

## 第2回特許情報シンポジウム

2012年11月30日

AAMT/Japio 特許翻訳研究会

## 第2回特許情報シンポジウム

AAMT/Japio 特許翻訳研究会

委員長 辻井潤一

日時：平成24年11月30日(金)

場所：京都大学東京オフィス

東京都港区港南2-15-1 品川インターシティA棟27階

<http://www.kyoto-u.ac.jp/ja/tokyo-office/>

主催者

- アジア太平洋機械翻訳協会 (AAMT)

<http://www.aamt.info/index.htm>

- (財) 日本特許情報機構 (Japio)

<http://www.japio.or.jp/index.html>

Program

13:00-13:10 Opening Remarks

- Toshimichi Moriya (Japio)

守屋敏道 ( (財) 日本特許情報機構専務理事 特許情報研究所所長)

- Jun'ichi Tsujii (Microsoft Research Asia)

辻井潤一 (AAMT/Japio 特許翻訳研究会委員長、マイクロソフトリサーチアジア研究所 首席研究員、  
東京大学 名誉教授)

13:10-13:50 Invited Talk 1

Dan WANG (China Patent Information Center)

Making Effective Use of Machine Translation for Patent Documents: Practice of CPIC

13:50-14:30 Invited Talk 2

Paul Schwander (Director External Products and Services, European Patent Office)

Machine Translation at the EPO

Removing language barriers from patent documentation

14:30-15:10 Invited Talk 3

Takashi INABA (Assistant Director, Patent Information Policy Planning Office, Japan  
Patent Office)

JPO's Approach for Machine Translation - To establish productive utilization method and create appropriate policies

15:10-15:40 Break

15:40-16:20 Meeting Report

Terumasa Ehara, Hiroshi Echizen'ya (AAMT/Japio Special Interest Group on Patent Translation):

Report of Review Meeting on Evaluation Methods for Machine Translation Results in Patent Document

\* Opening Remarks, Invited Talks and Meeting Report are in English.

16:20-17:40 一般セッション（4件の一般講演、講演15分、質疑5分）

16:20-16:40 依存関係を用いた特許分野のための日英中対訳フレーズの切り出しアルゴリズム

池田秀人、Nguyen Thanh Hung, Ze Zhong Li, Chong Zheng Zhong（立命館大）

16:40-17:00 特許明細書の翻訳者から基本的な誤訳の実例を示して対策を提案

吉川潔（フリー翻訳者）

17:00-17:20 特許翻訳の品質を向上するための形態素解析結果を利用した文書比較・日本語精査ツール  
－歌詠と鶯－の試作

楠本浩二（（株）クレステック）、山口日緒里、鈴木貴年、千引春菜（アイビー・システム（株））

17:20-17:40 技術調査のための特許情報抽出

原田綾花、太田貴久、小林暁雄、増山繁（豊橋技科大）、野中尋史（大分工業高専）、酒井浩之（成蹊大学）

17:40-17:50 まとめ

# **Invited Talk 1**

**Dan WANG**

**China Patent Information Center**

# Making Effective Use of Machine Translation for Patent Documents: Practice of CPIC

Dan WANG

Tokyo, November 30, 2012

## Overview

- Introduction
- Some reflections on MT
- CPIC's practice of MT for patents
- Possible cooperation in MT R&D

## Brief history of MT

- Until the 1980s → rule-based translations
- Currently, research dominated by corpus-based approaches:
  - 1) Statistical machine translation
  - 2) EBMT

## Rule-based MT vs. Statistical MT

- RBMT
  - More predictable and grammatically superior
  - Can be customized in different domains
  - Requires manual development of linguistic rules
  - Quality ceiling due to complexity of language
- SMT
  - Provides fluent translation results
  - Rapid development
  - Translation quality not predictable and consistent
  - Relies heavily on existing huge corpora

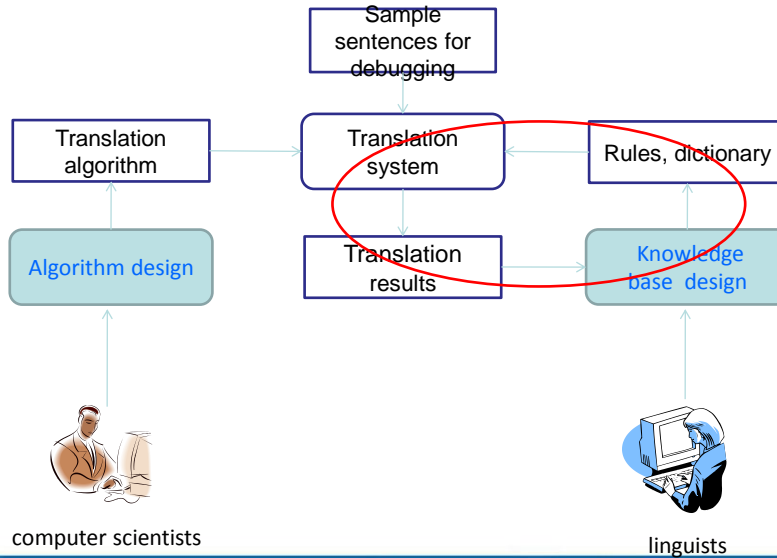
## Research vs. Real-world MT systems

- Research MT systems
  - Not commercially available
  - Goal: achieving highest evaluation score
  - Extreme measures may possibly be taken
- Real-world MT systems
  - Available commercially
  - Mostly rule-based
  - Speed and efficiency optimized for on-the-fly translation
  - Fit for unrestricted environment

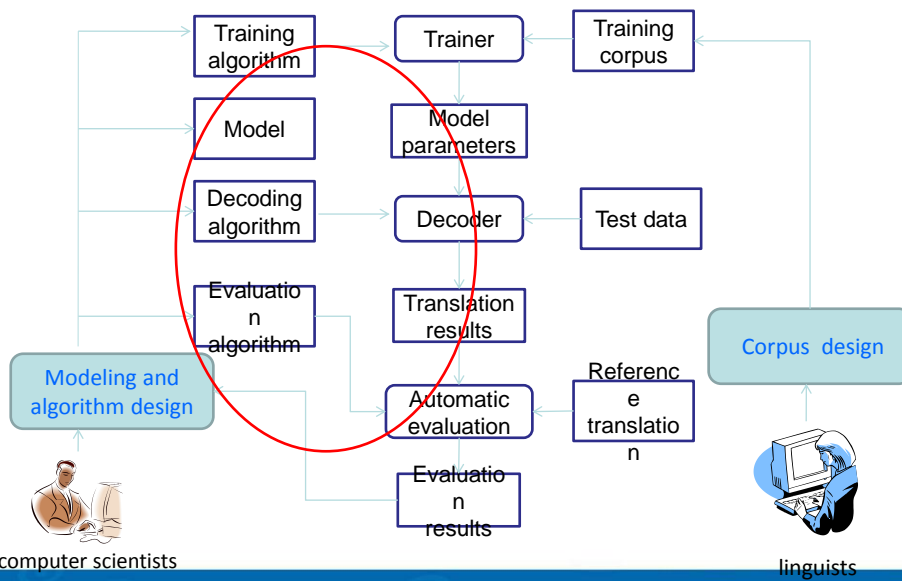
## What's the next step of MT research?

- RBMT: integrating statistically generated lexical entries
- Phrase-based statistical MT -> syntax-based statistical MT
- Hybrid approaches: effect still needs “wait and see”

## The research paradigm of RBMT

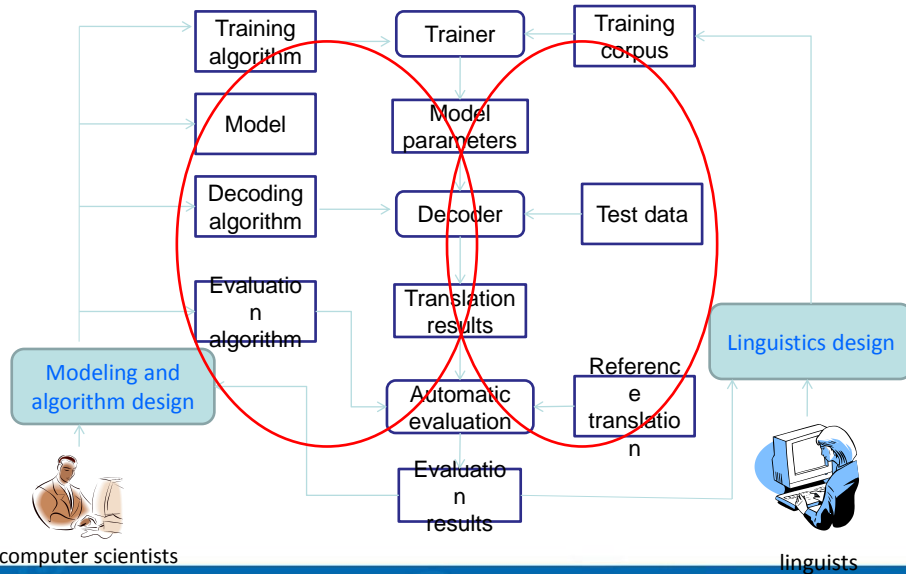


## The research paradigm of SMT





## New research paradigm combining SMT & RBMT



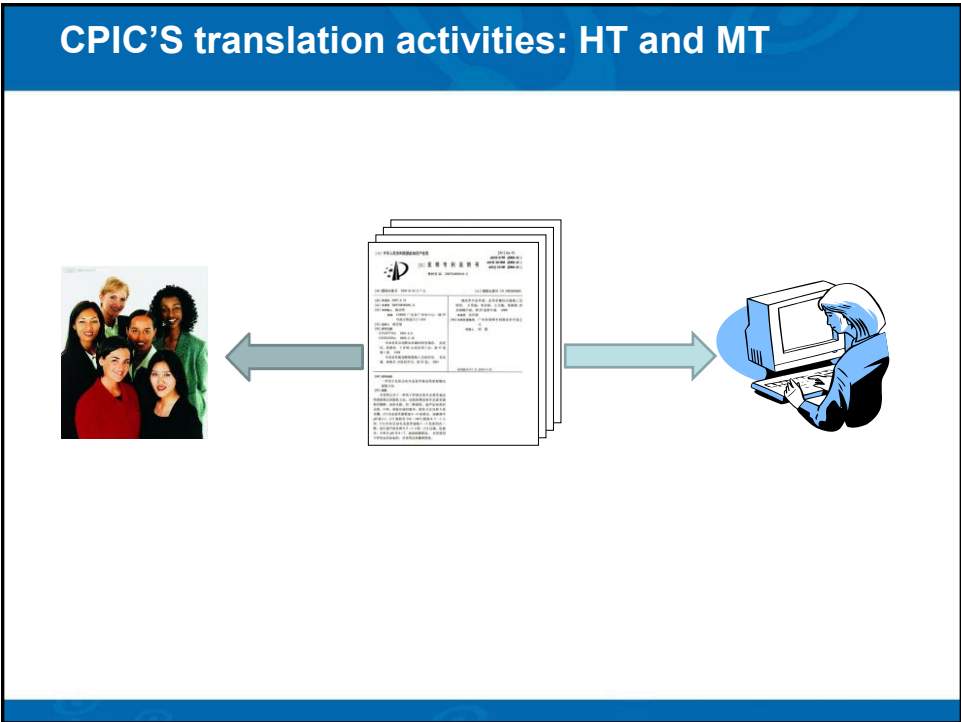
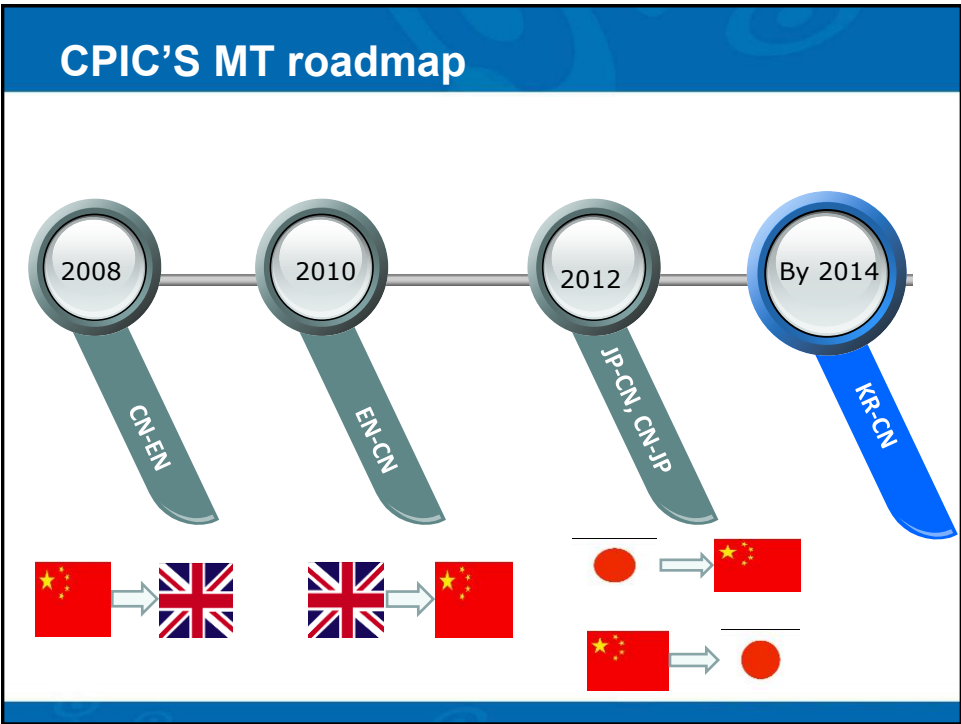
## Better MT evaluation framework needed

- Blue: not sensitive to comprehensibility or accuracy
- Automatic evaluation -> Human evaluation
- Adequacy only -> adequacy + Acceptability

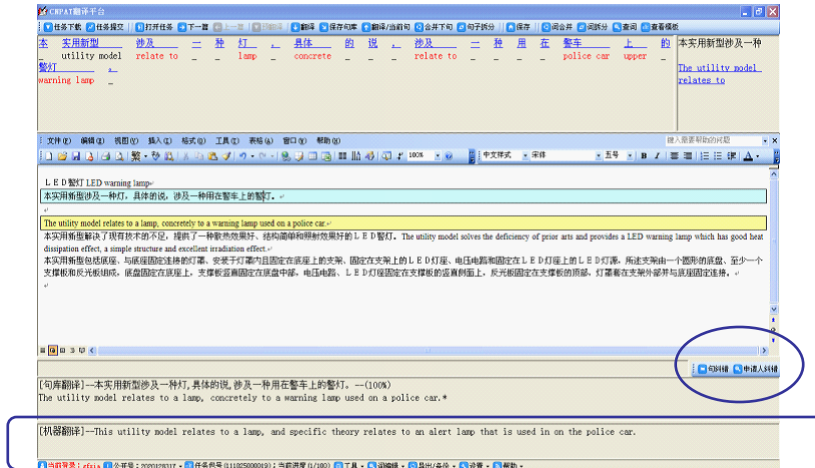
## What works for patent MT

- Patent documents contain mainly technical languages, grammar and parser tool adaption effective for quality improvement
- Patent-specific style and format can be handled by example-based approaches
- International Patent Classification (IPC) system makes it possible for adapting lexicons separately

- Introduction
- Some reflections on MT
- CPIC's practice of MT for patents
- Possible cooperation in MT R&D



# CN-EN MT --->CAT



# EN-CN MT @ pss-system

首页

专利检索

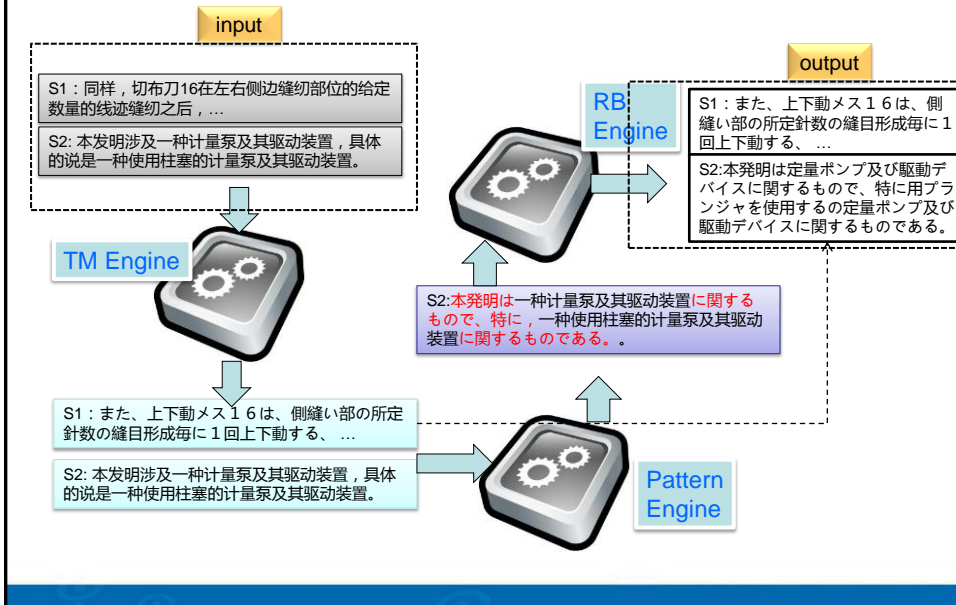
申请号	US201113036793
申请日	2011.02.28
公开(公告)号	US2011151904A1
公开(公告)日	2011.06.23
IPC分类号	H04W86/02; H04W94/
ECLA分类号	H04M1/32C; H04M1/
申请人(专利权人)	MOTOROLA INC;
	MOCK VON A;BAUDIF
优先权	US201113036793; U
优先权日	2011.02.28; 2006.

摘要 [支持多种翻译]

简体中文->英文    英文->简体中文

Abstract\_EN: A server includes a communications adapter; a controller and a data store. The controller receives a first user input via a user interface to associate at least one image attribute with a contact. Thereby, responsive to receiving an image from an electronic apparatus, the controller automatically processes the image to identify at least one feature.

## CN-JP MT: multi-engine strategy

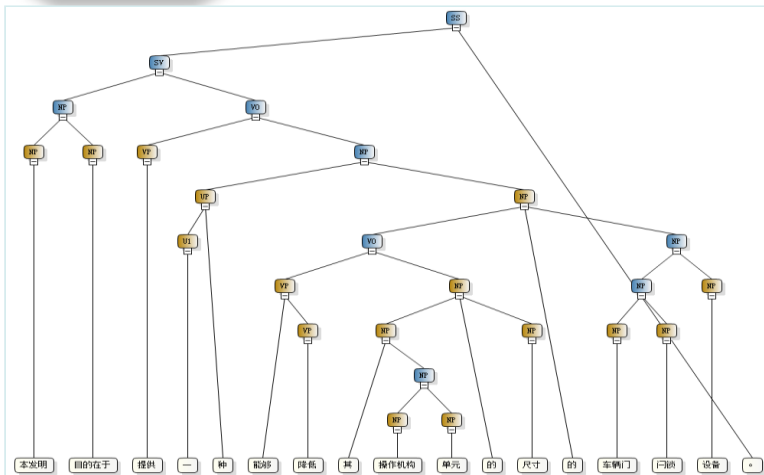


## CN-JP MT: syntax tree



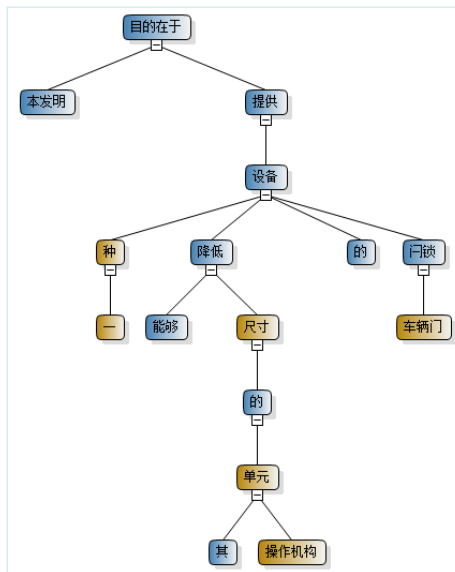
Source

本发明的目的在于提供一种能够降低其操作机构单元的尺寸的车辆门闩锁设备。



## CN-JP MT: dependency tree

transfer



本发明是操作机构部的サイズの減少を可能にした自動車用ドアラッチの装置を提供することを目的としている。

Pattern example

原文

本发明涉及 一种计量泵及其驱动装置 ，具体的说是一种使用柱塞的计量泵及其驱动装置。

载入模板

本发明は 一种计量泵及其驱动装置 に関するもので 、特に 一种使用柱塞的计量泵及其驱动装置 に関するものである。

译文

本发明は定量ポンプ及び駆動デバイスに関するもので、特に用プランジャを使用するの定量ポンプ及び駆動デバイスに関するものである。

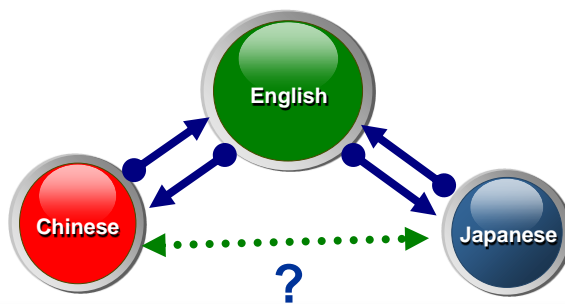
## JP-CN MT result

原文 原稿	Google	業界他社	日汉机译系统结果 本システム
押圧ロッド42の表面には、軟質材によって形成されたキャップ56が被冠されている。	下冠已上迫切杆42，第56章，这是由软质材料形成的表面。	押圧漁42写着軟質材料而形成的盖达56被冠さ。	在加压连杆42的表面中，根据软性材料形成的顶盖56被加上。

- Introduction
- Some reflections on MT
- CPIC's practice of MT for patents
- Possible cooperation in MT R&D

## MT between Japanese & Chinese: English as a pivot

- There's room for quality improvement for MT into English;
- By-products already very useful: dictionary and corpora



14/34

## MT between Japanese & Chinese: direct approach?

- Any possibility of MT evaluation cooperation?
- Any basis of collaboration on RBMT modules of analysis, generation and transfer?



14/34



Thank you!

## **Invited Talk 2**

**Paul Schwander**

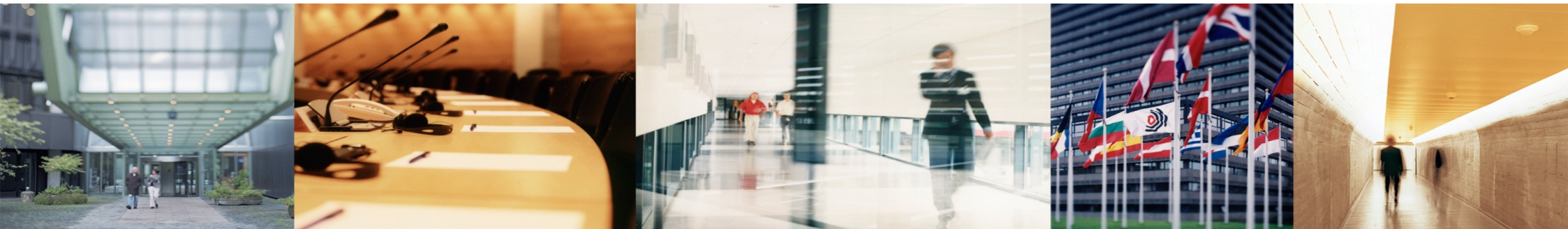
**European Patent Office**

# Machine Translation at the EPO

## Removing language barriers from patent documentation

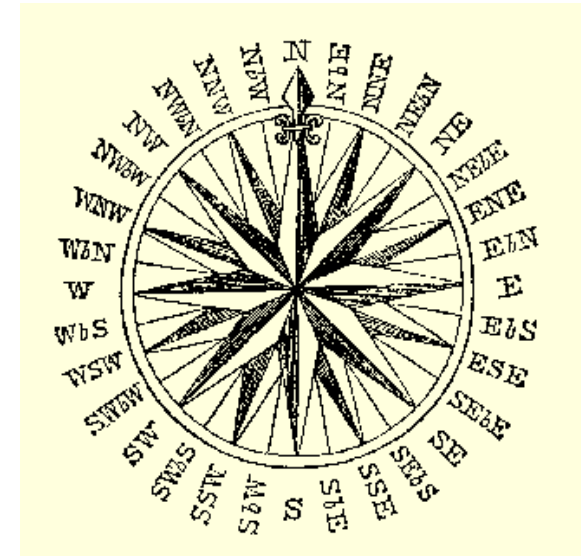
Paul Schwander  
Director External Products and Services  
European Patent Office

**Second Symposium on Patent Information Processing, Tokyo 30 November 2012**



# Roadmap

- The context: why is MT strategic?
- Machine Translation @ the EPO:  
Status and future plans



# Why Machine Translation?

- Reducing the language barrier in the European context: Unitary Patent to be supported by an MT solution.
- Access to global patent information for prior art searches.

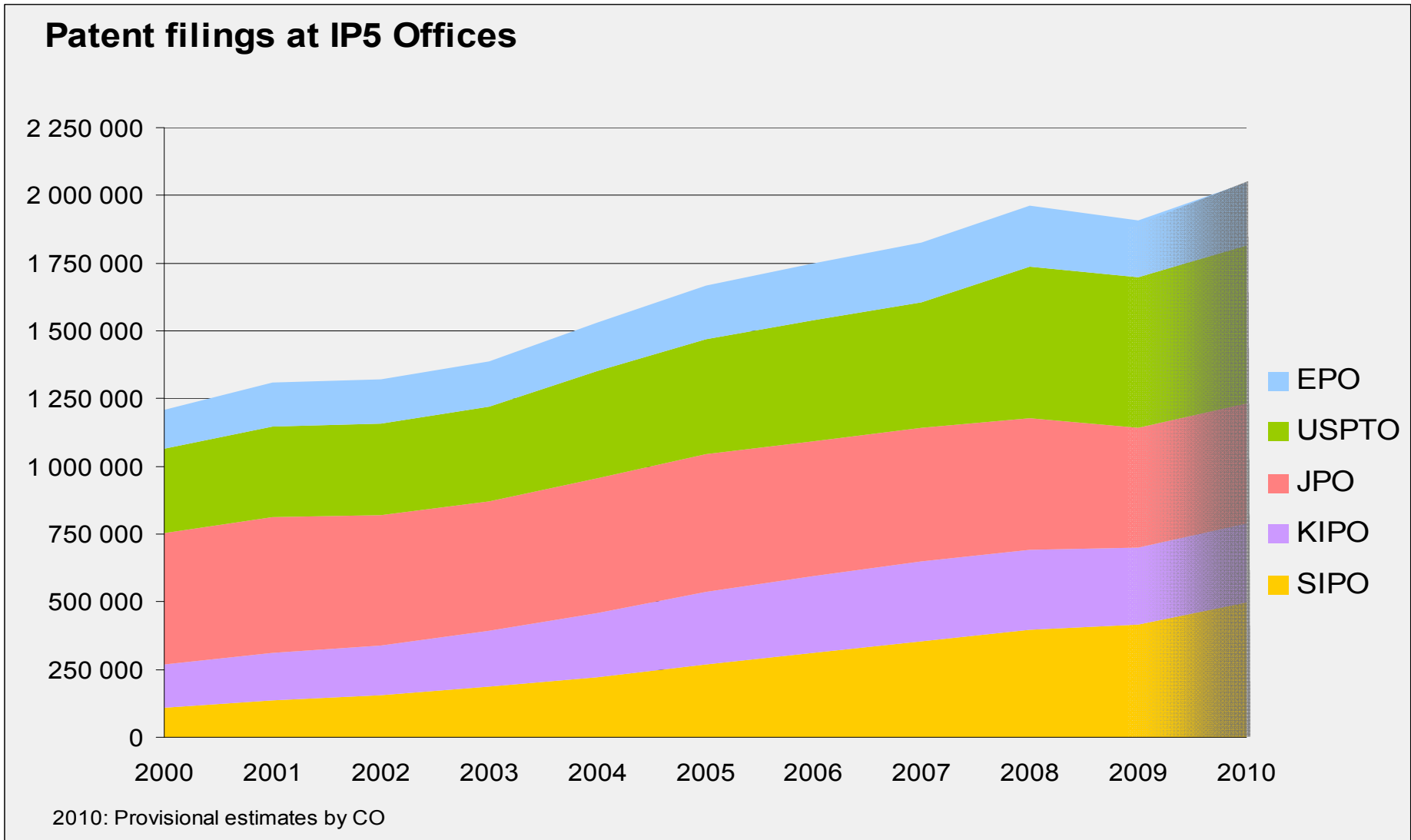


# The Unitary Patent

## ▪ Basic principles

- a European patent **granted under the EPC**
- **unitary effect** for the territories of the 25 EU member states currently participating, at the applicant's request
- **co-existence with the existing European patent and national patents**
- **validated in one single administrative step by the EPO** for all the participating states in the language in which it was granted
- **simplified language regime MT supported;** transition measures foreseen





**Global patent filings rising continuously, especially Chinese applications**  
**IP5= Europe, USA, Japan, China and Korea**

# The case of patents in Asian languages patent and EPO examiners



WO0217204 : CN1201256C CN1388935 EP2375700 US7020602 EP1312023

**© TXTCNT**

**CN1201256C C 20050511**

Automatically Translated by the Intellectual Property Publishing House of SIPO

Registered system and method of the domain name of native language

**Technical field**

This invention generally involves the communication network. Particularly, it is not a method and apparatus of naming system of English language that this invention involves a adaptation, addressing and visit the communication network, especially the entity of the WWW directly.

**Technological background**

Internet, especially is offering a lot of information resources in the WWW , every resource offers some useful information to vast computer network user, goods and service, its typical form is that hyperlink labels the language (HTML) page (webpage). As everyone knows, user use install in they net browser in the computer, such as NAVIGATOR of American Netscape, or INTERNET EXPLORER of American Microsoft, in order to visit their required webpage.

In order to visit a particular webpage, users' net browser must know in the network, namely the address of required targeted website in Internet. Especially in the WWW, its address means the IP address . What the IP address is used is digital format, such as 123.456.78.9. And the website server that every one is regarded as the WWW has a specialized IP address.

However, as webpage quantity in the WWW runs up, design a more direct site selection method, it has used the alias that convenient memorial letter and figure form. According to the site selection system designed newly, usually, the address means a common resource indicating device (URL), its typical grammar is: Protocol (agreement): //Www . Domain-name (domain name). Domain (land). To WWW, its agreement refers to the hypertext transfer protocol (HTTP). The land has pointed out which Network Entity NE more advanced classification this address belongs to, for example, the land can be com, org, net, or similar name (besides U.S.A., it was this country that showed in other countries in the land, it is the commercial land of Britain as com.uk represented). The domain name makes up the alias of the real IP address with land word separator.

And the domain name can be made up with many stature names, for instance: Sub name 1. Sub name 2 ....Son famous n.. For example, one URL, <http://www.xyz.com> Represent the server in the WWW as the host computer of the websites of xyz Company.

