A Method of Cross-Lingual Question-Answering Based on Machine Translation and Noun Phrase Translation using Web documents

Tatsunori MORI and Kousuke TAKAHASHI

Graduate School of Environment and Information Sciences

Yokohama National University
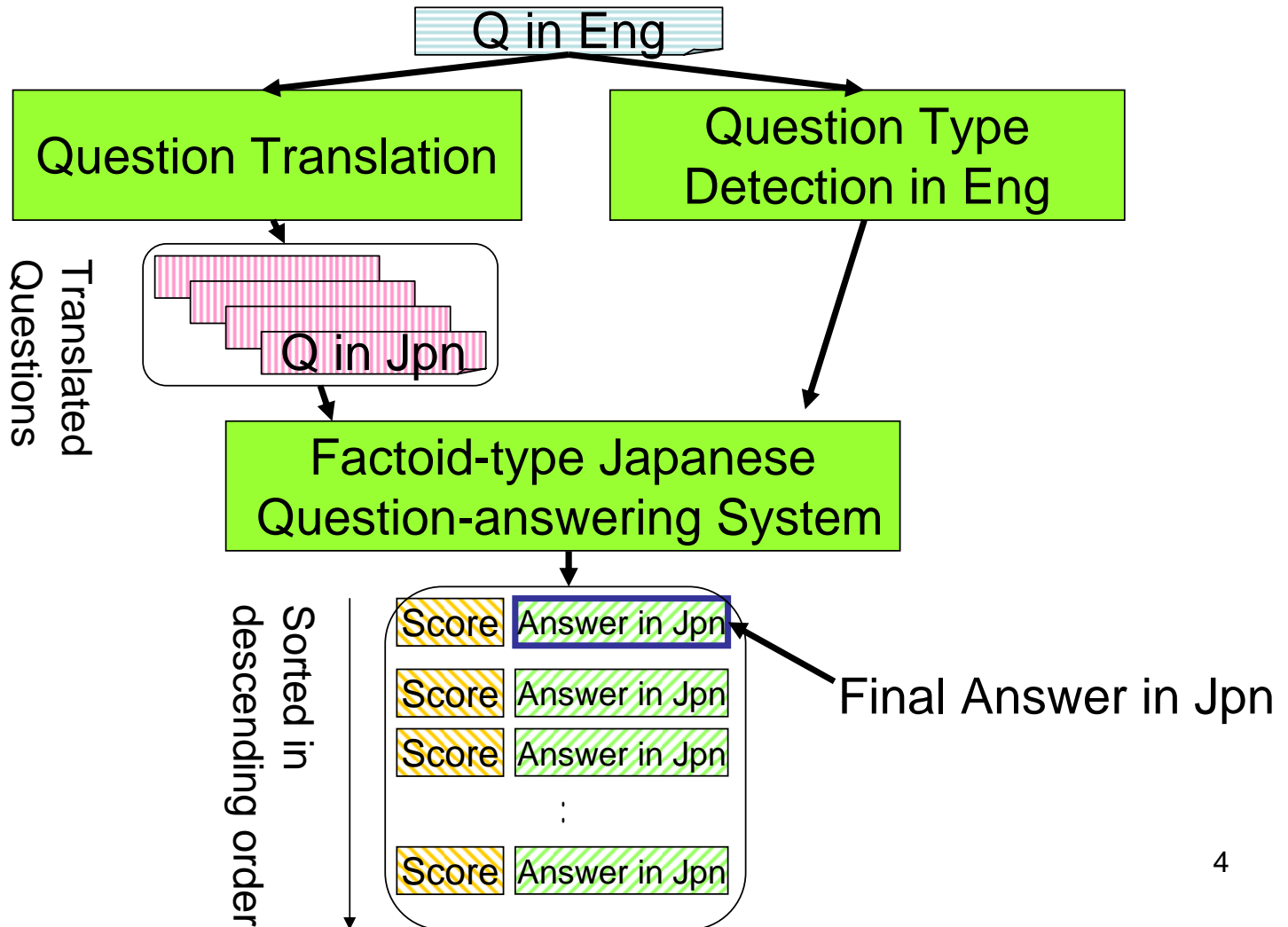
mori@forest.eis.ynu.ac.jp

Introduction and related work

- Cross-lingual Question Answering
  a. For each target language, one individual QA system is prepared. The CL process is achieved as the translation of Qs.
  b. One pivot language is assumed and one QA system is prepared. The CL process appears in the translation of Qs and/or documents.

- While some researches adopt the second approach [Bowden 06, Laurent 06, Shimizu 05, Mori 05], the majority adopts the first approach.

- One of main concerns is the improvement of translation accuracy.

