An Investigation of Machine Translation Evaluation Metrics in Cross-lingual Question Answering

Kyoshiro Sugiyama, Masahiro Mizukami, Graham Neubig, Koichiro Yoshino, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura
Graduate School of Information Science
Nara Institute of Science and Technology
Takayamacho 8916-5, Ikoma, Nara
{sugiyama.kyoshiro.sc7, neubig}@is.naist.jp

Abstract

Through using knowledge bases, question answering (QA) systems have come to be able to answer questions accurately over a variety of topics. However, knowledge bases are limited to only a few major languages, and thus it is often necessary to build QA systems that answer questions in one language based on an information source in another (cross-lingual QA: CLQA). Machine translation (MT) is one tool to achieve CLQA, and it is intuitively clear that a better MT system improves QA accuracy. However, it is not clear whether an MT system that is better for human consumption is also better for CLQA. In this paper, we investigate the relationship between manual and automatic translation evaluation metrics and CLQA accuracy by creating a data set using both manual and machine translations and perform CLQA using this created data set.[1] As a result, we find that QA accuracy is closely related with a metric that considers frequency of words, and as a result of manual analysis, we identify 3 factors of translation results that affect CLQA accuracy.

1 Introduction

Question answering (QA) is the task of searching for an answer to question sentences using some variety of information resource. Generally, documents, web pages, or knowledge bases are used as these information resources. When the language of the question differs from the language of the information resource, the task is called cross-lingual question answering (CLQA) (Magnini et al., 2004; Sasaki et al., 2007). Machine translation (MT) is one of the most widely used tools to achieve CLQA (Mori and Kawagishi, 2005; Fujii et al., 2009; Kettunen, 2009).[2]

In the realm of monolingual question answering, recent years have seen a large increase in the use of structured knowledge bases such as Freebase (Bollacker et al., 2008), as they allow for accurate answering of questions over a variety of topics (Frank et al., 2007; Cai and Yates, 2013). However, knowledge bases are limited to only a few major languages. Thus, CLQA is particularly important for QA using knowledge bases.

In contrast to the CLQA situation, where an MT system is performing translation for a downstream system to consume, in standard translation tasks the consumer of results is a human (Matsuzaki et al., 2015). In this case, it is important to define an evaluation measure which has high correlation with human evaluation, and the field of MT metrics has widely studied which features of MT results are correlated with human evaluation, and how to reflect these features in automatic evaluation (Machácek and Bojar, 2014).

However, translations which are good for humans may not be suitable for question answering. For example, according to the work of Hyodo and Akiba (2009), a translation model trained using a parallel corpus without function words achieved higher accuracy than a model trained using full sentences on CLQA using documents or web pages, although it is not clear whether these results will apply to more structured QA using knowledge bases. There is also work on optimizing translation to improve CLQA accuracy (Riezler et al., 2014; Haas and Riezler, 2015), but these methods require a large set of translated question-answer pairs, which may not be available in many

[1] All data used in the experiments will be released upon publishing of the paper.

[2] MT is also used in mono-lingual QA tasks when question sentences are translated into the formal language used to query the information resource (Andreas et al., 2013).

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 442–449,
Lisboa, Portugal, 17-18 September 2015. ©2015 Association for Computational Linguistics.

languages. Correspondingly, it is of interest to investigate which factors of translation output affect CLQA accuracy, which is the first step towards designing MT systems that achieve better accuracy on the task.

In this paper, to investigate the influence of translation on CLQA using knowledge bases, we create a QA data set in which each question has been translated both manually and by a number of MT systems. We then perform CLQA using this data set and investigate the relationship between translation evaluation metrics and QA accuracy. As a result, we find that QA accuracy is closely related to NIST score, a metric that considers the frequency of words, indicating that proper translation of infrequent words has an important role in CLQA tasks using knowledge bases. In addition, as a result of fine-grained manual analysis, we identify a number of factors of translation results that affect CLQA.

2 Data sets

To create data that allows us to investigate the influence of translation on QA, we started with a standard QA data set, and created automatic and manual translations. In this section, we describe the data construction in detail.

