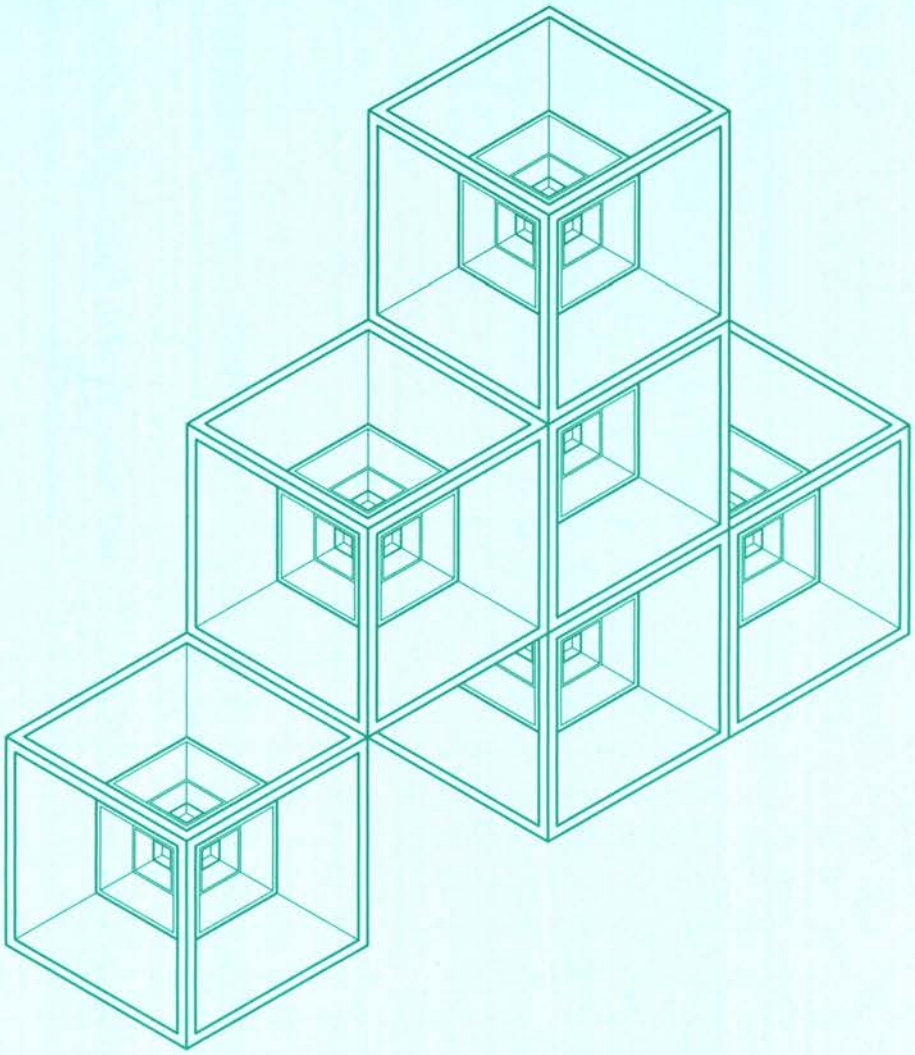
ISSN 1883-1818

AAMT Journal

Asia-Pacific Association for Machine Translation



June 2012 No.51

アジア太平洋機械翻訳協会

目　次

CONTENT

巻頭言

産業のグローバル化に伴う機械翻訳の必要性

株式会社高電社
代表取締役　岩城　陽子

創業以来 30 有余年、弊社は一貫して機械による言語処理技術の開発向上に研鑽を重ねて参りました。初期にはドットによる文字フォントの製作や、ワープロソフト等の開発を経て、1991 年に「日韓翻訳システム J・ソウル」を開発、発売致しました。（当製品は現在バージョン 8 まで展開中）

21 年を経て現在の翻訳精度の水準とは勿論比較に及びませんが、当時世界初の製品であったと自負致しております。この開発の原点は、創業者が少年時代に直面した「言葉の壁」であり、その壁をフリーにし「社会に資するものづくり」を目指す姿勢は、現在も変わることがありません。寄稿させていただくにあたり、近年の機械翻訳の目覚しい発展を支える開発研究者の挑戦、また貴協会を始めとする関係各位のご尽力に、心より敬意を申し上げます。

昨今、世界経済や人口の急速な流動化は、インターネットの普及と共にグローバル化に一層拍車をかけ、それに伴い機械翻訳の必要性や、応用分野、技術向上の重要性を痛感しております。

企業としての機械翻訳への取り組み

Ⅰ　開発・技術向上への取り組み

弊社開発室では長年機械翻訳技術の開発、向上に取り組んで参りました。この間、言語処理技術、言語資源、計算機パワー、インターネット環境、翻訳ニーズなど、機械翻訳を取り巻く環境が大きく変化したのに伴い、従来のルールベースに加え、用例翻訳、統計処理、コーパスよる専門用語辞書の半自動構築等々、様々な開発を実施して参りました。

しかし、言葉は世につれ人につれ（翻訳が難しそうですね）時代を反映する生き物であり変化を続けています。

弊社では先述の技術をベースに研究を重ねるとともに、日々得る顧客の要望や課題解決への挑戦が、翻訳エンジンの改良に繋がっています。

翻訳精度の向上には、大量のデータ収集と解析、未知語や時事用語、ユーザー辞書の拡充が必要であることに加え、都度、定性・定量評価を繰返し、品質向上の確認修正を行うなど、「熱意ある取組みの継続」が必須であることは言うまでもありません。

弊社の強みは、人力翻訳（人の手による翻訳）と機械による自動翻訳の両輪を擁していることです。弊社ではこの強みを活かし、翻訳精度向上やかかる時間と費用の削減のため、双方を融合させた「次世代翻訳機能の実現」を目指しています。これは全社プロジェクトとして取り組んでおり、幸いにも非常に有益な結果を得ています。この取組みは、今後も継続し、実用性も高まると思いますので、特許庁のプロジェクトに参加させて頂くなど、研究機関の皆様とも積極的にご協力させていただきたく願っております。

Ⅱ製品、サービス化への取り組み

技術も使われてこそ価値を生むものであり、顧客ニーズを満たす機械翻訳の活用を常に考えています。高機能端末や電子書籍の登場により従来のパッケージ版からインターネット翻訳が主流となって参りました。

弊社では 2000 年から多言語ソリューションによるサービスを開始しています。（ASP、エンジンレンタル）

○携帯電話翻訳　大手公式コンテンツとして採用
○ホームページ自動翻訳

○インターネット翻訳サイト

ポータルサイトでのサービス、EC サイトのグローバル化

○国立国会図書館文献検索等を実現する

さらに優れた汎用性の高い翻訳ソリューションによりユーザー様の利便性を高めて参ります。

今音声認識の実用化が注目されています。特に視覚障害者に喜ばれる技術であり、機械翻訳との相乗効果により役立つ製品、サービスが誕生するものと期待が高まります。

弊社のミッションは「言語の壁を超え、世界中の人々の心と心をつなぐ」です。国際ビジネスや多文化共生に貢献出来ますよう、今後も精進を重ねて参る所存です。

皆様のご指導をよろしくお願い申し上げます。

Project Report

European Patent Office and Japan Patent Office agree to co-operate on automated translation

Dr. Richard Flammer
Principal Director Patent Information
European Patent Office (EPO)

The President of the European Patent Office (EPO), Benoît Battistelli, and the Commissioner of the Japan Patent Office (JPO), Yoshiyuki Iwai, have signed an agreement on 6 February 2012 which will result in the enhancement of patent information services that incorporate machine translations of patents from Japanese into English and later into German and French. The body of Patent documentation published in the Japanese language is at once vast in its scale, and deep in the importance of the technical knowledge contained therein. This agreement represents a major step forward in opening up this information source to non-speakers of the Japanese language.

EPO President Battistelli further clarified that "Making Japanese patents available in English not only brings a wealth of technological information to users in Europe and elsewhere. It also offers an effective and reliable way for engineers, inventors and scientists to take account of the latest Japanese technology when defining their intellectual property strategies, and thus improve the focus and quality of their own work."

Open access to information is widely recognised as a corner stone of the innovation economy, and when it comes to public information, even a right of citizens. Therefore not only must the EPO fulfil its legal obligation for the publication of European Patent documents, but in the fast changing world of innovation and high technology the EPO must seize new opportunities as it strives to help Europe remain a leader in the global knowledge economy. Moving to a new level of machine translation is just such an opportunity where meaningful access to information, and especially the complex technical information found in patent documents, is greatly enhanced when it becomes available in a language with which users of the information are comfortable.

With the rapid improvements in "fit-for-purpose" (not legally binding) machine translation, President Battistelli has made support for innovation and the accessibility of the technical information contained in patent documents one of his top priorities. As President Battistelli explains "We have to find a balance to make patent information accessible to companies, especially SMEs, and lower the language threshold".

To this end, in October 2010 President Battistelli obtained the agreement of the EPO's Administrative Council to collaborate with national Patent Offices with the aim of raising the bar in machine translation. This was followed by another milestone in November 2010, when President Battistelli succeeded in bringing the technology giant Google onboard. The arrangements with Google are by no means exclusive, and the EPO looks forward to working with further suitable partners that can provide a valuable contribution.

Since the start of 2011, things have been moving very rapidly, and the benchmark of consistently "fit-for-purpose" machine translations is now close to reality for innovators and other users of patent information. Not only have the technical specialists continued to improve the underlying technology of machine translation, but a key new ingredient has been added - the incorporation of a massive corpora of "aligned" patent texts for training the translation systems. In essence these "aligned" patent texts, collected and aligned by the EPO, comprise patent documents that contain matching texts that were published in different languages.

With an eye to the future, the EPO continues to build up the corpora of aligned patent texts, and machine translation continues to evolve a relentless pace. Just take a look at the computers on your desk or your mobile phone. They have changed beyond recognition in the past 3 years. There is every reason to expect equally stunning changes in machine translation in the next 3 years. For 5 years down the road I will not even hazard a guess at the how much automatic translation quality will have advanced, and by how much translation costs will be reduced.

But of one thing we can be sure - "We will do everything we can to maintain our leadership in the use and development of machine translation technology for the good of the EPO and the European economy as a whole".

It is our hope that automatic, free-of-charge translation of patent documents which are consistently fit-for-purpose and provided through non-exclusive collaboration agreements such as that between the EPO and Google will form part of the standard expectations of patent information users within a few years. Furthermore, the EPO does not intend to stop at the language pairs (French to English, German to English etc) supported today. In a reflection of the great importance attached by both the European Union and the EPO to providing services to European citizens in their own language, the EPO is aiming at support for official languages of all 38 EPO member states.

Project Report

PATENTSCOPE CLIR
An empirical approach to applying SMT techniques to Cross Language Information Retrieval in the patent domain

Christophe Mazenc
World Intellectual Property Organization
34, chemin des Colombettes
CH-1211 Geneva 20
Christophe.Mazenc@wipo.int

Abstract

This paper presents the approach and the methodology used to implement WIPO's cross lingual information retrieval system for patents (subsequently called PATENTSCOPE CLIR). The system is freely available at http://www.wipo.int/patentscope/search/clir/clir.jsp.

The system allows a user to search is his native language (English, French, German, Japanese, Spanish, Chinese, Korean, Portuguese, Russian, Dutch, Italian and Swedish) and retrieve Patent documents published in other languages.

1 Introduction

WIPO is a specialized United Nations Agency responsible for promoting a balanced Intellectual Property system and one of its activities consists in disseminating Patent Information to facilitate knowledge transfer, see (Takagi and Czajkowski, 2012). In this context, WIPO puts a free patent search system called PATENTSCOPE (www.wipo.int/patentscope/search) at public disposal offering advanced search functionality for internationally published Patent Cooperation Treaty (PCT[1]) applications and for a growing number of national patent collections. At the time of writing, 10 million patent records can be searched, among which 4.5 million have description and claims searchable in full text[2].

A major share of these patent documents were published in English but the world patent landscape is shifting and more and more patents are published in other languages, with a growing share being in Asian languages (Chinese, Japanese and Korean, referred to as CJK). As the grant of a patent in many jurisdictions requires that it fulfills the criteria of novelty and/or inventive step, the need to search in many different languages is rising worldwide and WIPO developed the PATENTSCOPE CLIR system among other tools and services, see (Pouliquen and Mazenc, 2011) in response.

2 Background

Cross Lingual Information Retrieval has been an active research area in the field of Computational Linguistics since the mid 1990s (see, notably, the proceedings of the Text Retrieval Conferences (TREC[3]), of the Cross-Language Evaluation Forum (CLEF[4]) and the Japanese NTCIR workshops[5]). Research results have shown that it is possible to build multilingual search systems that reach monolingual search system performance in terms of recall and precision. See (Oard, 2009)

[1] WIPO. (2010) PCT The International Patent System-WIPO Publication No. 901(E)/09 June 2010

[2] Most Patent applications have only the title and a small abstract available in full text while the description and claims are usually more informative.

[3] http://trec.nist.gov/

[4] http://www.clef-initiative.eu

[5] http://research.nii.ac.jp/ntcir/outline/prop-en.html

and (Nasharuddin and Abdullah, 2010) for recent surveys of the state of the art.

Projects that have been successfully developed use one of the two following paradigms:

(1) Document translation: the documents are all translated into English and their translations indexed together with the documents written in English. Cross lingual searches are then achieved by running monolingual English queries.
(2) Query translation: the user enters his search keywords in his native language which are then translated into all languages. All document collections are searched in parallel; the results are then ranked and presented to the user in his own language using Machine Translation.

Each paradigm has advantages and drawbacks discussed briefly later.

PATENTSCOPE CLIR uses the second paradigm and leverages Statistical Machine Translation techniques that led to the development of tools like Google*Translate in the mid 2000s. PATENTSCOPE CLIR entered production in May 2010 initially covering five languages: English, French, German, Japanese and Spanish. Support for Chinese, Korean, Portuguese and Russian was added in July 2011 and for Dutch, Italian and Swedish in January 2012. As will be explained later, rapid development of the system and easy addition of new languages are possible as the approach is corpus based, statistical and unsupervised, and as the required corpora were available at WIPO.

The implementation of CLIR systems in the patent domain comes up against specific difficulties, see (Sarasúa, 2000):

- the wide range of technical keywords with many variations
- the ambiguity of technical keywords in different technical domains
- the vagueness and imprecision of technical keywords used to describe matters as generally as possible so as not to restrict the scope of protection sought by the patent applicant

To give an example, a jackhammer was found to be described as a "device for breaking the surface of roads" in a patent application published in German ("Strassendeckenaufbruchgerät").

3 Presentation of the system

We will start by presenting the interface of our system and with a brief explanation from a user's point of view.

3.1 Navigation

The PATENTSCOPE system can be accessed from the WIPO main site by using the "WIPO GOLD" [6] portal, an assembly of all WIPO's collections of searchable IP data. Clicking on PATENTSCOPE opens the simple search screen of PATENTSCOPE and you can navigate to the CLIR system by selecting the entry "Cross Lingual Expansion" in the Search drop down menu (See Figure 1). The interface is available in Chinese, English, French, Japanese, Korean, Portuguese, Russian and Spanish.

Figure 1: Navigation to PATENTSCOPE CLIR

3.2 Automatic expansion mode

A first Graphical User Interface, the so-called "Automatic expansion mode", offers the user the possibility to search for terms without any further input as shown in Figure 2. This mode is designed for the general public not wishing to drill into the complexity of multilingual query expansion and wanting quick search results.

[6] http://www.wipo.int/wipogold/en/

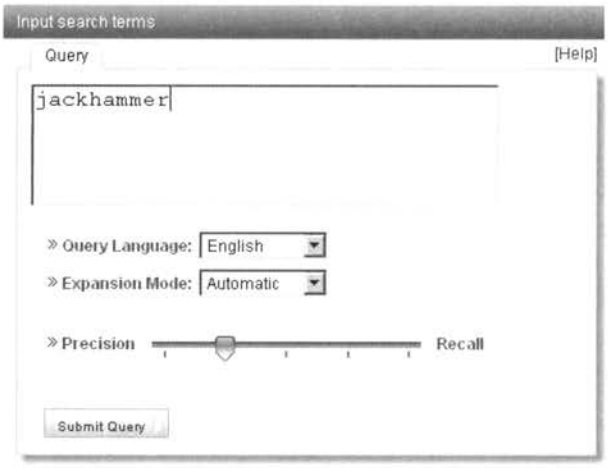


Figure 2: performing a quick multilingual search

Behind the scene, the system finds the technical domain(s) associated with the keywords, chooses relevant synonyms and translations in these domains, builds the multilingual Boolean query and executes it to give immediate results. In many cases where the entered keywords are technically precise, belong to a single technical domain and are non ambiguous, this provides excellent results. If the user judges that he does not get enough relevant results, he can execute the query again by requesting a greater recall using the displayed sliding bar. On the other hand, if the user judges that there are too many irrelevant results, he can execute the query again by requesting more precision using the same sliding bar. Indeed, the parameter indicated by the sliding bar sets the system a maximum level of freedom it can take in selecting close synonyms and translations for the entered keywords, enlarging and narrowing the scope of the search as required.

The expanded query in our example is shown below:

Results **1-10** of **575** for Criteria: **(EN_TI:("jackhammer" OR "breaking hammer") OR EN_AB:("jackhammer" OR "breaking hammer")) OR (DE_TI:("Bohrhammer" OR "Schlaghammer" OR "Brechhammer" OR "Brecherhammer") OR DE_AB:("Bohrhammer" OR "Schlaghammer" OR "Brechhammer" OR "Brecherhammer")) OR (ES_TI:("martillo rompedor") OR ES_AB:("martillo rompedor")) OR (FR_TI:("marteau de démolition" OR "marteau piqueur") OR FR_AB:("marteau de démolition" OR "marteau piqueur")) OR (RU_TI:("отбойный молоток") OR RU_AB:("отбойный молоток")) OR (ZH_TI:("拆除") OR ZH_AB:("拆除"))** Office(s):**all** Language:**EN** Stemming: **true**

We can see by examining the expanded query that the system searches in both titles and abstracts and has found translations in five other languages. 575 results have consequently been found and the first relevant results illustrate clearly the multilingual nature of the search, as seen below.

No	Ctr	Title
1.	EP	0790881 - HYDRAULISCHER BRECHHAMMER
2.	RU	02345881 - PNEUMATIC JACKHAMMER
3.	WO	WO/1996/015881 - HYDRAULIC BREAKING HAMMER
4.	EP	1722933 - HYDRAULISCHER BRECHHAMMER MIT GESCHMIERTER WERKZEUGFÜHRUNGSHÜLSE
5.	WO	WO/2008/139039 - BREAKING HAMMER AND METHOD OF FASTENING THE SAME
6.	ES	2297619 - MARTILLO ROMPEDOR
7.	EP	1722931 - HYDRAULISCHER BRECHHAMMER
8.	WO	WO/1995/022442 - HYDRAULISCHER SCHLAGHAMMER
9.	SU	01813177 - Отбойный молоток
10.	ES	2296139 - MARTILLO ROMPEDOR HIDRAULICO.
11.	ES	2331707 - ACCESORIO AMOVIBLE PARA EQUIPAR UN MARTILLO ROMPEDOR
12.	ES	1009917 - UTIL PERFORADOR PARA MARTILLOS ROMPEDORES
13.	EP	0884138 - Bohrhammer

Figure 3: Example of multilingual result list

To get an idea of the increase in recall that can be obtained in some cases, at the time of writing the corresponding monolingual query "jackhammer" retrieves only 24 results in PATENTSCOPE when searching patent titles and abstracts.

Finally, the user can request PATENTSCOPE to machine translate the titles of the results list into his native language (through integration with web MT engines like Google Translate or Bing Translate).

Expert users who wish to be more certain of the quality of the search results are invited to use the "Supervised mode", revealing in more depth the way the system works internally and allowing the user to drive the expansion process taking into account the context of their search.

3.3 Supervised expansion mode

We will reuse the example of "jackhammer" to explain the use of the supervised expansion mode. In a first step, the system asks for confirmation of the automatic selection of technical domains associated with the search terms entered by the user. PATENTSCOPE CLIR uses 31 domains which were defined using earlier work on terminology for PCT abstracts and which are mapped to the International Patent Classification[7]

[7] http://www.wipo.int/classifications/ipc/en/

(IPC) as shown in Table 1.

Technical domain	IPC mapping
[ADMN] Admin, Business, Management & Soc Sci	G09B
[AERO] Aeronautics & Aerospace Engineering	B64, C06B
[AGRI] Agriculture, Fisheries & Forestry	C05, C12N
[AUDV] Audio, Audiovisual, Image & Video Tech	G03, G09G, G09F, G10L, G10K, G11, H04
[AUTO] Automotive & Road Vehicle Engineering	B60, B62
[BLDG] Civil Engineering & Building Construction	B28, B66, C04, E0, F17
[CHEM] Chemical & Materials Technology	B01-05, B08, B28C, B82, C0-1, C25, C3-4, C9
[DATA] Computer Sci, Telecom & Broadcasting	H04, G11, G09C, G08, G06
[ELEC] Electrical Engineering & Electronics	H01-03, H05
[ENGY] Energy, Fuels & Heat Transfer Eng	B01B, F16P, F17, F22,
[ENVR] Environment & Safety Engineering	B03, B08-09, C02, F27, G08
[FOOD] Foods & Food Technology	A21-24
[GENR] Generalities, Language, Media & Info Sci	B03, B67, C12-13
[HOME] Home Contents & Household Maintenance	A45, A46B, A47, E05-06
[HORO] Precision Mechanics, Jewelry & Horology	A44C, B81, G04, G12
[MANU] Manufacturing & Materials Handling Tech	A46D, B07, B82B 3/, B23-29, B3, B65-66, C03-04, C06F 1/, C06F 3/, C14
[MARI] Marine Engineering	B63
[MEAS] Standards, Units, Metrology & Testing	C12Q, F16, G01, G05, G12
[MECH] Mechanical Engineering	B02-07, B25-26, B30, E02, F0, F15-16, F26
[MEDI] Medical Technology	A61, B82, C12P, C12N
[METL] Metallurgy	B21-22,C2, F27
[MILI] Military Technology	C06C, C06B, C06D, F4
[MINE] Mining, Oil & Gas Extraction & Minerals	E21, F42
[NANO] Nano Technology	B81-82
[PACK] Packaging & Distribution of Goods	B31-32, B65
[PRNT] Printing & Paper	B30B 3/, B31, B41-43, D21, G07B
[RAIL] Railway Engineering	B61, G09D
[SCIE] Optical Engineering	G02
[SPRT] Sports, Leisure, Tourism & Hospitality Ind	A63, F41B
[TEXT] Textile & Clothing Industries	A41, A43, A44B, B68, D0, F26
[TRAN] "Transportation"	G07-08

Table 1: PATENTSCOPE CLIR domains

For the purposes of the demonstration, we would like to search for jackhammers in a relatively general sense.

The system proposes the technical domain [MANU] which is fine and we also add the relevant technical domains [BUILD] and [MECH] (Up to five technical domains can be selected for the query expansion).

In the next step, the system searches for all the terms comprising a sub-list of consecutive keywords in the query entered by the user.

For each found term, the system then proposes a list of variant terms that the user needs to select or unselect. The user can also change the technical domains associated with the current term (allowing a query with terms from several different domains), add his own variant terms, or specify that the current term should not be translated in the expansion.

Figure 4: Technical domain(s) selection

By choosing different values on the sliding bar, more or less variants are displayed. As the unselected variants are used to disambiguate, the user needs to review all the proposed variants. In the example, we select all proposed variants ("paving breaker", "impact hammer", "rotary demolition hammer"), which are relevant.

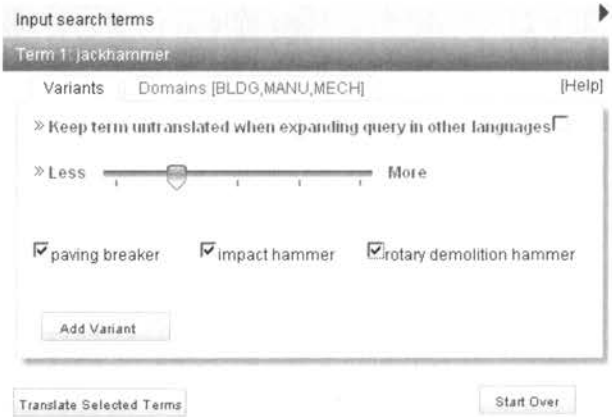


Figure 5: Selection of variants

In the next screen, the system finally displays the expanded queries in all target languages.

At this point in time, the user can influence the final query generation with three factors:

- Setting stemming[8] on or off. Stemming on is recommended to increase the recall
- Selecting the acceptable distance between words matched in the retrieved documents (proximity searching)
- Choosing on which fields the query should be performed (title, abstract, description, claims).

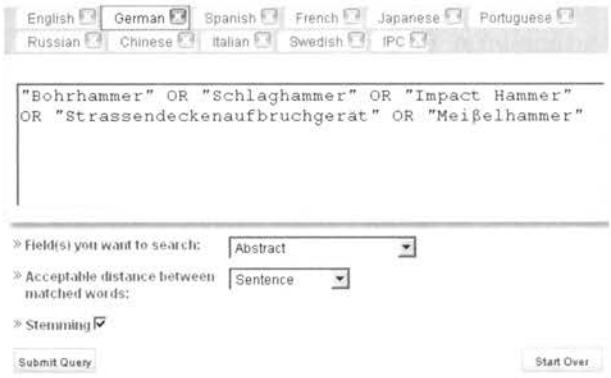


Figure 6: validation of language expansions

If the user wishes to search in description and claims, it is advised to use proximity searching so that document hits, in which related keywords are found far apart in the text, are discarded.

The user also has full editing rights on the expanded query for each language and can select or deselect the IPC filter computed according to the technical domains selected in the first step. Expanded queries use also the LUCENE syntax and can easily be copied, pasted and eventually rewritten to run on patent search systems other than PATENTSCOPE.

Results **1-10** of **940** for Criteria: **((DE_TI:(((Bohrhammer OR "Impact Hammer" OR Schlaghammer OR Strassendeckenaufbruchgerät OR Meißelhammer))) OR DE_AB:(((Bohrhammer OR "Impact Hammer" OR Schlaghammer OR Strassendeckenaufbruchgerät OR Meißelhammer)))) OR (EN_TI:((("rotary demolition hammer" OR jackhammer OR "paving breaker" OR "impact hammer"))) OR EN_AB:((("rotary demolition hammer" OR jackhammer OR "paving breaker" OR "impact hammer")))) OR (ES_TI:((percusión)) OR ES_AB:((percusión))) OR (FR_TI:((("marteau à percussion" OR "marteau de démolition" OR "marteau piqueur"))) OR FR_AB:((("marteau à percussion" OR "marteau de démolition" OR "marteau piqueur")))) OR (IT_TI:(("martello a percussione")) OR IT_AB:(("martello a percussione"))) OR (JA_TI:((("インパクトハンマー" OR "打撃ハンマー"))) OR JA_AB:((("インパクトハンマー" OR "打撃ハンマー")))) OR (PT_TI:((("martelo de impacto" OR "martelos de choque"))) OR PT_AB:((("martelo de impacto" OR "martelos de choque")))) OR (RU_TI:((("отбойный молоток" OR "мельница ударного действия"))) OR RU_AB:((("отбойный молоток" OR "мельница ударного действия")))) OR (SV_TI:(((slaghammares OR slaghammare))) OR SV_AB:(((slaghammares OR slaghammare)))) OR (ZH_TI:((("拆除" OR "路面破碎机用" OR "或镐" OR "或冲击锤" OR "撞击式锤片"))) OR ZH_AB:((("拆除" OR "路面破碎机用" OR "或镐" OR "或冲击锤" OR "撞击式锤片"))))) AND ICF:("B82B 3" OR "C06F 1" OR "C06F 3" OR A46D OR B02 OR B03 OR B04 OR B05 OR B06 OR B07 OR B23 OR B24 OR B25 OR B26 OR B27 OR B28 OR B29 OR B30 OR B3? OR B65 OR B66 OR C03 OR C04 OR C14 OR E02 OR E0? OR F0? OR F15 OR F16 OR F17 OR F26)** Office(s):**all** Language:**EN** Stemming: **true**

In our example, we see that using the supervised mode, we obtained many more adequate synonyms and translations (E.g. the French term "marteau à percussion"), leading to an improved recall when searching titles and abstracts with 940 documents found compared to the unsupervised mode with 575 documents found and to the original monolingual query with 24 documents found.

4 Implementation approach

4.1 Principles

The initial idea at the beginning of the project arose when looking at phrase tables generated by a first trial of the open source toolkit Moses (see Koehn et al., 2007) on titles of PCT applications published in both English and French. The obtained phrase tables were indeed showing very promising associations of technical terms in English and French.

[8] See http://en.wikipedia.org/wiki/Stemming

PATENTSCOPE CLIR is an implementation of an SMT decoder generating patent search Boolean queries. The main difference to the Moses decoder is simplicity, it relies mostly on in-memory Machine Readable Dictionaries (MRDs), and the fact that it seeks to find all alternative variants of words in order to improve recall instead of focusing on correct translations (no reordering or natural language sentences generation with the help of mono lingual language models are needed).

The second idea is to use English as a pivot language. Indeed, most of the available translated patent titles are from or to English and another language.

The third idea is to use a feature of the Moses toolkit called "Factored Translation Models" (see Koehn and Hoang 2007), to implement technical domain adaptation according to the domains defined in Table 1. As a result, the statistical pairing in the phrase tables are done in the context of their technical domain.

The fourth idea is to involve the patent searcher in order to drive the query expansion. Statistical pairings of groups of words in phrase tables indeed come with a probability parameter informing of the likelihood of the pairing. Those probability parameters are stored in the MRDs, allowing the definition of varying thresholds for selecting only the best translations and synonyms and ranking them. The patent searcher has control over the thresholds by using the displayed sliders.

4.2 Challenges and solutions

Preparation of the MRDs

The preparation of the MRDs is performed according to the following steps:

Step 1: Source data gathering: the patent application titles are gathered in a relational database, so that they can be cleaned and exported in the format suitable for the training of the Moses phrase tables. Our title database includes around 45 million unique titles for around 80 million different patent applications.

Step 2: Cleaning: this step is language specific and involves language checking, bad character filtering, character standardization, recasing, accents generation and tokenization for Asian languages. We use open source tokenizers from the LUCENE (see McCandless et al., 2010) sandbox and created our own Korean tokenizer.

Step 3: Export to Moses format: this step involves pairing cleaned titles, associating technical domains and writing them in the Moses training input format. Paired translated titles for the same patent application are selected in priority. Unfortunately, the obtained volume of parallel titles is not sufficient for most of the languages. It is therefore necessary to resort to family patent relationships[9], see (Simmons, 2009). However, titles for patents belonging to the same family can be far from exact translations, leading to erroneous statistical pairings of translated terms. It is therefore necessary to filter the pairings to keep only the most probable ones, which is achieved by computing matching scores using previous MRDs built from exact translations (bootstrapping).

Step 4: Moses training: this step is the usual Moses training step with factored models.