- Web as resource to translate Out-of-Vocabulary (OOV) words
  - Zhan et al. [Zhang 05] proposed a method to obtain translation candidates from the results of a search engine.
  - Bouma et al. [Bouma 06] extracted from English Wikipedia all pairs of lemma titles and cross-links to the corresponding link to Dutch Wikipedia.
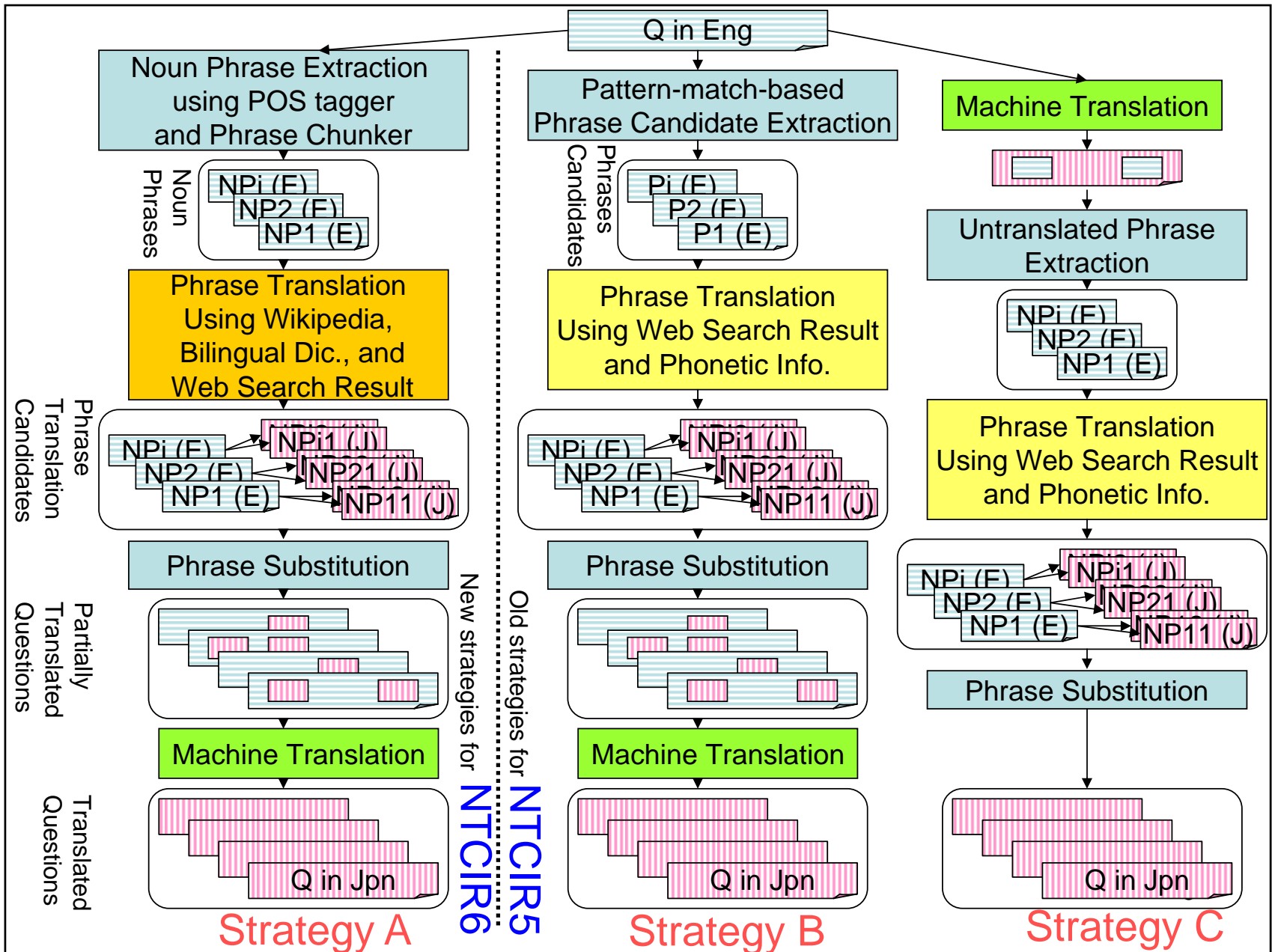
Our approach

- English-Japanese CLQA
- A question translation approach (next slide)
  1. Translate an English Q. into Japanese
  2. Detect the Q. type in the English Q.
  3. Perform Japanese QA with translated Qs.
- Points at issues
  - Treatment of OOV phrases in combination with MT
    - Many off-the-shelf MT products are available.
    - Translation of English Q. into Japanese by using MT.
    - Out-of-vocabulary (OOV) phrases
  - Management of multiple translation candidates in QA phase
    - Different translation strategies of OOV phrases yield different translated Q.