As our seed data, we used a data set called Free917 (Cai and Yates, 2013). Free917 is a question set made for QA using the large-scale knowledge base "Freebase," and is widely used in QA research (Cai and Yates, 2013; Berant et al., 2013). It consists of 917 pairs of question sentences and "logical forms" which are computer-processable expressions of the meaning of the question that can be fired against the Freebase database to return the correct answer. Following Cai and Yates (2013), we divide this data into a training set (512 pairs), dev set (129 pairs) and test set (276 pairs). In the remainder of the paper, we refer to the questions in the test set before translation as the original (OR) set.

Next, to investigate the influence of translation quality on the accuracy of QA, we created a question set with five different varieties of translation results. First we translated the question sentences included in the OR set into Japanese manually (the JA set). Then, we created translations of the JA set into English by five different methods:

Manual translation We asked a professional translation company to manually translate the questions from Japanese to English (the HT set).

GT and YT The questions are translated using Google Translate[3] (GT) and Yahoo Translate[4] (YT) systems, these commercial systems can be used via web pages. While the details of these systems are not open to the public, it is likely that Google takes a largely statistical MT approach, while the Yahoo engine is rule-based.

Moses The questions are translated using a phrase-based system built using Moses (Koehn et al., 2007) (the Mo set). A total of 277 million sentences from various genres are used in training.

Travatar The questions are translated using Travatar (Neubig, 2013) (the Tra set), a tool for forest-to-string MT that has achieved competitive results on the Japanese-English language pair. The training data is the same as Moses.

Table 1: A sample of translations and logical forms in the test set

Set	Question	Logical form
OR	what is europe 's area	(location.location.area en.europe)
JA	ヨーロッパの面積は	
HT	what is the area of europe	
GT	the area of europe	
YT	the area of europe	
Mo	the area of europe	
Tra	what is the area of europe	

3 QA system

To perform QA, we used the framework of Berant et al. (2013), as implemented in SEMPRE. [5] SEMPRE is a QA system that has the ability to use large-scale knowledge bases, such as Freebase.

In this section, we describe the framework briefly and consider how translation may affect each element of it. We show an example of how this system works in Figure 1.

Alignment A lexicon, which is a mapping from natural language phrases to logical predicates, is constructed using a large text

[3] https://translate.google.co.jp/
[4] http://honyaku.yahoo.co.jp/
[5] http://nlp.stanford.edu/software/sempre/

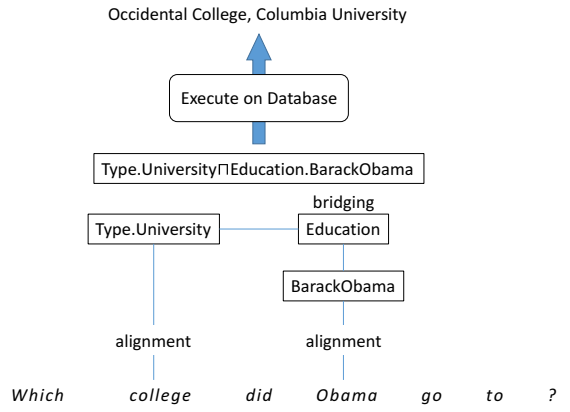

Figure 1: Framework of the SEMPRE semantic parsing system used to perform QA

corpus, which is linked to the knowledge base through the use of named entity prediction. By default, SEMPRE uses ClueWeb09[6] (Callan et al., 2009) as the large text corpus and Freebase as the knowledge base. During the QA process itself, this lexicon is used to convert entities into logical forms through a process called alignment.

Translation has the potential to affect this part by changing the words in the translation. Because the strings in the sentence are used to look up which logical form to use, a mistranslated word may result in a failure in lookup.

Bridging To create the query for the knowledge base, SEMPRE merges neighboring logical forms in a binary tree structure. Bridging is an operation that generates predicates compatible with neighboring predicates.

Translation has the potential to affect this operation by changing the word order in the translation. Because adjacent logical forms are combined in the bridging process, the different word order may cause changes in the combination of logical forms.

Scoring and learning The previous two steps are not deterministic, and thus the system must select the best of many candidates. Scoring evaluates candidates according to a scoring function, and learning is optimization of the weights used in the scoring function.

It is possible that translation also affects this process, with a different set of weights being ideal for CLQA than monolingual QA. On the other hand, to train these weights it is necessary to have a translated version of the QA training set, which represents a significant investment, and thus we do not examine this within the scope of this paper.

