Justsystem-Clairvoyance CLIR Experiments

NTCIR-4 Workshop
Tokyo, Japan

Yan Qu, Gregory Grefenstette, David A. Hull, David A. Evans, Toshiya Ueda, Tatsuo Kato, Daisuke Noda, Motoko Ishikawa, Setsuko Nara, Kousaku Arita

Clairvoyance Corporation, USA

Justsystem Corporation, Japan

June 2, 2004

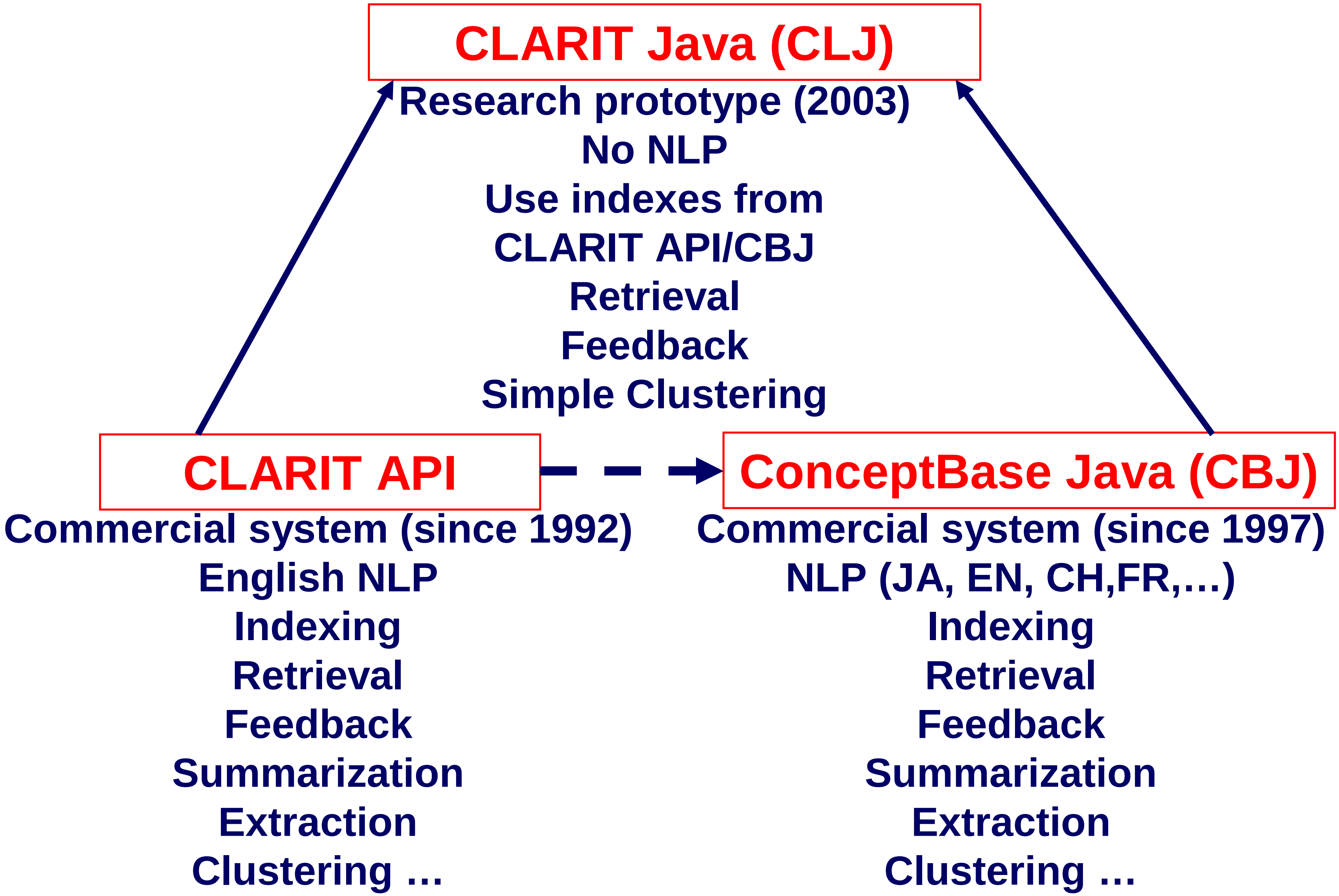
Overview of Participation

- **Clairvoyance (USA) and Justsystem (Japan) collaboration**
- **Single Language IR (SLIR)**
  - Japanese–Japanese
  - Chinese–Chinese
  - English–English
- **Bilingual CLIR (BLIR)**
  - Japanese–English
  - Chinese–English
- **Goal of participation:**
  - Evaluating performance and robustness of commercial-grade CLIR systems for English, Japanese, and Chinese

System Description

- **System framework**
  - Clairvoyance IM APIs (CLARIT)
  - Justsystem ConceptBase Java (CBJ)
- **Functionalities**
  - Natural language processing
  - Ad hoc retrieval
  - Feedback
  - Visualization
  - Cross-language IR
  - Etc.

Systems



CLARIT Java (CLJ)

Research prototype (2003)
No NLP
Use indexes from
CLARIT API/CBJ
Retrieval
Feedback
Simple Clustering