3

A question translation approach

Q in Eng

Question Translation

Question Type Detection in Eng

Translated Questions

Q in Jpn

Factoid-type Japanese Question-answering System

Sorted in descending order

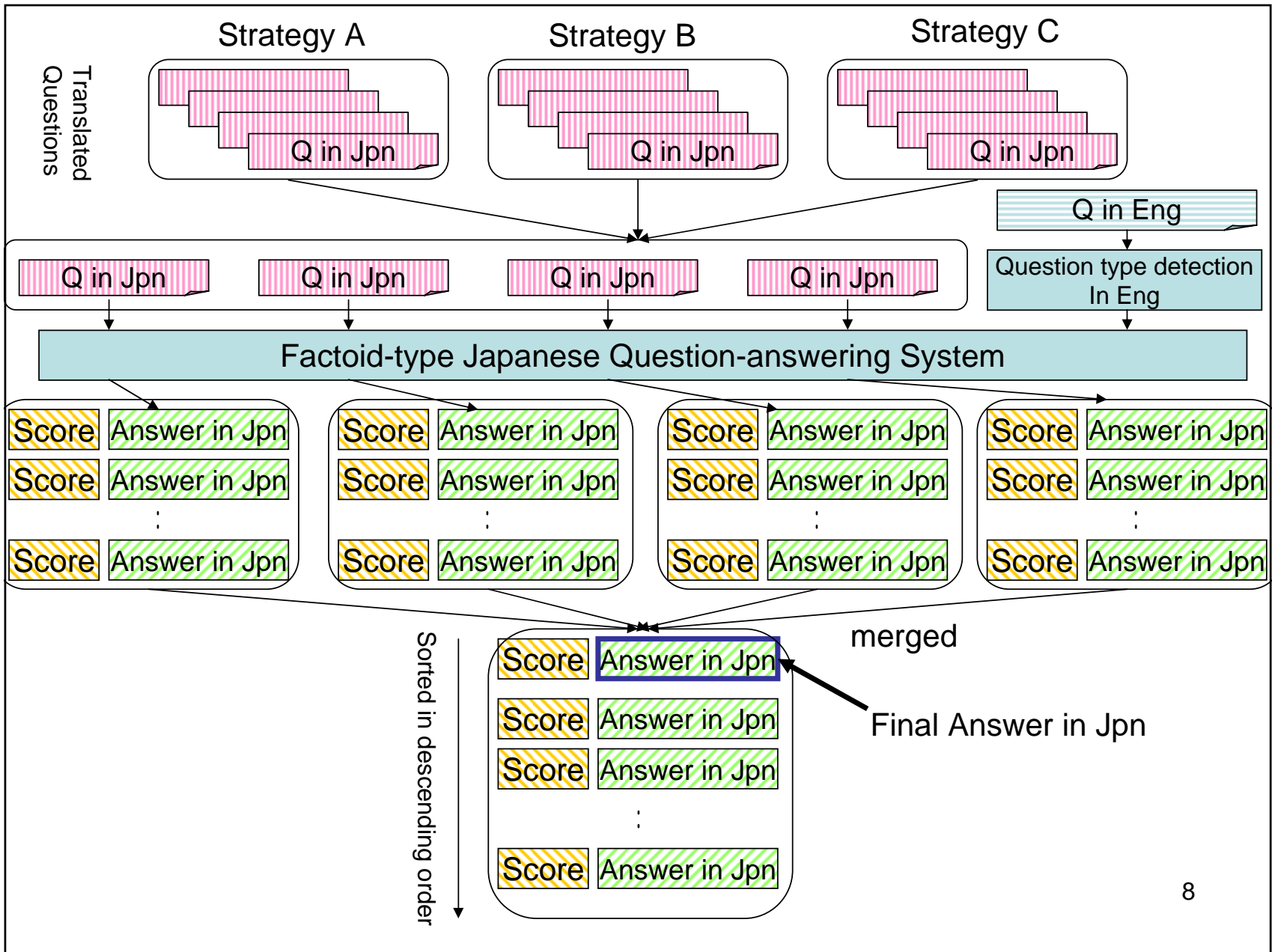| Score	Answer in Jpn
Score	Answer in Jpn
Score	Answer in Jpn
Score	Answer in Jpn

Final Answer in Jpn

4

Treatment of OOV phrase in combination with an MT

- Translation of OOV phrases using external resource
  - There are <u>several different approaches</u> that are worth employing (described later)
- Timing of combining translation of OOV phrases with an MT
  - As a pre-editing process of MT
    - Some of E-J MT systems can treat Japanese strings in an input English sentence as unknown noun phrases and outputs them as they are.
    - *Pre-translation*: originally a technique to utilize Translation Memory
    - Partial translation of noun phrases first, then perform MT
  - As a post-editing process of MT
    - MT first, then translate un-translated noun phrases.
    - We do not have ways to correct translation error in MT.

