

# ASPEC: 亚洲科学论文摘录语料库

1. 科学技术振兴机构 日本 102-8666; 2. 国家信息与通讯研究所 日本 619-0289;  
3. 京都大学 日本 606-8501; 4. 丰桥技术科学大学 日本 441-8580;  
5. 中国科学技术信息研究所 北京 100038; 6. 河北地质大学 石家庄 050031

Toshiaki Nakazawa<sup>1</sup> Manabu Yaguchi<sup>1</sup> Kiyotaka Uchimoto<sup>2</sup> Masao Utiyama<sup>2</sup>  
Eiichiro Sumita<sup>2</sup> Sadao Kurohashi<sup>3</sup> Hitoshi Isahara<sup>4</sup> 何彦青<sup>5</sup> 刘建辉<sup>6</sup>

**摘要** 本文详细介绍了 ASPEC (亚洲科学论文摘录语料库)。作为首个大规模的科学论文领域内的平行语料库, ASPEC 是由日-中机器翻译项目于 2006 年至 2010 年间利用科技促进专用协作基金构建起来的。它包含约 300 万条平行语句的日-英科学论文摘要语料库 (ASPEC-JE) 和约 68 万条平行语句的中-日科学论文摘录语料库 (ASPEC-JC)。ASPEC 被用作机器翻译评测研讨会 WAT (亚洲翻译研讨会) 的官方数据集。

**关键词:** 平行语料库, 科学论文, 亚洲语言

**中图分类号:** G355

## ASPEC: Asian Scientific Papers Excerpt Corpus

1. Japan Science and Technology Agency, Tokyo 102-8666, Japan;  
2. National Institute of Information and Communications Technology, Kyoto 619-0289, Japan;  
3. Kyoto University, Kyoto 606-8501, Japan;  
4. Toyohashi University of Technology, Aichi 441-8580, Japan;  
5. Institute of Scientific and Technical Information of China, Beijing 100038, China;  
6. Hebei GEO University, Shijiazhuang 050031, China

Toshiaki Nakazawa<sup>1</sup> Manabu Yaguchi<sup>1</sup> Kiyotaka Uchimoto<sup>2</sup> Masao Utiyama<sup>2</sup>  
Eiichiro Sumita<sup>2</sup> Sadao Kurohashi<sup>3</sup> Hitoshi Isahara<sup>4</sup> HE YanQing<sup>5</sup> LIU JianHui<sup>6</sup>

**Abstract** In this paper, we described 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

**作者简介:** Toshiaki Nakazawa (1982-), 男, 博士, 国立研究开发法人, 日本科学技术振兴机构(JST)研究员, 京都大学黑桥&河原研究室成员, 研究方向: 机器翻译。Email: nakazawa@pa.jst.jp。

**Keywords:** Parallel corpus, scientific paper, Asian languages

## 041

与之相应，日 - 汉平行语料库也逐步建成，最终开发了日 - 汉平行词典和机器翻译系统。

涉及多个科学领域的日 - 英平行摘要完成后，借助从摘要对中抽取平行语句构建了日 - 英平行语料库。因此，ASPEC-JE 包含了所有科学领域的平行语句。需要注意的是，ASPEC-JE 摘要对的源语言为日语，并由人工翻译成英语。这是由于版权问题。

另一方面，由于没有日 - 中的预备资源，ASPEC-JC 最初是通过由人工将日语文档翻译成中文而构建的。ASPEC-JC 只包括“医学”、“信息”、“生物学”、“环境学”、“化学”、“材料”、“农业”和“能源”等领域而无法涵盖所有的科学领域。这些领域是通过研究中国的重要科学领域和日本研究人员与工程师对于文献数据库的使用趋势选出的。

### 3 ASPEC具体内容

ASPEC 的统计描述见表 1。ASPEC 由 4 部分组成：训练数据，开发数据，开发测试数据和测试数据，这些数据将用于机器翻译研究。值得一提的是，ASPEC-JE 只包含论文摘要，而 ASPEC-JC 包含摘要和部分正文文本。ASPEC-JE 和 ASPEC-JC 的细节将在以下诸小节中加以描述。

