ASPEC: Asian Scientific Paper Excerpt Corpus

Toshiaki Nakazawa*, Manabu Yaguchi*,
Kiyotaka Uchimoto, Masao Utiyama**, Eiichiro Sumita****
Sadao Kurohashi†, Hitoshi Isahara††
*Japan Science and Technology Agency
5-3, Yonbancho, Chiyoda-ku, Tokyo, 102-8666, Japan
nakazawa@pa.jst.jp, yaguchi@jst.go.jp

**National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{uchimoto, mutiyama, eiichiro.sumita}@nict.go.jp

†Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
kuro@i.kyoto-u.ac.jp

††Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku, Toyohashi, Aichi, 441-8580, Japan
isahara@tut.jp

Abstract

In this paper, we describe the details of the ASPEC (Asian Scientific Paper Excerpt Corpus), which is the first large-size parallel corpus of scientific paper domain. ASPEC was constructed in the Japanese-Chinese machine translation project conducted between 2006 and 2010 using the Special Coordination Funds for Promoting Science and Technology. It consists of a Japanese-English scientific paper abstract corpus of approximately 3 million parallel sentences (ASPEC-JE) and a Chinese-Japanese scientific paper excerpt corpus of approximately 0.68 million parallel sentences (ASPEC-JC). ASPEC is used as the official dataset for the machine translation evaluation workshop WAT (Workshop on Asian Translation).

Keywords: parallel corpus, scientific paper, asian languages

1. Introduction

The parallel corpus is no doubt an essential resource for almost all the current machine translation systems. There are many parallel corpora available online[1] for various language pairs and domains. However, there is no available parallel corpus which includes a variety of *scientific paper* domains so far.

There are plenty of useful scientific and technical documents which are written in languages other than English, and are referenced domestically. Accessing these domestic documents in other countries is very important in order to know what has been accomplished and what is needed next in the science and technology fields. However, we need to surmount the language barrier to directly access these valuable documents. One obvious way to achieve this is using machine translation systems to translate foreign documents into the users' language.

ASPEC (Asian Scientific Paper Excerpt Corpus) is the first large-size parallel corpus of scientific paper domain in the world[2]. This corpus is constructed by the Japan Science and Technology Agency (JST) in collaboration with the National Institute of Information and Communications Technology (NICT). It consists of a Japanese-English scientific paper abstract corpus of approximately 3 million parallel sentences (ASPEC-JE) and a Chinese-Japanese scientific paper excerpt corpus of approximately 0.68 million parallel sentences (ASPEC-JC). ASPEC is free and available online[3] for non-commercial purposes.

There are some related work which try to automatically extract scientific paper domain parallel corpus using scientific paper Web portals such as MEDLINE (Jimeno Yepes et al., 2013) and Elsevier (Morin and Prochasson, 2011). However the corpus size is not large enough to achieve the satisfactory machine translation quality of the paper abstracts. In addition, the domain of the corpora are limited. They do not contain a variety of scientific paper domains.

In this paper, we describe the details of the ASPEC and briefly introduce the Workshop on Asian Translation (WAT) as an application of ASPEC.

2. Background of the Corpus Construction

This corpus is one of the fruits of the Japanese-Chinese machine translation project conducted between 2006 and 2010 using the Special Coordination Funds for Promoting Science and Technology. The goal of this project was to promote and evolve the science and technology exchange between Japan and Asian countries by 1) constructing a large-size parallel dictionary for technical terms and 2) developing a machine translation system applicable to scientific papers. The research topics included the parallel corpus construction for the machine translation system.

In the beginning of the project, Japanese-English parallel corpus was constructed because it is easier to col-

[1] http://www.statmt.org/moses/?n=Moses.LinksToCorpora
[2] It is released in January, 2014
[3] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

LangPair	Train	Dev	DevTest	Test
ASPEC-JE	3,008,500	1,790	1,784	1,812
ASPEC-JC	672,315	2,090	2,148	2,107

Table 1: Statistics of ASPEC.

lect the language resouces of Japanese-English than those of Japanese-Chinese. It was used for the investigation and evaluation of the parallel dictionary construction method and machine translation system. In parallel with this, Japanese-Chinese parallel corpus was gradually constructed, and finally Japanese-Chinese parallel dictionary and machine translation system was developed.

There had already been Japanese-English parallel abstracts of various scientific fields, and the Japanese-English parallel corpus was constructed by extracting the parallel sentences from the abstract pairs. Therefore, the ASPEC-JE contains parallel sentences of all the scientific fields. Note that the source language of the abstract pairs for ASPEC-JE is Japanese and it is manually translated into English. This is due to the copyright issue.