5

Q in Eng

Strategy A

Noun Phrase Extraction using POS tagger and Phrase Chunker

Noun Phrases: NPi (F), NP2 (F), NP1 (E)

Phrase Translation Using Wikipedia, Bilingual Dic., and Web Search Result

Phrase Translation Candidates: NPi (F), NP2 (F), NP1 (E) → NPi1 (J), NP21 (J), NP11 (J)

Phrase Substitution

Partially Translated Questions

Machine Translation

Translated Questions: Q in Jpn

Strategy B

Pattern-match-based Phrase Candidate Extraction

Phrases Candidates: Pi (F), P2 (F), P1 (E)

Phrase Translation Using Web Search Result and Phonetic Info.

NPi (F), NP2 (F), NP1 (E) → NPi1 (J), NP21 (J), NP11 (J)

Phrase Substitution

Machine Translation

Q in Jpn

Strategy C

Machine Translation

Untranslated Phrase Extraction

NPi (F), NP2 (F), NP1 (E)

Phrase Translation Using Web Search Result and Phonetic Info.

NPi (F), NP2 (F), NP1 (E) → NPi1 (J), NP21 (J), NP11 (J)

Phrase Substitution

Q in Jpn

New strategies for NTCIR6

Old strategies for NTCIR5

Management of multiple translation candidates in QA phase

- Multiple translation candidates of Q. from different translation strategies
  - Which is the best translation? ➔ No criterion
- "Cohesion with information source" approach.
  - Hypothesis 1: if the translation is performed well, some context similar to the translated Q. is likely found in information source.
  - "Answering a question" is finding objects whose context in the information source is coherent with the question.
  - Hypothesis 2: the degree of cohesion with information source is analogous to the appropriateness of the answer candidate.
    - E.g. Score of answer

Strategy A   Strategy B   Strategy C

Translated Questions

Q in Jpn   Q in Jpn   Q in Jpn

Q in Eng

Q in Jpn   Q in Jpn   Q in Jpn   Q in Jpn

Question type detection In Eng

Factoid-type Japanese Question-answering System

Score Answer in Jpn   Score Answer in Jpn   Score Answer in Jpn   Score Answer in Jpn
Score Answer in Jpn   Score Answer in Jpn   Score Answer in Jpn   Score Answer in Jpn
Score Answer in Jpn   Score Answer in Jpn   Score Answer in Jpn   Score Answer in Jpn

merged

Sorted in descending order

Score Answer in Jpn
Score Answer in Jpn
Score Answer in Jpn

Score Answer in Jpn

Final Answer in Jpn

8

Translation strategies

- Strategy A: newly introduced for NTCIR-6 CLQA
  - Performed as the pre-translation process.
  - SVM-based NP chunker to extract all possible NPs.
  - Phrase translation using Wikipedia
  - Phrase translation using Web search results

- Strategy B and C: introduced for NTCIR-5 CLQA
  - Translate loan words into the original Japanese words using Web and the information of pronunciation.
  - B is performed as the pre-translation process.
  - C is performed as the post-translation process.

Phrase translation using Wikipedia

- Wikipedia is a free content encyclopedia, and has a lot of articles in more than 200 languages.
- We can easily obtain multilingual translation of an entry term because of hyper-links [Bouma 06, Fukuhara 07].
  1. To perform the E-J translation, search for the target phrase in the English Wikipedia.
  2. Find out the link to the corresponding Japanese entry.
  3. The name of the Japanese entry is expected to be a proper translation.
- We may use not only English entries but also other entries in different languages that have similar alphabets.

Algorithm 4.1: WIKIPEDIATRANSEJ($PhraseE$)

for each $langCode \in \{$'en', 'de', 'fr', 'es', 'it',
'nl', 'pt', 'lt', 'no', 'pl'$\}$

do $\begin{cases} art \leftarrow Wikipedia(PhraseE, langCode) \\ \textbf{if } (art \text{ has a link to a Japanese entry } PhraseJ) \\ \quad \textbf{then return } (PhraseJ) \end{cases}$

return ('')

Phrase translation using Web search results (1)

- We propose a modification of Zhang's method [Zhang 05].

- Main idea: the case of E-J translation
  - Submit an English phrase to a Web search engine in order to retrieve Japanese documents.
  - Many of retrieved documents are expected to contain not only the English phrase but also Japanese phrases that related to the original English phrase.
  - Scoring method that estimate the appropriateness of the candidate in terms of translation.