Step 5: MRD generation: this step involves parsing the titles phrase tables and generating the dictionaries. It appears that paired translated terms are often prefixed or suffixed by parasitic words in the phrase tables. Those words are for example in English stop words like "the", "a", "by", "for", "and"… and also words like "apparatus", "compound", "comprising", "consisting", "improved", "produced", … These parasitic words are simply lost by translating from one language to the other where they are unnecessary. An empirical approach has been used to identify and list them so as to remove them as prefixes and suffixes when adding entries in the MRDs. For the less common languages, title phrase tables obtained as described above may be insufficient to produce a large enough dictionary. In these cases, phrases tables obtained from paired abstracts from WIPO's machine translation project are also used (see Pouliquen et al., 2011) but in order not to pollute the MRDs with groups of words not being technical terms, entries are added to the MRD only if the English group of words already pre-exists in the MRD and was obtained from a title

[9] A patent family is a single invention patented in more than one country. Usually the titles are translations of each others but they sometimes differ

phrase table.

Source language of the MRD	Number of entries(*)
English	17'350'000
French	8'497'000
Chinese	6'004'000
Russian	3'096'000
German	2'986'000
Korean	1'831'000
Spanish	1'730'000
Japanese	1'589'000
Portuguese (Brazil and Portugal)	680'000
Italian	632'000
Swedish	613'000
Dutch	439'000

Table 2: Volumetric of MRDs (January 2012)

(*)Number of entries per language are reported, an entry being a group of words in the source language that has been at least matched statistically with another group of words in another language (after removal of the parasitic prefix and suffix words). It is to be noted that these entries include technical terms but that a large percentage of them are not in the sense they would not be selected to be recorded in a terminology database.

Automated Language detection

Step 2 of the MRD generation uses automated language detection. Also when a user enters keywords for a CLIR expansion, the source language has to be known for certain in order to use the appropriate MRD. Automated language detection is implemented by computing a score for all supported languages based on word frequencies using the MRDs and using the character sets for CJK languages and Russian.

Automated technical domain detection

This is needed for the implementation of the unsupervised mode. It is achieved by computing a score per technical domain based on the probabilities stored in the MRDs and returning the technical domains ranked by score.

Word variants generation

Word variants are generated by using a pivot language and translating back into the source domain. For all languages but English, English is used as the pivot language and for English, French is used. The translations of translations are done in the context of the selected technical domains and the results are sorted to remove duplicates and ranked. Thresholds and filtering heuristics are applied to keep only the best variant candidates.

Another important aspect of variants is that missing variants can be added by expert patent searchers. They also need to indicate to which technical domain their added variant applies. As a consequence, the added variant will appear in the subsequent variants proposals for all users, adding a crowdsourcing component to the system that can dramatically improve the recall of specific searches.

Out Of Vocabulary (OOV)

The consequence of the Out Of Vocabulary problem in the case of PATENTSCOPE CLIR is that potentially relevant documents in a target language will not be returned by the expanded query. This is therefore less serious than in Machine Translation. PATENTSCOPE CLIR allows the user to indicate that the source term be added untranslated as-is in the translated queries, which solves some cases (like acronyms or brands).

Also, when the source language is English, an OOV means in practice that the technical term mostly does not appear in an English translated title for the target language, which indicates that probably not many relevant documents are missed (as the CLIR MRDs have been generated with the titles of the major patent collections in the world which are for the most part also translated into English).

When the source language is not English, it means indeed that relevant documents will probably be missed. Efforts are then made to improve the coverage of the MRDs and keep them up-to-date with the new terminology surfacing in the recently published patents, by gathering the new titles and regenerating from scratch all the MRDs once every six months.

Disambiguation

For polysemic keywords entered by the user, disambiguation aims at selecting the right meaning in the context of his/her search when selecting variants and translations. For example "pet" can mean "animal companion", "polyethylene terephtalate" or "positron emission tomography".

Disambiguation is achieved first by selecting correctly the technical domains associated with the search context. In the example, the "animal

companion" meaning can be associated with the [HOME] technical domain, "polyethylene terephtalate" with [CHEM] and/or [PACK] and "positron emission tomography" with [MEDI].

This will not solve all cases. CLIR implements a second disambiguation technique in supervised mode. Indeed, different meanings can be characterized by different variants. When a variant is not selected by the user, the system identifies its most probable translation in each target language. Then, if this translation appears in the translation proposals of the initial keywords or of the variants selected by the user and if it is not ranked as best probable translation, it is then removed from the list of candidate translations, before the generation of the final Boolean query.

When noticing irrelevant documents retrieved due to an unwanted meaning of a polysemic keyword in a language different to the source language, the expert user has also the possibility to edit the corresponding expansion to remove the irrelevant translations.

Precision loss

When executing a CLIR query, it happens from time to time that a large number of irrelevant documents published in the same target language appear in the result list. This is caused by the statistical nature of the building process of the MRDs and by the mistakes in the used corpora. As a result, a keyword in a source language is translated into the defective target language with a term that has a broader meaning and even sometimes an unrelated meaning. In these cases, the user is invited to regenerate his expanded query using the CLIR supervised mode and check and correct the query expansion in the defective target language. This can easily been achieved even if the user does not understand the target language by copying and pasting the defective language expansion in Google Translate, which allows to easily identify the defective translated term that has to be eliminated.

It is also worth noting that in these cases, defective translated terms are often exceptions when compared to relevant translated terms in the final expanded query. As a result, due to the LUCENE ranking algorithms on multiple fields searched concurrently, irrelevant documents usually are not well ranked and appear at the bottom of the results list, which is a lesser evil.

Finally, as the approach is corpus based, the system does not propose out of domain synonyms and translations that would harm the precision. For example, expanding "device coupling" will never propose variants like "wedding" or "reproduction", a "computer fan" will not be expanded and translated as a "geek" but as a blower or cooler. This is also valid for all languages where idioms are used in technical terms like "chien de fusil" in French that is correctly expanded in English as "rifle hammer" and not "rifle dog".

Boolean query generation

The last step in the CLIR expansion procedure concerns the Boolean query generation (following the standard LUCENE query syntax with the PATENTSCOPE index field names). CLIR generates expanded queries using two formats:

- *AND* queries
- Proximity search queries

PATENTSCOPE text field names are as follows:

LL_FF

Where LL is the two letters code of the language (DE, EN, ES, FR, IT, JA, KO, NL, PT, RU, SV, ZH) and FF is the field code: TI for Title, AB for Abstract, DE for Description and CL for Claims.

The assumption is that the keywords entered by the user are compulsory in the retrieved documents. AND queries ensure this, while proximity search queries also ensures that the match keywords are within a range of consecutive words.

Let us consider that the user has entered three keywords A, B and C and selected A^1 for A and C^1 and C^2 for C as variants. The expanded *AND* query in the source language is then as follows:

$$((A \text{ OR } A^1) \text{ AND } B \text{ AND } (C \text{ OR } C^1 \text{ OR } C^2))$$

It is then combined with the OR operator with a language sub-query for each target language:

$$((A_{tl}^1 \text{ OR } A_{tl}^2 \text{ OR } A_{tl}^3) \text{ AND } B_{tl}^1 \text{ AND } (C_{tl}^1 \text{ OR } C_{tl}^2))$$

where A_{tl}^1 is the best probable translation candidate for A or A^1 in target language tl, A_{tl}^2 is the second best probable translation candidate for A or A^1, and so on.

Proximity queries use LUCENE phrase queries where keywords to be found are specified between quotation marks, followed by a tilda and

the maximum allowed proximity expressed in number of words:

The expanded source language query is then obtained by developing all the combinations of variants as follows and the same applies for the target languages:

"A B C" ~5 OR "A B C^1" ~5 OR "A B C^2" ~5 OR "A^1 B C" ~5 OR "A^1 B C^1" ~5 OR "A^1 B C^2" ~5

The size of the expanded queries becomes very long as soon as there are several source keywords with many variants and translations and several indexed fields searched concurrently.

As a result, the system keeps only the combinations with the most probable variants and translations in the final query.

Language specific difficulties

In the implementation of this project, we were faced with specific language difficulties that are briefly presented below:

- Missing accents: most of the patent titles are available only in capital letters without accents, which is a particular problem for the Romance languages (French, Spanish, Italian).
- Lost casing: the fact that patent titles are in capitals is a problem with German where nouns are mixed with other words.
- Synthetic/Agglutinative languages: in these languages, new words are built by "sticking" smaller words together. This is the case for German, Dutch, Swedish and Korean. This causes OOV problems for very unusual combinations.
- Japanese and Chinese require tokenizers to set word boundaries. The possibility to pre-tokenize the searched keywords by adding spaces where need be is offered to the user.

Finally, the full support of a language in PATENTSCOPE CLIR requires that a corresponding LUCENE stemmer be available.

5 Results/evaluation

We discuss in this chapter the advantages and drawbacks of our system and present the pre-production test cases we defined and their results that authorized the deployment in production.

5.1 Advantages of the PATENTSCOPE CLIR approach

First of all, the system is not English centric in the sense it does not require end users to understand English and it effectively supports query translations from any of the 12 supported languages to the 11 others.

Secondly, the system is easily understandable by end users, it is WYSIWYG (there is no black box effect) and users have full control on the expansion mechanism and full editing rights on the results.

Thirdly, its development required a manageable amount of resources, significantly less than that required for a document translation approach.

Fourthly, due to the unsupervised training of the MRDs, the system is easy to maintain and can be updated regularly to take benefit of the new terminology appearing in recently published patent applications, of the quality correction applied to previous patent applications and of the quality progresses in the statistical aligning algorithms employed.

Fifthly, the system exhibits good disambiguation capabilities due to the adaptation to technical domains and to the involvement of the user in the query expansion process.

Sixthly, in most cases the system exhibits large recall gains compared to the associated monolingual queries due to its ability to propose relevant and in-domain variants (e.g. common typos) and to the increased scope of searched documents (monolingual documents in a language different than the source keywords are included in the search scope).

Finally, we believe the approach is generic and can be used in domains different than patent information where large parallel corpora are available.

5.2 Drawbacks of the PATENTSCOPE CLIR approach

First of all, the MRD linguistic coverage for a language depends heavily on the size of the preexisting corpus of translated titles that is available. CLIR cannot be implemented for under-represented languages. The system

therefore exhibits quality disparities among the languages.

Secondly, the generated query expansions are not free of noise. In these, hopefully not too frequent, cases, the user can, if necessary, amend the query expansion themselves in the target languages to remove the noise in the results, as explained in paragraph 4.2.6, especially if they wants to reuse their queries.

Thirdly, the generated queries are very heavy (they can contain dozens if not hundred of keywords) and therefore a very powerful search engine with adequate hardware sizing is necessary to execute them in order to maintain good response times.

Fourthly, while the maintenance of the system is relatively easy, painstaking cleaning and filtering procedures have to be put in place to cope with the quality deficiencies of the available patent titles, especially for the older titles.

Fifthly, due to the fact that English is used as a pivot language and has the largest MRD, multilingual expansion quality is better when starting from English.

5.3 Pre-production testing methodology and tests results

In addition to non-functional tests for robustness, performance and stress, a quality assessment was conducted prior to the production go-ahead.

The following test methodology was implemented:

- Technical keywords that appear often in PCT application abstracts in English were selected, making sure of a broad technical domain coverage
- Corresponding search queries were built, adding, where appropriate, collocated words so that a proportion of the queries have adequate semantic complexity to expose ambiguities problems (See Annex)
- WIPO internal translators were asked to furnish reference translations for these queries in French, German and Spanish
- The quality of the expansions in the target languages were evaluated in automatic mode against the reference translations and the same was done with expansions obtained using Google Translate.

The following grading system has been used:

- **Insufficient**: all or a part of the meaning is lost in the proposed keyword translations
- **Good/sufficient**: there exists one translation in the cross lingual expansion that accurately conveys the expected meaning (within the reference translations or close)
- **Excellent**: there exist more than one translation in the cross lingual expansion that accurately conveys the expected meaning (within the reference translations or close)

and following test results were obtained:

System	Insufficient	Good / Sufficient	Excellent
Google Translate	24%	76%	N/A
WIPO CLIR	7%	57%	37%

Table 3: Evaluation of query expansions in French

System	Insufficient	Good / Sufficient	Excellent
Google Translate	21%	79%	N/A
WIPO CLIR	9%	40%	51%

Table 4: Evaluation of query expansions in German

System	Insufficient	Good / Sufficient	Excellent
Google Translate	21%	79%	N/A
WIPO CLIR	10%	51%	39%

Table 5: Evaluation of query expansions in Spanish

6 Conclusion and future work

The PATENTSCOPE CLIR system developed by WIPO facilitates simultaneous patent searching in multiple collections in different languages, exhibits search recall improvements and brings patent searching capabilities to non-English speaking internet users. It demonstrates how the recent advances in Statistical Machine Translation can be put to work in implementing cross lingual information retrieval systems.

Some of the more popular spoken languages in the world like Hindi/Urdu and Arabic (According

to Wikipedia[10]) are not yet supported by CLIR. Also some of the languages already supported have a limited coverage. Therefore, WIPO is actively seeking additional technical parallel corpora in these languages. Other types of corpora, different from patents could also be used, such as technical documentation, manuals and scientific articles.

We also believe that it is impossible to say that document translation based CLIR approaches perform better than query translation based CLIR approaches as a matter of principle. Just as quality is very diverse in Machine Translation, so it is in query translation based CLIR systems and only specific systems at a certain point in time can be compared.

What is obvious is that the query translation approach is more economical and manageable and that no information is lost while making the documents searchable to the end users.

Future work includes: (a) Defining an approach for under-represented languages and find sponsors to build parallel corpora in these languages meeting minimum size requirements; (b) Defining a crowdsourcing interface for proofreading the translated title misalignments obtained by using patent family relationships; (c) consider word decompounding for synthetic/agglutinative languages.

Acknowledgments

I would like to thank the translators of the International Bureau of the PCT who have been building the English-French high quality patent titles corpus during the last 30 years that has made the CLIR system possible. Special thanks also to my colleagues Paul Halfpenny and Bruno Pouliquen for their useful comments and corrections on the draft of this article.

References

Douglas W. Oard. 2009. Multilingual Information Access. Encyclopedia of Library and Information Sciences, 3rd Ed., edited by Marcia J. Bates, Editor, and Mary Niles Maack, Associate Editor, Taylor & Francis.

Bruno Pouliquen, Christophe Mazenc, Aldo Iorio. 2011. TAPTA: a user-driven translation system for patent documents based on domain-aware Statistical Machine Translation. Proceedings of 15th annual conference of the European Association For Machine Translation (EAMT)., Leuven Belgium.

Bruno Pouliquen, Christophe Mazenc. 2011. COPPA, CLIR and TAPTA: three tools to assist in overcoming the Patent Language barrier at WIPO. Proceedings of the Machine Translation Summit XIII, Xiamen. China.

Edlyn S. Simmons. 2009. "Black sheep" in the patent family. World Patent Information. Volume 31, Issue 1, March 2009 Pages 11-18

Leo Sarasúa 2000. Cross Lingual issues in patent retrieval. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece. ACM 2000, ISBN 1-58113-226-3.

Michael McCandless, Erik Hatcher, Otis Gospodnetić. 2010. Lucene in Action. 2nd Edition. Manning Press.

Nurul Amelina Nasharuddin, Muhamad Taufik Abdullah. 2010. "Cross-lingual Information Retrieval State-of-the-art" electronic Journal of Computer Science and Information Technology (eJCSIT), Vol. 2, No. 1.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris C. Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. Proceedings of ACL 07. Morristown, NJ, USA, 177-180.

Philipp Koehn & Hoang, Hieu. 2007. Factored Translation Models. EMNLP.

Yo Takagi, Andrew Czajkowski. 2012. WIPO Services for access to patent information – Building patent information infrastructure and capacity in LDCs and developing countries. To appear in World Patent Information, Volume 34, Issue 1, pages 30-36

[10]http://en.wikipedia.org/wiki/Most_spoken_languages_of_the_world

Appendix A: Test data

English source queries	French reference	German reference	Spanish reference	Japanese reference
Linguistic level: usual terms (plain language)				
feedingstuffs	aliments pour animaux	Futtermittel	pienso	飼料
pain killer, pain reliever, pain alleviating	analgésique, antidouleur, antalgique, soulagement de la douleur	Schmerzmittel, Analgetikum	analgésico, agente de alivio del dolor	鎮痛剤, 鎮痛薬, 痛み止め
noise cancellation, canceling	annulation de bruit, suppression du bruit	Kompensation von Störungen, Geräuschreduktion, Störgeräuschausblendung	anulación/cancelación de ruido	雑音除去, ノイズキャンセル, ノイズキャンセリング, ノイズ除去
antibody	anticorps	Antikörper	anticuerpo	抗体, 免疫体
packaging box	boîte d'emballage	Packkarton / Packkiste	envase	包装箱, 梱包箱
translucent/transparent housing, translucent casing	boîtier translucide, boîtier transparent	durchscheinende(n) / lichtdurchlässig / transparent Gehäuse	carcasa translúcida, casetón transparente	透明の筐体, 半透明の筐体, 透明のケース, 半透明のケース
noise, background noise, ground noise	bruit de fond	Hintergrundrausch, Grundrausch, Rauschanteil,	ruido de fondo, interferencia de fondo	暗騒音, 背景雑音, バックグラウンドノイズ
spurious noise, interference noise, noise	bruit parasite, bruit interférentiel	Störschall, Rauschen	interferencia de ruido, ruido	雑音, 干渉雑音
padlock	cadenas	Bügelschloss, Vorhängeschloss	seguridad, candado	南京錠
waste heat	chaleur perdue, chaleur dissipée	Abwärme	calor de desecho, calor residual	廃熱
electrical charge, electric load	charge électrique	eine elektrische Last	carga eléctrica,servicio eléctrico	電荷
hydraulic circuit, fluid circuit, hydraulic system	circuit hydraulique, circuit de fluide	hydraulische Stecke, Hydraulikkreislauf, Fluidkreislauf	circuito hidráulico,circuito de fluido	油圧系, 油圧回路, 水圧系
AC, alternating current	courant alternatif, CA	Wechselstrom	corriente alterna	交流電流, 交流
marine current, water current, sea current	courant marin	Meeresströmung, Wasserströmung	corriente marina	海流, 水流
transmission belt	courroie de transmission	Getriebeband, Antriebsriemen	correa de transmisión	伝動ベルト
safety catch	cran de sûreté	Sicherung	gatillo	安全装置
trenching	creuser des tranchées	graben	abrir trinchera	溝掘り
cancer diagnosis	diagnostic du cancer	Krebdiagnose	diagnóstico de cáncer	がん診断, 癌診断, がんの診断, 癌の診断
news/data/information broadcasting/distribution/dissemination	diffusion/dissémination d'informations, de données, de contenu	Verbreiten/Zurverfügungstellen von Informationen, Versand von Nachrichten, Informationsverbreitungsvorrichtung	difusión de información, difusión de datos, de contenidos	情報配信, データ配信
distribution of advertisement	distribution de messages publicitaires	Verteilung von Reklame / Verteilung von Werbungsmaterial	reparto de publicidad	広告の配信
touch screen	écran tactile	Berührungsbildschirm, Tastbildschirm, Touch-Screens	pantalla táctil	タッチパネル
liquid crystal display screen	écrans à cristaux liquides	Flüssigkristallanzeige / Flüssigkristalldisplay	pantalla de cristal líquido	液晶画面, 液晶ディスプレイ
inflatable element	élément gonflable	aufblasbares Element	elemento hinchable	膨張要素, インフレータブル要素
disposal of hospital waste	élimination des déchets médicaux	Entsorgung von medizinischen Abfällen / Beseitung von medizinischen Abfällen	eliminación de los desechos hospitalarios	医療廃棄物の処理
wind energy	énergie éolienne	Windenergie	energía eólica	風力
cue sports training	entraînement au billard	Billiard-Training	entrenamiento de billar	ビリヤードの訓練
wind turbine	éolienne	Windturbin,	aerogenerador,	風車, 風力タービン

		Windenergieanlage, Windkraftanlage, Windkraftgetriebe	generador eólico,turbina de viento	
steam iron	fer à vapeur	Bügeleisen	plancha de vapor	スチームアイロン, 蒸気アイロン
pieces of paper, paper sheets	feuilles de papier	Papier-Blättchen, Papierblätte, Papierbögen	láminas de papel, hojas de papel,	紙, 紙片, 紙切れ
glass fiber, glass-fiber	fibre de verre	Glasfaser	fibra de vidrio	ガラス繊維, 硝子繊維
optical fiber, fiber optic	fibre optique	Lichtwellenleiter, Lichtleitfaser , Lichtleiter, optische Faser	fibra óptica	光ファイバー
protective film, coating film	film protecteur, pellicule protectrice	Schutzfolie, Schutzfilm	película protectora	保護フィルム, 保護膜
blood flow	flux sanguin	Blutfluss, Blutfluß	flujo sanguíneo, torrente sanguíneo	血流
homeopathic	homéopathique	homöopathisch	homeopática/homeopático	ホメオパシー
pest control	lutte contre les nuisibles	Bekämpfung von tierischen Schädlingen, Schädlingsbekämpfung	control de plagas	害虫駆除, 害虫防除
washing machine	machine à laver	Waschmaschine	lavadora	洗濯機
skin diseases/disorders/conditions, dermatosis	maladies de la peau, affection cutanées, dermatoses	Hauterkrankungen, Hautkrankheiten, Dermatosen	enfermedades de la piel, enfermedad cutanea, piel infectadas	皮膚疾患, 皮膚病
respiratory mask	masque respiratoire	Atemmaske / Beatmungsmaske	máscara respiratoria	呼吸マスク, マスク
composite material	matériau composite	Verbundwerkstoff, Verbundmaterial	material compuesto	組成物, 複合材料, 複合物質
medicament, drug, medicine, medication, pharmaceutical composition	médicament, composition pharmaceutique	Medikament, Arzneimittel, pharmazeutische Zusammensetzung, Pharmazeutikum	medicamento, composición farmacéutica, fármaco	治療薬, 医薬
RAM, random access memory	mémoire vive	Arbeitsspeicher	(memoria) RAM	ランダムアクセスメモリ, ラム, ＲＡＭ, RAM
eyeglass frames	monture de lunettes	Brillenfassung	montura de gafas	眼鏡のフレーム, メガネのフレーム, 眼鏡の縁, メガネの縁
DC motor,D.C. motor, direct current motor	moteur à courant continu, moteur DC	Gleichstrommotor	motor de corriente continúa, motor de corriente directa	直流モーター, 直流モータ, 直流電動機
internet search engine	moteur de recherche par internet	Internetsuchmaschine / Suchmaschine	motor de búsqueda por Internet	インターネット検索エンジン, インターネットサーチエンジン, 検索エンジン, サーチエンジン
cutting tool	outil de coupe	Schneidewerkzeug	herramienta de corte	切削工具, 刃物
photovoltaic panels	panneaux photovoltaïques	Photovoltaikpaneelen	paneles fotovoltaicos	光起電性パネル
solar panels	panneaux solaires	Solaranlage, Solar-Panels, Solarpanels, Solarkollektoren, Solarpaneelen	paneles solares	太陽光発電パネル, 太陽電池パネル, ソーラーパネル
blood plasma	plasma sanguin	Blutplasma	plasma sanguíneo	血漿
active ingredient, active principle	principe actif	Wirkstoff	principio activo, ingrediente activo	有効成分, 活性成分
vegetable protein	protéine végétale	Pflanzenprotein, pflanzliche Protein	proteína vegetal	植物性蛋白質, 植物性タンパク質, 植物性たん白質
wireless receiver	récepteur sans fil	Funkempfänger	receptor inalámbrico	無線受信機
telecommunications network, telecommunication network	réseau de télécommunications	Telekommunikationsnetz	una red telecomunicaciones	電気通信網, 通信網

transportation network, communication network, transport network, supply network	réseau de transport, réseau de distribution	Kommunikationsnetz, Transportnetz,Versorgungsnetz	red de transporte, red de distribución	輸送網, 交通網, 流通網
vacuum coating	revêtement sous vide	Vakuumbeschichtung	revestimiento al vacío	真空塗装, 真空めっき
mineral salts	sels minéraux	anorganische Salze, Mineralsalze, mineralische Salze	sales minerales	無機塩
DNA sequences, deoxyribonucleic acid	séquences ADN, acide désoxyribunocléique	DNA-Sequenz	ADN, ácido desoxirribonucleico, secuencias de ADN	ＤＮＡ配列, DNA 配列, デオキシリボ核酸
tolerance, tolerance threshold	seuil de tolérance	Toleranzgrenze, Toleranz	umbral de daño	許容範囲, 許容レベル
shelving system	système de rayonnage	Regalsystem	sistema de estantería	棚割りシステム
mobile phone, cell phone	téléphone mobile, téléphone portable	Mobiltelefon, tragbares Telefon	un teléfono móvil, teléfono portátil	携帯電話
breast cancer therapy	traitement du cancer du sein	Krebstherapie	tratamiento de cáncer de mama	乳癌治療, 乳がん治療
central processing unit, central unit, CPU	unité centrale (de traitement), CPU	Zentraleinheit, CPU	entidad central, unidad central, CPU	中央演算処理装置, 中央処理装置, CPU
vaccine	vaccin	Vakzine	vacuna	ワクチン
water vapor, water vapour	vapeur d'eau	Wasserdampf	vapor de agua	水蒸気
cooling fan	ventilateur	Kühlgebläse, Kühlerlüfter, Lüfterantriebe	ventilador de enfriamiento	クーリング・ファン, 冷却ファン, 冷却用ファン
lock, latch, bolt	verrou, serrure	Schloss, Riegel, Schloß, Verriegelung, Gesperre	pasador, pestillo, cerroyo, cerradura	ロック, 錠, 掛け金, 掛金, ラッチ, ボルト
VOD, video on demand	vidéo à la demande	Video on Demand	vídeo bajo demanda	ビデオ・オン・デマンド, VOD
Linguistic level: usual technical term				
acetic acid	acide acétique	Essigsäure	ácido acético	酢酸
sulphuric acid	acide sulfurique	Salpetersäure	ácido sulfúrico	硫酸
amino acids	acides aminés	Aminosäure	aminoácidos	アミノ酸
anti-viral agent	agent antiviral, anti-viral	antiviral Mittel	agente antiviral	抗ウィルス物質, 抗ウィルス薬
antifungal	antifongique	Antimykotikum, antifungish	antifúngica	抗真菌剤, 抗真菌薬, 抗真菌性
coupling apparatus	appareil de couplage	Kopplungsgerät / Kopplungsvorrichtung	aparato de acoplamiento	カップリング装置, 結合装置
rifle body	boîtier de fusil	Gewehrgehaüse	cuerpo de rifle	ライフル銃
transmission channel	canal de transmission	Übertragungskanal	un canal de transmision	送信チャネル, 伝送路, 通信路, 伝送チャネル
temperature sensor, thermal sensor	capteur de température, capteur thermique	Temperatursensor, Thermosensor	dispositivo sensor de temperatura, dispositivo sensor térmico, sensor térmico	温度センサー, 感温部, 温度検知器, 熱センサー
PCB, printed circuit board, circuit board, motherboard	carte de circuit imprimé, PCB, carte mère	Leiterplatte, Motherboard, Hauptplatine	placa de circuito impreso, tarjeta de circuito impreso, tarjeta madre	プリント基板, 印刷回路板, PCB, サーキットボード, 回路基板, マザーボード
stem cells	cellules souche	Stammzellen	célula madre	幹細胞
vacuum chamber	chambre à vide	Vakuumkammer	cámara de vacío	真空室, 真空槽, 真空箱, 真空チャンバー
vehicle frame structure, vehicle chassis	châssis de véhicule	Fahrwerk, Fahrzeugchassis, Fahrzeugrahmer	chasis de vehiculo, carroceria de vehiculo	車両本体フレーム構造, 自動車シャシ
heat-conductor, thermally conductive	conducteur de chaleur, conducteur thermique	Wärmeleiter, Heizleiter, wärmeleitende(n)	conductor de calor, conductor térmico	熱導体, 熱伝導性
thin layer, thin film	couche mince	Dünnfilm, eine dünne Schicht, Dünnschicht	capas finas, láminas delgadas	薄層, 薄い層, 薄膜
pneumatic cylinder	cylindre pneumatique	pneumatische Zylinder, Pneumatikzylinder	cilindro neumático	空気圧シリンダー
hard drive, HD	disque dur	Festplatte	disco duro	ハード・ドライブ, ハ

				ードディスク・ドライブ, 固定磁気ディスク装置, HD, HDD
heat dissipating structure for LED, heatsink for LED	dissipation thermique de DEL	Hitzeableitungsvorrichtung für Leuchtdioden	disipación térmica para DEL	LED の放熱構造, 発光ダイオードの放熱構造, LED のヒート・シンク, LED の放熱板
anti-theft lock-nut	écrou antivol	Diebstahlsicherung	tuerca antirrobo	盗難防止ロックナット
electromagnet	électroaimant	Elektromagnet	electroimán	エレクトロマグネット, 電磁石, 電動磁石
outer case, outer shell, outer housing	enveloppe extérieure	äußere Haut, Umhüllung, Außenhaut, Außenschal	carcasa externa, envolvente exterior, piel exterior	外箱
visual field test	examen du champ visuel	Gesichtsfelduntersuchung	examen del campo visual	視野検査, 視野テスト
plastic film	film plastique	Kunststofffolie	film plástico	プラスチックフィルム
particulate filter	filtre à particules	Rußfilter, Russfilter, Partikelfilter	filtro de particulas	粒子フィルタ
Greenhouse gas, GHG, CO2 gas, carbon dioxyde	gaz à effet de serre, GHG, dioxyde de carbone, CO2, gaz carbonique	Treibhausgas, Kohlendioxid, Kohlensäure, Kohlendioxidgas	gases de invernadero, dióxido de carbono, anhídrido carbónico	グリーンハウスガス, 温室ガス, 温室効果ガス, 温室化ガス, CO2 ガス, 二酸化炭素
liquid hydrogen	hydrogène liquide	flüssige Wasserstoff, Flüssigwasserstoff	hidrógeno liquido	液体水素
corrosion inhibitor	inhibiteur de corrosion	korrosionshemmender, Korrosionsinhibitor	inhibidor de corrósion	防蝕剤, さび止め剤, 錆止め剤, 腐食防止剤
graphical user interface	interface graphique utilisateur	graphische Benutzeroberfläche / grafische Benutzeroberfläche	interfaz usuario gráfica	グラフィカル・ユーザ・インタフェース, GUI
seal	joint d'étanchéité	Dichtung, Abdichtmanschette, Abdichtung	junta de estanqueidad	封印, シール
composite material	matériau composite	Verbundwerkstoff, Verbundmaterial	material compuesto	複合材料
flash memory	mémoire flash	Flash-Speicher	memoria flash	フラッシュメモリ
conductive metal	métal conducteur	leitfähiges Metall	conductoro metal	導電性金属, 導電性の金属, 導電性の高い金属
starch molecule	molécule d'amidon	Stärkemoleküle	almidón molécula	澱粉分子, でんぷん分子
fine particles	particules fines	feinster Partikel	partículas finas	細粒子
double-hulled tank ship	pétrolier double coque	Doppelhüllentanker	buque tanque doble casco	ダブルハル・タンカー, 二重船殻構造のタンカー
fuel cell	pile à combustible	Brennstoffzelle / Brennstoffelement	pila de combustible	燃料電池
potash	potasse	Kali / Pottasche	potasa	炭酸カリウム, カリ
hydraulic press	presse hydraulique	hydraulische Presse	prensa hidráulica	液圧プレス, 水圧プレス, 水圧機, 油圧プレス
free radicals	radicaux libres	freie Radikale	radicales libres	遊離基, フリーラジカル
control signal, switching signal	signal de commande	Schaltsignal	señal de control	制御信号, 調節信号, コントロール信号, 開閉信号
aqueous solution	solution aqueuse	wässrige Lösung	disolución/solución acuosa	水溶液
organic solvent	solvant organique	ein organisches Lösungsmittel, Lösemittel	solvente orgánico	有機溶媒, 有機溶剤
intake valve	soupape d'admission	Zylinderkammereinlassventil, Einlassventil	válvula de admisión	吸気弁
lumbar support	support lombaire	Lordosenstützen	soporte lumbar	ランバーサポート
actuator, controlling device	système de commande, actionneur	Leitsystem, Steuermodul, Bestätiger	actuator,sistema para el control	作動装置, アクチュエータ, 制御装置, 調整装