On the other hand, since there was no pre-existing resource for Japanese-Chinese, ASPEC-JC was constructed by manually translating the Japanese documents into Chinese from scratch. ASPEC-JC only includes "Medicine", "Information", "Biology", "Environmentology", "Chemistry", "Materials", "Agriculture" and "Energy" fileds because it was difficult to include all the scientific fields. These fields were selected by investigating the important scientific fields in China and the use tendency of literature database by researchers and engineers in Japan.

3. Details of ASPEC

The statistics of ASPEC is described in Table 1. ASPEC is composed of 4 parts: training data, development data, development-test data and test data on the assumption that it would be used for machine translation research. One thing which is worth mentioning here is that the ASPEC-JE only contains paper abstracts, but the ASPEC-JC contains both abstracts and some parts of the body texts. The details of ASPEC-JE and ASPEC-JC will be described in the following sections.

3.1. ASPEC-JE: Japanese-English Paper Abstract Corpus

The training data for ASPEC-JE was constructed by the NICT from approximately 2 million Japanese-English scientific paper abstracts owned by the JST. Because the abstracts are comparable corpora, the sentence correspondences are found automatically using the method from (Utiyama and Isahara, 2007). Each sentence pair is accompanied by a similarity score which is calculated by the method from (Utiyama and Isahara, 2007). The sentence pairs in ASPEC-JE are sorted by the similarity score. Note that the sentence pairs with the low similarity scores are not necessarily the clean parallel sentences, rather not parallel sentences because they are automatically extracted from the comparable abstracts, not the direct translations of each other.

Field symbol and name	Top1M	ALL # sents	ALL %
A: common science	9,863	28,608	0.95
B: physics	149,406	297,292	9.88
C: basic chemistry	36,233	116,400	3.87
D: space, earth science	19,949	59,083	1.96
E: biology	80,282	266,209	8.85
F: agriculture, forestry, and fishery	56,722	186,265	6.19
G: medicine	284,277	1,016,024	33.77
H: common engineering	8,274	39,279	1.31
I: system, control engineering	11,193	47,856	1.59
J: information engineering	38,659	131,298	4.36
K: industrial engineering	10,112	22,583	0.75
L: energy engineering	4,475	7,006	0.23
M: nuclear engineering	11,187	27,480	0.91
N: electrical engineering	89,232	197,269	6.56
P: thermal engineering	13,273	47,178	1.57
Q: mechanical engineering	33,776	102,157	3.40
R: construction engineering	37,817	128,710	4.28
S: environmental engineering	24,791	51,765	1.72
T: transportation, traffic engineering	6,706	16,967	0.56
U: mining engineering	1,844	4,635	0.15
W: metal engineering	26,939	86,789	2.88
X: chemical engineering	8,997	21,419	0.71
Y: chemical industry	33,749	98,835	3.29
Z: other industry	2,244	7,393	0.25

Table 2: Distribution of the scientific fields in ASPEC-JE.

Each sentence pair is also given a field symbol. The field symbol is a single letter of A-Z and show the scientific field for each document[4] where the sentence pair is extracted. The correspondence between the symbols and field names, along with the frequency and occurrence ratios for the training data, are given in Table 2. Top1M only considers the top 1 million sentence pairs sorted by the similarity score, and ALL considers all the sentence pairs in the training data.

Figure 1 shows the examples of the ASPEC-JE training data. The first example has the highest similarity score in the training data, the second example is the one millionth sentence pair and the third example is the two millionth sentence pair. Each sentence pair has the document ID (DID), sentence ID (SID) in the document and similarity score (sim). The first and the second examples are almost perfect parallel sentences, however the third example is not. For example, Japanese "22 日目頃 (*about the 22th day*)" is translated as "in 3 weeks" in English.

The development, development-test and test data were extracted from parallel sentences from the Japanese-English paper abstracts owned by JST that are not contained in the training data. Each data set contains 400 documents. Furthermore, the data has been selected to contain the same relative field coverage across each data set. The sentence alignment was conducted automatically and only documents with a 1-to-1 alignment are included. It is therefore possible to restore the original documents. The format is the same as for the training data except that there is no similarity score.

[4] http://opac.jst.go.jp/bunrui/index.html

DID: G-03A0568930	SID: 0	Sim: 0.881
Ja: 現在，筋ジストロフィー患者の移動介助において文書マニュアルを使用している。		
En: At present, the document manual is used in transfer assistance of the muscular dystrophy patient.		