Keep the global address book with special server which is called the name server (DNS) put in Internet, it can map if the domain name of " xyz.com " to regarding as the physical address of the server of the network station host of xyz Company correspondingly, such as 123.456.78.9. The name server has more than one computer separating physically, and carry

WO0217204 : CN1201256C CN1388935 EP2375700 US7020602 EP1312023 DE10193513T JP2004516531

Bib 1/1

[19] 中华人民共和国国家知识产权局 [51] Int. Cl.<sup>7</sup>  
C06F 17/00

[12] 发明专利说明书

[21] ZL 专利号 01802492.0

[45] 授权公告日 2005年5月11日 [11] 授权公告号 CN 1201256C

[22] 申请日 2001.8.17 [21] 申请号 01802492.0 [74] 专利代理机构 中科专利商标代理有限责任公司  
[30] 优先权 [32] 2000. 8.21 [33] US [31] 09/642,471 代理人 戎志敏  
[86] 国际申请 PCT/US2001/041785 2001. 8. 17  
[87] 国际公布 WO2002/017204 英 2002. 2. 28  
[85] 进入国家阶段日期 2002. 4. 22  
[71] 专利权人 金基锡  
地址 美国马里兰州  
共同专利权人 柳志烈  
[72] 发明人 金基锡 柳志烈  
审查员 蔡 萍

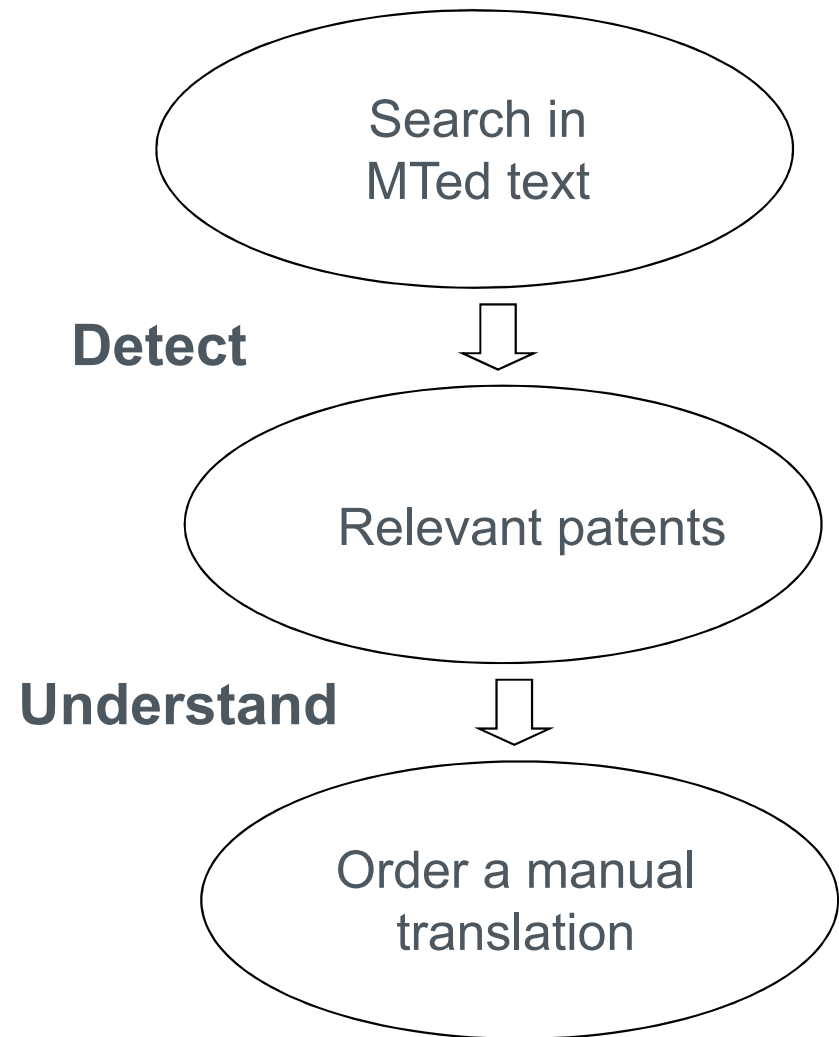
权利要求书2页 说明书11页 附图5页

[54] 发明名称 本族语域名的注册系统和方法  
[57] 摘要  
一种域名系统, 包括: 本族语域名注册器 (105), 它用来接收本族语域名, 这个本族语域名包括至少一个非字母数字字符; 申请注册并将本族语域名转换成一个字母数字域名。字母数字域名被用在域名服务器 (DMS) (103) 的域名/IP 地址入口中。

ISSN 1 0 0 8 - 4 2 7 4

知识产权出版社出版

- MT Full-text acquired, ca. 4 million documents.
- An on demand manual translation service offered to examiner
- 4 million patents: manual translation -> 1 day a patent -> 18 years of work for a team of 1000 translators.



# Patent Translate

## Patent Translate in Espacenet

1. Find a patent document in Espacenet.
2. Choose the part you want to translate - abstract, description or claims.
3. Select the target language, and click the Patent Translate button.

### Description of WO2011131922 (A2)

Translate this text into 

German



patenttranslate

powered by EPO and Google

4. Once you see the translation, you can mouse-over to see the original text sentence by sentence.

SYSTEM UND VERFAHREN Für eine Fläche von DMS

BESCHREIBUNG

HINTERGRUND DER ERFINDUNG

Die Erfindung betrifft ein System und Verfahren für ein DMS-Oberfläche, die eine intelligente intelligente Oberfläche DMS und welche einen direkten Anwendung der Turbine Strukturen und

Windtürb

Die Erfin und Flug umungsprofil Tragflächen

Insbeson intelligent smart surface strain gauge and which has

Eigensch direct application to wind turbine structures and her intelligenten Oberfläch


DMS biet wind turbine blades and wind turbine struts.

Gemäß der Erfindung kann das Material in die Oberfläche einer Struktur in einem Array integriert werden, umfassend getrennte Blätter aus kohlefaserverstärktem Epoxy, die zusammen verbunden sind und denen jeweils eine separat adressierbare Element wie beispielsweise ein Mikrocontroller.

5. Submit your feedback.

# Patent Translate launched on 29 February 2012

System integrated in Espacenet, the EPO Publication Server and EPOQUE



Europäisches Patentamt  
 European Patent Office  
 Office européen des brevets

**Espacenet**  
 Patent search

Deutsch English  
 Change

---

← About Espacenet Other EPO online services

Search
Result list
★ My patents list (0)
Query history
Settings
Help

EP2144485 (A1)  
**Bibliographic data**  
 Description  
 Claims  
 Mosaics  
 Original document  
 Cited documents  
 Citing documents  
 INPADOC legal status  
 INPADOC patent family  


---

**Quick help**  
 → What does A1, A2, A3 and B stand for after a publication number?  
 → What happens if I click on "In my patents list"?  
 → What happens if I click on the "Register" button?  
 → Why are some sidebar options deactivated for certain documents?  
 → How can I bookmark this page?  
 → Why does a list of documents with the heading "Also published as" sometimes appear, and what are these documents?  
 → What is a cited document?  
 → What are citing documents?  
 → What information will I find if I click on the link "View all"?  
 → Why do I sometimes find the

**Bibliographic data: EP2144485 (A1) — 2010-01-13**  
 ★ In my patents list → EP Register → Report data error

**Device for mounting components**

**Page bookmark** EP2144485 (A1) - Device for mounting components


**Inventor(s):** CHOWANIEC MICHAEL [DE] ±

**Applicant(s):** SIEMENS AG [DE] ±

**Classification:**  
 - international: H02B1/00; H05K7/14  
 - European: H02B1/052

**Application number:** EP20080012517 20080710

**Priority number(s):** EP20080012517 20080710



Europäisches Patentamt  
 European Patent Office  
 Office européen des brevets

**Patent Translate**  
 Powered by EPO and Google

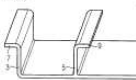
Contact

---

**Abstract of EP2144485 (A1)**  
 Translate this text into 


German  
 Spanish  
 French  
 Italian  
 Portuguese  
 Swedish

second brackets (3, 4) arranged on lateral edges of a FIG 2 (2), respectively and provided with first and second brackets. The third bracket has a third mounting unit (9) for receiving another component i.e. motor starter. A fourth bracket (6) is arranged between the second and third brackets and runs parallel to the second bracket. A T-shaped additional element partially rests on the second mounting unit.



**Hinweis**  
 Diese Übersetzung wurde maschinell erstellt. Es kann keine Gewähr für die Verständlichkeit, Richtigkeit, Vollständigkeit, Verlässlichkeit oder Zweckmäßigkeit übernommen werden. Wichtige Entscheidungen - wie geschäftliche oder finanzielle Entscheidungen - sollten nicht auf der Grundlage einer maschinellen Übersetzung getroffen werden. - Nutzungsbedingungen - Impressum - Hilfe -

**Abstract EP2144485**  
 Vorrichtung (1) erste und zweite Halterung (3, 4) an seitlichen Kanten einer in Längsrichtung verlaufenden Stange (2) angeordnet ist, jeweils und mit ersten und zweiten Bestückungseinheiten (7, 8), um eine Komponente, dh niedriger Spannung erhalten Schutzfolie.  
 Eine dritte Halterung (5) zwischen den ersten und zweiten Halterungen angeordnet ist.  
 Der dritte Träger einen dritten Montageeinheit (9) zur Aufnahme eines weiteren Komponente, dh Motorstarter.  
 Eine vierte Halterung (6) zwischen der zweiten und dritten Halterungen und verläuft parallel zu der zweiten Klammer angeordnet ist.  
 Ein T-förmiger zusätzliche Element teilweise auf dem zweiten Montageeinheit.



Europäisches Patentamt  
 European Patent Office  
 Office européen des brevets

**Patent Translate**  
 Powered by EPO and Google

Contact

---

French  
 German  
 Italian  
 Portuguese  
 Spanish  
 Swedish

**Hinweis**  
 Diese Übersetzung wurde maschinell erstellt. Es kann keine Gewähr für die Verständlichkeit, Richtigkeit, Vollständigkeit, Verlässlichkeit oder Zweckmäßigkeit übernommen werden. Wichtige Entscheidungen - wie geschäftliche oder finanzielle Entscheidungen - sollten nicht auf der Grundlage einer maschinellen Übersetzung getroffen werden. - Nutzungsbedingungen - Impressum - Hilfe -

**Abstract EP2144485**  
 Vorrichtung (1) erste und zweite Halterung (3, 4) an seitlichen Kanten einer in Längsrichtung verlaufenden Stange (2) angeordnet ist, jeweils und mit ersten und zweiten Bestückungseinheiten (7, 8), um eine Komponente, dh niedriger Spannung erhalten Schutzfolie.  
 Eine dritte Halterung (5) zwischen den ersten und zweiten Halterungen angeordnet ist.  
 Der dritte Träger einen dritten Montageeinheit (9) zur Aufnahme eines weiteren Komponente, dh Motorstarter.  
 Eine vierte Halterung (6) zwischen der zweiten und dritten Halterungen und verläuft parallel zu der zweiten Klammer angeordnet ist.  
 Ein T-förmiger zusätzliche Element teilweise auf dem zweiten Montageeinheit.

Print

**Please help us to improve our quality**

Your opinion on this translation:

Human translation

Very good

Good

Acceptable

Rather bad

Very bad

Reason for this translation

Overall information

Patent search

Patent examination

[EPO Home](#)

## Patent Translate: how does it work?

- Result of a collaboration between the EPO and Google
- Patent data represent a huge source of corpora.
- Patent documents and their translation/corresponding documents are prepared and stored in a corpora repository.
- Translation system is trained using this corpora.
- Translation quality check before launch: test fit for purpose level.

# Sentence Alignment: done for some part of the corpora

157	Upon failure of the live process B/L the recovery means causes the replicate process to take over as the re	1	Auf einen Ausfall des lebendigen Prozesses B/L hin bewirkt die Wiederherstell	g06f11/14
158	Finally, low molecular weight material (MW 200) was removed by membrane filtration and the product wa	1	Schließlich wurde niedrigmolekulares Material (Molekulargewicht &lt; 200) du	c08b37/00
159	Exemplary of the carboxylic acid protecting group represented by R 3 are allyl, benzyl, p-methoxybenzyl, p-	0	Wenn das durch R <sub>2</sub> dargestellte Aryl eine Naphthylgruppe ist, kann das Aryl 1 k	c07d501/59
160	Once the controller 50 has identified a particular command string, it outputs a control signal to activate a p	1	Sobald das Steuergerät &lt;rn> 50 &lt;/rn> eine bestimmte Befehlszeichenfolg	f21s8/10
161	Further, it is also effective for the positive photo resist composition to comprise an electron donor (D) havi	1	Darüber hinaus ist es auch effektiv, wenn die positiv arbeitende Photoresistzu	g03f7/004
162	A live monitor 52 is also connected to the computer 40 by way of the junction board 30 and displays a vide	1	Ein Live-Monitor 52 ist ebenfalls über die Anschlußplatine 30 mit dem Comput	g01n21/90
163	Performing a dot product calculation makes extensive use of the multiply accumulate operation where cor	0,5	Ergebnis = A <sub>i</sub> ·B <sub>i</sub> Beim Durchführen einer Skalarproduktberechnung wird die Mu	g06f7/00
164	Twenty grams of DBTDA were then placed in the catalyst tray and both samples placed on a paper towel o	1	Es wurden anschließend 20 g DBTDA auf das Katalysatorentablett aufgebrach	b27k3/15
165	Software control of potential conflicts between maintenance packets does not, however, present a seriou	1	Die Softwaresteuerung möglicher Konflikte zwischen Wartungspaketen stellt j	g06f15/163
166	It will be understood that the above description and the claim nomenclature is presented in a two-dimensi	1	Es ist selbstverständlich, daß die obige Beschreibung und die Anspruchsnumen	g06t9/00
167	Molecular cloning, recombination, mutagenesis and modeling studies of mAb 5C3 variable region indicated	1	Untersuchungen zur molekularen Klonierung, Rekombination, Mutagenese un	c07k16/28
168	Figure 6 is a view taken along the arrows of the B - B line in Figure 5.	1	Fig. 6 ist eine Ansicht entlang den Pfeilen der Linie B-B in Fig. 5.	b60h1/00
169	A number of sheets are contained in a paper feed cassette, for example, and a sheet feed unit is provided f	1	In einer Papierzuführkassette ist z. B. eine Anzahl von Blättern enthalten, und	b65h3/38
170	Thus, the "core region" covers the regions which will in use cover the body opening from which the exudate	1	Somit bedeckt die "Kernregion" die Regionen, die bei der Anwendung die Körp	a61f13/15
171	The material in Step 2 is monomethylamine, which may contain methanol and dimethyl ether along with ot	1	Das Material in Schritt 2 ist Monomethylamin, das Methanol und Dimethyleth	b01j29/06

# Quality level ranking

Assessment		Usable for PATENT public	Usable for PATENT search	Usable for PATENT examine
<b>5</b>	Accurate + consistence IPC vocabulary	Yes	Yes	Yes
<b>4</b>	Fluent - consistence IPC vocabulary	Yes	Yes	Yes/No
<b>3</b>	Actionable	Yes	Yes	-
<b>2</b>	May be actionable	Yes/No	-	-
<b>1</b>	Not useful	-	-	-

## Current achievements

- Corpora Repository contains corresponding patent documents for the following language pairs which was used for translation system training:
  - Batch1: EN-(FR,DE,PT,IT,ES,SV)
  - Batch 2 EN-(HU.PL.FI.NL.NO.DK.EL)
  - Asian languages: EN-CN and JP (partial)
- Peaks of usage: 35 000 requests per day for Patent Translate
- 188 million different machine translations of complete patent documents can be accessed 'on the fly', using the current language pairs offered = 940 years of work for 1000 translators if done manually



# Future plans

- Project to be completed end of 2014: 34 languages European plus Asian.

## 2013-2014

- Turkish, Czech, Slovak, Bulgarian, Estonian, Romanian, Icelandic, Croatian, Slovenian, Latvian, Lithuanian, Albanian, Macedonian, Serbian - Russian-Japanese-Korean

## 2014

- Languages to and from French and German.

# Patent Translate : illustrative example (Description)

US2006000930 (A1)
Bibliographic data
<b>Description</b>
Claims
Mosaics
Original document
Cited documents
Citing documents
INPADOC legal status
INPADOC patent family

## Description: US2006000930 (A1) — 2006-01-05

[★ In my patents list](#)
[➤ EP Register](#)
[➔ Report data error](#)

Print

### Servo valve for controlling an internal combustion engine fuel injector

#### Description of US2006000930 (A1)

Translate this text into **i**

powered by EPO and Google

The EPO does not accept any responsibility for the accuracy of data and information originating from other authorities than the EPO; in particular, the EPO does not guarantee that they are complete, up-to-date or fit for specific purposes.

#### Description of US2006000930

Translate this text into **i**

German  
 Spanish  
 French  
**Italian**  
 Portuguese  
 Swedish

[0002] As is known, an injector comprises a control chamber bounded outside the control chamber, comprising a control rod having a ball engaging the conical seat, and so vary the pressure inside the control chamber. The pressure of the fuel in the outlet hole is kept the outlet hole closed when the pressure of the fuel in the outlet hole is less than the pressure of the fuel in the control chamber.

[0003] Current market demand is for a servo valve which can be used in the known

[0004] When subjected to voltage, the control servo valve also comprises a shutter which in turn comprises a ball engaging the conical seat, and is activated by an electromagnet to move axially to and from the seat to open and close the outlet hole and so vary the pressure inside the control chamber.

French
German
<b>Italian</b>
Portuguese
Spanish
Swedish

#### Avviso

La presente traduzione è stata eseguita mediante sistema meccanizzato. Pertanto non si garantisce la sua intelligibilità, precisione, completezza, affidabilità o idoneità a impieghi specifici e se ne sconsiglia l'uso in sede di decisioni di natura commerciale o finanziaria. - [Condizioni di utilizzo](#) - [Avviso a norma di legge](#) - [Aiuto](#) -

#### Description US2006000930

[0001] La presente invenzione si riferisce ad una servovalvola di controllo di un motore a combustione interna iniettore di combustibile.

[0002] Come è noto, un iniettore comprende un corpo iniettore che definisce un ugello per iniettare combustibile nel motore, e ospita un'asta di comando mobile lungo un rispettivo asse di attivare una spina di chiusura dell'ugello.

Il corpo iniettore ospita anche una valvola elettromagnetica servo controllo comprendente una camera di controllo delimitata assialmente da un lato dall'asta di comando, e dall'altra da una parete di fondo avente un foro di uscita che, all'esterno della camera di controllo, esce assialmente all'interno di una sede conica.

La servovalvola di controllo comprende inoltre un otturatore che a sua volta comprende una sfera che si impegna nella sede conica, ed è attivato da un elettromagnete per spostarsi assialmente da e verso il sedile per aprire e chiudere il foro di uscita della camera di controllo.

Più specificamente, l'otturatore comprende una sfera che si impegna nella sede conica, e la spinta assiale di una molla che agisce sulla sfera. L'elettromagnete non è eccitato quando la sfera è nella sede conica.

[0003] domanda di mercato attuale per un servovalvola di controllo che può essere usato nelle soluzioni note sopra descritte.

[0004] Quando sono sottoposti a tensione, tuttavia, attuatori piezoelettrici possono esercitare spinta ma non tirare, e pertanto non può essere usato nelle soluzioni note sopra descritte.

[0005] Inoltre, attuatori piezoelettrici produrre uno spostamento relativamente piccolo, così che, per raggiungere la necessaria di carburante tratti di flusso di scarico, sistemi di amplificazione viaggio devono essere forniti, o parzialmente il foro di uscita zona di tenuta aumentata.

Da un lato, sistemi di amplificazione di viaggio sono indesiderabili, principalmente essendo complessi ed ingombranti, e, dall'altro, un aumento nella zona di sigillatura aumenterebbe la forza assiale esercitata dalla pressione del combustibile sull'otturatore nella posizione chiusa, in modo che il precarico della molla dovrebbe essere aumentato per mantenere l'otturatore chiuso, e una maggiore

The control servo valve also comprises a shutter which in turn comprises a ball engaging the conical seat, and is activated by an electromagnet to move axially to and from the seat to open and close the outlet hole and so vary the pressure inside the control chamber.

the EPO; in

uses a control rod

control servo valve

outlet hole which,

ch in turn comprises

se the outlet hole and

exerted on the ball by

bringing preloaded to

used in the known

# MT Example

14.

Dispositif selon la revendication 13, comprenant un dispositif de commutation permettant d'activer ledit dispositif lors de l'insertion de ladite clé dans une serrure.

15.

Dispositif selon la revendication 14, comprenant une indication visuelle de l'état de la serrure comme mis à jour par le dispositif de traitement.

A device as claimed in claim 13 including a switch arrangement used to activate said device upon insertion of said key in a lockset.

age comprend une indication

# MT Example

1.  
Ce qui est revendiqué est la suivante: 1.  
Procédé de synchronisation d'une représentation d'affichage de données avec des changements associés aux données, le procédé comprenant:  
déterminer un changement de données, les données associées à une représentation graphique des données;  
calculer une valeur de temps préféré pour la synchronisation de la représentation graphique de la variation déterminée de données, la valeur de temps préféré basé au moins partiellement sur une durée de temps nécessaire pour effectuer au moins une synchronisation préalable de la représentation graphique;  
la comparaison d'une valeur de temps actuelle à la valeur de temps préféré, et  
la détermination d'un temps de synchronisation sur la base de résultats de la comparaison, le temps de synchronisation identifiant un temps d'initier une synchronisation de la représentation graphique.
2. determining a synchronization time based on results of the comparison, the synchronization time identifying a time to initiate a synchronization of the graphical representation.  
Procédé selon la revendication 1, dans lequel les données sont stockées dans des fichiers dans un système de fichiers.
2.  
Procédé selon la revendication 1, dans lequel les données comprennent des attributs d'objets dans une banque de propriétés.

# Conclusion

MT is more than ever a must in the context of the global patent documentation:

- The size of patent collections is increasing
- Systematic manual translation is not an option
- MT has proven to be fit for purpose

# Thank You

[www.epo.org](http://www.epo.org)

[pschwander@epo.org](mailto:pschwander@epo.org)

## **Invited Talk 3**

**Takashi INABA**

**Japan Patent Office**

# JPO's Approach for Machine Translation

**- To establish productive utilization method and  
create appropriate policies**

Takashi Inaba

Assistant Director  
Research and Policy Planning Section  
Patent Information Policy Planning Office  
Information Dissemination and Policy Planning Division  
General Affairs Department  
Japan Patent Office

0

## Contents

1. JPO's Current Translation Utilization
  - 1-1. Translation Role in Examination Work
  - 1-2. Mutual Utilization of Examination Information with Foreign Offices through Machine Translation
  - 1-3. Industrial Property Information Dissemination to the Public in English
2. Approach for Quality Improvement of Machine Translation (Japanese to English)
3. Approach for Foreign Document Access Improvement through Machine Translation
4. Quality Assessment of Machine Translation

1



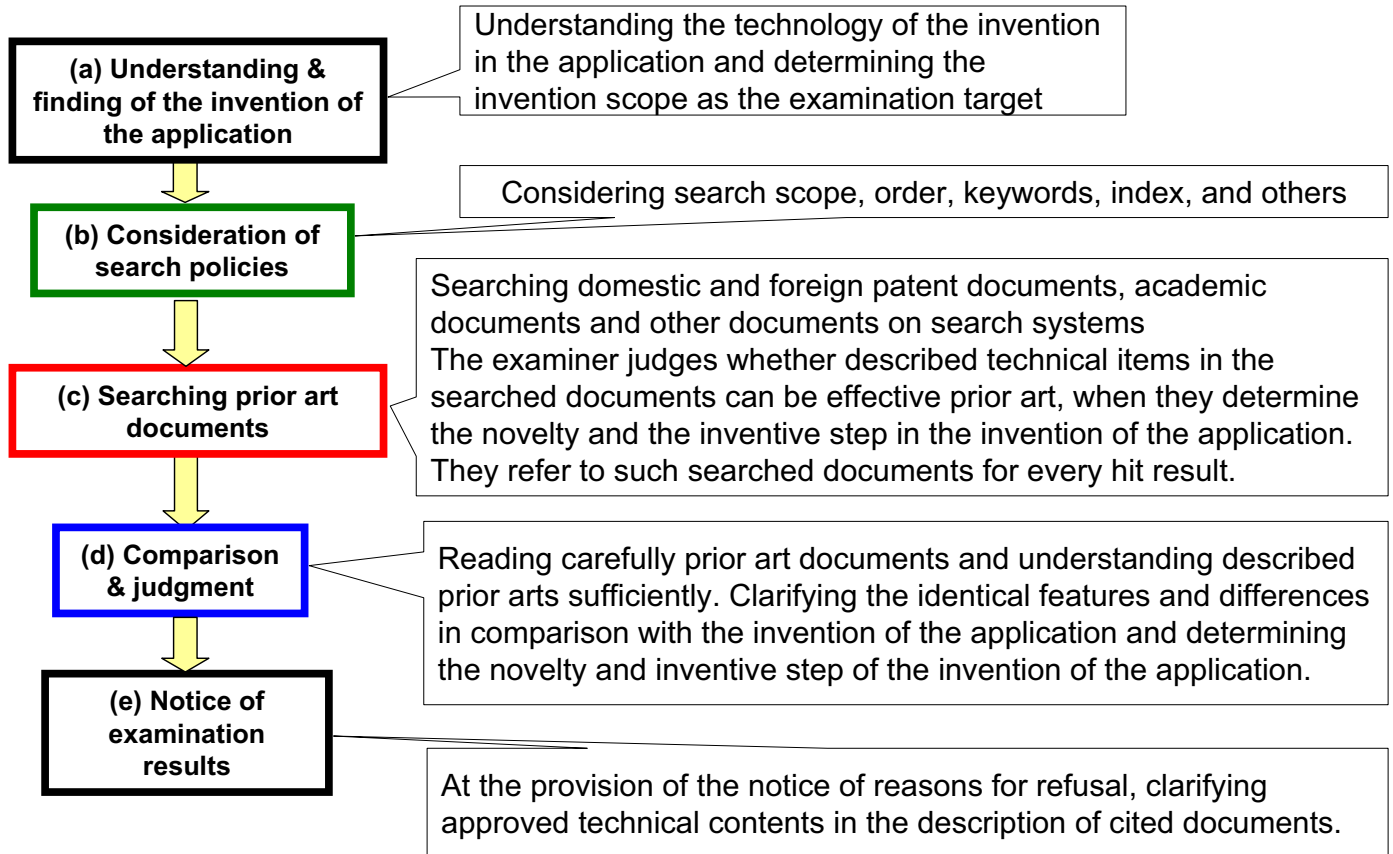


## *1. JPO's Current Translation Utilization*



### *1-1. Translation Role in Examination Work*

# 1-1. (1) Outline of Patent Examination Operation



# 1-1. (2) Stored Data of Domestic and Foreign Patent Documents (English & Japanese)

	JP	US	EP	WO	KR	CN	DE	FR	CA	GB	CH	Others (*)
Japanese full text	○				□	□						
Japanese abstract		○	○			△						
English full text		○	○	○								
English abstract	○	○	○	○	○	○	○	○	○	○	○	○

○ Original sentences

○ Human-translated sentences

\*: DE utility models and RU, AU and other country full texts and abstracts

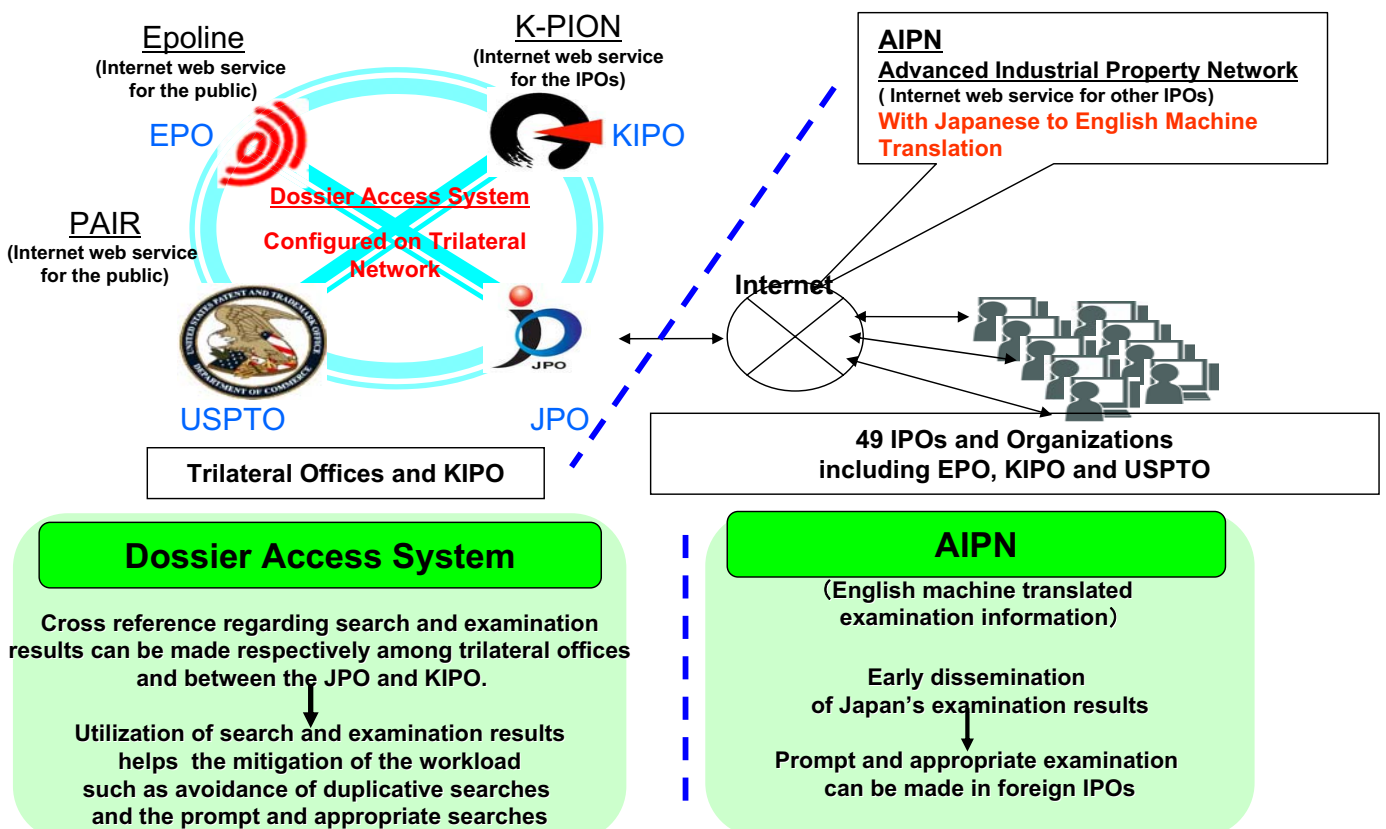
△ Start the creation through human translation from FY2012

□ Will create through machine translation in FY2014

# 1-2. Mutual Utilization of Examination Information with Foreign Offices through Machine Translation

## 1-2. (1) Dossier Access System·AIPN

### - Infrastructure to access to examination information -



## 1-2. (2) Provision Example of Japanese to English Machine Translated Examination Information

- Provision of examination information to other Offices through AIPN
- Examination information is **machine translated into English**.

