Effect of Cross-Language IR in Bilingual Lexicon Acquisition from Comparable Corpora

Takehito Utsuro
Graduate School of Informatics, Kyoto University, Japan
utsuro@pine.kuee.kyoto-u.ac.jp

July 4-5, 2003,
German-Japan WS on NLP

Background

Translation Knowledge Acquisition
from Parallel/Comparable Corpora

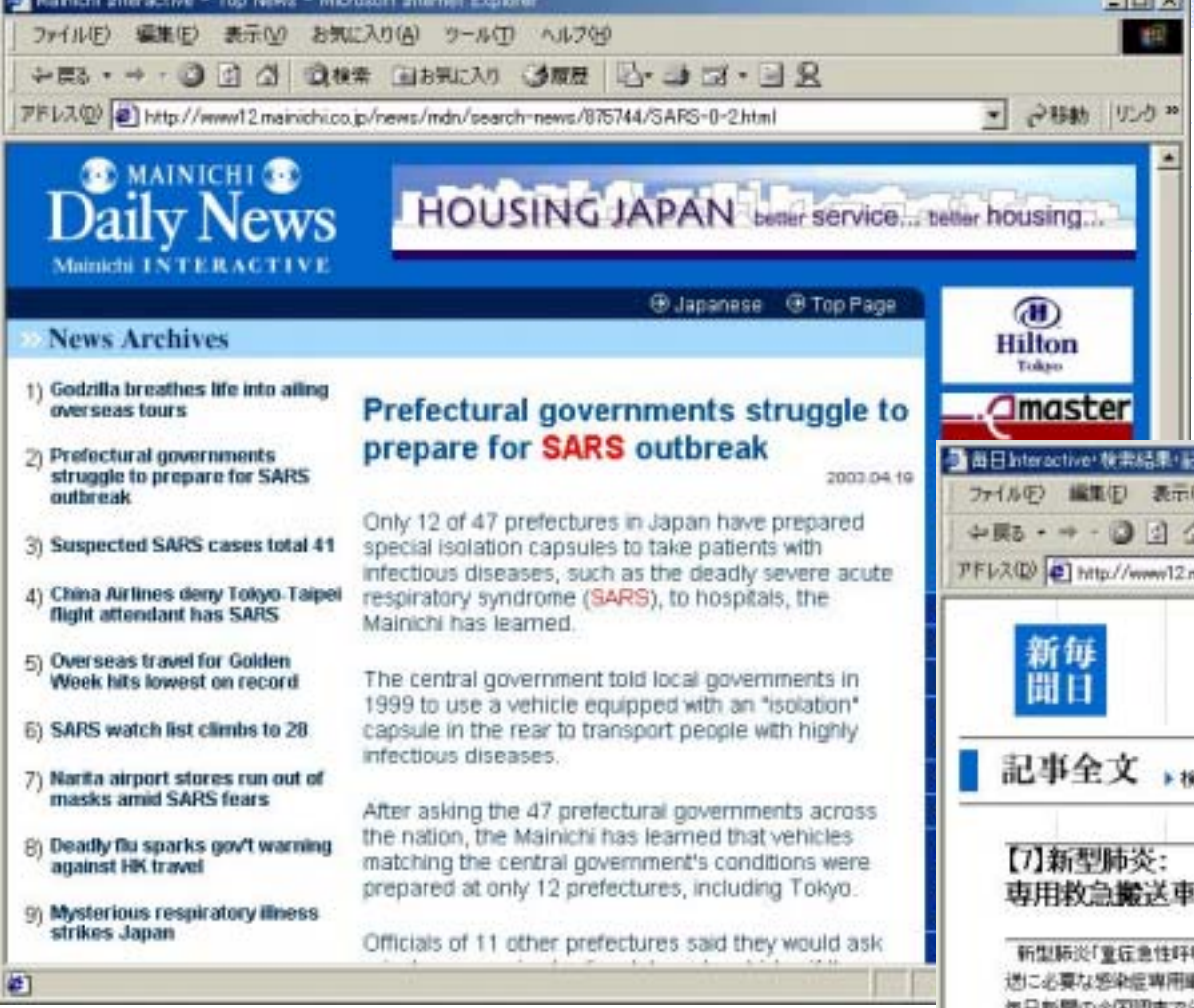
- From Parallel Corpora
  - translation knowledge acquisition: relatively easier
  - resource: less available
- From Comparable Corpora
  - translation knowledge acquisition: relatively harder
  - resource: more available

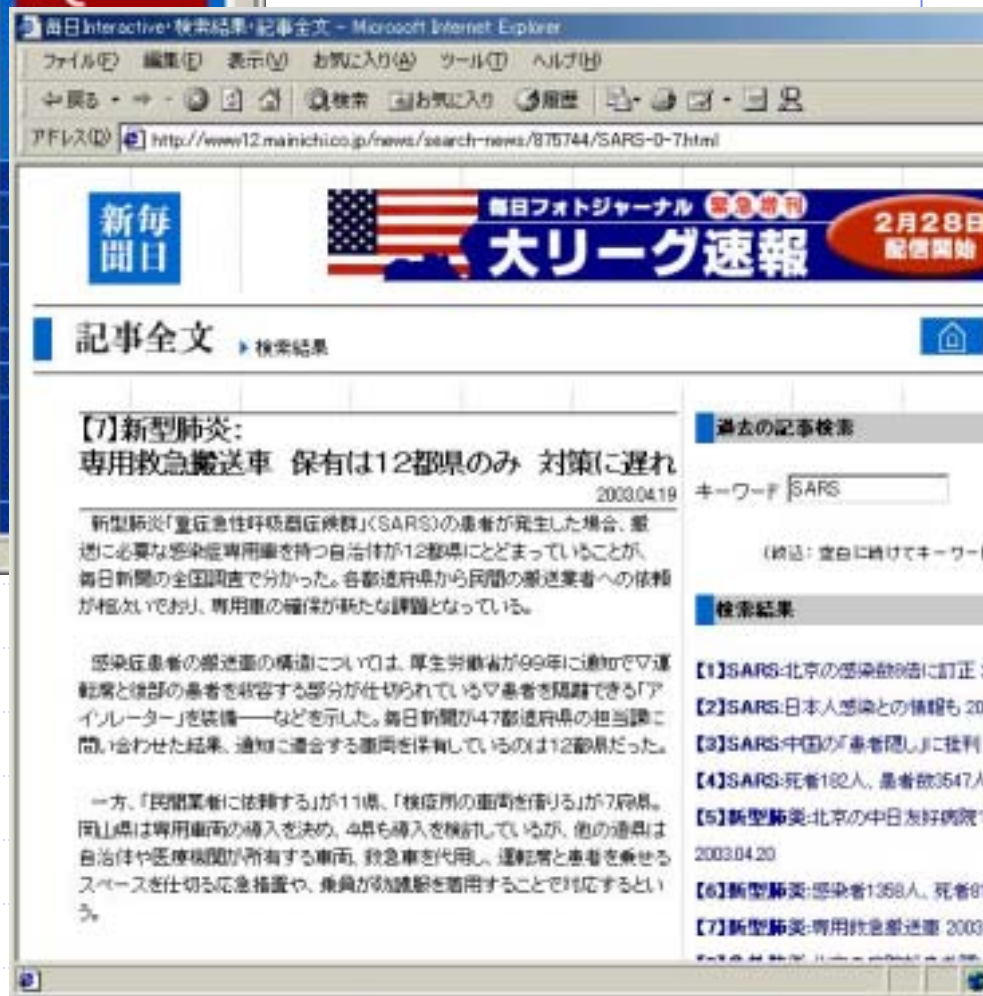
Translation Knowledge Acquisition: Our Approach

- Source of Translation Knowledge Acquisition:
  Cross-Lingually Relevant News Articles
  (on WWW news sites)
  - Updated everyday → enabling efficient acquisition of up-to-date translation knowledge
- Techniques:
  Integration of CLIR and
  Translation Knowledge Acquisition
  from Parallel/Comparable Corpora

Translation Knowledge Acquisition: Our Approach

- Source of Translation Knowledge Acquisition:
  Cross-Lingually Relevant News Articles
  (on WWW news sites)
  - Updated everyday → enabling efficient acquisition of up-to-date translation knowledge
- Techniques:
  Integration of CLIR and
  Translation Knowledge Acquisition
  from Parallel/Comparable Corpora