				置
measuring system, measurement arrangement, measuring instrument	système de mesure, instrument de mesure	Messvorrichtung, Messsystem, Messanordnung, Messgerät	sistema de medida, systema de medición, systema para medirm,sensor para la medida, instrumento y método	測定系, 測定システム, 測定装置, 計測器, 測定機器
audio data processing	traitement de données audio	Audiodatenverarbeitung	procesamiento de datos audio	音声データ処理, 音声情報処理
signal transmission, transmitting signal	transmission de signal	Signalübertragung	una transmision de senal/señal	信号伝送, 送信信号
neurological disorders	troubles neurologiques	neurologischen Störungen	desordenes neurológicos	神経学的障害, 神経障害, 神経疾患, 神経学的異常
worm, worm screw, screw conveyor	vis sans fin	Förderschnecke	tornillo sin fin, transportador helicoidal , transportador sin fin	ねじコンベヤ, ウォームねじ
Linguistic level: technical term (common in popular science journals)				
alkyl	alkyle	Alkyl	alquilo	アルキル
power analysis attack	attaque par analyse de consommation	Angriff durch Stromanalyse	ataque por análisis de consumo	電力解析攻撃
induction coil, inductor	bobine d'induction	Induktionsspule	inducido, bobina de induccion	誘導コイル, インダクションコイル
drillling fluid	boue de forage	Bohrflüssigkeit	lodo	掘削流体, 掘削泥水, ボーリング泥水
milling process	broyage, processus de broyage, mouture, fraisage	Fräsverfahren, Fräsprozess, Mahlprozess	fresado, proceso de molienda, Procedimiento de fresado/molido	製粉過程, 製粉工程, フライス加工, 精錬工程
control channel	canal de commande	Steuerkanal, Steuerungskanal	canal de control	制御通信路, コントロールチャネル
pressure sensor	capteur de pression	Drucksensor	un sensor de presión	圧力センサー, 圧力センサ, 圧力感知器, 圧力検知器
internal combustion chamber	chambre de combustion interne	Brennraum, Brennkammer	Cámara de combustión interna	内燃機関, IC
VoIP communication	communication VoIP	VoIP-Kommunikation	comunicación VoIP	VoIP 通信
pressure switch	commutateur de pression	Druckschalter / Druckwächter / Druckluftschütz / pneumatisches Schütz	interruptor de presión	圧力スイッチ
air duct	conduit d'air	Luftkanal	conducto de aire	エアダクト, 空気孔, 送風ダクト
drive torque	couple d'entraînement	Antriebsdrehmoment	torque de impulsion	駆動トルク
scanning direction	direction de balayage	Scanrichtung	direccion de exploracion	スキャン方向
tunneling device	dispositif de tunnellisation	Tunnelierer (med.)	dispositivo de tunelización	掘進機, トンネリング機器, トンネリング装置, 掘削装置
notch filter	filtre coupe-bande	Kerbfilter / Bandsperrfilter / Bandsperre	filtro de muesca	ノッチフィルター, ノッチ・フィルター, ノッチフィルタ
cryogenic fluid	fluide cryogénique	kryogene Flüssigkeit / kryogenische Flüssigkeit	fluido criogénico	低温流体
magnetic flux	flux magnétique	magnetische(s) Fluss, Fluß, Magnetfluss	flujo magnético	磁束
motive power, motive force	force motrice	Antriebskraft, motorische Kraft	una fuerza motriz	原動力
hemostatic, antihemorrhagic, styptics	hémostatique, antihémorragique, styptique	Hämostasisch, Alaunstift, Blutstillstift, Rasierstift	hemostático/a	収斂剤, 止血薬, 抗出血性の, 収れん性の, 止血性の, 収斂性の
inhibition of cell proliferation	inhibition de prolifération cellulaire	Hemmung von Zellproliferation / Hemmung von Zellvermehrung	inhibición de proliferación celular	細胞増殖抑制, 細胞増殖調節, 細胞増殖阻害
plasma arc, plasma jet	jet de plasma	Plasmastrahl, Plasma-	haz de plasma, chorro	プラズマアーク, プラ

		Jet	de plasma	ズマジェット
discharge lamp	lampe à décharge	Entladungslampe, Hochdruckentladungslampe	ámpara de descarga	放電ランプ
waterline	ligne d'eau/ de flottaison	Wasserlinie	línea de agua/flotación	水線, 水位線
multi-phase mixture, mix	mélange à phases multiples	Mehrphasengemisch	multifase mezcla , mezcla de fases multiples	混合, マルチフェーズ混合, 多重相混合組織
injection molding	moulage par injection	Spritzguss, Spritzgieß	inyección por moldeo, moldeo de inyección	射出成形
revolving/rotating/rotary core	noyau rotatif	verdrehbare Kern	núcleo giratorio/rotatorio	回転コア, 回転核
antisense nucleotides	oligonucléotides antisens	Antisense-Nukleotid	oligonucleótidos antisentido	アンチセンスヌクレオチド
thermal grease	pâte thermique, pâte thermoconductible	Wärmeleitpaste	pasta de transferencia térmica	熱グリース
bulk memory device	périphérique de mémoire de masse	Massenspeichereinheit	almacenamiento masivo	大容量記憶装置
voltage range	plage de tension	Spannungshub	escala de voltaje, intervalo de voltajes	電圧範囲
porous polymer	polymère poreux	poröse(s) Polymer	polimero poroso	多孔質高分子, 多孔質ポリマー
sugarcane juice preservation	préservation de jus de sucre de canne	Konservierung von Zuckerrohrsaft	preservación de jugo de caña (de azúcar)	サトウキビ汁の保存, サトウキビ汁の防腐, 砂糖黍汁の保存, 砂糖黍汁の防腐
OCR, optical character recognition	reconnaissance optique de caractères	optische Zeichenerkennung	OCR, reconocimiento óptico de caracteres	光学式文字認識, OCR
insulating resin	résine isolante	Isolierharz	resina aislante	絶縁樹脂
thermosetting resin, thermohardening resin, duroplastic	résine thermodurcissable, duromère	wärmehärtbare Harz, Duroplasten, thermisch härtendes Harz	resina termoendurecedora, material duraplastico	熱硬化性樹脂
abrasion resistance	résistance à l'abrasion	Abriebfestigkeit, Kratzfestigkeit, Scheuerbeständigkeit	resistencia a la abrasión	耐摩耗性, 磨耗抵抗, 磨耗強度, 磨耗耐久性
rolling resistance	résistance au roulement	Rollwiderstand	resistencia al rodadura, resistencia al deslizamiento	転がり抵抗, 走行抵抗
transgenic rodents	rongeurs transgéniques	transgene Nagetiere / transgene Nager	roedores transgénicos	トランスジェニックマウス, トランスジェニック齧歯動物, 遺伝子移植用の齧歯動物, 遺伝子組み換え齧歯動物, 遺伝子組換え齧歯動物, 遺伝子組換齧歯動物, 遺伝子組み換えマウス, 遺伝子組換マウス
siphon valves	soupapes de siphon	Siphonverschluss / Heberverschluss / Siphonventil / Siphon-Ventil / Heberventil / Heber-Ventil	válvula de sifón	サイフォンバルブ
hollow structure, tubular structure	structure creuse, structure tubulaire	Hohlstruktur,eine rohrförmige Struktur	estructura hueca,estructura tubular	空洞構造, 管腔構造, 管状構造物
crystalline structure	structure cristalline	Kristallstruktur	una estructura cristalina	結晶構造
data structure	structure de données	Datenstruktur	una estructura de datos	データ構造, データ・ストラクチャ, データストラクチャー, データストラクチャ
aqueous slurry, aqueous suspension, water suspension	suspension aqueuse	wässrige Suspension	suspensión acuosa	水性スラリー, 水性スラリ, 水性懸濁液, 水懸濁
DSP, digital signal processing, digital processor	traitement numérique du signal, DSP, processeur numérique, processeur de signaux	digitale Signalverarbeitung, DSP, digitale Signalprozessor	procesamiento digital de señal/senal,DSP, procesador digital	DSP, デジタル信号処理, デジタルプロセッサ, デジタルプロセッ

	numériques			サー, デジタル処理装置
Allen screw	vis Allen	Allen Schrauben	tornillo Allen	アレンスクリュー
trim hatch	volet d'ajustement d'assiette	Trimmklappe		
Patent technical term				
fat-soluble flavoring	aromatisant liposoluble	fettlöslicher Aromastoff	aromatizante liposoluble	脂溶性調味料, 脂溶性香味料
piezo-resistive pressure sensor	capteur de pression à piézorésistance / piézorésistif	piezoresistiver Drucksensor	sensor de presión del tipo piezorresistivo	ピエゾ抵抗圧力センサー, ピエゾ抵抗圧力センサ, ピエゾ抵抗圧力感知器, ピエゾ抵抗圧力検知器
a cleat for fastening a rope	clayette de fixation d'attache	Klemme zur Sicherung eines Seils / Klemme zur Befestigung eines Seils	taco de sujeción	綱留め, 綱止め
nozzle body, nozzle housing, nozzle element	corps de tuyère, corps de buse, corps de buselure	Düsengehäuse, Düsenkörper	cuerpo de boquilla,un cuerpo de tobera	ノズル体, ノズル筐体, ノズル部
peritoneal dialysis	dialyse péritonéale	Bauchfelldialyse / Peritonealdialyse	diálisis peritoneal	腹膜透析
swirling flow	écoulement tourbillonnant, écoulement tourbillonnaire	Drallströmung	flujo del remolino	渦巻き流, 渦巻流
bacterial endotoxin	endotoxine bactérienne	bakterielles Endotoxin	endotoxina bacteriana	細菌内毒素
rotary/rotating sintering	frittage rotatif	Rotationssinterverfahren	giratorio sinterizado	回転焼結
refractive index	indice de réfraction	Brechungsindex	indice de fraccion	屈折率
welded joints	joints soudés, à soudure	Schweißverbindungen	juntas soldadas	溶接継手, 溶接継ぎ手
polyethylene glycol	polyéthylène glycol	Polyethylenglykol	polietilenglicol	ポリエチレングリコール
cross-linked polymer, crosspolymer, crosslinked polymer	polymère réticulé	vernetzte(s) Polymer	polimero reticulado	架橋重合体, 架橋ポリマー, 架橋されたポリマー, 架橋された重合体, クロスポリマー
polyurethane foam	polyuréthane expansé	Polyurethanschaum	espuma de poliuretano	ポリウレタンフォーム, ポリウレタン・フォーム
cavitation process	procédé de cavitation	Kavitationsprozess	cavitación	キャビテーションプロセス, キャビテーション工程
water absorbent resin	résine hydrophile	hydrophile Harz	resina hidrófila	吸水性樹脂
DNA segregation	ségrégation d'ADN	DNS-Trennung / DNA-Trennung	segregación de ADN	DNA 分離
E. Coli strain	souche d'Escherichia Coli	E. Coli-Stamm	cepa E. Coli	大腸菌株, エシェリキア菌株, E. Coli 菌株
semiconductor substrate	substrat semi-conducteur	Halbleitersubstrat, Halbleiterträger	sustrato semiconductor	半導体基板
PET scan, positron emission tomography	tomographie par émission de positrons, TEP, PET scan	Positronen-Emissions-Tomographie, PET-Untersuchung	tomografía por emisión de positrones, PET, TEP	PET スキャン, 放射断層撮影法スキャン, ポジトロン CT, 陽電子放出断層撮影, 陽電子放出型断層撮影, ポジトロン放出断層撮影, ポジトロン放出型断層撮影

プロジェクト報告

特許翻訳における統計翻訳とルールベース翻訳の比較 —NTCIR-9 特許機械翻訳タスクでの分析—

情報通信研究機構　後藤 功雄

City University of Hong Kong / Hong Kong Institute of Education　Bin Lu

Hong Kong Institute of Education　Ka Po Chow

情報通信研究機構　隅田 英一郎

Hong Kong Institute of Education / City University of Hong Kong　Benjamin K. Tsou

1 はじめに

特許翻訳は，外国語で書かれた特許の理解や外国への特許出願のために，実用上において大きなニーズがある．そのために特許の機械翻訳の研究は重要であり意義がある．この研究の促進のために評価型ワークショップである NTCIR-9 特許機械翻訳タスクを主催した[4]．NTCIR-9 特許機械翻訳タスクでは，特許の機械翻訳の研究に利用できる共通のデータの構築，最新の機械翻訳システムの評価，自動評価手法の信頼性の評価を実施した．このタスクは過去 2 回の特許翻訳タスク[2,3]を基にしている．今回は新たに中英翻訳と acceptability に基づく人手評価も行った．本稿では，NTCIR-9 特許機械翻訳タスクの概要，評価手法，翻訳システムと自動評価の評価結果について述べる．

2 特許機械翻訳タスクの概要

特許機械翻訳タスクでは，共通のデータを用いて複数の参加グループのシステムおよび主催者によるベースラインシステムを評価することで，特許翻訳におけるシステムの比較評価を可能にするとともに，特許機械翻訳の実用に向けての評価を実施した．

タスク実施の流れは次のとおりである．

- 主催者が特許文からなる訓練データおよびテストデータを参加グループに提供する．
- 参加グループはそれぞれの手法でテストデータを機械翻訳して提出する．
- 主催者は提出された翻訳結果を評価して，評価結果を参加グループへ返送する．
- ワークショップで参加グループが研究成果を発表する．

実施した翻訳の言語対および翻訳方向は，日本語から英語（日英），英語から日本語（英日），中国語から英語（中英）である．訓練データとして，日英・英日翻訳は約 320 万文対の日英対訳コーパス，中英翻訳は 100 万文対の中英対訳コーパス，翻訳先言語の単言語コーパスとして，日英・中英翻訳には英語の特許文 3 億文以上，英日翻訳には日本語の特許文 4 億文以上を提供した．開発データには 2,000 文対，テストデータには 2,000 文を用いた．対訳コーパス，開発データ，テストデータは，請求項の文を含まず，主に発明の詳細な説明部分の文からなる．訓練データは 2005 年以前，開発データおよびテストデータは 2006−2007 年の特許から構築した．テストデータは自動的に抽出した対訳文対から，人手で正しい対訳の文対を選択して構築した．テストデータは対訳文対として自動抽出された文から構築しているため，実際の特許文の傾向が

表 1 ベースラインシステム

システム ID	システム	種類	日英	英日	中英
BASELINE1	Moses hierarchical phrase-based SMT system	SMT	✓	✓	✓
BASELINE2	Moses phrase-based SMT system		✓	✓	✓
RBMTx	SYSTRAN 7 Premium Translator	RBMT			✓
RBMTx	Huajian Multilingual EasyTrans version 3.0				✓
RBMTx	The Honyaku 2009 premium patent edition		✓	✓	
RBMTx	ATLAS V14		✓	✓	
RBMTx	PAT-Transer 2009		✓	✓	
ONLINE1	Google online translation system	SMT	✓	✓	✓

完全には反映されていない．この点は次の NTCIR-10[1] での改善を検討中である．

参加グループ数は全体で 21，日英翻訳は 12，英日翻訳は 9，中英翻訳は 18 であった．タスクに参加した翻訳エンジンの種類は，大きく分類すると統計翻訳（SMT），ルールベース翻訳（RBMT），用例翻訳（EBMT），複数手法の組み合わせ（HYBRID） の 4 種類である．

さらに，主催者がベースラインシステムの結果として 2 種類の SMT（フレーズベース SMT，階層フレーズベース SMT）[2]，5 つの RBMT，Google 翻訳による翻訳結果を追加した．利用したベースラインシステムを**表 1** に示す．商用 RBMT には各言語対でよく知られているシステムを選択した．SMT ベースラインシステムは一般に公開されているソフトウェアで構築し，システムの構築と翻訳の手続きは特許機械翻訳タスクの Web ページ[3]で公開されているため，参加者は同じシステムを構築し，その結果と比較することができる．ベースラインシステムの商用 RBMT のシステム ID は，本稿では匿名にしている．

[1] http://research.nii.ac.jp/ntcir/ntcir-10/
[2] http://www.statmt.org/moses/
[3] http://ntcir.nii.ac.jp/PatentMT/

3 評価手法

人手による文単位の訳質評価を実施した．各システムあたり 3 人の評価者がそれぞれ 100 文，合計 300 文を評価した．評価基準として，adequacy と acceptability の 2 つを用いた．これらは，主に翻訳結果を読んで内容を理解するという情報アクセスにおける有用性の評価基準である．情報発信のために翻訳結果を下訳として用いた場合における有用性の評価は，今回は実施しなかった．以下に本評価で用いた評価基準について説明する．

3.1 Adequacy

翻訳の適切さ（adequacy）の評価を 5 段階（1-5）で実施した．この評価の目的は，システム間の比較である．本評価では，節レベルの訳質までを考慮して評価した．

3.2 Acceptability

図 1 に示す 5 段階の acceptability 評価を実施した．この評価の目的は，文レベルの意味が正しく伝わる文の割合を明らかにすることである．acceptability は入力文の意味が正しく伝わらない（たとえば，重要な情報が一つでも欠ける）と F になる。この評価は，adequacy と比べて，より実用に近い評価を目指している．例え

ば，システムへの要求水準が「入力文の意味が分かればよい」であれば，C 以上の評価となった訳文が有用であり，システムへの要求水準が「入力文の意味が分が分かり，かつ文法的に正しい」であれば，A 以上の評価となった訳文が有用である．このように，要求水準に応じた訳質の文の割合を明らかにすることができる．adequacy 評価の結果からは，このような文の割合は分からない．

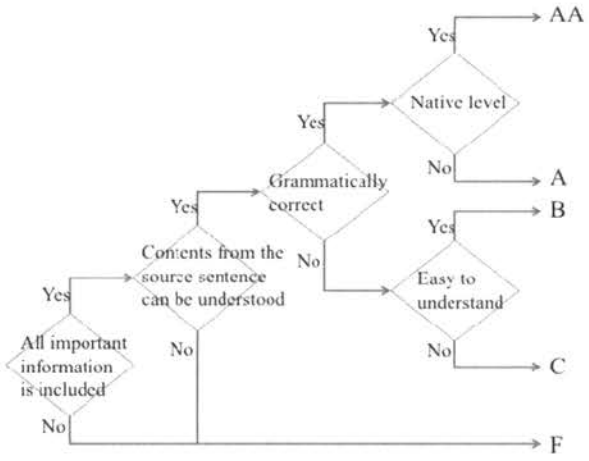


図 1: Acceptability

4 翻訳システムの評価結果

紙面の都合上，評価結果の要点のみ報告する．詳細な報告[4]および各グループの成果報告はオンラインで入手可能である[4]．図 2ー図 4 に adequacy 評価結果，図 5ー図 7 に acceptability 評価結果を示す．図中のシステム名は，グループ ID（またはシステム ID）とプライオリティ番号の組で表示されている．順位は，adequacy は平均値，acceptability は一対比較に基づいている．

機械翻訳における主な課題は訳語選択と語順推定である．SMT は訳語選択の性能は高いため，SMT の現状の大きな課題は語順推定である．以下，言語対毎に結果を分析する．

4.1 日英翻訳

図 2 および図 5 に評価結果を示す．JAPIO-1 と RBMTx-1 は RBMT，EIWA-1 は RBMT と SMT の HYBRID，KYOTO-1 は EBMT，それ以外は SMT である．RBMT の訳質が SMT より高いことが分かる．acceptability で C 以上の割合は，RBMT のトップのシステムが 6 割程度であったのに対し，SMT のトップのシステムは 2.5 割程度であった．日英翻訳は，言語間の語順が大きく異なる（日本語の語順は SOV，英語の語順は SVO）ため，語順の推定が難しい．それが SMT で高い翻訳品質を達成できなかった主な原因と思われる．日英の RBMT の性能が高いため，結果的に RBMT が SMT よりも良い結果となった．

4.2 英日翻訳

図 3 および図 6 に評価結果を示す．RBMTx-1 と JAPIO-1 は RBMT，KYOTO-1 は EBMT，それ以外は SMT である．トップの SMT（NTT-UT-1）の訳質がトップの RBMT と同等以上であることが分かる．英日の特許翻訳で SMT がトップレベルの RBMT と同等以上の訳質を達成したことが明らかになったのはこの評

[4] http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/

価が初めてである．英日翻訳は，日英翻訳と同様に語順が大きく異なるため，SMT での語順の推定が難しい．そのため，過去の NTCIR-7 では，RBMT の人手評価が SMT より高かった．今回，NTT-UT グループが翻訳の前処理で英語の語順を日本語の語順に入れ替える手法を用いる[8]ことによって，語順の推定精度を大幅に向上させることに成功した．トップの SMT と RBMT のシステムはテスト文の 6 割程度で acceptability が C 以上という結果を得た．

4.3 中英翻訳

図 **4** および図 **7** に評価結果を示す．RBMTx-1 は RBMT, EIWA-1 は RBMT と SMT の HYBRID, KYOTO-1 は EBMT，それ以外は SMT である．SMT の訳質が RBMT より高いことが分かる．中英翻訳は，日英翻訳に比べて言語間の語順が似ている（どちらも語順が SVO の言語）ことと，RBMT の性能が低かったために，SMT が RBMT よりも良い結果になったと思われる．トップの BBN グループは，特許翻訳の品質向上のために，低頻度な数値表現の汎化，中国語単語分割の最適化，言語モデルの適応，特徴量の追加，英語依存構造の利用などを行い[6]，テスト文の 8 割程度で acceptability が C 以上という結果を得た．

5 自動評価に対する評価結果

訳質の評価には自動評価が重要である．この重要な役割を担う自動評価の信頼性を人手評価結果に基づいて評価した．自動評価スコアとして BLEU[7], NIST[1], RIBES[5]を用いた．詳細なスコア計算方法は，特許機械翻訳タスクの Web ページに示されている．テストデータ 2,000 文の訳から計算した各システムの自動評価スコアと各システムの adequacy 平均値とを比較した．人手評価と標準化した自動評価値の散布図を図 **8**ー図 **10** に示す．横軸が adequacy 平均値，縦軸が標準化した自動評価値である．**表 2** にスピアマン順位相関係数とピアソン積率相関係数，**表 3** に RBMT を除いた日英と英日の各係数を示す．

図 **8**, 図 **9** の緑枠部分は RBMT の結果である．日英・英日では，RBMT の自動評価は人手評価との相関が低い．日英では RBMT を除くと人手評価との相関が高くなるが，図 **8** からいずれの自動評価も中英での人手評価との相関（図 **10**）と比べると人手評価との相関は高くない．英日では RBMT を除くと図 **9** から RIBES と人手評価との相関が高いと言える．中英では図 **10** からすべての自動評価と人手評価との相関が高いと言える．

6 まとめと今後の予定

NTCIR-9 特許機械翻訳タスクでの分析により，特許翻訳における最新の統計ベース翻訳と規則ベース翻訳の比較および自動評価に対する評価を示した．NTCIR-10 では，より実用的な評価および実際の特許文の傾向を十分に反映した評価の実施を検討中である．

[1] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of HLT, pp. 138–145, 2002.

[2] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In Proceedings of NTCIR-7, 2008.

[3] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizenya, and Sayori Shimohata. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In Proceedings of NTCIR-8, 2010.

[4] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the Patent Machine

Translation Task at the NTCIR-9 Workshop. In Proceedings of NTCIR-9, pp. 559–578, 2011.

[5] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic Evaluation of Translation Quality for Distant Language Pairs. In Proceedings of EMNLP, pp. 944–952, 2010.

[6] Jeff Ma and Spyros Matsoukas. BBN's Systems for the Chinese-English Sub-task of NTCIR-9 Patent MT Evaluation. In Proceedings of NTCIR-9, 2011.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of ACL, pp. 311–318, 2002.

[8] Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Masaaki Nagata, Xianchao Wu, Takuya Matsuzaki, and Jun'Ichi Tsujii. NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT. In Proceedings of NTCIR-9, 2011.

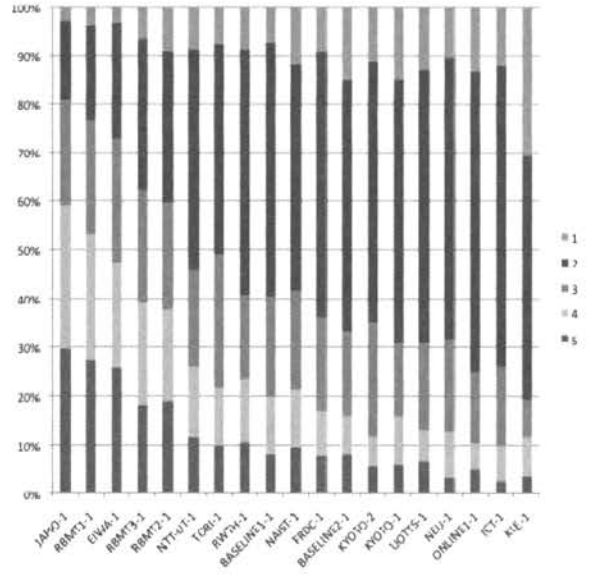


図 2: 日英 adequacy

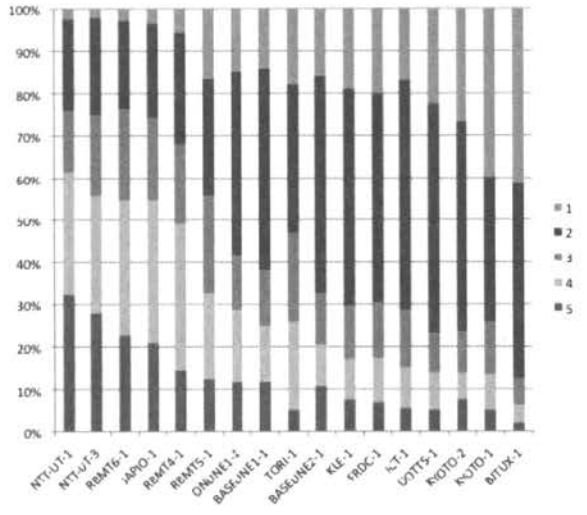


図 3: 英日 adequacy

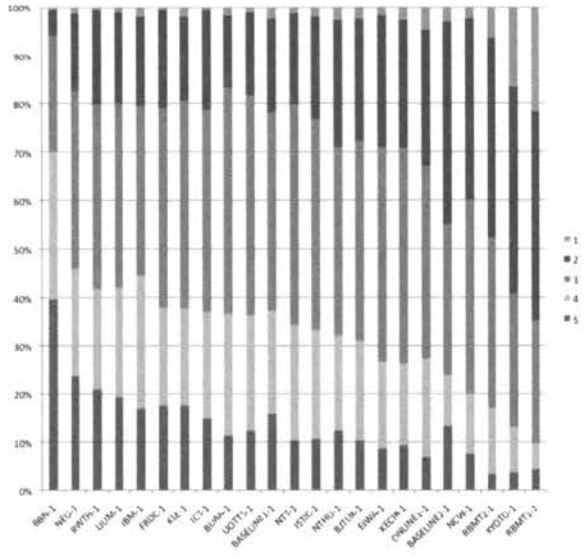


図 4: 中英 adequacy

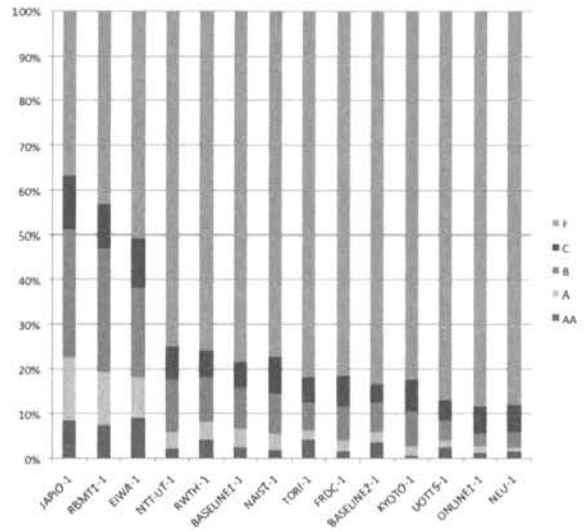


図 5: 日英 acceptability

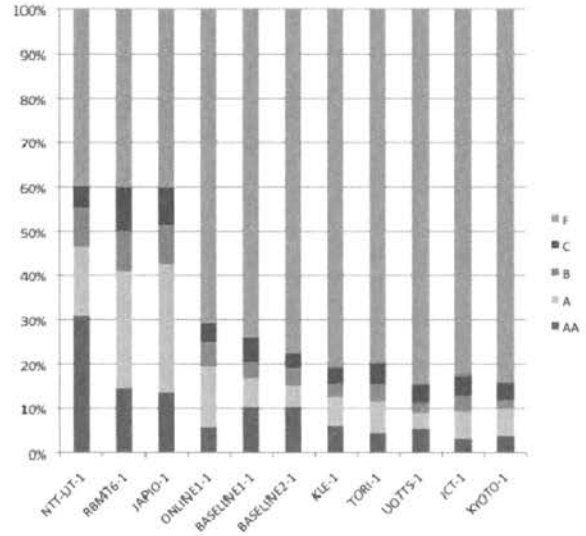


図 6: 英日 acceptability

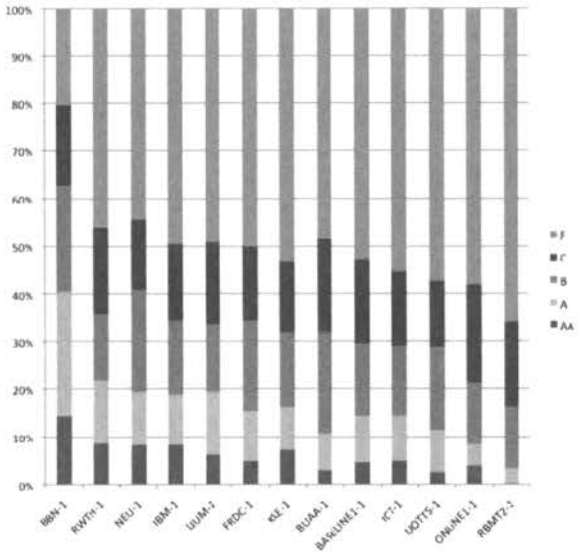


図 7: 中英 acceptability

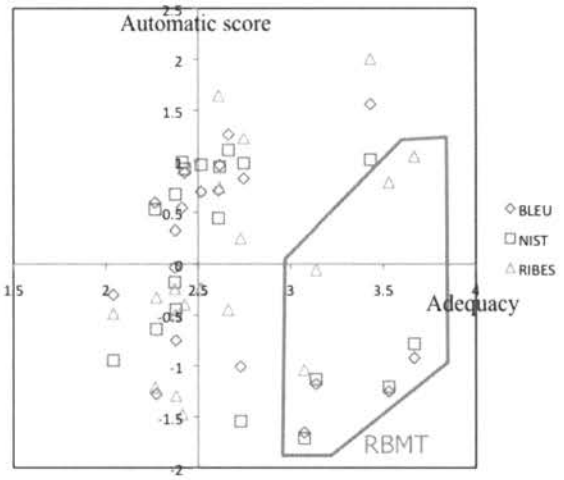


図 8: 日英 Adequacy と自動評価との相関

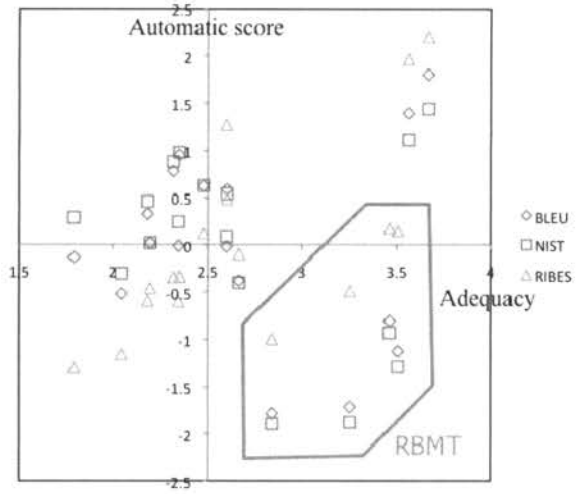


図 9: 英日 Adequacy と自動評価との相関

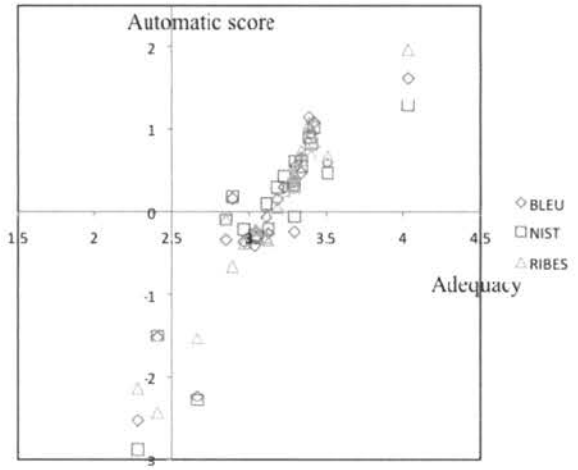


図 10: 中英 Adequacy と自動評価との相関

表 2: Adequacy と自動評価の相関係数

		Spearman	Pearson
日英	BLEU	-0.042	-0.241
	NIST	-0.114	-0.286
	RIBES	0.632	0.579
英日	BLEU	-0.029	-0.032
	NIST	-0.074	-0.209
	RIBES	0.716	0.683
中英	BLEU	0.931	0.915
	NIST	0.911	0.891
	RIBES	0.949	0.967