**ENGLISH JAPANESE**  
**Note: Japanese environment is required to properly display Japanese characters.**  
**You must install and use a TIFF image plug-in on your system in order to view image files directly.**

**ENGLISH JAPANESE**  
**Note: Japanese environment is required to properly display Japanese characters.**  
**You must install and use a TIFF image plug-in on your system in order to view image files directly.**

**Disclaimer:**  
This English translation is produced by machine translation and may contain errors. The JPO, the INPIT, and those who drafted this document in the original language are not responsible for the result of the translation.

**Notes:**  
1. Untranslatable words are replaced with asterisks (\*\*\*\*).  
2. Texts in the figures are not translated and shown as it is.

Translated: 16:32:55 JST 06/05/2008  
Dictionary: Last updated 05/30/2008 / Priority:

**Specification**

[Document Name] Description

[Title of the Invention] Flexible copper-clad sheet

[Claim(s)]

[Claim 1] In the flexible copper-clad sheet with which the copper layer was formed on the flexible polymer base material (1) The surface of a flexible polymer base material is mostly dotted with the independent minute metal membrane at homogeneity. (2) The part which is not dotted with the metal membrane with the minute surface of a flexible polymer base material has average depth (d)0.1-2.0micrometer impression structure from the surface, and covers a minute metal membrane and impression structure on the surface of (3) flexibility polymer base material. The flexible copper-clad sheet characterized by forming the intermediate metal layer and the copper layer in this order.

**ENGLISH JAPANESE**  
**Note: Japanese environment is required to properly display Japanese characters.**  
**You must install and use a TIFF image plug-in on your system in order to view image files directly.**

produced by machine translation and may contain errors. The JPO, the INPIT, and those who drafted this language are not responsible for the result of the translation.

replaced with asterisks (\*\*\*\*).  
not translated and shown as it is.

/05/2008  
/30/2008 / Priority:

**Notification of reasons for refusal**

Notification of Reasons for Refusal

Application for patent 2001-123456  
sei 15(2003) August 12  
GAMI, Nobuhiro 9341 3S00  
licant: NISHIKAWA, Shigeaki  
Patent Law Section 29(2)

should be refused for the reason mentioned below. If the applicant has any argument  
such argument should be submitted within 60 days from the date on which this  
notified.

Reason

(s) in the each claim listed below of this patent application should not be granted a  
vision of Patent Law Section 29 (2) for the reason that the claimed invention(s)  
was made by persons who have common knowledge in the technical field to which the

8

## 1-3. Industrial Property Information Dissemination to the Public in English

9

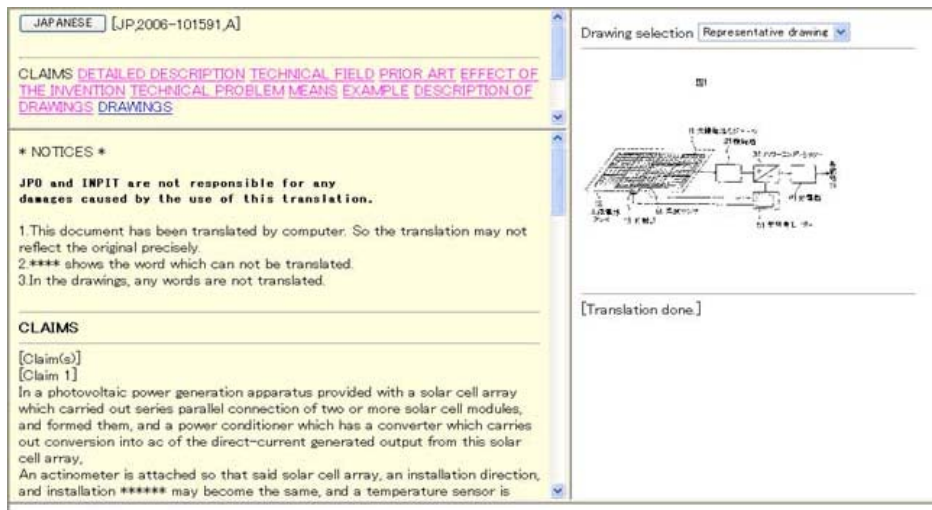


**A list of IPDL's English Database**


<p><b>1. Patent &amp; Utility Model Database</b>  <u>Database with Machine Translation</u>                  Patent &amp; Utility Model DB                  Patent &amp; Utility Model Concordance                  FI/F - term Search  <u>Database with human translation</u>                  PAJ (Patent Abstracts of Japan) Search                  Patent Map Guidance</p>	<p><b>2. Design Database</b>  <u>Database with Machine Translation</u>                  Design DB</p>	<p><b>3. Trademark Database</b>  <u>Database</u>  <u>(Bibliographic Data including some pieces of information such as Dates and Numbers)</u>                  Japanese Trademark Database                  Japanese Figure Trademarks                  Japanese Well-Known Trademark                  List of Goods and Services</p>
--	---	--

**Patent & Utility Model Gazette Database**

- Patents and utility model gazettes can be searched by number and classification (FI/F term).
- All pieces of information (except drawings) is translated by Machine Translation system.



An example of publications of unexamined patent applications



## 2. Approach for Quality Improvement of Machine Translation (Japanese to English)

12

### 2. (1) Routine Operation

#### Collection and registration of unknown words

Unknown and untranslatable words are collected in IPDL and AIPN and added to the user dictionary (5,000 words/year).  
As of October, 2012, 80,000 words were included.

#### Feedback from foreign IPOs

Mistranslation feedback provided by AIPN users (examiners of foreign offices such as the EPO and the USPTO) is analyzed and registered in the dictionary.

#### Upgrade of machine translation engine

Enrichment of translation knowledge and increase in the number of technical and IP terms (December, 2011)

#### Establishment of translation memories

Fixed expressions used for notification of reasons for refusal are registered in translation memories.

- (a) The survey was conducted for the quality improvement of Japanese to English machine translation on AIPN in 2003, 2007 and 2008. Based on the results, fixed expressions were extracted and about 1,110 sentences were registered.
- (b) Fixed expressions (about 460 sentences) which examiners use in drafting notification of reasons for refusal were registered in 2009, 2010 and 2012.



13

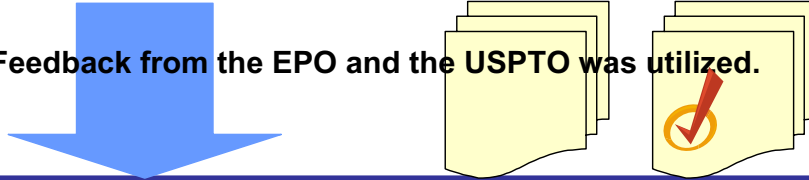
## 2. (2) Approach in IP5 Project

### ○ Mutual Machine Translation Project

#### Error review feedback process (September to December, 2011)

- English speaking offices provided specific feedback for English machine-translated sentences provided by non-English speaking offices, based on standards regarding selected words, the degree of understanding, grammar and other items.

Feedback from the EPO and the USPTO was utilized.



#### Upgrade of machine translation engines by non-English speaking offices (From January, 2012)

- Based on the feedback results, non-English speaking offices upgraded machine translation engines and took other measures.
- Non-English speaking offices including the JPO reported improvement results to IP5.
- In the future, the quality assessment will be conducted for the examination of improvement effects

## 3. Approach for Foreign Document Access Improvement through Machine Translation

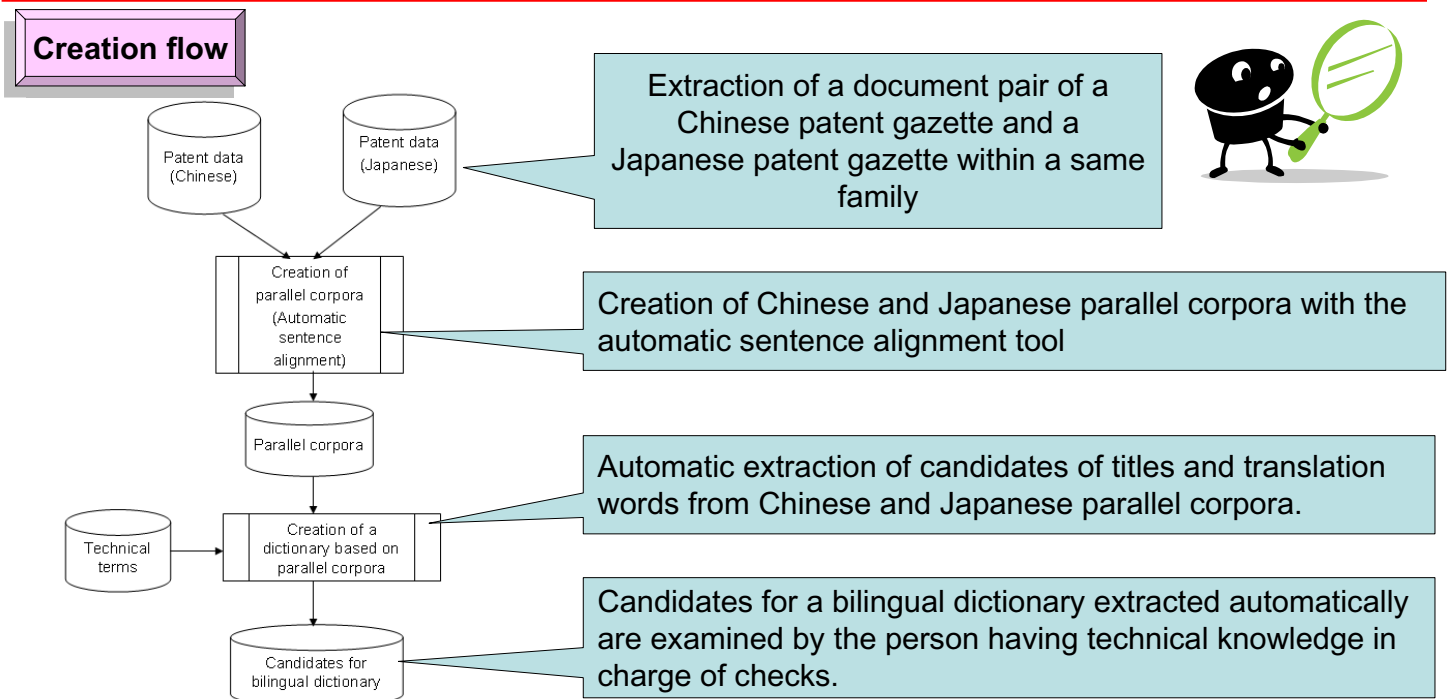
### 3. (1) Approach for Chinese and Korean Documents using Machine Translation

- The number of Chinese and Korean patent documents have been increasing. The urgency is required for arranging environments, where such documents are easily searchable.
- To implement such environments earlier, the development of Japanese search system with Machine Translation is required.

	Roadmap (planned)					
	FY2011	FY2012	FY2013	FY2014	FY2015	From FY2016
Japanese abstracts	development & data creation	Start of provision: Search of Japanese abstracts of Chinese utility models (machine translation)				Duration not yet determined
		Development/data creation	Start of provision: Search of Japanese abstracts of Chinese Patents (human translation and other translation methods)			Duration not yet determined
Development of Chinese to Japanese dictionaries	Research of dictionary	Development of Chinese to Japanese machine translation dictionaries	Creation and addition of dictionaries based on Japanese abstracts		System to be released in FY2014	
Full text search through Chinese and Korean to Japanese machine translation system		Efficient utilization for full text search	Procurement support and design/development	Japanese full text search of Chinese and Korean patent and utility model documents		

### 3.(2) Development of Chinese to Japanese Dictionaries

- The JPO aims for the creation of the bilingual dictionaries with a million pairs of Chinese and Japanese words including technical terms and other terms used in patent documents. The purpose is for the quality improvement of Chinese to Japanese machine translation of documents.





- Machine Translation of Chinese and Korean documents will enable the creation of full-text document of Japanese patents and the search of Japanese full text.
- Finally, cross-lingual search system in Japanese will be realized.
- Examiners and system users will be able to search and access Chinese and Korean documents in Japanese.



## ***4. Quality assessment of Machine Translation***

From the following perspectives, results of quality assessment are important.

### (1) Selection of a machine translation system with appropriate quality

- Examination about whether the system to be introduced meets the required quality level.
- Appropriate assessment of the high quality system

### (2) Consideration about how machine translation is utilized in the work

- Identification about whether machine translation should be introduced in the work
- Consideration of the utilization method based on specific advantages and disadvantages

### (3) Creating of policies relating to machine translation

- Basic materials for the consideration of multilingual approach methods
- Assessment of effects of the quality improvement approach

## 4. (2) Examples of Quality Assessment Which was Implemented in the Past

### Quality assessment of Chinese to Japanese machine translation (Survey at 2011)

- Human assessment was conducted about the translation results of Chinese gazette abstracts provided by multiple Chinese to Japanese Machine Translation software/services
- Translation quality and search quality were assessed.

#### Translation quality

- (a) With or without grammatical mistranslation
- (b) With or without oversight of translation and unnecessary words
- (c) With or without use of unclear or uncommon translation words
- (d) With or without contrivance as Japanese sentence structure

It cannot be helped that evaluators objectively assess problem areas to be focused on and weigh the importance of such problems. Thus, it is highly subjective assessment.

Following is 5-scale assessment for items (a) to (d)

Score	Decision Criteria (Standard for error rate)
5	No problem is found (0%)
4	Few problems are found (10% or less)
3	Only a few problems are found (30% or less)
2	Some problems are found (50% or less)
1	Many problems are found (Exceeding 50%)

## 4. (2) Examples of Quality Assessment which Was Implemented in the Past

**Search quality (e) With or without selected keywords based on the translation results**  
**(f) Appropriateness of selected keywords**

The existence or non-existence of preselected keywords and mistranslation were identified and those numbers were counted. Thus it is highly objective assessment method.

As for (e), the following 5 scale assessment was conducted. As for (f), the following 5 scale assessment was conducted.

Score	Decision Criteria
5	All selected keywords are included.
4	15% or less of selected keywords are missing.
3	30% or less of selected keywords are missing.
2	45% or less of selected keywords are missing.
1	Exceeding 45% of selected keywords are missing.

Score	Decision Criteria
5	All Translation words contained in translation results of (e) are accurate.
4	15% or less of translation words contained in the translation results of (e) are inaccurate.
3	30% or less of translation words contained in the translation results of (e) are inaccurate.
2	45% or less of translation words contained in the translation results of (e) are inaccurate.
1	Exceeding 45% of translation words contained in the translation results of (e) are inaccurate.

• Search quality indicates “to which degree the technical terms corresponding to search keywords used in the text search are translated accurately without omission”.

• About 10 words were selected as “selected keywords” in advance. The content rate of selected words and translation accuracy of translation words were evaluated in the translation results.

22

## 4. (3) List of Quality Assessment which Were Implemented in the Past

Survey and Meeting	Translation direction	Assessment perspectives (human assessment)						Notes (Other perspectives, etc.)
		Technical terms		Content transfer		Grammar/sentence structure		
		Objectivity (*1)	Subjectivity (*2)	Objectivity	Subjectivity	Objectivity	Subjectivity	
Survey at 1998	Japanese to English	Ratio of correct translation words	3 scale assessment	-	3 scale assessment	Deduction of points in relevant assessment items	3 scale assessment	Assessment only of translation sentences and assessment of expressions
Survey at 2008	Chinese to Japanese	-	5 scale assessment	-	5 scale assessment	-	2 scale assessment	Availability and fluency of translation results
Survey at 2009	Korean to Japanese	-	-	-	5 scale assessment	-	-	
Survey at 2010	Japanese to English	Calculation of errors/sentences	-	-	5 scale assessment	Calculation of errors/sentences	-	
IP5 Meeting at 2010	Japanese to English	-	-	-	5 scale assessment	-	-	Linking the usage and the score
Survey at 2011	Chinese to Japanese	Omission and mistranslation rates of selected words	5 scale assessment	-	5 scale assessment	-	5 scale assessment	Contrivance of translation as Japanese

\*1. Highly objective assessment method

\*2. Highly subjective assessment method

As in the above list, various assessments of machine translation were implemented. However, the definite method has not been established.

23

## 4. (4) To Establish Quality Assessment Method

To establish quality assessment method, following items need to be considered.

### Standpoints and Standards of Evaluation

- Clarification of intended use of Machine Translation and its required quality
- Standpoints and standards of evaluation to secure the required quality

### Evaluation Work

- Comparative target in evaluation works
- Support tool for evaluation works

### Sentences to be evaluated

- requirements and selection methods of sentences

## 4. (5) Standpoints and Standards of Evaluation

### Utilization of machine-translated sentences

- Utilization at search of prior art documents (ref. step (c) in 1-1.(1) ) is expected.

### Required quality

1. Search results are obtained through keyword searches.
2. Documents in hit results enable users to grasp the point (to the degree the user can identify the necessity of intensive reading of such documents).

### Assessment method

Establishment of the assessment method is on-going in the following direction.

Technical terms (corresponding to the above mentioned Item #1)	Content transfer (corresponding to the above mentioned Item #2)	Grammar/sentence structure
Objective assessment (count of errors of translation words)	Subjective assessment (5 scale assessment)	Objective assessment (check of the corresponding assessment items)

As for grammar/sentence structure, understanding more specific problems is expected to help creating future policies.

### Specific assessment method (example)

1. About 150 sentences to be evaluated are considered.
2. Evaluators list technical terms in advance from sentences to be evaluated.
3. Whether such technical terms are mistranslated is confirmed.
4. Evaluators judge mistranslation, deciding whether the translation in question is within the same technical scope of the appropriate translation.
5. Evaluators count the number of mistranslations and calculate the ratio.

### Problem

How do we ensure the appropriateness and objectivity for judging mistranslation?

- Utilization of the presentation tool of synonyms?

### Specific assessment method (example)

1. About 150 sentences to be evaluated are considered.
2. Evaluators compare the results of human translation (or the original sentences in case of English) and the results of machine translation and conduct 5 scale assessment in response to the degree of the meaning transfer.
3. As 5 scale assessment, the following ranges are considered: From the level where “all critical information accurately transferred” to the level where “little or no information is transferred accurately”.

### Problem

- How do we ensure the objectivity for allocating the degree of meaning transfer into each scale degree?

- Creation of a criteria or collection of case examples?

### Specific assessment method (example)

1. About 150 sentences to be evaluated are considered.
2. It is checked whether those sentences are corresponding to predetermined assessment items.

#### Examples of assessment items

##### Errors relating to sentence structure

- ✓ Translation errors of redundant dependencies
- ✓ Incorrect dependent direction of adverbs in case of multiple verbs

##### Grammar errors

- ✓ Confusion of parts of speech
- ✓ Inappropriate relation between a modifying word and a modified word

##### Errors relating to process of symbols and others

- ✓ Incorrect identification of mathematical formula and chemical formula
- ✓ Inappropriate process of itemized expressions

### Problem

- Effective assessment items to be organized

In the past quality assessments, sentences to be evaluated were selected randomly. However, for more effective assessment, there is a room of consideration in requirements and selection methods of such sentences.

Items to be studied are as follows.

- What sentences are appropriately chosen from perspectives of sentence structures, grammar, technical terms, sentence length and other perspectives?
- Can we consider the selection of appropriate sentences to be evaluated based on the level of machine translation as the assessment target?

### Standpoints and Standards of Evaluation

- Clarification of intended use of Machine Translation and its required quality  
: Utilization at search of prior art documents, keyword searches, grasping the point
- Standpoints and standards of evaluation to secure the required quality  
: Standpoints of technical terms, content transfer and grammar/sentence structure

### Evaluation Work

- Comparative target in evaluation works: results of human translation
- Support tool for evaluation works  
: (example) presentation tool of synonyms, criteria or collection of case examples

### Sentences to be evaluated

- Requirements and selection methods of sentences  
: consideration from perspectives of sentence structures, grammar, technical terms, sentence length and level of machine translation



*Thank you for your attention*

# 特許庁における機械翻訳に関する取組

—機械翻訳の有効な活用方法の確立と適切な施策立案に向けて

特許庁総務部普及支援課特許情報企画室調査第二係長  
稲葉 崇

0

## 目次

1. 特許庁における翻訳活用の現状
  - 1-1. 審査業務における翻訳の役割
  - 1-2. 機械翻訳を活用した海外産業財産庁との審査情報の相互利用
  - 1-3. 英語による産業財産権情報の公衆への普及
2. 機械翻訳(日英)の精度向上のための取り組み
3. 機械翻訳を活用した外国文献へのアクセス向上の取組
4. 機械翻訳の品質評価

1



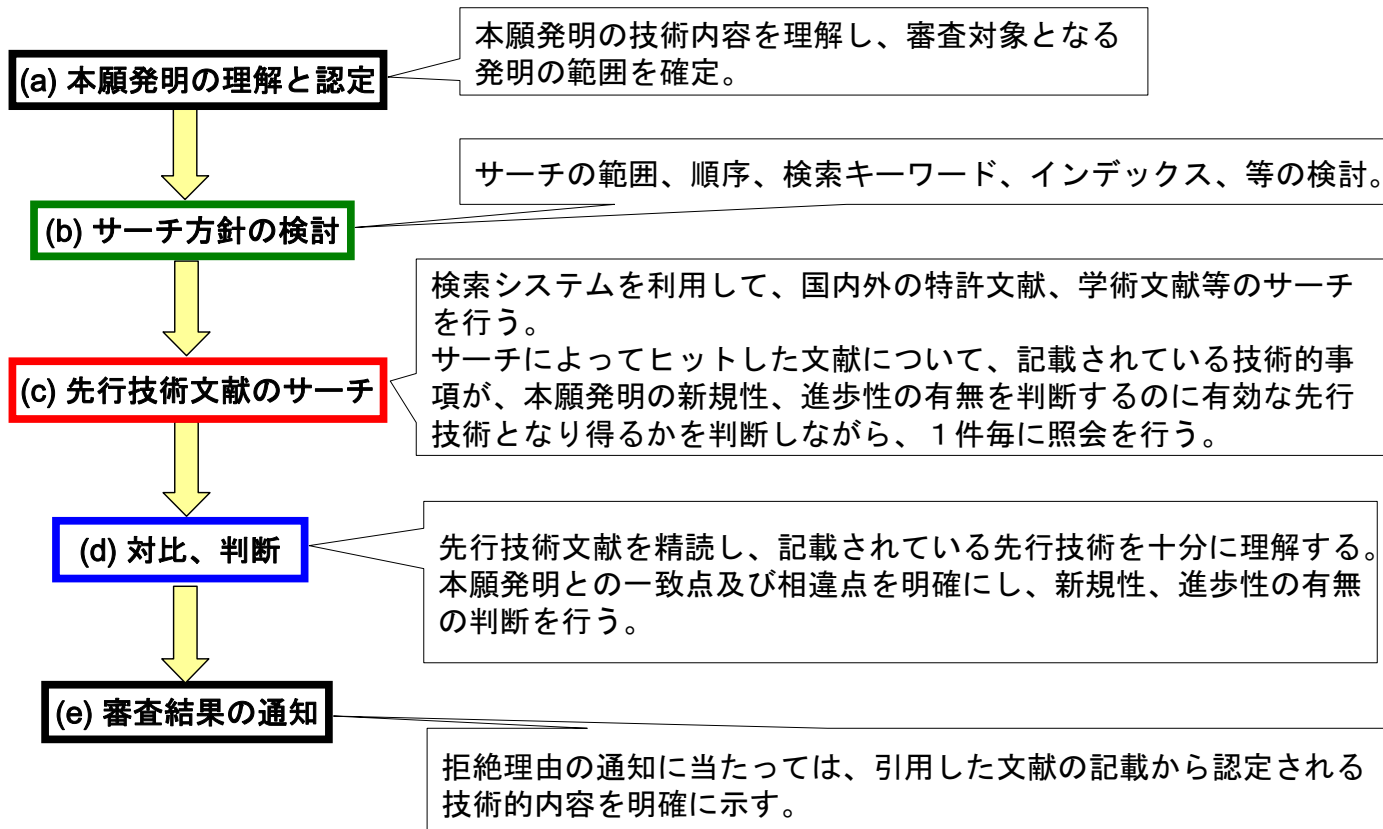


## 1. 特許庁における翻訳活用の現状



### 1-1. 審査業務における翻訳の役割

## 1-1. (1) 特許審査業務の概要



4

## 1-1. (2) 内外国の特許文献蓄積データ(和文、英文)

	JP	US	EP	WO	KR	CN	DE	FR	CA	GB	CH	他 (※)
和文全文	○				□	□						
和文抄録		○	○			△						
英文全文		○	○	○								
英文抄録	○	○	○	○	○	○	○	○	○	○	○	○

※例:DE実用やRU、AU等

○ 原文    ○ 人手翻訳文

△ 2012年度より人手翻訳により作成開始

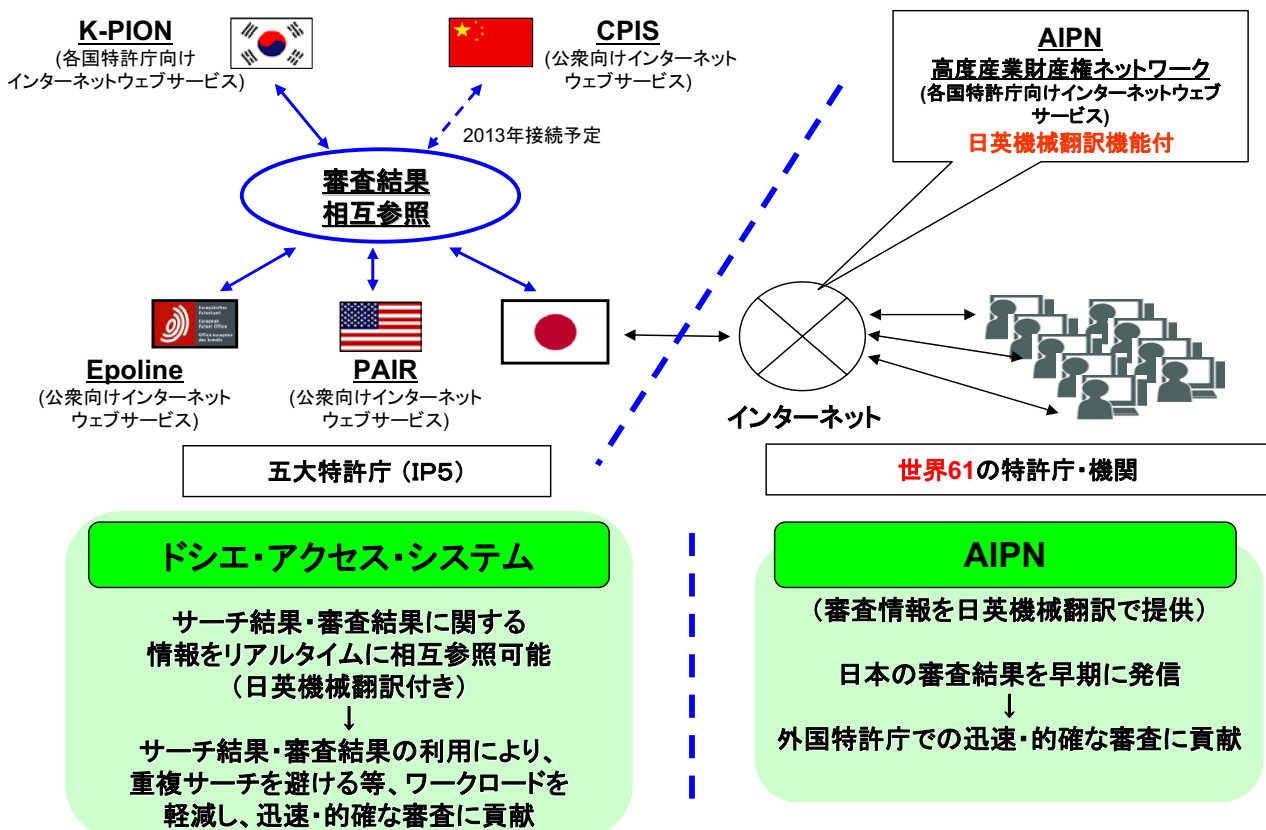
□ 2014年度に機械翻訳により作成予定

5

# 1-2. 機械翻訳を活用した海外産業財産件庁との審査情報の相互利用

## 1-2. (1) ドシエ・アクセスシステム/AIPN

～審査情報にアクセスするための基盤～



## 1-2. (2) 審査情報の日英機械翻訳による提供例

- AIPNにより審査情報を他庁へ提供
- 審査情報は**機械翻訳システムによって英語に翻訳**

**ENGLISH JAPANESE**

**Note: Japanese environment is required to properly display Japanese characters.  
You must install and use a TIFF image plug-in on your system in order to view image files directly.**

**ENGLISH JAPANESE**

**Note: Japanese environment is required to properly display Japanese characters.  
You must install and use a TIFF image plug-in on your system in order to view image files directly.**

**拒絶理由通知**

**明細書**

Disclaimer:  
This English translation is produced by machine translation and may contain errors. The JPO, the INPIT, and those who drafted this document in the original language are not responsible for the result of the translation.

Notes:  
1. Untranslatable words are replaced with asterisks (\*\*\*\*).  
2. Texts in the figures are not translated and shown as it is.

Translated: 16:32:55 JST 06/05/2008  
Dictionary: Last updated 05/30/2008 / Priority:

[Document Name] Description

[Title of the Invention] Flexible copper-clad sheet

[Claim(s)]

[Claim 1] In the flexible copper-clad sheet with which the copper layer was formed on the flexible polymer base material (1) The surface of a flexible polymer base material is mostly dotted with the independent minute metal membrane at homogeneity. (2) The part which is not dotted with the metal membrane with the minute surface of a flexible polymer base material has average depth (d)0.1-2.0micrometer impression structure from the surface, and covers a minute metal membrane and impression structure on the surface of (3) flexibility polymer base material. The flexible copper-clad sheet characterized by forming the intermediate metal layer and the copper layer in this order.

**ENGLISH JAPANESE**

**Note: Japanese environment is required to properly display Japanese characters.  
You must install and use a TIFF image plug-in on your system in order to view image files directly.**

produced by machine translation and may contain errors. The JPO, the INPIT, and those who drafted this language are not responsible for the result of the translation.

replaced with asterisks (\*\*\*\*).  
not translated and shown as it is.