表1 ASPEC统计量

| 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 |

### 3.1 ASPEC-JE: 日-英论文摘要语料库

ASPEC-JE 的训练数据是由 NICT 从 JST 所属的大约 200 万日 - 英科学论文摘要中构建起来的。因为摘要是可比較的语料库，句子对齐是使用 (Utiyama 和 Isahara, 2007)<sup>[3]</sup> 的方法自动实现的，每个句对附着了一个利用 (Utiyama 和 Isahara, 2007)<sup>[3]</sup> 的方法计算而得的相似度得分。ASPEC-JE 中的句子对皆按照相似性得分排序。需要注意的是，由于这些语句是从可比較的摘要中自动提取的，而不是彼此直译的结果，那些具有低相似度得分的句子对不一定是整齐的平行语句，甚至不一定是并行语句。

每个句子对还被赋予一个领域标记。领域标记是从 A 到 Z 的单个字母以显示抽出该句子对的文档<sup>④</sup>所属的科学领域。标记与领域名之间的对应关系以及训练数据的频率和出现率在表 2 中给出。Top 1M 仅考量按相似性分数排序的前 100 万个句子对，ALL 考量训练数据中的所有句子对。图 1 显示了 ASPEC-JE 训练数据的实例。例一在训练数据中具有最高的相似度得分，例二是第 100 万个句子对，例三是第 200 万个句子对。每个句子对都有在文档中的文档 ID 号 (DID)、句子 ID 号 (SID) 和相似度得分 (sim)。例一和例二几乎是完美的平行句子，但例三不是。例如，日语中的“22 日目頃 (约 22 日)”在英语中翻译为“在 3 周内”。

开发数据、开发测试数据和测试数据抽取自 JST 所属的日 - 英论文摘要的平行语句，都不包含在训练数据中。每个数据集包含 400 个

<sup>④</sup> <http://opac.jst.go.jp/bunrui/index.html>

文档。此外，每个数据集覆盖了在同一相近领域范围内的数据。自动进行了句子对齐且仅包括了具有 1-1 对齐的文档。这使得原始文档可以复原。除却无相似度得分之外，格式与训练数据相同。

表2 ASPEC-JE的科学领域分布

| 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  |

|                                                                                                       |        |            |
|-------------------------------------------------------------------------------------------------------|--------|------------|
| 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0.5～1.0%，0.1～1.0ml（1.0～6.0mg）であった。                                                       |        |            |
| En: The use concentration and the amount of lidocaine were 0.5～10% and 0.1～1.0ml (1.0～6.0mg) 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. |        |            |

图1 ASPEC-JE训练数据实例

## 3.2 ASPEC-JC:日-中论文摘要语料库

ASPEC-JC 是经所需的学术协会批准，将日文的科学论文人工翻译成中文而构建起来的。日文论文取自文献数据库 JDreamII<sup>⑤</sup>和电子杂志网站 J-STAGE<sup>⑥</sup>中。

整个摘要和正文中的完整段落用作翻译单元。待翻译的文本需迭代地提取以便于尽可能地覆盖词汇表，无论其为名词、动词亦或格框(case frames)。因为经过了逐句翻译，所以句子对完全平行而不像 ASPEC-JE 那样。需要注意的是，摘要或段落可以使用句子 ID 号来重构，但是整个正文文本无法被重构。

开发数据、开发测试数据和测试数据包括 400 个翻译单元，它们是从翻译数据中随机提取的。被抽取出来的翻译单元都不与翻译数据中的单元在同一篇论文中。因此，没有任何句子对在训练数据、开发数据、开发测试数据和测试数据共享同一篇论文。这是着眼于未来的科学论文机器翻译的一个实用设置，适用于输入语句不在训练数据之中的情况。