表 3: Adequacy と自動評価の相関係数（RBMT を除く）

		Spearman	Pearson
日英	BLEU	0.618	0.525
	NIST	0.543	0.362
	RIBES	0.679	0.741
英日	BLEU	0.511	0.753
	NIST	0.412	0.603
	RIBES	0.929	0.943

プロジェクト報告

オフショアソフトウェア開発における異言語文書理解支援システム

程　祥瑞，松本真佑，中村匡秀

神戸大学大学院システム情報学研究科　〒657-8501 神戸市灘区六甲台町 1-1

1. はじめに

ソフトウェア開発の全体，または，一部を海外の企業に委託するオフショアソフトウェア開発(以降，オフショア開発と呼ぶ) [1][2][3]が盛んである．オフショア開発において，開発作業を委託される側をオフショア側，委託する側をオンショア側と呼ぶ．

一般に，オフショア側とオンショア側とでは文化や言語が異なるため，コミュニケーションにおいて様々な問題が生じる．典型的な問題として，文化の違いによる相互理解の難しさや，言語の違いによる意図の誤解などがあげられる[4]．オフショア開発における様々な問題の中から，我々の研究グループでは，特に言語の違いより発生する「ソフトウェア文書の理解」という問題に焦点をあて，その解決策を求めている．

より具体的には，(a) オフショア側の言語能力不足による文書理解の問題と，(b) オンショア側の相手側の理解状況把握の問題，という 2 つの問題に着目する．問題(a) は，オンショア側から提示される仕様書や指示書を，オフショア側が読むとき，オフショア側の言語能力の不足によりその内容を完全に理解できないという問題である．一方，問題(b) は，提示した文書がオフショア側にどこまで理解されたかを，オンショア側が把握しにくいという問題である．

これら 2 つの問題を改善すべく，本論文ではオフショア開発のための異言語文書理解を支援するシステムを開発する．提案システムでは，問題(a) を解決すべく，言語グリッド Web サービス[5] [6]による機械翻訳機能を採り入れる．オンショア側が登録した文書は，機械翻訳と専門辞書によってオフショア側の言語に翻訳され，オリジナルの文書に併記して訳文が表示される．また，問題(b)に対応すべく，オフショア側が文書の各文に対してどれだけ理解したかを入力するフィードバック機能を採り入れる．これらの機能を実装したシステムのプロトタイプを開発した．

また開発したシステムを用いて 2 種類の評価実験を行った．まず，オフショア側機能の評価実験では，日本語能力を持つ 8 名の中国人に被験者になってもらい，酒屋在庫管理システム[7]の要求仕様を理解してもらう実験を行った．その結果，提案システムを用いた場合，特に日本語経験の浅い被験者の文書理解がより向上することがわかった．さらに，翻訳によって母国語への安心感が得られること，文書全体の俯瞰 (斜め読み) に役立つことがわかった．

一方，オンショア側機能の評価実験では，2 名の日本人に被験者になってもらい，システムに入力されたオフショア側被験者 (中国人) の理解度情報に基づいて，オフショア側が実際にどれだけ文書を理解したのかを推測してもらった．この推測結果と理解度テストの結果を比較したところ，おおむね結果が一致し，提案システムがオンショア側ユーザの理解状況把握に有効であることがわかった．

2. 準備

2.1 オフショア開発

オフショア開発とは，ソフトウェア開発の全体，または，一部を海外の企業に委託する開発形態を指す[1][2][3]．オフショア開発において，開発作業を委託される側をオフショア側，委託する側をオンショア側と呼ぶ．

典型的な例としては，先進国の企業が開発コスト

の削減を狙って，開発の一部を途上国の企業に委託する形態がある．この場合の最大のメリットは，海外の人件費の安さであり，開発全体の劇的なコストダウンを図ることができることにある．代表的なオフショア側国としては，中国やインド，ベトナムといった国が有名である．

図 1 にオフショア開発の例を示す．この例は，ある日本企業がシステムを新たに開発する際に，設計と試験は本社で行うが，開発は中国企業に委託するという例である．ソフトウェア開発では開発費の大部分が人件費であるため，人件費の安い中国に開発を委託することで開発全体の劇的なコストダウンを図ることができる．

2.2 オフショア開発における問題点

オフショア開発では，前節で述べたメリットが受けられる反面，文化や言語の違う人間同士が協調して開発を進めなければならないという独特の難しさがある．そのため，コミュニケーション問題と文化差異という 2 つの大きな問題が生じる[4].

コミュニケーション問題は，言語の違いから相手に意思を伝えにくい，あるいは，相手の意思を理解しにくいという問題である．同じ国の人同士なら簡単に伝えられることでも，外国人に伝えるのは何倍もの時間がかかる．さらに，相手はその一部しか理解できていない可能性もある．文化差異は，文化や習慣の違いから発生する問題である．同じ国の人同士では言わなくてもわかる事でも，外国人には明確に言わないと分からないといったことがある． これにより，オンショア側で重視していることが，オフショア側で気にされないという問題が起こる．

本研究では，オフショア開発における文書を介した異言語コミューニケーションに注目して，そこで起こる下記の 2 つの問題に焦点をしぼる．

問題(a): (オフショア側の文書理解問題) オンショア側から提示される仕様書や指示書をオフショア側が読むとき，オフショア側の言語能力の不足により内容を完全に理解できない問題．

問題(b): (オンショア側の理解度把握問題) 提示した文書がオフショア側にどこまで理解されたかを，オンショア側が把握しにくいという問題．

2.3 オフショア開発のメンバー

通常のソフトウェア開発では，メンバーとして，顧客，開発部隊と SE が存在する．オフショア開発では，また両組織の言語を理解して橋渡しをするブリッジ SE (以下，BSE) が存在する[1][2]．BSE はオフショア側とオンショア側を仲介する役割をし，全てのコミュニケーションは BSE を通して行われる．BSE はオフショア側に滞在することが多く，オンショア側からの指示や文書を，オフショア側にわかるような形で伝達する．また，仕様検討段階からプロジェクトに参加し，プロジェクトの目的や計画を熟知することが望まれる．

しかしながらこの従来のやり方は，BSE に依存しすぎる問題がある．全てのコミュニケーションが BSE 経由となるため，コミュニケーションの効率も悪い．予算的な観点からもあらゆるプロジェクトで常に優秀なブリッジ SE を確保することは難しい．さらに，大規模なプロジェクトでは大量の文書のやり取りが発生し，これらを完璧に橋渡しすることは相当な労力である．以上の理由から，本研究では BSE になるべく頼ることなく，問題(a)(b) の解決を支援する IT システムを開発し，従来手法の改善を目指す．

2.4 言語グリッド

言語グリッドとは，辞書や機械翻訳などの言語資源を，インターネット上から利用できる多言語サービス基盤である[4][5]．言語グリッド運営組織は言語資源提供者が提供している分散的な言語資源をプラットフォーム上に集約し，これらを Web サービスとして公開する．言語グリッド利用者は，公開された Web サービスを自分の用途に合わせてアプ

リケーションに組み入れて利用することができる．また複数のサービスを組み合わせ，新たなサービスを構築することも容易となっている．例えば，機械翻訳サービスとユーザが作成した用例集や，専門辞書等を連携して，より精度の高い機械翻訳サービスを実現できる．本研究では，この言語グリッドをコミュニケーション問題の解決手段として積極的に採り入れる．

3. 提案手法

3.1 概要

前節で述べた問題(a)(b) を解決するため，本稿では図 2 に示すコミュニケーション支援手法を提案する．

具体的には，オンショア側からオフショア側へ文書(図 2 においては仕様書)を送る際，原文に併記して，オフショア側の母国語に機械翻訳した文書を送る．この機械翻訳は言語グリッドを用いて行う．仮に機械翻訳が正確でなくても，原文と一緒に参照することで，オフショア側は内容を理解しやすくなると考えられる．また，ソフトウェア開発の専門辞書を機械翻訳に導入することで，機械翻訳の品質を改善する．この手法を利用して，問題(a) の解決をねらう．

次にオフショア側の開発者は，翻訳が併記された文書を閲覧し，文書の各項目をどれだけ理解したのかを「理解度」のスコアとして記入し，**コーディネータ**(オフショア側の取りまとめ役．BSE が兼任してもよい)に報告する．またスコアだけでなく，どのように理解したのかのコメント，質問なども母国語で記入する．さらに，機械翻訳された文の誤りや欠点を見つけた場合，それらの修正案・改善案をコーディネータに送る．コーディネータは，複数の開発者から集めた翻訳文をとりまとめて清書し，以後該当文書を組織で閲覧する際には，この清書した母国語文書を提示する．理解度のスコア，コメント，質問等は，コーディネータによってオンショア側に報告・共有される．オンショア側は文書のどの項目がどのように理解されたかをオンデマンドに知ることが出来るようになるため，問題(b) を解決できると考えられる．

提案法におけるコーディネータの役割は 2.2 の BSE の役割と類似している．しかし，コーディネータに課される仕事も，求められる能力も，BSE に比べて軽減されている． BSE の場合，オンショア側とオフショア側両方の言語への精通が必須であるが，コーディネータの場合必須ではない．また，BSE の場合は，ソフトウェア仕様と関連専門知識を詳細に理解している必要があるが，コーディネータの場合は，必ずしもその必要は無い．

以降，問題(a) の解決であるオフショア側の理解支援手法と，問題(b) の解決である理解度把握支援について詳細に述べる．

3.2 オフショア側の理解支援手法：言語グリッドウェブサービスを用いた機械翻訳の導入

オフショア側の開発者は，基本的にオンショア側の言語を学習・習得していることが多い．しかしながら，長い時間勉強したとしても，専門用語が多く含まれるソフトウェアの文書を外国語で読むには労力がかかるし，理解するまでの時間もかかる．また，開発者間での言語の習得レベルもばらつきもあり，必ずしも全員が外国語で文書を読むことは難しい．

したがって，外国語で書かれた原文に併記する形で，母国語の翻訳を載せることで，開発者にとっての異言語理解に対するハードルを下げることができると考える．ただし，手動で翻訳を行うのはコストがかかりすぎるため，言語グリッド Web サービスを用いて機械翻訳を行う．この部分を図 2 の実線部分に示す．

機械翻訳の技術は現在かなりの進歩が見られるものの，人間の翻訳品質・精度にはまだまだ及ばない．しかしながら，オフショア側の開発者にとって，

品質が悪くても母国語の訳文があることで，安心感が得られると考える．また，訳文を原文の補助として一緒に読むことで，原文の理解補助になると考える．また，質問がある場合，コーディネータを中心で，開発者の間で話し合って，解決することも考えられる．

機械翻訳の品質を上げるためには，言語グリッドの翻訳サービスに専門辞書サービスを組み合わせて使うことが考えられる．ソフトウェア開発には，ソフトウェア開発に関する一般的な専門用語と，そのプロジェクトに特化した専門用語の 2 種類の専門辞書を用意する必要がある．後者は開発するシステムを適用する業務に関連する専門用語であり，プロジェクトのデータディクショナリ[8]として定義されることがある．例えば，「酒屋在庫管理システム」[7]の仕様書を翻訳する際には，ソフトウェア工学に関する一般的な専門辞書と，酒屋の在庫管理に必要な専門辞書の 2 つが必要となる．

3.3 オンショア側の理解度把握支援：理解度フィードバックによる理解状況の把握

オフショア側に送った文書がどれだけ理解されたかを，何のフィードバックも無しにオンショア側が把握することは難しい．従来ではオフショア側に常駐する BSE に状況を尋ねることが一般的であった．しかし，BSE の主観的な評価・判断が入るため，開発者個人個人の理解状況を，細かく把握することは難しい．オフショア側の理解状況を把握せずにオンショア側がプロジェクトを進めていくと，ついていけなくなって問題が発生しやすくなる．

この問題を解決するために，開発者個人が自らの理解度をフィードバックできる仕組みを考える．この部分を図 2 の点線部分に示す．オフショア側が文書を読む時に，文書の各文をどれだけ理解したのかを「理解度」のスコアとして提案システム画面で記入する．またスコアだけでなく，どのように理解したのかのコメント，質問なども母国語で記入する．さらに，機械翻訳のまずい部分を見つけたら，それらの修正・改善した文書をコーディネータに送る．コーディネータは，複数の開発者から集めた翻訳文をとりまとめて清書する．理解度のスコア，コメント，質問と清書等は，コーディネータによってオンショア側に報告・共有される．

4. 異言語文書理解支援システムの試作

前節で述べた提案手法を言語グリッドツールボックス[9] [10]を拡張する形で試作した．言語グリッドツールボックスは言語グリッド利用のための多言語コミュニケーションツールの集合であり，XOOPS と呼ばれるポータルサイト構築システム上で動作する．以降では試作したシステムの概要について説明する．

4.1 オフショア側支援機能

3.2 で説明したオフショア側の理解支援手法を実現するシステム画面を説明する．

オフショア側の開発者は，まずシステムにログインする．ログイン後，ユーザは文書一覧画面にて，オンショア側がシステムに登録した文書の一覧を確認する．閲覧したい文書を選択すると，原文に併記される形で機械翻訳された文書が表示される．

図 3 に文書閲覧画面を示す．この例では，オンショア側が日本語で書かれた「酒屋在庫管理システム」の要求仕様書を，オフショア側である中国人開発者が閲覧するという設定である．仕様書の原文は一文ごとに分解され，項目 ID が振られる（画面の最左端）．各項目に対して，左上がオリジナルの文章であり，左下が機械翻訳された文章である．各ユーザはこの訳文を編集することが可能となっている．右上はコーディネータが提示するコメントであり，右下は開発者が入力するコメントである．これらのコメント欄は，該当項目に関する質問や自分の理解を，開発者とコーディネータの間でやり取りするために利用され，開発者全員で共用することもで

きる．最右端の欄が理解度情報となっており，開発者それぞれが文章ごとの理解度を 5 段階で入力する．

次にコーディネータが使用する画面を説明する．コーディネータがシステムにログインし，作業する文書と項目番号を選択すると，図 4 に示すコーディネート画面に入る．この例では，ある文書の項目 7 に関する 10 名の開発者の理解度とコメントをまとめる画面となっている．

画面上半分が 10 人の開発者全員の理解情報である．左側からそれぞれ，項目 ID，修正した訳文，理解度スコア，コメント，開発者 ID を示している．下半分は左側から順に，日本語の原文，コーディネータが開発者の修正訳文に基づいて清書する訳文，コーディネータのコメントと全体の理解度を入力する場所である．

コーディネータは開発者全員のコメントを読んで，皆の理解に役に立ちそうなコメントを母国語で作成する．さらに，開発者の理解度スコアに基づき，コメントを考慮した上で，オフショア側全体の理解度を入力する．開発者全員の文書閲覧が完了すると，コーディネータはこれらの理解情報をオンショア側に提出する．その際に，清書した訳文とコーディネータのコメントが，システムよりオンショア側の母国語に逆翻訳される．

4.2 オンショア側支援機能

オンショア側のユーザは，はじめに，オフショア側に委託するソフトウェア文書をシステムに登録する．システムは登録された文書を一文ごとパースし，言語グリッド Web サービスを利用してオフショア側の母国語に翻訳する．

またオンショア側のユーザは，登録した文書のオフショア側の閲覧状況を確認できる．理解度の確認画面を図 5 に示す．各項目に関して，原文と逆翻訳された文書を比較し，ドキュメントの各文章が意図したとおりに読まれたかを確認する．また，コメントと理解度スコアも参考して，オフショア側の理解情報を推測することができる．これにより，オフショア側の理解情報を把握する．

オフショア側の理解度が低い文章については，コメントに基づいてオンショア側で文章を最修正し，再びオフショア側に送付される．

4.3 システム実装

試作システムは言語グリッドツールボックス上で動作している．サーバー側のプログラムミング言語としては PHP を用い，各文章や理解度を保存するデーターベースサーバーは MySQL を利用した．また HTTP サーバーには Apache2 を利用した．クライント側のプログラミング言語は HTML と JavaScript を利用した．

5. 評　価

提案システムの有効性を確認するため，日本－中国間のオフショア開発を模擬した被験者実験を行った．オンショア側である日本企業が「酒屋在庫管理システム」[7]の仕様書（原文は日本語）を，オフショア側である中国に送付し理解してもらうという設定である．実験は 2 種類行う．提案システムのオフショア側支援機能（4.1 参照）を評価する実験と，オンショア側支援機能（4.2 参照）を評価する実験の 2 つである．また本章では，実験から得られた知見を元に，浮かび上がった課題についても考察する．

5.1 オフショア側支援機能の評価実験

5.1.1 実験概要

この実験では，酒屋在庫管理システムの仕様書を中国人被験者に読んでもらい，その理解度を試験問題により評価する．仕様書には，倉庫に保管された酒の在庫管理，及び顧客からの発注に対する配送処理が書かれている．仕様書は全て日本語で記述されており，その規模は A4 で 7 ページ程度である．

被験者は8 名の日本語能力を持つ中国人であり，その内訳は 4 名の留学生と 4 名の IT 系企業の社員である．日本語能力に関しては，4 名がビジネスレベルであり 4 名が日常会話レベルであった．

各被験者には同一の仕様書を，提案システムを用いる場合と用いない場合の 2 通りの方法で読んでもらい，理解度をチェックする試験問題を実施してもらう．またそれぞれの読解に要した時間も計測する．実験においては同一の仕様書を二度読むことになるため，慣れに対する対処が必要である．よって，2 回目の実験を 1 回目から 1 週間以上空けて行い，試験問題の内容を少し変える対処を行った．また，実験順序による偏りをなくすため，4 名の被験者には1回目システムあり，2回目システムなしの順で，残りの4名は逆の順序で実験を行ってもらった．各被験者の能力，及び実験順序を表 1 に示す．

また仕様書の理解度を評価するため，6 問の試験問題を用意した．具体的な内容は，倉庫の最大容量に対する簡単な問題から，ある倉庫の状況に対し，ある発注依頼が発生した際にどのような伝票がどの係に発行されるかといった，複雑な問題まで含まれる．問題 1 から問題 6 になるほど高い理解度を要する複雑な問題となっている．各問題は 5 点満点で採点を行った．提案システムの定性的な評価を得るため，実験の最後にアンケートを実施した．

5.1.2 実験結果

提案システムを用いた場合の理解度試験の結果を表 2 に，システムを用いなかった場合の結果を表 3 に示す．システムありとなしの理解度総平均点は，それぞれ 4.6 と 4.1 となった．問題難易度ごとの理解度支援の効果を見るために，問題ごとの理解度平均をまとめたグラフを図 6 に示す．問題 1 は全ての被験者が誤りなく正答しているが，それ以外の問題についてシステムを用いることで理解度が向上していることが読み取れる．

また，ビジネスレベルと日常会話レベル別の理解度平均の比較を図 7 に示す．概ねビジネスレベルの被験者の方が理解度は高い傾向にある．提案システムは日常会話レベルの被験者に対し，より高い効果が得られたことが分かる．

一方で，システムを利用することで文書の読解時間が増加する傾向が見られた．図 8 に仕様書読解に要した時間を示す．読解時間は被験者の能力や性格に依存しやすく，被験者ごとのばらつきが激しい．全体としては，提案システムを利用することで読解に要する時間が増加する傾向にあった．その具体的な値は，システムありで約 50 分，システムなしで約 40 分であった．

5.1.3 考察

実験結果から，提案システムを用いることで日本語の仕様書の理解に対して，ある程度の効果が得られることが分かった．図 6 においては特に問題 3 と 4 で理解度の差が大きくなっており，中程度の複雑な問題に対して支援効果が高かった．また図 7 の結果から，日本語経験の浅い被験者に対して文書理解の効果が得られやすいと言える．

実施したアンケートから得られたコメントの代表的なものについて紹介する．まずシステムに対する肯定的な意見として，「機械翻訳であっても母国語に変換されていると安心である」，「翻訳があるので，難しい単語と複雑な表現が理解し易い」，「システムを使うことで，文書全体の俯瞰に役立つ」といった意見が得られた．機械翻訳による訳文の併記は，訳文自体が読みづらいものであっても，理解の手がかりとして有用であったといえる．

一方システムに対する否定的な意見として，「翻訳精度によっては誤解も発生しやすい」，「長文の翻訳精度が特に低い」といった機械翻訳の精度に対するコメントが得られた．

5.2 オンショア側支援機能の評価実験

5.2.1 実験概要

この実験では，オフショア側の開発者が仕様書をどこまで理解したかを，システムに入力された理解度情報に基づいて，オンショア側で正確に推測できるかどうかをチェックする．

実験の流れは以下の通りである．まず，オフショア側開発者である中国人被験者に，システムを用いて酒屋在庫システムの仕様を読んでもらう．その後，各開発者は，理解度スコアと訳文の修正，およびコメントをシステムに入力し，コーディネータに報告する．同時に5.1の実験と同様，理解度に関する試験問題も受けてもらう．コーディネータは開発者全員の理解度情報に基づき，訳文とコメントを清書し，全体の理解度をシステムに入力，オンショア側に報告する．オンショア側はフィードバックされた理解度情報と逆翻訳された訳文・コメントから，オフショア側の実際の理解状況を推測する．推測した結果を，開発者の理解度試験の結果と照らし合わせて，理解状況推測の精度を評価する．

この実験では，5.1の実験の仕様書を2ページに短縮したものを用いた．オフショア側の被験者は，6名の中国人であり，全員が日常会話レベルの日本語能力を持つ留学生であった．オンショア側の被験者は2名のソフトウェア工学の博士号を持っている大学教員である．理解度の推測においては，仕様書を7つのセクションに分け，部分ごとに5点満点で理解状況を推測した．

5.2.2 実験結果

実験結果を表4に示す．表の各行は，前節で述べた仕様書の7つのセクションを表す．表の左半分は，オフショア側で取得した開発者の理解情報を表す．2列目は開発者がシステムに入力した理解度スコアを，セクションごとに平均を取ったものである．3列目は，セクション内の項目の理解度の最小スコアを表している．4列目は，実施した理解度試験問題の得点の平均を表しており，これが開発者の実際の理解度を反映したものになっている．

表の右半分は，オンショア側で推測した（オフショア側の）理解状況を，被験者2名分記入している．各エントリには推測値に加えて，括弧内に試験得点との差分を表す．この差分は見積もりと実際の誤差を表す．14件の推測値（7項目×2被験者）のうち，誤差なく見積もれたのは4件，誤差1が6件，誤差2が4件であった．誤差の平均は-1，分散は0.83であった．

5.2.3 考察

表4の結果から，被験者1,2の理解度の把握は比較的精度が高く，本実験の範囲内ではオンショア側支援機能の有効性が示された．

見積もりの誤差が2となった4件の内訳を見ると，すべてのケースでオンショア側がオフショア側の理解度を少なく（つまり厳しく）見積もっていることが分かる．また，オンショア側の見積もりは，オフショア側がシステムに入力した理解度の最低値と相関が高くなっており，1人でも理解度が低いと「理解が甘い」と厳しく判断したようである．これらについて，オンショア側被験者への事後インタビューから，以下のようなことがわかった．

・業務そのものが難しいところは日本語表現も複雑になりがちだが，コメントが少ないと本当に理解できたのか不安であった．

・コメントがあっても，本質をついたものでない場合，本当に理解したのかを疑った．

・修正・清書されなかった機械翻訳文が，そのまま日本語に折り返し翻訳され，品質劣化が起きた文を見て不安になった．

こうしたオンショア側の不安は，オフショア側とのコミュニケーションが不十分であることを反映している．オンショア側にこうした不安がある場合には，その不安をオフショア側にもう一度フィードバックし，意識のすり合わせをすることで解消できると考える．この機能については，協調型翻訳[11]のような技術を参考にして，将来的に実装していき

たい.

5.3 課　題

評価実験のアンケートや事後インタビューから，ユーザが提案システムに求める重要課題として，機械翻訳の精度向上があることがわかった．この問題に対して現在2つの対処方法を検討している．一つは専門辞書による翻訳精度の改善である．要求仕様書などのソフトウェア文書には，固有の名詞や単語が多く出現するが，機械翻訳で処理しにくい特徴がある．例えば今回の実験では「在庫不足票控」という固有名詞が「在庫不足票の待機」という意味に翻訳されており，誤解や理解の妨げとなっていた．プロジェクト固有の表現を翻訳する専門辞書を用意することで改善できると考える．

もう一つの対処は，翻訳対象である日本語の原文を機械翻訳しやすい形に整形することである．日本語は主語や補語の抜けを許容しやすい特徴があるが，機械翻訳の精度低下の原因となりやすい．折り返し翻訳を利用して機械翻訳しやすさの尺度を計測するアイデアがある [11]．これを用いて，オンショア側文書の作成者やコーディネータが文章を整形して，翻訳精度を改善できると考える．

また，提案システムのインターフェースの改善案として「検索機能とハイライト機能」，「翻訳インターフェースの改善」等の要望意見が寄せられた．これらのインターフェースの改善を行うことで，システムのユーザビリティを確保し読解時間増加の問題を解決できると考えられる．これらは今後の研究課題としたい．

6. まとめ

本研究では，オフショアソフトウェア開発における2つの問題，すなわち，(a)オフショア側の文書理解と(b)オンショア側の理解度把握を解決することを目的として，異言語文書の理解支援手法を提案した．また，提案手法を言語グリッド Web サービスを用いて実装し，評価実験を行った．実験では，日本と中国間のオフショア開発を模擬し，提案システムの効果を確認した．オンショア側とオフショア側の2つの実験より，提案システムが上記2つの問題の解決に役立つことが明らかとなった．

今後の課題としては，専門辞書による翻訳精度の向上，および，システムインターフェースの改善が挙げられる．

謝辞

本研究の一部は，科学技術研究費（基盤研究 C 24500079，基盤研究 B 23300009），および，関西エネルギー・リサイクル科学研究振興財団の助成を受けて行われている．

参考文献

[1] 齊藤邦夫，“中国オフショア開発におけるコミュニケーション・マネジメント: オフショア開発成功の鍵，” Journal of the Society of Project Management, vol.9, no.1, pp.26–31, 2007.

[2] 齊藤邦夫，“中国オフショア開発のプロセス改善への取り組み，” 日立 TO 技報, vol.15, no.1, pp.72−76, 2007.

[3] 韓旭，阿部智和，“層別リテンション・マネジメントに関する探索的研究－日系在中ソフトウェア企業における中国人ソフトウェア技術者の事例分析－，” 經營と經濟，vol.90(4), pp.115-146, 2011.

[4] 林田尚子，石田亨，“日本－中国共同ソフトウェア開発の観察異文化コラボレーション支援に向けて，” 情報処理学会研究報告．vol.49, pp.25−30,2005.

[5] 石田亨，“異文化コラボレーション研究の構想，” 異文化コラボレーション研究グループ，2006. http://langrid.nict.go.jp/report/report1.htm

[6] Toru Ishida. “Language Grid: An Infrastructure for Intercultural Collaboration,” Proc. IEEE/IPSJ Symposium on Applications ad Internet,

pp.96-100, 2006.

[7] 山崎利治, 共通問題によるプログラム設計技法解説, 情報処理, Vol. 25, No. 9, pp. 934−935 (1984).

[8] Will Tracz, Lou Coglinese, Patrick Young, "A Domain-Specific Software Architecture Engineering Process Outline", ACM SIGSOFT SOFTWARE ENGINEERING NOTES vol.18, no. 2, pp40-49, 1993.

[9] 田仲正弘, 村上陽平, 林冬惠, 石田亨, "言語グリッド Toolbox : 多言語コミューニケーション支援ツールのための開発フレームワーク(言語グリッドと異文化コラボレーション)," 電子情報通信学会技術研究報告. 人工知能と知識処理, pp.13-18, 2011.

[10] 田仲正弘, 村上陽平, 林冬惠, 石田亨, "1H-4 Language Grid Toolbox:多言語コミュニティ支援のためのオープンソースソフトウェア(グループウェアー般, 一般セッション, インターフェース, 情報処理学会創立 50 周年記念)," 全国大会講演論文集, pp."4-59"-"4-60" ,2010.

[11] Daisuke Morita and Toru Ishida. "Designing Protocols for Collaborative Translation," Pacific Rim International Conference on Multi-Agents, 2009.

