Statistical Machine Translation based Passage Retrieval — Experiment at NTCIR-7 IR4QA Task

Tatsuhiro Hyodo Tomoyosi Akiba
Dept. of Information and Computer Sciences,
Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi,
441-8580, JAPAN
hyodo@cl.ics.tut.ac.jp

Abstract

In this paper, we apply the statistical machine translation based passage retrieval, which was proposed at the last NTCIR-6 CLQA subtask, to the IR4QA Task. The experimental evaluation shows that the method is more effective for the relation and event type questions, which are longer and including relatively mane common keywords, than the definition and biography type questions, which are shorter and often including only named entities.

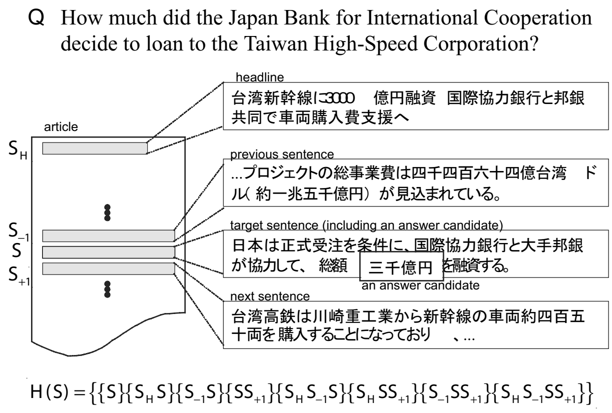
Q How much did the Japan Bank for International Cooperation decide to loan to the Taiwan High-Speed Corporation?
headline
台湾新幹線に3000　億円融資　国際協力銀行と邦銀共同で車両購入費支援へ
article
S_H
previous sentence
...プロジェクトの総事業費は四千四百六十四億台湾　ドル（ 約一兆五千億円）が見込まれている。
S_{-1}
S
S_{+1}
target sentence (including an answer candidate)
日本は正式受注を条件に、国際協力銀行と大手邦銀が協力して、 総額 三千億円 を融資する。
an answer candidate
next sentence
台湾高鉄は川崎重工業から新幹線の車両約四百五十両を購入することになっており　、...
$H(S)=\{\{S\}\{S_H S\}\{S_{-1}S\}\{SS_{+1}\}\{S_H S_{-1}S\}\{S_H SS_{+1}\}\{S_{-1}SS_{+1}\}\{S_H S_{-1}SS_{+1}\}\}$

Figure 1. An examples of a question and the corresponding passage candidates.

1 Introduction

In the previous NTCIR-6, we introduced the novel approach for CLQA, in which the statistical machine translation (SMT) is deeply incorporated into the question answering process, instead of using it as the preprocessing of the mono-lingual QA process (Akiba et al., 2008). Our approach consists of the cross-language document retrieval method and the cross-language passage retrieval method, both of which use the word translation model trained by the parallel corpus between Japanese and English. In the current NTCIR-7, we applied such methods to IR4CLQA Task, i.e. a cross-language document retrieval task.

In the rest of this paper, Section 2 overviews our methods for document and passage retrieval for CLQA used in the previous NTCIR CLQA subtask. Section 3 describes the experimental evaluation conducted to see the performance of the proposed method by comparing it to some reference methods. Section 4 describes our conclusion.

2 Our Method

2.1 Document Retrieval

Given an English question sentence, the document retrieval subsystem of our proposed CLQA system retrieves Japanese documents directly. In order to do so, each Japanese document in the target collection has been indexed by English terms by using the word translation probability used in the SMT framework.

The expected term frequency $tf(e, D)$ of an English term e that would be used as an index to a Japanese document D can be estimated by the following equation.

$$tf(e, D) = \sum_{j \in D} t(e|j) tf(j, D) \qquad (1)$$

where $tf(j, D)$ is the term frequency of a Japanese term j in D and $t(e|j)$ is the word translation probability that j is translated into e. The probability $t(e|j)$ is trained by using a large parallel corpus as the SMT framework. Because the expected term frequency $tf(e, D)$ is consistent with $tf(j, D)$ that is calculated from the statistics of D, the conventional vector space IR model based on the TF-IDF term weighting can be used for implementing our IR subsystem. We used *GETA* [1] as the IR engine in our system.