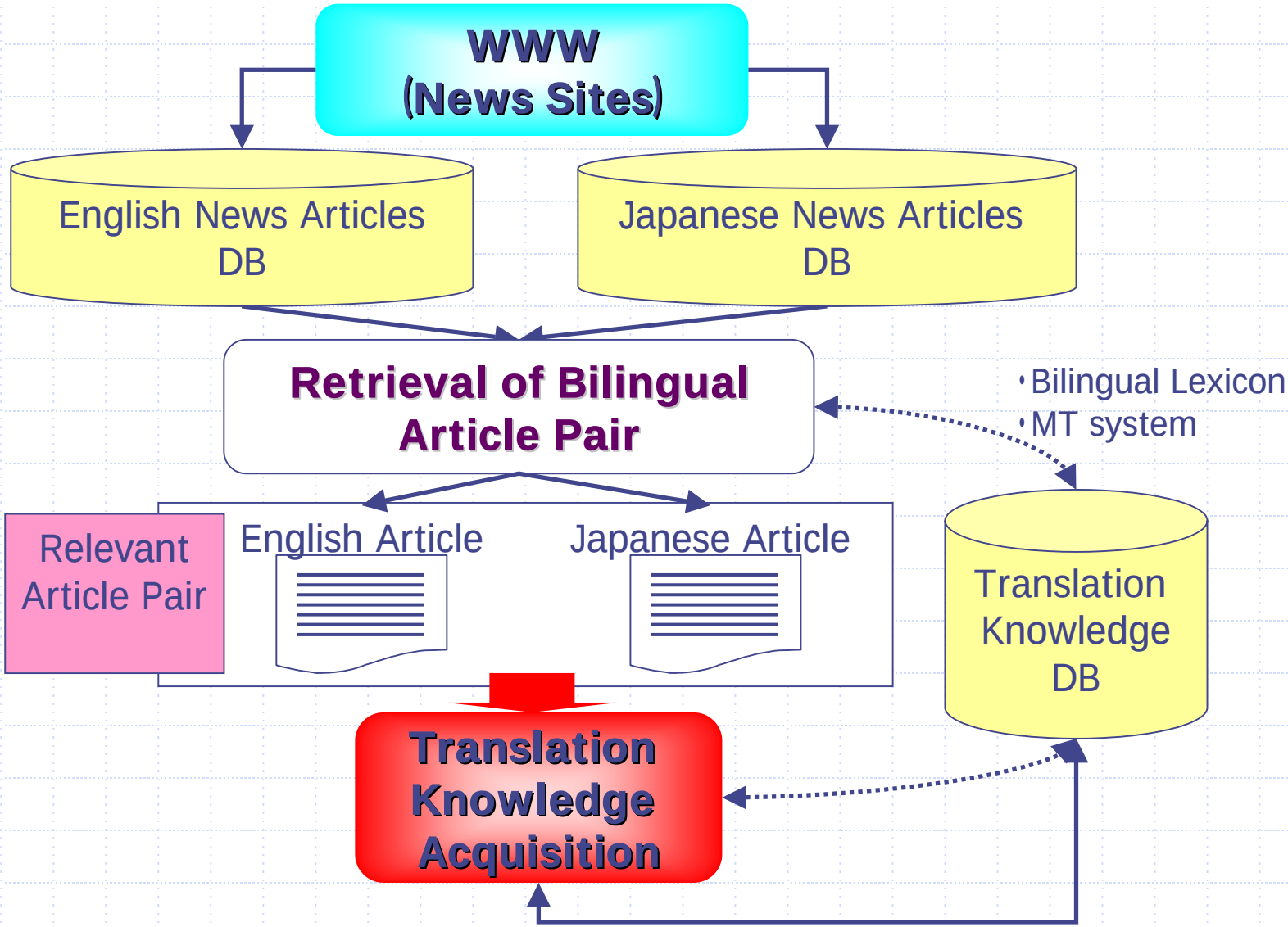

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り 履歴
アドレス(D) http://www12.mainichi.co.jp/news/mdn/search-news/875744/SARS-0-2.html
移動 リンク
MAINICHI Daily News
Mainichi INTERACTIVE
HOUSING JAPAN better service... better housing...
Japanese Top Page
News Archives
1) Godzilla breathes life into ailing overseas tours
2) Prefectural governments struggle to prepare for SARS outbreak
3) Suspected SARS cases total 41
4) China Airlines deny Tokyo-Taipei flight attendant has SARS
5) Overseas travel for Golden Week hits lowest on record
6) SARS watch list climbs to 28
7) Narita airport stores run out of masks amid SARS fears
8) Deadly flu sparks gov't warning against HK travel
9) Mysterious respiratory illness strikes Japan
Prefectural governments struggle to prepare for SARS outbreak
2003.04.19
Only 12 of 47 prefectures in Japan have prepared special isolation capsules to take patients with infectious diseases, such as the deadly severe acute respiratory syndrome (SARS), to hospitals, the Mainichi has learned.
The central government told local governments in 1999 to use a vehicle equipped with an "isolation" capsule in the rear to transport people with highly infectious diseases.
After asking the 47 prefectural governments across the nation, the Mainichi has learned that vehicles matching the central government's conditions were prepared at only 12 prefectures, including Tokyo.
Officials of 11 other prefectures said they would ask
Hilton Tokyo
master


毎日Interactive 検索結果・記事全文 - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り 履歴
アドレス(D) http://www12.mainichi.co.jp/news/search-news/875744/SARS-0-7.html
新毎聞日
毎日フォトジャーナル 緊急増刊
大リーグ速報
2月28日 配信開始
記事全文 検索結果
【7】新型肺炎:
専用救急搬送車 保有は12都県のみ 対策に遅れ
2003.04.19
新型肺炎「重症急性呼吸器症候群」(SARS)の患者が発生した場合、搬送に必要な感染症専用車を持つ自治体が12都県にとどまっていることが、毎日新聞の全国調査で分かった。各都道府県から民間の搬送業者への依頼が相次いでおり、専用車の確保が新たな課題となっている。
感染症患者の搬送車の構造については、厚生労働省が99年に通知で▽運転席と後部の患者を収容する部分が仕切られている▽患者を隔離できる「アイソレーター」を装備――などを示した。毎日新聞が47都道府県の担当課に問い合わせた結果、通知に適合する車両を保有しているのは12都県だった。
一方、「民間業者に依頼する」が11県、「検疫所の車両を借りる」が7府県。岡山県は専用車両の導入を決め、4県も導入を検討しているが、他の道県は自治体や医療機関が所有する車両、救急車を代用し、運転席と患者を乗せるスペースを仕切る応急措置や、乗員が防護服を着用することで対応するという。
過去の記事検索
キーワード SARS
検索結果
【1】SARS:北京の感染者8倍に訂正
【2】SARS:日本人感染との情報も
【3】SARS:中国の「患者隠し」に批判
【4】SARS:死者182人、患者数3547人
【5】新型肺炎:北京の中日友好病院
2003.04.20
【6】新型肺炎:感染者1358人、死者
【7】新型肺炎:専用救急搬送車 2003

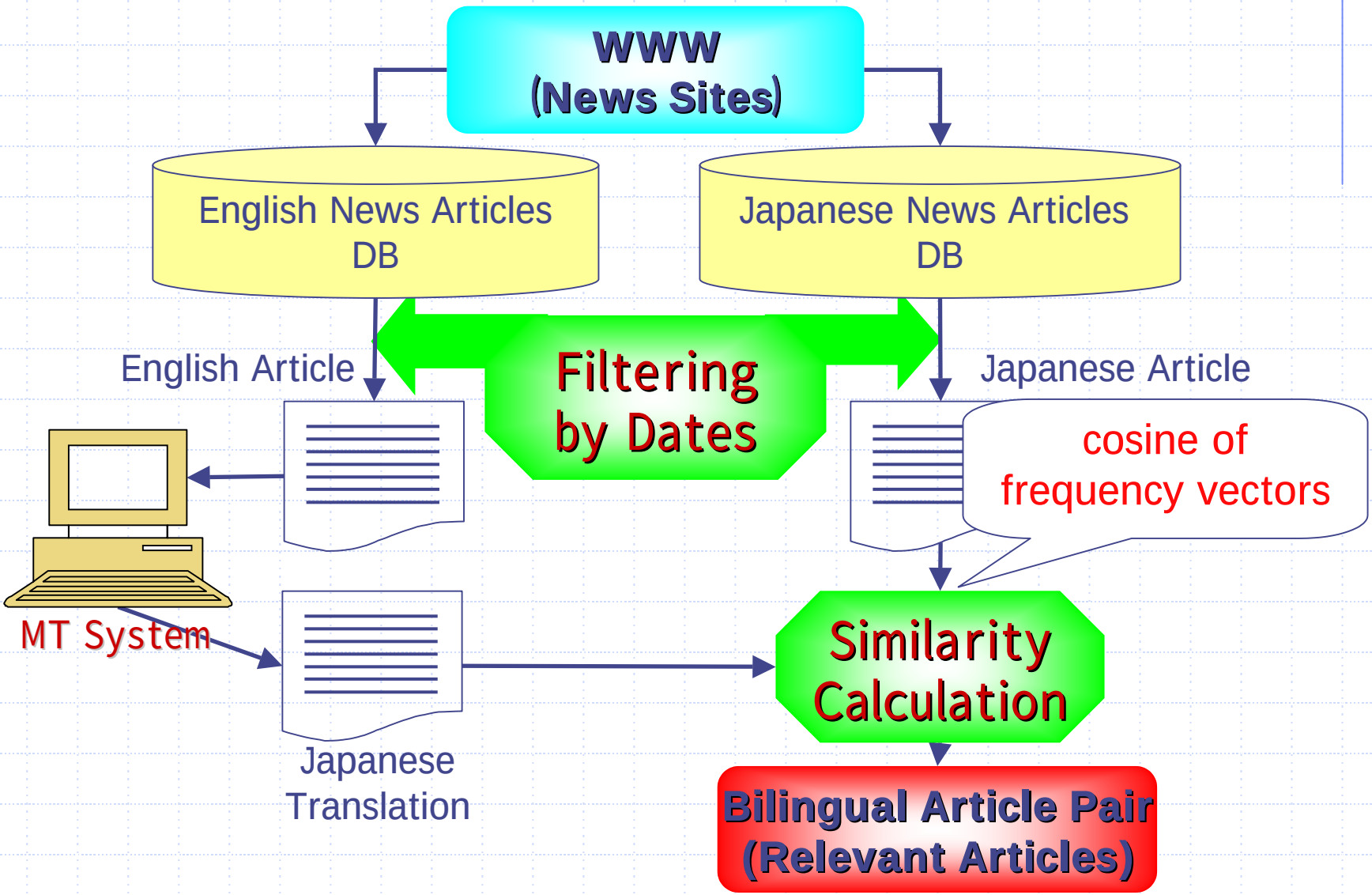
Translation Knowledge Acquisition: Our Approach