/05/2008  
/30/2008 / Priority:

**拒絶理由通知**

Notification of Reasons for Refusal

Application for patent 2001-123456  
sei 15(2003) August 12  
GAMI, Nobuhiro 9341 3S00  
licant: NISHIKAWA, Shigeaki  
Patent Law Section 29(2)

should be refused for the reason mentioned below. If the applicant has any argument  
such argument should be submitted within 60 days from the date on which this  
notched.

Reason

(s) in the each claim listed below of this patent application should not be granted a  
vision of Patent Law Section 29 (2) for the reason that the claimed invention(s)  
in made by persons who have common knowledge in the technical field to which the

8

## 1-3. 英語による産業財産権情報の公衆への普及

9



## IPDLの英語データベース一覧

### 1. 特許・実用新案データベース

#### 機械翻訳を活用したデータベース

特許・実用新案公報DB  
 特許・実用新案文献番号索引照会  
 FI/Fターム検索

#### 人手翻訳によるデータベース

PAJ (Patent Abstracts of Japan) 検索  
 パテントマップガイダンス

### 2. 意匠データベース

#### 機械翻訳を活用したデータベース

意匠公報DB

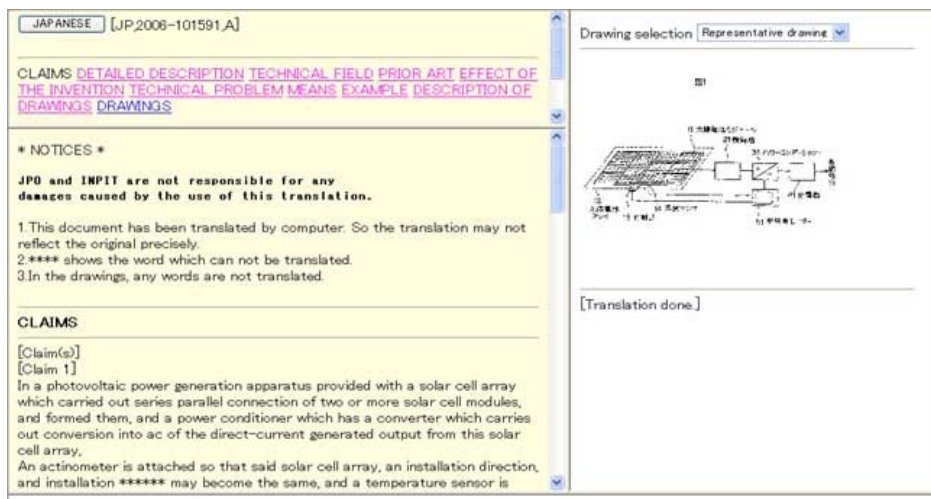
### 3. 商標データベース

#### データベース(日付や番号等一部の書誌事項のみ)


日本国周知・著名商標検索  
 商品・サービス国際分類表

## 特許・実用新案公報データベース

- 特許・実用新案の各種公報を番号照会・分類検索 (FI/Fターム) 可能
- 全ての情報(図面を除く)を機械翻訳システムによって翻訳



公開特許公報を英語に機械翻訳したものの一例



## 2. 機械翻訳(日英)の精度向上のための取り組み

12

### 2. (1) 定常的な取組

#### 未知語の収集・登録

IPDLおよびAIPNにおいて、翻訳不可能な単語(未知語)のログを収集し、ユーザー辞書に追加登録(5,000語/年)  
2012年10月時点、8万語を収録

#### 海外特許庁からの誤訳フィードバック

AIPNユーザー(EPO、USPTOをはじめとする海外特許庁審査官)からの誤訳フィードバックを分析の上辞書登録



#### 機械翻訳エンジンのバージョンアップ

翻訳知識の強化や専門用語・知財用語数の増加(2011年12月)

#### 翻訳メモリの構築

拒絶理由通知に利用される定型表現を翻訳メモリに登録

- ①AIPN日英機械翻訳の翻訳精度向上に向けた調査(2003, 2007, 2008年度実施)の結果、抽出した定型表現の登録(約:1110文登録)
- ②審査官が拒絶理由通知書の起案時に利用する定型表現(汎用文例)の登録(2009, 2010, 2012年度:約460文登録)

13

## 2. (2) 5大特許庁 (IP5) プロジェクトにおける取組

### ○相互機械翻訳プロジェクト (Mutual Machine Translation)

#### エラーレビューフィードバックプロセス (2011年9月～12月)

- ・非英語圏の庁が提供している英語機械翻訳文について、英語圏の庁が語の選択、理解度、文法等の基準に基づき具体的なフィードバックを行う



#### 非英語圏の各庁による機械翻訳エンジンのアップグレード(2012年1月～)

- ・フィードバック結果に基づき、非英語圏の庁は機械翻訳エンジンのアップグレード等を行った。
- ・2012年9月にJPOを含む非英語圏の庁から、機械翻訳エンジンの改善結果を五庁に報告した。
- ・今後は、改善効果の検証のため、品質評価を実施予定。

## 3. 機械翻訳を活用した外国文献へのアクセス向上の取り組み

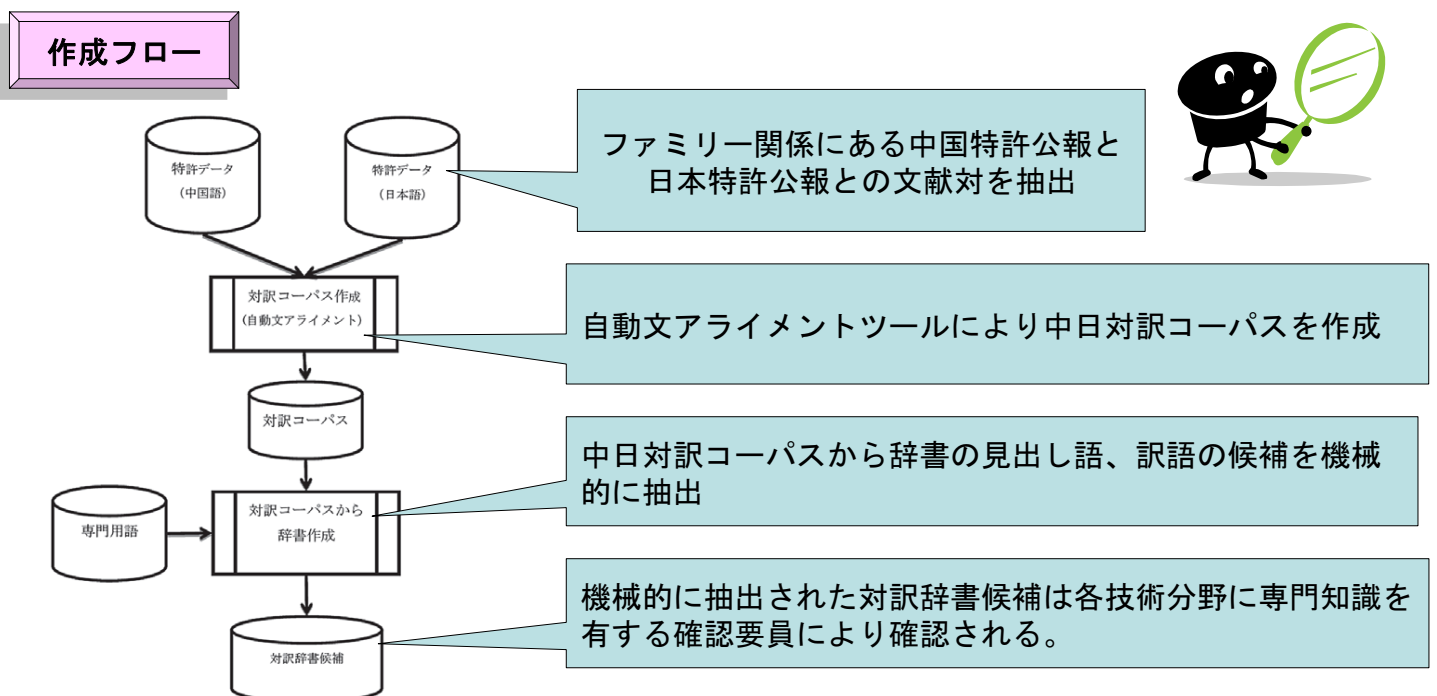
### 3. (1) 機械翻訳を利用した中韓文献への対応

- 増大する中韓特許文献等を容易に調査できる環境整備の必要が高まっている。
- 当該環境を早期に実現するためには、機械翻訳を活用した日本語による検索システムの開発が必要。

	ロードマップ(予定)						
	平成23年度	平成24年度	平成25年度	平成26年度	平成27年度	平成28年度～	
和文抄録	開発&データ作成	提供開始	中国実用和抄(機械翻訳)の検索			継続期間は未定	
		開発・データ作成	提供開始	中国特許和抄(人手翻訳等)の検索			継続期間は未定
中→日辞書開発	辞書調査研究	中日機械翻訳辞書開発	和抄から辞書作成・追加				
中韓→日への機械翻訳を利用した全文検索システム			全文検索システムに有効活用	調達支援・設計開発	平成26年度中にシステムリリース予定。	中韓文献の特許・実用 日本語全文検索	

### 3. (2) 中日辞書開発

- ・ 中国文献の中日機械翻訳の精度向上に資することを目的として、特許文献で使用されている技術用語等について100万語の中日対訳辞書データの作成を目指す





- 中韓文献の機械翻訳により、日本語全文テキストを作成し、日本語による全文テキスト検索を可能化。
- 最終的に、日本語による多言語文献の横断的検索を実現。
- 審査官及び一般ユーザーの双方が、中韓文献を日本語によって調査・閲覧可能に。

## 4. 機械翻訳の品質評価



品質評価の結果は、以下の観点から重要である。

### (1) 適切な品質の機械翻訳システムの選択

- ・導入システムが求める品質レベルに達しているかの検証
- ・高い品質を提供するシステムの適正な評価

### (2) 機械翻訳の業務での活用の形の検討

- ・業務への機械翻訳の導入可否の判断
- ・具体的な強みや弱みを踏まえた活用の形の工夫

### (3) 機械翻訳に関する施策立案

- ・多言語へのアプローチ方法検討の基礎資料
- ・品質向上のための取組の効果の検証

## 4. (2) 過去に実施した品質評価の例

### 中日機械翻訳の品質評価（H23年調査）

- ・複数の中日機械翻訳ソフトウェア／サービスについて、中国公報の要約部分の翻訳結果を人手により評価。
- ・**翻訳精度、検索精度**について、評価を実施。

#### 翻訳精度

- (a) 文法的な誤訳の有無
- (b) 訳漏れ、不要な語句の有無
- (c) 意味不明又は一般的でない訳語の使用の有無
- (d) 日本語文章構成としての不自然の有無

着目する問題箇所や、その問題の軽重の評価は主観的にならざるを得ない面があり、**主観性の高い評価方法**といえる。

(a)-(d)の各々について以下の5段階評価

配点	判定基準（誤りの割合の目安）
5	問題箇所なし（0%）
4	ほぼ問題箇所なし（～10%）
3	若干の問題箇所を含む（～30%）
2	かなりの問題箇所を含む（～50%）
1	多くの問題箇所を含む（51%～）

#### 4. (2) 過去に実施した品質評価の例

##### 検索精度

- (e) 翻訳結果における選定キーワードの有無
- (f) 選定キーワードの訳語の適確性

予め選定キーワードの有無や誤訳を判断して、その数をカウントするというもので、**客観性の高い評価方法**といえる。

(e) については以下の5段階評価

配点	判定基準
5	選定キーワードを全て含んでいる
4	選定キーワードの15%以下が欠落している
3	選定キーワードの30%以下が欠落している
2	選定キーワードの45%以下が欠落している
1	選定キーワードの45%以上が欠落している

(f) については以下の5段階評価

配点	判定基準
5	(e)で翻訳結果に含まれた選定キーワードの全ての訳語が的確である
4	(e)で翻訳結果に含まれた選定キーワードの15%以下の訳語が不的確である
3	(e)で翻訳結果に含まれた選定キーワードの30%以下の訳語が不的確である
2	(e)で翻訳結果に含まれた選定キーワードの45%以下の訳語が不的確である
1	(e)で翻訳結果に含まれた選定キーワードの45%以上の訳語が不的確である

- ・ 検索精度とは、「テキスト検索において用いられる検索キーワードに対応する技術用語が、どの程度、漏れなく正確に翻訳されているかという観点の精度」を指す。
- ・ 技術用語を「選定キーワード」としてあらかじめ10語程度選定し、翻訳結果におけるこれら選定キーワードの含有率と、訳語の的確性とを評価した。

#### 4. (3) 過去に実施した品質評価の一覧

事業	翻訳方向	評価観点（人手評価）						備考（その他の観点、等）
		技術用語		内容伝達		文法・構文		
		客観 ※1	主観 ※2	客観	主観	客観	主観	
H10年調査	日英	正しい訳語の割合	3段階	-	3段階	評価項目該当による減点	3段階	訳文のみでの評価、表現の評価
H20年調査	中日	-	5段階	-	5段階	-	2段階	翻訳結果としての利用性、流暢さ
H21年調査	韓日	-	-	-	5段階	-	-	
H22年調査	日英	誤り数/文を算出	-	-	5段階	誤り数/文を算出	-	
H22年IP5	日英	-	-	-	5段階	-	-	用途とスコアの対応付けの実施
H23年調査	中日	選定ワードの欠落、誤訳率	5段階	-	5段階	-	5段階	日本語としての不自然さ

- ※1 客観性の高い評価方法
- ※2 主観性の高い評価方法

上表の通り、これまで機械翻訳の品質評価は種々実施されてきたが、定まった方法は確立されていない。

品質評価方法の確立に向けて、以下の項目について検討する必要がある。

### 評価の観点、基準

- 機械翻訳文の用途、求められる品質の明確化
- 求められる品質を担保するための評価観点、基準

### 評価作業

- 評価作業における比較対象
- 評価作業の支援ツール

### 評価用の問題文

- 問題文の要件や選定方法

### 機械翻訳文の用途

- ・ 先行技術文献のサーチ（1-1. (1) のステップ(c)）における活用が想定される。

### 求められる品質

1. キーワードによる検索でヒットすること
2. ヒットした文献の大まかな内容が把握できること（精読の必要性が判断できる程度に）

### 評価方法

以下のような方向で評価方法の具体化を進めている。

技術用語 (上述の1. に対応)	内容伝達 (上述の2. に対応)	文法・構文
客観評価 (訳語の誤り数のカウント)	主観評価 (5段階)	客観評価 (評価項目への該当性の チェック)

文法・構文については、より具体的な問題点を把握して、その後の施策等に役立てることを想定。

### 評価の具体的方法（例）

1. 評価文として、150文程度を想定。
2. 評価者が評価文の中から技術用語を予めリストアップ
3. 当該技術用語の翻訳が誤訳となっていないか確認
4. 誤訳の判断は、適切な訳に対して技術的に同義の範囲か否かにより判断
5. 誤訳のものをカウントし、その割合を算出する。

### 課題

- 誤訳の判断についての、適切性、客観性の担保
  - － 同義語の提示ツールの活用？

### 評価の具体的方法（例）

1. 評価文として、150文程度を想定。
2. 評価者が人手翻訳結果（又は、原文が英語の場合は原文）と、機械翻訳結果を比較し、意味の伝達度に応じて5段階評価を行う。
3. 5段階としては、“全ての重要情報が正確に伝達されている“レベルから、”正確に伝達されている情報はほとんどない”レベルまでの、広いレンジを想定。

### 課題

- 意味伝達度の各段階への対応付けの客観性の担保
  - － 各段階への当てはめ基準や事例集の作成？

### 評価の具体的方法（例）

1. 評価文として、150文程度を想定。
2. 各評価文について、予め設定した評価項目に該当するかをチェックする。

#### 評価項目の例

##### 構文解析に関する誤り

- ✓ 重複した係り受けの訳出を誤っている
- ✓ 動詞が複数の場合に、副詞の係り受け先が誤っている

##### 文法に関する誤り

- ✓ 品詞の取り違え
- ✓ 修飾語と被修飾語の位置関係が不適切

##### 記号等の処理に関する誤り

- ✓ 数式、化学式の部分の認識誤り
- ✓ 箇条書き表現の処理が不適切

### 課題

- 有効な評価項目の設定

過去の品質評価では、評価用の問題文はランダムに選定されたが、有効な評価の実施のため、問題文の要件や選定方法についても検討の余地がある。

検討事項として、以下の点が挙げられる。

- 構文・文法、技術用語、文の長さ、等の観点から、どのような問題文を用いるのが適切か？
- 評価対象の機械翻訳のレベルを踏まえた適切な問題文の設定は考えられないか？

### 評価の観点、基準

- 機械翻訳文の用途、求められる品質の明確化  
: 先行技術文献のサーチにおける活用、キーワード検索、大まかな内容把握
- 求められる品質を担保するための評価観点、基準  
: 技術用語、内容伝達、文法・構文の観点から具体化

### 評価作業

- 評価作業における比較対象 : 人手翻訳結果
- 評価作業の支援ツール : (例) 同義語の提示ツール、評価の当てはめ基準や事例集

### 評価用の問題文

- 問題文の要件や選定方法  
: 構文・文法、技術用語、文の長さ、機械翻訳レベルを踏まえた検討

ご静聴ありがとうございました。

# Meeting Report

**Terumasa EHARA, Hiroshi ECHIZEN'YA**

**AAMT/Japio Special Interest Group on**

**Patent Translation**



# Report of Review Meeting on Evaluation Methods for Machine Translation Results in Patent Documents

---

Terumasa Ehara and Hiroshi Echizen'ya  
AAMT/Japio Special Interest Group on Patent  
Translation

## Table of contents

- Overview of the meeting
- What is MT evaluation?
- Human judgments and automatic evaluations
- Problems of judgments/evaluations
- Challenges of new human judgments
- Realization of new automatic evaluation
- Combination of human judgment and automatic evaluation

## Review Meeting in Evaluation Methods for Machine Translation Results on Patent Documents

- Date: September 7, 2012 (Fri.) 1 : 00 p.m.- 5 : 00 p.m.
- Location: Faculty of Engineering Bldg. 11, The Univ. of Tokyo
- Focus of discussion
  1. Evaluation of Machine Translation used to investigate patent documents written in a foreign language
  2. Present situation in automatic evaluation and decision of most high-quality automatic evaluation
  3. Difference between human judgment and automatic evaluation
  4. Test sets for patent translation
  5. Future of translation evaluation
- Number of participants : 96

## Program

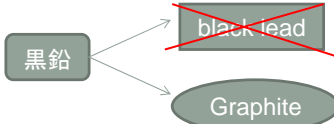
- Hiroshi Echizen-ya (Hokkai-Gakuen Univ.):  
Delight, Disappointment and Wish Automatic Evaluation brings
- Hideki Isozaki (Okayama Prefectural Univ. ):  
Recent Research Trends in Automatic Evaluation of Translation Quality and RIBES
- Hirokazu Suzuki (Toshiba Corp.):  
An Evaluation Method of 'Atmosphere-Sensitive Machine Translation'
- Isao Goto (NICT):  
Human Evaluations at the NTCIR-9 and 10 Patent Machine Translation Tasks
- Yohsuke Morita (Toyota Technical Development Corp.):  
Necessary of Machine Translation Accuracy for Patent Search  
~A Case Study of Chinese Patent Search~
- Tomoki Nagase (Fujitsu Laboratories LTD.):  
A Method for Japanese-Chinese MT Evaluation with AAMT Test-Sets

## Why do we need evaluation for machine translation?

### ➤ Purpose

- Researchers ➡ Confirmation of effectiveness of new methods, Participation in Workshop competition
- Developers ➡ Confirmation of performance improvement, Benchmark tests, Feedback to develop
- Users ➡ Reference of introduction of new system, Guidepost for improvement of operations

## What is the point of evaluation for machine translation?

- It depends on **an evaluator** and **the purpose of evaluation**.
  - For technical investigation ➡ correctness of equivalent
    - Example: Chemistry; electrode material for lithium-ion battery
- 
- ```
graph LR; A[黒鉛] --> B[black lead]; A --> C([Graphite]);
```
- The diagram illustrates the translation of the Japanese term '黒鉛' (black lead). It shows two possible translations: 'black lead' (which is crossed out with a red 'X') and 'Graphite' (which is enclosed in an oval). This highlights the importance of technical context in machine translation evaluation.
- Grammatical correctness is more important than correctness of equivalent
    - In translation task data in NTCIR-7, equivalent error is not so important

## What methods for machine translation evaluation are available?

- Human judgments
  - Adequacy → evaluation for correctness of equivalent
  - Fluency → evaluation for grammatical correctness
  - Acceptability: evaluation by question (NTCIR-9)
    - evaluation whether evaluators can understand the source sentence meaning through the translated sentence
- Automatic Evaluation Metrics
  - Metrics that do not use external linguistic information and which are rapidly computed
    - BLEU → evaluation for correctness of equivalent
    - RIBES → evaluation for correctness of word sequence based on words
    - IMPACT → evaluation for correctness of word sequence based on chunks

## Problems of current human judgments

- Adequacy → evaluation criterion is not clear
- Fluency → evaluators can understand the source sentence through the translated sentence
  - Source sentence: 今日は晴れです。(It is fine today.)
  - Translated sentence: Hello! → Adequacy = 1
  - Fluency = 5

↓

Relative evaluation
- Acceptability: evaluation by question
  - Evaluation for low or middle quality translated sentence is insufficient
  - Evaluation for each item is insufficient

↓

It is impossible to use it to analyze error of MT systems

## Ideal human judgments

- Realization of evaluation on an absolute scale (evaluation for attainment level)
- Improvement of evaluation for low-quality or middle-quality translated sentences
- Realization of evaluation that can be used to analyze error of MT systems

## Problems of current automatic evaluation

- Level of achievement is not clear (*i.e.*, meaning of the score is not clear)
- It cannot be used to analyze error of MT systems

## Ideal automatic evaluation

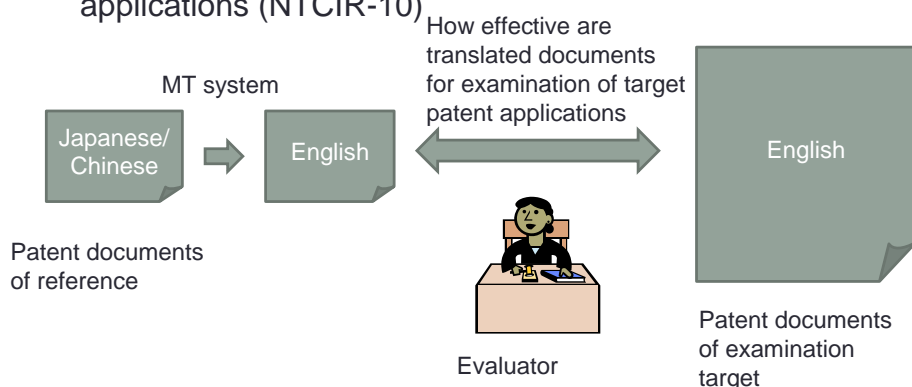
- Low-cost metric (*i.e.*, the number of references is 0 or 1, ideally)
- High-speed metric
- Usable metric for SMT tuning
- **Meaningful metric**
- Metric that has high correlation with human judgments

## Challenge of New Human Judgments

- Toward evaluation on an absolute scale
  - Focusing on an evaluation method for English teaching
  - Use of Common European Framework of Reference (CEFR): CEFR is evaluation on an absolute scale
    - Step 1 - Criteria Setting
    - Step 2 - Item Setting
    - Step 3 - Standard Setting
    - Step 4 - Agreement/Consistency
    - Step 5 - Validation
    - Step 6 - Test Set

## Challenge of New Human Judgments

- Toward practical evaluation for examination of patent applications
  - Concept of evaluation for examination of patent applications (NTCIR-10)



## Challenge of New Human Judgments

- Toward analysis of error of MT systems
  - Question-based evaluation using test sentences
  - Test sets mean sentence examples of questions for grammatical items
  - Evaluator answers 'Yes' or 'No'

| Item       | Source sentence                                 | Translated sentence (reference) | Question                                                 |
|------------|-------------------------------------------------|---------------------------------|----------------------------------------------------------|
| comparison | 彼は君より高い<br>(He is taller than you.)             | 他比你高                            | 比較文に「比」を使っていますか<br>(Is “比” used in comparison sentence?) |
| comparison | これはあれより大きい<br>(This is larger than that.)       | 这个比那个大                          | 比較文に「比」を使っていますか                                          |
| comparison | これはあれより大きくない<br>(This is not larger than that.) | 这个没有那个大                         | 比較文の否定は「没有」になっていますか                                      |
| comparison | 彼女と同じくらい綺麗だ<br>(You are as beautiful as her.)   | 跟她一样漂亮                          | 「同じくらい」が「跟...一样」になっていますか                                 |

## Challenge of New Human Judgments

- Advantage of using test sentences
  - It enables feedback to MT systems
  - Realization of automatic judgments
  - High correlation with human judgments

## Problems of New Human Judgments

- Evaluation on an absolute scale
  - Domain, number of sentences, length of sentence, etc.
- Analysis of error of MT systems
  - Deal with grammatical items that are special representations between two languages
    - Example: 「渋滞が自然解消する」 (Traffic jams solve itself.)
    - It is difficult to translate “自然(naturally)” into “by itself” between Japanese and English
    - “自然解消” is used in both Japanese and Chinese
  - Modification of questions fit for the MT system
  - Construction of efficient test sets
  - Avoidance of tuning of systems to the test sets



## Activity to Solve Problems of New Human Judgments

- Analysis of MT system error
  - Construction of test sets for patent
  - In that case, it is effective to construct test sets each filed (*e.g.*, electronic, mechanics, chemistry etc.) or each structure of patent documents (*e.g.*, claim, example of working, etc.).
  - Use of patent family to reduce construction costs

## Important point for Realization of New Automatic Evaluation

- Use of language resource
  - Tunable metric → BLUE, NIST, WER, PER, TER etc.
  - Tunable and higher-quality metric → METEOR, TER-plus, MaxSim, TESLA, AMBER, MTeRater, etc.
  - Highest-quality metric → RTE, DCU-LFG, MEANT, etc.

## Toward Realization of New Automatic Evaluation

- Metric without language resource
  - Parameters that are effective to address various languages
- Metric using language resources
  - Realization of high-quality automatic evaluation that depends on a specific language pair

## Toward Realization of New Automatic Evaluation

- Analysis of error of MT systems
  - Indication of calculation processes of scores

|                 |                                                                    | human | IMPACT with lemma |
|-----------------|--------------------------------------------------------------------|-------|-------------------|
| source sentence | これらのガスは、所定の割合で混合して用いてもよい。                                          | 4     | 0.3917            |
| system          | you may use these gases mixing it by the given percentage.         |       |                   |
| reference       | these gases might be used by mixing at a predetermined percentage. |       |                   |

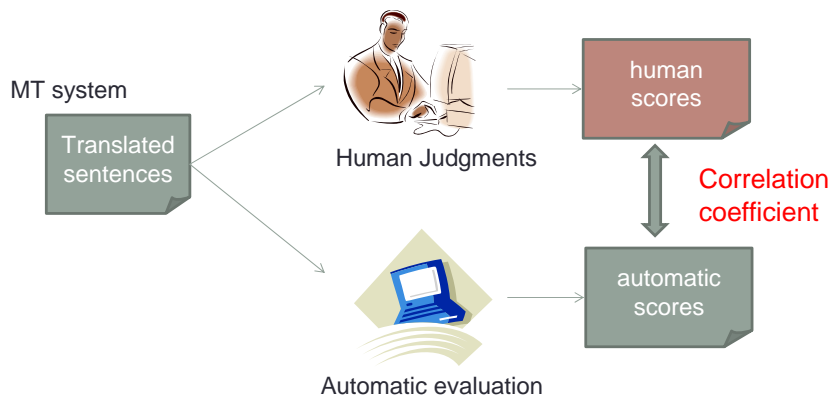
Chunks determined using IMPACT with lemma

Source: you may [[use]] [these gases] [mixing] it [[by]] the given [percentage.]

Ref. : [these gases] might be [[used]] [[by]] [mixing] at a predetermined [percentage.]