如第 2 节所述，ASPEC-JC 仅包括 8 个科学领域。领域的分布如表 3 所示。图 2 显示了 ASPEC-JC 的实例。平行语句的质量在整个数据上几乎始终如一，不像逐句翻译的 ASPEC-JE 那样。

表3 ASPEC-JC的科学领域分布

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

<sup>⑤</sup>现在升级为 JDreamIII (<http://jdream3.com>)

<sup>⑥</sup><https://www.jstage.jst.go.jp/browse>

|                                                                               |
|-------------------------------------------------------------------------------|
| 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・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。                 |

图2 ASPEC-JC实例

## 4 应用：亚洲翻译研讨会(WAT)

### 4.1 WAT回顾

亚洲翻译研讨会(WAT)是由JST、NICT和京都大学主办的一个新的开放式评测活动，重点关注亚洲语言。首届研讨会在2014年举行<sup>[4]</sup>，其以ASPEC作为科学论文翻译子任务的官方数据集。ASPEC在2015年的研讨会上再次被使用<sup>[5]</sup>，与新添加的数据集一起用以观测机器翻译技术的连续性发展。WAT将作为亚洲机器翻译技术开发的领导者继续成长。WAT在所有亚洲国家致力于机器翻译的实际应用。WAT尝试理解机器翻译的实质并通过收集和分享在研讨会中获得的知识来解决问题。WAT在以下几点上是独一无二的：

- 开放式创新平台

测试数据固定并开放，因此评测可以在相同的数据集上重复进行，以确认翻译准确性跨时段的变化。WAT没有翻译质量的自动评测(连续评测)的截止日期，因此可以随时提交翻译结果。

- 领域与语言对

WAT是世界上第一个限定在科学论文领域内的研讨会，并以中↔日和韩→日作为语言对，将来会有更多的亚洲语言添加进来。

- 评估方法

研讨会采用了自动评价和人工评价。对于人工评价，WAT使用了众包作为第一阶段的评估，这样既降低了成本又兼容多个评估。而且，根据众包评估结果对所选的提交结果进行了JPO充分性评价。

所有的评测结果和翻译结果可以在WAT主页<sup>⑦</sup>中找到，WAT2014<sup>⑧</sup>和WAT2015<sup>⑨</sup>的文件也列于ACL文集中。

### 4.2 人工评价

WAT2015实施2种人工评估：成对众包评价和JPO充分性评价。

#### 4.2.1 成对众包评价

从测试集中随机选择400个测试语句进行成对众包评估。输入句和两个预翻译(基线和提交)提供给众包工作人员，工作人员判断哪个翻译更好，或者是否具备相同的质量。两个预翻译的供给顺序是随机的。

众包工作人员不是专家，因此判断的质量未必特别精确。为了保证评估的质量，每个句子由5个不同的工作人员评估，最终的决定取决于5个判断。每个判断定义 $J_i(i=1,\dots,5)$ 为：

$$J_i = \begin{cases} 1 & \text{如果好于基线} \\ -1 & \text{如果差于基线} \\ 0 & \text{如果质量相同} \end{cases}$$

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>



最终的决策  $D$  使用公式  $S = \sum j_i$  来定义:

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

假设  $W$  是与基线相比的 win 的数量,  $L$  是 loss 的数量,  $T$  是 tie 的数量。Crowd 得分可以通过以下公式计算:

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

根据定义, Crowd 的得分范围在 -100 和 100 之间。

#### 4.2.2 JPO 充分度评价

在每个子任务的成对众包评价结果中位列最高分前 3 个的参与系统, 再使用 JPO 充分性评价。JPO 充分度评价由翻译专家进行, 他们执行的是由日本专利局 (JPO) 核定的专利文件翻译的质量评估标准。对于每个系统, 都由两个标注者评估测试语句以保证质量。JPO 充分度评价的测试句数为 200。这 200 个测试句是从成对评价的 400 个测试句中随机选择的。表