[6]http://www.lemurproject.org/clueweb09.php/

4 Experiments

In our experiments, we examine the effect of various features of translation quality on CLQA. To do so, we use the data sets described in Section 2, and we performed QA with the system described in Section 3. In the experiments, we suppose a situation in which Japanese question sentences are translated into English and inputted into an English-language QA system.

4.1 Result 1: Evaluation of translation quality

First, we evaluate translation quality of each system using 4 automatic evaluation measures BLEU+1 (Lin and Och, 2004), WER (Leusch et al., 2003), NIST (Doddington, 2002) and RIBES (Isozaki et al., 2010) and manual evaluation of acceptability (Goto et al., 2013).

BLEU+1 BLEU (Papineni et al., 2002) is the most popular automatic evaluation metric of machine translation quality, and BLEU+1 is a smoothed version that can be used with single sentences. It is based on n-gram precision, and the score is from 0 to 1, where 0 is the worst and 1 is the best.

WER Word error rate (WER) is the edit distance between the translation and reference normalized by the sentence length. The formula of WER is as follows:

$$WER = \frac{S+D+I}{N}$$

where

- S is the number of substitutions.
- D is the number of deletions.
- I is the number of insertions.
- N is the number of word in the reference.

The score is a real number more than 0, and can be over 1 when the length of the output is larger than the reference. Like BLEU, WER focuses on matches between words, but

is less lenient with regards to word ordering, having a strong performance for linear matches between the two sentences. WER is an error rate, thus lower WER is better. To adjust direction of axis to match the other measures, we use the value of $1 - WER$.

RIBES RIBES is a metric based on rank correlation coefficient of word order in the translation and reference, and thus focuses on whether the MT system was able to achieve the correct ordering. It has been shown effective for the evaluation of language pairs with greatly different structure such as Japanese and English. The score is from 0 to 1, where 0 is the worst and 1 is the best.

NIST NIST is a metric based on n-gram precision and each n-gram's weight. Rarer n-grams have a higher weight. Therefore, less frequent words such as content words are given more importance than function words such as "of," "in," and others. The score is a real number more than 0.

Acceptability Acceptability is a 5-grade manual evaluation metric. It combines aspects of both fluency and adequacy, with levels 1-3 evaluating semantic content, and 3-5 evaluating syntactic correctness.

Figure 2 shows the result of the evaluation for each system. Note that NIST and Acceptability have been normalized between 0 and 1 by dividing by the highest possible achievable value.

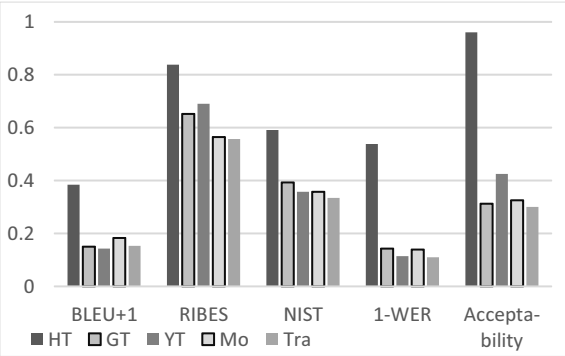

Figure 2: Evaluation scores (mean)

From this, we can see that HT has the best score on all metrics. Indicating that human translation is still more accurate than machines in this language pair and task. Next comes commercial systems, with GT being the 2nd best on BLEU and NIST, while YT is higher than GT on RIBES and manual evaluation. This confirms previous reports (Isozaki et al., 2010) that RIBES is well correlated with human judgments of acceptability for Japanese-English translation tasks. In the next section, we examine whether this observation also holds when it is not a human but a computer doing the language understanding.

4.2 Result 2: QA accuracy

Next, we performed QA using the created data sets. We found that for 12 questions in the test set even the correct logical form did not return any answer, so we eliminate these questions and analyze the remaining 264 questions.

Figure 3 shows QA accuracy of each data set.