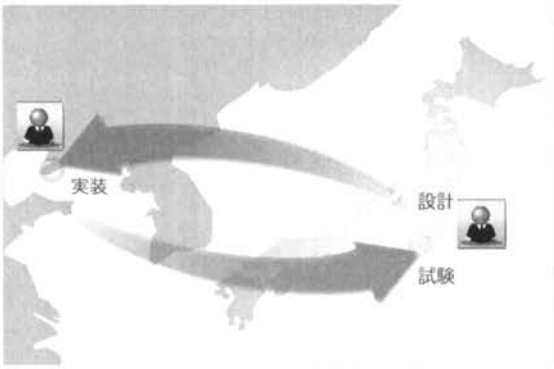


図 1 オフショア開発

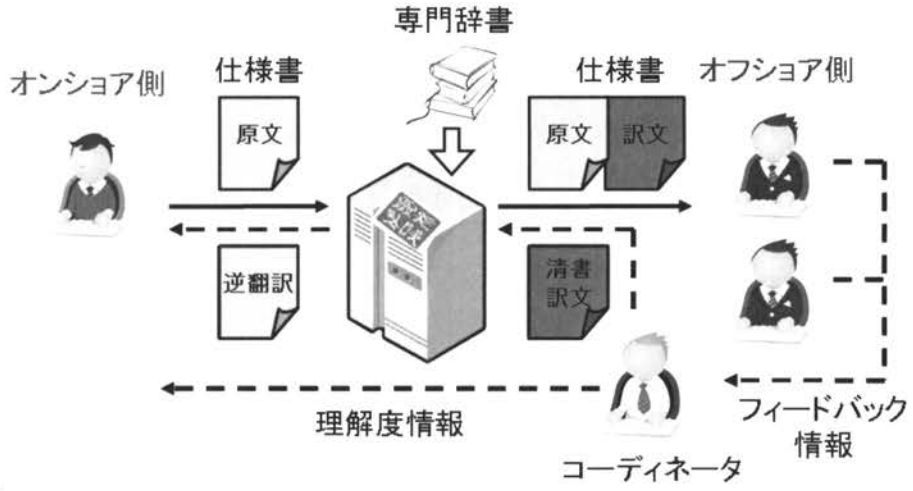


図 2 提案手法全体図

DOCUMENT READING

ID	Orignal/Translated Text	Recommendation/Comment	Understanding
1	酒管理プログラム要求仕様書	全体OK	
	酒管理程序要求说明书	完全理解	5
2	1．要求仕様	推荐翻译 要求说明	
	1.要求配置	完全理解	5
3	ある酒屋では，すべての在庫(酒）を幾つかのコンテナに載せて，一つの倉庫に保管している．	某酒店，所有的酒都存放在多个货柜上，在一个仓库里保是管 container是集装箱或货柜	
	在某酒店,为几个的集装箱登载全部的库存(酒)，为一个仓库保管着．	完全理解	5
4	「受付係」の仕事は，倉庫で在庫の積み下ろしを行う「倉庫係」，及び，客の注文を聞いて客に酒を届ける「配送係」との間で数種類の伝票をやりとりしながら，倉庫内の在庫を管理することにある．	「接待员」的工作是在负责仓库进行库存的装载卸的「仓库员」和听客人的订购送交酒给客人的「发送员」之间传递票，并管理仓库内的库存係的翻译为部门比较恰当	
	「接待员」的工作,在在仓库进行库存的装载卸的「仓库员」,和,听客人的订购送交酒给客人的「发送员」间一边交换数种类传票,一边在于管理仓库内的库存．	完全理解	5

図3　オフショア側の文書閲覧画面

DOCID : 0115185530　ITEMID : 7

ITEMID	TRANSLATION	UNDERSTANDING	COMMENT	READER
7	并且仓库员和发送员的输出不被区别,并被输入到程序.	5		admin
7	还有,库存科和配送科的输出不被区别,被输入到程序.	5	OK	test1
7	还有,仓库员和发送员的输出不被区别,被输入到程序.	5	完全理解	test2
7	仓库员和发送员的输出不被区别,被输入到程序.	5	ok	test3
7	还有,仓库员和发送员的输出不被区别,被输入到程序.	5	ok	test4
7	还有,被输入到程序仓库员和发送员的输出不被区别.	5	OK	test5
7	还有,仓库员和发送员的输出不被区别,被输入到程序.	5	改成“不区别仓库员..., 并被....“	test6
7	还有,仓库员和发送员的输出不被区别,被输入到程序.	0		test7
7	还有,仓库员和发送员的输出不被区别,被输入到程序.	0		test8
7	还有,仓库员和发送员的输出不被区别,被输入到程序.	0		test9

Orinal Text	Translated Text	Comment	Understanding
また，倉庫係と配送係の出力は区別されず，プログラムへ入力される．	还有,仓库员和发送员的输出不被区别,被输入到程序.	改成“不区别仓库员..., 并被....“	5

submit

図4　オフショア側のコーディネート画面

DOCID : 0115185530　FILENAME : 酒管理プログラム要求仕様(一部).txt

ITEMID	TEXT/RETRANSLATION	COMMENT	UNDERSTANDING	DATE
1	酒管理プログラム要求仕様書 酒の管理プログラムは説明書を求める	全体OK	5	0000-00-00 00:00:00
2	1. 要求仕様 1.配置を求める	翻訳(通訳)を推薦する(薦める) 説明を求める	5	0000-00-00 00:00:00
3	ある酒屋では，すべての在庫(酒) を何つかのコンテナに載せて，一つの倉庫に保管している． 某ホテルで，いくつ(か)のコンテナのために全部の(全て)在庫(手持ち資金)(酒)を掲載して，1つの倉庫のために保管している(保証する)．	某ホテル，すべての酒はすべて今 財のコンテナの(商品陣列台)上で	4	0000-00-00 00:00:00
4	「受付係」の仕事は，倉庫で在庫の積み下ろしを行う「倉庫係」，及び，客の注文を聞いて客に酒を届ける「配送係」との間で就種類の伝票をやりとりしながら，倉庫内の在庫を管理することにある． 「接待員」の作業(仕事)，倉庫が在庫(手持ち資金)のを行って下ろす(外す)「倉庫員」を積載して，，と客のを聞いて(任せて)酒に客「発送(送信)員」の間で一方でに交換することを種類の召喚状(伝票)を就える引き渡すこ	「接待員」の作業(仕事)は倉庫の を担当して在庫(手持ち資金)のを	5	0000-00-00 00:00:00
5	1.1 受付係への入力 1.1接待員の入力(輸入)に向って	1.1接待員の入力(輸入)	5	0000-00-00 00:00:00
6	受付係への入力は，プログラム外部から倉庫係と配送係の出力として，ファイルまたは標準入力より与えられる． 接待員の入力(輸入)に向って，プログラムの(順序)外部から倉庫員と発送(送信)員の出力(輸出)にして，ファイル(文書)はあるいは標準的な入力(輸入)から与えることができる．	標準的な入力(輸入)からファイル (文書)を与えられるあるいは	4	0000-00-00 00:00:00
7	また，倉庫係と配送係の出力は区別されず，プログラムへ入力される． ある，倉庫員と発送(送信)員の出力(輸出)は区別されないで，入力(輸入)にプログラム(順序)まで(へ)．	“に変えて倉庫員を区別しな い...., そして(合わせて)掛け布	5	0000-00-00 00:00:00
8	(A1) 積荷票 (A1)積載する品物の切符(チケット)	=オランダの(ハス)切符(チケット) は積荷明細書に類似して入る	5	0000-00-00 00:00:00
9	酒を積んだコンテナが倉庫に搬入されると，倉庫係から「積荷票」が送られてくる． もし酒のコンテナを積載して倉庫に運ばれて(引っ越しされて)入る(入れる)ならば，倉庫員「積載する品物の切符(チケット)」に郵送される．	「積載する品物の切符(チケット)」 は倉庫員発行から。	5	0000-00-00 00:00:00
10	1 台のコンテナに酒を10 銘柄まで混載できるため，積荷票には，搬入されたコンテナの番号(コンテナに付けられた一意なシリアル番号)，搬入年月，搬入日時と共に，積載銘柄就が記入されている． (中降)10品種の様量の1台のコンテナの酒まで(へ)混じって(混ぜて，無為に過ごして)載せて，積載する品物の切符の(チケット)上で，入る(入れる)コンテナの番号(コンテナの取るただ1つの直列の番号に)を運ばれること(引	1つのコンテナ(商品陣列台)は10 種類の酒を詰める(設置する，装	3	0000-00-00 00:00:00
11	さらに，積載されている銘柄と銘柄毎の就量(積載量) が積載銘柄就だけ繰り返し記入されている． しかも，品種と品種の就量(量を積載する)を積載してただ品種就の再現だけを積載して書き込むと．	しかも，＝柄就は品種と就量の繰り 返し(重複)だ	4	0000-00-00 00:00:00
12	積荷票|コンテナ|番号|搬入年月|搬入日時|銘柄就(銘柄，就量) の繰り返し 積載する品物の切符(チケット)|コンテナ|番号|運んで(引っ越しして)歳月に入る(入れる)|運んで(引っ越しして)期日と時間に入る(入れる)|品種は(品種，就量)の繰り返しを就える	全従業員OK	4	0000-00-00 00:00:00
13	(A2) 出庫依頼票 (A2)委託の切符(チケット)を出す	全従業員OK	5	0000-00-00 00:00:00
14	客から酒の注文があると，配送係から「出庫依頼票」が送られてくる． もし客から酒の予約積入があるならば，発送(送信)員「出して切符(チケット)を頼む」に移される(移られる)．	客が酒の予約積入がある時，発送 (送信)員は願いの切符(チケット)	5	0000-00-00 00:00:00
15	出庫依頼票には，出庫すべき酒の銘柄(1 銘柄のみ)，就量，及び，送り先名が記入されている． 出して切符(チケット)を頼む上に，ただ酒の品種(1品種)を出すだけべきで，就量，，と配達して点呼して(指名して)書かれる．	出して切符(チケット)を頼む上 に，ただ記録してすべきに酒の品	5	0000-00-00 00:00:00

図5　オンショア側の理解度確認画面

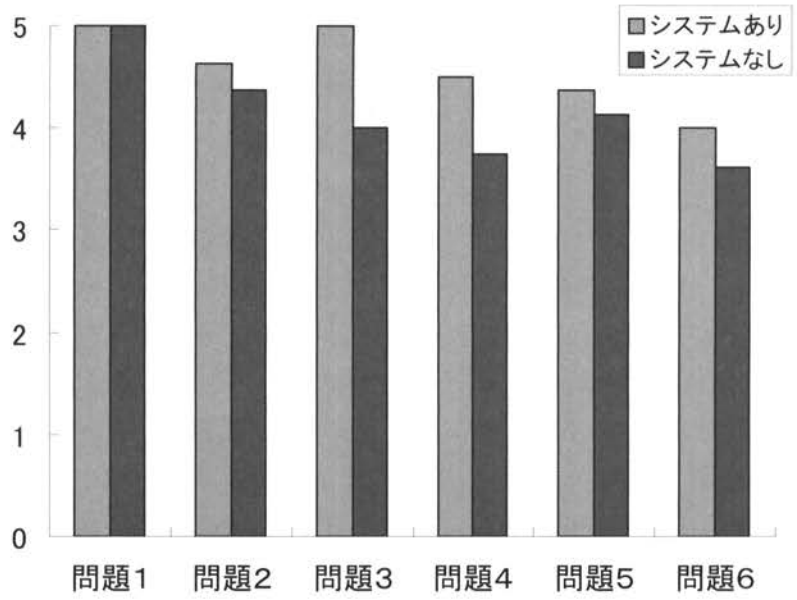


図6　問題別の理解度平均

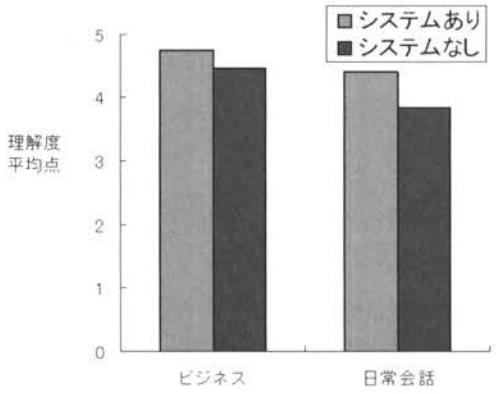


図7　日本語能力別の理解度平均

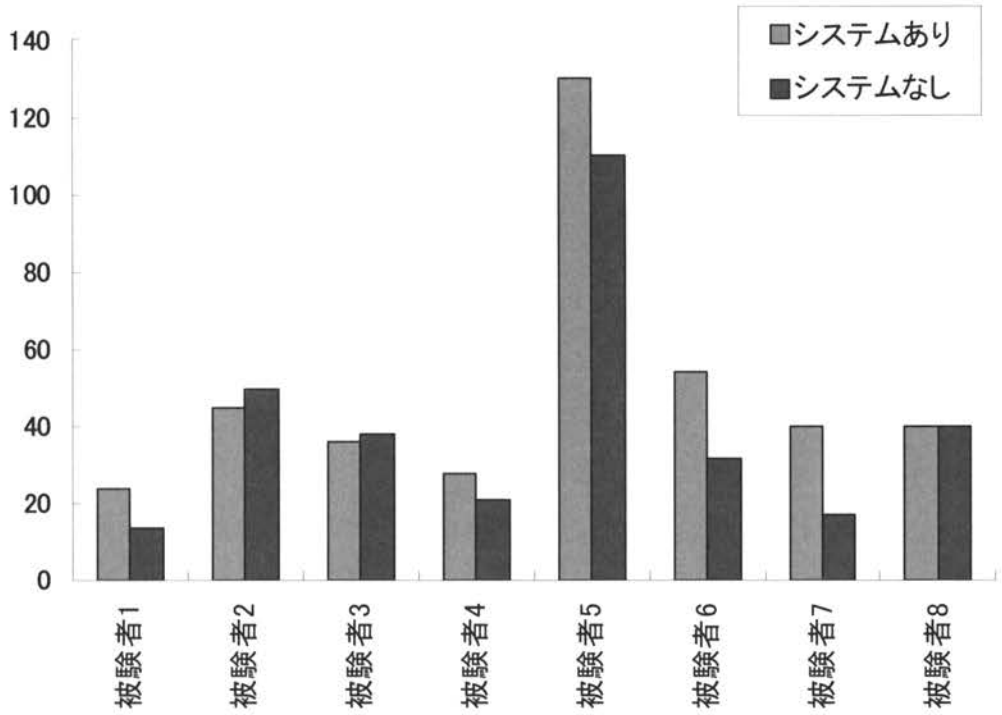


図 8 実験結果：使用時間

表 1　被験者のプロフィール

	実験順序	日本語能力	学生/社員
被験者1	システムなし先行	ビジネス	学生
被験者2	システムなし先行	日常会話	社員
被験者3	システムなし先行	ビジネス	学生
被験者4	システムなし先行	ビジネス	学生
被験者5	システムあり先行	日常会話	社員
被験者6	システムあり先行	日常会話	社員
被験者7	システムあり先行	日常会話	学生
被験者8	システムあり先行	ビジネス	社員

表 2　理解度試験の結果（提案システムあり）

	問題1	問題2	問題3	問題4	問題5	問題6	平均
被験者1	5	5	5	5	4	5	4.8
被験者2	5	5	5	5	4	5	4.8
被験者3	5	4	5	5	4	5	4.7
被験者4	5	5	5	5	5	5	5.0
被験者5	5	3	5	3	4	3	3.8
被験者6	5	5	5	5	5	3	4.7
被験者7	5	5	5	3	4	3	4.2
被験者8	5	5	5	5	5	3	4.7

表3　理解度試験の結果（提案システムなし）

	問題1	問題2	問題3	問題4	問題5	問題6	平均
被験者1	5	5	5	5	5	3	4.7
被験者2	5	5	5	5	3	3	4.3
被験者3	5	5	3	3	5	5	4.3
被験者4	5	5	5	3	5	3	4.3
被験者5	5	2	2	2	3	4	3.0
被験者6	5	3	5	5	5	5	4.7
被験者7	5	5	2	2	3	5	3.7
被験者8	5	5	5	5	4	1	4.2

表4　オンショア側評価実験の結果

仕様書セクション	オフショア側で取得した理解情報			オンショア側で推測した理解状況	
	理解度平均	理解度最小	試験得点	被験者 1	被験者 2
1. 積荷票の様式	4.2	3	5	4 (-1)	3 (-2)
2. 出庫依頼票の様式	5.0	5	5	5 (0)	4 (-1)
3. 在庫不足表の様式	4.9	4	4	4 (0)	4 (0)
4. 出庫指示票の様式	4.0	3	5	3 (-2)	3 (-2)
5. 依頼番号の付加業務	5.0	5	4	5 (+1)	4 (0)
6. 在庫不足票の作成業務	5.0	5	5	4 (-1)	3 (-2)
7. 出庫指示票の作成業務	4.3	3	3	2 (-1)	2 (-1)

Project Report

The potential and significance of readability research

SHIBASAKI, Hideko

Nagaoka University of Technology

1. An overview of Japanese readability research

Readability is the ease with which texts can be read, and the purpose of measuring readability is to objectively rate the difficulty of a text using a numeric scale. Readability research began in the late 19th century and reached its peak in the 1940s in the United States when criteria were needed to correlate reading materials with grade levels, taking into account the English literacy rates of immigrant children. Currently, there are more than 200 methods for determining readability in various languages, including English, Chinese, Korean, Danish, French, Spanish, Hebrew, and Vietnamese, and many readability measurement tools can be downloaded for free from the Internet.

In Japan, Ichiro Sakamoto and Sukeyori Shiba began readability research in the 1960s. Sakamoto's measurement method was to repeat manual calculations with many steps, and the method was never presented as a single coherent formula. Following the rapid spread and advancement of computer technology, corpus linguistics and natural language processing were born, and it became possible to analyze large amounts of text quickly. Three recent studies use this technique.

The first of these studies is TxReadability, which uses the formula of Tateishi, Ono, and Yamada (1988) and is currently available online at http://www.utexas.edu/disability/ai/resource/readability/. This study is constructed as a database of 77 journal papers, translated documents, magazine articles, and primers with a numeric scale from zero to 100. The variables include the average length of one sentence, the relative frequency of the same character type in a series (kanji, hiragana, katakana, romaji), the average length of each successive casing, and the ratio of the number of commas to the number of punctuations. Additionally, the formula specifies that (1) the longer the sentence, the more difficult cognitive processing will be, and (2) reading is difficult when the same character type is repeated. With the average value of the 77 texts in the database set to 50, the relevant texts are rated closer to 100 when the texts are easier and are rated closer to zero when the texts are more difficult. Because research shows that it is harder to read Japanese when the same character casing continues (Shibasaki, 2010), the variables in this formula are considered valid; however, Shibasaki and Tamaoka (2010) claim that materials with many hiragana for lower elementary school grades cannot be evaluated because the database comprises reading materials for adults.

The second study is called Kotoba Fushigibako, which implements the outcomes of Sato, Matsuyoshi, and Kondoh (2008) and is accessible to the public online at http://kotoba.nuee.nagoya-u.ac.jp/. This method is a 13-level letter model based on the probability of the characters' (unigram) occurrence and the likelihood of these 13 levels is plotted according to difficulty for the relevant text; subsequently, the optimal estimate value is output. Although this method accurately determines the grade level for lower elementary materials, the grade determination for middle or higher elementary materials may be extremely high because only a limited number of kanji by grade are used for the lower elementary materials, whereas kanji characters other

than those commonly used appear in the middle and higher elementary materials.

The third study is the Readability Research Laboratory, which is based on the work of Shibasaki and Tamaoka (2010) and is available online at http://readability.nagaokaut.ac.jp/. In the model of Shibasaki and Tamaoka (2010), a first-order equation for grade determination is denoted where the variables are the ratio of hiragana in all characters of a text and the average number of predicates in one sentence. This equation includes the following variables: (1) the longer the sentence, the more difficult cognitive processing tends to be; (2) cognitive processing with more propositions takes more time if the length of the sentence is the same; and (3) cognitive processing with more complicated grammatical structures is more difficult. In this formula, the determined grade does not deviate significantly from the actual grade, and the accuracy of the grade determination is good, especially from the middle grades of elementary school to the third year of junior high school. However, only the range from the first grade of elementary school to the third year of junior high school can be evaluated, and the results for the reading materials for adults are inconsistent.

Thus, each method has strengths and weakness, and it is necessary to use the methods properly according to their different purposes. The first method should be used to measure the difficulty of reading materials for adults, with an emphasis on academic papers. The second method should be used to rate the difficulty of texts based on the types of kanji used, and the third method should be used to examine the correspondence between language textbooks and the text elements. Hence, there is no universal method as long as a readability formula is formulated with specific variables and measures. The same conclusion applies to readability formulas in other languages, which should be employed according to the specific variables used. At conference presentations, certain questions regarding readability are often posed, such as, "My child is in the fourth grade of elementary school and is able to read Tolstoy. How does the significance of readability research apply?" This type of question reflects a misunderstanding or lack of awareness by the person who poses the question and may indicate a wider misunderstanding by many people. Readability researchers are not recommending specific books for grade levels, and we must be cautious when interpreting results. For example, when a text is rated at the fourth-grade level according to a certain readability formula, this result does not mean that fourth graders can understand the text. The implication is merely that the quantitative analysis of the relevant text's components is close to that of fourth-grade texts analyzed from corpus data. In any case, no text should be proposed for all fourth-grade children to understand equally. Readability research is not a quest for truth; rather, the methods apply the difficulty of texts to certain criteria and set standards to be accepted by as many people as possible, if not everyone.

2. The significance of readability research

How should we determine the validity of a readability formula? In the pioneering days of readability research, Sukeyori Shiba stated that the "difficulty scale of a text is not an unchangeable absolute element that belongs to the text, but it is a relative index of whether text A is easier or more difficult than text B." In other words, if a relative grade-level differentiation is achieved by a formula, then that formula is successful. One example is the rating outcomes of two texts determined by English readability formulas. Table 1 shows

the readability results of *The Adventures of Tom Sawyer* by Mark Twain and *Alice's Adventures in Wonderland* by Lewis Carroll, which were scored by six methods. Only the Flesch reading ease score contains scales from zero to 100, whereas the other five formulas use grades as a measurement. The results reveal that the determined grades do not match and that they bear a large discrepancy because *The Adventures of Tom Sawyer* is rated at the second year of the junior high level (7.8) by the Coleman-Lian index and at the first year of the college level (13.5) by the Gunning Fog index. Similarly, *Alice's Adventures in Wonderland* illustrates a significant difference because the Flesch-Kincaid grade level scores this text at the fifth-grade level (4.8), whereas the Gunning Fog index places it at the second year of the junior high level (8.0). However, because *The Adventures of Tom Sawyer* is rated between the grades of 7.8 and 13.5 and *Alice's Adventures in Wonderland* is between the grades of 4.8 and 8.0, the former is evaluated by both procedures as having higher difficulty than the latter. Moreover, according to the Flesch reading ease score, which indicates difficulty by points, *The Adventures of Tom Sawyer* is also more difficult, with a score of 72.4, whereas *Alice's Adventures in Wonderland* is scored at 88.4. These ratings may be used as a guide for teaching reading by designating all grades above a certain level as recommended rather than recommending only one grade. For example, in some cases, all grades above the fourth elementary grade are suggested for *Alice's Adventures in Wonderland*, whereas in other cases, the book is specified from the fourth grade of elementary to the second year of junior high school.

What are the research outcomes for the readability of Japanese texts? Expanding the results of Shibasaki and Tamaoka (2010), Shibasaki and Hara (2010) constructed three types of readability formulas for 12 grades based on a corpus of language textbooks. Let us select one of the formulas and examine its grade determination results with another set of grade determination results produced by Kotoba Fushigibako. The books chosen were bestsellers published relatively recently and were well known to the general public; specifically, the books were published between 2005 and 2007 with more than 100 million copies in circulation as of March 2008. The selected six books are *Homuresu Chugakusei* (Kirin, Hiroshi Tamura, 2007), *Saodakeya wa naze tuburenainoka?* (Masaya Yamada, 2005), *Kokka no hinkaku* (Masahiko Fujiwara, 2005), *Josei no hinkaku* (Mariko Bandoh, 2006), *Tanin o mikudasu wakamono* (Toshihiko Hayami, 2006), and *Donkanryoku* (Junichi Watanabe, 2007).

Table1. Readability of two novels

readability formulas	The Adventures of Tom Sawyer	Alice's Adventures in Wonderland
Flesch reading ease score	72.4	88.4
Automated readability index	11.1	5.7
Flesch-Kincaid grade level	9.5	4.8
Coleman-Liau index	7.8	6.4
Gunning Fog index	13.5	8.0
SMOG index	11.0	7.5

Additionally, the so-called 'Keitai novels', which were written using cell phone text messages, were evaluated. These novels are a new form of Japanese literature that has appeared since 2006, and many 'Keitai novels' have become bestsellers with more than 100 million copies in circulation. We thought it would be interesting to evaluate the difficulty of these texts that are currently popular among young people because the authors and readers are primarily high school students. Two 'Keitai novels' were selected: *Kurianesu* (Towa, 2007), which won the Japanese Keitai novel first grand prize, and the second grand prize winner, *Shiroi jaji sensei to watashi* (reY, 2008).

In addition, we thought it would be interesting to rate multiple translations of the same book to examine the differences in readability depending on the translators. Because *Le Petit Prince* by Saint-Exupéry has been translated into many different Japanese versions as *Hoshi no ohjisama,* we selected three translations by Aroh Naito (1953), Masahiro Mita (2006), and Yumiko Kurahashi (*Shinyaku Hoshi no ohjisama, 2005*) and analyzed the difficulty of the texts.

These 11 works were assessed for grade determinations with the formula of Shibasaki and Hara (2010), and the results from Kotoba Fushigibako were simultaneously compared. Although there is no correct outcome for this judgment, we can likely confirm whether relative differentiation was accomplished. The results are shown in Table 2.

First, in the category of the bestsellers, the formula by Shibasaki and Hara (2010) rated *Donkanryoku* as the easiest, *Tanin o mikudasu wakamono* as the most difficult, and *Saodakeya wa naze tuburenainoka?, Homuresu Chugakusei, Kokka no hinkaku*, and *Josei no hinkaku* as the medians. Kotoba Fushigibako scored *Donkanryoku* as the easiest, and *Tanin o mikudasu wakamono* and *Kokka no hinkaku* as the most difficult, with *Kokka no hinkaku* presenting largely dissimilar results. Whereas the formula of Shibasaki and Hara (2010) rated *Kokka no hinkaku* at the eighth-grade level, Kotoba Fushigibako judged this text to be at the tenth-grade level. This result is attributable to the fact that kanji characters other than those for common use, such as 唖, 股, 綻, 訊, 蔓, 汎, 眉, 綻, 瞞, 傲, 靡, were used in the books, which presumably led to the higher rating results from Kotoba Fushigibako. For the two 'Keitai novels', both the formulas of Shibasaki and Hara (2010) and Kotoba Fushigibako determined that *Kurianesu* was easier than *Shiroi jaji sensei to watashi.* The translated versions of *Le Petit Prince* revealed similar results in that both methods measured degrees of difficulty, with Naito's translation rated as the most difficult, followed by the translations of Mita and Kurahashi. Excerpts from these three translations are appended at the end of this paper for reference.

The results of both the analysis of the probability of the characters' occurrence and a polynomial with the variables of the average number of characters, the predicates and clauses in one sentence, and the proportion of hiragana in the text show similar tendencies, and it is safe to conclude that there is no significant difference between the two methods. With results such as these, in which one formula rates *Donkanryoku* as the most difficult, it is difficult to prove the formula's validity.

Table2. Readability of 11 Novels by Two Formulas

Title	Author/Translater	Shibasaki/Hara Formura	Kotoba Fushigibako
Homuresu Chugakusei	Kirin・Hiroshi Tamura	8	8
Saodakeya wa naze tuburenaino ka?	Masaya Yamada	7	9
Kokka no hinkaku	Masahiko Fujiwara	8	10
Josei no hinkaku	Mariko Bandoh	8	8
Tanin o mikudasu wakamono	Toshihiko Hayami	9	10
Donkanryoku	Junichi Watabane	6	8
Kurianesu	Towa	8	6
Shiroi jaji sensei to watashi	reY	9	8
Hoshi no ohjisama	Aroh Naito	2	2
Hoshi no ohjisama	Masahiro Mita	4	3
Shinyaku Hoshi no ohjisama	Yumiko Kurahashi	6	5

3. The future of readability research

Since we opened the Readability Research Laboratory to the public in 2009, we have received responses from many people. Many users have commented that the model has helped them to determine the grade levels of texts for creating language exams. This tool is useful for determining whether any kanji exceed the ranges of specific grade levels because the model outputs not only readability results but also colored kanji characters differentiated by grade allocation within texts. For example, although the Japanese Federation of the Bar Association has been attempting to revise legal terminology since the jury system began in May 2009 to help laypersons understand certain terms, the highest level of difficulty of such a text should be at the third year of junior high school, which is the final year of compulsory education. Given the above results, we can conclude that the purpose of determining a text's scale of difficulty is to provide an index for the text's creation rather than the initial purpose of assigning reading materials appropriate for grade levels. Because we expect that many measurement tools will be developed in the future, we believe the essential element of readability formulas is the target readers. Therefore, the scales and variables should be determined after the selection of the base corpus data, depending on what types of readers are targeted, whether the readers are young students of various development stages or adults, whether they are native speakers of Japanese or second/foreign language learners, and whether those learners are from regions where kanji characters are used.

Reference

Sato, S., Matsuyoshi, S., and Kondoh, Y. (2008) Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. LREC-08.

Shibasaki, H. (2010) Difference at Cognitive Processing Speed of Sentences by Character-types: Basic Study for Development of the Readability Scale of Japanese Texts, *Research in Experimental Phonetics and Linguistics,* 18-31.

Shibasaki, H. and Hara, S. (2010) The Readability Formula to Predict School Grades 1-12 based on Japanese Language School Textbooks, *Mathematical Linguistics,* 215-232.

Shibasaki, H. and Tamaoka, K (2010) Constructing a Formula to Predict School Grades 1-9 based on Japanese Language School Textbooks, *Japan Journal of Educational technology,* 449-458.
建石由佳，小野芳彦，山田尚勇 (1988) 日本文の読みやすさの評価式,『文書処理とニューマンインターフェース 』18(1), 1-8.