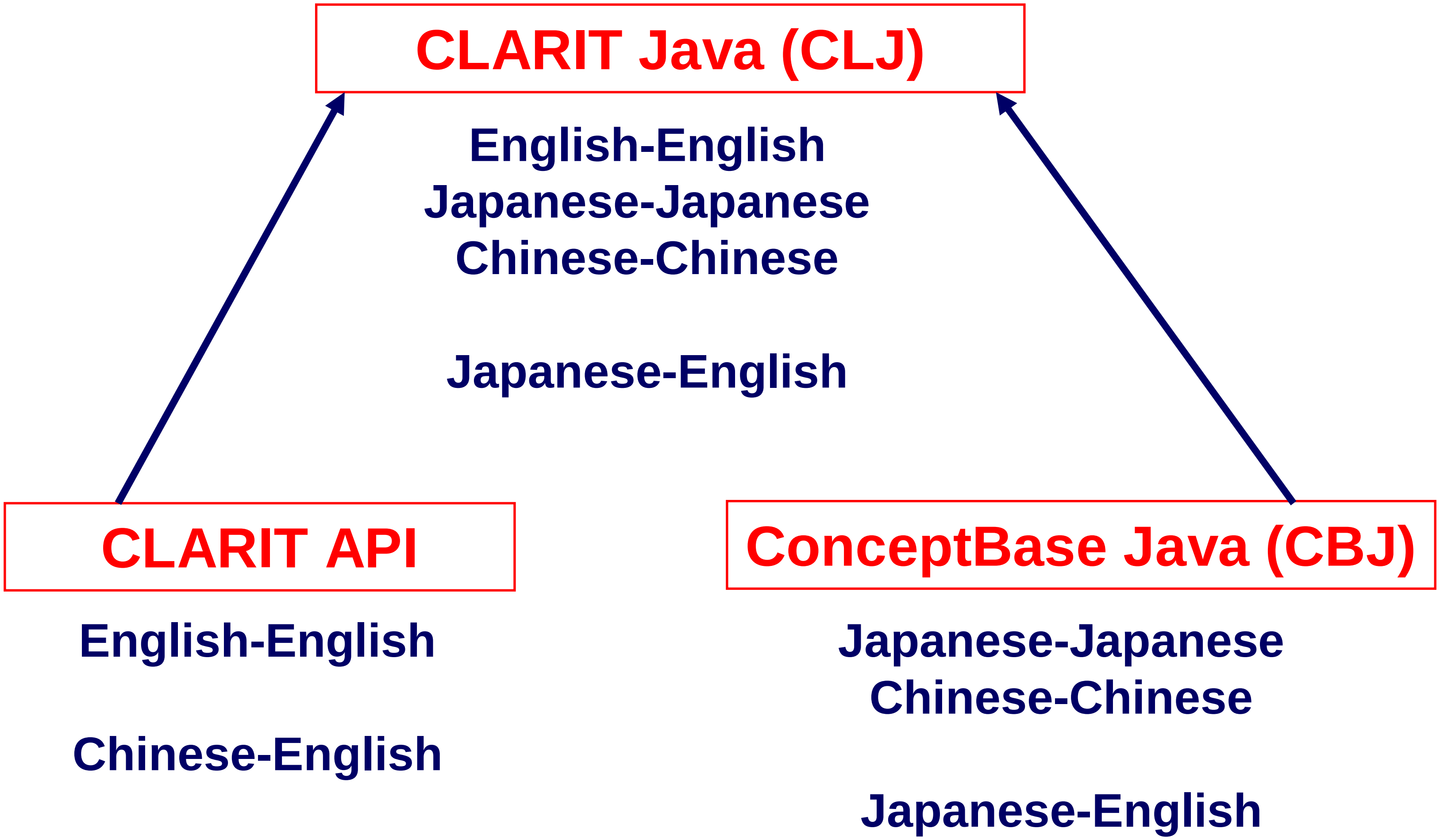
CLARIT API

Commercial system (since 1992)
English NLP
Indexing
Retrieval
Feedback
Summarization
Extraction
Clustering …

ConceptBase Java (CBJ)

Commercial system (since 1997)
NLP (JA, EN, CH,FR,…)
Indexing
Retrieval
Feedback
Summarization
Extraction
Clustering …

Submission Distributions



CLARIT Java (CLJ)

English-English
Japanese-Japanese
Chinese-Chinese

Japanese-English

CLARIT API

English-English

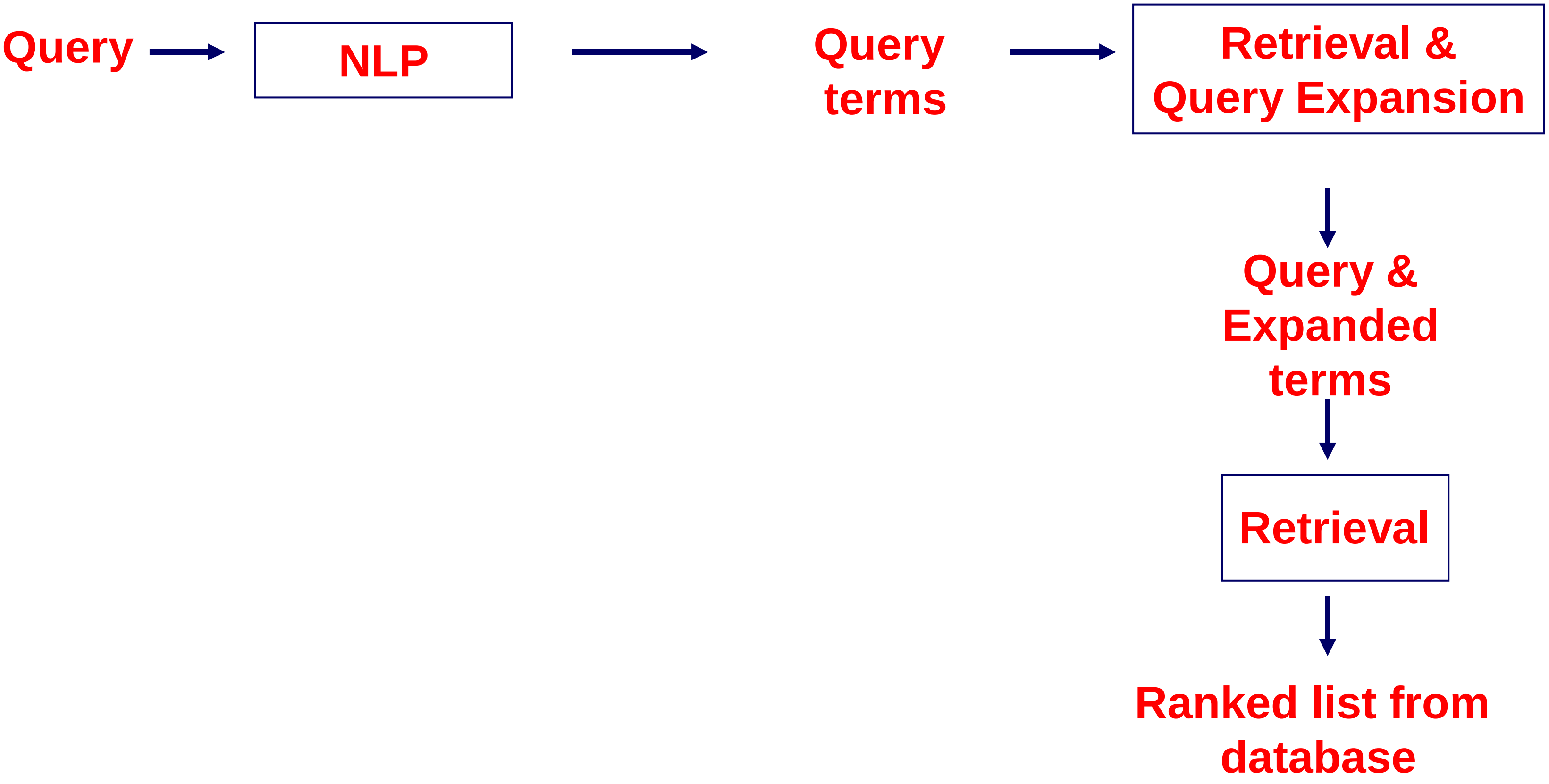
Chinese-English

ConceptBase Java (CBJ)