Phrase translation using Web search results (2)

Algorithm 4.2: $\text{WebTransCand}(Phrase)$

main

$\langle T, S \rangle \leftarrow CallExternalSEngine(Phrase, N_d)$

comment: T and S are the arrays of titles and snippets, respectively. The index corresponds to the rank in the search result. N_d is the number of document to be retrieved.

$TC \leftarrow \{\}$

for $i \leftarrow 1$ **to** $n-1$

do $\begin{cases} \textbf{for } j \leftarrow i+1 \textbf{ to } n \\ \textbf{do} \begin{cases} LCSS_T \leftarrow LCCS(T_i, T_j) \\ LCSS_S \leftarrow LCCS(T_i, S_j) \\ TC \leftarrow MergeFreq(TC, LCSS_T) \\ TC \leftarrow MergeFreq(TC, LCSS_S) \end{cases} \end{cases}$

return (TC)

comment: $TC = \{\langle C_1, freq_1 \rangle, \langle C_2, freq_2 \rangle, \ldots\}$, where C_i and $freq_i$ are a translation candidate of $Phrase$ and its frequency, respectively.

Candidates: *Longest Common Contiguous Substring of Japanese characters*

Title 1
Snippet 1
Title 2
Snippet 2
Title 3
Snippet 3

Search Result

12
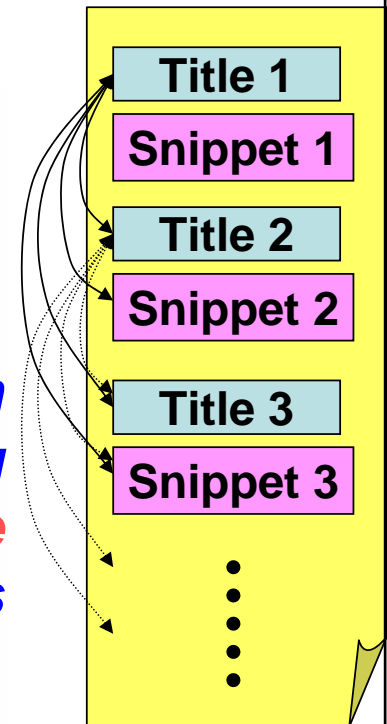
Phrase translation using Web search results (3)

- Assigning score to each candidate
  - Zhang's original score
    - $ITF(C_i)$: Inverse of translation freq. that represents how many times the translation candidate C_i appears in different candidate lists.

$$Sc_{org}(C_i) = \alpha \cdot \frac{freq_i}{maxFreq(TC)} \cdot ITF(C_i)$$

$$+ (1 - \alpha) \cdot \frac{1}{Rank(C_i) + \beta}$$

ITF is properly calculated only when we want to translate a number of phrases simultaneously.

  - Our modification

$$Sc_{rev}(C_i) = \log_2(length(C_i) + \gamma)$$

$$(\alpha \cdot \frac{\log_2(freq_i)}{\log_2(maxFreq(TC))} + (1 - \alpha) \cdot \frac{1 + \beta}{Rank(C_i) + \beta})$$