- Source of Translation Knowledge Acquisition:
  Cross-Lingually Relevant News Articles
  (on WWW news sites)
  - Updated everyday → enabling efficient acquisition of up-to-date translation knowledge
- Techniques:
  Integration of CLIR and
  Translation Knowledge Acquisition
  from Parallel/Comparable Corpora

Translation Knowledge Acquisition from WWW News Sites: Overview



Cross-Language Retrieval of Relevant News Articles



Related Research Issues: Translation Knowledge Acquisition

- Acquisition from Parallel Corpora
  - statistical MT models: e.g., [Brown 90, 93]
  - term correspondences estimation based on contingency tables of cross-language co-occurrence frequencies: e.g., [Gale 91, Kumano 94, Haruno 96, Smadja 96, Kitamura 96, Melamed 00]

- Acquisition from Comparable Corpora: contextual similarities of words across languages
  - without the help of existing bilingual lexicons: earlier works [Fung 95]
  - exploiting existing bilingual lexicons as initial seed: later works [Rapp 95,99, Kaji 96, K.Tanaka 96, Fung 98, T.Tanaka 02]

- Collecting Partially Bilingual Texts from WWW with Internet Search Engines: [Nagata 01]

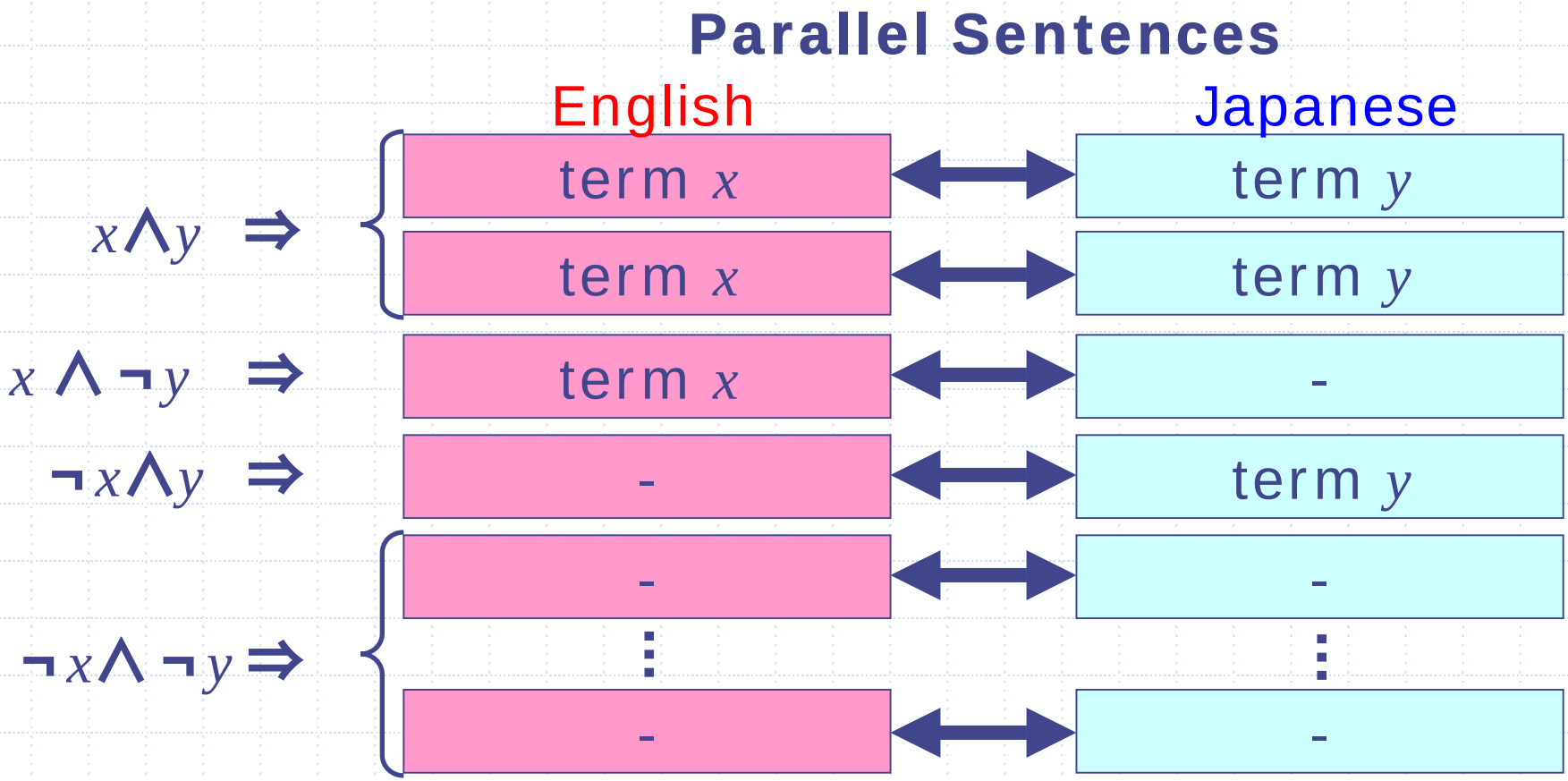
Related Research Issues: Translation Knowledge Acquisition

- Acquisition from Parallel Corpora
  - statistical MT models: e.g., [Brown 90, 93]
  - term correspondences estimation based on contingency tables of cross-language co-occurrence frequencies: e.g., [Gale 91, Kumano 94, Haruno 96, Smadja 96, Kitamura 96, Melamed 00]

- Acquisition from Comparable Corpora: contextual similarities of words across languages
  - without the help of existing bilingual lexicons: earlier works [Fung 95]
  - exploiting existing bilingual lexicons as initial seed: later works [Rapp 95,99, Kaji 96, K.Tanaka 96, Fung 98, T.Tanaka 02]

- Collecting Partially Bilingual Texts from WWW with Internet Search Engines: [Nagata 01]

Estimating Bilingual Term Correspondences from Parallel Sentences



Measures for Estimating Bilingual Term Correspondences from Contingency Table

	y	$\neg y$
x	$freq(x, y) = a$	$freq(x, \neg y) = b$
$\neg x$	$freq(\neg x, y) = c$	$freq(\neg x, \neg y) = d$

- mutual information (MI)

$$I(x ; y) = \log_2 \frac{aN}{(a+b)(a+c)}$$

- Φ^2 statistic

$$\phi^2(x, y) = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

- dice coefficient

$$Dice(x, y) = \frac{2a}{2a+b+c}$$

- log-likelihood

$$\text{Log-like} = f(a)+f(b)+f(c)+f(d)-f(a+b)-f(a+c)-f(b+d)-f(c+d)-f(a+b+c+d)$$

Where $f(x) = x \log x$

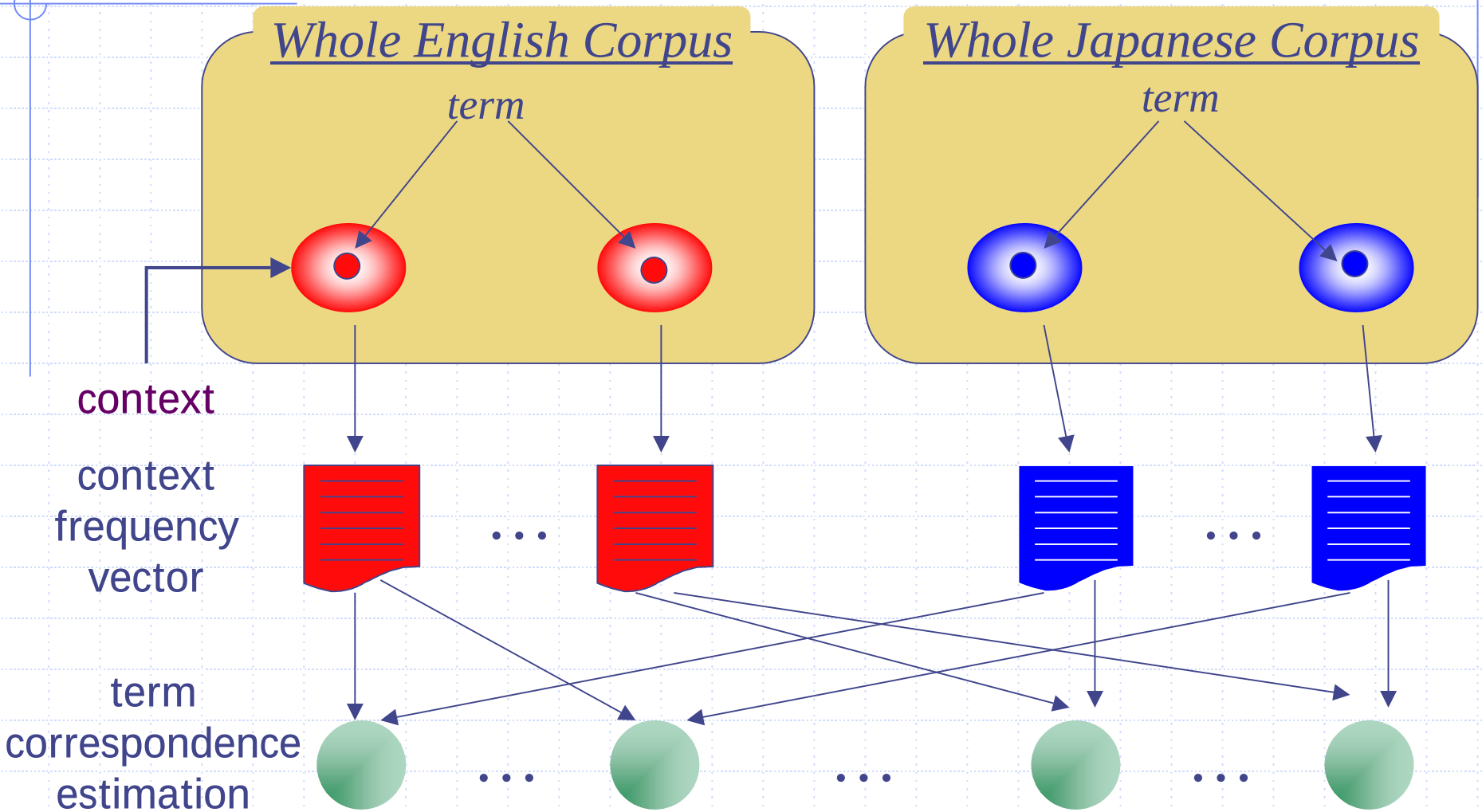
Related Research Issues: Translation Knowledge Acquisition

