft, 1997; McNamee and Mayfield, 2002) based on pseudo relevance feedback (PRF) using the output of CLQS, MT and DT. PRF takes top 10 terms from top 25 retrieved documents.

CLIR Performance without Post-translation Expansion

- Average precision

IR method	Average Precision	% of Monolingual IR
Monolingual	0.266	100%
MT	0.217	81.6%
DT	0.186	69.9%
CLQS	0.233	87.6%

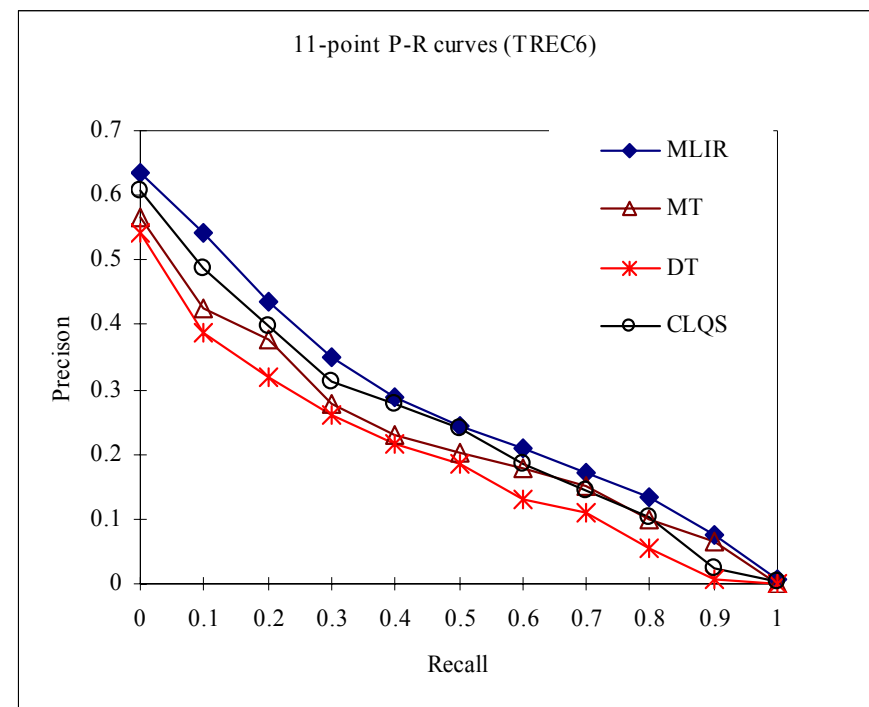
Monolingual: monolingual IR;

MT: CLIR based on machine translation;

DT: CLIR based on dictionary translation;

CLQS: CLQS-based CLIR

- 11-point interpolated precision-recall curve



11-point P-R curve without post-translation expansion

CLIR Performance with Post-translation Expansion

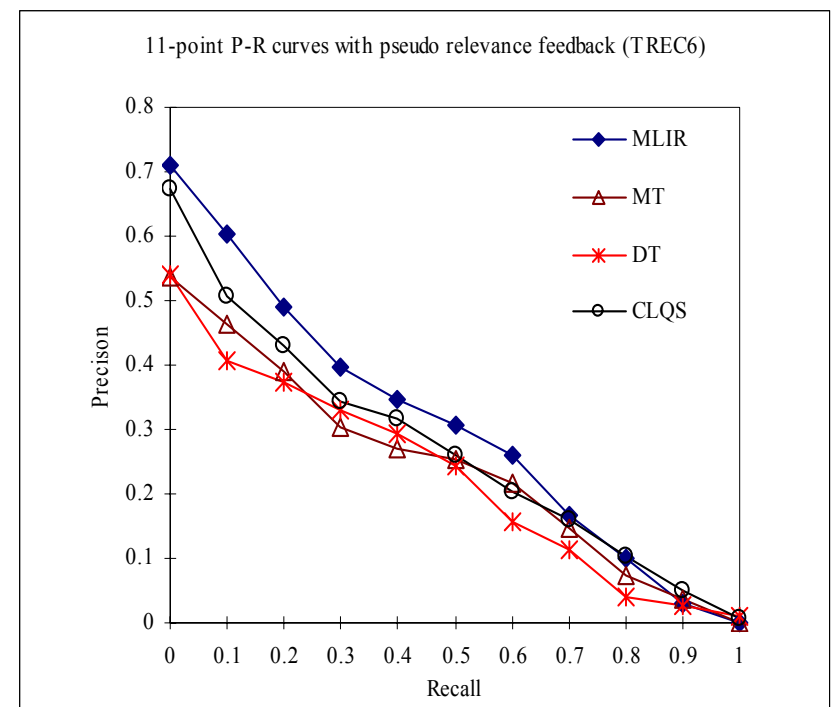
- Average precision comparisons before and after post-translation expansion

IR method	AP without PRF	AP with PRF	Change
Monolingual	0.266 (100%)	0.288 (100%)	+8.27%
MT	0.217 (81.6%)	0.222 (77.1%)	+2.30%
DT	0.186 (69.9%)	0.220 (76.4%)	+18.3%
CLQS	0.233 (87.6%)	0.259 (89.8%)	+11.2%

- Significant test (t-test): p-value < 0.05 is regarded significant

	MT	DT	MT+PRF	DT+PRF
CLQS	0.0298	3.84e-05	0.1472	0.0282
CLQS+PRF	0.0026	2.63e-05	0.0094	0.0016

- 11-point P-R curve with post-translation expansion





Outline

- Introduction
- Discriminative Model for Cross-Lingual Query Suggestion (CLQS)
- Mono-/Cross-Lingual Features
- CLIR with CLQS
- Performance Evaluation
- **Conclusions**



Conclusions

- Summary:

- Present a principled approach to estimate cross-lingual query similarity
- Build a CLQS system by mining query logs of different languages
- CLIR based on CLQS significantly out-perform other approaches
- CLQS and post-translation expansion are complementary to CLIR

- Future Work:

- Investigate CLQS between language pairs which are loosely correlated