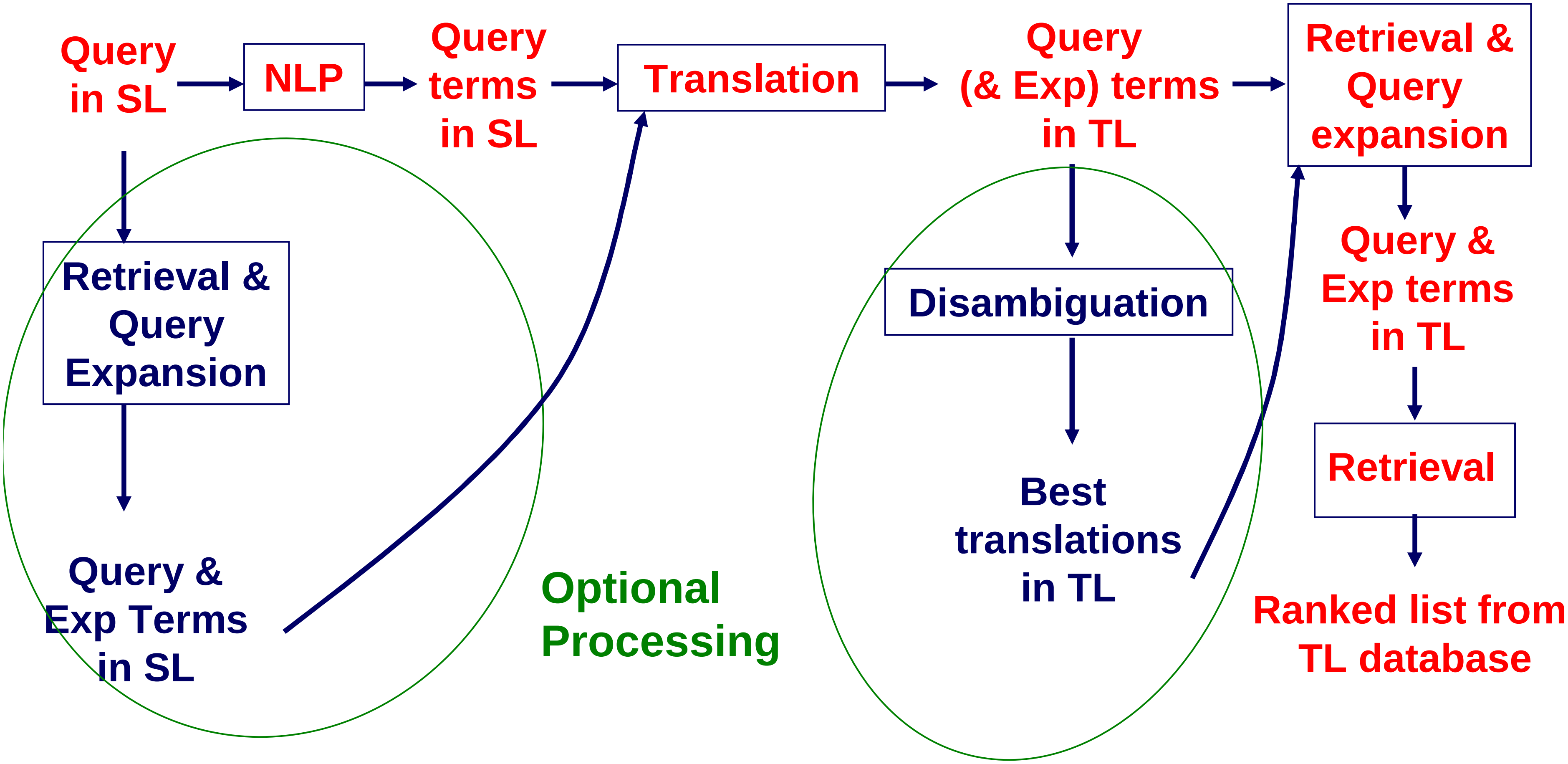
Japanese-Japanese
Chinese-Chinese

Japanese-English

Single-Language IR (SLIR)



Bilingual CLIR (BLIR)



Query in SL

NLP

Query terms in SL

Translation

Query (& Exp) terms in TL

Retrieval & Query expansion

Retrieval & Query Expansion

Query & Exp Terms in SL

Optional Processing

Disambiguation

Best translations in TL

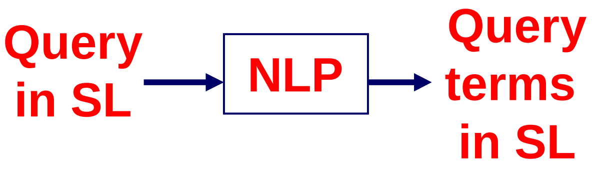
Query & Exp terms in TL

Retrieval

Ranked list from TL database

SL: source language TL: target language

Bilingual CLIR (BLIR)

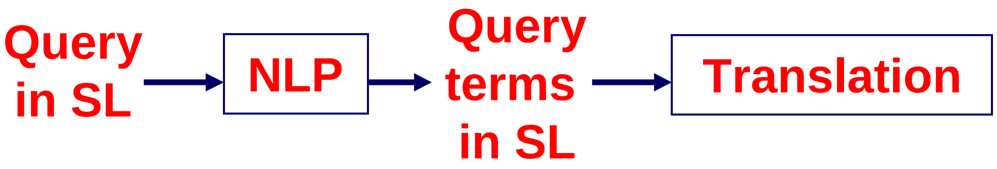


Word segmentation for Chinese/Japanese

- Statistical part of speech tagging
- NLP for phrase identification

SL: source language TL: target language

Bilingual CLIR (BLIR)


Query in SL → NLP → Query terms in SL → Translation

Missing translation

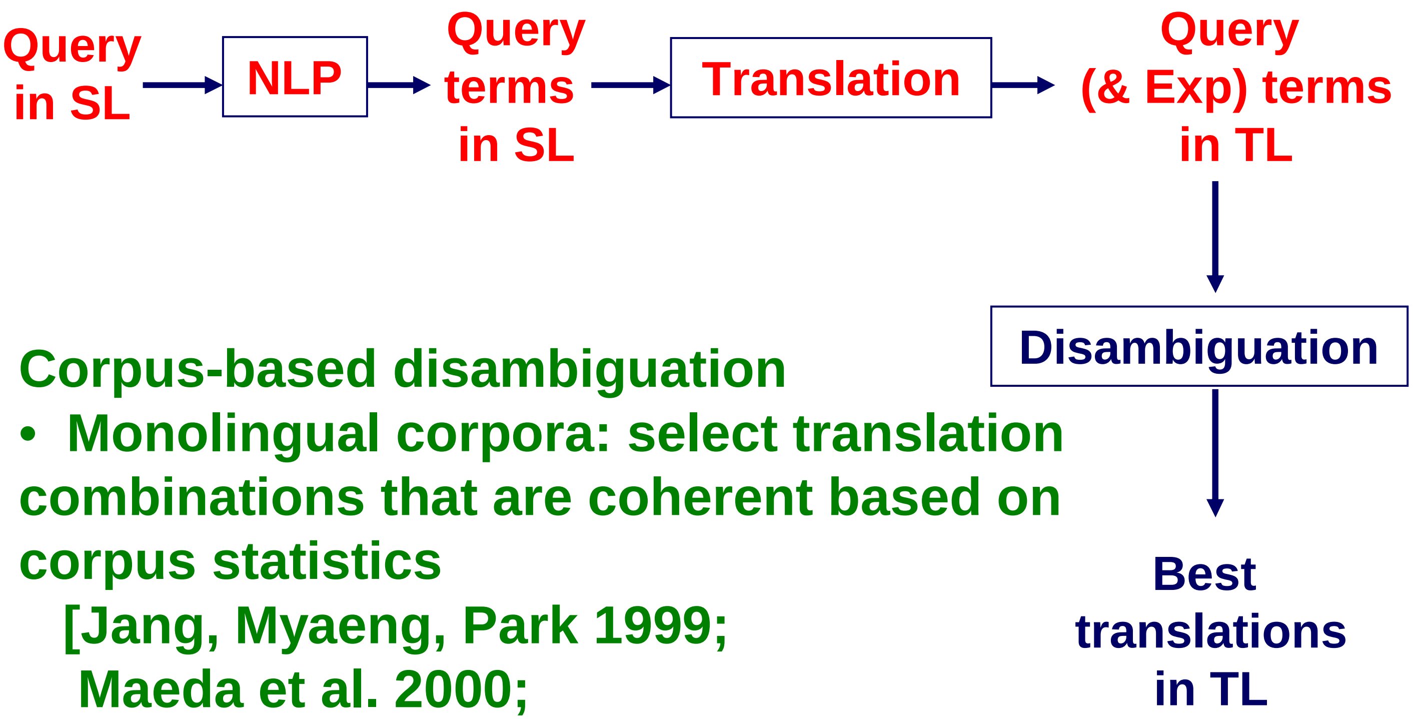
- Proper name translation (transliteration)
- Multi-word terms
- General unknown terms