DID: G-01A0204677	SID: 1	Sim: 0.137
Ja: リドカイン使用濃度，使用量は０．５～１０％，０．１～１．０ｍｌ（１０～６０ｍｇ）であった。		
En: The use concentration and the amount of lidocaine were 0.5～10% and 0.1～1.0m l(10～60mg) respectively.		

DID: G-93A0370292	SID: 0	Sim: 0.048
Ja: 症例は４３歳の女性で，心臓弁膜症手術後２２日目頃より，発熱と共に全身に紅斑が出現した。		
En: A 43 - year - old female was seen at our clinic with complaints of high fever and erythroderma like skin rashes, which have developed in 3 weeks after her heart operation.		

Figure 1: Examples of ASPEC-JE training data.

3.2. ASPEC-JC: Japanese-Chinese Paper Excerpt Corpus

The ASPEC-JC was constructed by manually translating the Japanese scientific papers into Chinese with permission from the necessary academic associations. The Japanese papers were retrieved from the literature database JDreamII[5] and electronic journal site J-STAGE[6].

The whole abstract or paragraph in the body text was used as the translation unit. The texts to be translated were so iteratively extracted as to cover the vocabularies as much as possible considering the nouns, verbs and the case frames. They were translated sentence-by-sentence, so the sentence pairs are perfectly parallel sentences unlike the ASPEC-JE. Note that the abstract or paragraph can be reconstructed using the sentence IDs, but the whole body text cannot be reconstructed.

The development, development-test and test data consist of 400 translation units which are randomly extracted from the translated data. The translation units which do not have another unit belonging to the same paper in the translated data are extracted. Therefore, there is no sentence pairs sharing the same paper across the training, development, development-test and test sets. This is a practical setting of the machine translation for scientific papers in the future where the input sentences are not in the training data.

As described in Section 2., ASPEC-JC includes only 8 scientific fields. The distribution of the fields is shown in Table 3. Figure 2 shows the example of the ASPEC-JC. The quality of the parallel sentences are almost the same over the whole data unlike ASPEC-JE because they were translated sentence-by-sentence.

[5]It is upgraded to JDreamIII now (http://jdream3.com)
[6]https://www.jstage.jst.go.jp/browse

Field name	Ratio
Medicine	28%
Information	28%
Biology	14%
Environmentology	12%
Chemistry	6%
Materials	5%
Agriculture	4%
Energy	3%

Table 3: Distribution of the scientific fields in ASPEC-JC.

ID: JST_JC_AGR-jcs-74_149.txt-par27-sen1
Ja 精米タンパク質含有率は，５．６９～６．６０％の範囲で，主茎および第４・５節の１次分げつが第７節の１次分げつや第５節の２次分げつに比べ有意に低かった。
Zh 精米蛋白质含量在５．６９～６．６０％的范围内，主茎及第４ Â· 5 节的 1 次分蘖比第 7 节的 1 次分蘖和第５节的２次分蘖明显低。

Figure 2: Example of ASPEC-JC.

4. Application: Workshop on Asian Translation (WAT)

4.1. Overview of WAT

The Workshop on Asian Translation (WAT) is a new open evaluation campaign focusing on Asian languages hosted by JST, NICT and Kyoto University. The first workshop was held in 2014 (Nakazawa et al., 2014) where the ASPEC was centered as the official dataset for the scientific paper translation subtasks. ASPEC was again used in the workshop in 2015 (Nakazawa et al., 2015) to observe the contiguous development of machine translation technologies together with the newly added dataset. WAT will keep growing as the leader of the machine translation technology development in Asia.

WAT is working toward the practical use of machine translation among all Asian countries. WAT tries to understand the essence of machine translation and the problems to be solved by collecting and sharing the knowledge acquired in the workshop. WAT is unique in the following points:

- Open innovation platform
  The test data is fixed and open, so evaluations can be repeated on the same data set to confirm changes in translation accuracy over time. WAT has no deadline for automatic translation quality evaluation (continuous evaluation), so translation results can be submitted at any time.

- Domain and language pairs
  WAT is the world's first workshop that uses scientific papers as the domain, and Chinese↔Japanese and Korean→Japanese as language pairs. More Asian languages will be added in the future.