2.2 SMT based Passage Retrieval

In order to enable the direct passage retrieval, where the query and the passage are in different languages,

[1] http://geta.ex.nii.ac.jp

the statistical machine translation is utilized to calculate the similarity between them. In order words, we calculate the similarity between them as the probability that the Japanese passage is translated into the English question.

The similarity $sim(Q, S|A)$ between a question Q and a sentence S including an answer candidate A is calculated by the following equation.

$$sim(Q, S|A) = \max_{D \in H(S)} P(Q|D - A) \qquad (2)$$

where $P(Q|D - A)$ is the probability that a word sequence D except A is translated into a question sentence Q, and $H(S)$ is the set of the candidate passage (term sequences) that are related to a sentence S. The set consists of S and the power set of S_H, S_{-1}, and S_{+1}, where S_H is the headline of the article that S belongs, S_{-1} is the previous sentence of S, and S_{+1} is the next sentence of S (Figure 1).

In this paper, we use IBM model 1 (Brown et al., 1993) in order to get the probability $P(Q|D - A)$ as follows,

$$P(Q|D - A) = \frac{1}{(n+1)^m} \prod_{j=1}^{m} \sum_{i=1,\cdots,k-1,k+l+1,\cdots,n} t(q_j|d_i) \qquad (3)$$

where $q_1 \cdots q_m$ is a English term sequence of a question Q, $d_1 \cdots d_n$ is a Japanese term sequence of a candidate passage D, $d_k \cdots d_{k+l}$ is a Japanese term sequence of an answer candidate A. Therefore, the Japanese term sequence $d_1, \cdots, d_{k-1}, d_{k+l+1}, \cdots, d_n$ (= D - A) is just D except A. We exclude the answer term sequence A from the calculation of the translation probability, because the English terms that corresponds to the answer should not be appeared in the question sentence as the nature of question answering.

Moreover, we interpolate the translation probability and the prior of the question terms for smoothing as follows,

$$P(Q|D - A) = \frac{1}{(n+1)^m} \prod_{j} \sum_{i} \{\lambda t(q_j|d_i) + (1-\lambda) u(q_j)\} \qquad (4)$$

where $u(q)$ is a uni-gram probability trained from the question (English) side of the parallel corpus.

3 Experimental Evaluation

The experimental evaluation was conducted to see the document retrieval performance by using our proposed method.

3.1 Task and Data

The NTCIR-7 ACLIA formal run data (Sakai et al., 2008) for English to Japanese CLIR task were used for the evaluation. This data contains 98 question in English. The target documents are four years newspaper articles from “MAINICHI SHINBUN” (1998-2001).

We used either **Question** or **Narrative** for queries.

3.2 Translation Model

The translation model used for our method was trained from the following English-Japanese parallel corpus; 206,782 sentence pairs from newspaper articles obtained by the automatic sentence alignment (Utiyama and hitoshi Isahara, 2003).

Before training the translation model, both English and Japanese sides of the sentence pairs in parallel corpus were normalized. For the sentences of Japanese side, the inflectional words were normalized to their basic forms by using a Japanese morphological analyzer. For the sentences of English side, the inflectional words were also normalized to their basic forms by using a Part-of-Speech tagger and all the words were lowercased. For each side, we deleted the words that has the functional word class. In Japanese side, we deleted “conjunction”, “particle”, “auxiliary verb” and “interjection”. In English side, we deleted “conclusion”, ”definite article” and “postfix”. GIZA++ (Och and Ney, 2003) was used for training the IBM model 4 from the resulting normalized parallel corpus. The trained Japanese-to-English word translation model $t(e|j)$ was used for our proposed document retrieval (Section 2.1) and passage similarity calculation (Section 2.2).

3.3 Compared methods

We submitted three runs (TA-EN-JA-01-D, TA-EN-JA-02-D and TA-EN-JA-03-T) to the IR4QA Task. Then, we tried additional 2 different runs (DR300-D and PR1000-D).