Methods

- Automatic name transliteration
  [Fujii & Ishikawa 2001; Qu, Grefenstette, Evans 2003]
- Corpus-based mining of translations
  pre-translation feedback [Ballesteros & Croft 1996]
  context vectors [Fung & Yee 1997]
- Translation of multi-word terms [Grefenstette 1999]

SL: source language TL: target language

Bilingual CLIR (BLIR)

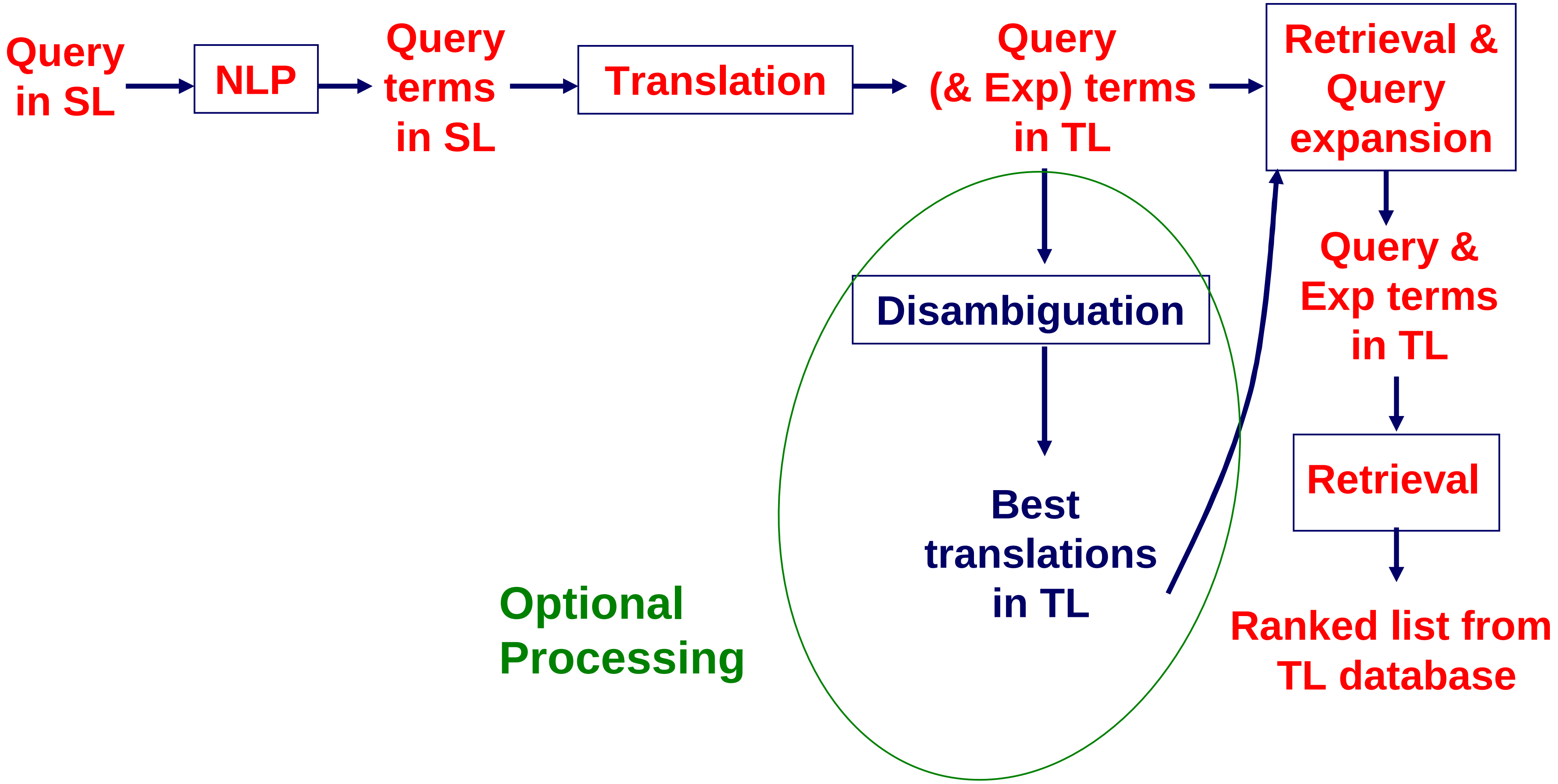


Corpus-based disambiguation

- **Monolingual corpora: select translation combinations that are coherent based on corpus statistics**
  [Jang, Myaeng, Park 1999;
  Maeda et al. 2000;
  Qu, Grefenstette, Evans 2002]
- **Parallel corpora: select translations found in aligned pairs**

SL: source language TL: target language

Bilingual CLIR (BLIR)



Query in SL

NLP

Query terms in SL

Translation

Query (& Exp) terms in TL

Retrieval & Query expansion

Disambiguation

Best translations in TL

Optional Processing

Query & Exp terms in TL

Retrieval

Ranked list from TL database

SL: source language TL: target language

Indexing and Retrieval

- **Indexes based on sub-documents**
- **Vector space model**

“importance coefficient” for phrase down-weighting

$$sim(Q,\ D) = \sum_{t \in Q \cap D} W_Q(t) \cdot W_D(t).$$

term frequency

Inverse document frequency

$$W_Q(t) = C(t) \cdot TF_Q(t) \cdot IDF(t)$$

$$W_D(t) = TF_D(t) \cdot IDF(t)$$

term freq smoothing factor

$$W_D(t) = \frac{(k_1 + 1) \ast TF_D(t)}{k_1[(1 - b) + b \ast (d/\Delta d/ + TF_D(t)}$$

BM25

doc length smoothing factor **doc length** **average doc length**

Query Expansion

Rocchio

$$Rocchio(t) = IDF(t) \bullet \frac{\sum_{D \in DocSet} TF_D(t)}{NumDoc}$$

IDF(t) **= the inverse document frequency of term *t***
NumDoc **= the number of documents in the given set of documents**
TF(t) **= the term frequency score for term *t* in document *D***