- Evaluation method
  Evaluation is done both automatically and manually. For human evaluation, WAT uses crowdsourcing,

which is low cost and allows multiple evaluations, as the first-stage evaluation. Also, JPO adequacy evaluation is conducted for the selected submissions according to the crowdsourcing evaluation results.

All the evaluation results and findings can be found in the WAT homepage[7], and the papers of WAT2014[8] and WAT2015[9] are listed in the ACL Anthology.

4.2. Human Evaluation

WAT2015 conducted 2 kinds of human evaluations: *pairwise crowdsourcing evaluation* and *JPO adequacy evaluation*.

4.3. Pairwise Crowdsourcing Evaluation

The pairwise crowdsourcing evaluation was conducted on the randomly selected 400 test sentences from the testset. The input sentence and two translations (the baseline and a submission) are shown to the crowdsourcing workers, and the workers are asked to judge which of the translation is better, or if they are of the same quality. The order of the two translations are at random.

The crowdsourcing workers are not specialists, and thus the quality of the judgments are not necessarily precise. To guarantee the quality of the evaluations, each sentence is evaluated by 5 different workers and the final decision is made depending on the 5 judgements. Each judgement $j_i (i = 1, \cdots, 5)$ is defined as:

$$j_i = \begin{cases} 1 & \text{if better than the baseline} \\ -1 & \text{if worse than the baseline} \\ 0 & \text{if the quality is the same} \end{cases}$$

The final decision D is defined as follows using $S = \sum j_i$:

$$D = \begin{cases} win & (S \geq 2) \\ loss & (S \leq -2) \\ tie & (otherwise) \end{cases}$$

Suppose that W is the number of *wins* compared to the baseline, L is the number of *losses* and T is the number of *ties*. The Crowd score can be calculated by the following formula:

$$Crowd = 100 \times \frac{W - L}{W + L + T}$$

From the definition, the Crowd score ranges between -100 and 100.

4.4. JPO Adequacy Evaluation

The participants' systems, which achieved the top 3 highest scores among the pairwise crowdsourcing evaluation results of each subtask, were also evaluated with the JPO adequacy evaluation. The JPO adequacy evaluation was carried out by translation experts with a quality evaluation criterion for translated patent documents which the Japanese Patent Office (JPO) decided. For each system, two annotators evaluate the test sentences to guarantee the quality. The number of test sentences for the JPO adequacy evaluation was 200. The 200 test sentences were randomly selected from the 400 test sentences of the pairwise evaluation.

5	All important information is transmitted correctly. (100%)
4	Almost all important information is transmitted correctly. (80%–)
3	More than half of important information is transmitted correctly. (50%–)
2	Some of important information is transmitted correctly. (20%–)
1	Almost all important information is NOT transmitted correctly. (–20%)

Table 4: The JPO adequacy criterion

Table 4 shows the JPO adequacy criterion from 5 to 1. The evaluation is performed subjectively. "Important information" represents the technical factors and their relationships. The degree of importance of each element is also considered to evaluate. The percentages in each grade are rough indications for the transmission degree of the source sentence meanings. The detailed criterion can be found on the JPO document (in Japanese) [10].

4.5. Evaluation Results

Figure 3 shows the summary of automatic and human evaluations for the selected submissions (the team names are anonymized). In the results, the best system achieved about 4 points of the JPO adequacy evaluation scores for all the language pairs except ASPEC-JC. This result supports that the ASPEC is the high quality parallel corpus.

5. Conclusion

In this paper, we introduced the ASPEC which is the first large-size parallel corpus of scientific paper domain in the world. The history and characteristics of the ASPEC is described. The purpose of releasing the ASPEC is to contribute to the improvement of not only the machine translation but also all the natural language processing technologies for scientific papers.

Currently, ASPEC only contains 3 languages (Japanese, Chinese and English), however we want to include more Asian languages such as Korean, Vietnamese, Thai, Indonesian, Myanmar, etc. We believe that the ASPEC plays an important role in promoting and evolving the science and technology exchange between Asian countries.

Jimeno Yepes, A., Prieur-Gaston, É., and Névéol, A. (2013). Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 14(1):1–10.

Morin, E. and Prochasson, E., (2011). *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, chapter Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora, pages 27–34. Association for Computational Linguistics.