- Acquisition from Parallel Corpora
  - statistical MT models: e.g., [Brown 90, 93]
  - term correspondences estimation based on contingency tables of cross-language co-occurrence frequencies: e.g., [Gale 91, Kumano 94, Haruno 96, Smadja 96, Kitamura 96, Melamed 00]

- Acquisition from Comparable Corpora: contextual similarities of words across languages
  - without the help of existing bilingual lexicons: earlier works [Fung 95]
  - exploiting existing bilingual lexicons as initial seed: later works [Rapp 95,99, Kaji 96, K.Tanaka 96, Fung 98, T.Tanaka 02]

- Collecting Partially Bilingual Texts from WWW with Internet Search Engines: [Nagata 01]

Term Correspondence Acquisition from Comparable Corpora



Related Research Issues: Translation Knowledge Acquisition

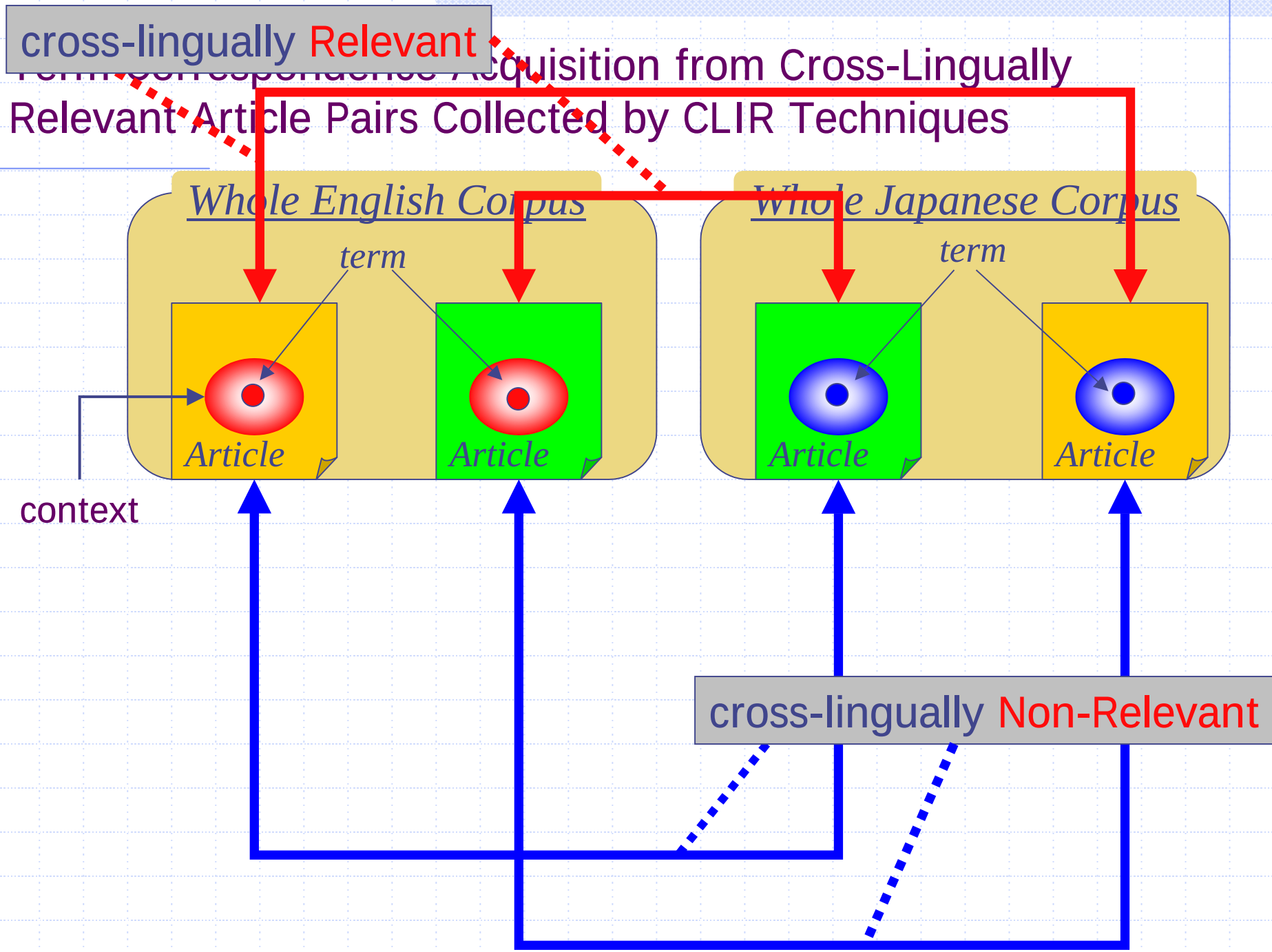
- Acquisition from Parallel Corpora
  - statistical MT models: e.g., [Brown 90, 93]
  - term correspondences estimation based on contingency tables of cross-language co-occurrence frequencies: e.g., [Gale 91, Kumano 94, Haruno 96, Smadja 96, Kitamura 96, Melamed 00]

- Acquisition from Comparable Corpora: contextual similarities of words across languages
  - without the help of existing bilingual lexicons: earlier works [Fung 95]
  - exploiting existing bilingual lexicons as initial seed: later works [Rapp 95,99, Kaji 96, K.Tanaka 96, Fung 98, T.Tanaka 02]

- Collecting Partially Bilingual Texts from WWW with Internet Search Engines: [Nagata 01]

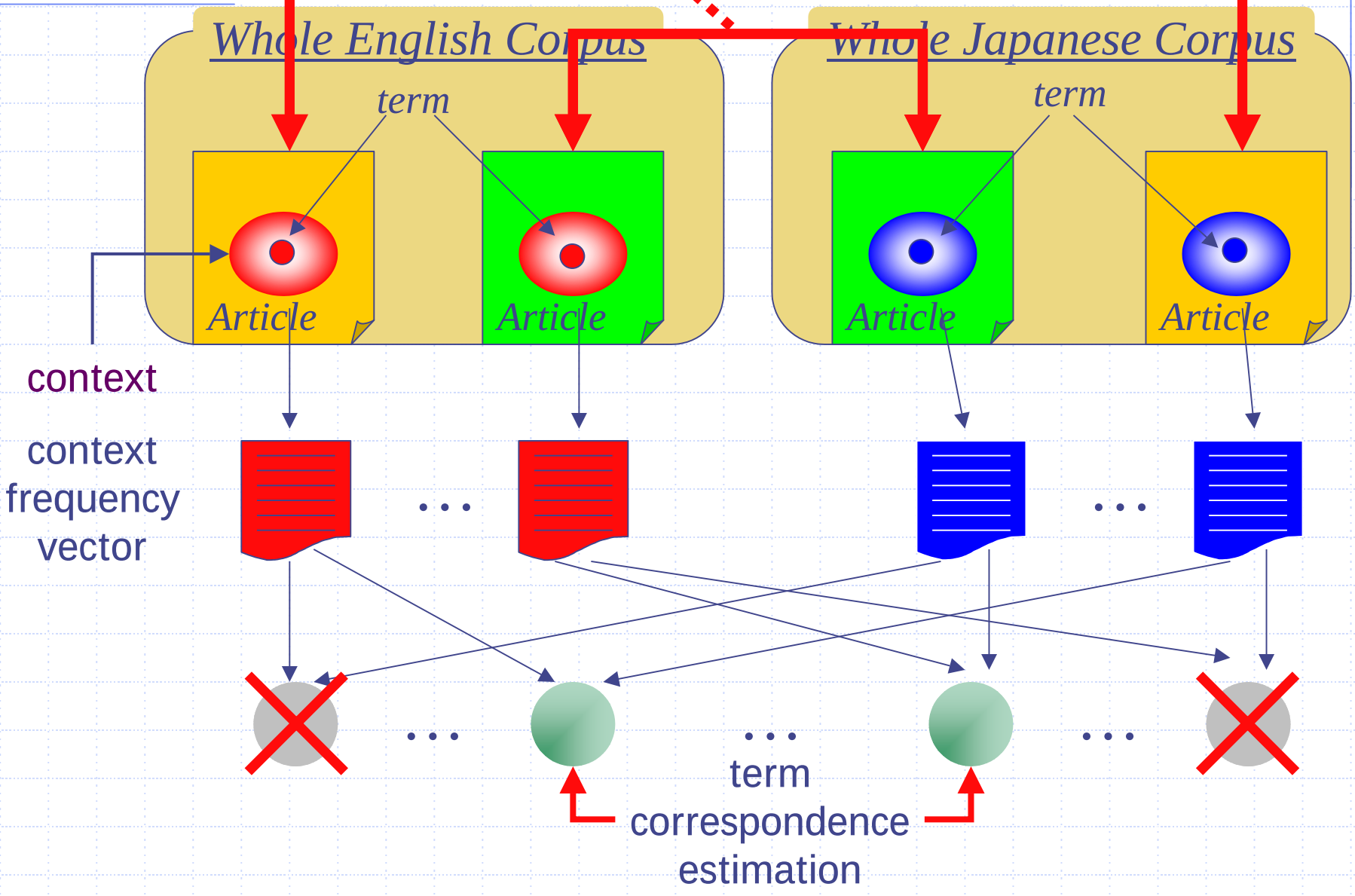
Translation Knowledge Acquisition: Our Approach