Query Expansion

Prob2

$$Prob2(t) = log(R_t + 1) \; x \left(log(\frac{N - R + 2}{N_t - R_t + 1} - 1) - log(\frac{R + 1}{R_t} - 1) \right)$$

N = the number of documents in the target corpus

N$_t$ = the number of documents in the corpus that contain term ***t***

R = the number of documents for feedback that are (presumed to be) relevant to the topic

R$_t$ = the number of documents that are (presumed to be) relevant to the topic and contain term ***t***

Query Expansion

Term Weighting

$$Q_{new} = k \times Q_{orig} + Q_{exp}$$

- **Term weight options for Q_{exp}**
  - Uniform: $W(t) = 1.0$ for all terms
  - Normalized: $W(t) = W_{prob2}(t) \, / \max W_{prob2}(t)$
  - Scaled: $W(t) = W_{prob2}(t) \, / \, \Sigma \, W_{prob2}(t)$

    applies to Q_{exp} and Q_{orig} (using original weights)
- **Parameters in green are tunable**

Example Prob2 Weights

Topic 5: PRC's economic reform (Ntcir-3) **Uniform**

Prob2 expansion: ndocs = 10, ntrms = 10

k=1, Q_{orig} query weight = 2.0, Q_{exp} expansion term weight = 1.0

			Query Merge Strategy	
	Orig	Prob2	Unif	K Q_{orig} Q_{exp}
prc	1.0	12.66	3.0	←1*2.0 + 1.0 = 3.0
reform	1.0	9.66	3.0	←1*2.0 + 1.0 = 3.0
economic_reform	1.0	8.94	3.0	←1*2.0 + 1.0 = 3.0
economic	1.0	8.55	3.0	←1*2.0 + 1.0 = 3.0
future_status	0.0	7.98	1.0	←1*0.0 + 1.0 = 1.0
political_reform	0.0	7.50	1.0	←1*0.0 + 1.0 = 1.0
political	0.0	6.70	1.0	←1*0.0 + 1.0 = 1.0
accordance	0.0	5.97	1.0	←1*0.0 + 1.0 = 1.0
strait	0.0	5.14	1.0	←1*0.9 + 1.0 = 1.0
continue	0.0	4.64	1.0	←1*0.0 + 1.0 = 1.0

Example Prob2 Weights

Topic 5: PRC's economic reform (Ntcir-3) **Normalized**

Prob2 expansion: ndocs = 10, ntrms = 10

k=0.6, Q_{orig} query weight = 2.0, Q_{exp} expansion term weight = prob2/12.66

			Query Merge Strategy	
	Orig	Prob2	**Scale**	**K Q_{orig} Q_{exp}**
prc	1.0	12.66	**3.00**	**←1*2.0 + 1.00 = 3.00**
reform	1.0	9.66	**2.76**	**←1*2.0 + 0.76 = 2.76**
economic_reform	1.0	8.94	**2.71**	**←1*2.0 + 0.71 = 2.71**
economic	1.0	8.55	**2.68**	**←1*2.0 + 0.68 = 2.68**
future_status	0.0	7.98	**0.63**	**←1*0.0 + 0.63 = 0.63**
political_reform	0.0	7.50	**0.59**	**←1*0.0 + 0.59 = 0.59**
political	0.0	6.70	**0.53**	**←1*0.0 + 0.53 = 0.53**
accordance	0.0	5.97	**0.47**	**←1*0.0 + 0.47 = 0.47**
strait	0.0	5.14	**0.41**	**←1*0.0 + 0.41 = 0.41**
continue	0.0	4.64	**0.37**	**←1*0.0 + 0.37 = 0.37**

Example Prob2 Weights

Topic 5: PRC's economic reform (Ntcir-3) **Scaled**

Prob2 expansion: ndocs = 10, ntrms = 10

k=0.6, Q_{orig} query weight = prob2 / 4
Q_{exp} expansion term weight = prob2 / 77.74

Query Merge Strategy

	Orig	Prob2	Scale	K ↓ Q_{orig} ↓ Q_{exp} ↓
prc	1.0	12.66	0.31	←0.6*0.25 + 0.16 = 0.31
reform	1.0	9.66	0.27	←0.6*0.25 + 0.12 = 0.27
economic_reform	1.0	8.94	0.26	←0.6*0.25 + 0.11 = 0.26
economic	1.0	8.55	0.26	←0.6*0.25 + 0.11 = 0.26
future_status	0.0	7.98	0.10	←0.6*0.00 + 0.10 = 0.10
political_reform	0.0	7.50	0.10	←0.6*0.00 + 0.10 = 0.10
political	0.0	6.70	0.09	←0.6*0.00 + 0.09 = 0.09
accordance	0.0	5.97	0.08	←0.6*0.00 + 0.08 = 0.08
strait	0.0	5.14	0.07	←0.6*0.00 + 0.07 = 0.07
continue	0.0	4.64	0.06	←0.6*0.00 + 0.06 = 0.06
	Sum = 4	Sum = 77.74		

Parameter Calibration

- **Based on NTCIR-3 topics and data collections**
- **Stable performance with parameter variation**

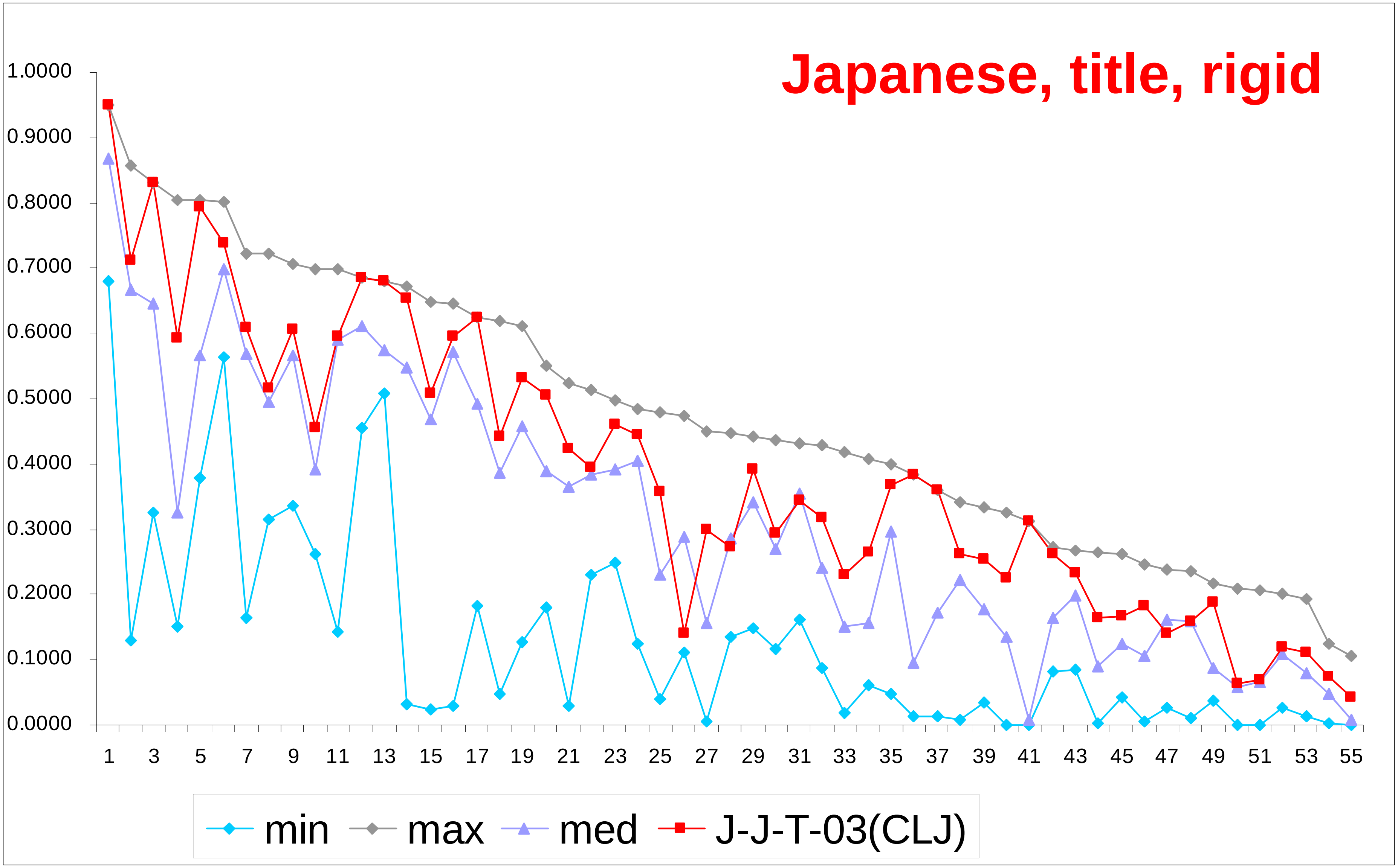
Parameters	Submission	Range
BM25 (k_1, b)	0.311	0.306–0.311
Phrase weight (0.0–1.0)	0.389	0.361–0.389
# docs (5–20) # terms (20–40)	0.389	0.361–0.391
Query weight (0.0–1.0)	0.389	0.366–0.389