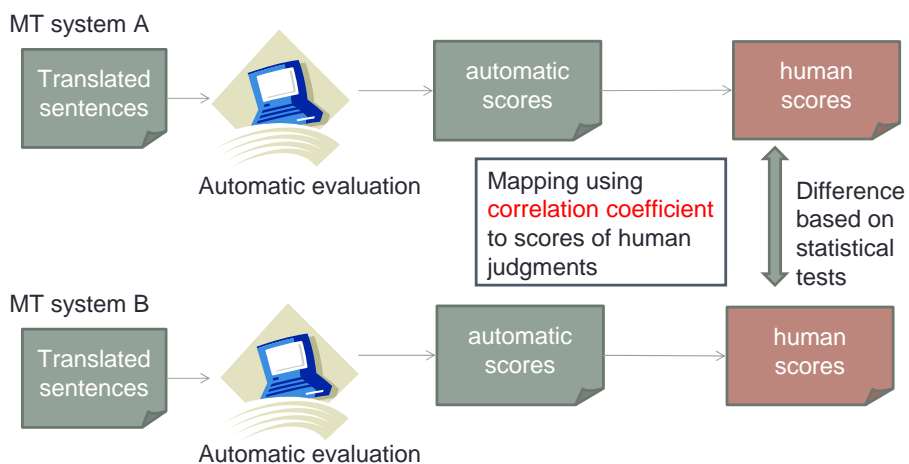
## New Evaluation using Human Judgments and Automatic Evaluation

➤ Meta-evaluation phase ➡ correlation coefficient



## New Evaluation using Human Judgments and Automatic Evaluation

➤ Evaluation phase ➡ MT system comparison



Thank you for your attention

## Session 1

依存関係を用いた特許分野のための日英中対訳  
フレーズの切り出しアルゴリズム

## 依存関係を用いた特許分野のための日英中対訳フレーズの切り出しアルゴリズム

池田秀人  
Ze Zhong Li

立命館大学 情報理工学部

Nguyen Thanh Hung  
Chong Zheng Zhong

抄録—この論文は、対訳フレーズに基づく特許文書の機械翻訳システムの基本になる対訳フレーズの自動抽出の方法を提案する。フレーズは、特許分野従属依存関係に基づく対訳フレーズである。ここで対訳フレーズと呼ぶのは、語と節の中間構造としての句だけでなく、語自身も文型も含んでいる。このアルゴリズムでは、すべての対訳フレーズは抜き出せないが、84%の対訳フレーズは正確に抽出でき、手作業による対訳辞書開発の工数を大幅に減少させることができる。

キーワード：分野従属従属関係に基づくフレーズ抽出、機械翻訳、事例ベース翻訳、日英フレーズ対応アルゴリズム、対訳フレーズ辞書開発

## I. はじめに

句対応問題(Phrase Alignment Problem)は、機械翻訳の品質を向上させる上、重要な位置を占める。これまでいろいろな句対応アルゴリズムが提案されているが、それらは2つに分類できる。1つは、2段階抽出を行うもので、初めに語対応(word alignment)を GIZA++などで自動的に行っておき、その対応を使ってその上位構造である句の対応を人間の手で実現しようというもので、[Koehn, 2003; Chiang, 2007]などの論文に見られる。他の1つは、[Marcu and Won] が最初に提案したもので、統計的類似性に基づき句対応を行うものである。DeNero (2008) は、フレーズベースモデルに基づき句対応を行っている。しかし、これらの結果は、かなりの偽フレーズ対応が発生し、結局最終的には人手によって修正するとしている。重要なことは、最終的に人手で見直すにしても、その修正結果が、自動切り出しアルゴリズムの修正につながり、精度が向上するかどうかである。この性質を、「成長性」と呼ぶことにする。

フレーズと呼ばれるものには、いくつかの種類がある。文中の意味を持つ連続文字列である「線形フレーズ(LP: Linear Phrase)」、構文木の間節点に対応している「構文木階層フレーズ(SHP: Syntax-based Hierarchical Phrase)」、係り受け関係(依存関係)階層木の間節点に対応する「依存関係階層フレーズ(Dependency-based Hierarchical Phrase)」がその代表的なものである。更に、機械翻訳を目的として考える場合、これらのフレーズを言語間で対応させた「対訳フレーズ(ParaPhrase)」が重要となる。言語ごとに抽出したフレーズ間には、部分的にしか対応がなく、対応させる相手を見ながら、フレーズ

の再構成を行わなければ、言語間で完全な対応は見いだせない。その場合、フレーズの連続性はむしろ邪魔で、DHP を基本にした対訳フレーズが有用である。これを「依存関係対訳フレーズ(DHPP: Dependency-based Hierarchical ParaPhrase)」と呼ぶことにする。

この論文では、DHPP を更に進化させた、分野従属 DHPP(DDHPP: Domain-specific DHPP)」を提案し、それを特許翻訳に適用した例を紹介する。更に、そのアルゴリズムを使って構築した DDHPP 対訳フレーズ辞書とその応用としての機械翻訳システム、および翻訳品質の保証された文作成支援システムを紹介する。DDHPP 対訳フレーズ辞書のサイズは、1000 万フレーズに及ぶ。問題は、この巨大なデータベースをどのように構築するかである。このデータベース構築のために NCIR-10[15]で提供された日英対訳特許コーパスを使用した。

## II. 各種のフレーズ

## A. 線形フレーズ(LP)及び構文木階層フレーズ(SHP)

フレーズにはいろいろな定義がある。Oxford Dictionary は、フレーズを「接続関係に基づく最少語グループで節の構成要素となるもの。"A small group of words standing together as a connective unit, typically forming a component of a clause"」としている。すなわち、文の文法的構成要素で接続(connection)によって文を構成することができるのもというわけである。文をその要素の接続で構成されているというモデルは、1957年 Teniere [17] によって提案されている。このモデルに基づく、文は木構造で表すことができる。例えば、次の日本語文：

JS0="なお、上述の各実施形態においては、コイルを備えたモータを駆動するための回路を例に説明したが、この発明はこれに限定されるものではない。"

を考えてみる。これを最少線形フレーズに切ると、

JS0="なお、/上述の/各実施形態においては、/コイルを/備えた/モータを/駆動するための/回路を/例に/説明したが、/この/発明は/これに/限定されるものではない。"

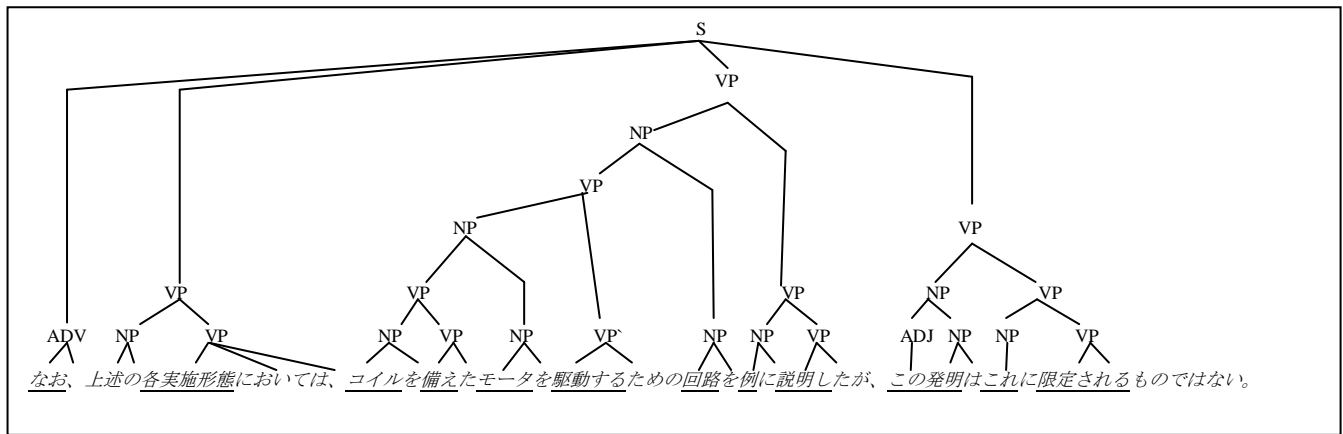


図1 線形フレーズに基づく階層構造

となる。ここで最少線形フレーズと呼んだのは、「上述の各実施形態においては、」も文中の意味を持つ連続語列であるから線形フレーズであり、最少フレーズ「上述の」と「各実施形態においては、」の2つのフレーズを接続したものであることと、この2つのフレーズは、それぞれ1つだけ内容語（「上述の」は形容詞、「実施形態」は名詞）を含んでおり、これ以上小さなフレーズに分解できないことによる。

この文は、図1のように階層的に構造化することができる。この構造に基づくと、文NS0は、次のようにフレーズ関数の列で表現できる。

JS0=なお、\_においては、@v連用タ接続:たが、\_ものではない。(JN1],[JS2],[JS3])  
 JN1=上述の\_(JN4);  
 JS2=\_を例に説明する(JN5);  
 JS3=\_はこれに限定される(JN6);  
 JN4=各\_(JN7);  
 JN5=@v基本形ための\_(JP8],[JN9);  
 JN6=この\_(JN10);  
 JN7=実施形態  
 JP8=\_を稼働する(JN11);  
 JN9=回路  
 JN9=@v連用タ接続:た\_(JP11],[JN12);  
 JN10=発明  
 JP11=\_を備える(JN13);  
 JN12=モータ  
 JN13=コイル

図2：文の線形フレーズ関数分解

ここで、JS2の「例に」というフレーズは、「例」を名詞と考えれば1つの独立フレーズであるが、「例に説明する」を複合動詞として扱ったフレーズ分解となっている。

### B. 依存関係階層フレーズ(DHP)

これに対し、依存関係階層フレーズは、語間の係り受け関係を使った階層構造(図1)を使って切り出した階層構造である。例文の、依存関係フレーズは、つぎのようになる。

JS0=なお、@v連用形:たがこの\_は@v未然レル接属:れるものではない。(JN1),(JN2],[JP3])  
 JP1=上述の各\_においては、\_を例に説明する(JN4],[JN5])  
 JN2=発明  
 JP3=これに限定する()  
 JN4=実施形態  
 JN5=@v基本形:ための回路(JP6)  
 JP6=\_を駆動する(JN7)  
 JN7=@v連用形:たモータ(JP8)  
 JP8=\_を備える(JN9)  
 JN9=コイル

図2 文の依存関係フレーズ分解

文のフレーズ分解としては、SHPとDHPは、大差ないように見えるが、係り受け関係を使うDHPでは、係り語とその係り先の語が離れた場合もその関係を維持してくれるという長所を持っている。例えば、「上述の各実施形態においては、」という副詞節は、「説明する」に係っているが、線形フレーズとしては、隣接していないため、この関係を認識できておらず、図2のJS0のような文関数が切り出されているが、DHPでは、この関係が認識されているため、図2のJP1のような文関数として表現されている。

### C. 対訳フレーズ(DHPP)

上述のDHP分解は、別の言語で行うことも可能である。ここでは、日本語の例文JS0の翻訳文の例をES0としてあげる。

ES0="While the above embodiments are described as examples in which the circuits are used for driving the motor with coils, this invention is not limited to those examples."

これを DHP 分解したものを図 3 で示す。

|                                                           |
|-----------------------------------------------------------|
| ES0=While the above __, this __.([EN1],[EP2],[EN3],[ES4]) |
| EN1=embodiments                                           |
| EP2= are described as _(EN5)                              |
| EN3=invention                                             |
| ES4= is not limit to those _(EN10)                        |
| EN5=examples in which the __([EN6],[EP7])                 |
| EN6=circuits                                              |
| EP7=are used for driving the _ with _([EN8],[EN9])        |
| EN8=motor                                                 |
| EN9=coils                                                 |
| EN10=examples                                             |

図 3 英文の DHP 分解の例

英文の分解としては、かなり自然な分解がなされていると考えられるが、前出の日本語文の DHP 分解の関数との関連を見ると、かなりの食い違いが見て取れる。例えば、文レベルの関数 JS0 と ES0 とでは、関数の数ばかりでなく、英文では、「example」という語が 2 回使われているのに対し、日本語では、「例」という語は 1 回しか使われていないこと、英文では、節「ES4」を引数にしているが、日本文では、節引数はなく、句引数のみであること、また引数の数も同じではない。

これを、何らかの関数の再構成をして、英文のフレーズ分解と日本文のフレーズ分解が、意味的にも、構文的にも 1 対 1 対応がつくようにしたものが、対訳フレーズである。実際、以下のようにフレーズを再構成すれば、それは可能である。

表 1. 日英対訳フレーズ(DHPP)の例

|                                                                   |                                                                                                                         |
|-------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| JS0=なお、上述の各_においては、_を例に説明したがこの_はこれに限定されるものではない。([JN1],[JN2],[JN3]) | ES0=While the above _ described as examples in which the __, this _ is not limit to those examples. ([EN1],[EN2],[EN3]) |
| JN1=実施形態                                                          | EN1=embodiments                                                                                                         |
| JN2=_を備えたモータを駆動するための回路([JN4])                                     | EN2=circuits are used for driving the motor with _([EN4])                                                               |
| JN3=発明                                                            | EN3=invention                                                                                                           |
| JN4=コイル                                                           | EN4=coils                                                                                                               |

ここで、疑問になるのは、すべての翻訳対にたいして、いつでもこんなうまく言語間でフレーズ対応がつけられるかという問題である。日本語に多い、省略語の問題や、対応する語はないため、説明的に翻訳されている例や、長い文を複数の文に分割して翻訳した例や、文脈依存表現などが、その候補であるが、たとえそのような文対に対してでも、対訳フレーズは構築できる。ただ、その場合、単語は句や文に対応させ、分解しないで、全体として対応させるという工夫だけである。特許のような、

部分を含めて一致さなければいけない場合は、うまく対応させられる。

ここで、注目してほしいのは、JS0 と ES0 である。本論文では、このようなものも、「フレーズ」と呼ぶ。

(あえて言えば、「文レベルのフレーズ」である。) また、単語もフレーズの一つと考える。このようにすべてを「フレーズ」として捉えることで、異なる言語の文の間の「フレーズ」対応が可能となるのである。

#### D. 分野従属 DHPP(DDHPP)

最後に、DHPP を更に変形する。それは、分野の専門用語を、(引数でなく) 関数名の部分として明示するものである。前出の文レベルのフレーズをもう一度見てみよう。この文に、引数としての、「実施形態(embodiments)」と「発明(invention)」を、埋め込んで、

JS0="なお、上述の各実施形態においては、\_を例に説明したがこの発明はこれに限定されるものではない。([JN2])"

ES0="While the above embodiments described as examples in which the \_\_, this invention is not limit to those examples.([EN2])"

とすると、これは、特許文書で頻出する文型である。この文型のキーワードは、「実施形態(embodiments)」や「発明(invention)」であり、この部分になにか別の語を埋め込んで使うことは、あまりない。このようなフレーズは、1 つのフレーズとして記憶しておくことが、翻訳品質を向上させるばかりでなく、翻訳の標準化にも有効である。実際、JS0 の文に対する英訳は 1 つではなく、多くの異なる文型が存在する。

ここで、我々はやっと思つて理想とすべきフレーズに行き着いた。今後の問題は、このフレーズ対をどのように見つけ出すかである。我々の簡単な試算によると、このような対訳フレーズの数は、1,000 万件を超える。しかし、このうち 90% は名詞フレーズであり、既に商用機械翻訳システムのいくつかは、それを集めている。動詞フレーズも切り出すのはそんなに難しくない。問題は、「文レベルのフレーズ」である。これは、100 万件程度と考えている。この文レベルのフレーズの知識が、翻訳家の「ノウハウ」となっていると考えている。機械翻訳が翻訳専門家の品質に近づくためには、この「文レベル」のフレーズのコレクションがその鍵になると、筆者は考えている。

この問題を解決するために、どのように文レベルのフレーズ対も含めてフレーズ対を、自動的に抽出するかについて、次の節で述べる。



### III. 対訳フレーズの自動抽出

#### A. 日本語文のDHP抽出

ここで、例として使うのは、前出のJS0である。DHPを切り出すために日本語係り受け解析器 CaBoChaを使った。その手順は以下の通りである。

##### 1) 日本語係り受け解析

上記の文に日本語係り受け解析を行うと、図4のような結果が得られる。

##### 2) 全フレーズ品詞パターン抽出

全フレーズの品詞パターンを抽出する。この時、一緒に具体的なフレーズも抽出しておく。

##### 3) フレーズ処理パターンの付加

フレーズパターンをどのように処理するかを示すフレーズ処理パターンを付加する。例えば、上述の第3フレーズの「各実施形態においては、」というフレーズは、

JF2=各\_においては、([JN18]);

JN18=実施形態;

I19=、;

という2つの関数に分解すべきであるため、処理パターンとしては、「F1:N1:F1:I1」というパターンが付加される。これは、「実施形態」を名詞フレーズとして抜き出して処理せよという意味になる。1つのフレーズが与えられた時、この処理パターンを決める方法があるかという問題が発生するが、フレーズの品詞情報だけを使ったのでは、完全ではないことがわかっている。例えば、次の2つのフレーズ:「手で」と「手で」はいずれも

名詞-一般:助詞-格助詞-一般

という品詞構造を持つが、「手で」は、「N1:F1」という処理パターンになり、

JP=\_で([JN]); JN=手

という2つの関数に分解されるのに対し、「手で」は、

JP=手で();

と1つの関数になる。いずれも副詞句であるが、それは、その英語対訳が「手で」は「by hand」と2語であり、「手で」は「manually」と1語になることから、対訳フレーズを構築するときの都合によるものである。このようなきめ細かい処理を行うためには、「手」も「手動」も「名詞-一般」とする解析では不十分で、そのため、「手動」という特別な名詞(この場合は、手段・方法を表す名詞で、「動かす」という動詞の派生語の一種)は、「名詞-一般(手動で)」と具体的な値を書いてパターン化した。一般に機能語は、それぞれの役割が異なる場合が多いため、すべての機能語には、具体的な値の付加してパターン化した。

実際、フレーズの品詞列パターンは、NTCIR-10の300万件の文対のうちの387,500件の特許文のフレーズの品詞列パターンは、916,000件もある。この1つ1つの処理パターンを手で割り付けるのは大変な作業であるが、品詞だけを使ったパターンでパターン化すると、97,832件のパターンが出てくる。これを代表的なパターンの例から処理パターンを作成したら、96.2%は正しく分割されることが分かった。残りの4.8%(具体的なフレーズ件数としてはそれでも約44,000件)は、その規則以外の分割となった。こうして作成した処理パターンのうち、例文で使われているものが表1に上げてある。

##### 4) 依存関係を使ったフレーズの再統合

|             |                  |                                  |   |
|-------------|------------------|----------------------------------|---|
| *0 15D 0/0  | なお               | 接続詞                              | F |
| 、           | 記号-読点            |                                  | I |
| *1 2D 0/1   | 上述の              | 名詞-サ変接続:助詞-連体化                   | F |
| *2 10D 2/4  | 各                | 接頭詞-名詞接続                         | F |
| 実施形態        | 名詞-サ変接続:名詞-一般    |                                  | N |
| においては       | 助詞-格助詞-連語:助詞-係助詞 |                                  | F |
| 、           | 記号-読点            |                                  | I |
| *3 4D 0/1   | コイル              | コイル 名詞-一般                        | N |
| を           | 助詞-格助詞-一般        |                                  | F |
| *4 5D 0/1   | 備える              | 動詞-自立                            | V |
| @v連用形:た     | 助動詞              |                                  | P |
| *5 6D 0/1   | モータ              | 名詞-一般                            | N |
| を           | 助詞-格助詞-一般        |                                  | F |
| *6 7D 1/1   | 駆動する             | 名詞-サ変接続:動詞-自立                    | V |
| *7 8D 0/1   | ための              | 名詞-非自立-副詞可能:助詞-連体化               | F |
| *8 10D 0/1  | 回路               | 名詞-一般                            | N |
| を           | 助詞-格助詞-一般        |                                  | F |
| *9 10D 0/1  | 例に               | 名詞-一般:助詞-格助詞-一般                  | F |
| *10 15D 1/3 | 説明する             | 名詞-サ変接続:動詞-自立                    | V |
| @v連用形:たが    | 助動詞:助詞-接続助詞      |                                  | F |
| 、           | 記号-読点            |                                  | I |
| *11 12D 0/0 | この               | 連体詞                              | F |
| *12 15D 0/1 | 発明               | 名詞-サ変接続                          | N |
| は           | 助詞-係助詞           |                                  | F |
| *13 14D 0/1 | これに              | 名詞-代名詞-一般:助詞-格助詞-一般              | F |
| *14 15D 1/2 | 限定する             | 名詞-サ変接続:動詞-自立                    | V |
| @v未然レル接属:れる | 未然レル接属動詞-接尾      |                                  | P |
| *15 -10 0/3 | ものではない           | 名詞-非自立-一般:助動詞/連用形:助詞-係助詞:助動詞/基本形 | P |
| 。           | 記号-句点            |                                  | I |
| EOS         |                  |                                  |   |

図4 日本語係り受け解析の結果

つぎに、依存関係を使ってフレーズを再統合する。このアルゴリズムはつぎのとおりである。

#### 4-1) 対象フレーズの発見と関数化

係り受け解析の結果は、対象文の中心フレーズを指示してくれる（フレーズヘッダーの第2項の「-1D」がその識別子）。例文の中心フレーズは以下のものである。これを対象フレーズとすると、その処理パターンは、V:P:I であるから、次の関数ができる。

*P15=ものではない。*

4-2) 対象フレーズに直接係っているフレーズの抽出例の場合は、#0, #10, #12 および #14 の4つのフレーズが抽出される。これは、対象フレーズ識別番号（例では「15D」）を第2項に持つフレーズである。

4-3) その各フレーズに対して、後出のものから関数化を行う。例では、まず、最後のフレーズ#14 に対して、このフレーズの処理パターンは、「V:P」であるから、次の関数が見つかる。

*P14=限定する()*

*P15=@v 未然レル接属:れるものではない。([P14])*

ここで、「活用語」に続く F タイプの関数には、前出の語の活用型（ここでは、「未然レル接属」を明示して関数を作成している。更に、依存関係から、F タイプ関数は、それが係るフレーズの中心関数に接続させる。こうして、4-1) で既にできていた関数 P5 は、上記の P5 に置き換えられる。

同様に、フレーズ#12 からは、

*N12=発明*

*P15=\_は@v 未然レル接属:れるものではない。([N12],[P14])*

フレーズ #10 からは

*P10=説明する()*

*P15=@v 連用形:たが\_は@v 未然レル接属:れるものではない。([N10],[N12],[P14])*

フレーズ#0 からは、

*P15=なお、@v 連用形:たが\_は@v 未然レル接属:れるものではない。([N10],[N12],[P14])*

が作成される。

4-3) つぎに未処理のフレーズのうち最後尾のものから、4-2)と同様な処理を繰り返し、未処理フレーズがなくなるまで繰り返す。

例文では、次の対象フレーズは、#14 となり、それに係るフレーズ #13 に対して、

表2 フレーズの品詞パターンと処理パターン

| フレーズの品詞パターン                                                | 処理パターン      |
|------------------------------------------------------------|-------------|
| 接続詞:記号-読点                                                  | F1:I1       |
| 名詞-サ変接続(上述):助詞-連体化(の)                                      | F2          |
| 接頭詞-名詞接続:名詞-サ変接続:名詞-一般<br>:助詞-格助詞-連語(において):助詞-係助詞(は):記号-読点 | F1:N2:F2:I1 |
| 名詞-一般:助詞-格助詞-一般(を)                                         | N1:F1       |
| 名詞-一般(例):助詞-格助詞-一般(に)                                      | F2          |
| 動詞-自立/連用形:助動詞(た)                                           | V1:P1       |
| 名詞-サ変接続:動詞-自立/基本形(する)                                      | V2          |
| 名詞-サ変接続:動詞-自立/連用形(する):助動詞(た)<br>:助詞-接続助詞(が):記号-読点          | V2:F2:I1    |
| 名詞-非自立-副詞可能(ため):助詞-連体化(の)<br>連体詞(この)                       | F2          |
| 名詞-サ変接続:助詞-係助詞(は)                                          | F1          |
| 名詞-代名詞-一般(これ):助詞-格助詞-一般(に)                                 | N1:F1       |
| 名詞-サ変接続:動詞-自立(する):未然レル接属動詞-接尾(が)                           | F2          |
| 名詞-非自立-一般(もの):助動詞/連用形(だ):助詞-係助詞(は)<br>:助動詞/基本形(ない):記号-句点   | V2:P1       |
|                                                            | P4:I1       |

*P14=これに限定する()*

更にそれに続く対象フレーズ #12 に係るフレーズ#11 に対して、

*P15=なお、@v 連用形:たがこの\_は@v 未然レル接属:れるものではない。([N10],[N12],[P14])*

を得る。対象フレーズ #10 に係るフレーズは、#9, #8 および#2 で、ここから、次の関数を得る。

*N8=*

*N2=実施形態*

*P10=各\_においては、\_を例に説明する([N2],[N8])*

対象フレーズ #8, we have a modifier #7 and

*N7=ための\_([N8])*

対象フレーズ #7 に係るフレーズ #6 から、

*P6=駆動する()*

*N8=@v 基本形:ための回路([P6])*

対象フレーズ #6 に係るフレーズ#5 から、

*N5=モータ*

*P6=\_を駆動する([N5])*

対象フレーズに係るフレーズ #4 から

*N5=@v 連用形:たモータ([P4])*

対象フレーズ #4 に係るフレーズ #3 から

*N3=コイル*

*P4=\_を備える([N3])*

対象フレーズ #2 に係るフレーズ #1 から

*P10=上述の各\_においては、\_を例に説明する([N2],[N8])*

を得る。こうして、次の関数列を得ることができる。

JN3=コイル  
 JN2=実施形態  
 JP4=\_を備える(JN3)  
 JN5=@v 連用形:たモータ(JP4)  
 JP6=\_を駆動する(JN5)  
 JN8=@v 基本形:ための回路(JP6)  
 JP10=上述の各\_においては、\_を例に説明する(JN2, JN8)  
 JN12=発明  
 JP14=これに限定する()  
 JP15=なお、@v 連用形:たが本\_は@v 未然レル接属:れるものではない。  
 (JN10),(JN12),(JP14)

5) フレーズ関数と内容語のクロス表

最後に、抽出され合成された関数と、文中の内容語(名詞、動詞、副詞、形容詞)とのクロス表を作成する。例文では、「なお、上述、説明する、例、実施形態、コイル、モータ、駆動する、回路、備える、発明、限定する」の12語が内容語である。これらの内容語が出現する関数に対しては「1」で、出現しなければ空白で示したのが、表2である。

表2 日本語フレーズ関数と内容語とのクロス表

|      | JP1<br>5 | JP1<br>0 | JN<br>2 | JN<br>3 | JN<br>4 | JN<br>5 | JN<br>6 | JN<br>8 | JN<br>12 | JN<br>14 |
|------|----------|----------|---------|---------|---------|---------|---------|---------|----------|----------|
| なお   | 1        |          |         |         |         |         |         |         |          |          |
| 上述   |          | 1        |         |         |         |         |         |         |          |          |
| 説明する |          | 1        |         |         |         |         |         |         |          |          |
| 例    |          | 1        |         |         |         |         |         |         |          |          |
| 実施形態 |          |          | 1       |         |         |         |         |         |          |          |
| コイル  |          |          |         | 1       |         |         |         |         |          |          |
| モータ  |          |          |         |         |         | 1       |         |         |          |          |
| 駆動する |          |          |         |         |         |         | 1       |         |          |          |
| 回路   |          |          |         |         |         |         |         | 1       |          |          |
| 備える  |          |          |         |         | 1       |         |         |         |          |          |
| 発明   |          |          |         |         |         |         |         |         | 1        |          |
| 限定する |          |          |         |         |         |         |         |         |          | 1        |

B. 英文の対訳フレーズの抽出

英文の対訳フレーズに抽出も、日本語文と同様に次に手順で行う。例として使うのは、上述のES0である。

1) 英文の構文・従属解析

Standord Parserは、構文解析と係り受け解析を同時に行ってくれる。ただし、構文解析の結果は、線形フレーズ分解ではなくて、構文木として出力され、係り受け解析は、フレーズ間の係り受け関係ではなくて、語の従属関係が、その関係の種類も含めて出力される。その結果は、図文に日本語係り受け解析を行うと、図5に示す。図5の上部は、構文構造で、下部は従属解析の結果である。

2) 最小フレーズと、その依存関係の切り出し

図5の構文木を、関数型に書き換えると、図6のようになる。各関数の型は、構文木の節点の品詞の種類から作成した。名詞句(NP)は、「N」、動詞句(VP)は、「P」、節(SBAR)または文(S)は、「S」である。それ

以外のもの(PP など)は、それを含むフレーズに合体した。

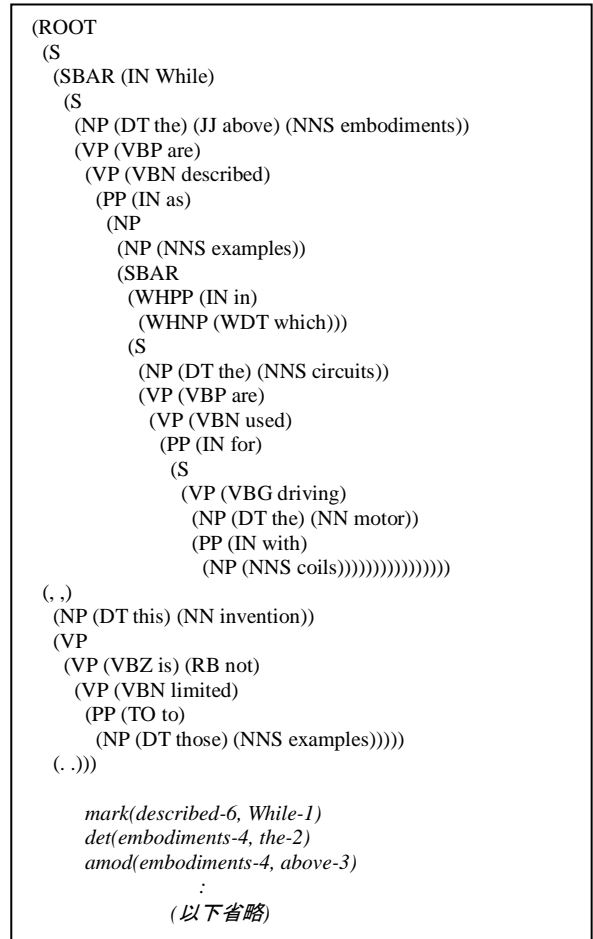


図5 英文の構文・従属解析の結果

3) 従属関係に基づくフレーズ関数の再構成

図5の従属関係を使って、フレーズ関数を再構成する。その場合、内容語を含まないフレーズは、そのフレーズが係っているフレーズに埋め込むことによってなくすことが、処理の基本となる。こうしてできたのが、前述の図3である。

これは、(関数の番号は異なるが、)図3で示したDHPである。従って、DHPの作成法を示したことになる。

4) 英語フレーズと内容語のクロス表の作成

日本語の場合と同様に、内容語と上記DHPのクロス表を作成すると、表3のようになる。

C. 日英DDPの対応と対訳フレーズ

A.およびB.で作成したフレーズ列に対して、どのように対応させられるかを示そう。それは、次の手順で行う。

表3 英語フレーズ関数と内容語のクロス表

|             | ES0 | EN5 | EP4 | EN7 | EN9 | EP10 | EN13 | EN14 | EN16 | ES2 | EN20 |
|-------------|-----|-----|-----|-----|-----|------|------|------|------|-----|------|
| While       | 1   |     |     |     |     |      |      |      |      |     |      |
| Describe    |     |     | 1   |     |     |      |      |      |      |     |      |
| Example     |     |     |     | 1   |     |      |      |      |      |     | 1    |
| Embodiments |     | 1   |     |     |     |      |      |      |      |     |      |
| Coils       |     |     |     | 1   |     |      |      | 1    |      |     |      |
| Motor       |     |     |     |     |     | 1    |      |      |      |     |      |
| Driving     |     |     |     |     |     |      |      |      |      |     |      |
| Circuit     |     |     |     |     | 1   |      |      |      |      |     |      |
| limited     |     |     |     |     |     |      |      |      | 1    |     |      |

1) 出現内容語の対応

出現内容語（ここでは、日本語12語、英語11語）の関係をまず調べる。そのために、いろいろな用語辞書からあらかじめ言語間用語関連表を作成しておく。実際本システムでは、NiCT-EDRの英日対訳辞書、英語単語辞書、日英対訳辞書、専門用語辞典、およびJapio機械翻訳辞書を使ってこれを作成した。1語の訳語は100以上あるものも少なくないが、これを12×9の行列に制約すると、その候補は極めて限定され、どの訳語に対応しているかを見つけ出すことができる。こうして作成したのが、表4の内容語同士のクロス表である。このクロス表にセルの数値は、例えば「Motor」がカタカナの「モータ」に訳される可能性が60%位あることを示している。この数値が50%以上あると、ほぼ正解で、20%以下であると、訳語として不適当であることが多い。上の例では、「備える」に対応する英単語はない可能性が高い。

2) 日英DHP同士の対応

この内容語の対応関係（表4）を使って、DHPフレーズの対応を作成する。その場合、日本語フレーズ関数と内容語のクロス表（表2）英語フレーズ関数と内容語のクロス表（表3）も一緒に使うと、次のような対応表（表5）を作ることができる。

3) DHPのグループ化

表5でわかるように、日英のDHPフレーズは1対1には対応していない。これを1対1に対応させるため、DHPフレーズのグループ化を行う。表6は、そのグループ化を行った結果である。