- Translation knowledge acquisition from cross-lingually relevant article pairs collected by CLIR techniques


cross-lingually Relevant
Relevant Article Pairs Collected by CLIR Techniques
Whole English Corpus
Whole Japanese Corpus
term
term
Article
Article
Article
Article
context
cross-lingually Non-Relevant

cross-lingually Relevant

Term Correspondence Acquisition from Cross-Lingually Relevant Article Pairs Collected by CLIR Techniques



Translation Knowledge Acquisition: Our Approach

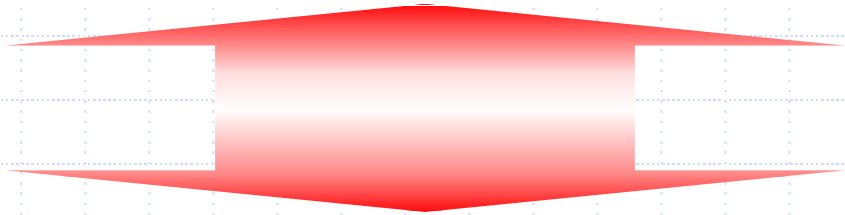
- Translation knowledge acquisition from cross-lingually relevant article pairs collected by CLIR techniques

Translation Knowledge Acquisition: Our Approach

- Translation knowledge acquisition from cross-lingually relevant article pairs collected by CLIR techniques
- Techniques for parallel corpora become applicable to translation knowledge acquisition from comparable corpora

Translation Knowledge Acquisition: Our Approach

- Translation knowledge acquisition from cross-lingually relevant article pairs collected by CLIR techniques
- Techniques for parallel corpora become applicable to translation knowledge acquisition from comparable corpora



Related Work: Translation Knowledge Acquisition from Comparable Corpora

- Estimating term correspondences based on contextual similarities across languages
- Contextual vectors: averaged over the whole corpus
- No use of CLIR techniques for restricting relevant documents across languages

Cross-Language Retrieval of Relevant News Articles: Evaluation Issues

- Availability of Cross-Lingually Relevant Articles
  - Query articles should be English rather than Japanese
  - Cross-Lingually relevant articles are available for more than 60% English query articles
- Recall/Precision of Cross-Language Retrieval of Relevant News Articles
  - precision: 50% or more when article similarities $\geqq$ 0.4

Term Correspondence Acquisition: Evaluation Issues

- Source of Translation Knowledge Acquisition:
  Cross-Lingually Relevant News Articles
  on WWW News Sites
- Effect of CLIR
  in Reducing Bilingual Term Pair Candidates
- Accuracy of
  Bilingual Term Correspondence Estimation

Term Correspondence Acquisition: Evaluation Issues

- Source of Translation Knowledge Acquisition: Cross-Lingually Relevant News Articles on WWW News Sites
- Effect of CLIR in Reducing Bilingual Term Pair Candidates
- Accuracy of Bilingual Term Correspondence Estimation

Statistics of Article Pairs with Similarity Values above Lower Bound

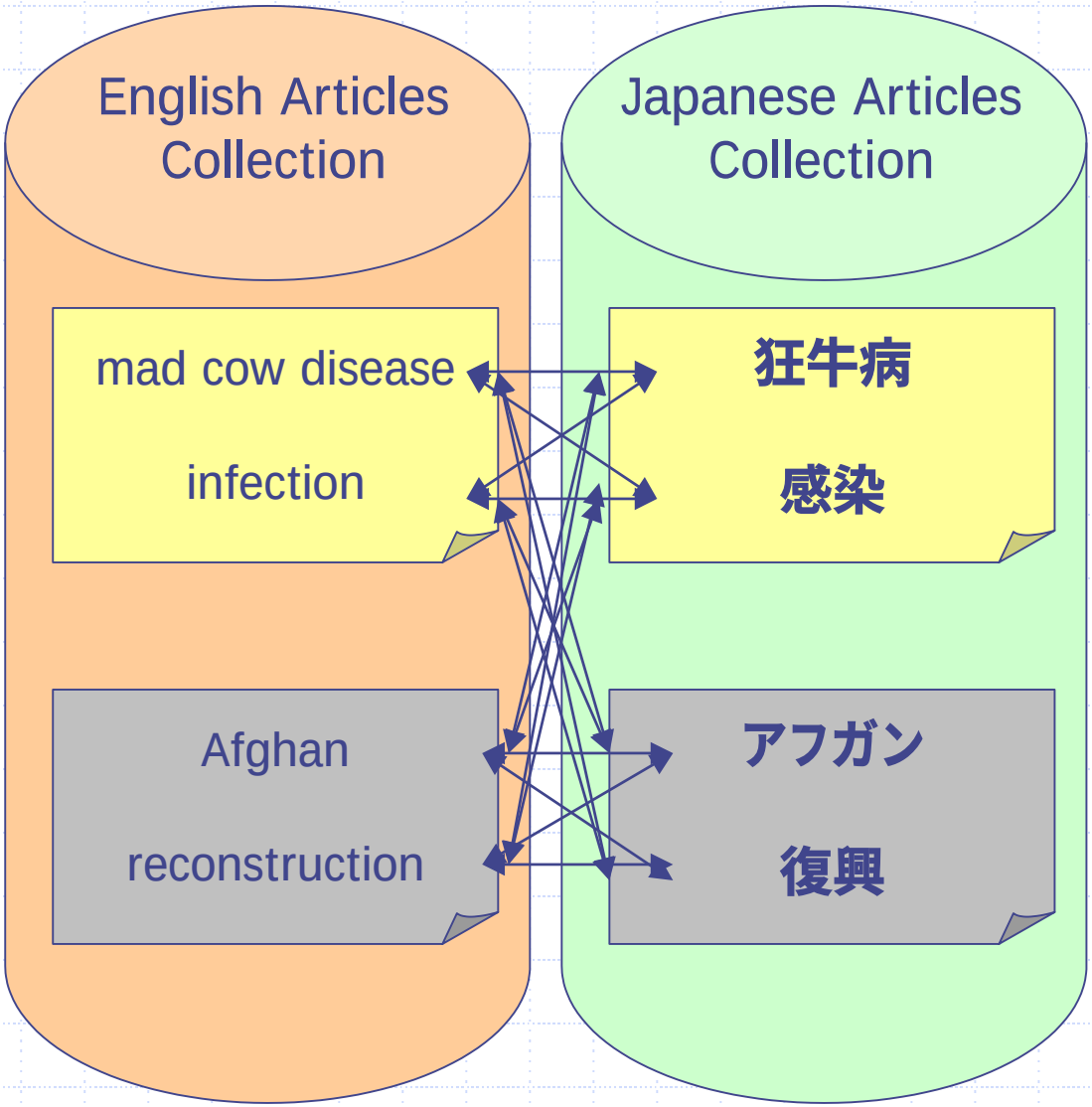
Site		A	B	C
Total # of Days	English	562	162	162
	Japanese	578	168	166
Total # of Articles	English	607	2910	3435
	Japanese	21349	14854	16166

Site	A				B		C	
Lower Bound L_d of Article's Sim	0.25	0.3	0.4	0.5	0.4	0.5	0.4	0.5
Difference of dates (days)	±4				±3		±2	
# of English Articles	473	362	190	74	415	92	453	144
# of Japanese Articles	1990	1128	377	101	631	127	725	185

Term Correspondence Acquisition: Evaluation Issues

- Source of Translation Knowledge Acquisition: Cross-Lingually Relevant News Articles on WWW News Sites
- Effect of CLIR in Reducing Bilingual Term Pair Candidates
- Accuracy of Bilingual Term Correspondence Estimation

Bilingual Term Pair Candidates: *Full* pairs vs. *Reduced* pairs



<"full"> 16 pairs

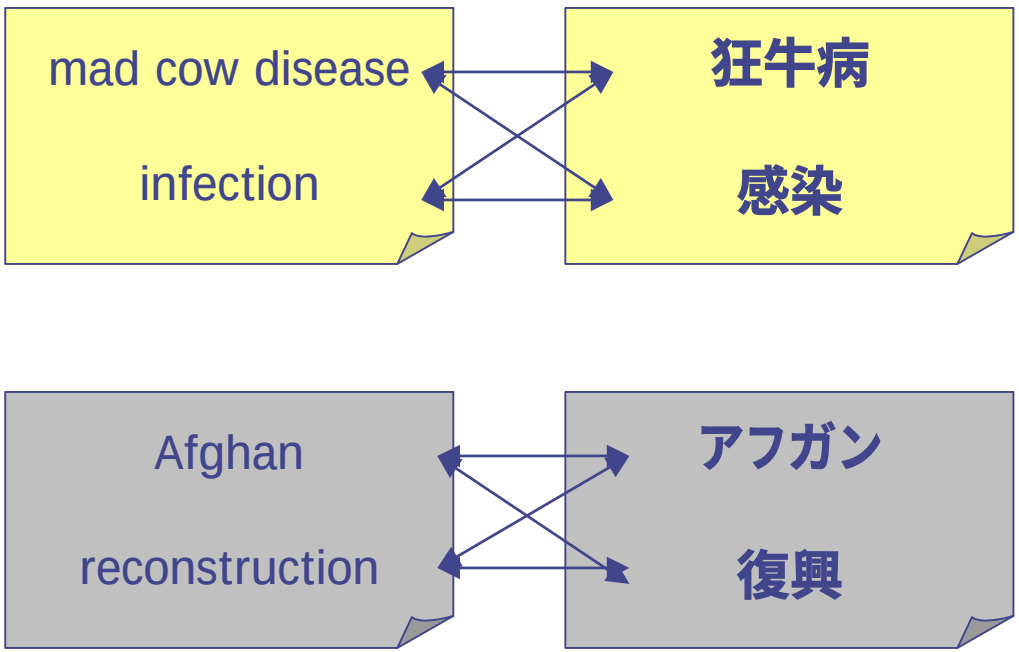
- mad cow disease-**狂牛病**
- mad cow disease-**感染**
- mad cow disease-**アフガン**
- mad cow disease-**復興**