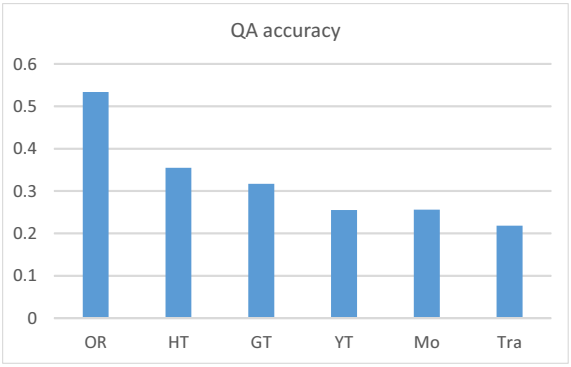

Figure 3: QA accuracy of each data set

Here, we can see that accuracy of the OR set is about 53%. Accuracy of the HT set is the highest of the translated data sets. However, although HT has high translation quality, its accuracy is significantly ($p < 0.01$ according to the Student's t-test) lower than OR. YT is the second for acceptability but its accuracy is lower than GT and Mo. This indicates that there is, in fact, a significant difference between translations that are good for humans, and those that are good for QA systems.

In the next section, we analyze these phenomena in detail.

5 Discussion

5.1 Correlation between translation quality and QA accuracy

First, we analyze the sentence-level correlation between evaluation scores and QA accuracy to attempt to gain more insights about the features of translation results that affect QA accuracy, and potential implications for evaluation. One thing to

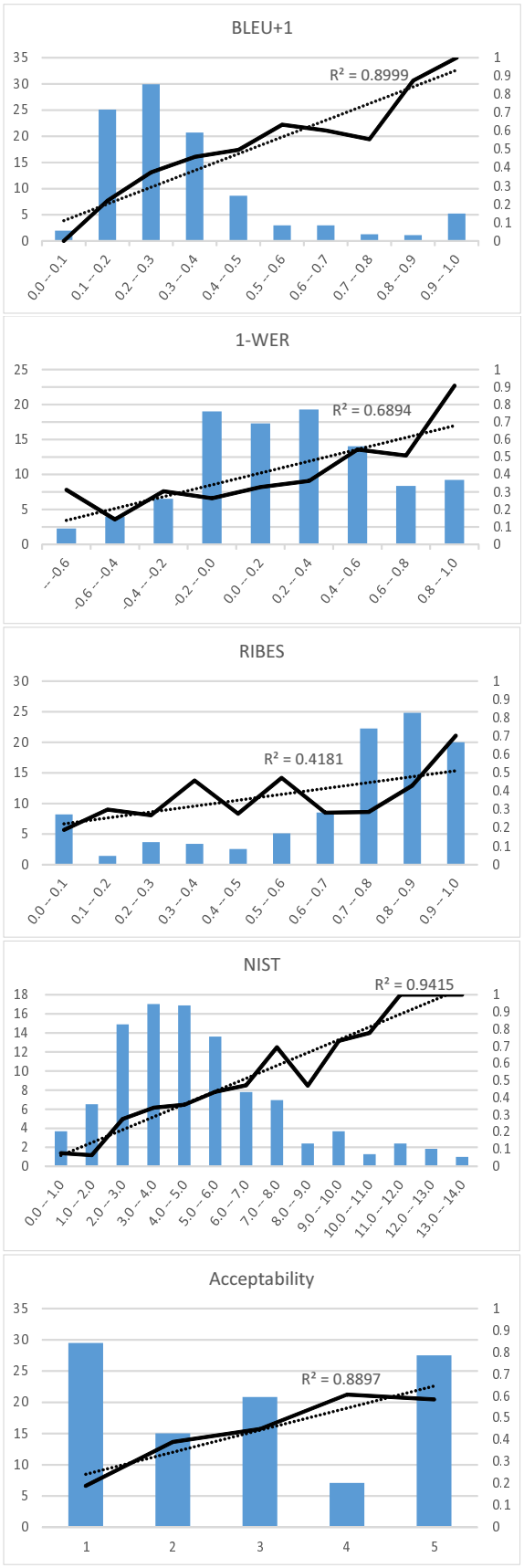

Figure 4: Correlation between QA accuracy and evaluation score (correct group)
Horizontal axis: Range of evaluation score
Bar (left axis): Percentage of # questions
Line (right axis): Rate of QA accuracy (average in the range)