この表を作るには、まず表5を見る。例えば、JN15、JN10は、ES0に関係しており、JP10は、ES0、ES4、ES20に関係している。従って、日本語フレーズを{JN15,JN10}をグループ化し、英語フレーズを{ES0,ES4,ES20}をグループ化すれば2つのグループは対応することになる。この操作は行列の行と列の順序を入れ替えて、直交化することに対応している。しか

表4 内容語同士のクロス表

|             | なお | 上述 | 例 | 説明する | 実施形態 | コイル | 備える | モータ | 駆動する | 回路 | 発明 | 限定 |
|-------------|----|----|---|------|------|-----|-----|-----|------|----|----|----|
| While       | 5  |    |   |      |      |     |     |     |      |    |    |    |
| Describe    |    | 4  |   | 7    |      |     | 1   |     |      |    |    |    |
| Example     |    |    | 9 |      |      |     |     |     |      |    |    |    |
| Embodiments |    |    |   |      | 8    |     |     |     |      |    |    |    |
| Coils       |    |    |   |      |      | 9   |     |     |      |    |    |    |
| Motor       |    |    |   |      |      |     |     | 9   |      |    |    |    |
| Driving     |    |    |   |      |      |     |     |     | 6    |    |    |    |
| Circuit     |    |    |   |      |      |     |     |     |      | 9  |    |    |
| used        |    |    |   |      |      |     | 1   |     |      |    |    |    |
| invention   |    |    |   |      |      |     |     |     |      |    | 9  |    |
| limited     |    |    |   |      |      |     |     |     |      |    |    | 6  |

表5 日英DHP同士の対応表

|      | ES0 | EN5 | EP4 | EN7 | EN9 | EP10 | EN13 | EN14 | EN16 | ES2 | EN20 |
|------|-----|-----|-----|-----|-----|------|------|------|------|-----|------|
| JP15 | 1   |     |     |     |     |      |      |      |      |     |      |
| JP10 | 1   |     | 1   |     |     |      |      |      |      |     | 1    |
| JN2  |     | 1   |     |     |     |      |      |      |      |     |      |
| JN3  |     |     |     | 1   |     |      |      |      |      |     |      |
| JN4  |     |     |     |     |     |      |      |      |      |     |      |
| JN5  |     |     |     |     |     |      | 1    |      |      |     |      |
| JN6  |     |     |     |     |     |      |      |      |      |     |      |
| JN8  |     |     |     |     | 1   |      |      |      |      |     |      |
| JN12 |     |     |     |     |     |      |      |      | 1    |     |      |
| JN14 |     |     |     |     |     |      |      |      |      | 1   |      |

し、表5をみただけでは、対応するフレーズグループがないフーズがない。この問題を解決するためには、グループを再構成する必要が出てくる。例えば、JP4は、対応する英語フレーズに変更しなければいけない。実際、単独グループJP4をなくすためには、JN5と合体させればいいことが、引数を見るとわかる。EN7も同様に、グループ{{ES0,ES4,ES20}}に含めてしまえばいいことがわかる。更に、(JN8,EN9)および(JN14,ES2)も引数の数が異なる。これも、再グループ化で、対応できる。こうして再グループ化を繰り返してできた対応が、表7である。フレーズを当てはめてみると、表8のようになる。

4) フレーズ合成

最後に、各グループを1つのフレーズに合成すれば、完全に1対1対応のついたフレーズ対、すなわち前述の表1の対訳フレーズが完成する。こうして作成された対訳フレーズを使って、NTCIR-10の300万件の文対から任意に抽出した日英文対3000件に対し、フレーズ対応の精度を計算した。フレーズ対応が正しく行われているかどうかは、人手による方法をとった。その結果が表9である。

表6 日英 DHP フレーズのグループ化

| J-Phrase          | Parameters            | E-Phrase             | Parameters               |
|-------------------|-----------------------|----------------------|--------------------------|
| G1(JP15,J<br>P10) | JN12,JP14,<br>JN2,JN8 | G2(ES0,EN4,<br>EN20) | EN5,EP4,EN16,<br>ES2,EN7 |
| JN2               |                       | EN5                  |                          |
| JN3               |                       | EN9                  |                          |
| JP4               | JN3                   |                      |                          |
| JN5               | JP4                   | EN13                 |                          |
| JN6               | JN5                   | EN12                 | EN13,EN14                |
| JN8               | JP6                   | EN9                  |                          |
| JN12              |                       | EN16                 |                          |
| JN14              |                       | ES2                  | EN20                     |
|                   |                       | EN7                  | EN9,EP10                 |

表7 グループ化によるフレーズ関数の対応

| J-Phrase               | Parameters    | E-Phrase                    | Parameters                |
|------------------------|---------------|-----------------------------|---------------------------|
| JP15,JP10,<br>JN14     | JN12, JN2,JN8 | ES0,EN4,<br>EN20,EN7<br>EN2 | EN5,EP4,EN16,<br>EN9,EN10 |
| JN2                    |               | EN5                         |                           |
| JN3                    |               | EN14                        |                           |
| JN 4, JN 5,<br>JN6,JN8 | JN2           | EN9,EN13,EN1<br>2, EN9      | EN5                       |
| JN12                   |               | EN16                        |                           |

表8 完成したフレーズグループ対応

|                                                                                                                                       |                                                                                                                                                                                                               |
|---------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| JP15=なお、@v 連用形:たがこの_は@v 未<br>然レル接属:れるものではない。<br>([JN10],[JN12],[JP14])<br>JP10=上述の各_においては、_を例に説明<br>する([JN2],[JN8])<br>JP14=これに限定する() | ES0=While the above _ , this _<br>_([EN5],[EP4],[EN16],[ES2])<br>EP4= are described as _([EN7])<br>EN7=examples in which the _<br>_([EN9],[EP10])<br>EN20=example<br>ES2= is not limit to those _<br>([EN20]) |
| JN2=実施形態                                                                                                                              | EN5=embodiments                                                                                                                                                                                               |
| JN3=コイル                                                                                                                               | EN14=coils                                                                                                                                                                                                    |
| JP4= を備える([JN3])<br>JN5=@v 連用形:たモータ([JP4])<br>JP6=_を駆動する([JN5])<br><br>JN8=@v 基本形:ための回路([JP6])                                        | EN13=motor<br>EP12=driving the _ with<br>_([EN13],[EN14])<br>EN9=circuits                                                                                                                                     |
| JN12=発明                                                                                                                               | EN16=invention                                                                                                                                                                                                |

表9 対訳フレーズ抽出の精度

|                                              |                             | 文数          |
|----------------------------------------------|-----------------------------|-------------|
| 正しく対応フレーズが抽出された文                             |                             | 2411(84%)   |
| 正しく<br>対応フ<br>レーズ<br>が作成<br>されな<br>かった<br>もの | Stanford Parser のエラーに起因するもの | 242( 8%)    |
|                                              | Cabocha のエラーに起因するもの         | 93(3%)      |
|                                              | 日本語の関数化に起因するもの              | 452( 15%)   |
|                                              | 英語の関数化に起因するもの               | 43(1%)      |
| 日英フレーズ対応に起因するもの                              |                             | 123(4%)     |
| 対象文全体 (テスト用)                                 |                             | 3,000(100%) |

D) 英中対訳フレーズの抽出

日英コーパスと同様に、中英に関しても同様の処理を行った。英中にはどちらも Stanford Parser があり、処理は日英に比べて簡単であるが、どちらも「SHP フレーズ抽出→語対応→フレーズ対応」の手順は同様である。

I. 評価と結論

この方法で、NCIR-10 で提供された訓練用日英コーパス 300 万件、中英コーパス 100 万件的文対の一部である対訳フレーズを抽出した。この結果から、80%以上の文から自動的に完全フレーズ対応が抽出できることが分かった。同様の基本的には、対訳フレーズ辞書は人手にたよらなければ完成させられないが、この論文で提案した対訳フレーズ抽出アルゴリズムによって、大幅に人力に頼る部分が削減でき、大きな効果を発揮することが分かった。

参考文献

[1] Alagin(Advanced Language Information Forum) . 2009. <http://www.alagin.jp/purpose-e.html>

[2] Barkley Aligner. 2009. *A word alignment software package for machine translation.* <http://code.google.com/p/berkeleyaligner/>

[3] Chiang, Devid. 2007. *Hierarchical Phrase-Based Translation.* Computational Linguistics, Volume 33, Number 2, Association for Computational Linguistics.

[4] Daniel Marcu and Daniel Wong. 2002. A Phrase-based, Joint Probability Model for Statistical Machine Translation. In Proceedings of EMNLP, pp. 133-139, USA.

[5] John DeNero, Alexandre Bouchard Cote, Dan Klein. 2010. Sampling Alignment Structure Under a Bayesian Translation Model. In Proceedings of EMNLP, pages 314-323, USA.

[6] Finch, G. 2000. Linguistic terms and concepts. New York: St. Martin's Press.

[7] Koehn, P., Och, F. J., and Marcu, D. Statistical phrase based translation. In Proceedings of HLT-NAACL, 2003.

[8] Mel'čuk. 2003. *Levels of Dependency in Linguistic Description: Concepts and Problems.* In V. Agel, L. Eichinger, H.-W. Eroms, P. Hellwig, H. J. Herringer, H. Lobin (eds): Dependency and Valency. An International Handbook of Contemporary Research, vol. 1, Berlin - New York, W. de Gruyter, 188-229.

[9] Miller, J. 2011. A critical introduction to syntax. London: continuum.

[10] NiCT(National Institute of Information and Communications Technology). 2010. *Nict-EDR.*

[11]NiCT.“NTCIR-9”, [http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/toc\\_ntcir.html](http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/toc_ntcir.html)

[12]NCTIR-10: <http://www.nii.ac.jp/cscenter/idr/en/index.html>

[13] Oxford University Press: Oxford Dictionary of English. Second edition, revised 2005.

[14] SNLPG (The Stanford Natural Language Processing Group) 2003. *Stanford Parser: A statistical Parser.* <http://nlp.stanford.edu:8080/parser/index.jsp>

[15] Taku Kudoh, Yuji Matsumoto (2000) Japanese Dependency Analysis Based on Support Vector Machines,EMNLP/VLC 2000

[16] Tesniere, Lucien 1957. Elements de syntaxe structural. Paris, Klincksieck.

## Session 2

特許明細書の翻訳者から基本的な誤訳の実例を  
示して対策を提案

## 特許明細書の翻訳者から基本的な誤訳の実例を示して対策を提案

Indication of Example of Basic Erroneous Translation and Proposal for Countermeasure  
from Translator of Patent Specification

吉川 潔 (Kiyoshi Kikkawa)

## ① 概要

新潟の田舎で東京の特許事務所から明細書の原稿を電子メールで受信し、翻訳後に返信という仕事を30年近く行ってきた。特許事務所は、情報処理技術（ICT）を用いた文献調査に関心が多い。私の体験から得た特許関係者の意向を特許情報シンポジウムに反映できたらと願っている。

翻訳中に基本語句とその逸脱語句に出会うとパソコンに保存していたら約2千に達した。12年前から、8社の翻訳ソフトを用いて上記の語句を試訳していた。2社が撤退し、他の2社はOEM。そこで、4社のソフトを試訳対象にしている。

各社の廉価版と高価版の翻訳ソフトを購入し差異も調べた。それらは5年前に発売のモデルであるが、最新版といえどもワード1000に対応するために調整しただけで、実態は同じという情報を得ている。

2年前の第一回特許情報シンポジウムで「翻訳ソフト実用化の提案」を発表し、様々な誤訳の実例を示した。対策として、長文の翻訳は誤訳しやすいので、翻訳対象の単語数を制限し、短文で数量表現に特化した翻訳ソフトの開発を提案した。

## ② 今回（第二回）の発表の目的

コンマ（,）と（and）と（to）が入り混じる語句の誤訳の実例と対策

25年前にMTの試訳依頼を受けた際に、それは文章の末尾のピリオド（.）と数字の小数点（.）を区別できなかった。今は解決している。しかし“コンマ（,）と（and）と（to）”が混じると、いま市販のMTは、正訳できない場合が多い。

例えば、

**Claims 1, 2 can not be granted a patent because documents A, B describe the same technique.**

○ 文書A、Bが同じテクニックを説明するので、主張1、2は特許を与えられることができない。

× クレーム1とクレーム2は特許を与えることができません、ので、Aを文書化する、Bは同じ技術について記述します。

○ ドキュメントA、Bが同じテクニックについて説明するので、請求項1、2に特許を与えることができません。

× 1例は主張します、2例は特許を受けることができません、文書A、Bは同じ技術を記載します。

**② The inventions of claims 1 to 3 and 6 can not be granted a patent because documents 1, 2 describe the same technique.**

× 文書1のため、主張1から3と6の発明品は特許を与えられることができず、

2は同じテクニックを説明する。

- ドキュメント 1 および 2 が同じ技術について記述するので、クレーム 1~3 と 6 の発明は特許を与えることができません。
- ドキュメント 1、2 が同じテクニックを説明するので、請求項 1~3 と 6 の発明品に特許を与えることができません。
- 文書 1、2 が同じ技術を記載するので、主張 1~3 と 6 の発明は特許を受けることができません。

発明者や特許事務所が外国特許庁に出願すると、上記の返書がくる。関係者は文献 1 と 2 を調べ、補正して再出願するか又は諦める。従って、この誤訳は致命傷である。この文体は、法律書や貿易の契約書にも現れるので、絶対に解決すべき課題である。

そこで、私は、翻訳の作業中に、上記の類似文を見つけたらメモして、試訳していた。それらを次に示す。

### ③ (,) (and), (to) が混じる語句の誤訳

3-1. A width between pins A, B is 5m.

- × ピンA間の幅、Bは5mです。
- × ピンAの間の幅ゆえ、Bは5m。
- ピンA、B間の幅は5mです。
- × 点Aの間の幅、Bは5mです。

英文を下記のように変えると、4社が○

A width between pins A and B is 5m.

3-2. A width between pins 1, 2 is 5m.

- × ピン1の間の幅、2は5mである。
- ピン1と2の間の幅は5mです。
- ピン1、2の間の幅は5mです。
- ピン1、2の間の幅は、5mです。

英文を次のように変えると、4社が○

A width between pins 1 and 2 is 5m.

3-3. The prices of samples A, B are cheap.

- サンプルA、Bの値段は安い。
- × サンプルAの価格、Bは安い。
- サンプルA、Bの価格は安いです。
- × サンプルAの価格、Bは安いです。

3-4. A difference between samples A, B is big.

- × サンプルAの違い、Bは大きい。
- × サンプルAの間の違いゆえ、Bは大きい。
- サンプルA、Bの違いは小さいです。
- × サンプルAの違い、Bは小さいです。

英文を次のように変えると、4社が○

Samples A, B are expensive.

3-5. A difference between samples A, B and C is small.

4社が○： サンプルA、B、およびCの違いは小さい。

3-6. Buttons 10, 11 have visible surfaces (12, 13).

4社が○  
ボタン10、11は

可視の表面(12、13)を持っている。

3-7. Each of buttons 10, 11 has visible surfaces (12, 13).

- × ボタン10について、11は可視の表面(12,13)をそれぞれ持っている。
- 各のボタン10及びに11は可視の表面(12と13)があります。



○ 各のボタン 10、11 には、目に見える表面が (12、13) あります。

○ ボタンの各には、可視表面 (12、13) を持ちます。

3-8. In the equations (1, 2), G indicates a gain, z1, z2 and z3 indicate zero points, and d indicates a pole.

○ 方程式(1、2)中で、G は、増加、z1、z2、および z3 が 0 ポイントを示すことを示し、d はポールを示す。

○ 方程式(1 と 2)では、G は利得を示します。z1、z2 および z3 はゼロ点を示します。また、d は柱を示します。

× 方程式(1、2)で、G は利得を示します、そして、z1、z2、および z3 はポイントを全く示しません、そして、d はポールを示します。

○ 方程式 (1、2) では、G は増加、z1、z2 と z3 がゼロ点を示すことを示します、そして、d は棒を示します。

3-9. A device D, which inputs values 1, 2, calculates a difference between the values 1, 2 and outputs the difference, is located at a position A.

○ 値 1、2 を入力し、値 1、2 の違いを計算し、違いを出力する機器 D はポジション A に置かれる。

× 装置 D(それは値 1、2 を入力する)は、値 1 の間の違いを計算します、2、また違いを出力する、位置 A に位置します。

○ 装置 D(値 1、2 を入力して、値 1、2 の違いについて計算して、違いを出力します)は位置の A に位置しています。

× 装置 D (それは価値 1 を入力します)、2 つは位置 A に位置すると、価値 1、2 の違いは計算します、そして違いを出力します。

3-10. The values 1, 2 and 3 from the points H, I and J in the block are sent to devices 4, 5 and 6, respectively.

× ポイント H からの価値 1、2、および 3、私とブロックの中の J は機器 4、5、および 6 にそれぞれ送られる。

× ポイント H からの値 1、2 および 3、私およびブロック中の J は、装置 4、5 および 6 へそれぞれ送られます。

× それぞれブロックのポイント H からの値 1、2、および 3、I、および J を装置 4、5、および 6 に送ります。

× 点 H からの値 1、2 と 3、それぞれ、私とブロックの J は装置 4、5 と 6 に送られる。

3-11. Fig. 8 shows a state in which a material 15 is compressed between the upper and lower dies, 30, 40.

○ 図 8 は、素材 15 が上部と下のダイス、30、40 の間で圧縮される状態を示す。

× 図 8 は、材料 15 が上部で、より低いもの間で圧縮される状態が死ぬことを示します、30 と 40。

× 8 が死ぬのを物質的な 15 上下の間に圧縮される状態に案内している図、30、40

△ 図 8 は、材料 15 が上下の型、30、40 の間で圧縮される州を表します。

3-12. The coating separation parts 6, 7 corresponding to ridges 1, 2 serve as the protective covers 8, 9, respectively.

× コーティング分離パート 6(それぞれ尾根 1(防護カバー8、9 としての 2 サーブ)と一致している 7)。

△ 尾根 1 および 2 に対応するコーティング分離部分 6 および 7 は、掩護戦闘機 8、9 としてそれぞれ役立ちます。

○ 尾根 1、2 に対応するコーティング分離パート 6、7 は保護的なカバー8、9 としてそれぞれ機能します。

○ それぞれ、峰 1、2 と一致しているコーティング分離部品 6、7 は、保護カバー8、9 として用いられます。

3-13. The plated layers A, B on the surfaces C, D are removed from the device.

× メッキされたレイヤーA、表面 C の上の B、D は機器から除去される。

× メッキ層 A、表面 C の上の B、D、装置から取り除かれます。

× 装置からメッキ層 A、表面 C、D の B を取り除きます。

× 装甲層 A、表面 C に関する B、D は装置から削除されます。

3-14. The plated layers A, B and E on the surfaces C, D and F are removed from the device.

○ 表面 C、D、および F の上のメッキ層 A、B、および E は機器から取り除かれる。

○ 表面 C、D および F の上のメッキ層 A、B および E は装置から取り除かれます。

○ 装置から表面 C、D、および F のメッキ層 A、B、および E を取り除きます。

△ 表面 C、D と F の装甲層 A、B と E は、装置から削除されます。

3-15. The plated layers A and B on the surfaces C and D are removed from the device.

○ 表面 C と D の上のメッキ層 A と B が機器から取り除かれる。

○ 表面 C および D の上のメッキ層 A および B が、装置から取り除かれます。

○ 装置から表面 C と D のメッキ層 A と B を取り除きます。

△ 表面 C と D の装甲層 A と B は、装置から削除されます。

3-16. The plated layers A, B and E to F on the surfaces C, D and G to H are removed from the device.

○ 表面 C、D、および G から H の上のメッキされたレイヤーA、B、および E から F は機器から除去される。

× H への表面 C、D および G の上の F へのメッキ層 A、B および E は、装置から取り除かれます。

○ 装置から表面 C、D、および G から H のメッキ層 A、B、および E から F を取り除きます。

× 装甲層 A、表面 C に関する F への B と E、D と H への G は装置から削除されます。

3-17. Figs. 23 to 26 show examples of a stopper 100 for a fastener.

× 図 ファスナーのためのストッパー 100 の 23 から 26 ショー例。

○ 図 23~26 は、ファスナー用の栓 100 の例を示します。

○ 図 23~26 はファスナーのために栓 100 の例を示しています。

○ 図 23~26 は、ファスナーのためにストッパー100の例を表します。

3-18. Figs. 1, 3 to 5, 8 and 10 to 26 show examples of a stopper.

× 図 1、5時3分前、ストッパーの8つ、および10から26ショー例。

× 1と3は、5、8および10まで26まで栓の例を示します。

○ 図 1、3~5、8、および10~26 は栓の例を示しています。

○ 図 1、3~5、8と10~26は、ストッパーの例を示します。

3-19. Refer to paragraphs [0003] and [0017] to [0074] and [FIG.1] to [FIG.3] and the like.

× パラグラフ [0003]と [0017]から [0074]を、そして [FIG.3]とそのようなもの [FIG.1]を参照する。

× パラグラフ [0003]、 [0074]への [0017]および [FIG.3]などへの [FIG.1]を参照してください。

× パラグラフを参照してください。0003と0017年の対0074と図3、および同様のものへの図1。

× パラグラフ [0003]と [0017]に言及します

[0074]、そして、[図1][図3]、そして、その他。

上記の英文の末尾の

(and the like)を省略。

Refer to paragraphs [0003] and [0017] to [0074] and [FIG.1] to [FIG.3].

○ パラグラフ [0003]と [0017]から [0074]を、そして [FIG.3][FIG.1]を参照する

× パラグラフ [0003]、 [0074]への [0017]および [FIG.3]への [FIG.1]を参照してください

× パラグラフを参照してください。0003と0017年の対0074と図3への図1

× パラグラフ [0003]と [0017]に言及し、[0074]そして [図1][図3]

3-14. The documents 4, 5 describe an example in which gap is placed between a tip and a member.

△ 文書 4、5 は、ギャップが秘訣とメンバーの間で置かれる例を説明する。

○ ドキュメント 4 および 5 は、ギャップが先端とメンバーの間で置かれる例について記述します。

○ ドキュメント 4、5 はギ

チップがチップとメンバーの間に置かれる例について説明します

○ 文書 4、5 は、隙間が先端とメンバーの間に置かれる例を記載します。

3-15. Claims 1 to 8, 13 to 16, 18 and 22 to 26 of the specification have no new idea.

× 主張仕様の 26 への 1 から 8、16 時 13 分前、18、および 22 は新しいアイデアを全然持っていない。

× 8 と 13 までクレーム 1 は 16、18 および 22 に対して明細のうちの 26 まで新しい考えを持っていません。

× 請求項 1~8、13~16、18、および 26 の 22~仕様には、どんな新しいアイデアもありません。

× 主張 1~8、13~16、18 と 22~仕様のうちの 26 には、新しい考えがありません。

3-16. Addresses of terminals 2, 3 are referred to as B, C, respectively.

× ターミナル 2 のアドレス、3 は B、C と各称される。

× ターミナル 2 と 3 のアドレスは各々 B(C)と呼ばれます。

○ 端末 2、3 のアドレスは各々 B、C と呼ばれます。

× ターミナル 2、3 のアドレスは、それぞれ B(C)と呼ばれます。

3-17. The addresses of the terminals 1, 2 are referred to as C, D, respectively.

× ターミナル 1 のアドレス、2 はそれぞれ C と称される、D。

× ターミナル 1 および 2 のアドレスはそれぞれ C(D)と呼ばれます。

○ 端末 1、2 のアドレスはそれぞれ C、D と呼ばれます。

× ターミナル 1、2 のアドレスは、それぞれ C (D) と呼ばれます。

3-18. The inventions in the claims 1, 2 and 14 to 17 of the application should not be granted under Patent Law Section 29.

× アプリケーションの 17 個への主張 1、2、および 14 における発明品は特許法セクション 29 の下で特許を与えられるべきでない。

△ クレームこの出願のうちの 1~17、2 および 14 での発明は、特許法セクション 29 の下の特許を与えられるべきではありません。

○ このアプリケーションの請求項 1、2、および 14~17 における発明品は Patent 法セクション 29 の下に特許を与えるべきではありません。

× 主張 1、2 と 14~このアプリケーションのうちの 17 の発明は、特許法第 29 節の下で特許を受けてはいけません。

3-19.

4社○ ピン1から10のプレッシャは高い。

Pressures of pins 1 to 10 are high.

Pressures of pins 1 to n are high.

△ 1からnピンのプレッシャーは高い。

× nへのサンプル1の圧力は高い。

○ サンプル1~nの圧力は高いです。

× nへのサンプル1の圧力は、高いです。

3-20.

There are books under samples 1, 2.

4社が○ サンプル1,2下に本があります。

3-21.

Books under samples 1, 2 are red.

× サンプル1下の本、2は赤い。

○ サンプル1および2の下の本は赤い。

○ サンプル1,2の下における本は赤いです。

○ サンプル1,2の下の本は、赤いです。

3-22.

Figs. 1A, 1B show perspective views.

○ 図1A、1Bは透視図を示します。

× 1Aと1Bは透視図を示します。

○ 図1A、1Bは斜視図を示しています。

○ 図1A、1Bは斜視図を示します。

#### ④ 誤訳のまとめ

4-1. 例えば、(samples 1 and 2)は正訳するが、(samples 1, 2)は誤訳が多い。

しかし、(samples 1, 2 and 3)は正訳する。

4-2. 上記に前置詞(to)が混じると、誤訳が多い。(to)は(～に、～へ)の他に、例えば、(integers 2 to 5)は「整数2~5」の意味。(ratio of 1 to 1)は「1対1の比」の意味である。(～に、～へ)の意味でも、目的語の部分が長いと、動詞と(to)の関係が不鮮明になり、誤訳する。

例えば、

It sends a signal

whose amplitude is low to a receiver.

× それは、振幅がレシーバーに低いシグナルを送ります。

× それは、その振幅がレシーバーに低い信号を送信する。

× それは振幅が受信機に低い信号を送信する。

× それは振幅がレシーバに低い信号を送る

4-3. 短文の場合に正訳で、長文で誤訳というわけでない。関係代名詞の混じった長文でも正訳する場合もある。

4-4. 誤訳に規則性が、特にあるわけでない。

#### ⑤ 対策としての提案

前述のように、誤訳と正訳のあいだに規則性がない。自然現象は論理的な原理があり数式で表現できるが、言語文法は、数式で表現できないからと考える。

私は、対策として、前述のように類似文を最大限にインプットして、共通部分と規則性を見だし、フローチャート化するか用例翻訳として、MTにプログラミングすることを、誤訳解決の一つのアプローチとして提案する。

市販の家庭用製品の故障や苦情は、上位の約5項目を解決できれば、その80%を解決できるといわれている。そこで

5-1. 翻訳ソフトのメーカーにある誤訳例を調査し、上位の項目をリストし体系化して、誤訳対策の対象を絞る。

5-2. 対象の誤訳例に類似の語句や文章を、今回の発表で述べたように、最大限にインプットして、上記のように規則性を見いだして、フローチャート化する。

5-3. そのために、翻訳の作業中に、上記の類似の語句や文章を見つけたらメモして試訳していた。この作業に実務翻訳者の協力が必須である。この作業に協力するには、翻訳の実務経験とMTの使用経験が長いことが不可欠であるが、そういう翻訳者は見あたらない。

5-4. そこで、(---)の誤訳に関連した類似文を集めたい場合、私、吉川に連絡してほしい。最大限度に協力する。

5-5. MTの問題を解決する優れた新理論が現れたら、翻訳ソフトの誤訳例に対して正訳しているか確認する必要がある。

## ⑥ 終わりに

2年前の[「第一回特許情報シンポジウム」]で述べたように、市販の誤訳の問題は、翻訳ソフトを変更すれば解決できて実用化直前とプログラミングの素人が感じて、彼方を立てれば此方が立たず、実際は難しいらしい。

しかし、私が今回指摘した「コンマ(,)と(a n d)と(t o)が混じる語句の誤訳」は、他を犠牲にしても解決すべき課題である。

とにかく、今の問題点の全ての解決が無理ならば、数量表現だけでも正確に訳してほしい。そのために、前回と同様に、「短文で数量表現に特化した翻訳ソフト」の開発を要請する。

誤訳の問題点は、私が今まで発表してきたレポートを一読すれば分かるように、1社が正訳で、他社が誤訳の場合もある。お互いに連携して解決したらと、第三者は考える。私も各社に暗に提案したが、独自に直すと各社からいわれた。しかし、7年たっても、市販品から解決したという気配は感じられない。従って、独自解決は無理と見なすべきである。

この状況を解決するには、翻訳ソフトのメーカーだけでなく関係者が提携する必要がある。今の翻訳ソフトのレベルで可能なこと、不可能なこと、妥協レベル(ユーザが許容するか?)について、技術者、研究者、翻訳者、販売者、言語学者を含めて総括する必要がある。

本稿の土台となる「翻訳ソフトの試訳」を12年前から徐々に進めてきた。多くの研究者や技術者に、電子メールで質問し或いは直接訪問したこともあった。そのつど、浅学非才の私の、唐突で、時に失礼な愚問に対して意見や助言を承った。ここに御礼を申し上げると共に、翻訳ソフトの問題が解決し有意義な存在になることを、あらためて願っている。

## Session 3

特許翻訳の品質を向上するための形態素解析  
結果を利用した文書比較・日本語精査ツール

— 歌詠と鶯 —

の試作





# 特許翻訳の品質を向上するための形態素解析結果を利用した 文書比較・日本語精査ツール－歌詠と鶯－の試作

楠本 浩二<sup>†</sup>, 山口 日緒里<sup>††</sup>, 鈴木 貴年<sup>††</sup>, 千引 春菜<sup>††</sup>  
<sup>†</sup>株式会社 クレステック    <sup>††</sup>アイビー・システム株式会社

E-mail: k-kusumoto@crestec.co.jp

## Application for Comparing and Checking Document Contents to Enhance Quality of Patent Specifications on Language Translations

Koji Kusumoto<sup>†</sup>, Hiori Yamaguchi<sup>††</sup>, Takatoshi Suzuki<sup>††</sup>, Haruna Chibiki<sup>††</sup>  
<sup>†</sup>CRESTEC Inc.,    <sup>††</sup>Ivy System Co., Ltd.