⋮

- reconstruction-**狂牛病**
- reconstruction-**感染**
- reconstruction-**アフガン**
- reconstruction-**復興**

Bilingual Term Pair Candidates: *Full* pairs vs. *Reduced* pairs

Reduced pairs:
collected from relevant articles



<“full”> 16 pairs

- mad cow disease-**狂牛病**
- mad cow disease-**感染**
- mad cow disease-**アフガン**
- mad cow disease-**復興**

⋮

- reconstruction-**狂牛病**
- reconstruction-**感染**
- reconstruction-**アフガン**
- reconstruction-**復興**

<“reduced”> 8 pairs

- mad cow disease-**狂牛病**
- mad cow disease-**感染**

⋮

- reconstruction-**アフガン**
- reconstruction-**復興**

of Monolingual Terms and Bilingual Term Pairs

site	article sim lower bound	# of monolingual terms		# of candidate bilingual term pairs		
		English	Japanese	reduced	full	ratio (full/ reduced)
A	0.5	780	737	52,435	574,860	**11.0**
	0.4	2,684	3,231	427,889	8,672,004	**20.3**
	0.3	5,463	8,119	1,639,714	44,354,097	**27.1**
B	0.5	2,468	2,158	494,544	5,325,944	**10.8**
	0.4	11,968	8,658	4,074,980	103,618,944	**25.4**
C	0.5	3,760	2,612	638,089	9,821,120	**15.4**
	0.4	13,200	9,433	4,367,775	124,515,600	**28.5**

Term Recognition Criteria: (preliminary)

- No statistics-based nor grammar-based intelligent criteria

- English terms:
  every word sequences (5 words or less)
- Japanese terms:
  (noun|verb)+ (5 words or less)
- Frequency Lower Bounds

of Monolingual Terms and Bilingual Term Pairs

site	article sim lower bound	# of monolingual terms		# of candidate bilingual term pairs		
		English	Japanese	reduced	full	ratio (full/ reduced)
A	0.5	780	737	52,435	574,860	**11.0**
	0.4	2,684	3,231	427,889	8,672,004	**20.3**
	0.3	5,463	8,119	1,639,714	44,354,097	**27.1**
B	0.5	2,468	2,158	494,544	5,325,944	**10.8**
	0.4	11,968	8,658	4,074,980	103,618,944	**25.4**
C	0.5	3,760	2,612	638,089	9,821,120	**15.4**
	0.4	13,200	9,433	4,367,775	124,515,600	**28.5**

Effect of CLIR in Bilingual Lexicon Acquisition

- # of bilingual term pair candidates
  reduced to 1/10〜1/30
- most of filtered-out pairs are not correct translation
- reducing computational complexity of
  bilingual term correspondence estimation

Term Correspondence Acquisition: Evaluation Issues

- Source of Translation Knowledge Acquisition:
  Cross-Lingually Relevant News Articles
  on WWW News Sites

- Effect of CLIR
  in Reducing Bilingual Term Pair Candidates

- Accuracy of
  Bilingual Term Correspondence Estimation

Bilingual Term Correspondence Estimation with Statistical Measure

Maximum Estimated Values

English term	Japanese term	freq(t_E)	freq(t_J)	freq(t_E,t_J)	ϕ^2
Tokyo District Court	東京地裁	11	9	7	0.486
	救済	11	3	3	0.268
	被告	11	9	4	0.151
	地方裁判所	11	3	2	0.116
	:	:	:	:	:

higher rank

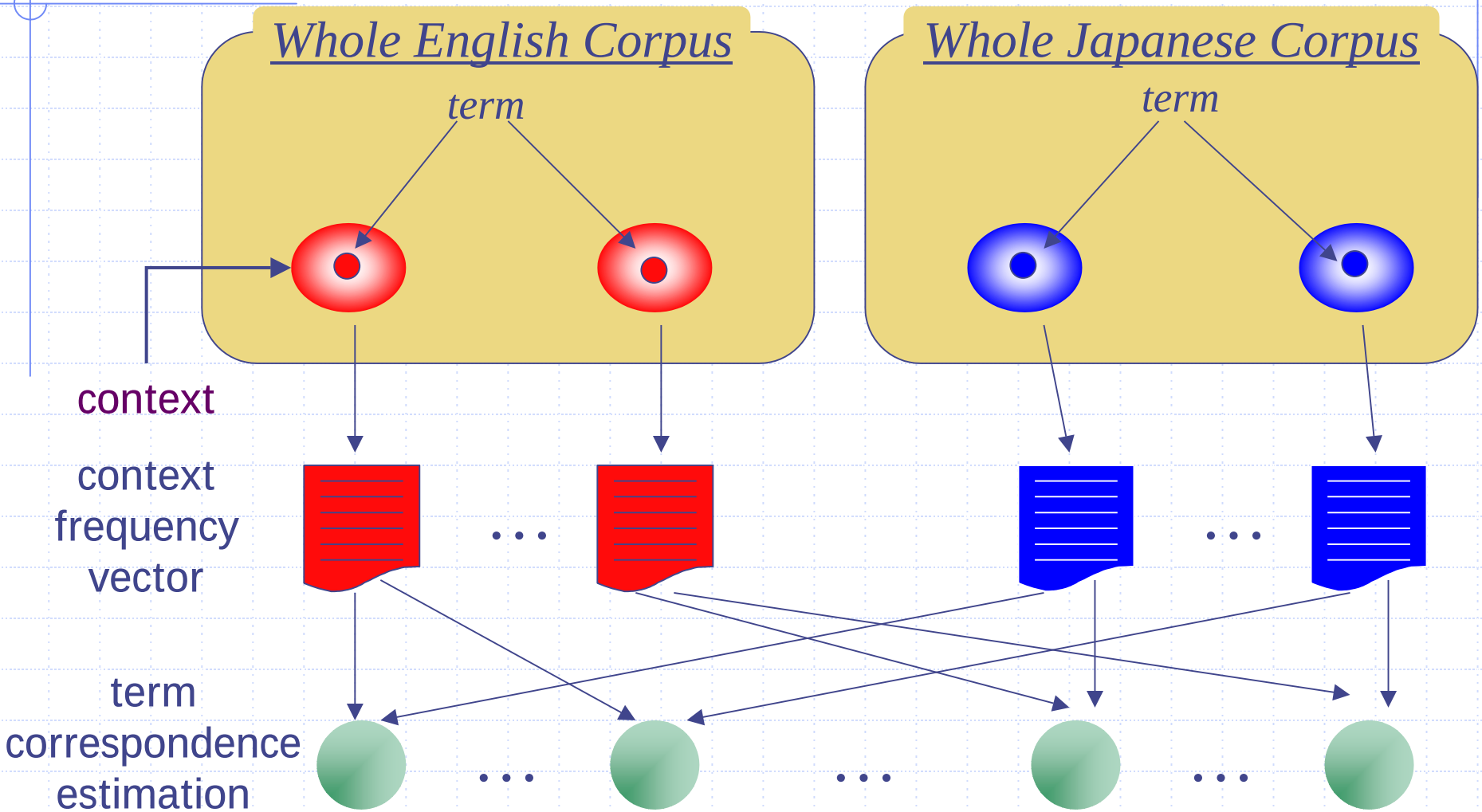
Japanese translation candidates

Comparison of Measures for Bilingual Term Correspondence Estimation

(site A, Sim LBD=0.4, for 200 English terms with highest max estimation values)

- contextual similarities: reduced vs. full

Term Correspondence Estimation based on Contextual Similarity across Languages



Comparison of Measures for Bilingual Term Correspondence Estimation

(site A, Sim LBD=0.4, for 200 English terms with highest max estimation values)