NTCIR-4 Max-Med-Min Ranges

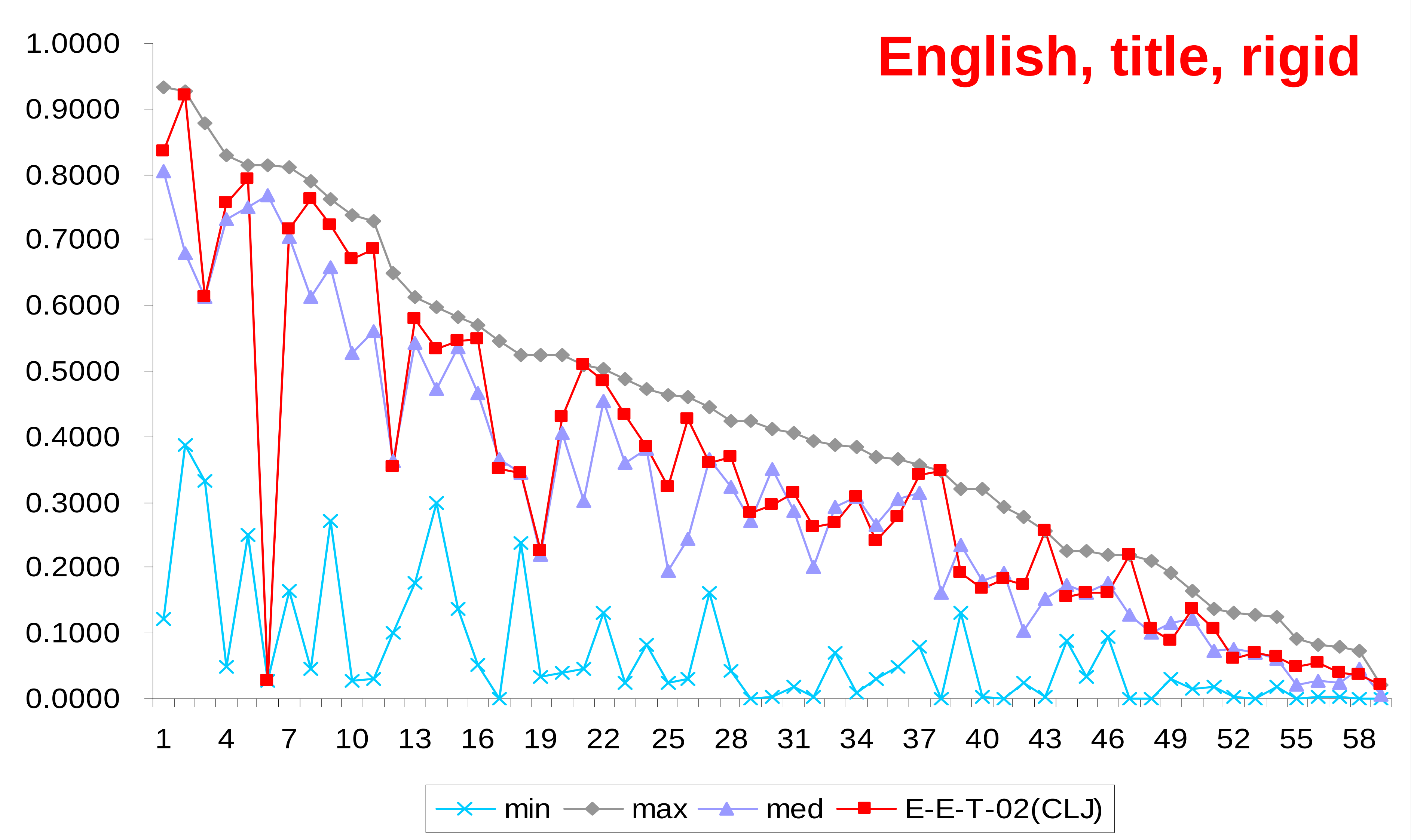
Single-Language IR Results, Rigid, Automatic

subtask	min	max	med	JSC-CC best
E-E-T	0.0802	0.3576	0.3145	0.3412
E-E-D	0.0342	0.3469	0.3026	0.3382
J-J-T	0.1966	0.3890	0.3135	0.3890
J-J-D	0.2130	0.3804	0.3352	0.3747
C-C-T	0.1327	0.3146	0.1881	0.1899
C-C-D	0.1251	0.3255	0.1741	0.1886