Since the algorithm tends to produce shorter candidate, we give "reward" to longer one.
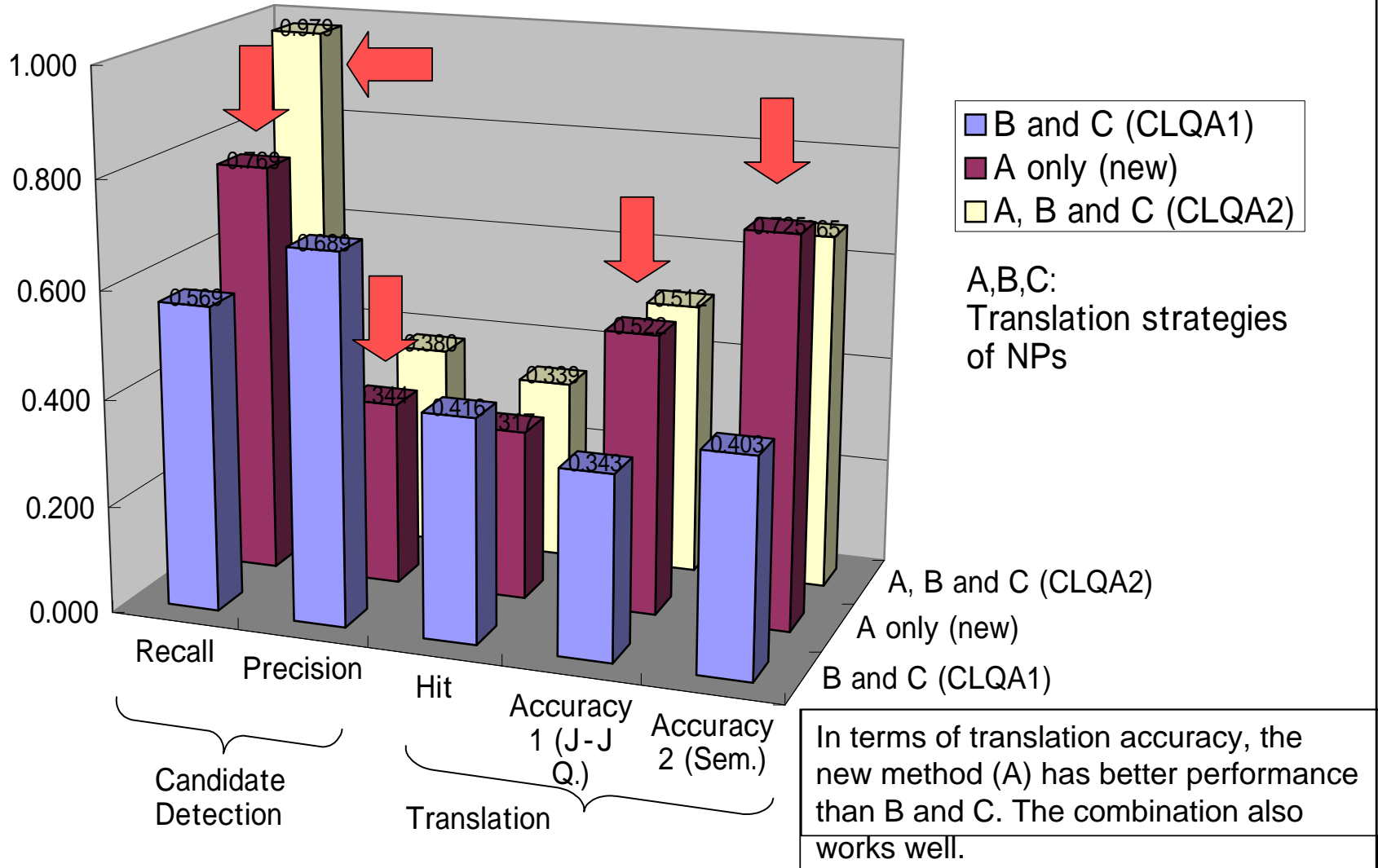
13

Runs at NTCIR-6 CLQA

- Participated in the English-Japanese task.
- Settings
  - An off-the-shelf MT product that has "pre-translation" function (IBM Japan, *Hon'yaku-no Ousama*)
  - EDR E-J translation dictionary
  - A Japanese QA system for factoid Qs. [Mori 05]
  - Strategy A