- contextual similarities: reduced vs. full

Comparison of Measures for Bilingual Term Correspondence Estimation

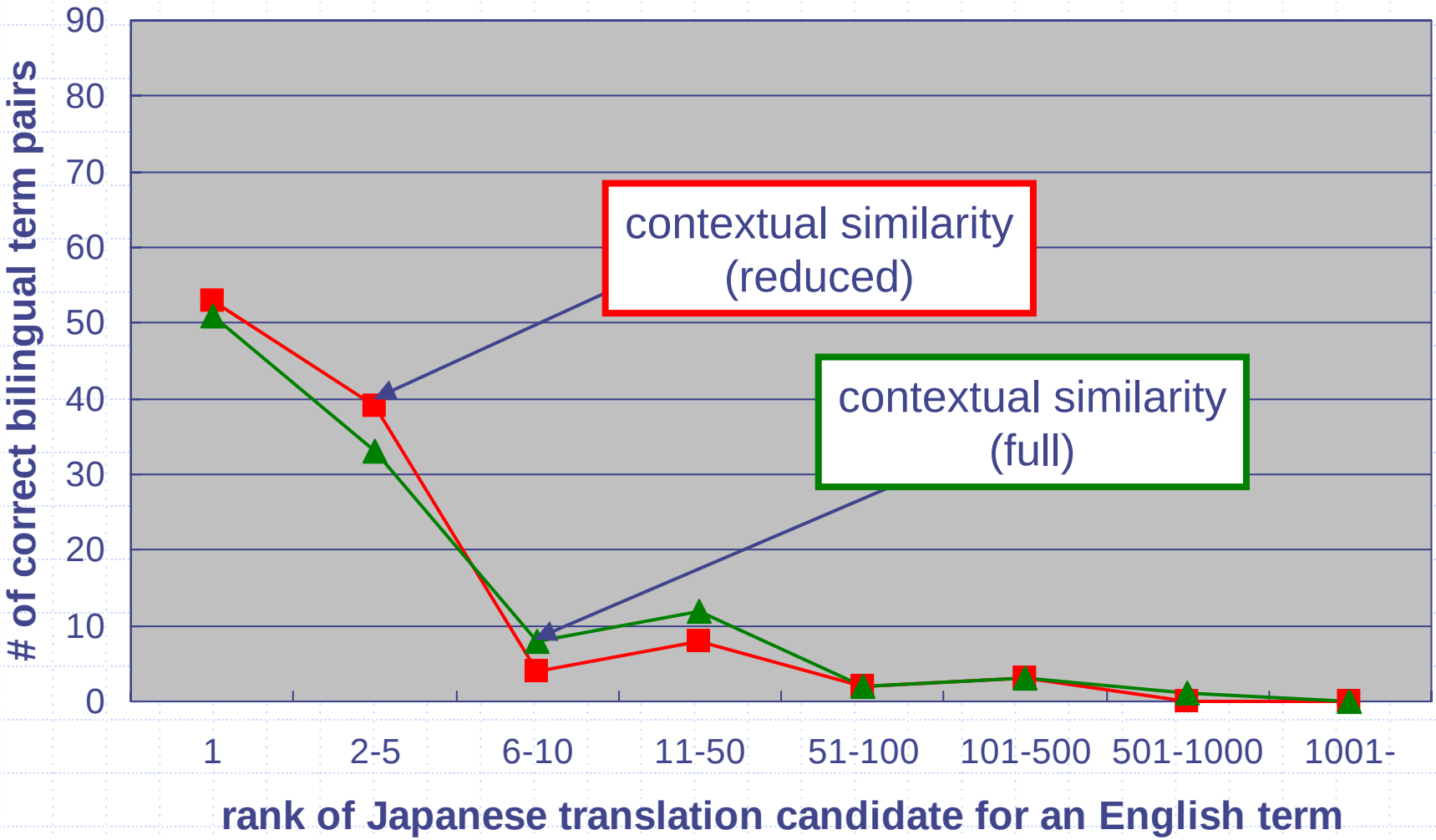
(site A, Sim LBD=0.4, for 200 English terms with highest max estimation values)

- contextual similarities: reduced vs. full
  - estimated bilingual term pairs *mostly overlap*

Numbers of Correct Bilingual Term Pair Pairs

(Manual Evaluation)

(site A, Sim LBD=0.4, for 200 English terms with highest max estimation values)



Bilingual Term Correspondence Estimation with Statistical Measure

Maximum Estimated Values

English term	Japanese term	freq(t_E)	freq(t_J)	freq(t_E,t_J)	ϕ^2
Tokyo District Court	東京地裁	11	9	7	0.486
	救済	11	3	3	0.268
	被告	11	9	4	0.151
	地方裁判所	11	3	2	0.116
	:	:	:	:	:

higher rank

Japanese translation candidates

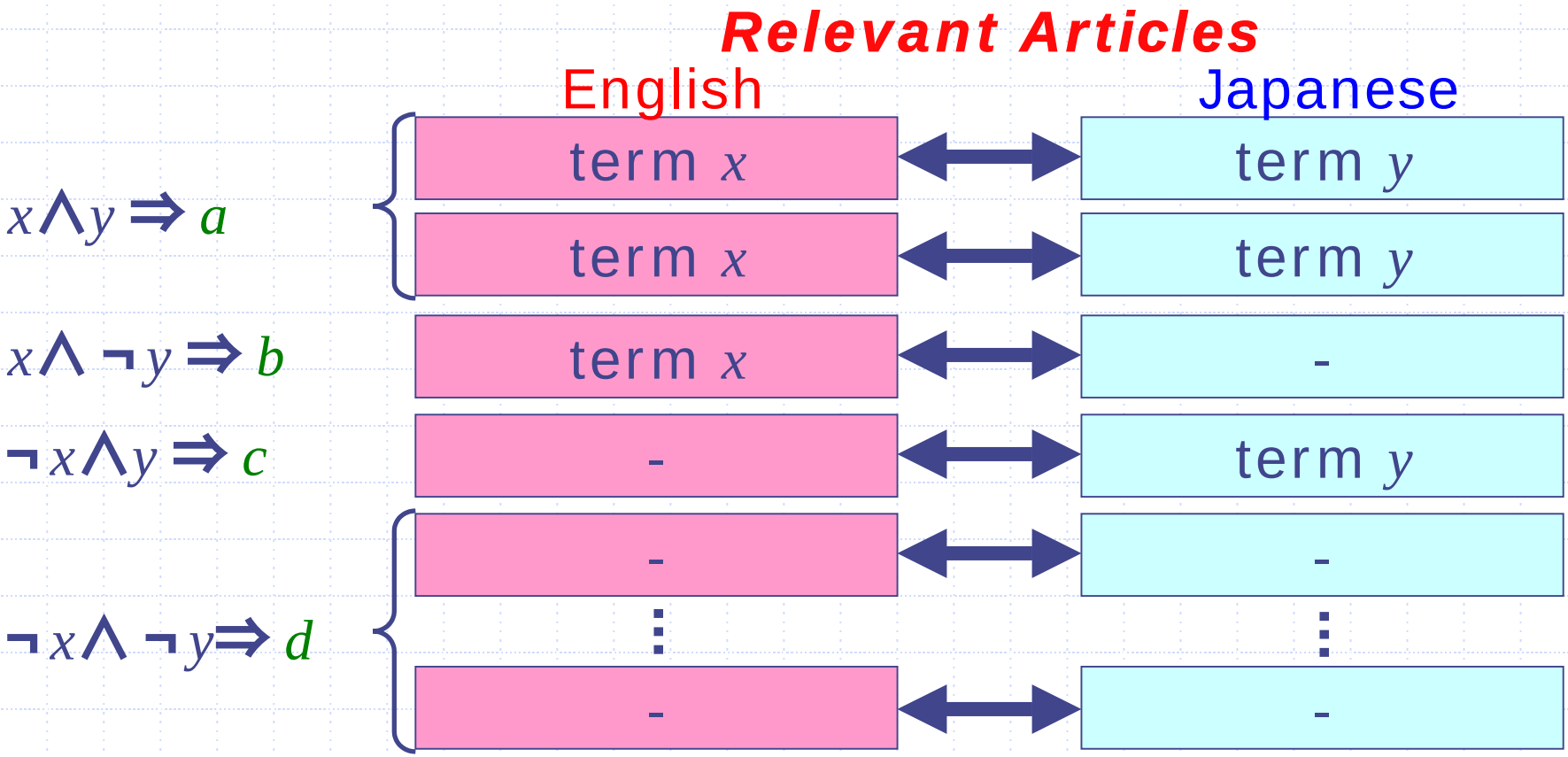
Comparison of Measures for Bilingual Term Correspondence Estimation

(site A, Sim LBD=0.4, for 200 English terms with highest max estimation values)

- contextual similarities: reduced vs. full

- contextual similarities (reduced) vs. ϕ^2 (contingency table)

Applying Φ^2 statistic to Bilingual Term Correspondences Estimation from *Relevant Articles*



$$\Phi^2(x,y) = \frac{(a \cdot d - b \cdot c)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Comparison of Measures for Bilingual Term Correspondence Estimation

(site A, Sim LBD=0.4, for 200 English terms with highest max estimation values)