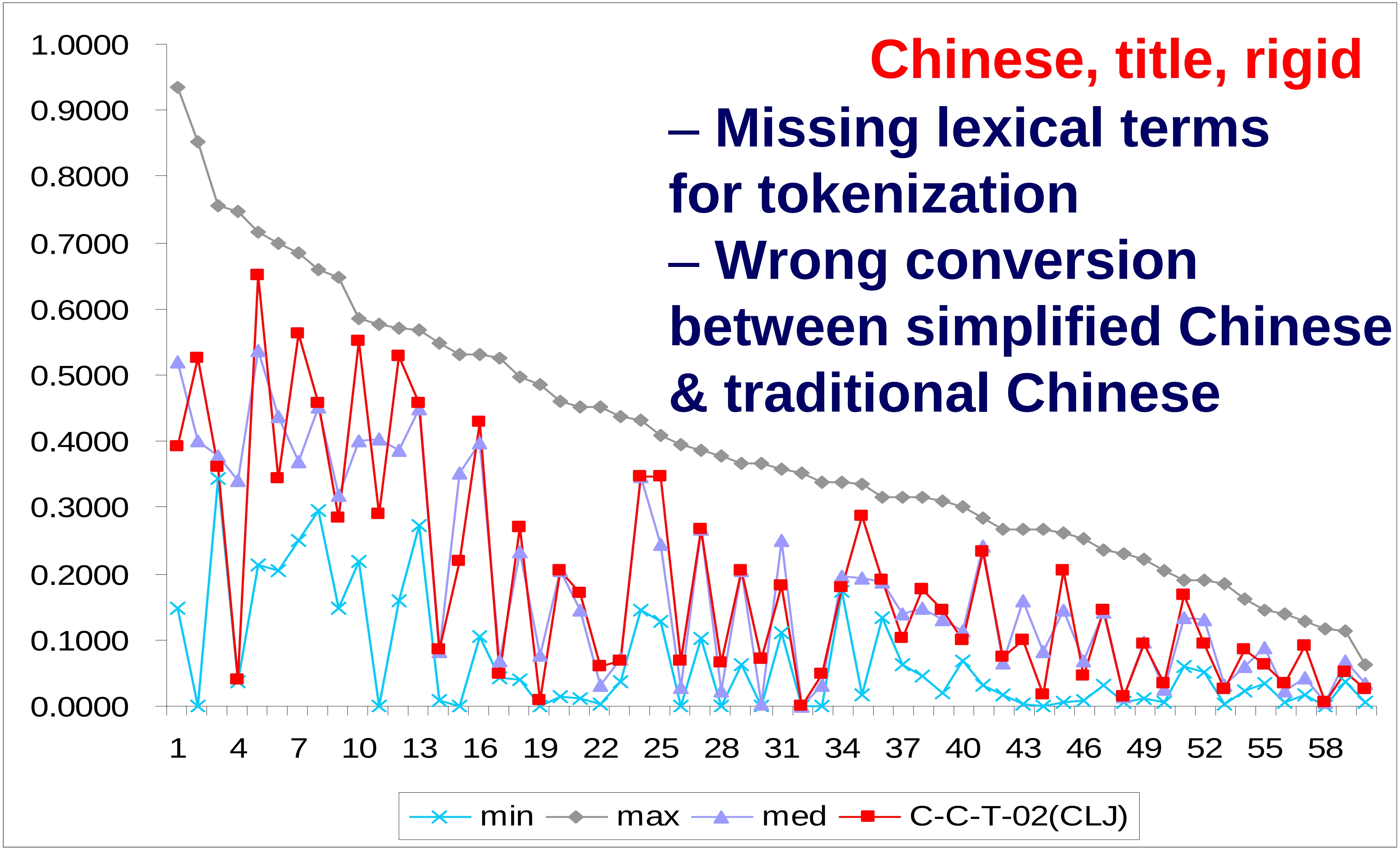
Japanese-Japanese Performance



English-English Performance



Chinese-Chinese Performance



Chinese, title, rigid

– Missing lexical terms for tokenization

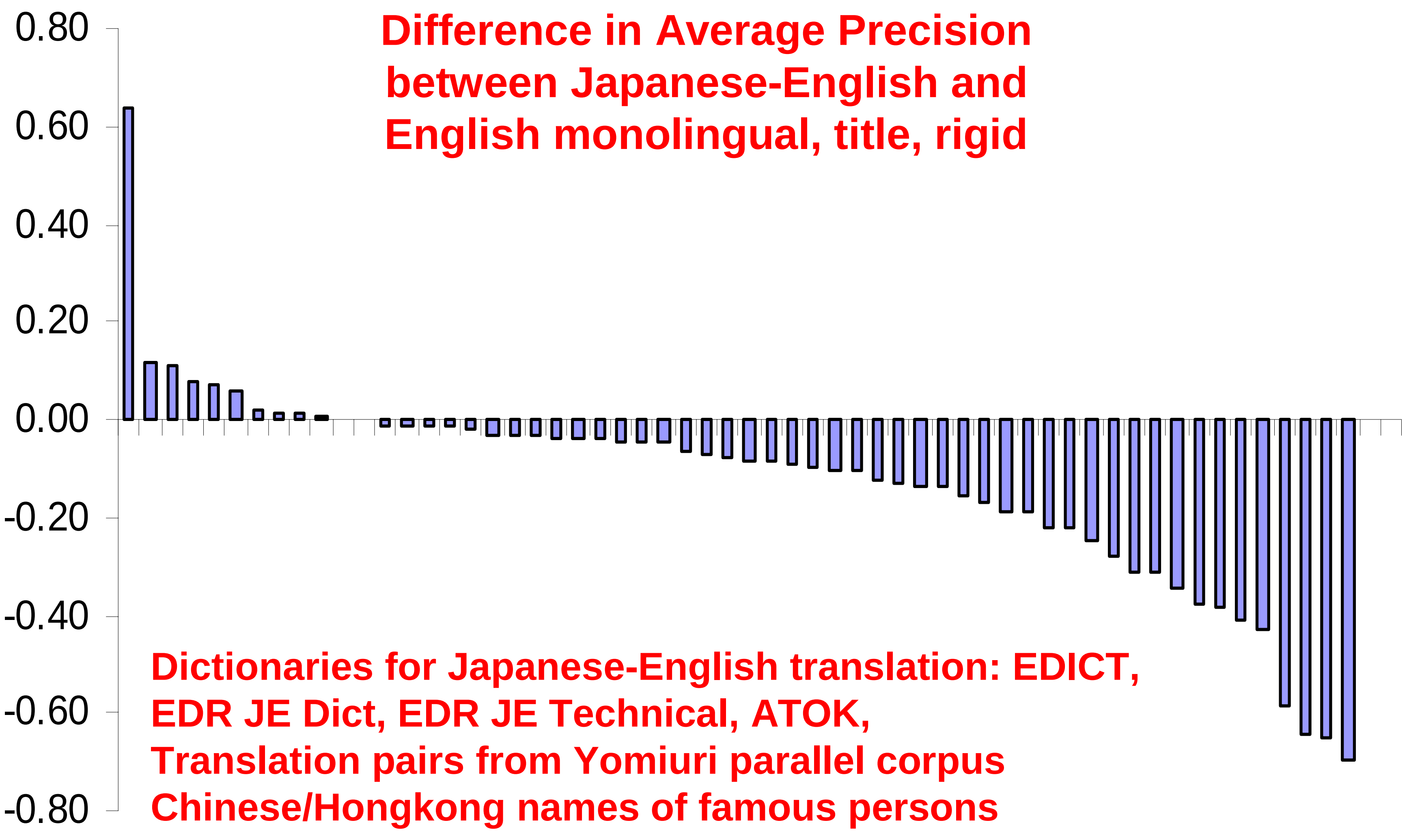
– Wrong conversion between simplified Chinese & traditional Chinese