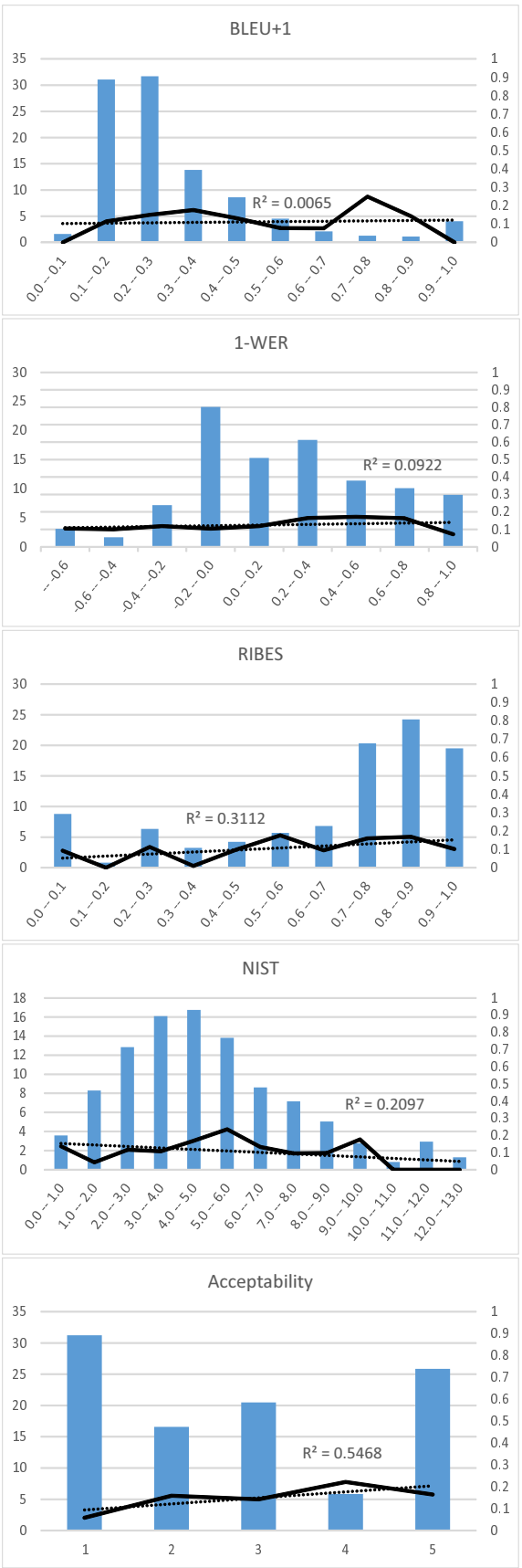

Figure 5: Correlation between QA accuracy and evaluation score (incorrect group)
Horizontal axis: Range of evaluation score
Bar (left axis): Percentage of # questions
Line (right axis): Rate of QA accuracy (average in the range)

be noted first is that even with the original set OR, only approximately half of the questions were answered correctly, and thus in some cases the question might be difficult to answer even with the correct translation result. To take this effect into account, we divide the questions in two groups. The "correct" group consists of $141 * 5 = 705$ translated questions of the 141 question answered correctly in OR and the "incorrect" group consists of $123 * 5 = 615$ translated questions of the remaining 123 questions.

Figure 4 shows correlation between QA accuracy and evaluation score of the correct group. The bar graphs indicate the percentage of the number of the questions in each range of evaluation scores. From these figures, we can first note that there is some correlation between all investigated evaluation metrics and QA accuracy, demonstrating that translation accuracy is, in fact, important for CLQA. We can also see that QA accuracy is most closely related to NIST score. Recall that NIST is a metric that considers the frequency of each word, resulting in content words being treated as more important than function words. According to this result, it seems that content words are important for translation in CLQA tasks, which is natural given the importance of matching entities in the alignment step of Section 3. It is also encouraging that NIST score also seems to be effective at assessing this automatically.

On the other hand, RIBES, which has higher correlation with human evaluation as shown in Section 4, has the lowest correlation with CLQA accuracy. Thus, we can see that the overall order of words might not be as important in translation for CLQA. In other words, looking back at the QA framework in Section 3, this means that the "alignment" process is likely more sensitive to errors than the "bridging" process, which may not be affected as heavily by word order.

Figure 5 shows correlation between QA accuracy and evaluation score of the incorrect group. In contrast to the correct group, in the incorrect group, QA accuracy has very little correlation with all of the scores. Even the manually evaluated adequacy score has only moderate correlation. These results show that if the reference sentences cannot be answered correctly, the sentences are not suitable, even for negative examples. Thus, when evaluating MT systems for CLQA, we may benefit from creating a set of references that are answered correctly by the system before performing evaluation.[7]

5.2 Case studies

In this section, we show some examples of QA results that changed as a result of translation. In addition, we consider what causes the change and implications for evaluation.