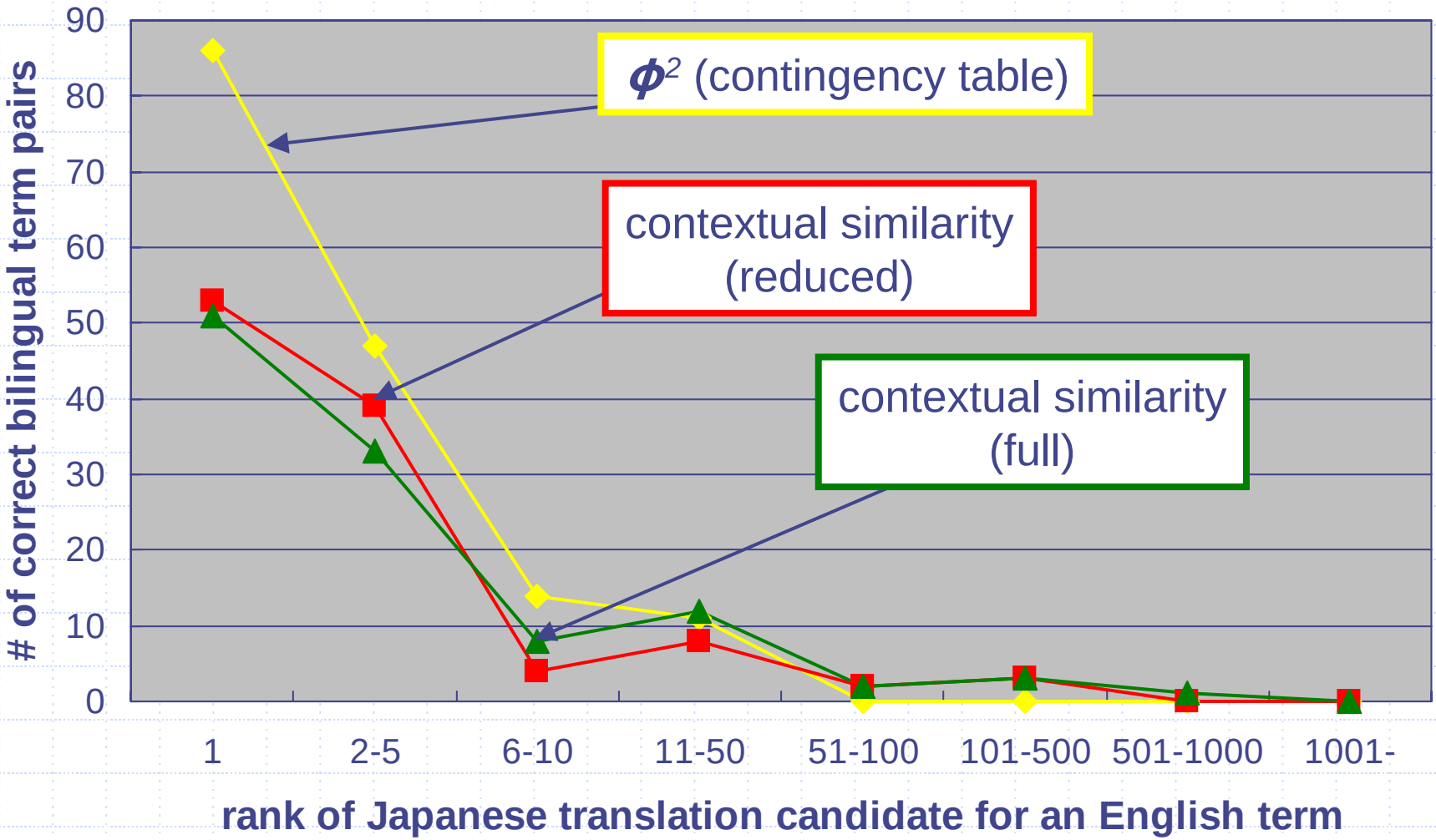
- contextual similarities: reduced vs. full
  - estimated bilingual term pairs mostly overlap

- contextual similarities (reduced) vs. ϕ^2 (contingency table)
  - overlap of the estimated bilingual term pairs is less than 30%

Numbers of Correct Bilingual Term Pair Pairs

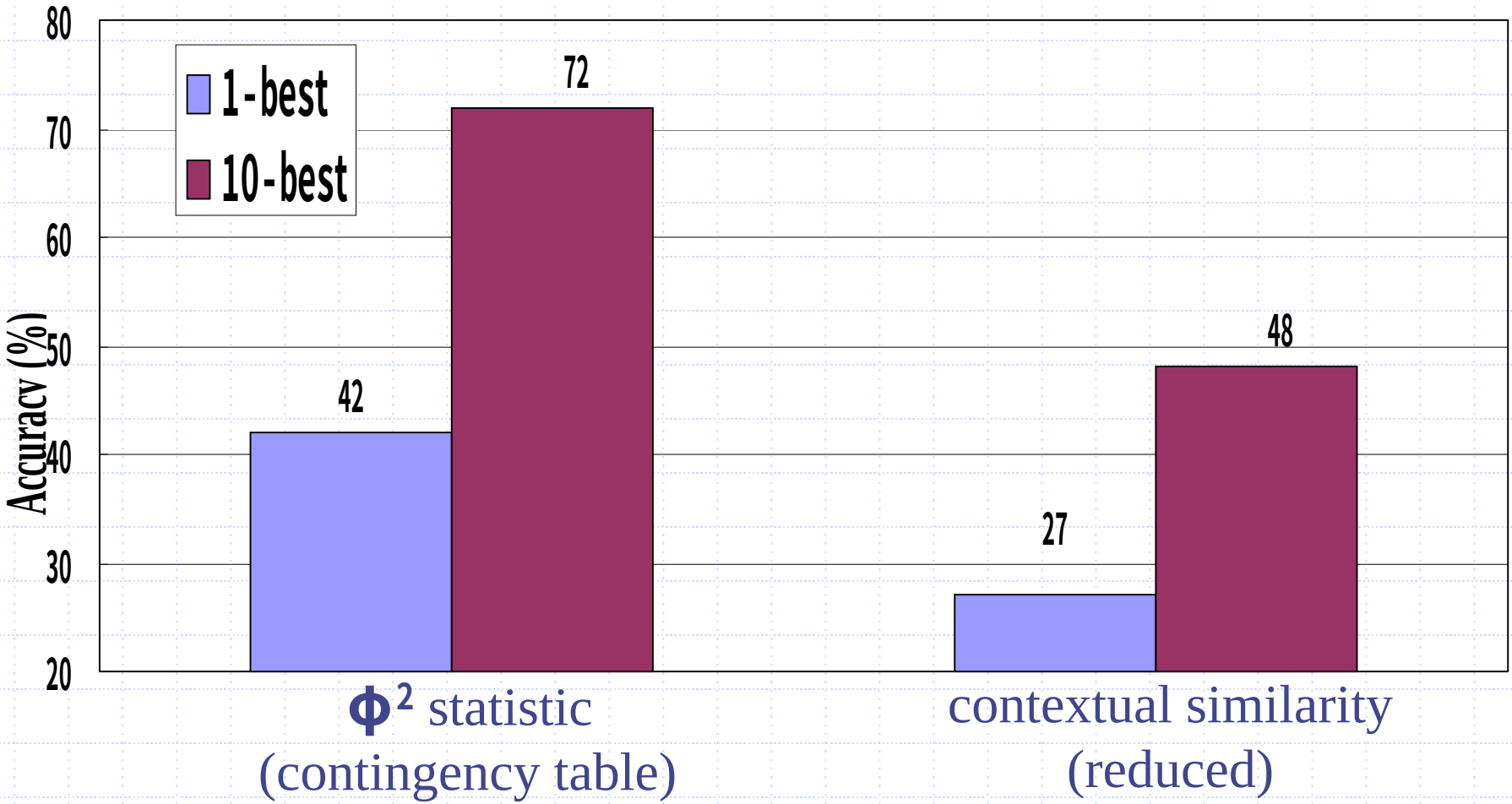
(Manual Evaluation)

(site A, Sim LBD=0.4, for 200 English terms with highest max estimation values)



Accuracy of N-best Bilingual Term Pair Candidates (Manual Evaluation)

(site A, Sim LBD=0.4, for 200 English terms with highest max estimation values)



Bilingual Term Correspondence Acquisition: Current Results Summary

- Effect of CLIR in Bilingual Lexicon Acquisition
  - # of bilingual term pair candidates reduced to 1/10〜1/30
  - most of filtered-out pairs are not correct translation

- Metrics for Bilingual Term Correspondence Estimation
  - accuracy: ϕ^2 (contingency table) *:*
    42% (1-best) and 72% (10-best)
    contextual similarity (reduced):
    27% (1-best) and 48% (10-best)
  - overlap of the estimated bilingual term pairs is less than 30%

Conclusion

- Source of Translation Knowledge Acquisition:
  Cross-Lingually Relevant News Articles
  (on WWW news sites)
- Techniques:
  Integration of CLIR and
  Translation Knowledge Acquisition from
  Parallel/Comparable Corpora

- **Novel** Term Correspondences Can Be Discovered
  - 1.4 times those found
    in an existing bilingual lexicon (0.85M entries)
  - Demo of Semi-automatic Acquisition Tool
    at ACL-2003 Exhibition

Examples of Discovered Correct Term Correspondences

t_E	t_J	f_E	f_J	f_{EJ}	ϕ^2	rank	$TP(t_E)$	$\hat{\phi}^2(TP(t_E))$
～Found in an existing bilingual lexicon～								
Hansen's disease	ハンセン病	3	3	3	1.000	1	36	1.000
mad cow disease	狂牛病	13	16	12	0.681	1	84	0.681
Cabinet Office	内閣府	5	7	3	0.249	4	101	0.357
Yasukuni Shrine	靖国神社	10	15	7	0.310	11	105	0.432
～Not found in an existing bilingual lexicon～								
conference on Afghan reconstruction	アフガン復興会議	3	3	3	1.000	1	43	1.000
New job offers	新規求人	3	3	3	1.000	1	64	1.000
Prime Minister Yoshiro Mori	森総理大臣	16	7	5	0.210	3	45	0.482
Kato faction	加藤派	3	4	3	0.748	2	70	1.000

Current and Future Works

- Incorporating Sophisticated Term Recognition Criteria

- Integrating
  - Article Similarities
  - Several Primitive Measures (contingency table and contextual similarities)

  for more accurate

  Bilingual Term Correspondences Estimation

Publications

- Takehito Utsuro, Takashi Horiuchi, Yasunobu Chiba, and Takeshi Hamamoto.
  Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-Lingually Relevant News Articles on WWW News Sites.
  In S.D.Richardson,editor,Machine Translation: From Research to Real Users, Lecture Notes in Artificial Intelligence:Vol. 2499, pp. 165-176. Springer, October 2002.

- Takehito Utsuro, Takashi Horiuchi, Takeshi Hamamoto, Kohei Hino, and Takeaki Nakayama.
  Effect of Cross-Language IR in Bilingual Lexicon Acquisition from Comparable Corpora. Proc. 9th EACL, pp. 355-362. April 2003.