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.
A question translation approach
Treatment of OOV phrase in combination with an MT

• Translation of OOV phrases using external resource
  – There are several different approaches 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.
Noun Phrase Extraction using POS tagger and Phrase Chunker

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

Phrase Translation Candidates

Phrase Substitution

Machine Translation

Untranslated Phrase Extraction

Phrase Translation Using Web Search Result and Phonetic Info.

Phrase Substitution

Old strategies for NTCIR5

New strategies for NTCIR6

Pattern-match-based Phrase Candidate Extraction

Noun Phrases

NP1 (E)
NP2 (E)
NP1 (E)

NP1 (J)
NP2 (J)
NP1 (J)

NP1 (E)
NP2 (E)
NP1 (E)

NP1 (J)
NP2 (J)
NP1 (J)

NP1 (E)
NP2 (E)
NP1 (E)

NP1 (J)
NP2 (J)
NP1 (J)

NP1 (E)
NP2 (E)
NP1 (E)

NP1 (J)
NP2 (J)
NP1 (J)

NP1 (E)
NP2 (E)
NP1 (E)
Management of multiple translation candidates in QA phase

• Multiple translation candidates of Q. from different translation strategies
  – Which is the best translation? \(\rightarrow\) 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
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 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)

\[
\begin{align*}
& \text{art } \leftarrow \text{Wikipedia}(\text{PhraseE, langCode}) \\
& \text{do } \{ \text{if (art has a link to a Japanese entry PhraseJ)} \} \\
& \text{then return (PhraseJ)} \\
& \text{return (**)}
\end{align*}
\]
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: WEBTRANSAND(Phrase)

main
\[ \langle T, S \rangle \leftarrow \text{CallExternalSEnsemble}(\text{Phrase}, N_d) \]
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 \( n - 1 \)
for \( j \leftarrow i + 1 \) to \( n \)
do
\[ \text{LCSS}_T \leftarrow \text{LCSS}(T_i, T_j) \]
\[ \text{LCSS}_S \leftarrow \text{LCSS}(T_i, S_j) \]
do
\[ TC \leftarrow \text{MergeFreq}(TC, \text{LCSS}_T) \]
\[ TC \leftarrow \text{MergeFreq}(TC, \text{LCSS}_S) \]
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 \text{Phrase} and its frequency, respectively.

Candidates: **Longest Common Contiguous Substring of Japanese characters**
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.
      \[
      S_{org}(C_i) = \alpha \cdot \frac{freq_i}{\maxFreq(TC)} \cdot ITF(C_i) + (1 - \alpha) \cdot \frac{1}{\text{Rank}(C_i) + \beta}
      \]

  – Our modification
    \[
    S_{rev}(C_i) = \log_2(\text{length}(C_i) + \gamma) - \alpha \cdot \frac{\log_2(freq_i)}{\log_2(\maxFreq(TC))} + (1 - \alpha) \cdot \frac{1 + \beta}{\text{Rank}(C_i) + \beta}
    \]

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

  Since the algorithm tends to produce shorter candidate, we give "reward" to longer one.
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)
  - Baseline: MT only
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)
Since the newly introduced method (A) detects all NP candidates, the recall is higher but the precision is lower in the detection. The combination method A+B+C can detect almost all proper noun. In terms of translation accuracy, the new method (A) has better performance than B and C. The combination also works well.
The new strategy has better coverage in translation than the strategy in CLQA1 (B+C). Combination of translation strategies improves the coverage of proper noun translation. MT system works well for Questions in NTCIR-6 E-J.
22 proper nouns are newly correctly translated in the case of combination A+B+C.
### Performance in E-J CLQA

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Strategy</th>
<th>Acc+UTOP5+MRR+U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.325</td>
<td>MT+A+B+C</td>
<td>0.180</td>
</tr>
<tr>
<td>0.320</td>
<td>MT+A</td>
<td>0.195</td>
</tr>
<tr>
<td>0.325</td>
<td>MT+B+C</td>
<td>0.192</td>
</tr>
<tr>
<td>0.315</td>
<td>MT only</td>
<td>0.175</td>
</tr>
</tbody>
</table>

**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” (芥川龍之介賞)
  - *More popular translation*: “akutagawa shou” (芥川賞)
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” (ハワイ大学)
  – Correct one: “hawai daigaku manoa kou” (ハワイ大学マノア校)

• It also tends to translate a phrase into a related phrase.
  – NP: “FIFA president”
  – Translated: “sakkaa” (football, サッカー)
  – Correct one: “FIFA kaichou” (FIFA会長)
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.