レポート

制限言語と機械翻訳

石川ドキュメンテーション株式会社
石川　諭

初めに- STE との関わり

石川ドキュメンテーション株式会社では、制限言語のひとつであるシンプリファイド・テクニカル・イングリッシュ（STE）による英文のリライトサービスや STE ライティングのトレーニングを通じてお客様が STE を導入するお手伝いをしております。

個人的にはこの約7年にわたる STE との関わりの中で、約 10 名のネイティブイングリッシュスピーカーを含む 100 名以上の受講者の方々を対象に、2 - 3 days のトレーニングコースを通じて、リライトのコツやテクニックを解説してまいりました。1day セミナーや 3 日間で 1 セットのセミナーを含めると、さらに多くの方に Basic な STE のルールを解説してきました。

1. STE から Basic STE へ

STE の仕様書（ASD-STE100）には、約 60 の英文作成ルールと用語集（約 2,000 の禁止語と約 900 の言い換えのための承認語からなります）が収録されています。ルールはいいとして、この用語法を覚えるのが大変ですから、用語法違反などを検出するツールを使用します。

しかし、しばらくこのツールを使っているとこの用語法もだいたい頭の中に入ってしまい、「この用語は使わない方がいいな」などという感が働くようになります。そうなれば、ツールなしでも STE の仕様に沿った英文作成が可能になります。そもそもあまり厳格にルールを適用する必要もなく（法律ではありませんから）、70%～80%でも仕様書のルールや用語法に沿っていれば、分かりやすい英語としては格段の進歩をします。

そこで、STE に準拠して英文ライティングをするといっても、ASD-STE100 の厳格な適用ではなく、状況に応じて時にはルールや用語法に違反しても STE の重要な目的を達成している範囲で英文を書くことが、実践的な取り組みとされています。そこで、この指標を簡単にまとめたのが、以下に紹介する Basic STE です。

2. Basic simplified technical English (Basic STE)

以下をできるだけ使用します:

- 明確で曖昧さのない表現
- 短く歯切れのいい表現、文、段落
- 説明的な文、手順を表す文、警告を表す文それぞれにふさわしい表現
  - 説明文は 25 ワード以内で書く
  - 手順文は 20 ワード以内で書き、命令法を用いる
  - 警告文は他と切り離し、命令法を用いる
- 一文一トピック
- やさしい構文
- やさしい文法
- やさしい語句
- 能動態
- 箇条書き
- 承認語（STE の承認語および用語集などを含む）
- STE の承認語については、厳格な適用をしなくともよい

以下をできるだけ使用しません:

- 現在完了形
- “can be used”などの助動詞と過去分詞による複合動詞
- 動詞の-ing形による分詞構文や動名詞構文
- 禁止語(STEの非承認語および用例集を含む)

STEの非承認語については、厳格な適用をしなくともよい

3. 日本語の「構造」が大切

さて、英文のライティングと言っても、日本では、多くの場合、日本語から英語に翻訳して英文ドキュメントを作成しています。その場合、日本語の「構造」がきちんとしていれば、それだけ、翻訳した英文も分かりやすくなります。要は「構造」です。表現や用語は翻訳のプロセスでなんとかなりますが、「構造」は翻訳では変えられないのです。

そこで、Basic STEに連動して、Basic STJとしてグローバル時代の最低限のライティング指標をまとめたのが以下です。

4. Basic simplified technical Japanese (Basic STJ)

以下をできるだけ使用します:

- 明確で曖昧さのない表現
- 短く歯切れのいい表現、文、段落
- 説明的な文、手順を表す文、警告を表す文それぞれにふさわしい表現
  - 説明文は50字以内で書く
  - 手順文は40字以内で書く
  - 警告文は他と切り離す
- 一文一トピック
- 単純な構文
- 簡単な文法
- 論理的な説明
- 分かり易い表現
- 能動態(可能な限り)
- 箇条書き
- 承認語(用語集など)

以下をできるだけ使用しません:

- 文の一部の省略
- 複雑な修飾関係
- 禁止語(用例集など)

5. ルールにあわせて書くテクニック

ここまで、ルールについて簡単にご紹介しました。しかし、ルールはルールです。実際にBasic STJのルールにあうように、文を書くにはテクニックが必要です。たとえば、説明文は50文字以内で書くというルールですが、ではどうしたらいいのでしょうか。文を短くするためのテクニックはたくさんありますが、ここでは、条件節を含んだ文を短くするテクニックのひとつをご紹介します。

たとえば、

「AのBがCで、かつCがDの場合、EがFになると、GはHになる。」

というような文はよく見かけます。この文には4つの節があり、A～Hに技術用語が入ると複雑で長く、非常に分かりにくい文になります。分かりやすい文にするため、これらの節を単文化して、短いセンテンスに分解するには、ひとつのテクニックを使います。すなわち、次のようにいたします。

以下の条件により、GはHになる。

- AのBがCである
- CがDである
- EがFになる

では、実際の応用例をご覧いただきましょう。

6. 実際の応用例

(例文)

このとき、もし昇降油圧シリンダのロッドの動きが、何らかの外因等でロックされた場合、油圧ポンプのギヤポンプ部から昇降油圧シリンダのシリンダロッド側の間の作動油圧力が6.9～7.4 MPa以上となると、作動油がリリーフバルブポペットを押し開き、作動油は昇降制御バルブアッシのT2 ポートから出て、油圧ポンプのT ポートへ入って、トランスミッションケースへ戻る。

これは、1 文ですが 6 つの節からなり、トータルで 175 文字もあります。STJ の指標の 50 文字どころではありません。これをそのまま英文にすると、以下のような翻訳例が考えられます。

(翻訳例)

If the movement of the hydraulic cylinder rod is locked due to some external causes and the hydraulic oil pressure between the gear pump of the hydraulic pump and the cylinder rod side of the raise/lower hydraulic cylinder exceeds 6.9 to 7.4 MPa, the hydraulic oil pushes open the relief valve poppet. Here, the hydraulic oil leaves through the T2 port of the raise/lower control valve assembly, enters the T port of the hydraulic pump and returns to the transmission case.

訳文では 2 つのセンテンスに分けていますが、それでも最初のセンテンスは 52 単語もあり、Basic STE が指示する 25 単語の倍以上です。これがベースになり、さらに多言語化されるのではたまりません。そこで少なくともこの段階で STE にリライトする必要があります。しかし、これは 2 度手間です。翻訳する前の元の日本語の「構造」がシンプルであれば、このような複雑な英文にならなかったはずです。

それでは、最初から日本語を STJ で書きます。先ほどの応用です。

(STJ で書いた日本語)

このとき、以下の条件により、作動油がリリーフバルブポペットを押し開く。

- 昇降油圧シリンダのロッドの動きが、何らかの外因でロックされる
- 油圧ポンプのギヤポンプと、昇降油圧シリンダのシリンダロッドの間の作動油力が 6.9～7.4 MPa 以上になる

これにより、作動油は昇降制御バルブアセンブリの T2 ポートから出て、油圧ポンプの T ポートへ入って、トランスミッションケースへ戻る。

2 番目の条件部の文字数は、53 文字で、STJ の指標の 50 文字をわずかに超えていますが、この程度のルール違反を気にしないのが Basic たる所以です。ずっと読みやすくなり、英文に翻訳しても同じ構造ですから、多言語に翻訳する前のリライトは最小限ですみます。

7. Basic STJ と MT

STE や STJ により書いた文は、人間の翻訳者にも読み手にも分かりやすいばかりでなく、MT との親和性も高まります。複雑な修飾関係が是正され、シンプルな「構造」になるからです。

上の和文 (STJ で書いた日本語) を、Web 上で提供されている翻訳@nifty で翻訳した結果を下載します。

(MT による翻訳の結果)

At this time, the hydraulic fluid pushes the relief valve poppet open according to the following conditions.

- The movement of the rod of the going up and down oil pressure cylinder is locked by some external causes.

- The hydraulic fluid power between cylinder rods of the gear pump and the going up and down oil pressure cylinder of the oil-hydraulic pump is 6.9-7.4 It becomes MPa or more.

As a result, the hydraulic fluid goes out of the T2 port of the going up and down control valve assembly, enters T port of the oil-hydraulic pump, and returns to the gearbox casing.

いかがでしょうか。2番目の条件部の英語が少し変ですが、ここをリライトして、技術用語を置き換え、全体としてはいくつか表現を修正すれば、最初に掲げた（翻訳例）の英文よりずっと読みやすい英語になります。なぜMTがここまで理解でき、英語として利用できるかというと、原文の日本語の構造がSTJ（およびSTE）に沿ったものだからです。

石川ドキュメンテーションでは、このSTJよりさらに踏み込んだ和文のエディティング方法 - EJ（Englic Japanese: 英語のような日本語）法 - を用いて従来の人間翻訳のスピードを2〜3倍に高めたサービスを提供しています。STJとこのEJ法の違いなど、詳細については機会がありましたら、別稿でご紹介したいと思います。

8. まとめ

本稿では、はじめにBasicなSTEのルールをご紹介しました。そして、日本語から英語を作成することの多い現状では、日本語の「構造」が大切であること、そしてSTJのルールをご紹介しました。しかし、ルールを提示されただけではどのようにしてルールにあわせたらよいか分かりません。そこで、ルール通りに書くにはテクニックが必要なことを説明いたしました。本稿では、そのテクニックのひとつ、条件を箇条書きにするテクニックを例としてご紹介いたしました（他にもテクニックはたくさんあります）。最後に、このようなテクニックを用いて書いたSTJの文では、MTが十分使えるということをご覧いただきました。

レポート

クラウドソーシング型翻訳サービスの機械翻訳に与える影響

株式会社 anydooR
山田　尚貴

近年、インターネットの普及により、従来の翻訳会社が提供してきた人力での翻訳の形が変わって来ている。今までは翻訳会社が翻訳者を募集し、実際に試験を課し、試験に合格した翻訳者をコーディネータが管理し翻訳をクライアントの案件毎に提供するというスタイルが主流であった。しかし、インターネットが人々の生活の中に浸透するにつれ、海外との接点が増え、言語を跨いだコミュニケーションや他言語ソースの情報が増え続けるとともに、翻訳のニーズが多様化することとなった。

個人レベルの翻訳のニーズが増えるにつれ、従来の翻訳会社の提供する高価格高品質の翻訳から、個人でも手に入れる事ができる中価格、中品質の翻訳へのニーズが 2008 年頃から高まってきた。それとともに、クラウドソーシング（crowd sourcing）によて翻訳をインターネット上の誰かに対してアウトソースするサービスがいくつか誕生した。

元々クラウドソースは技術的な物が多く、例えばエンジニアリング等をインターネット上で賃金の安い途上国等にアウトソースするサービスなどが多く見られた。

日本でのクラウドソース翻訳の先駆けとして、弊社が提供するソーシャル翻訳 Conyac (http://www.conyac.cc) と myGengo (http://www.mygengo.com)がある。基本的には従来の翻訳会社が行っていた、「翻訳者に対する試験」「クライアントに対する見積り」「翻訳者のコーディネート」等、人的コストがかかっていた作業等をシステム化し自動化する事により、翻訳コストを下げることで、今までの翻訳会社が提供出来なかったプライスで翻訳を提供できるようになった。

一方でコーディネーターの入らないクラウドソーシング型の翻訳サービスにもデメリットはあり、その中でも品質の確保があげられる。

Conyac の場合は品質の確保の為に、評価システムを取り入れている。サービス提供側で翻訳を提供する代わりに、翻訳者間で結果の良否を判別する仕組みを取り入れている。この仕組みにより、サービス提供側が翻訳の善し悪しを判断出来ない言語に関しても確実な評価ができる為、多言語の翻訳の提供が可能となった。

クラウドソーシング翻訳の一番のメリットは居住地、国籍に関わらずインターネット環境がある場所ではどこにいても翻訳者となれ、自身のスキルを活かしたタスク（仕事）をこなす事ができ、そのタスクに見合った報酬を手に入れる事ができる。そのため、空き時間を利用して翻訳者としての仕事を行う事が可能となる。

弊社 Conyac での翻訳結果は、翻訳を依頼した依頼者と弊社に帰属する。そのため、翻訳依頼とその依頼に対してついた翻訳結果は対訳としてデータベースの中に貯まっていく。

長文から短文までの対訳が数十言語で貯められていき、年間 1,000 万以上の対訳が貯まっていく。ユーザー数と依頼数が年々増加していく為、10 年で数億から数十億の対訳が貯まる事になり、その対訳データベースを使用し、機械翻訳を生成する。統計型や検索型の機械翻訳とは違い、対訳データベースから生成された機械翻訳である為、高品質な機械翻訳を可能とする。

Conyac は機械翻訳を作る為のサービス運営をしている訳ではなく、個人レベルで使用出来る翻訳プラットフォームを目指しているので、最終的には Google 等、機械翻訳を提供しているサービスプロバイダ等にデータの提供を行う事を検討している。

レポート

機械翻訳の実用とポストエディット

斎藤　玲子

はじめに

翻訳を発注する立場から見た、機械翻訳導入のメリットは何でしょうか？コスト削減、納期短縮、人手不足の解消 — まるで魔法のように、すべてを解決してくれる素晴らしい存在であって欲しいと、期待は膨らみます。しかし、現実問題として、そこまでの道のりは平坦ではなさそうです。困ったときに打つ手は 3 つあります。

1. 環境が整うのを待って、耐える。
2. 今あるもので何とかする。
3. 新しい手段を探して投資する。

ここでは、2 の「今あるもので何とかする」方法で、現段階の機械翻訳を実際のビジネスにどうやって生かしていくことができるか、機械翻訳を導入する側の視点と作業する側の視点から考察してみたいと思います。

機械翻訳の 4 つのオプション

機械翻訳を使う場合、まず機械翻訳の結果をそのまま使う、機械翻訳した結果をポストエディットする、という 2 つの選択肢があります。後者の場合、ポストエディットを社内の人材を使って行うのか、社外に発注するのかで道は分かれます、社外に発注する場合は、さらにライトエディットをするのか、フルエディットをするのかに分かれます。つまり、以下のようなツリー構造になります。

「フルエディット」と「ライトエディット」という言葉について説明します。「フルエディット」は「正確で読みやすい内容に編集する」ことで、「ライトエディット」は、「正確な内容に編集する」ことを指します。どのように違うのか、例を挙げてみます。

英語: View the different location groups to which a location is associated.

MT: ロケーションが異なる場所に関連するグループを表示

ライト: ロケーションが関連する、異なるロケーショングループを表示

フル: ロケーションが関連付けられている、さまざまなロケーショングループを表示

ライトエディットでは、語順を入れ替えて読点を追加しただけです。フルエディットでは、それに追加して associated と different の訳を変更しています。フルエディットでは、ゼロから翻訳する場合と結果はほとんど変わりません。わかりやすい反面、時間の節約についてはあまり期待できません。

社内でエディットする場合は、基本的に「ライトエディット」になるだろうと想定しています。これは、作業にあたるのが選任の翻訳者ではなく、翻訳される内容の分野で実務に携わっている人であることが多いためです。

上記をふまえて、4 つのオプションをスピード、品質、コストの 3 点から比較し、それぞれに適した用途を考察してみます。

1. MT のみの場合

	1	2	3	4
	MT	社内ライト	社外ライト	社外フル
スピード	□	□	○	□
品質	□	○	○	□
コスト	□	□	○	□

1 つめの MT のみでは、スピードとコストはベストです。一方、まだ品質については難しい面があります。あるテストで Moses というオープンソースの統計ベースの機械翻訳エンジンの品質評価をしてみると、現在は「そのままでは、半分ぐらいは理解できる」レベルという結果になりました。このオプションに適しているのは「オンラインで大量で無償で提供するもの」で、提供方法に工夫が必要です。「これは機械翻訳です」などと、ユーザーに事前に知らせ、元の英語が参照できるようにリンクを設けておくなどして、ユーザーが「英語でしか情報がないよりは便利」と思ってもらえることを目指します。

2. MT + 社内でエディットする場合

	1	2	3	4
	MT	社内ライト	社外ライト	社外フル
スピード	□	□	○	□
品質	□	○	○	□
コスト	□	□	○	□

2 つめの MT + 社内でライトエディットの場合、品質は「正確な翻訳」レベルになり、コストも外注費がかからないという点では◎です。一方、スピードについては、社内で選任でない人が作業にあたるため、そのトレーニングやスケジュールを考えると、メリットがあるとは言えません。このオプションに適しているのは、1 件あたり 10,000 words 前後と量が少なめで、内容は作業者が専門とするものです。テクニカルノートなどの技術情報が一例です。

3. MT + 社外でライトエディットする場合

	1	2	3	4
	MT	社内ライト	社外ライト	社外フル
スピード	□	□	○	□
品質	□	○	○	□
コスト	□	□	○	□

3 つめの MT + 社外でライトエディットの場合、すべてについて合格点であると言えます。外注費はかかりますが、通常の翻訳費用よりは低く抑えられ、品質は「正確で理解できる内容」になります。作業量が通常翻訳より少ない分、スピードも上がります。このオプションに適しているのは、「定型的な文章の多い、大量でオンラインのドキュメント類」です。技術分野でいえば、管理ガイドやリファレンスマニュアルなどが一例です。

4. MT + 社外でフルエディットする場合

	1	2	3	4
	MT	社内ライト	社外ライト	社外フル
スピード	□	□	○	□
品質	□	○	○	□
コスト	□	□	○	□

4 つめの MT + 社外でフルエデットの場合、品質はゼロからの人による翻訳とほぼ変わりません。スピードとコストも、人による翻訳と同じぐらいかかることが多いものです。この 2 点を改善するには、分野に特化した MT エンジンへの教育、編集環境の効率化、および編集方法の効率化が必要です。それらが改善されれば、オンラインだけでなく、紙で出版されるドキュメントや、わかりやすさを求められるものにも、正確でわかりやすい内容の翻訳を割安なコストでより早く提供できることになります。

編集方法の効率化

上記で挙げた 3 点の改善方法うち、「今すぐ何とかできるもの」は何でしょうか。エンジンの教育と編集環境には特殊なツールや環境および時間が必要です。そこで、ガイドラインなどのソフト面で対応できる「編集方法の効率化」に着目してみます。

ポストエディットは、「翻訳し直し」ではありません。かといって、「どんなに悪い翻訳でも、修正して使わなければならない」わけでもありません。では、どのようなガイドラインがあればいいのでしょうか。

ポストエディットのガイドライン

ガイドラインの有無で作業の効率は大きく変わります。ガイドライン作成の前後で効率が 2 倍になったケースもあります。

ガイドラインには「目標」と「手法」の 2 点が必要です。つまり、「どのような品質にするのか」と「どのようにして編集するのか」です。品質については、「正確でわかりやすい翻訳」なのか「正確な翻訳」なのかを特定します。もちろんこれは、作業者だけが目指しても仕方ないので、発注する側との合意が事前に必要になります。具体例を挙げて、作業する前に相互で確認しておきます。正確な翻訳を目指すのであれば「スタイルの不統一や読みやすさは求めない」ことを明確にするべきです。

「手法」については、どうすればスピードを上げることができるかに着目し、具体的に方法を提案することが重要です。私が気をつけている点は以下です。

心構え

1. 「MT は正しい翻訳を出力するべき」という考えから離れる。
2. 「使えるものは使おう」という精神でのぞむ。
3. それでも「自分で翻訳したほうがはやい」と思えば、そうする。

「心構え」をするのは、MT に対して中立的な立場を保つためです。MT に対してネガティブな気持ちのある人ほど、効率は下がる傾向があります。

実作業

1. 日本語を読み、英語を読む
2. 短い時間で「理解できる内容になるか」チェックする。
3. Yes なら編集、No なら再利用またはゼロから翻訳する。

実作業の 1 と 2 をすばやく行うことで、ゼロから翻訳するだけよりも、少しでも効率を上げられるようになります。3 をより具体化すると次のようになります。

1. MT を修正して使う
2. MT を参考にする
3. MT を使わない

2 の「参考にする」とは、自分で翻訳するが、その中に MT にある表現を使うということです。さらに、1、2、3 の使い分けの目安も示すとよいでしょう。たとえば、以下のようにします。

作業ガイドラインの例

1 以下の場合は MT を修正して使う
- 1.1 一度読むだけで 80% 以上が理解可能
- 1.2 修正箇所を決めるために日英を比較する回数は 1 回
- 1.3 必要な編集作業は 2 回まで

2 以下の場合は MT を参考にする
- 2.1 一度読むだけで半分強が理解可能
- 2.2 参考にする箇所を決めるために日英を比較する回数は 1 回

3 以下の場合は MT を使わない
- 3.1 一度読むだけで半分以上が理解不能
- 3.2 参考にする箇所を決めるために日英を比較する回数が 2 回以上

まとめ

機械翻訳を発注する側としては、MT の使い方(オプション) とそれぞれに適した翻訳対象を見極め、期待する品質を決めて、作業者に伝えることが重要です。最初は「正確な翻訳」で良いと思っていても、出来上がったものを見ると「正確でわかりやすい翻訳」に修正したくなることがあります。しかしそれでは、低コストと短納期で通常翻訳を求めていることになります。このようなことが起こらない

ように、作業を開始する前に、発注側と作業側で目標とする品質について丁寧に共有し、合意しておくことが大切です。作業側も、目指す「品質」の具体例を挙げ、積極的に発注側に提示するべきです。そしてその内容を「作業方法」とともに、ガイドラインとして作業者間で共有することが、ポストエディットの品質の一定化と作業の効率化につながると考えます。

Report

From life-long enemy to new market – a translation service provider view of MT

Don Shin

1-Stop Translation USA, LLC

don@1stoptr.com

I still remember the day clearly—it was in July of 2009. I was standing in front of 150 translation service provider owners. My presentation was about the future of translation industry. Knowing that machine translation had been a life-long enemy of our industry, I tried to be as reserved and as subtle as possible. I simply expressed that MT was not the ultimate evil or 'he who must not be named.' It might increase our efficiency and someday we may even use MT just like we are using translation memory every day. However, the response was very cold. Apart from for a couple of companies, the rest were saying that they were neither using MT nor were even thinking of using it.

But the changes have been dramatic. Within just one year, I received several calls from other translation company owners who received inquiries regarding their MT capacity and experience.

The real ice breaker was the IMTT's conference in Denver in October 2010. As far as I know, it was the first time that ATA (American Translators Association) was officially invited to the conference. For the convenience of ATA members, it was held in the same city right after the closing day of the ATA conference. I participated in the conference along with my friend and client, Jiri Stejskal, who was the president of ATA at that time.

It was a real eye opener. These are a few points I learned in that conference: 1) There would not be just one MT engine that could translate better than the rest in all languages and all subjects. Instead, there would be thousands of MT engines. Someday, each company may possibly even have their own MT engine per language pair. 2) Instead of simply becoming one of the tools to be used in a translation process, it would generate many different jobs, businesses and even create new industries; just like the invention of a car had led to the creation of a giant car industry. 3) Further developments of MT would not only be the jobs of engineers or scientists. It would heavily depend on linguists or translation companies like us.

Since then, in almost every conference MT has been one of the most common and popular topics. More and more translation companies are proudly saying 'we handle MT.' At the Globalization and Localization Association conference held in Lisbon, Portugal in March 2011, almost one third of the companies were doing or had done MT related projects. For our company in particular, although we're only working with Asian languages, which is still a lot harder than European languages for machines to translate, almost every month, we perform MT editing jobs.

The majority of the jobs that we're currently handling are post editing of MT output. Some companies provide MT engine training services or cleaning corpus; other companies started to use MT to save costs and improve speed. One SLV (Single Language Vendor) owner with 30 full time translators told me that he was expecting 20-30 percent savings once he finished his custom MT engine training. The most notable sign to me was the birth of a new service company like Pangeanic (http://www.pangeanic.com/). They provide training of MT engine service for clients. I do believe that this is a very promising new 'blue ocean' in our industry.

Very soon we will see more and more new services related to MT. As a translation company specializing in Asian languages, we're planning to develop an MT engine for Simplified and Traditional Chinese. In our industry, currently most materials that need to be translated into Traditional Chinese are already available in

Simplified Chinese. Then why not translate from Simplified Chinese directly instead of English? In the coming years, more and more MT engines will be developed and sold by translation companies rather than software companies. Also, within the translation companies ourselves, we will be able to see new job titles such as MT editor, monolingual translator, MT trainer, Corpus miner and translation memory cleaning or specialist.

I do not know exactly where the industry will go, but one thing for sure is that the changes will come even faster compared to the last 3-4 years.

For more information on machine translation and the latest innovations in the industry, go to the Globalization and Localization Association website, www.gala-global.org. GALA is the largest localization industry association worldwide and is a great resource with articles, explanations and webinars on machine translation.

シンポジウム参加報告

第 11 回日中自然言語処理共同研究促進会議に参加して

北陸先端科学技術大学院大学
白井　清昭

2011 年 10 月 29 日(土)・30 日(日)の両日に宮崎市のホテルプラザ宮崎において第 11 回日中自然言語処理共同研究促進会議(The 11th China-Japan Natural Language Processing Joint Promotion Conference)が開催された．日本からは 14 名，中国から 10 名，韓国から 1 名の研究者が参加した．

会議は長尾真先生(国立国会図書館)の基調講演で始まった．講演では，2006 年度から 2010 年度にかけて実施された科学技術振興調整費プロジェクト「日中・中日言語処理技術の開発研究」による日本語・中国語を対象とした機械翻訳研究の成果について触れ，現状の日中機械翻訳システムを本格的に実用化可能なレベルまで精緻化するためには日本と中国の研究機関の協力体制が必要不可欠であると述べた．その後，日中両国からの参加者による研究発表が 2 日間に渡って行われ，活発な議論が展開された．以下，中国および韓国からの参加者の発表タイトルと内容について簡単に紹介する．

WANG Huilin 氏 (Institute of Scientific and Technical Information of China)

Chinese-Japanese Meta Grammar and Parallel Treebank

Meta grammar とは，言語，さらには様々な文法枠組(文脈自由文法，LTAG，GPSG)などに共通して利用可能な文法的知識の総称である．機械翻訳における中間言語の文法版ともいうべきもので，チャレンジングな研究であると感じた．

ZONG Chengqing 氏 (Institute of Automation, Chinese Academy of Sciences)

Parse Reranking Based on Higher-Order Lexical Dependencies

中国語の単語分割の新しい手法として，generative model と discriminative model によるスコアを組み合わせる手法を紹介した．さらに，中国語の構文解析の曖昧性解消のためのリランキングモデルについて講演した．このモデルでは、まず構文解析を行い，そのスコアの上位 n 件の構文木を取得する．次に，同じ文の依存構造を別途解析し，得られた依存構造と矛盾する構文木のスコアを低くすることで構文木のリランキングを行う．木構造と依存構造という 2 つの異なる構文解析の結果を利用することで解析精度を向上させている．

ZHU Jingbo 氏 (Northeastern University)

Some Issues of Syntax-based Statistical Machine Translation

タイトルにある Syntax-based SMT とは部分木を単位とした統計的機械翻訳である．フレーズや木を単位とした統計的機械翻訳は近年盛んに研究が行われているが，同氏は自身の研究も含めたその最先端の研究について講演された．

荒瀬由紀 氏 (Microsoft Research Asia)

An Overview of J-E machine translation research at Mircosoft

まず，日英の対訳文対および対訳単語対をウェブから収集する 2 つの手法を紹介した．1 つは対訳関係にあるウェブページを自動収集し，DOM のアライメントや文のアライメントを同定することで文の対訳対を得る手法，もう 1 つは同一ページにある日

本語と英語の対訳文の組や訳語の組を自動的に検出する手法である．実験データとして検索エンジンBingのクローリングデータを使用するなど，大規模な実験によって提案手法の有効性を確認していた．次に Syntax-oriented Pre-Reordering Model と呼ばれる日英統計的機械翻訳の手法を紹介した．日本語と英語の構文木において，木構造内のノード(非終端記号)の出現順序が入れ換わる確率を学習し，これを統計的機械翻訳に組み込む．最後に Engkoo (http://research.microsoft.com/en-us/projects/engkoo/) と呼ばれるオンラインの英語学習システムのデモを紹介した．

MENG Yao 氏 (FUJITSU Research and Development Center)

Research for the named entities resolution

固有名詞とその属性(例えば人名とその住所や会社など)をウェブから自動的に収集する手法を紹介した．提案手法は 3 つのステップで構成される．standardization では，大規模な人名辞書を使い，異表記の人名をまとめる前処理を行う．disambiguation では，クラスタリングによって固有名詞の曖昧性解消を行う．人名，組織名，地名を対象にした実験での正解率は 0.75ー0.8 程度であった．integration では，抽出した固有名を含む関係をウェブから抽出し，統合する．ここでの「関係」とは住所や会社など固有名に関する属性を指す．

ZHOU Ming 氏 (Microsoft Research Asia)

Tweet sentiment classification

Twitter などのマイクロブログにおける書き込みを対象とした感情判定について講演した．具体的には，あるトピックを与え，そのトピックに関する tweet の感情を positive, negative, neutral のいずれかに分類する．機械学習の手法を適用し，感情を表わす語やアイコン，ハッシュタグの他に，tweet とトピックとの関係を考慮し，トピックと統語的関係にある語も学習素性として利用する点に特徴がある．さらに，ツイートをノード，リプライやリツイートなどをリンクとしたグラフを作成し，グラフ全体で感情の分類を最適化する手法も紹介した．

SUN Maosong 氏 (Tsinghua University)

Social Tag Computation and its Application in Microblog

ソーシャルタグに関する研究について講演された．「ソーシャルタグ推薦」では，あらかじめ決められたタグのセットの中から，記事にふさわしいタグを自動的に選択する．タグを選択する際には，記事中の単語とタグとの関連度を計算したり，LDA(Latent Dirichlet Allocation)によって推定された潜在的トピックを利用する．一方，「ソーシャルタグ補完」では，あらかじめタグを定義せず，記事に適した新しいタグを推薦する．具体的には，記事の中から重要なキーワードをソーシャルタグとして推薦する．Weibo (中国におけるマイクロブログサービス)における書き込みに対して上記 2 つの手法を適用した結果についても紹介した．

ZHAO Hai (Shanghai Jiao Tong University)

Detection of Deceptive or Valuable Information from Internet Reviews

ウェブ上のレビュー記事を worthless, valuable, deceptive(pseudo) の 3 つに分類する研究について講演した．deceptive は，広告など，評価自体は当てにならないがレビュー対象に関する何らかの客観的な事実が書いてある記事を指す．分類器として決定木を用い，学習素性としてレビュー長，感情語の有無，単語とその TF，レビューがリンクを含むか，feature term dependency tree などを利用する．feature term dependency tree とは，A compared to B など，レビュー記事の有用性の判定に有効と考えられる関係の集合である．

LIU Qun 氏 (Institute of Computing Technology, Chinese Academy of Sciences)

Multilingual Machine Translation Research in CAS-ICT

中国語と，中国におけるマイナーな言語の間の機械翻訳について講演した．同氏は，中国語とモンゴル語，チベット語，ウイグル語，韓国語，ロシア語，タイ語，ベトナム語，日本語の機械翻訳のシステム開発を進めており，一部の言語については中国語とのパラレルコーパスも構築している．また，機械翻訳の手法として Multigrain MT と呼ばれる手法を提案している．これは，語，語幹，形態素を単位とした統計的機械翻訳の出力を組み合わせて翻訳精度を向上させる手法である．

ZHANG Yujie (Beijing Jiaotong University)

Research Activities in BJTU

言語処理技術をフルに活用した学習支援システムについて講演した．ここでの学習支援とは，研究に関する学習を支援することである．例えば，機械翻訳の研究を始めたい学生に対し，機械翻訳に関する重要なキーワード，著名な手法やシステム，過去の研究論文などを自動的に収集して提示する．この他に，日本語で書かれた満州鉄道の関連文書のドキュメントを収集してデジタルアーカイブを構築するプロジェクトを紹介しており，興味深かった．

Key-Sun Choi 氏 (Korea Advanced Institute of Science and Technology)

Ontology-based Feature Selection

同氏は韓国からのゲスト参加者である．講演では，電子メイルから会議の情報(会議名，日時，場所，参加者，開始時間，終了時間)を自動抽出する研究を紹介した．会議の情報の抽出は，従来の情報抽出と同様にパターンマッチにより行う．さらに，時間については，それが開始時間もしくは終了時間を表わすのかを機械学習により分類する．このとき，学習素性のオントロジーを構築し，オントロジーの構造やオントロジーにおける上位-下位関係の情報を利用し，AdaBoost によって素性選択を行う．

一方，日本側からも 10 件の研究発表が行われた．本稿ではタイトルと講演の概要のみ紹介する．

梶博行 氏 (静岡大学)

Term Translation Using Comparable Corpora and the Web

日本語の複合語の英訳をコンパラブルコーパスから獲得する手法．

小谷克則 氏 (関西外国語大学)

Compiling Learner Corpus Data of Speaking, Listening, Writing, and Reading

I-Learning Corpus(英語学習者コーパス)の構築．

安田圭志 氏 (情報通信研究機構)

Introduction to the Field Experiments of Speech-to-speech Translation Systems

観光案内ドメインの音声機械翻訳システムの実地検証．

深澤信之 氏 (JST)

Report from Beijing Symposium

中国機械翻訳技術協力シンポジウム(詳細は後述)の報告．

奥村明俊 氏 (日本電気)

Information Value Creation Platform -- Integrating Media, Text, and Sensor Technologies --

テキスト翻訳・音声翻訳，ビジネスミーティングサポート，自然言語をクエリとした画像検索，Computer Aided Engineering Tool など．

黒橋禎夫 氏 (京都大学)

Efficient Retrieval of Tree Translation Examples for Syntax-Based Machine Translation

用例に基づく機械翻訳における部分木の翻訳メモリの効率的な検索手法．

永田昌明 氏 (NTT)

Basic Research of Natural Language Processing at NTT

日本語の文節係り受け解析，日英統計的機械翻訳，機械翻訳システムの評価基準.