    - Web search engine: Web service by Yahoo! Japan
  - Strategy B and C
    - The setting is same as our formal run in NTCIR-5 CLQA.
    - Web search engine: Google SOAP Search API.
- Runs
  - Forst-E-J-01: Strategy A, B, and C with MT
  - Forst-E-J-02: Strategy A with MT
  - Forst-E-J-03: Strategy B and C with MT (NTCIR-5 CLQA)
  - Forst-J-J-01: Mono-lingual run. An upper bound.
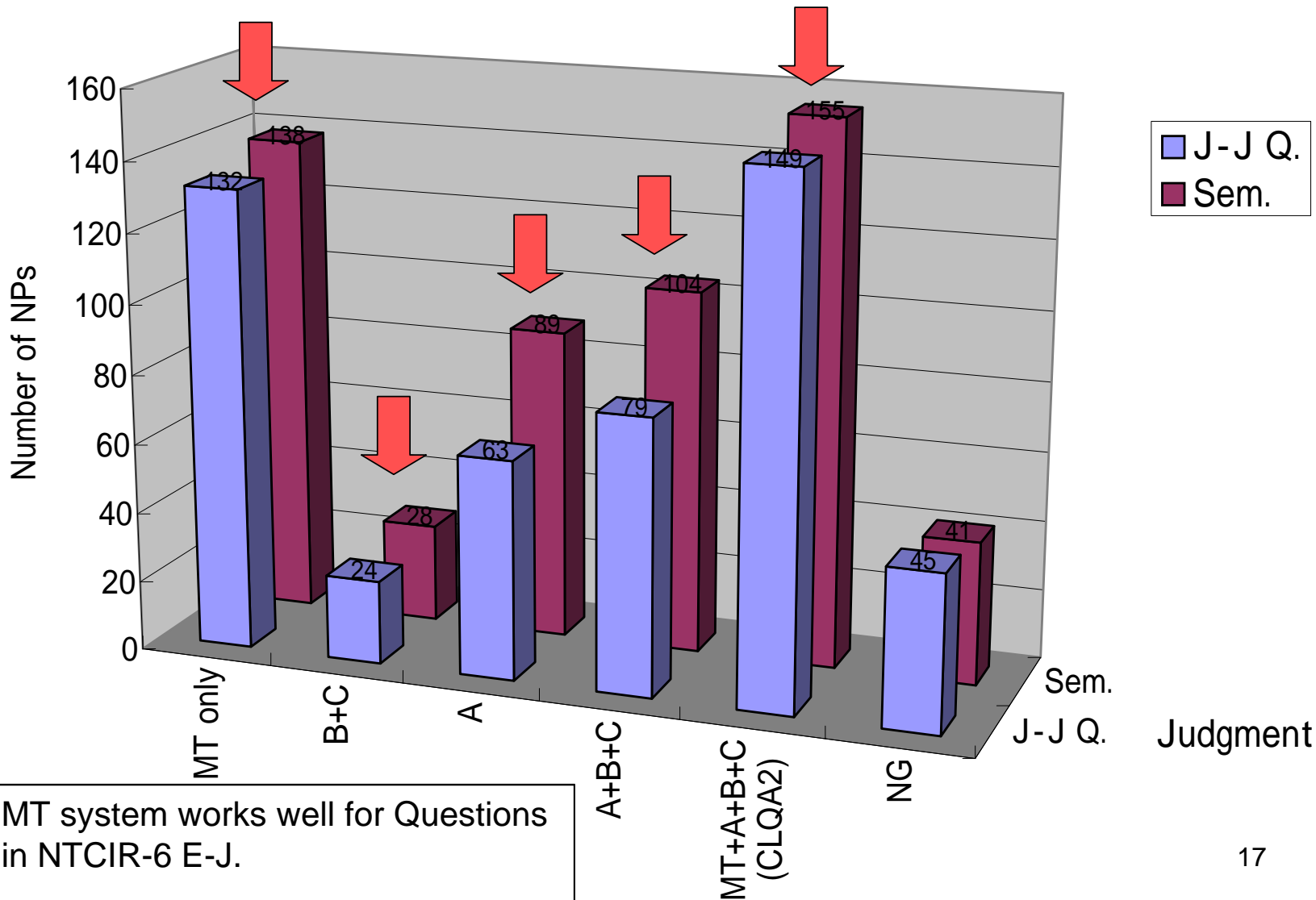 - Baseline: MT only