[7] http://lotus.kuee.kyoto-u.ac.jp/WAT/

[8] http://aclanthology.info/events/ws-2014#W14-70

[9] http://aclanthology.info/events/ws-2015#W15-50

[10] http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

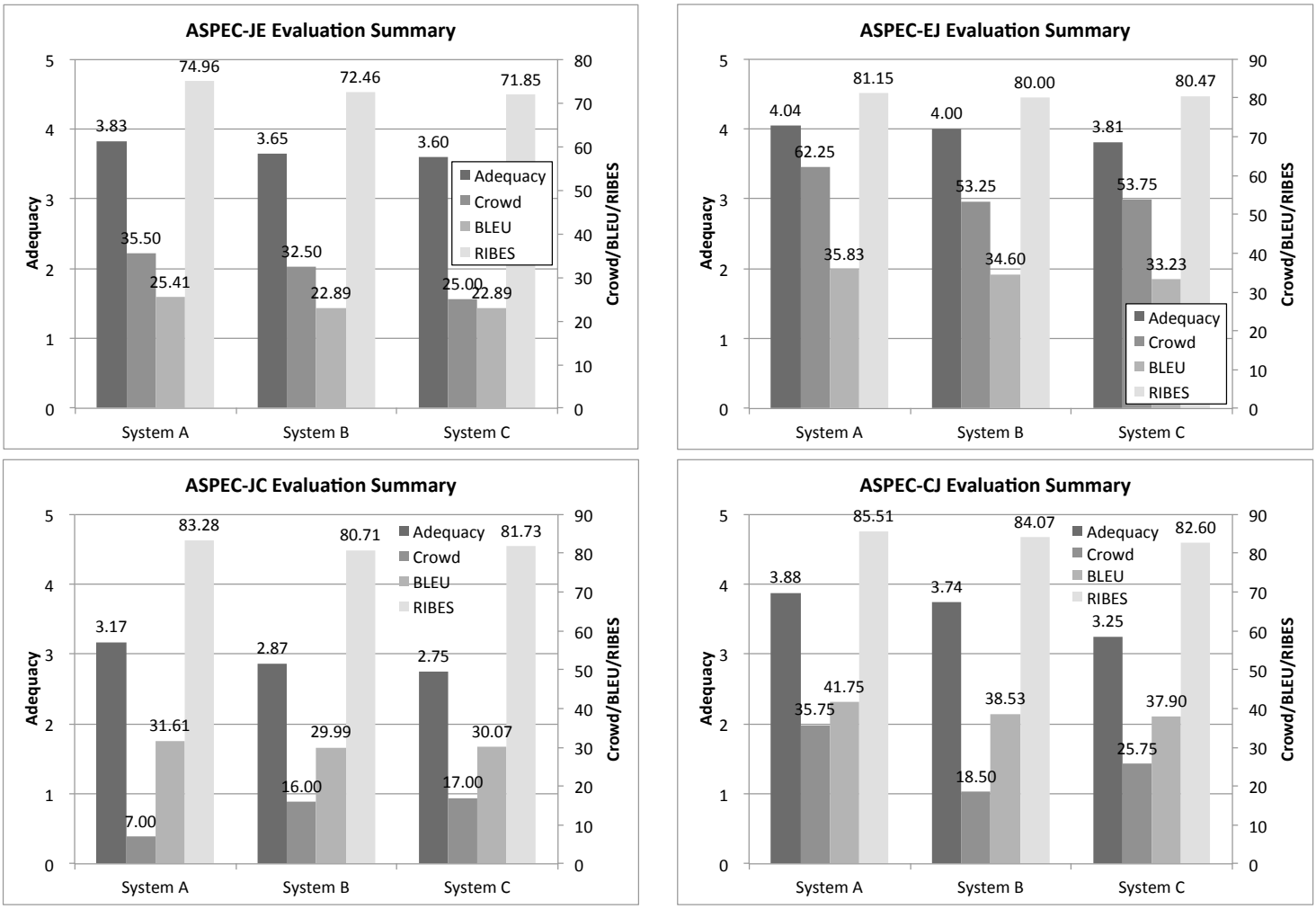

Figure 3: Summary of automatic and human evaluations of ASPEC data in WAT2015.

Nakazawa, T., Mino, H., Goto, I., Kurohashi, S., and Sumita, E. (2014). Overview of the 1st Workshop on Asian Translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*, pages 1–19, Tokyo, Japan, October.

Nakazawa, T., Mino, H., Goto, I., Neubig, G., Kurohashi, S., and Sumita, E. (2015). Overview of the 2nd Workshop on Asian Translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 1–28, Kyoto, Japan, October.

Utiyama, M. and Isahara, H. (2007). A japanese-english patent parallel corpus. In *MT summit XI*, pages 475–482.