Table 2: Examples of changes in content words

○	OR when was interstate 579 formed
-	JA 州間高速道路 579 号が作られたのはいつですか
×	HT when was interstate highway 579 made
×	GT when is the interstate highway no. 579 has been made
×	YT when is it that expressway 579 between states was made
×	Mo interstate highway 579) was made when
○	Tra when interstate 579) was built
○	OR who was the librettist for the magic flute
-	JA 魔笛の台本を作成したのは誰ですか
×	HT who wrote the libretto to the magic flute
×	GT who was it that created the script of the magic flute
×	YT who is it to have made a script of the the magic flute
×	Mo the magic flute scripts who prepared
×	Tra who made of magic script
○	- who librettist magic flute

Table 2 shows the examples of change of content words. In the first example, the phrase "interstate 579" has been translated in various ways (e.g. "interstate highway 579," "expressway 579," ...). Only OR and Tra have the phrase "interstate 579" and have been answered correctly. The output logical forms of other translations lack the entity of the highway "interstate 579," mistaking it for another entity. For example, the phrase "interstate highway 579" is instead aligned to the entity of the music album "interstate highway." Similarly, in the second example, the translations that don't have "librettist" were answered incorrectly. Here, we created a new sentence, "who librettist magic flute," which was answered correctly.

These observations show that the change of content words to the point that they do not match entities in the entity lexicon is a very important problem. To ameliorate this problem, it may be possible to modify the translation system to consider the named entity lexicon as a feature in the translation process.

Next, we show examples of another common cause of mis-answered questions in Table 3. In the

[7] It should be noted that the shapes of the translation accuracy distributions of two groups are similar, therefore, it is difficult for MT evaluation metrics to help to choose better datasets.

Table 3: Examples of mis-translated question words

○ OR how many religions use the bible
- JA 聖書を使う宗教はいくつありますか
× HT how many religions use sacred scriptures
○ GT how many religions that use the bible
○ YT how many religion to use the bible are there
○ Mo how many pieces of religion, but used the bible
× Tra use the bible religions do you have

○ OR how many tv programs did danny devito produce
- JA ダニー・デヴィートは何件のテレビ番組をプロデュースしましたか
○ HT how many television programs has danny devito produced
× GT danny devito or has produced what review television program
× YT did danni devito produce several tv programs
× Mo what kind of tv programs are produced by danny devito
× Tra danny devito has produced many tv programs

first example, the sentence of Tra has all the content words of OR, but was answered incorrectly. Likewise, in the second example, "tv (television) programs," "danny devito," and "produce(d)" have appeared in all translations. However, these translations have been answered incorrectly, other than HT. It can be seen that to answer these questions correctly, the sentence must include a phrase such as "how many," which indicates the question type. This demonstrates that correct translation of question words is also important. It should be noted that these words are frequent, and thus even NIST score will not be able to perform adequate evaluation, indicating that other measures may be necessary.

Table 4: Examples of translations with mistaken syntax

○ OR what library system is the sunset branch library in
- JA サンセット・ブランチ図書館はどの図書館システムに所属しますか
○ HT to what library system does sunset branch library belong
○ GT sunset branch library do you belong to any library system
○ YT which library system does the sunset branch library belong to
○ Mo sunset branch library, which belongs to the library system
○ Tra sunset branch library, belongs to the library system?

× OR what teams did babe ruth play for
- JA ベイブ・ルースはどのチームの選手でしたか
× HT what team did babe ruth play for
○ GT did the players of any team babe ruth
○ YT was babe ruth a player of which team
○ Mo how did babe ruth team
○ Tra babe ruth was a team player

Table 4 shows examples regarding syntax. In the first example, all of the sentences were answered correctly, while GT, Mo, and Tra are grammatically incorrect. On the other hand, in the second example, the sentences of OR and HT are grammatically correct, but were answered incorrectly. The OR and HT translations resulted in the QA system outputting Babe Ruth's batting statistics, probably because "babe ruth" and "play" are adjacent in sentences. These cases indicate that, at least for the relatively simple questions in Free917, achieving correct word ordering plays only a secondary role in achieving high QA accuracy.