江原暉将 氏 (山梨英和大学)

Machine Translation System for Patent Documents Combining Rule-based Translation and Statistical Post-editing

複数の中間言語を用いる機械翻訳方式，ルールベース機械翻訳と統計ベース後編集の組み合わせ手法など.

井佐原均 氏 (豊橋技術科学大学)

Research Activities in Toyohashi University of Technology

豊橋技科大とマイクロソフトによる多言語機械翻訳システムの共同研究プロジェクトの紹介，機械翻訳，e-learning など.

白井 清昭 (北陸先端科学技術大学院大学)

Word Sense Induction based on Clustering with Multiple Feature Vectors

複数の素性に基づく語義推定の手法.

日中自然言語処理共同研究促進会議は過去 11 年間に渡って毎年開催され，日本と中国の研究者間の交流や両国における自然言語処理研究の情報交換に大いに貢献してきた．また，本会議のおよそ 1 ヶ月前の 2011 年 9 月 26・27 日に，北京にて「中国機械翻訳技術協力シンポジウム」が開催された．このシンポジウムは，前述の「日中・中日言語処理技術の開発研究」プロジェクトの終了を受けて，同プロジェクトの研究成果を中国の研究者に報告し，また日本および中国における機械翻訳技術の現状について情報交換を行い，両国による共同研究や技術協力の可能性を探ることを目的として開催された．今回の会議では，参加者の多くがこのシンポジウムにも参加していたことから，2 日目の朝に同シンポジウムの議論を踏まえたディスカッションの場が設けられた．中国側の参加者に中国語を対象とした自然言語処理技術の現状についての見解を伺うとともに，日中両国による共同研究プロジェクト実現の具体化について話し合いを進めた．今後，日中両国の連携を深める場として，また，今回韓国から Choi 氏が参加したようにアジア諸国との研究交流の輪を広げるための場として，本会議が果たす役割はますます大きくなるだろう．次回は 2012 年 7 月に中国のハルビンで開催される予定である.

委員会活動報告

設問に基づく日中機械翻訳テストセットの自動評価に向けた拡張

機械翻訳課題調査委員会ワーキンググループ 1

1. はじめに

アジア太平洋機械翻訳協会の機械翻訳課題調査委員会（以下、AAMT 調査委員会）では、今後益々重要性が増すことが予想される日中機械翻訳システムの翻訳品質を評価する手法の確立を目的として活動を推進してきた。具体的な活動内容は、日本語および中国語の文法的特徴（評価項目）を翻訳システムが正しく処理することができるかどうかをチェックするための例文と設問のセット（AAMT 機械翻訳文テストセット）の開発である。評価方法に Yes/No での回答が可能な設問形式を取り入れることにより、客観性が高く、かつ、人間による主観評価に近い評価を低コストで実現することを狙ったものである。

2009 年度に行った「AAMT 機械翻訳文テストセット第 1 版」の公開に引き続き、2010 年度は、設問による評価の信頼性と客観性をより高めるために、評価項目の追加と設問内容の見直しを行い「AAMT 機械翻訳文テストセット第 2 版」を作成した。2011 年度は、「AAMT 機械翻訳文テストセット第 2 版」をもとに、プログラムが自動的に設問に回答することが可能な、「設問自動評価用テストセット」を作成した。本稿では、自動評価用テストセットによる自動評価について、人間による設問評価との比較、従来の自動評価手法との比較という観点で調査を行った結果について報告する。そして、代表的な日中翻訳エンジンに対して設問に基づく自動評価を実施した結果に基づき、文法項目、翻訳方式等の観点からエラー分析を行った結果について述べる。

2. AAMT 機械翻訳文テストセットの拡張

「AAMT 機械翻訳文テストセット」は、日本語文（原文）、中国語文（参照訳）、文法カテゴリ、設問の組を基本レコードとし、第 1 版は 325 組、第 2 版は 576 組のレコードから成るテストセットである。AAMT 機械翻訳文テストセットの概要は、AAMT Journal No.49(2011)の解説を参照されたい。

本節では、テストセット第 2 版の設問をもとに作成した自動評価用設問の概要について述べる。ここで導入する自動評価用設問は、人が評価するための設問の近似であり、両者が完全に一致することは理論的に保証されていない。しかしながら、3 節で述べる実験により、両者が同一とみなせること、さらには自動評価用設問の方が評価の揺れがないために、人間による設問評価よりよい可能性があることを示唆する。

自動評価用設問作成の基本的な考え方は、設問を満たすために必要なキーワードや表現（指定語句）が訳文中に含まれるか（含まれないか）をチェックするものである。ただし、指定語句の同義語や置き換え可能な表現を考慮しておく必要がある。このため、表 1 のように、各設問を見直して同義語を網羅的に記述するようにした。

プログラムの例(pseudo code)を図 1 に示す。訳文に含まれていなければならない（含まれていてはいけない）文字列を正規表現で記述している。同義語や置き換えが可能なフレーズを OR 条件にすることで、自動評価を人間による評価に近付けている。

表 1　拡張された設問の例

Original Questions	Expanded Questions
「病気で」が「因为生病」になってますか？	「病気で」が、「因为生病」または「由于生病」または「因为毛病」または「由于毛病」または「因为病」または「由于病」または「因病」になっていますか？
「木で」が「用木头」になってますか？	「木で」が、「用木头」または「拿木头」になっていますか？

```
read a line;
if the line match /严|严厉|严格|厉害/ then
    print "Yes";
else
    print "No";
endif

read a line;
if the line match /去买东西|去购物|去逛街|去逛商店/ then
    print "Yes";
else
    print "No";
endif
```

図 1　自動評価のためのプログラムコード例

3. 評価実験の実施

本節では 2 節で述べた設問回答の自動化プログラムによる機械翻訳評価の有効性を確認するための実験について述べる。今回の実験では、中国語に特有の文法項目について追加した 251 の評価例文を利用した。

実験を通して、設問回答の自動化手法が従来の自動評価手法と比較してどの程度の信頼性を持っているかについて、伝統的な人間による主観評価手法との相関を調べることによって分析する。また、設問回答の自動化手法が、人間による設問評価とどの程度一致した結果を導けるかを設問回答の一致度を調べることによって分析する。

これらの分析に必要なデータを集めるために、以下の要領で実験を実施する。

① 人間による評価：

・評価者：2 名（日本語検定 1 級を持つ中国語ネイティブ）

・評価観点

－正確さ（Adequacy）：原文と訳文を見て 1～5 で評価

－流暢さ（Fluency）：訳文のみを見て 1～5 で評価

－設問：訳文と設問を見て Yes/No 評価

② プログラムによる自動評価

－従来の自動評価尺度：BLEU、NIST、WER、PER

－自動化された設問評価

二人の評価者は、日本語のスキルが高く、かつ、言語学的な素養のある中国語ネィティブ（中国人）である。評価対象とした翻訳システムは、Web 上でアクセスすることができる６つの代表的な日中翻訳システムを選択した。６つの翻訳システムを翻訳方式で分類すると、４つが RBMT、２つが SMT である。

4. 実験結果と考察

設問による機械翻訳の評価方法が従来の自動評価尺度と比較してどの程度の信頼性を持っているかを検証するために、人間による主観評価（正確さ、流暢さ）との相関係数(Pearson product-moment correlation coefficient)を求める。Adequacy、Fluency は、評価文ごとに評価者１と評価者２のつけたスコアを平均して、これを人間の評価値とした。同様に人間による設問評価についても、ふたりの評価者のつけたスコアの平均をとった。

６つの日中翻訳システムの従来自動評価（BLEU,NIST,WER,PER）によるスコア、設問評価（人間、設問）の平均値と、正確さ、流暢さの平均値の間の相関係数を表 2 に示す。全体的に相関係数が高い値になっているのは、通常のテストセットに比べて文が短いためと思われる。

表２　各評価尺度と従来型主観評価との相関係数

	正確さ	流暢さ
BLEU	0.9453	0.9588
NIST	0.9783	0.9123
1-WER	0.9801	0.9267
1-PER	0.9707	0.8986
Questions（by Human）	0.9793	0.9334
Questions(Automatic)	0.9902	0.9208

表２から、いずれの評価尺度、評価方法においても、正確さ、および、流暢さとの間で高い相関を示している。相関の有意性を判定するために、表２の相関係数を用いて t 検定を行った。その結果、いずれの組み合わせにおいても、相関は有意水準 1%で有意であった。従来自動評価と設問評価の相関係数を比較すると、正確さにおいては設問評価の数値が従来自動評価尺度の数値を上回っている。このことから、設問による評価手法は人間による主観評価の代替手法として有効であり、かつ、従来自動評価尺度に劣らない信頼性を持っていると言える。

表３、表４は、自動評価と人間評価の間でのスコアの組合せの件数を表している。表５は、二人の評価者の間でのスコアの組み合わせの件数を表している。設問評価では、スコアは Yes または No のどちらかであり、スコアの組み合わせが Yes-Yes または No-No の件数の合計が「評価一致」それ以外が「評価不一致」である。自動評価と二人の評価者との評価一致率は、それぞれ、78.4%,78.3%となった。二人の評価者の間の一致率は 83.9%であり、

表３　回答パターンの分布　（ 自動 －　人間評価者１）

Same	Yes-Yes	425	1181	78.4%
	No-No	756		
Different	Yes-No	222	325	21.6%
	No-Yes	103		

表４　回答パターンの分布（ 自動 －　人間評価者２）

Same	Yes-Yes	473	1179	78.3%
	No-No	706		
Different	Yes-No	174	327	21.7%
	No-Yes	153		

表５　回答パターンの分布　（ 人間評価者１　－　人間評価者２）

Same	Yes-Yes	456	1264	83.9%
	No-No	808		
Different	Yes-No	170	242	16.1%
	No-Yes	72		

自動評価との一致率を多少上回った。表６は、表３-５のデータを元に計算した Kappa 係数である。いずれの組み合わせにおいても、Kappa 係数は 0.5 以上の高い数値を示している。

表６　　自動評価と人手評価の間の Kappa 係数

自動評価:人手評価 (評価者 1)	0.5494
自動評価: 人手評価(評価者 2)	0.5552
人手評価(評価者１) : 人手評価(評価者 2)	0.6616

人間評価と自動評価が各文の評価が一致していることを検証するために、表３、表４の分布に対して、有意水準 1%で検定を行った。その結果、評価者１、評価者２のどちらにおいても自動評価のつけるスコアと一致しているという結果になった。このことより、各設問の評価についても自動評価は人間による評価と同等であることが検証された。

5. 日中翻訳システムへの適用による翻訳誤りの評価

設問に基づく評価法は、ドキュメントの訳文品質を数値化できるだけでなく、文法項目毎に翻訳システムの得手、不得手を知ることができることが特徴である。設問による自動評価の Yes 回答を１、No 回答を０として、評価対象の６つの日中翻訳システム(A~F)について項目別の平均値を求めた。各項目の６つのシステムの平均値と対比してグラフ化したものが図２である。

図２で文法項目ごとにシステムのスコアの平均値を見ると、文法項目のカバー率は、項目によってかなりばらつきがあることがわかる。たとえば、No.10（形容詞述語文）やNo.9(「有」文)などでは全体の平均値が0.7を超え、ほぼすべてのシステムで対応できている項目がある一方で、No.32(自発)やNo.31(可能)などは、いずれのシステムでも対応できていない。カバー率が低い項目には、日中の機械翻訳における共通の課題が存在する可能性がある。全体的には、39項目のうち６つのシステムの平均値が0.5を超えた項目は10しかなく、現状の日中翻訳システムは開発余地が大きいと考えられる。

システム間の比較では、最もスコアの高いシステムDと最もスコアの低いシステムEでは、性能的に大きな差があることがグラフから読み取れる。６つのシステムのうち下位の２つ（E,Fエンジン）は統計ベース翻訳であり、方式の違いがスコアに表れている可能性がある。E,Fの２つエンジンは、まったく対応できていない文法項目(0点)がそれぞれ15項目、12項目あり、そのうち11項目が互いに重複している。これらの中で、RBMTと対比して特にスコアが悪い項目であるNo.30（被を用いない受身文）、No.15（願望）、No.22（「過」（～たことがある））、No.37（「～ば～ほど」と「越～越～」構文）などは、学習データ中の出現頻度が小さいか、SMT特有の課題を包含している可能性がある。

表７　文法項目

	文法項目
1	「の」と「的」
2	「Vた」と「的」
3	「もの」と「的」
4	方位詞
5	疑問詞
6	否定文（否定副詞の位置）
7	「不」と「没」
8	「在」文
9	「有」文
10	形容詞述語文
11	兼語文
12	二重目的語をとる動詞
13	できる・可能（「会・能・可以」）
14	許可（可以）
15	願望
16	前置詞
17	比較文
18	「Vたばかり」と「刚刚」
19	「～ている」と「了」「着」
20	「了」（変化）

	文法項目
21	「～ている」が無標になる場合
22	「過」（～たことがある）
23	自他動詞
24	～で
25	のだ文
26	「把」字文
27	補語
28	使役文
29	受身文
30	被を用いない受身文
31	可能
32	自発
33	尊敬
34	敬語
35	ようだ・そうだ
36	「う・よう」と「吧」
37	「～ば～ほど」と「越～越～」
38	介詞
39	決まり文句

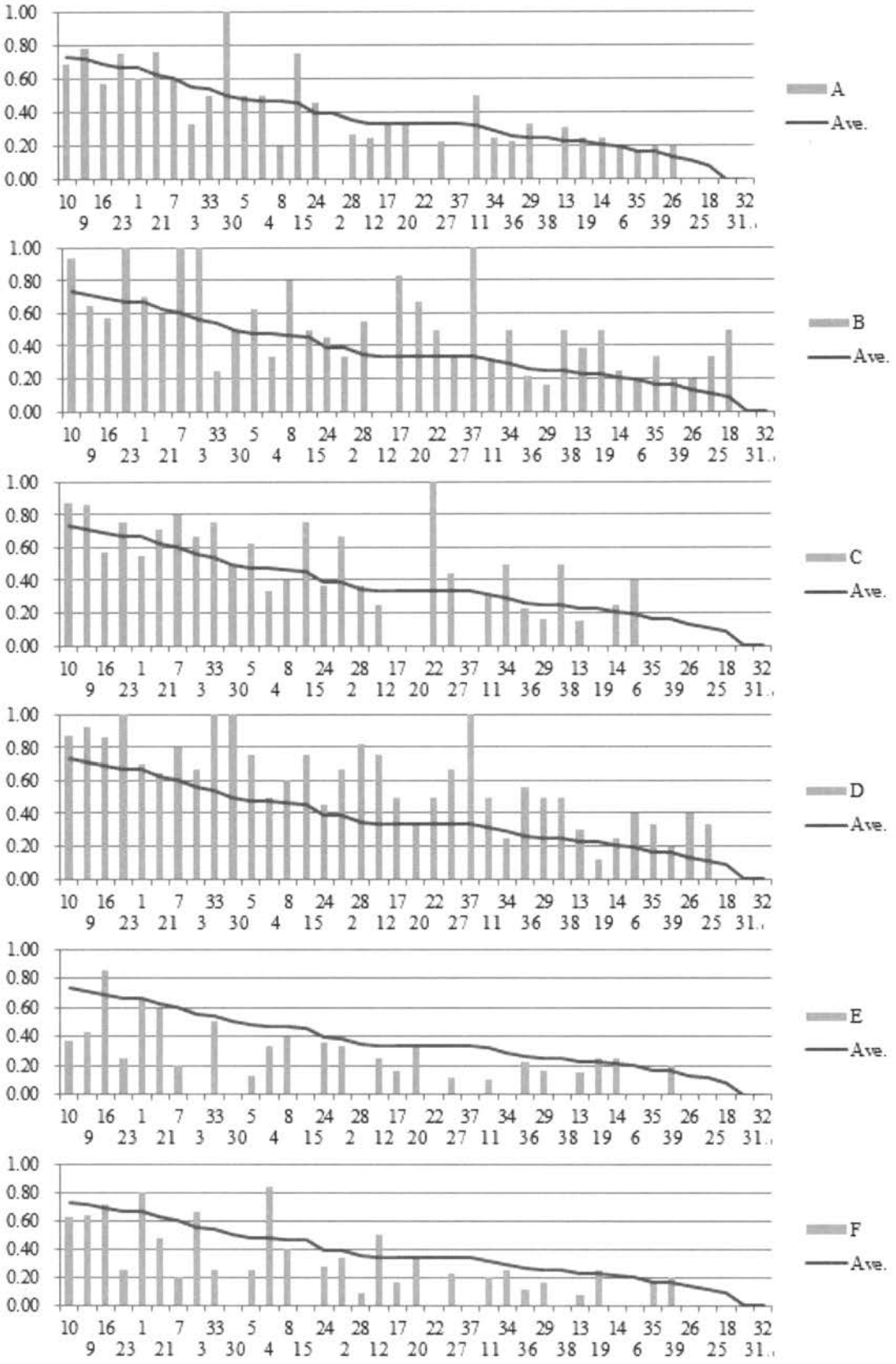


図２　文法項目毎の設問に基づく自動評価スコア（平均値）

図2を個々のシステム毎に見てみると、システムDではほぼすべての項目で相対的に高いスコアを示しているが、No.34（敬語）、No.14（「~ている」と「了」）については全体の平均値を下回っていることがわかる。一方、システムE,Fではほとんどの項目で芳しくない成績を示しているが、例外的にNo.16（前置詞）とNo.14（「~ている」と「了」）、No.39（決まり文句）では全体平均スコアを上回っている。これらの情報は、これまでの自動評価ではできなかった情報であり、ベンダの開発者にとって非常に有用であると思われる。開発者は、文法項目ごとの自社システムの強み、弱みについての情報や、グラフから得られる様々な知見や仮説を、開発項目の優先度づけや開発のアプローチにフィードバックすることができる。

上記のように、本論文の提案手法を用いれば、文法項目という新しい切り口を利用して従来手法ではできなかった観点からの分析ができることが確認できた。特にグラフ等で可視化することによって、文法項目ごとに各システムの特性についての情報を容易に得られることが確認できた。

6. おわりに

本稿では、AAMT 調査委員会の活動報告として、日中機械翻訳文テストセットをもとに自動評価用に作成した「設問自動評価用テストセット」の概要を説明し、その有効性について検討した。設問自動評価は、単独で用いても人間の主観評価と非常に高い相関があり、従来の自動評価尺度や、設問に基づく人間による評価と比較しても、これらを上回るほどであることを示した。また、設問毎の回答分布に基づき、自動評価と人間による評価の結果が十分に一致していることを検証し、設問自動評価が文法項目毎の評価に使えて、エラー分析にも活用できることを明らかにした。

設問に基づく評価を自動化することは、評価者を必要としないという効率的メリットのほかに、評価結果の客観性が向上するという効果がある。語句の有無を問うタイプの設問であっても、二人の人間で評価を行うと16.1%の設問で評価スコアの不一致が見られた。自動評価では、プログラムが評価するため評価による評価の揺れを排除することができる。

AAMT 調査委員会では、機械翻訳の開発者やユーザが誰でも「設問自動評価用テストセット」を利用できるように、2012年度中にWebサービス化してAAMTホームページ上で公開する予定である。

委員会活動報告

機械翻訳・翻訳メモリに関するアンケート結果報告

機械翻訳課題調査委員会ワーキンググループ2

1. はじめに

AAMT機械翻訳課題調査委員会WG2（調査・広報・啓蒙）では、その前身である市場動向調査委員会の時代から調査活動の一環として、機械翻訳ユーザを対象としたアンケートを実施している。本稿では、昨年11月29日に開催されたJTF翻訳祭で、翻訳者の方たちを対象に実施したアンケートの結果について報告する。WG2では、年一回、AAMTのWebサイトを活用して不特定の方を対象にアンケートを実施しているが、今回のアンケートは翻訳者という特定業種のユーザを対象にしたものとなっている。課題調査委員会の委員が翻訳祭の会場に足を運び、来場者に直接、アンケートへの協力を呼びかけるという形をとった。

JTF翻訳祭は、JTF（日本翻訳連盟）が主催する年に1度の翻訳業界最大のイベントであり、昨年の翻訳祭には約700名が参加し、数多くの講演やデモ展示が行われた。AAMTはこの翻訳祭において「展示・デモコーナー」のブースを1コマ契約し、アンケート（他にAAMTの活動紹介、会員勧誘など）を実施した。

以下にアンケートの実施概要を示す。

- 実施日
  - 2011年11月29日 9:00～17:00
- 実施方法
  - 調査員が来場者にアンケートへの回答を依頼。その場で記入してもらい即回収。
  - 機械翻訳と翻訳メモリについてのアンケートを別々に用意し、いずれか使用頻度の多い方のアンケートにのみ回答。
- 進呈品
  - ①1000円のQUOカードを抽選で10名に進呈（その場でくじ引き） ②抽選で1名に10000円のギフト券を進呈(後日送付)。
- 回答者数
  - 70人

2. 設問項目

以下の9項目について質問を行った。機械翻訳と、翻訳メモリの質問項目は問 1-3 を除き基本的に同じ内容である。

問 1-1：使ったことがあるツール（サイト）は?

問 1-2：現在最も良く使うツール（サイト）は?

問 1-3：[機械翻訳ユーザ向け質問]ツール（サイト）にどのような翻訳品質を求めていますか?

[翻訳メモリユーザ向け質問]どのような機能を最も重視していますか?

問 1-4：ツール（サイト）の満足度は?

問 1-5：ツール（サイト）の利用頻度は?

問 1-6：ツール（サイト）の主な利用目的は?

問 1-7：ツール（サイト）で主に使用する翻訳方向は?

問 1-8：ツール（サイト）で何を翻訳していますか?

問 1-9：ツール（サイト）の利用分野は?

この後に回答者の属性（年齢、性別、職業、英語力など）についての設問（5問）が続く。

以下では、今回実施した機械翻訳と翻訳メモリに関するアンケート結果について、回答者の属性を明らかにした後、翻訳ソフトと翻訳メモリを対比させながら説明を行う。

3. 回答者について

70名のアンケート回答者のうち、機械翻訳を主に利用する人が36名、翻訳メモリを主に利用する人が34名でほぼ同数であった。したがって、機械翻訳のアンケート、翻訳メモリのアンケートの回答者数は、それぞれ36名、34名である。

回答者の年齢、性別、職業、英語力についてはアンケ

ートの質問項目に含めて調査を行った。

年齢：40 代が 39%で最も多く、以下 30 代 50 代と続く。30 代から 50 代で全体の 8 割を占める。機械翻訳ユーザよりも翻訳メモリユーザの方がやや高齢であった。

男女比：機械翻訳ユーザ、翻訳メモリユーザとも、男女比はおよそ 7 対 3 であった。

職業：会社員が 58%で最も多く、次が翻訳業であった。翻訳メモリユーザは翻訳業の比率が高く、学生や無職と回答した人はいなかった。

英語力：翻訳祭への来場者ということで英語力は高く、回答者の 7 割以上（翻訳メモリユーザに限れば 8 割以上）が「会話の内容理解や応答が問題なくできる」レベル以上であった。

4. 機械翻訳、翻訳メモリの利用状況

問 1-1：使ったことがある機械翻訳ツール（サイト）、翻訳メモリは?

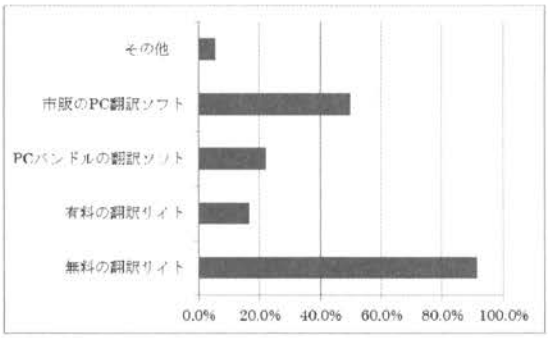


図 1 使ったことがある翻訳ツール

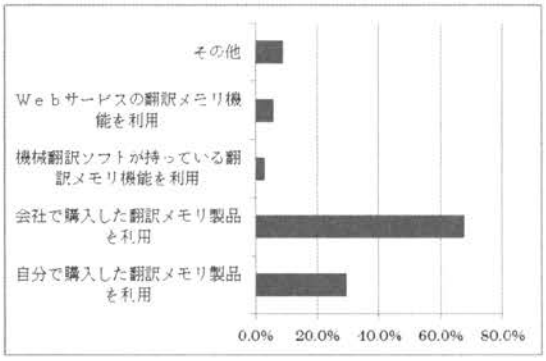


図 2 使ったことのある翻訳メモリ

翻訳メモリは会社で購入して使っているユーザが多いようである。

問 1-2：現在最も良く使う機械翻訳ツール（サイト）、翻訳メモリは?

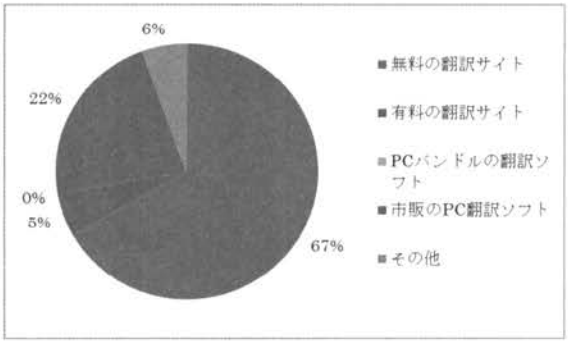


図 3 よく利用する翻訳ツール

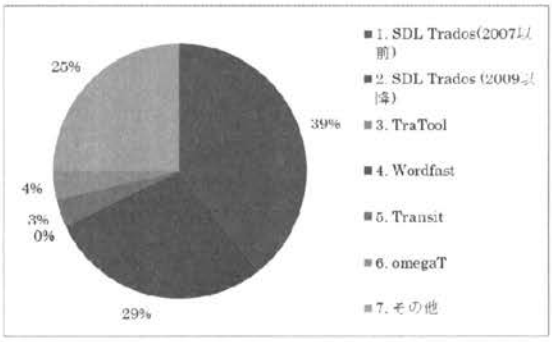


図 4 よく利用する翻訳メモリ

よく利用する翻訳ソフトの約 7 割が無料の翻訳サイトであり、翻訳ソフトを手軽に使うユーザが多いことがわかる。翻訳メモリでは、約 7 割ユーザが最も多く利用するツールとして SDL Trados をあげており、Trados が事実上の標準ツールと言えそうである。

問 1-3：（機械翻訳向け）機械翻訳ツール（サイト）にどのような翻訳品質を求めていますか?

（翻訳メモリ向け）翻訳メモリでは、どのような機能を最も重視していますか

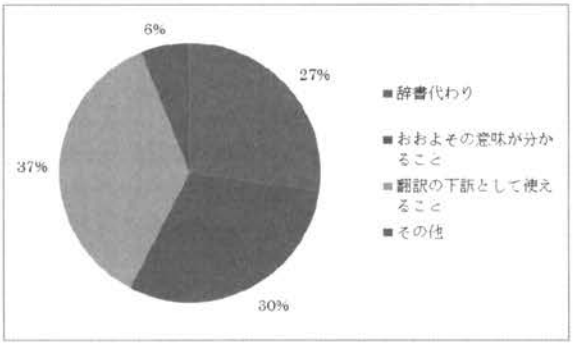


図５　機械翻訳に期待する品質

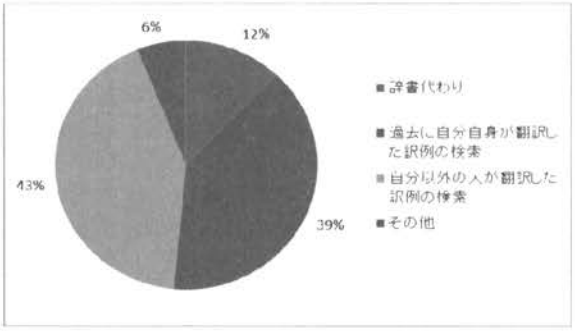


図６　翻訳メモリで重視する機能

翻訳メモリユーザの約４割もの人が、「自分の訳例を検索委する機能」を最も重視する機能と考えている点が興味深い。

問 1-4：機械翻訳ツール（サイト）、翻訳メモリの満足度は?

最も多く利用する機械翻訳ソフト・サイト、または、翻訳メモリの満足度について、５つの切り口からで６段階で評価した結果が図７、図８である。

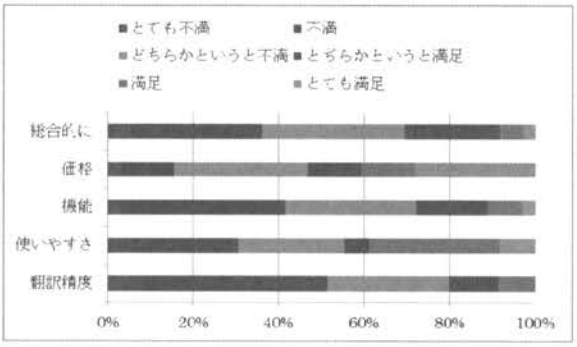


図７　機械翻訳の満足度

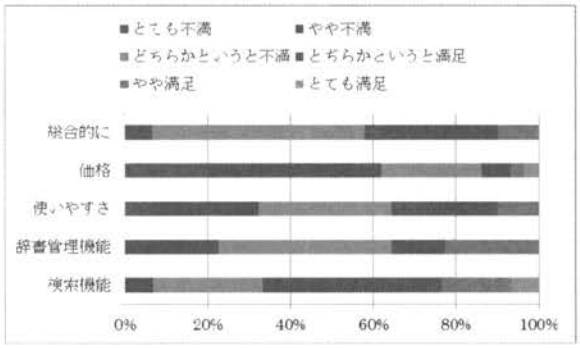


図８　翻訳メモリの満足度

翻訳メモリの基本機能である検索機能に対する満足度が高いのに比べて、機翻翻訳の基本機能である翻訳精度の満足度が低いのが特徴的である。価格面では、無料翻訳が普及している翻訳ソフトの満足率が高い。

問 1-5：機械翻訳ツール（サイト）、翻訳メモリの利用頻度は?