14

Performance of proper noun translation

- Measures for evaluation of proper noun detection
  - Recall and precision
- Measures for evaluation of proper noun translation
  - *Hit*: ratio of # of phrases to which the system can find at least one translation candidate.
  - *Trans. Accuracy 1*: ratio of # of phrases for which the system can find at least one "correct" translation. "correct" when the translation is the correspondent phrase in J-J Q. (strict)
  - *Trans. Accuracy 2*: same as *1*, but the correctness is judged semantically. (lenient)

Performance in translation of proper NPs

Number of correctly translated proper NPs

Number of NPs

Judgment

Legend: J-J Q. | Sem.

Data labels:
- MT only: 132, 138
- B+C: 24, 28
- A: 63, 89
- A+B+C: 79, 104
- MT+A+B+C (CLQA2): 149, 155
- NG: 45, 41

MT system works well for Questions in NTCIR-6 E-J.

17

Number of correctly translated proper NPs which the MT cannot translate

22 proper nouns are newly correctly translated in the case of combination A+B+C.

18
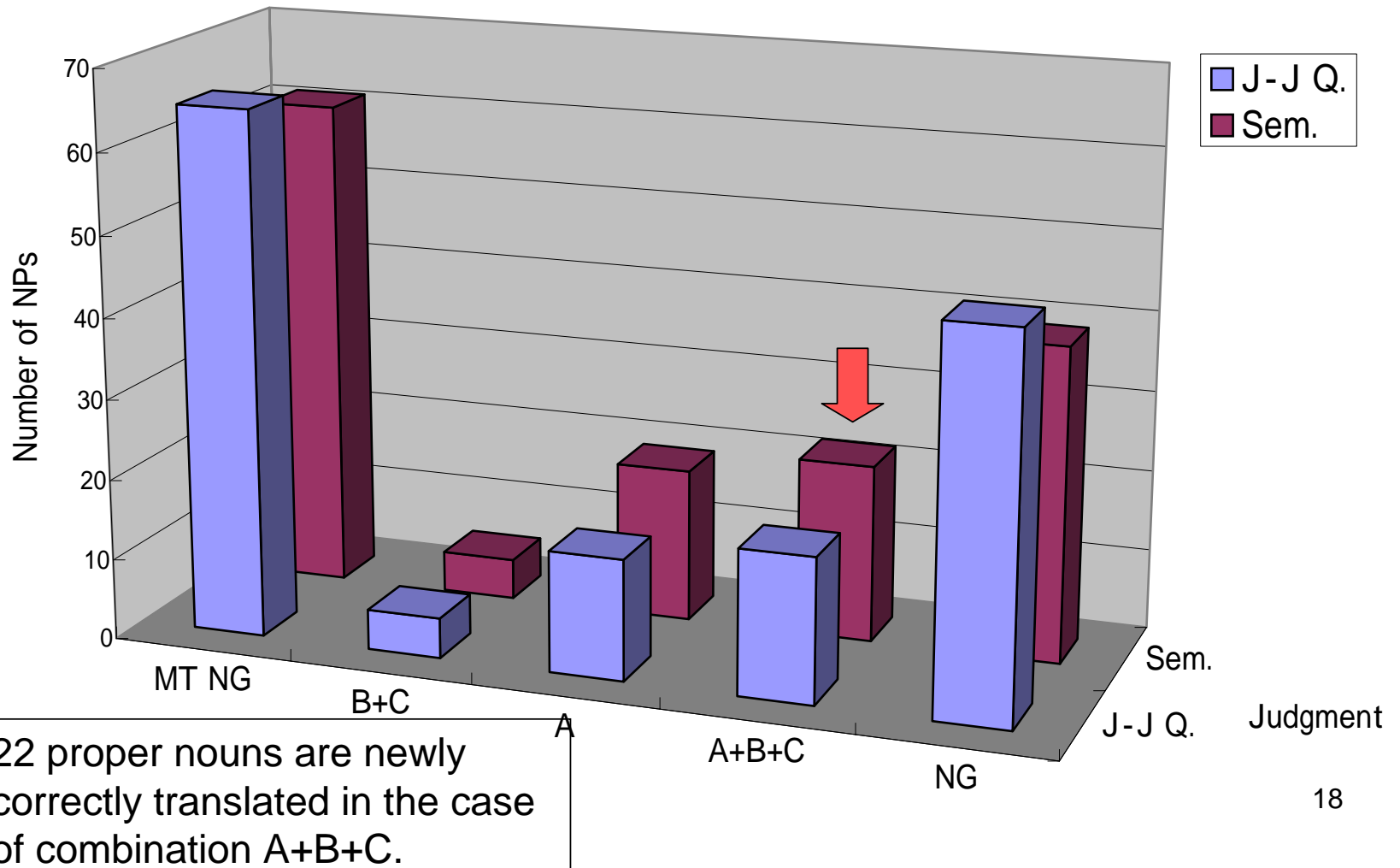
Performance in E-J CLQA

Strategy	Run ID	Acc	MRR	TOP5	Acc+ U	MRR +U	TOP5 +U
MT only		0.175	0.195	0.230	0.185	0.230	0.315
MT+B+C (CLQA1)	Forst-E-J-03	0.170	0.193	0.235	0.180	0.229	0.325
MT+A	Forst-E-J-02	0.170	0.192	0.230	0.180	0.231	0.325
MT+A+B+C	Forst-E-J-01	0.175	0.197	0.230	0.195	0.244	0.320
JJ QA	Forst-J-J-01	0.310	0.361	0.440	0.335	0.410	0.525

Acc: Accuracy
+U: Unsupported answers are allowed
JJ QA: Japanese monolingual QA system with correct Japanese questions.

- Although "MT+A+B+C" has better performance than others, the difference between it and "MT only" is not significant.

- MT system works well and the actual improvement by phrase translation is small.

Failure in extracting NPs.

- Adjacent proper nouns are extracted as one phrase
  - *Question*: "Where did former Spice Girl Posh Spice hold her wedding ceremony?"
  - *Extracted NP*: "Spice Girl Posh Spice"
  - *Correct NPs*: "Spice Girl" and "Posh Spice"

Failure in phrase translation by using Wikipedia

- Translation using Wikipedia mostly works well, when it is applicable.

- It has unwilling tendency to translate a NP into an official name of translation instead of a popular translation.

  – *Phrase*: "Akutagawa Prize"

  – *Translated*: "akutagawa ryunosuke shou" (
)

  – *More popular translation*: "akutagawa shou"(   )

Failure in phrase translation by using Web search result

- The method tends to fail in translation of longer NPs.
  - NP: "University of Hawaii at Manoa"
  - Translated: "hawai daigaku" (   )
  - Correct one: "hawai daigaku manoa kou" (
)

- It also tends to translate a phrase into a related phrase.
  - NP: "FIFA president"
  - Translated: "sakkaa" (football,   )
  - Correct one: "FIFA kaichou" (FIFA   )

Concluding remarks

- English-Japanese (E-J) task with three systems.
  - Basis of approach: MT + an existing Japanese QA system.
  - Methods for noun phrase translation using the Web.
- The combination works well.
- MT system also works well for Qs in NTCIR-6 E-J.