DR1000-Q Using **Question** as a query, our **document retrieval** is applied to obtain the top 1000 documents. (submit run ID is TA-EN-JA-03-T)

DR1000-N Using **Narrative** as a query, our **document retrieval** is applied to obtain the top 1000 documents. (submit run ID is TA-EN-JA-02-D)

PR300-N Using **Narrative** as a query, our document retrieval is applied to obtain the top 300 documents. Then, our **passage retrieval** is applied to rerank the result. (submit run ID is TA-EN-JA-01-D)

DR300-N Using **Narrative** as a query, our document retrieval is applied to obtain the top 300 documents.

PR1000-N Using **Narrative** as a query, our document retrieval is applied to obtain the top 1000 documents. Then, our **passage retrieval** is applied to rerank the result.

Table 1. The performances of the proposed and reference CLQA systems with respect to IR for QA task.

Method	AP	Q	nDCG
DR1000-Q	0.0127	0.0155	**0.0446**
DR1000-N	**0.0141**	0.0155	0.0337
PR300-N	0.0115	0.0119	0.0268
DR300-N	0.0131	0.0134	0.0241
PR1000-N	0.0102	0.0118	0.0353

Table 2. The performances of each question type in AP.

Method	DEF	BIO	REL	EVE
DR1000-Q	6.6e-05	0.0016	0.0220	0.0202
DR1000-N	1.9e-05	0.0017	**0.0293**	0.0176
PR300-N	3.1e-04	0.0021	0.0065	**0.0318**
DR300-N	9.5e-06	0.0005	**0.0293**	0.0150
PR1000-N	**4.4e-04**	**0.0030**	0.0046	0.0284

For TA-EN-JA-01-D and DR300-N, we used only top 300 documents instead of 1000 documents because of the time limitation.

3.4 Results

Table 1 shows the results of our system.

Firstly, we compared the results obtained by DR1000-Q and DR1000-N. In terms of AP, **Narrative (N)** is better than **Question (N)**. But in terms of nDCG, **Question (Q)** is better.

Secondly, we compared the results obtained by PR300-N and DR300-N. The result of passage retrieval get worse in terms of AP and Q. But nDCG result get better.

As a whole, our methods perform poorly for the NTCIR-7 IR4QA data, while they performed better for the past NTCIR-5 CLQA1 and NTCIR-6 CLQA2 data. it seems the difference on performance comes from the difference on question types; the past CLQA series targeted factoid questions, while the IR4QA targets nonfactoid questions.

Table 2 shows the results type by type. It shows that the performance is worse on DEF (Definition) and BIO (Biography) types, while better on REL (Relation) and EVE (Event). The DEF and BIO questions tend to be short (Table 3) and to include only named entities as the available key words to find the relevant documents. Currently, our method cannot deploy the unseen named entities by itself, as it only depends on the training parallel corpus to find their translations. Therefore, it fails to find any keyword translations for such questions. On the other hand as the REL, EVE and factoid questions are longer and have much common keywords, our method works better on them.

Table 3. The proportion and the average length for each type of questions.

Type	Number of questions	Average length(Q)	Average length(N)
DEF	21	4.71	14.61
BIO	20	5.10	10.90
REL	29	12.58	23.55
EVE	28	12.17	24.50
TOTAL	98	9.07	18.90

4 Conclusion

In this paper, we applied the statistical machine translation based passage retrieval to the NTCIR-7 IR4QA Task, which had been proposed at the last NTCIR-6 CLQA subtask. The experimental evaluation showed that the method was more effective for the relation and event type questions, which were longer and including relatively many common keywords, than the definition and biography type questions, which were shorter and often including only named entities.

Because our method cannot deploy the unseen named entities by itself, it should be incorporate with the method that handles the unseen named entities for document retrieval, in future work.

References

Tomoyosi Akiba, Kei Shimizu, and Atsushi Fujii. 2008. Statistical machine translation based passage retrieval. In *Proceedings of 3rd International joint Conference on Natural Language Processing*, pages 751–756.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 18(4):263–311.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

T. Sakai, N. Kando, C.-J Lin, T. Mitamura, D. Ji, K.-H. Chen, and E. Nyberg. 2008. Overview of the ntcir-7 aclia subtask. In *Proceedings of NTCIR-7*.

Masao Utiyama and hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 72–79.