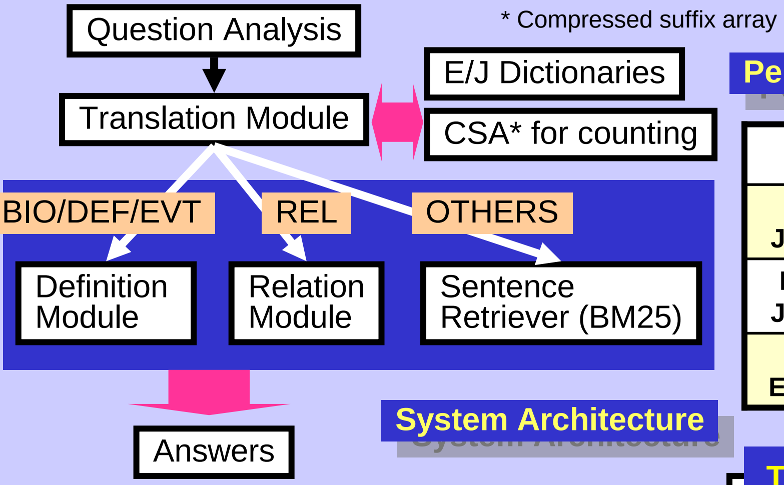
NTT's CCLQA System for NTCIR-7 ACLIA

Ryuichiro Higashinaka and Hideki Isozaki
NTT Communication Science Laboratories, NTT Corporation, Japan

Overview

- We built our CCLQA (EN-JA/JA-JA) system based on the technologies used in our past NTCIR systems.
  - DEFINITION, BIOGRAPHY, and EVENT questions ⇒ we reused our definition module for QAC-4.
  - RELATIONSHIP questions ⇒ we developed a new module based on our why-QA approach for QAC-4.
  - Other questions ⇒ we used a simple sentence retriever based on BM25.
  - English questions are translated into Japanese using translation dictionaries.
- Our EN-JA system performed rather poorly, but our JA-JA system showed reasonable performance.



* Compressed suffix array

System Architecture

Performance of our JA-JA/EN-JA Systems

	DEF	BIO	REL	EVT	ALL
Our JA-JA	0.289	0.179	0.221	0.092	0.187
Best JA-JA	0.420	0.190	0.233	0.094	0.220
Our EN-JA	0.170	0.093	0.048	0.002	0.068

Definition Module

Definition module extracts adnominal/adverbial modifiers for target X using TGREP2.

Example: "President Suharto (スハルト大統領)"
⇒ 32 年間にわたって，人口約2 億人のインドネシアを牛耳ってきたスハルト大統領
(President Suharto, who has been ruling Indonesia with the population of 200 million for 32 years)

Relation Module

We train a classifier that detects the mention of relationship from the EDR corpus using BACT (classifier of trees).

Positive Examples: sentences that have semantic categories corresponding to "関係 (relationship)"
Negative Examples: others

Scoring of an answer candidate (C):

$$\mathrm{score}(C) = \mathrm{score}_{\mathrm{rel}}(C) + \mathrm{score}_{\mathrm{sim}}(C)$$

BACT's score indicating how likely relationship is expressed in C

Word-overlap-based similarity between the question and C

Translation Module

If the dictionary does not have the target, we use the left-to-right longest match.

Example: "Next Generation Network"
(1) "Next Generation Network" is consulted
(2) If the dictionary does not have this entry, "Next Generation" is consulted.
⇒ We get "次世代".
(3) The remaining "Network" is consulted
⇒ we get "ネットワーク".
As a result, we get "次世代ネットワーク".

Failure Analysis

Failures are mainly caused by the fragility of our English question analyzer and dictionary-based translator.

(1) Inappropriate dictionary look-up results
"Martina Navratilova"
⇒ "マルティナナヴラティロワ"
(no occurrence in the target corpus.)
(2) No entry in E/J dictionaries
"Suharto" ⇒ "スハルト"
E-J transliteration module can save this case.
(3) Transliteration fails in some cases
"embryonic stem cells" ⇒ "ES 細胞"