NTCIR-4 Max-Med-Min Ranges

Bilingual CLIR Results, Rigid, Automatic

subtask	min	max	med	JSC-CC best
C-E-T	0.0389	0.2380	0.1860	0.1660
C-E-D	0.0412	0.2238	0.1819	0.1575
J-E-T	0.0189	0.3407	0.2131	0.2131
J-E-D	0.0075	0.3340	0.2427	0.2620

Japanese–English Performance



Difference in Average Precision between Japanese-English and English monolingual, title, rigid

Dictionaries for Japanese-English translation: EDICT, EDR JE Dict, EDR JE Technical, ATOK, Translation pairs from Yomiuri parallel corpus Chinese/Hongkong names of famous persons

Japanese–English: Error Analysis

- **Missing translations of proper nouns**

  Topic 10: 胡錦涛(Hu JinTao)

  Topic 22: 起亜(Kia)

- **Missing multi-word terms because of stop words**

  Topic 20: 西暦2000年問題(Y2K problem)

  ⇒ 西暦(A.D.), 2000 (問題 is a stop word)

- **Improper low distribution translations**

  Topic 37: 対応(correspondence)

  ⇒ "対応" has many translations and they contain improper and low distribution word, such as "pentagon".

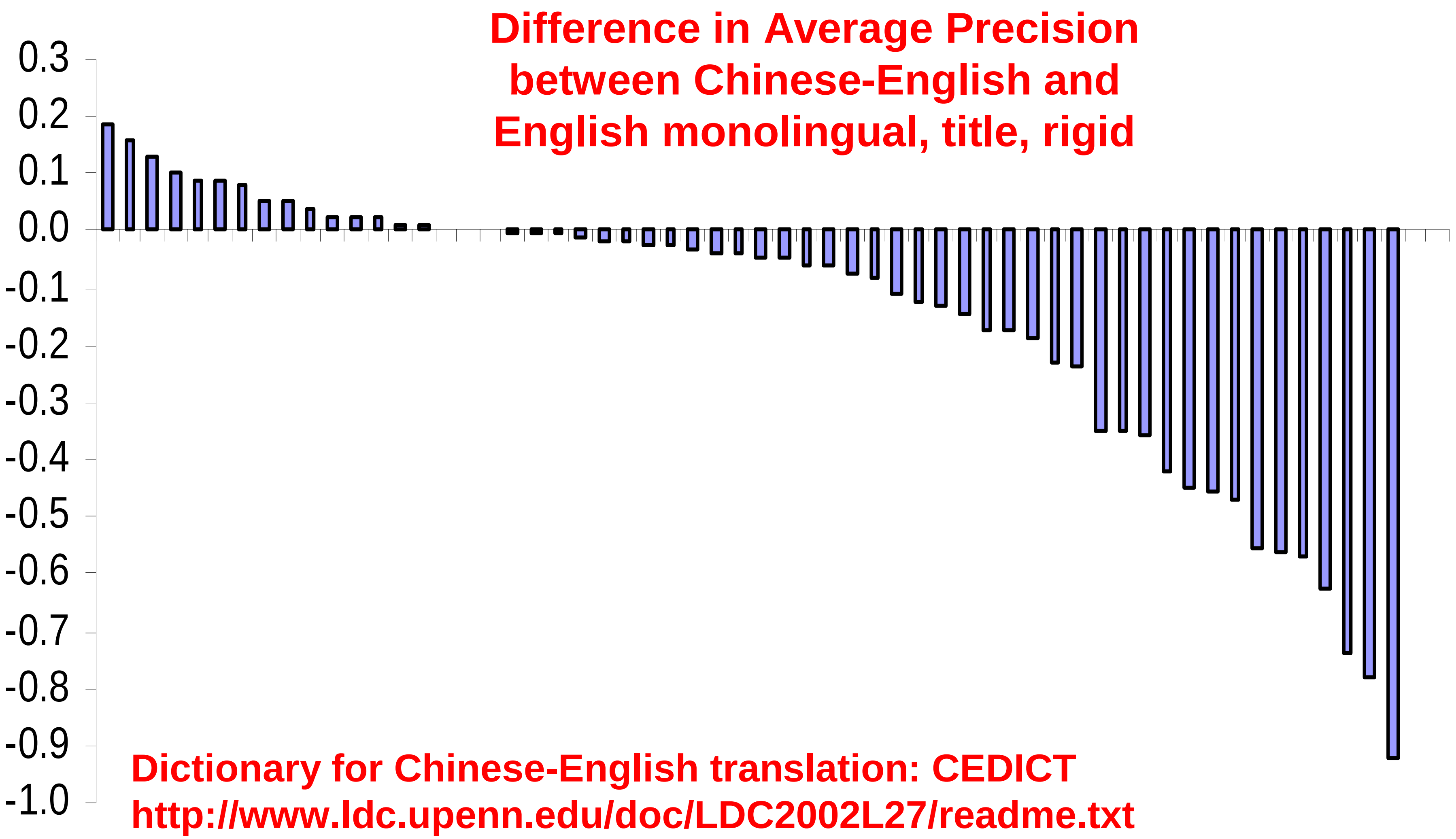
- **Using only noun phrases in retrieval**

  Topic 43: デリバティブ(derivative)

  ⇒ "derivative" is parsed as an adjective, so it isn't used in retrieval.

- **Insufficient tuning of weighting and scoring algorithms. The higher the number of expansion terms, the more influential they become.**

Chinese–English Performance



Difference in Average Precision between Chinese-English and English monolingual, title, rigid

Dictionary for Chinese-English translation: CEDICT
http://www.ldc.upenn.edu/doc/LDC2002L27/readme.txt

Chinese–English: Error Analysis

- **Improper segmentation of words**

  Topic 43: 衍生性商品 (derivative)

  ⇒ 衍生 (derivative) 性 (sex) 商品 (commodity)

- **Improper segmentation of names**

  Topic 2: 约翰走路 (Jonnie Walker)

  ⇒ 约 (partake) 翰 (pen)走 (walk) 路 (path)

  Topic 4: 佛罗伦斯·葛瑞菲丝．乔纳，花蝴蝶 (Florence Griffith Joyner, FloJo)

  ⇒ 佛 (buddha) 罗 (surname) 伦 (human relationship) 葛 (coarse grass linen) 瑞 (auspicious) 菲 (Philipines) 丝 (silk) 乔 (tall) 纳 (pay) 花 (blossom) 蝴蝶 (butterfly)

  Topic 8: 威而钢 (Viagra) → 威 (power) 钢 (steel)

  Topic 10: 胡锦涛 (Hu Jintao) → 胡 (beard) 锦 (bright) 涛 (big wave)

  Topic 12: 黑 泽明 (Akira Kurosawa) → 黑 (black) 泽 (beneficence) 明 (bright)

- **Absence of disambiguation causes unbalanced translations**

Summary

- Our systems have demonstrated stable and good performance for English and Japanese monolingual IR.
- Chinese word segmentation and lexicon coverage need improvement for better performance.
- Missing translation of proper names and terminology is still a big problem in CLIR.
- Our ongoing work focuses on automatic name translation and transliteration
  (Qu, Grefenstette, Evans 2003;
  Qu & Grefenstette 2004)

The End