E-mail: k-kusumoto@crestec.co.jp

### 概要

企業間、企業と個人との間で通常、交換される特許明細書の翻訳原稿は、以前から広く利用されている Microsoft Word、Adobe PDF のようなファイル形式である。一般的なユーザー環境において、このようなファイルを対象に処理可能な、特許翻訳の品質向上を支援するためのパーソナルツールを提供する。2つ以上の任意のファイルの中から類似した段落または文の対(ペア)を決定し、その差分を対比表形式でわかりやすく表示する。翻訳者は、変更、削除、挿入された文章の正確な箇所を容易に確認できる。例えば、明細書の原文と複数の対訳文書と翻訳された全文との差分を随時表示し、原文の忠実性を確認しながら翻訳できる。また、単一の明細書内にある類似した文をすべて検索し、差分を表示すると、反復した文、修飾句が追加された文や省略された文、対比文、並列文、表記ゆれの文、などを一覧できる。この内容から従来は見逃されていた誤り箇所も視覚的に確認でき、明細書の品質改善や効率的な翻訳の支援ができる。日本語の明細書では、形態素解析結果から用語だけを抽出し、この抽出用語リストから用語の適切性を判断する。その後、用語として登録すべきものだけを用語辞書として蓄積し、この辞書を別の明細書に対する用語の精査時に再利用する。更に、形態素を用いた精査ルールを定義することによって明細書の日本語精査も一部可能である。本稿では翻訳時の明細書にこれらの機能を適用した例を紹介し、その可能性と課題について述べる。

### 1. はじめに

歌詠(UTAYOMI)と鶯(UGUHISU)は、法令、条例、契約書、業務規定書、保険約款のような重要文書を対象とし、このような文書の自動化処理、比較、精査といったソリューションを提供するために開発された。比較モジュールである歌詠の利用実績としては、標準となる保険約款と各保険商品の約款との条文比較、改定された条文に類似した別の条文のリストアップ、差分情報からの改め文自動生成などがある。日本語精査モジュールである鶯を利用したものは、日本語としての一般的な規則 [1]、電子化文章のルール [2]、各業界のルール [3, 4, 5, 6] に従った日本語精査、紙から電子化の際の OCR 実行結果の誤認識チェック、入力時のタイプミスによる誤字の検出がある。歌詠および鶯双方を利用したものは、条の繰り上げと繰り下げに伴う条番号を引用している箇所の修正もれのチェックがある。

特許明細書は、法律文書であり、技術文書である。正確な技術内容を国内外に迅速に登録するための特許明細書の作成業務の支援は、技術立国である日本にとって大きな課題で

ある。ここで日本国内の特許出願件数を見ると、ピークだった2001年の約43万件から2011年には約34万件に減少した。しかし、国内出願を海外にも出願する割合は、20%から40%に増加し、PCT (Patent Cooperation Treaty) 出願件数は、米国に次ぐ2位である。この統計資料は、企業が出願すべき特許を厳選するようになり、重要特許を海外に漏れなく出願する傾向を示している。したがって海外出願に目を向けた特許明細書の作成、翻訳の重要性が増しており、特に和文英訳の需要は、今後、ますます高まると予想されている。海外での特許出願では出願国での先行技術調査も必要なことから、多言語に特化した特許検索システムも研究、開発されている[7]。

また特許明細書の翻訳は、他の産業翻訳との性質の違いも多く、誤訳があった場合、国益も損なうと言われている。こうしたことから情報システムを利用して特許明細書を効率的かつ正確に作成する支援環境やツールがいくつも開発されている。本稿では、重要文書ソリューションである歌詠や鶯が特許明細書の翻訳において、どのような場面で利用機会があり、翻訳者に対してどのような支援が可能か検証する。

## 2. 特許明細書の翻訳

### 2.1 特徴

特許翻訳は、マニュアル翻訳やソフトウェアのローカライズなど、その他の産業翻訳業務と比べると以下の点において異なる。

- ・明細書には、起承転結のストーリー性がある
- ・翻訳文は最長 20 年に渡る権利書であって修正が困難
- ・明細書の記載事項には新規性がある
- ・原文に忠実であるべきであって意識は許されない
- ・翻訳文が原本との差異を提示するための資料となる

平成6年の特許法改正では「外国書面出願の係る審査の運用方針」の中で「翻訳文とは、日本語として適正な逐語訳 (word by word translation) による翻訳文 (外国語書面の語句を一对一に文脈に沿って適正な日本語に翻訳した翻訳文) をいう」と定義され、PCT 出願特許の翻訳では日本語として不自然なものは特許翻訳ではないとされた。翻訳者は、常に明細書の原文と比較しながら、かつ、構造の異なる言語へ自然な表現をする必要がある。

### 2.2 業務形態・環境

特許の翻訳業務形態を見ると SOHO 環境での翻訳作業も多い。一般に企業の重要文書の作成は、共通のシステムを利用し、複数の担当者で共同執筆することが多いが、一方、特許翻訳の場合は、ひとりの翻訳者が担当し、その内容は第三者の目に触れないようにクローズド環境で厳密に管理される。

発明者と弁理士、弁理士と翻訳者、発明者と翻訳者との間には多くのコミュニケーションが必要な場合がある。しかし、翻訳者と依頼者の文書のやり取りでは、特定の共通システムを利用することは少ない。MS Word ファイル、Adobe PDF ファイル、紙媒体を交換するため、これらの形式のファイルを直接処理可能なことが要求される。このためインターネットを介したウェブベースのシステムよりもスタンドアロン型のソフトウェアが望まれる。

PCT 出願の場合、原文が出願されて初めて内容を精査する人は翻訳者であり、原文中の誤りを最初に発見することも多い。誤り箇所を MS Word のコメント機能を使用して、随時指摘できることは、テキストファイルの利用よりも便利である。また、特許庁への提出フォーマットが MS Word 形式であることからこれを対象として処理できることはメリットがある。

特許翻訳の場合、原言語で記載された特許原稿をひとりの翻訳者が担当して翻訳することがほとんどである。特許単体で起承転結構成になっている 1 件の特許をひとりの担当者が担当し、独力で翻訳する。更に、周辺技術を押さえるために、関連特許がしばしば同時に複数出願されることがある。この場合も複数の関連特許を同一の翻訳者が担当しない場合もあって、翻

訳者は一連の複数特許の全体象が見えないまま翻訳することになる。これらの特許の間には、背景や構成など、共通した内容が多い。最終的には、特許事務所や翻訳事務所が、これらを取り纏め、異なる翻訳者による翻訳結果の間のフレーズの統一、用語の統一など、整合性をとる必要もある。

### 2.3 翻訳に向けた和文明細書の改善

産業機器マニュアルにおいては、和文英訳時の品質を向上するための一手段として英訳を視野に入れて和文ドキュメントを作成する手法[8]が提案されている。特許明細書も一度、産業日本語[9]に置き換えると翻訳時の品質が向上すると報告されている。翻訳者が一旦、和文明細書を産業日本語のようなわかりやすい文章に置き換えて翻訳した場合、産業日本語と原文と翻訳文を発明者および翻訳者が常に共有していることが重要となる。

### 2.4 翻訳支援利用時の問題

特許翻訳作業でも TRADOS など翻訳メモリーが採用されている。複数の翻訳支援ソフトを組み合わせることも多い。上下左右の対訳形式表示は、翻訳者の目線移動が少ないので疲労が少なく、致命的な誤抜けが防止できるので、多くの翻訳システムではこの形式を採用している。しかし、翻訳支援ツールを用いたときの最大の問題点は、記憶装置にあるデータが文単位のものであり、翻訳者の注意が単文ごとの処理に集中することになるため、明細書全般を通じて流れる「技術思想」を翻訳者が見失いがちなことである[10]。翻訳者は、原文を見てから日本語を見るので置き換えに注意がいく。この結果、不自然な日本語になっても見過ごすことになる。これを防ぐには、定期的に日本語文章だけを読み、読み手がよく理解できるか、伝わりやすい文章になっているかを考える必要がある。また、連続して考えると思い込みで縛られ、ミスや不自然さに気が付かないことが多くなるため、少し時間を置いて見直すことも効果的であると言われている。

## 3. 比較機能

### 3.1 従来の比較ツール

シェアウェア、フリーウェアを含めて数多くの差分検出ツール[11]がある。また MS Word に搭載されている標準機能の1つである文書の比較機能は、比較対象を詳細に設定できるオプションもあって便利である。しかし、変更箇所がインラインで表示されるため、直感的にわかりにくく読解の流れを疎外してしまう。更に、その表示結果は第三者への提出文書としては適していない。更に、例えば文の入れ替え、ページ単位の入れ替え、章や節の位置関係の変更といった文章構成の変更を認識することは難しい。

### 3.2 歌詠の特徴

歌詠の比較機能の特徴を以下に述べる。歌詠は、重要文書ソリューションにおいて文章内容の変更を別の文書にも反映する必要から、類似した箇所が別の文書にも存在するか検索し、存在した場合はその差異をわかりやすく表示するために利用された。歌詠は、以下の要件も満たすことができる。

- 2つ以上の文書を比較して同時に表示
- 論理構造が異なる文書の比較
- 和文と英文の混在した文書の比較
- 異フォーマット文書間の比較
- XMLの文脈属性条件で比較
- 文書内の表同士の比較
- 同一内容を含む文書の検出
- 比較結果表示の切り替え
- ヘッダ、フッターの内容比較
- 空白を無視したコンテンツ比較
- 画像の位置変更
- 実行履歴の後日参照

2つ以上の文書の比較例とし

て Figure 1 のように、原文と原文の対訳、産業日本語の対訳と翻訳文など、これらを同時に比較して表示することが考えられる。

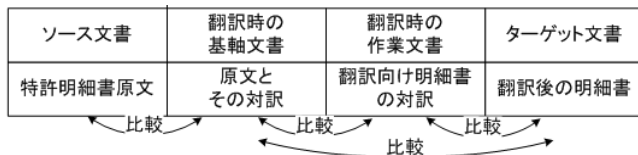


Figure 1 2つ以上の文書の比較例

### 3.3 歌詠の比較処理

歌詠の処理手順を簡単に説明する。文書内容を順次読み出す比較単位をまず指定する。通常は、段落か文か、またはその双方である。日本語の場合、句点「。」が存在するため一文の判定は比較的容易であるが、一方、英語の場合、必ずしも「。」が文の終わりとは限らないため、文末例外パターンリストを用意して文の区切りを認識する。このリストはユーザーが定義できる。

基軸となる文書と比較対象の文書双方をロードし、比較単位に分割する。分割単位の文字列の長さは不定なので Levenshtein の編集距離[12]を求めて類似度を求める。例えば、分割単位1の中の文字列を String1 とし、分割単位2の中の文字列を String2 とする。分割単位が段落の場合、String1 と String2 は双方とも1つ以上の文を包含している。このとき String1 と String2 の類似度  $\text{similarity}(\text{String1}, \text{String2})$  は、以下の式で求められる。

$$\frac{\max(\text{length}(\text{String1}), \text{length}(\text{String2})) - (\text{String1} \text{と} \text{String2} \text{の編集距離})}{\max(\text{length}(\text{String1}), \text{length}(\text{String2}))} \times 100$$

類似度が一定の閾値以上の場合、その差分を対比表の同一行に表示する。ユーザーは、実行時に任意の類似度閾値と、文字ごとに比較するか形態素ごとに比較するか一方を指定できる。文章内容の差分を見るときは形態素ごと、校閲のときは文字ごとの比較が適している。以下に分割単位を文とし、形態素ごとに比較した対比表の表示例を示す。

| 校正前.docx                 | 校正後.docx                 | 文書間類似度<br>75.63%<br>文書包含率<br>100% |
|--------------------------|--------------------------|-----------------------------------|
| 1 抵抗Rを端点Aに並列に接続する。       | 1 抵抗Rを端点Bに直列に接続する。       | 66.66%                            |
| 2 コンデンサCを端点Bの一方に直列に接続する。 | 2 キャパシタCを端点Bの一方に並列に接続する。 | 84.61%                            |

Figure 2 形態素ごとの一文単位の比較例

文番号1の校正前の形態素数は12、校正後の形態素数は11、編集距離は4なので類似度は、 $(12 - 4) / 12 \times 100 = 66.66\%$ となる。以下に文字ごとに比較した例を示す。

| 校正前.docx                 | 校正後.docx                 | 文書間類似度<br>76.98%<br>文書包含率<br>100% |
|--------------------------|--------------------------|-----------------------------------|
| 1 抵抗Rを端点Aに並列に接続する。       | 1 抵抗Rを端点Bに直列に接続する。       | 81.25%                            |
| 2 コンデンサCを端点Bの一方に直列に接続する。 | 2 キャパシタCを端点Bの一方に並列に接続する。 | 72.72%                            |

Figure 3 文字ごとの一文単位の比較例

文番号1の文字数は校正前、校正後それぞれ16、編集距離は3なので類似度は、 $(16 - 3) / 16 \times 100 = 81.25\%$ となる。

Figure 2 と Figure 3 の「文書間類似度」は、各行で算出された類似度の相加平均である。「文書包含率」は、比較文書の要素と類似した要素の基軸文書の要素すべてに対する割合である。この例では校正前の2つの文が、校正後の文すべてと類似していると判定され、100%になっている。

#### 比較方法

歌詠では2通りの比較方法がある。例えば、文書 A 中の m 個の要素集合を順序付リスト  $A < a_1, a_2, a_3, \dots, a_m >$  と表し、文書 B 中の n 個の要素集合を順序付リスト  $B < b_1, b_2, b_3, \dots, b_n >$  と表す。第1の比較方法は、文書 A と B の相対する要素間の類似度  $\text{similarity}(a_1, b_1)$ ,  $\text{similarity}(a_2, b_2)$ ,  $\text{similarity}(a_3, b_3) \dots$  を先頭から順に求めていく比較である。したがって比較の回数は  $\max(n, m)$  となる。第3の比較文書 C がある場合は  $\text{similarity}(a_1, c_1)$ ,  $\text{similarity}(a_2, c_2)$ ,  $\text{similarity}(a_3, c_3) \dots$  の計算が追加される。

第2の比較方法は、文書 A を基軸として文書 B の要素全体を比較対象と捉え、類似性が最も高い要素を検索する総当たりの比較である。例えば、 $a_1$  に関する類似度は  $\max(\text{similarity}(a_1, b_i)) (i = 1, \dots, n)$  となる。したがって比較の回数は  $n \times m$  となる。この場合も類似度が一定の閾値以上の場合、その差分を

対比表の同一行に表示する。この比較方法によると、比較回数が多いために実行速度は低下するが、文書 A と文書 B において、要素の表示順序や文書の論理構成が全く異なっているにもかかわらずその内容の差異を正確に表示できる利点がある。

### 言語を区別した比較

日本語のように分かれ書きのない言語と、スペースで区切られた言語との比較処理方法を切り替える。英文は、半角スペースを無視して編集距離を求めないと正しい類似度を計算できない。英文と和文が交互に記載される翻訳メモリの対訳形式にも考慮し、比較時に言語の判別をする。こうして和文と英文との比較計算の実行を回避する。

### 画像の比較

MS Word 内の図はすべて、共通フォーマットのビットマップ画像ファイル(.png)として作成される。歌詠は、このビットマップ画像をバイナリー比較し、その画像が一致か不一致の二者択一判定をする。したがって画像の差分は認識できない。MS Word 内の図が同一である場合、比較処理でも同一と見なされる。ビットマップ画像はファイル参照として対比表に表示される。

## 3.4 歌詠の表示

### 比較結果の表示

指定された閾値以上の場合、表の同一行のセルにそれぞれ差分を表示する。ペアごとの行は削除されたか挿入された内容を意味する。デフォルトの差分表示として(追加または変更された) 差異のある箇所には下線で示し、削除された箇所は、文字“□”とオーバーラインで示している。

### 比較結果表示の切り替え

比較結果は XHTML 形式の表である。CSS ファイルによって、表示を切り替える。再実行せずに目的に応じたビューイングが可能である。WEB ブラウザのデフォルトは Internet Explorer である。Google chrome, Firefox の利用を選択できる。XHTML 形式なのでエクセル文書や XSLT を介し、Adobe PDF、MS Word のようなオフィス文書としても出力できる。

## 4. 明細書中の類似文のリストアップ

特許明細書は、権利化したい内容について類似した表現を繰り返すことが多い。歌詠を実行して同一文書を比較すると、このような表現箇所を検出し、その差分を表示できる。

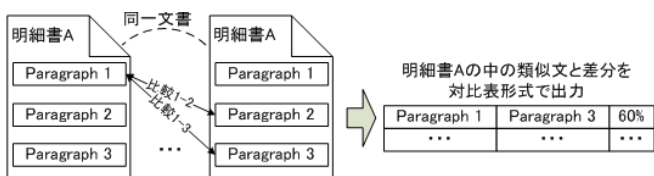


Figure 4 文書内の文を比較し類似文をリスト

具体的には、Figure 4 に示すように、文や段落を単位として比較した結果のうち、類似度が最も高いペアをリストすることによってこれを実現できる(同一要素は比較しない)。この例では、Paragraph1 と Paragraph3 が極めて類似した文であったことを示している。

### 4.1 類似文リストアップの効用

翻訳時に類似性の高い文をリストすることによって、以下のような文章の存在がわかる。

- (1) 反復されている重要文
- (2) 修飾句、修飾節の有無
- (3) 別の語句による表現
- (4) 並列文、対比文
- (5) 表記上のゆれ

(1)～(4)は、翻訳時の支援になる。(5)は、文書作成時や翻訳において発生する「表記のゆれ」であって、以下のパターンがある。

- ・句点、ピリオドの欠落、読点位置の差異
- ・名詞の区切り(名詞の間を繋ぐ「の」の有無)、順序の差異
- ・語尾、送りがなの差異
- ・助詞(“は”, “が”, “の”, “を”, “に”, “へ”, “と”等)の差異
- ・カタカナの表記ゆれ
- ・不要な空白やミスタイプ文字の混入

## 5. 日本語精査機能

### 5.1 日本語明細書における審査項目

日本特許庁に出願された日本語明細書は、日本特許法第 29 条、第 32 条、第 36 条、第 37 条、第 39 条に基づいて審査される[13]。明細書の原文または翻訳に問題があって拒絶される場合がある。特許庁審査基準は以下である。

- ・日本語として正確でない(主語述語、修飾関係の不備、前記箇所不明瞭、誤字、脱字、当て字)
- ・用語の不統一
- ・用語が技術用語ではなく、定義されていない
- ・一般名称ではない商標名を使用
- ・単位が計量法規定に沿っていない
- ・図面および符号説明に不備がある

鶯は、利用シーンに応じた個別の精査ルールを組み込むフレームワークである。以下に日本語特許明細書翻訳の場合の精査支援を考察する。

### 5.2 特許明細書における用語

特許は、新しい技術を記載しているため、発明者がその特許の中で初めて定義する用語や最新の用語を利用することも多い。一方、翻訳者にとっては、正しい最新用語集をどのように

充実させるかが常に課題となる。翻訳者は、複数の単語からなる造語だと思えるような名詞にしばしば悩まされる。そのまま日本語に置き換えると、一般的ではない用語という理由から特許法第 36 条違反で拒絶されることが多くある。用語に関してまとめると以下の場合があると考えられる。

- (1) 用語として適切
- (2) 別の一般的な用語にすべき
- (3) 用語辞書にひびく新しい概念
- (4) 単純なスペルミスまたは誤字

(2)の例として電子分野では「コンデンサー」は、英文では“capacitor”であって“condenser”ではない。このように用語辞書は特許の出願分野ごとに必要である。(3)は、本来は明細書の中で定義されている必要がある。(4)は、翻訳者から明細書の作成者にコメントすべき事項となる。

### 5.3 用語候補の抽出と辞書作成

#### 用語抽出ルール

形態素解析の結果、形態素の一連の並びが特定のパターンに一致する場合、複数の形態素をまとめて単一の用語と見なして外部ファイルに出力する。篤ではこれを「用語抽出処理」と呼ぶ。特定のパターンをカスタマイズ可能なように「用語抽出ルール」として記述する。その例を一部以下に示す。

#### (1) 複合名詞とするパターン

名詞と連結して出現するものうち用語とするものを定義。

- <接頭詞> + <名詞>
  - 「主<接頭詞>」+「成分<名詞>」⇒「主成分」
  - 「再<接頭詞>」+「起動<名詞>」⇒「再起動」
- <名詞> + <名詞, 接尾, 一般>
  - 「汎用<名詞>」+「(的 | 化 | 性)<接尾>」⇒「汎用(的 | 化 | 性)」
  - 「電子<名詞>」+「媒体<名詞, 一般>」⇒「電子媒体」
- <名詞, 固有名詞, 地域> + <名詞, 一般>
  - 「日本<地域>」+「固有<名詞>」⇒「日本固有」

#### (2) 複合名詞としない除外パターン

名詞と連結して出現するが、用語としないものを定義。

- (など | 等 | ごと) <名詞, 接尾>
  - 「ブロック(など | 等 | ごと)」⇒「ブロック」
- \* <名詞, 副詞可能>
  - 「構成物(すべて | それぞれ)」⇒「構成物」
- 前記 <名詞, 一般>
  - 「前記計算機」⇒「計算機」

#### 抽出用語リスト

明細書の本文から用語抽出ルールにマッチした用語を抽出し、「抽出用語リスト」として作成する。抽出した用語が用語辞

書になかった場合、未登録であることを精査結果の1つとして表示する。Figure 5 にその流れを示す。

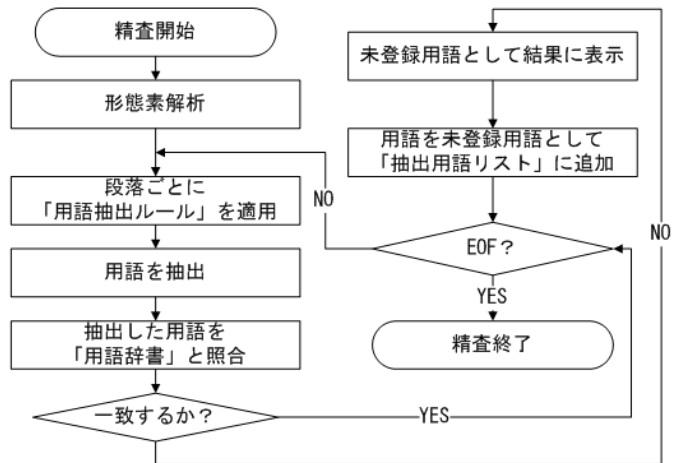


Figure 5 明細書からの用語候補の抽出

#### 使用用語の適性チェック

文章の中の用語だけを判定できると、より多角的な精査ができる。用語を単独で表示することによって、文章の中では気付かない誤りも発見できる。また上記リストされた用語には明細書中の出現頻度も併記している。頻度が少ない名詞は、表記ゆれや誤字脱字の可能性の指標にもなる。抽出用語リストの目視を必要とするために、利用者には負担もあるが、このリストをチェックすることは効果的である。

ユーザーは、この抽出用語リストから正しい用語を選択し、篤の専用辞書「用語辞書」に登録する。Figure 6 に登録例を示す。出現回数が多い未登録の「光 CPU」と「クラウド」を調査し、新しい技術用語であることを確認した場合、用語辞書に登録する。一方、明細書に統一に「コンピューティングクラウド」が多数出現していても調査した結果、一般的ではない用語であることを確認した場合、これを用語辞書に登録しない。その特許明細書の中に限って固有に使用されている用語と判断する。

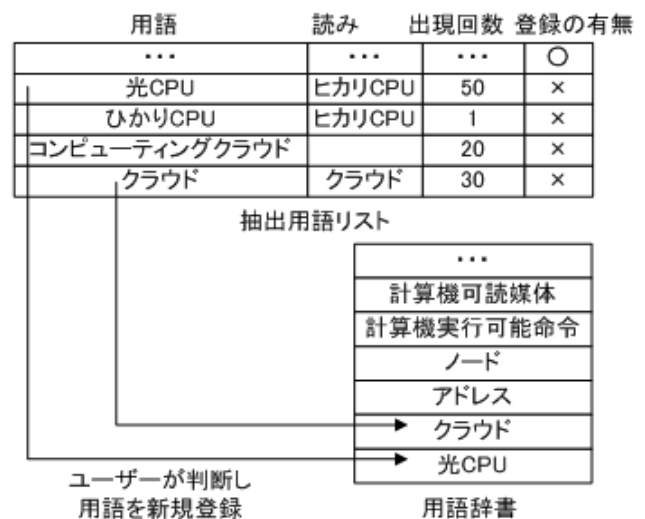


Figure 6 登録する用語の決定

## 用語辞書

この用語辞書は、用語チェックの際の正規用語として、その後の精査を通すための辞書となる。ユーザーは、常に文書を精査しつつ、この用語辞書を充実させていくことによって、正しい用語、新出の用語、誤った用語をチェックできる。

### 5.4 鶯による日本語明細書の精査項目

日本語特許明細書に対し、試作版において精査可能な機能を以下に列挙する。

#### 明細書の形式

2009年に日米欧間で明細書書式を統一した共通出願様式(Common Application Format)が採用された。この特許庁指定フォーマットを対象とする。

#### 引用精査

特許請求項では、特に厳密性が求められる「前記」の使用箇所を品詞情報から名詞、名詞句を判定してチェックする。具体的には、「前記」に対応する名詞が請求項内で既出かチェックする。翻訳過程で言語の語順の違いから、「前記」を逆の順序で記載している可能性がある。また請求項以外では通常、「前記」を使用しない。

翻訳原稿の図番や部品番号の不整合も多い。特許内の図番に重複がないこと、明細書の本文で引用されていること、記号の説明箇所の番号が本文で引用されているかチェックする。

#### 日本語精査

試作中の鶯が明細書を精査後に指摘する項目を以下にまとめる。

- ・「の」を連続して使用
- ・所有を示す「の」以外の可能性
- ・誤字・脱字の可能性
- ・「ように～ない」の指摘
- ・二重否定の表現の指摘
- ・接続助詞や助詞の「で」を多用しない
- ・間違いやすい同音異義語を使用
- ・「ならびに」の前には「および」が必要
- ・「もしくは」の前には「または」が必要
- ・一文中に「または」と「と」を両方使用しない
- ・漢字または平仮名の使い分けが不適切
- ・対応する開き(閉じ)括弧が存在しない
- ・い抜き言葉の可能性
- ・「動作性名詞」に「機能動詞」を続ける表現は冗長
- ・「形容詞＋名詞十の＋名詞」は、形容詞の係り先が不明瞭
- ・「全く」は否定表現を伴う必要がある
- ・登録商標の可能性

## 6. 実装

### 6.1 モジュール構成

本ツールのモジュール構成を以下に示す。日本語辞書、精査ルール、用語抽出ルール、用語辞書はすべて適用分野ごとに利用者が定義可能なアーキテクチャとなっている。

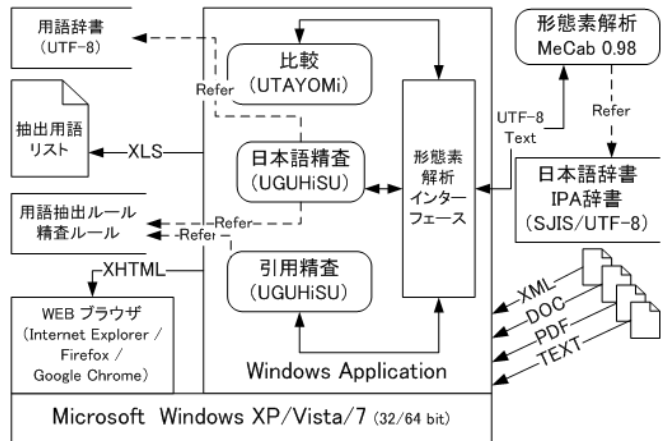


Figure7 歌詠と鶯の構成

### 6.2 形態素解析エンジンと日本語辞書

歌詠と鶯は、形態素解析エンジンとして京都大学情報学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所との共同研究ユニットプロジェクトを通じて開発されたオープンソースである形態素解析エンジン MeCab[14]を利用している。日本語の精査機能を実現するにあたり、MeCab 用の辞書の1つである IPA 辞書[15]を採用し、これを拡張している。拡張方法について以下にまとめる。

- ・登録済の単語において解析結果が望ましくない場合は、コストパラメーター[16]を変更する。
- ・未登録の単語を追加する場合、同様の文脈(左文脈 ID および右文脈 ID)において使用される登録済の別の語を辞書から探し、それを基に拡張する。
- ・未登録の単語について、意味が同一の単語のうち表記(漢字、仮名使い、送りかな等の表記)が異なるものが辞書に存在する場合は、それを基に拡張する。

現在、2,547 語の追加実績がある。また商標もいくつか登録しており、明細書中の商標利用を検出する。

### 6.3 動作条件

本実装は、Table 1 に示すようにスタンドアロンツールとして Windows OS 上で作動する。

|          |                                               |
|----------|-----------------------------------------------|
| OS       | Microsoft Windows XP / VISTA / 7 (32/64bit)   |
| 入力ファイル形式 | .doc / .docx / .rtf / .txt / .xml / .pdf      |
| その他の環境   | MeCab 0.98, Microsoft Word 2003 / 2007 / 2010 |

Table 1 動作環境

Linux や UNIX のサーバー上で比較や精査を実行し、インターネットを介してその結果をクライアントへ返却可能なアーキテクチャである。

## 6.4 処理速度

比較や精査には一定の時間を要するため、一度実行した結果を後で参照し、呼び出し可能にしている。また MeCab は解析が高速なため、形態素解析処理を経由しても比較回数が少ない形態素ごとの比較の方が文字ごとの比較よりも高速である。

|       |                                                    |
|-------|----------------------------------------------------|
| 実行マシン | Microsoft Windows 7 64bit CORE i5 (2.5GHz) RAM 4GB |
| 比較速度  | 日本語約 1800 文字／頁の間の比較に平均 1 秒                         |
| 精査速度  | 日本語約 1800 文字／頁の日本語精査に平均 2 秒                        |

Table 2 実行速度

## 7. 適用例

### 7.1 比較結果の例

類似文のリストアップと3つの文書を比較した表示例を示す。

288 If □, at step 635, the relevant PCR values are □ equivalent to those specified for release of the key, processing can proceed with step □ 640 and the provision of the key by the TPM to the requesting process.

292 If, however, at step 635, the relevant PCR values are not equivalent to those specified for release of the key, processing can proceed with step □ 650, at which point the TPM can refuse to provide the key to the requesting process.

Figure 8 類似文のペア(対比) (類似度 70.21%)

369 One or more items are chosen by the computing device for concurrent display with the first and second axes that correspond to a first □ parameter in the first axis and a second □ parameter in the second axis (block 606).

537 One or more items are chosen by the computing device for concurrent display with the first and second axes that correspond to a first □ one of the parameters of the first axis and a second □ one of the parameters of the second axis.