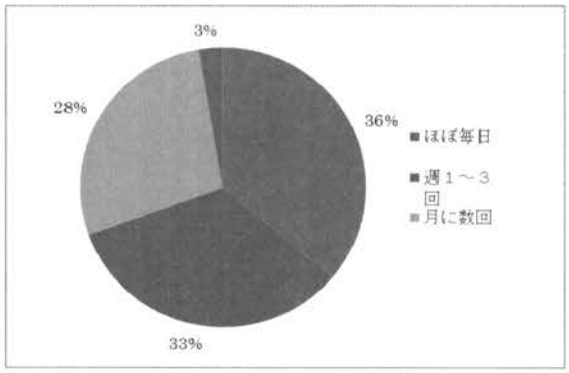


図９　機械翻訳の利用頻度

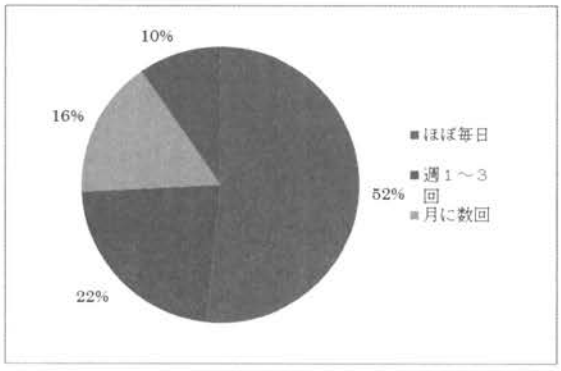


図１０　翻訳メモリの利用頻度

図９，１０を見ると、業務利用が多いと思われる翻訳メモリユーザの半数以上が「ほぼ毎日利用する」と答えており、翻訳メモリ機能のニーズの高さが窺える。一方、機械翻訳ユーザも、約３割が「ほぼ毎日」、７割近くが「週１~３回」と答えており、それなりの利用頻度がある。

問 1-6：機械翻訳ツール（サイト）、翻訳メモリの主な利用目的は?

機械翻訳、翻訳メモリ共に、主に仕事に使われており、プライベートに使うと答えた人は機械翻訳ユーザの 25%、翻訳メモリユーザの 3%のみだった。（図１１）

問 1-7：機械動翻訳ツール（サイト）、翻訳メモリで主に使用する翻訳方向は?

昨今、アジア各国の経済発展が目覚ましいとはいえ、依然として翻訳の対象の大部分（8 割以上）が英語であることがわかる。（図１２）

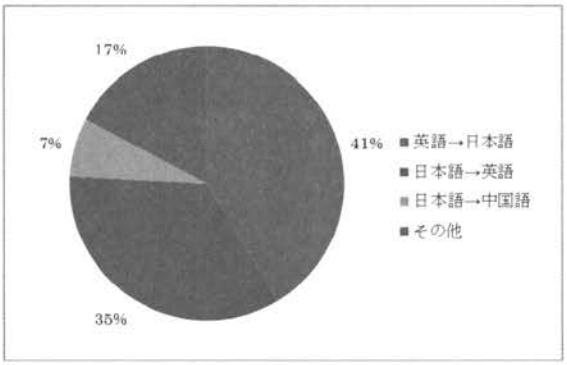


図１１　機械翻訳利用時の翻訳方向

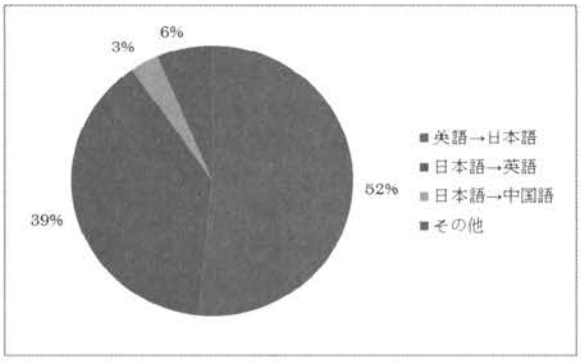


図１２　翻訳メモリ利用時の翻訳方向

日英と英日の比較では、機械翻訳、翻訳メモリ共に英日方向の翻訳で利用されること多いようである。

問 1-8：機械翻訳ツール（サイト）、翻訳メモリで何を翻訳していますか?

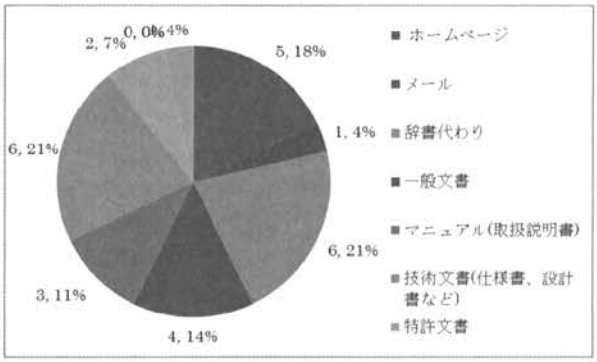


図１３　機械翻訳で翻訳する文書の文種

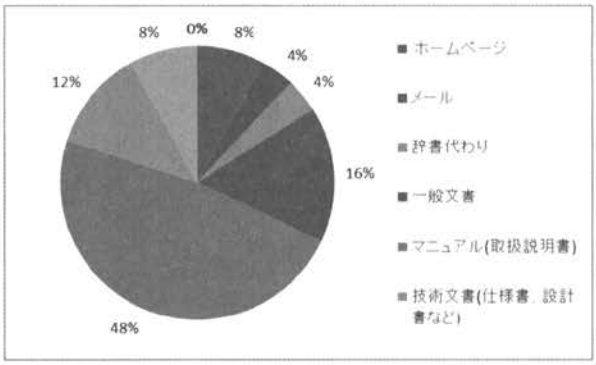


図１４　翻訳メモリ適用対象の文種

機械翻訳は、あらゆるジャンルの文書の翻訳で平均的使われているのに対して、翻訳メモリの適用はマニュアル翻訳時が突出して多いことがわかる（図１３、図１４）。マニュアルの改版では翻訳メモリの例文ヒット率が高くなり、翻訳効率が格段に向上するためと思われる。

問 1-9：機械翻訳ツール（サイト）、翻訳メモリの利用分野は?

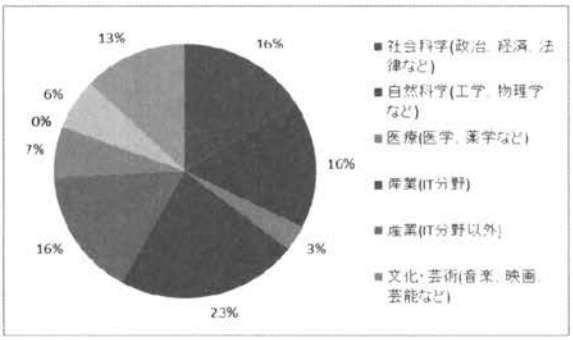


図１５　機械翻訳で翻訳する文書の分野

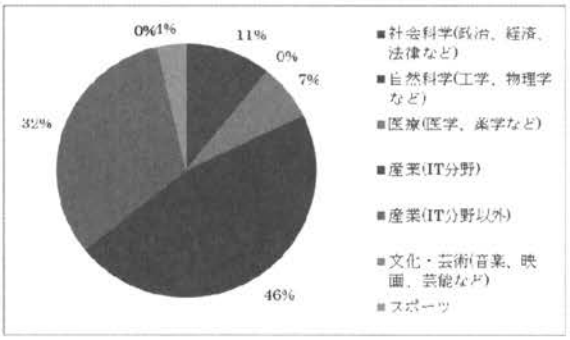


図１６　翻訳メモリ適用対象の分野

図１６より翻訳メモリが適用される翻訳対象文書の分野については、産業分野が８割弱を占めていることがわかる。しかも、その半分以上がIT分野である。問1-8の回答を合わせて見ると、翻訳メモリはITマニュアルの翻訳において最も多く使われているということが言えそうである。

5.　まとめ

本稿では、2011年度のJTF翻訳祭で実施した機械翻訳および翻訳メモリに関するアンケート結果について、アンケートの設問ごとに機械翻訳ユーザと翻訳メモリユーザを対比できる形でまとめたものである。これまで継続的に行ってきたインターネット上のアンケートとは異なり、今回は翻訳業者や「翻訳家」と呼ばれる人たちが回答者の中心であり、機械翻訳や翻訳メモリを日常的に扱うユーザに関して貴重なデータが採取できたと考えている。

たとえば、今回の被験者は非常に英語力の高い集団であるが、そのうちの７割が「ほぼ毎日」「週1~3回」機械翻訳を使っていると回答しており、想像以上に機械翻訳（しかも無料翻訳が大部分）のニーズが高いことがわかった。また、翻訳メモリは機械翻訳に比べてユーザの総合的な満足度は高いが、価格への不満が高いという実態が明らかになった。アンケート結果は設問ごとにグラフ化して掲載したので、他にもいろいろな傾向が読み取れると思う。

機械翻訳課題調査委員会では、今回のアンケート結果を踏まえて、今後、機械翻訳および翻訳メモリのユーザに対して、より具体的な活用事例の調査などを行っていきたいと考えている。

会員のひろば

日英機械翻訳を用いた英作文学習支援ツールのインターネットサービス

株式会社ラングテック
宮崎　正弘

株式会社ラングテック（本社：東京都世田谷区、代表取締役社長：宮崎正弘）[1]は、新潟大学工学部情報工学科・旧宮崎研究室（宮崎正弘教授は2011年3月末で新潟大学を定年退職し、現在、新潟大学名誉教授）の研究成果を活用して自然言語処理応用システムの製品開発を行なう新潟大学発ベンチャーです。このたび当社で研究開発した日英翻訳エンジンLaurelに翻訳過程を透明化・可視化する機能などを組み込んで、日英機械翻訳を積極的に利用した、インターネット上で英語学習者が英作文を学習するのを支援する英作文学習支援ツールを開発し、6月中旬から試行サービス（無料）、8月から本サービス（有料）を開始します。

1．日英翻訳エンジンLaurel

新潟大学の旧宮崎研究室では、時枝誠記が提唱し、三浦つとむが発展的に継承した言語モデルをベースに意味と親和性のある文構造を出力する高度な日本語解析技術[2]とそれを支える言語知識データ（文法、辞書、文型パターン、シソーラスなど）などの自然言語処理基盤技術とその応用システムの研究に20年あまり取り組んできました。

当社では、このような高度な日本語解析技術と日英間の発想の差を吸収する日本語書き換え技術を統合し、意味・ニュアンスの差を可能な限り正しく表現した自然な英語文に翻訳する、規則・用例融合型の日英翻訳エンジンLaurelを研究開発しました。対象は英作文や英会話の基礎となる中学英語の文がその典型である一般的分野の日本語文です。Eメールやツイッターなどの文字で表現される文では話し言葉風の表現も多いことを考慮して、書き言葉だけでなく話し言葉にも対応しています。定型的表現には用例ベースの用例翻訳機能を適用し、用例翻訳機能ではカバーできない非定型的表現にはパターンベースの規則翻訳機能を利用して、高品質な翻訳を可能としています。

英作文や英会話の基礎となる中学英語の文は、構造が複雑な長文はほとんどなく、文構造を解析する上の困難さは回避できるにも関わらず、既存の翻訳ソフトでは案外うまく翻訳できません。これは、英語に翻訳しにくい表現が多様で微妙なニュアンスをもつ文や日本語的発想で書かれ日英間で文構成要素間の対応がとれないこなれた日本語文が多いことだけでなく、態、時制、相、様相、比較、否定、仮定法、同形語・多義語、前置詞、冠詞・所有格、数、省略要素の補完など文法的・意味的に細かな点にも気を配って翻訳する必要があるためです。また、既存の翻訳ソフトでは、入力である原文と出力である訳文だけしか外部から見えないブラックボックスと化しているため、なぜそのような訳文が出力されたのか、翻訳過程でどのような処理が行われ、どのような曖昧さが発生し、それをどのように解消したかなどがわかりません。翻訳ユーザに対して、翻訳過程を透明化・可視化し、翻訳過程で生ずる様々な曖昧性を保持し、選択肢を提示・再試行する機能が望まれます。

日英翻訳エンジンLaurelは、現在、中学英語の文や会話基本表現を翻訳するのに必要な単文、連体節を含む文、従属節・並列節を含む文・名詞句・複合名詞の基本翻訳機能が完成し、既存の翻訳ソフトより高い品質で中学英語の文を翻訳することが可能となっています。また、翻訳過程を透明化・可視化する機能も組み込んでいます。

今後、以下のような機能を順次、日英翻訳エンジンLaurelに組み込んで行く予定です。

- 並列した連体節を含む文、2つ以上の従属節・並列節を含む文、主節や従属節に連体節を含む文など構造が複雑な長文の翻訳機能。
- 構造が複雑な名詞句・複合名詞の翻訳機能。
- よりよい英語文への自動書き換え機能。

以下の中長期的研究課題にも取り組み、その研究成果を順次、日英翻訳エンジンLaurelに組み込んで行く予定です。

- 高度な意味処理に必要な知識を効率的かつ的確に検索で

きる連想機能と多くの分類観点をもった多次元シソーラスを融合した連想型多次元シソーラスの構築。

・連想型多次元シソーラスや文脈を利用したより高度な意味解析や構文・訳語の選択機能の実現。

2．英作文学習支援ツール

英作文学習支援ツールでは、入力文である日本語文の表現を、日英翻訳の観点から大分類・中分類・小分類・細分類の最大４階層に分類・体系化し、末端の分類項目に関連するいくつかの対訳例文対を配置した独自の日本語表現分類体系を構築・利用することによって、各分類項目を学習内容の単位である単元として学習を進めていけるようになっています。本ツールを用いて英作文学習する画面の例を図１に示します。

英作文学習支援ツールをどのように使って学習を行うのか、図１の例を用いて説明します。

・本ツールを使用開始時、最初に日本語表現分類体系の大分類が表示されます。

・「従属節・並列節を含む文の翻訳」をクリックすると、日本語表現分類体系の中分類が表示されます。

・「接続詞による翻訳」をクリックすると、日本語表現分類体系の小分類が表示されます。

図１　英作文学習支援ツールの画面例

・「従属接続詞による翻訳」をクリックすると、日本語表現分類体系の細分類が表示されます。
・「原因・理由」をクリックすると、本項目に関連する対訳例文対が表示されます。
・対訳例文対の日本語文をそのまま翻訳（翻訳過程をみるため）、一部語句を追加・修正のうえ翻訳する場合：そのまま、または一部語句を追加・修正のうえ翻訳してみたい日本語文をクリックして、日本語文入力欄に選択した日本語文が入力されていることを確認のうえ、日本語文入力欄中で必要に応じて一部語句を追加・修正するなどの編集を行ったうえで、「英作文開始」ボタンをクリックすると翻訳を実行し、訳文を英作文翻訳結果に出力すると共に、翻訳過程を出力します。
・翻訳したい文（４８文字以内）を自由入力する場合：日本語文入力欄に翻訳したい日本語文を入力し、「英作文開始」ボタンをクリックすると翻訳を実行して訳文を英作文翻訳結果に出力すると共に、翻訳過程を出力します。
・「ＭｙＤＢへの保存」ボタンをクリックすると、日本語入力文と訳文を利用者ＤＢであるＭｙＤＢへ出力します。
・「クリア」ボタンをクリックすると、日本語文入力欄中の入力済み情報、英作文翻訳結果の訳文、および翻訳過程の出力をクリアします。

翻訳過程の透明化・可視化のため、以下の例に示すような情報を出力します。なお、ことわざ、定型表現など日英間でこなれた文表現同士を直接対応づけた対訳用例によって文全体を翻訳した場合、翻訳過程は出力しません。

－入力日本語文－

彼は風邪をひいたため、学校を休んだ。

・単語分割結果

彼/は/風邪/を/(ひ/い)/た/ため/、
/学校/を/(休/ん)/だ/。

・日本語書き換え結果

彼は風邪をひいたため、彼は学校を休んだ。

・日本語書き換え後基本構造

節１ ため(接続:原因) 節２

・節の翻訳結果

節１：彼が風邪をひく。{時制:過去}
＝＞ He caught a cold.
節２：彼が学校を休む。{時制:過去}
＝＞ He was absent from school.

・翻訳文の基本構造

節２, because 節１.

－英作文　翻訳結果－

He was absent from school, because he caught a cold.

３．英作文学習支援ツールのインターネットサービス

本ツールは、当社内に設置されたサーバ上で動作し、利用するにはＵＩＤとパスワードが必要です。利用者はインターネットから本ツールが動作するＷＥＢを呼び出し、パスワード認証後、本ツールを利用できます。なお、有料サービス契約に先立ち１カ月間の無料の試用期間があり、試用希望者には仮のＵＩＤ・パスワードを配布します。詳細は当社ＷＥＢ[1]を参照下さい。

本ツールは、小・中学校、高校、大学などで英語を学んでいる生徒・学生、英語を教えている先生だけでなく、英語で情報発信するため英語を基礎から学び直したり、英語力をスキルアップしようとしたりする一般社会人にもインターネットを利用した手軽で有効な英作文学習支援ツールとして幅広くご利用いただけるもので、以下のような３種のインターネットサービスがあります。

３．１　ローレルパーソナルサービス

英作文学習を主たる目的として利用する個人向けイータネットサービスです。

＜想定する利用形態＞

利用者個人に１つ配布されるＵＩＤを用い、利用者個人が利用する個人向けのインターネットサービスで、１つのＵＩＤを１人の利用者で利用するためＭｙＤＢの使用が可能です。学校・塾の生徒・学生が自宅のパソコンで英作文の学習演習を行ったり、一般社会人が自宅のパソコンで英作文の学習演習やスキルアップをはかったりできます。

＜基本料金＞

18,000 円／年(1,500 円／月に相当)、年払いです。

３．２　ローレルアカデミックサービス

教育・研究を主たる目的として利用する教育機関向けサービスで、小・中学校、高校、高専、短大・大学・大学院（学部・学科・コース・研究科・専攻・研究室などの単位

でも利用可能）や塾・予備校などでご利用いただけます。

＜想定する利用形態＞

学校や塾にまとめてｎ個配布されるＵＩＤを用いて、学校・塾の構成員が共同利用する教育機関向けのインターネットサービスで、利用者個人に１つずつ配布されるＵＩＤを用いて１つのＵＩＤを１人で利用したり、特定のパソコンに１つずつ割当てられたＵＩＤを用いて複数の利用者で共同利用したりできます。利用者個人にＵＩＤを配布すると、１つのＵＩＤを１人で利用するためＭｙＤＢの使用が可能となり自宅のパソコンからも本サービスを利用できます。学校・塾の教員が学校・塾の教室・コンピュータ室、研究室などのパソコンを用いて英語教育への活用法を研究したり、英作文の問題作成に活用したり、英語の授業に利用したりできます。学校・塾の生徒・学生が教室・コンピュータ室、研究室などのパソコンを用いて、英作文の学習演習を行ったり、通信教育受講生が自宅のパソコンで英作文の学習演習を行ったりできます。

＜基本料金＞

基本料金は、ＵＩＤ数が１、２の契約はそれぞれ18,000円、36,000円を年払い、他は原則として月払いです。ＵＩＤ数が１、２の場合、１ＵＩＤ当たりの基本料金はローレルパーソナルサービスと同じとなりますが、ＵＩＤ数が３以上の場合、ローレルパーソナルサービスより割安な基本料金となります。

－基本料金の算出例－

・学校や塾の生徒５０人に１つずつＵＩＤを配布し、学校・塾や自宅のパソコンで本サービスを利用する場合、５０ＵＩＤ分の料金42,000円／月（１ＵＩＤ当たり840円）が基本料金となります。

・学校の教室やコンピュータ室内の４０台のパソコンに１つずつＵＩＤを付与し、本サービスを利用する場合、４０ＵＩＤ分の料金 37,000 円／月（１ＵＩＤ当たり 925 円）が基本料金となります。

３．３　ローレルコーポレートサービス

英文作成業務、英作文に関する社内教育を主たる目的に利用する企業・団体向けのインターネットサービスです。

＜想定する利用形態＞

企業・団体にまとめてｎ個配布されるＵＩＤを用い、企業・団体の構成員が共同利用するインターネットサービスです。社員に１つずつＵＩＤを配布し、社内や自宅（在宅勤務者の場合）のパソコンを用いて英文作成業務や英作文能力の向上を目指す社内教育に利用したり、社内の特定のパソコンに１つずつＵＩＤを付与し、英文作成業務や英作文能力の向上を目指す社内教育に利用したりできます。

＜基本料金＞

基本料金・お支払いについては、個別にご相談させていただきます。詳細は、当社ＷＥＢ[1]を参照下さい。

４．英作文学習支援ツールの今後の展開

一般に翻訳では複数通りの訳出が可能であり、入力文「彼は風邪を引いたため、学校を休んだ。」でも接続詞「since」「as」などを用いた訳文、無生物主語をたてて全体を単文化した訳文などが考えられます。このような変換・生成過程で生じる構文・訳語の曖昧性の他に、文解析過程では様々な統語的・意味的曖昧性が生じます。このような様々な曖昧性を保持し、必要に応じてその選択肢と選択基準を提示し、ユーザの選択により別な訳文を出力する機能も英語学習者にとって有益です。今後、このような機能も本ツールに組み込んでいきます。また、英語文音声出力機能を組み込んで翻訳文を読み上げる機能を実現すること、英語学習者が犯しやすい翻訳誤り例を収集・データベース化して利用すること、英日機械翻訳の利用なども予定しています。なお、本システムのアドバンス版として、技術英語などの理系英語やビジネス英語を対象とする英作文学習支援ツールにも取り組んでいきたいと考えています。

参考文献

[1] http://www.languetech.co.jp/　（当社のＷＥＢ）

[2] 武本, 宮崎：意味と親和性のある統語構造を出力する日本語文パーザ, 自然言語処理, Vol.14, No.1, PP.19-42, 2007-1.

[3] 宮崎, 武本, 五百川：機械翻訳を用いた英作文・英会話学習支援システム, 信学技報, TL2010-49, PP.19-24, 2011-2.

協会活動報告

(2011 年 11 月～2012 年 5 月)

機械翻訳課題調査委員会

2011 年 11 月 21 日（2011 年度　第 7 回）

① 前回委員会の議事録の確認

② 「第 21 回 JTF 翻訳祭」JTF 翻訳祭参加のための準備

・配布物の確定（AAMT 紹介資料、UTX 資料など）

・アンケート関連の検討（アンケート形式&景品など）

③ まとめと次回委員会について

2012 年 1 月 10 日（2011 年度　第 8 回）

① 前回委員会の議事録の確認

② 今年度の機械翻訳アンケートの実施についての検討

③ まとめと次回委員会について

2012 年 2 月 6 日（2011 年度　第 9 回）

① 前回委員会の議事録の確認

② 機械翻訳アンケートの実施についての検討

・アンケート実施サイトの検討

・アンケート項目の検討

・アンケート収集・集計の検討

③ まとめと次回委員会について

2012 年 3 月 26 日（2011 年度　第 10 回）

① 前回委員会の議事録の確認

② アンケートの実施検討

③ 今年度のまとめと来年度計画の検討

④ 次回委員会について

2012 年 4 月 25 日（2012 年度　第 1 回）

① 前回委員会の議事録の確認

② アンケート実施結果のまとめに向けての検討

③ 「利用者のための評価」として何を行うべきかの議論

④ 「機械翻訳の使い方に関するインタビュー」に関する検討

⑤ 次回委員会について

2012 年 5 月 25 日（2012 年度　第 2 回）

① 前回委員会の議事録の確認
② AdWords とアクセス解析の成果報告
③ UTX の FAQ について
④ UTX 仕様と BOM の規定について
⑤ UTX 変換ツール公開
⑥ 次回委員会について

編集委員会

2011 年 11 月 28 日（2011 年度　第 3 回）

① AAMT Journal No.50 の 進捗状況について
② AAMT Journal No.51 の 記事の執筆依頼計画について
③ AMT Journal No.51 のスケジュールについて

2012 年 4 月 26 日（2012 年度　第 1 回）

① AAMT Journal No.51 の 進捗状況について
② AAMT Journal No.52 の 記事の執筆依頼計画について
③ AMT Journal No.52 のスケジュールについて

インターネット WG

① AAMT ホームページの更新（毎月）
② AAMT Forum メイリングリストの管理
③ AAMT Forum への情報発信
④ 委員会活動におけるメイリングリストの管理
⑤ 会員専用ホームページ更新にむけた準備
⑥ その他事務局ネットワークのインフラ管理全般

AAMT/Japio 特許翻訳研究会

2011 年 12 月 16 日（金）（2011 年度　第 6 回）

1. 前回議事録の確認
2. LREC 関連報告（横山副委員長）
3. 年次報告書関連
4. 研究会ホームページ関連（事務局）
5. NTCIR 関連報告（江原副委員長、安田委員）
6. 次回の開催について

2012年1月27日（金）（2011年度　第7回）

1. 前回議事録の確認
2. 安藤先生のコメント報告（江原副委員長）
3. 年次報告書関連
4. 第2回特許情報シンポジウム関連
5. 第3回産業日本語研究会・シンポジウム関連（Japio）
6. 『中国語文法資源CSSGを用いたSSG木変換式中日翻訳』（Japio）
7. 次回の開催について

2012年3月9日（金）（2011年度　第8回）

1. 前回議事録の確認
2. 第3回産業日本語研究会・シンポジウム報告（Japio）
3. 特許文書の機械翻訳結果の評価手法に関する検討会関連（江原副委員長）
4. 『パテントファミリーにおける対訳文対非抽出部分を利用した専門用語訳語推定』（宇津呂委員）
5. 年次報告書関連（事務局）
6. ご挨拶（Japio・森藤部長、横山副委員長、江原副委員長）
7. 次回の開催について

2012年4月20日（2012年度　第1回）

1. 新年度のご挨拶
2. 新委員のご挨拶（NICT･後藤委員、NTT・須藤委員）
3. 帰国のご挨拶（南カリフォルニア大学より：越前谷委員）
4. 平成23年度納品報告（事務局）
5. 報告書寄稿内容説明（各著者）
6. 特許文書の機械翻訳結果の評価手法に関する検討会関連（江原副委員長）
7. AAMT総会での発表関連（江原副委員長）
8. 第2回特許情報シンポジウム関連（議論）
9. 次回の開催について

編集後記

AAMT ジャーナル編集委員会委員長
筑波大学 システム情報系 知能機能工学域
宇津呂　武仁

AAMT ジャーナル 51 号をお送りします。

今号の巻頭言は、高電社社長の岩城陽子様よりご寄稿頂きました。

今号では、まず、近年、機械翻訳の重要な適用先分野の一つとなっている特許翻訳分野のプロジェクト報告として、海外より 2 件、国内より 1 件のご寄稿を頂きました。海外からは、EPO の Richard Flammer 氏より、欧日間の特許庁の協力関係について、また、WIPO の Christophe Mazenc 氏より、統計的機械翻訳技術を利用した言語横断特許検索アプローチについて、それぞれご寄稿いただきました。一方、国内からは、NTCIR-9 特許翻訳タスクの主催者である NICT の後藤功雄様、隅田英一郎様より、統計的機械翻訳と規則に基づく機械翻訳の比較についてご寄稿頂きました。その他、機械翻訳分野周辺に関するレポートとして、大学関係の研究者の方から 3 件、ならびに産業界から 4 件、それぞれご寄稿いただきました。

一方、AAMT 内の活動報告として、機械翻訳課題調査委員会から、日中機械翻訳テストセットの自動評価について、および、機械翻訳・翻訳メモリに関するアンケート結果報告についてご寄稿していただきました。その他、「AAMT 会員のひろば」の企画におきましては、個人会員の紹介文 1 件を掲載しました。

Memo

Memo

Memo

AAMT

Asia-Pacific Association for Machine Translation

AAMT 入会のご案内

AAMT は、機械翻訳の発展を目的として、機械翻訳の研究者、開発者、製造者、利用者が集まった任意の組織です。委員会による定期的な調査研究をはじめ、機関誌の発行、シンポジウム、セミナー等各イベントの開催など幅広く活動を行っています。
機械翻訳にご関心のあるすべての方にご入会をお勧めします。

＊＊AAMT 会員の特典＊＊

1.AAMT Journal の購読ができます。

会員には、機関誌である AAMT Journal （年 2～3 回発刊予定）が送付されます。購読料は年会費に含まれています。

2.機械翻訳関連の最新情報をメールでお届け

会員専用メーリングリストで、最新の機会翻訳関連の情報をお届けします。
MT 新製品、新サービスの紹介、国際会議、シンポジウムのお知らせ、WEB での MT 関連記事の紹介など盛りだくさんです。

3. AAMT が組織する委員会や調査活動に参加し、機械翻訳や翻訳に関心のある方との交流を深め、知見を広めることができます。

機械翻訳に関する言語資料の調査、広報、標準化活動に参加したり、AAMT Journal や会員専用メーリングリストで、自社製品、サービスの紹介を行うことができます。

4.関連機関の主催する国際会議に参加できます。

IAMT の主催で隔年開催される MT Summit をはじめ、AAMT、AMTA*、EAMT**の主催する会議やワークショップに参加できます。

AMTA* ：Association for Machine Translation in the Americas
EAMT** ：European Association for Machine Translation

年会費は以下の通りです。

法人会員：入会金　1 口　10,000 円　　年会費 1 口 50,000 円
個人会員：入会金　　1,000 円　　年会費　　5,000 円（学生は学生会費 1,000 円）

ご関心のある方は、事務局までお問い合わせください。

アジア太平洋機械翻訳協会（AAMT）
ホームページ：http://www.aamt.info
電子メール：aamt-info@aamt.info

AAMT

Asia-Pacific Association for Machine Translation

入会申込書

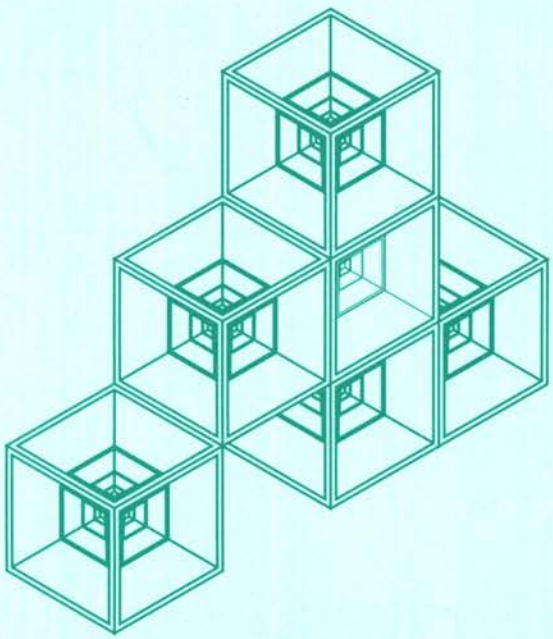
以下の通り、アジア太平洋機械翻訳協会の会員申し込みを致します。

申込日　２０　　年　　月　　日

氏名（ローマ字）	
氏名（漢字）	
電話番号	
メールアドレス	
所属先	
所属先住所	〒
種別	ユーザ　　研究開発者　　その他
機械翻訳に関するお知らせメールの配信	希望する　　希望しない

コメント

AAMT



AAMTジャーナル No.51

発行：アジア太平洋機械翻訳協会（AAMT）
ホームページ：http://www.aamt.info
住所：〒171－0014　東京都豊島区池袋2－55－2
鈴木ビル3階（株）日本システムアプリケーション内
phone：03-5951-3961　fax：03-5951-3966
編集委員会：宇津呂 武仁　小谷 克則　大倉 清司
鈴木 博和　三浦 貢　村上 嘉陽
事　務　局：神崎 享子　荻野 孝野
印　刷　所：株式会社ナビックス

Asia-Pacific Association for Machine Translation
c/o Japan System Applications Co., Ltd.
3F, Suzuki Bldg., 2-55-2 Ikebukuro, Toshima-ku, Tokyo #171-0014 Japan
Phone:+81(0)3-5951-3961　Fax:+81(0)3-5951-3966
URL:http://www.aamt.info