4 显示了从 5 到 1 的 JPO 充分度标准。该评估为主观评估。“重要信息”表示技术因素及其关系。每个要素的重要程度在评估中也被考虑。每个等级中的百分比是源语言句子的语义传输度的粗略指示。详细标准可以在 JPO 文档 (日语)<sup>⑩</sup> 中找到。

表4 JPO充分度标准

|   |                                                                          |
|---|--------------------------------------------------------------------------|
| 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%)    |

#### 4.3 评价结果

图 3 显示了选择提交 (团队名称是隐匿的) 的自动评价和人工评价的概要。在结果中, 除了 ASPEC-JC 之外的所有语言对, 最佳系统达到约 4 分的 JPO 充分度评价分数。这个结果说明 ASPEC 是高质量的平行语料库。

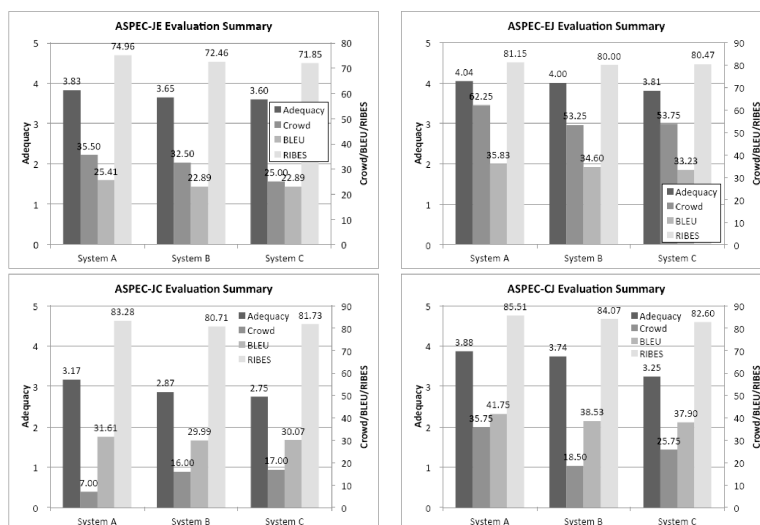


图3 WAT2015中ASPEC自动评估和人工评估总结

<sup>⑩</sup> [http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku\\_hyouka.htm](http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm)

## 5 结论

在本文中，我们介绍了 ASPEC 这一世界上首个大规模的科学论文领域内的平行语料库，涉及 ASPEC 的历史和特性。发布 ASPEC 的目的不仅仅助力改进机器翻译，而且益于改进面向科学论文的所有自然语言处理技术。目前，ASPEC 仅包含 3 种语言（日语、汉语和英语），但我们希望涵括更多的亚洲语言，如韩语、越南语、泰语、印度尼西亚语、缅甸语等。我们相信 ASPEC 在促进和发展亚洲国家之间的科技交流中会扮演一个重要的角色。

---

### 参考文献

---

[1] Yepes A J, Priourgaston E, Névél A. Combining MEDLINE and Publisher Data to Create Parallel

Corpora for the Automatic Translation of Biomedical text[J]. BMC Bioinformatics, 2013, 14(1):146.

[2] Morin E, Prochasson E. Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora[C]// The Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. 2011: 27-34.

[3] Utiyama M, Isahara H. A Japanese-English Patent Parallel Corpus[J]. Proc of Mt Summit XI, 2007.

[4] Nakazawa T, Mino H, Goto I, Kurohashi, and Sumita E. Overview of the 1<sup>st</sup> Workshop on Asian Translation[C]// In Proceedings of the 1<sup>st</sup> Workshop on Asian Translation (WAT2014), 1–19 October, Tokyo, Japan, 2014: 1-19.

[5] Nakazawa T, Mino H, Goto I, et al. Overview of the 2<sup>nd</sup> Workshop on Asian Translation[C]// In Proceedings of the 2<sup>nd</sup> Workshop on Asian Translation (WAT2015), October, Kyoto, Japan. 2015: 1-28.