Figure 9 類似文のペア(表記に差異) (類似度 70.45%)

Figure 10 2つの対訳文と明細書の比較

| 日本語明細書原文の対訳.docx                                                                                                                                     | 翻訳向けに修正した明細書の対訳.docx                                                                                                                                                 | 59.08% | 英文明細書.doc                                                                                                                                                                                     |
|------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 図1は、本発明の実施形態による計算機の制御装置の欠点を改善するための構成を示すブロック図である。                                                                                                     | 図1は、本発明の実施形態による、問題のない改善された制御装置を有する計算機を示すブロック図である。                                                                                                                    | 58.06% |                                                                                                                                                                                               |
| Fig.1 is a block diagram showing the construction to improve problems of the control unit of the computer as an embodiment of the present invention. | Referring a block diagram shown in Fig.1, a computer has an improved control unit which suppresses the problems according to an embodiment of the present invention. | 58.92% | Referring a block diagram shown in Fig.1, a computer has an improved control unit which eliminates or suppresses the described problems according to an embodiment of the present invention.□ |
| 本実施形態は、第1の例の制御装置を光CPUで構成した場合を示している。                                                                                                                  | 本実施形態は、第1の例において光CPUとして利用されている制御装置の改良を示している。                                                                                                                          | 51.85% |                                                                                                                                                                                               |
| The embodiment represents an example in which a control unit in the first example is constituted by an optical CPU.                                  | The embodiment represents a modification of the first example in which a control unit is used as an optical CPU.                                                     | 67.50% | The embodiment shows a modification of the first example, in which a control unit is used as an optical CPU.                                                                                  |

72 In operation □, □ the map □ may be utilized to route packets between the first endpoint and the second endpoint based on communications exchanged between the virtual presences within the virtual network overlay.

337 In operation, the map (e.g., utilizing the map □ 320 of FIG. 3) may be employed to route packets between the first endpoint and the second endpoint based on communications exchanged between the virtual presences within the □ overlay.

Figure 11 類似文のペア (動詞に差異) (類似度 60.86%)

### 7.2 日本語精査結果の例

米国特許明細書から日本語に翻訳された初校の明細書をチェックした実例を以下に掲載する。

① 陰影のある矩形部が接触し ▲ 始める。

- 所有を示す「の」以外の可能性:>>別の表現に書き換えてください:【陰影のある】⇒【陰影がある】
- 漢字・ひらがなの使い分けが不適切:【始める】⇒【はじめる】

Figure 12 助詞と仮名漢字の使い分け

① 制限が与えられたこの ① 現象は式11に示した ▲ ように説明がつかない。

- 「動作性名詞」に「機能動詞」を続ける表現は冗長である可能性があります:【制限が与え】
- 「は」の後に読点を打った方が分かりやすい可能性があります:【現象は】⇒【現象は、】
- 「～よみこ～でない」という表現を使用すると、複数解釈の余地を生じます:【ように説明がつかない】

Figure 13 日本語ルールからの指摘

相対的に ① 高い一定温度の水位が保たれるが、時間 ▲ がとともに変化する場合がある。

- 「形容詞+名詞+の+名詞」という表現は、形容詞の係り先が不明瞭になります:【高い一定温度の水位】
- 誤字・脱字の可能性:【かとともに】

Figure 14 日本語ルールと誤字可能性

計算装置によってユーザーインターフェースに出力される第1 ▲ の軸 ▲ の複数のパラメーターのうち第1のパラメーターである。

- 「の」を連続して使用しない:【の軸の複数のパラメーターの】
- 「の」を連続して使用しない:【の複数のパラメーターのうち第1の】

Figure 15 「の」の連続を指摘

## 8. 課題

ユーザー自身によって定義可能な辞書を用いて形態素解析を実行し、明細書を比較、精査した結果をいくつか示した。本システムが表示した結果のいくつかは、翻訳者に有用な情報を提供可能なことが検証された。しかし解析の限界もあり、不十分な結果もある。今後は以下のような技術を取り入れ、翻訳時の支援機能を拡張する。

### 産業日本語を対象とした精査ルールの追加

産業日本語やシンプリアイド・ジャパニーズなど、翻訳に適した明細書を作成するための、特許明細書向けの精査ルールを追加する。

### 類似文リストアップの精度向上

本手法は、文字列の類似性を利用しているので言語に依存しない利点がある反面、段落や文の情報量が小さい場合は、類似文とは判断されない不適切なペアを作る可能性がある。係り受けを利用した手法[17, 18]や、文と箇条書きなどの区別をするなど、同種の文体を判別してペアを作成する。

### 類似検索と処理の高速化

従来のキーワード検索だけではなく、任意の文の類似度検索も有効性があると思われる。大容量の明細書に対して類似度を高速に算出するために、形態素のフィルタリングおよび形態素を固定長で数値化することによってデータの一致判定速度を改善できる。更に、類似度閾値を下回ることが判明した時点で比較の実行を中止して次の比較をすることによって比較回数を減らすことができる。こうした性能改善方法が、複数の自治体の条例をすべて比較し、類似条例をリストアップする比較エンジンとして適用されてその効果を実証した。この手法を特許明細書データベースにも採用し、類似特許の内容検証をする。

## 9. おわりに

本システムを利用すると、一般的な環境上で翻訳作業中の2つ以上の文書比較、明細書中の類似文のリストアップ、明細書からの用語だけの抽出、用語辞書作成、引用箇所のチェック、および日本語精査が実行でき、特許翻訳時の品質を向上するためのいくつかの支援が可能であった。しかし、特許翻訳時には、本来克服すべき多くの課題がある。技術を理解し、法律的な配慮をして、自然で明確な外国語や日本語を記述しなければならない。特許の請求範囲を広くし、かつ、当事者が実施可能なまで具体的に技術を表現するといふ、相反する目標を両立させ、最終的に発明者にとって利益となるような翻訳が求められる。したがって翻訳品質は、その多くが翻訳者の力量による。しかし、残念ながら人間は参照番号の間違い、訳抜け、

用語の不一致、誤記などにしばしば気付かない。情報処理技術を用いた翻訳者に対する支援の1つに、明細書を多角的に処理して人間が実行しているチェックを代行し、翻訳者の負担を少しでも軽くすることがある。これによって翻訳者は、法的に翻訳期限がある中で思考に集中できる時間を少しでも確保できる。特許明細書の翻訳者がパーソナルツールとして手軽に利用できるような支援機能の充実、性能改善をしていきたい。

## 参考文献

- [1] 岩淵悦太郎編著 “悪文” 第三版 日本評論社
- [2] 永山 嘉昭編 “説得できる文章・表現 200 の鉄則” 日経 BP 社
- [3] 保険約款のわかりやすさ向上ガイドライン  
[http://www.sonpo.or.jp/about/guideline/pdf/index/yakkan\\_guideline.pdf](http://www.sonpo.or.jp/about/guideline/pdf/index/yakkan_guideline.pdf)
- [4] 石毛正純 “法制執務詳解” 新版 株式会社ぎょうせい 三版
- [5] 日本語スタイルガイド 第2版 一般財団法人テクニカルコミュニケーター協会
- [6] 記者ハンドブック: 新聞用字用語集 共同通信社, 2008
- [7] 亀谷 展 “多言語に特化した特許検索システム(仮称 atarikon)の構築” 第1回特許情報シンポジウム論文集 2010 年
- [8] 小倉 英里, 工藤 真代, 柳 英夫 “シンプリアイド・テクニカル・ジャパニーズ英訳を視野に入れて日本語を作る” 情報処理学会研究報告・デジタル・ドキュメント 2010-DD-78(5), 1-8
- [9] 産業日本語プラットフォーム  
<http://www.japio.or.jp/kenkyu/kenkyu01.html>
- [10] 浜口 宗武 “特許明細書翻訳の行方についての一考察” 日本知的財産翻訳協会発行 日本知的財産翻訳ジャーナル 33-34 号
- [11] <http://www.monjunct.ne.jp/ChawChaw>
- [12] Dan Jurafsky, James H. Martin “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”
- [13] 佐藤 亜古 “特許明細書翻訳～和訳の視点から～” 日本知的財産翻訳協会発行 日本知的財産翻訳ジャーナル 50 号
- [14] MeCab: Yet Another Part-of-Speech and Morphological Analyzer  
<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>
- [15] <http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz>
- [16] John Lafferty, Andrew McCallum, Fernando Pereira “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”
- [17] 工藤 拓, 松本 裕治 “チャンキングの段階適用による日本語の係り受け解析” 情報処理学会論文誌 Vol.43 No. 6 June 2002
- [18] 小林 幹門, 篠崎 政久, 加納 敏行 “係り受け構造アライメントを用いた文間の差異箇所検出” 情報処理学会第 74 回全国大会



## Session 4

技術調査のための特許情報抽出

## 技術調査のための特許情報抽出

原田 綾花<sup>\*</sup>(豊技大) 太田 貴久(豊技大) 小林暁雄(豊技大) 増山 繁(豊技大)  
野中 尋史(大分工業高等専門学校) 酒井 浩之(成蹊大学)

## Extraction of Patent Information for Technology Survey

Ayaka HARADA (Toyohashi University of Technology), Takahisa OTA<sup>\*</sup> (Toyohashi University of Technology), Kobayashi Akio (Toyohashi University of Technology), Shigeru MASUYAMA (Toyohashi University of Technology), Hirofumi NONAKA (Oita National College Of Technology), Hiroyuki SAKAI (Seikei University)

## 1. はじめに

研究開発や特許出願を行う際には、特許文書などを対象とした技術調査を行う必要がある。本研究では、この技術調査に必要な、発明の技術と効果に関する情報抽出を行う。ここで、技術は、発明を実現するための技術に該当し、効果は、発明が解決しようとする課題に該当する。特許情報を扱うユーザにとって、当該特許発明がどのような技術構成で、どのような効果が得られるものなのかを知ることは有用である。したがって、本研究では、これらの技術と効果に関する情報として、図1に示す4要素の抽出を行うことを目的とする。ここで、A. 技術の総称とは、発明全体を表す語のことをいう。技術の総称としては、例えば、物に関する特許であれば「音響装置」、方法に関する特許であれば「情報再生方法」などが挙げられる。

- A. 技術の総称
- B. 技術構成要素
- C. 構成要素の機能/作用
- D. 発明の効果

図1. 本研究で抽出対象の4要素  
Fig.1. Four elements to extract

図2は、消しゴム付きの筆記具に関する特許を対象に、上記に挙げた4つの要素をそれぞれ抽出した例である。Aによりユーザは発明の技術分野の概略を同定することができ、B、C、およびDを組み入れることで、各技術要素によってユーザにどのような利益が得られるかを知ることができる。なお、Cは概念的に技術構成要素と発明の効果との間に位置するものであり、Dのみならず、Cの情報も取得することで、ユーザは構成要素の機能がどのようにして発明の効果と結びついているのかを知ることができる。しかしながら、CとDを厳密に区別することは難しいため、以降では両者をまとめて発明の作用・効果として扱い、両者を抽出することとする。一方、発明の効果に該当する表現を取得する方法として、酒井らの手法[1]が挙げられるが、この手法では、まだ十分

な再現率を獲得できていない。したがって、本研究では、これら4つの要素のうち、Aの技術の総称と、CおよびDの発明の作用・効果の3つの要素を抽出する研究を行った。

- A. 筆記具
- B. 鉛筆と、消しゴムと、上記鉛筆の端と上記消しゴムの端とを連結する連結具を備える。
- C. 鉛筆等の芯の先と消しゴムとが触れないので、
- D. 消しゴムが汚れるのを防ぐことができる。

図2. 上記4要素の例

Fig.2. An example of the above-mentioned four elements

## 2. 従来手法

発明の効果に相当する表現(以下、効果表現)を抽出する手法として、西山らの手法([1],[2])や、Nanba[3]らの手法がある。西山らの手法[1]では、元来その製品または技術が持っている好ましくない点を抑えて特長とすることを示唆する表現(本書では特長対象と定義)と、「できる」や「向上する」などの発明の効果抽出するための手がかりとなる表現(以下、手がかり表現)を利用することで、効果表現を抽出していた。Nanbaら[3]の手法でも、特許文書中の「軽減」や「効果」などの手がかり表現を用いることで、技術、および、効果のラベル付けを自動的に行っている。石川らの手法[4]では、「ことにより」形式の文献を対象に、手段、および、効果の記述部分に現れる高頻度用語を用いて因果関係を抽出している。しかしながら、これらの手法では、手がかり表現は人手か、あるいは、半自動的に抽出しなければならない。また、体言のみからなる効果表現には対応できていない等、網羅性に欠けていた。

同じく、効果表現を抽出する手法である酒井らの手法[5]では、発明の効果タグに該当する文集合から、「ができる。」と「が可能である。」といった表現を、手がかり表現の種として与えることで手がかり表現を自動的に得ることができる。そして、西山らの手法[1]と同様に、抽出した手がかり表現を利用することで、効果表現を抽出

する。しかしながら、この手法では、例えば、効果に相当する表現の中でも、「抽出できる」や「増大可能である」などのように、助詞「が」のつかない表現や、「強化される」などのように、受身の表現となっている文は抽出することができず、まだ十分な再現率が得られていない。

したがって、本研究の作用・効果の抽出では、これらの表現を漏らさずに、すなわち、再現率よく取得する手法を提案し、その他の技術の総称の取得についても検討を行った。

### 3. 提案手法

#### 3.1 提案手法の概要

本研究では、特許請求項第一項の末尾の文末に技術の総称が書かれるという特徴を利用して技術の総称を抽出する手法を提案する。また、発明の作用・効果に関する内容については、主に、特許明細書中の「発明の効果」、「解決手段」、「課題を解決するための手段」の3つのタグに該当する文集合に書かれることが多い。ここで、タグというのは、図3に示すように、明細書に書かれている見出しのことをいう。図3のとおり、タグ名は通常、【タグ名】の形式で書かれる。しかしながら、これらのタグに該当する文集合には、作用・効果以外の内容が書かれることも多い。作用・効果以外の内容としては、例えば、図3に示すような、「本発明によって、以下のような効果が得られる。」といった、作用・効果の情報が含まれていない文のことを示す。したがって、これら3つのタグ中の文集合から、作用・効果以外について書かれている内容を省くことで、作用・効果を抽出する。

【解決手段】・・・ (省略) ...  
 【従来の技術】  
 従来の装置では・・・その後方位置にて歩行しながら各種作業を行うようにしている。  
 ... (省略) ...  
 【発明の効果】本発明によって、以下のような効果が得られる。  
 ... (省略) ...

図3. 特許明細書中のタグの例  
 Fig.3. An example of a tag in specifications

#### 3.2 技術の総称の抽出

本手法では、まず、技術の総称の抽出を行う。特許請求の範囲(請求項)は、書き手の書きやすい方法にしたがって記載されるのが通常である。しかしながら、誤解のない記載方法にするため、例えば図4のように、構成要素を先に列挙した上で最後に技術の総称を記載することが多い。したがって、技術の総称は、多くが各請求項の末尾の文末から取得できることが分かる(図4)。さらに、第一請求項は、どの請求項にも依存しないため、その発明の核となることが多い。実際、2002年度の特許における、300件の請求項第一項を対象に、末尾の文末が技術の総称になっているか否かを調査したところ、末尾の文末が技術の総称になっているものは94%であった。そこで、第一請求項の末尾の文末に出現する表現を、技術の総称として抽出する。

また、技術の総称の前には、技術の総称の特徴について述べた文との区切りを示す読点、または、技術の総称に係る、「する」や「含む」などの形態素列が頻出する(図5)。そこで、これらの形態素列を区切りとして、該当する文から技術の総称の抽出を行った。上記の手法で技術の総称を取得した結果、技術の総称の取得について、精度93%、再現率94%を達成することができた。

原始データを変換手段により別種のデータに変換して目的の種別のデータを得るようにしたデータ変換システムにおいて、前記原始データを受けて、その発生順序が特定できる情報とともに蓄積する。蓄積したデータは読出して前記情報を除去し、前記変換手段に与えると共に、この変換手段からの変換結果のデータは、原始データとの対応関係を特定する情報とともに保持する蓄積手段を備えたことを特徴とするデータ蓄積変換システム。

図4. 技術の総称の例 (下線:技術の総称)  
 Fig.4. An example of general names for technology (underlined parts : general names for technology)

|     |     |     |
|-----|-----|-----|
| 、   | なる  | 構成の |
| する  | 成る  | 設けた |
| いる  | できる | 優れた |
| させる | おける | 行う  |
| される | である | 持つ  |
| された | 備える | 有する |
| した  | 備えた | 含む  |

図5. 技術の総称の前に出現する形態素列  
 Fig.5. morphological sequences that appears before a general names for technology

### 3.3 発明の作用・効果の抽出

次に、発明の作用・効果の抽出を行う。特許請求項中には通常、発明の作用・効果は書かれないことから、前記3つのタグ中の文集合から、特許請求項と一致する内容の文/句を省くことで、作用・効果以外の内容の文/句を除去できると考える(図6)。また、請求項と一致する内容を省いた後の文/句集合から、除去されずに残った作用・効果以外の文/句を、SVMで分類を行うことにより省く。これにより、最終的に残った文/句から作用・効果を抽出する。この手法において、特許請求項と一致する内容を除く処理は現時点では未実装のため、人手で行った。

|                                                                                                                                       |
|---------------------------------------------------------------------------------------------------------------------------------------|
| <p>特許請求項</p> <p>…(省略)…続いて、<u>単語判定部で、英文字列の単語/非単語を文字列辞書を引いて判定すること</u>を特徴とする音声合成装置。</p>                                                  |
| <p>発明の効果タグ</p> <p>以下に、本発明の効果について述べる。請求項1にあるように、<u>単語判定部4で、英文字列の単語を文字列辞書7を引いて判定する。</u>こうして、<b>英文字列の文字長によらずに、上記英文字列の単語を正しく判定する。</b></p> |

図6. 特許請求項の例(上枠)、

発明の効果タグの例(下枠)

(下線:特許請求項の内容と一致する部分

太字:発明の作用・効果)

Fig.6. An example of patent claims (upper box), and an example of the tag of an effect of the invention (lower box)

(underlined parts : A part that is in agreement with the claim

bold parts : action and effect of the invention )

#### 3.3.1 発明の作用・効果の抽出の予備実験

そこで、まず、予備実験として、前記3つのタグを対象に、作用と効果に関する文/句、および、その他の文/句の出現する種類の数を調査した。その結果、図7に示す全8種類の文/句の項目に分類された(図7)。また、これらの項目について、前記3つのタグに該当する文集合中における、文字数の割合を調べた(図8,9,表1,2)。その結果、作用・効果に関する文/節は、解決手段、および、課題を解決するための手段タグに該当する文集合の

文字数においては約40%を占め、発明の効果タグにおいては74%と大部分を占めていることが分かった。

しかしながら、それ以外の項目として、(a)接続詞、文頭、文末や、(o)技術要素に関する項目は、解決手段、および、課題を解決するための手段タグに該当する文集合においては、それぞれ23%と33%、発明の効果タグに該当する文集合においては、それぞれ12%と1%と、(e)作用・効果ほどではないものの、各タグ中を占める割合が多いことが分かった。

#### ●発明に関する情報を含む文/句

(e) 作用・効果:

(例:強度を上げることができる。、汚れるのを防ぐ。)

(f) 作用・効果の条件など

(例:旅行中などに、)

(p) 従来例

(例:従来の携帯翻訳装置は、…といった問題点があった。)

(c) ハードウェア構成

(例:入力部は、マウス、キーボードを用いる。)

(d) 単語の定義

(例:ここで述語とは、…のことを指す。)

(t) その他、技術要素に関する項目

(例:モデルは、検索対象音声データに依存して作成する。)

#### ●発明に関する情報を含まない文/句

(a) 接続詞、文頭、文末など

(例:さらに、したがって、を示す。)

(o) 参照を示す文

(例:以下で説明する。、効果を次に挙げる。)

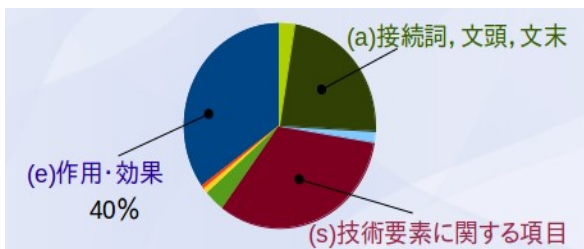
図7. 特許明細書におけるタグ中の文集合における項目

Fig.7. The item in the sentence set in the tag in specifications

表1. 解決手段,および,課題を解決するための手段タグ中の文集合における項目数と文字数

Table 1. The number of items and characters of each item in the sentence set in the tag of a solution and the tag of a means for solving the object

| 項目の種類          | 項目数  | 文字数   |
|----------------|------|-------|
| (e).作用・効果      | 288  | 11499 |
| (f).作用・効果の条件など | 6    | 194   |
| (p).従来例        | 3    | 275   |
| (c).ハードウェア構成   | 15   | 1146  |
| (s).技術に関する項目   | 197  | 10689 |
| (d).単語の定義      | 8    | 574   |
| (a).接続詞、文頭、文末  | 485  | 7673  |
| (o).参照を含む文     | 32   | 821   |
| 合計             | 1034 | 32871 |



- (e). 作用・効果
- (f). 作用・効果の条件など
- (p). 従来例
- (c). ハードウェア構成
- (s). その他, 技術要素に関する項目
- (d). 単語の定義
- (a). 接続詞、文頭、文末
- (o). 参照を含む文

図8. 解決手段,および,課題を解決するための手段タグ中の文集合における項目の文字数の割合  
Fig.8. The rate of the number of characters of each item in the sentence set in the tag of a solution and that means for solving the object

表2. 発明の効果タグ中の文集合における項目数と文字数

Table 2. The number of items and characters of each item in the sentence set in the tag of effects of the invention

| 項目の種類          | 項目数 | 文字数   |
|----------------|-----|-------|
| (e).作用・効果      | 443 | 17430 |
| (f).作用・効果の条件など | 3   | 202   |
| (p).従来例        | 3   | 147   |
| (c).ハードウェア構成   | 0   | 0     |
| (s).技術に関する項目   | 34  | 2521  |
| (d).単語の定義      | 1   | 131   |
| (a).接続詞、文頭、文末  | 148 | 2710  |
| (o).参照を含む文     | 15  | 326   |
| 合計             | 647 | 23467 |

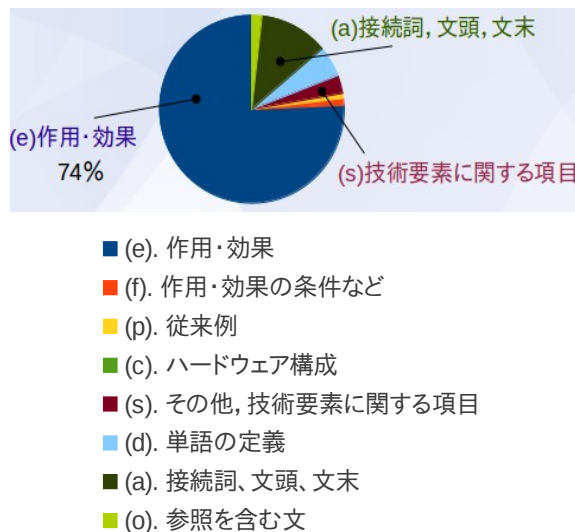


図9. 発明の効果タグ中の文集合における項目の文字数の割合

Fig.9. The rate of the number of characters of each item in the sentence set in the tag of effects of the invention

### 3.3.2 発明の作用・効果を表す項目の抽出の実験結果

正例を作用・効果に関する項目,負例を作用・効果以外の項目とし,SVM-Light で分類することにより,これらの項目を省く.データは,2002 年度の特許文書からランダムに選んだ 100 件を用いた.素性の単位は形態素とし,特徴ベクトルは,(a1)は形態素が出現する場合を1としたもの,(a2)は(a1)にさらに品詞情報を加えたもの,(a3)は形態素の頻度,(a4)は(a3)にさらに品詞情報を加えたものとして,それぞれの特徴ベクトルを用いた際の SVM の分類精度を調べた(図 10,11,表 3).

また,その際,SVMに入力する文は,各項目中の文,または,節とし,訓練データは  $1-n/10$ ,テストデータは  $n/10$  を用いて( $n$ :データ数),計 10 回 SVM を実行した.図 10 の結果より,(a1)と(a3)との間では,精度,および,再現率にあまり差は見られなかったが,いずれも品詞情報を付与することで,多くの場合,精度,再現率を上げることができた.特に(a4)の特徴ベクトルを用いた際は,「解決手段」および「課題を解決するための手段」タグに対しては精度 81%,再現率 73%,「発明の効果」タグにおいては精度 95%,再現率 95%を実現することができた.

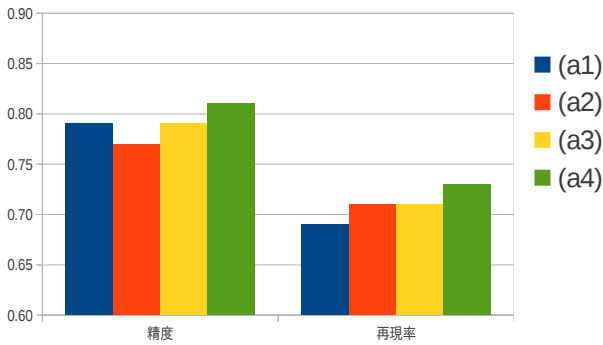


図 10. SVM による分類精度 (左), および, 再現率 (右)  
 (対象タグ: 解決手段, および, 課題を解決するための手段)  
 Fig.10. Classification accuracy( left ) and recall( right ) by SVM  
 (target tag : the tag of a solution and that means for solving the object)

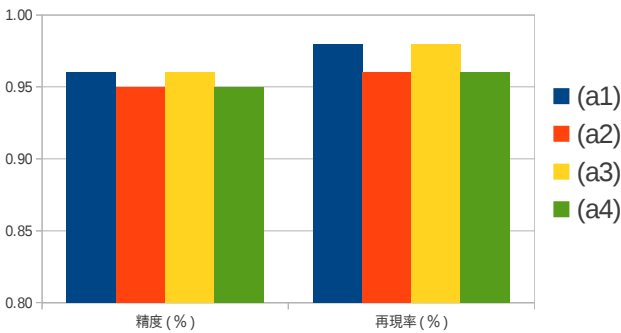


図 11. SVM による分類精度 (左), および, 再現率 (右)  
 (対象タグ: 発明の効果)  
 Fig.11. Classification accuracy( left ) and recall( right ) by SVM  
 (target tag : the tag of effect of the invention )

表 3. SVM による分類精度, および, 再現率  
 Table 3. Classification accuracy and recall by SVM

| 対象タグ                              |     | (a1) | (a2) | (a3) | (a4) |
|-----------------------------------|-----|------|------|------|------|
| 「解決手段」,<br>「課題を解決<br>するための<br>手段」 | 精度  | 0.79 | 0.77 | 0.79 | 0.81 |
|                                   | 再現率 | 0.69 | 0.71 | 0.71 | 0.73 |
| 「発明の効<br>果」                       | 精度  | 0.96 | 0.95 | 0.96 | 0.95 |
|                                   | 再現率 | 0.98 | 0.96 | 0.98 | 0.95 |

#### 4. 考察

表 4 は, 3.3.5 の実験において, 各項目の種類に対し, (a4) の特徴ベクトルを用いた際に誤って分類された数と割合を示している。これらの結果から, 特に精度が良かった項目としては, (a) 接続詞, 文頭, 文末, および, (o) 参照を含む文が挙げられる。これらの分類精度が良かった理由としては, 「そして」や「したがって」などの接続詞や, 「次に」や「以下」などの特有の語が多いことから, 分類が容易であったものと解釈できる(図 12~図 13)。一方, 誤りが多かったものとしては, (s) 技術要素に関する項目が挙げられる。この理由としては, 図 14 に示している例のように, 分類精度が良かった項目と比べ, 項目特有の語が少なく, 作用・効果と多くの語が共通であることから, 作用・効果への誤分類が多かったものと解釈できる。その他の, 精度が悪かった項目の理由としては, 事例の数がまだ少なく, 特徴を捉えきれていないためと考えられる。

表 4. SVM によって分類を誤った項目の数と割合  
 Table 4. The number and rate of an item which were accidentally classified by SVM

| 項目の種類           | 各項目の種類に対し, 誤って分類された数/総数 (割合) |
|-----------------|------------------------------|
| (e) 作用・効果       | 74/601 (0.12)                |
| (f) 作用・効果の条件など  | 7/9 (0.78)                   |
| (a) 接続詞, 文頭, 文末 | 4/525 (0.01)                 |
| (s) 技術要素に関する項目  | 52/140 (0.37)                |
| (o) 参照を含む文      | 2/46 (0.04)                  |
| (c) ハードウェア構成    | 6/11 (0.55)                  |
| (d) 単語の定義       | 3/8 (0.38)                   |
| (p) 従来例         | 4/6 (0.67)                   |

・また, この発明は,  
 ・さらに,  
 ・といったものがある。

図 12. 図 7 における項目(a)の例  
 Fig.12. An example of items of (a) in Figure 7.

- ・そのための方法を以下に示す。
- ・上記の通り、本発明には下記のような効果がある。

図 13. 図 7 における項目(o)の例

Fig.13. An example of items of (o) in Figure 7.

- ・SPS 音響モデルを、検索対象音声データ、ユーザ音声データのそれぞれに依存して作成する。
- ・請求項1に記載の機械翻訳装置においては、連続しているテキストが一文毎に分割されて翻訳単位として切り出される。

図 14. 図 7 における項目(s)の例

Fig.14. An example of items of (s) in Figure 7.

## 5. まとめ

本研究では、技術調査に必要な技術の総称、発明の作用・効果の情報を抽出することで、ユーザにとって有益な情報抽出を行った。その中でも、技術の総称を高精度に抽出することに成功した。また、発明の作用・効果を抽出する手法において、請求項と一致する項目を省いた後の文から、SVM による分類を行うことで、作用・効果を精度・再現率よく抽出できることが分かった。今後は、現段階では実現できていない、請求項と一致している内容を省く処理、および、技術構成要素の抽出について、手法を検討し、評価を行う。

## 6. 参考文献

- [1] 西山 莉紗, 竹内 広宜, 渡辺 日出雄, 那須川 哲哉, 武田 浩一:技術文書マイニングのための特長表現抽出, 第 22 回人工知能学会全国大会, pp. 3K3-2 (2008)
- [2] 西山 莉紗 他:未来技術動向予測のための技術文書マイニング, 第 21 回人工知能学会全国大会予稿集, No. 2H5-3 (2007)
- [3] H. Nanba, T. Kondo and T. Takezawa: Hiroshima City University at NTCIR-8 Patent Mining Task, Proceedings of NTCIR Workshop 8 Meeting, 2010.
- [4] 石川 大介, 石塚 英弘, 宇陀 則彦, 藤原 譲:特許文献における因果関係の抽出と統合, 情報知識学会誌, Vol. 14, No. 4, pp. 105-118 (2004)
- [5] 酒井 浩之, 他, 特許明細書からの技術課題情報の抽出, 人工知能学会論文誌, vol.24, no.6, pp.531-540, 2009.

————— 禁 無 断 転 載 —————

平成24年度AAMT/Japio特許翻訳研究会  
第2回特許情報シンポジウム 資料集

発行日 平成24年11月

発行 一般財団法人 日本特許情報機構 (Japio)  
〒135-0016 東京都江東区東陽4丁目1番7号  
佐藤ダイヤビルディング  
TEL:(03) 3615-5511 FAX:(03) 3615-5521

編集 AAMT/Japio特許翻訳研究会  
アジア太平洋機械翻訳協会 (AAMT)

印刷 株式会社 ナビックス