6 Conclusion

To investigate the influence of translation quality on QA using knowledge bases, we created question data sets using several varieties of translation and compared them with regards to QA accuracy. We found that QA accuracy has high correlation with NIST score, which is sensitive to the change of content words, although these results only hold when evaluating with references that actually result in correct answers. In addition, by analysis of examples, we found 3 factors which cause changes of QA results: content words, question types, and syntax. Based on these results, we can make at least two recommendations for the evaluation of MT systems constructed with cross-lingual QA tasks in mind: 1) NIST score, or another metric putting a weight an content words should be used. 2) References that are actually answerable by the QA system should be used.

We should qualify this result, however, noting the fact that the results are based on the use solely of the SEMPRE parsing system. While SEMPRE has shown highly competitive results on standard QA tasks, we also plan to examine other methods such as Berant and Liang (2014)'s semantic parsing through paraphrasing, which may be less sensitive to superficial differences in surface forms of the translation results. We also plan to to optimize machine translation systems using this analysis, possibly through incorporation into the response-based learning framework of Riezler et al. (2014).

Acknowledgment

Part of this work was supported by the NAIST Big Data Project and by the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

Jacob Andreas, Andreas Vlachos, and Stephen Clark. 2013. Semantic parsing as machine translation. In *Proc. of ACL*, pages 47–52.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proc. of ACL*, volume 7, pages 1415–1425.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proc. of EMNLP*, pages 1533–1544.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. of SIGMOD*, pages 1247–1250.

Qingqing Cai and Alexander Yates. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In *Proc. of ACL*, pages 423–433.

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. Clueweb09 data set.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of HLT*, pages 138–145.

Anette Frank, Hans-Ulrich Krieger, Feiyu Xu, Hans Uszkoreit, Berthold Crysmann, Brigitte Jörg, and Ulrich Schäfer. 2007. Question answering from structured knowledge sources. *Journal of Applied Logic*, 5(1):20–48.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2009. Evaluating effects of machine translation accuracy on cross-lingual patent retrieval. In *Proc. of SIGIR*, pages 674–675.

Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K Tsou. 2013. Overview of the patent machine translation task at the NTCIR-10 workshop. In *Proc. of NTCIR-10*, pages 260–286.

Carolin Haas and Stefan Riezler. 2015. Response-based learning for machine translation of open-domain database queries. In *Proc. of NAACL HLT*, pages 1339–1344.

Tatsuhiro Hyodo and Tomoyosi Akiba. 2009. Improving translation model for smt-based cross language question answering. In *Proc. of FIT*, volume 8, pages 289–292.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proc. of EMNLP*, pages 944–952.

Kimmo Kettunen. 2009. Choosing the best mt programs for clir purposes–can mt metrics be helpful? In *Proc. of ECIR*, pages 706–712.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL*, pages 177–180.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proc. of MT Summit IX*, pages 240–247.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: A method for evaluating automatic evaluation metrics for machine translation. In *Proc. of COLING*, pages 501–507.

Matouš Machácek and Ondrej Bojar. 2014. Results of the WMT14 metrics shared task. *WMT 2014*, pages 293–301.

Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Penas, Víctor Peinado, Felisa Verdejo, and Maarten de Rijke. 2004. The multiple language question answering track at CLEF 2003. In *Comparative Evaluation of Multilingual Information Access Systems*, pages 471–486. Springer.

Takuya Matsuzaki, Akira Fujita, Naoya Todo, and Noriko H Arai. 2015. Evaluating machine translation systems with second language proficiency tests. In *Proc. of ACL*, pages 145–149.

Tatsunori Mori and Masami Kawagishi. 2005. A method of cross language question-answering based on machine translation and transliteration. In *Proc. of NTCIR-5*.

Graham Neubig. 2013. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proc. of ACL*, pages 91–96.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.

Stefan Riezler, Patrick Simianer, and Carolin Haas. 2014. Response-based learning for grounded machine translation. In *Proc. of ACL*.

Yutaka Sasaki, Chuan-Jie Lin, Kuang-hua Chen, and Hsin-Hsi Chen. 2007. Overview of the NTCIR-6 cross-lingual question answering (CLQA) task. In *Proc. of NTCIR-6*, volume 6.