第 4 回特許情報シンポジウム

論文資料集

2016 年 11 月 25 日

グランパークカンファレンス　401 ホール

アジア太平洋機械翻訳協会

一般財団法人日本特許情報機構

実行/プログラム委員会

委員長：	宇津呂 武仁	筑波大学
副委員長：	須藤 克仁	NTT コミュニケーション科学基礎研究所
顧問：	辻井 潤一	産業技術総合研究所
	守屋 敏道	日本特許情報機構
委員：	磯崎 秀樹	岡山県立大学
	今村 賢治	情報通信研究機構
	越前谷 博	北海学園大学
	江原 暉将	元・山梨英和大学
	大塩 只明	日本特許情報機構
	木下 聡	日本特許情報機構
	熊野 明	東芝ソリューション
	黒橋 禎夫	京都大学
	後藤 功雄	NHK 放送技術研究所
	下畑 さより	沖電気工業株式会社
	白土 博之	日本特許情報機構
	綱川 隆司	静岡大学
	中澤 敏明	科学技術振興機構 /京都大学
	二宮 崇	愛媛大学
	横井 巨人	日本特許情報機構
	横山 晶一	元・山形大学
事務局：	小松 浩平	インターグループ
	佐藤 伶奈	インターグループ
	大久保 あかね	インターグループ

まえがき

特許情報の翻訳は、商用機械翻訳システム開発の初期の段階から重要なターゲットと考えられてきた。1987 年に箱根で開催された第 1 回機械翻訳サミットにおいて、日本特許情報機構（Japio）から「発明の名称」の日英機械翻訳の試行について報告されている。その後、システムの改良や専門用語対訳辞書の整備が進み、現在では、「明細書」や「審査書類」の機械翻訳がそれぞれ公知例調査や審査結果の参照などに利用されている。言語対も日英、英日だけでなく日中、中日へと広がってきている。

アジア太平洋機械翻訳協会（AAMT）では、Japio の依頼と支援を受け、2003 年度から AAMT/Japio 特許翻訳研究会を設置し、辻井潤一委員長のリーダーシップの下、特許の機械翻訳に関わるさまざまな技術や事例の調査研究を行ってきた。特許情報シンポジウムはその活動の一環として 2010 年、2012 年、2014 年に開催され、今回が第 4 回目である。前 3 回と同様に、研究者、開発者、利用者、あるいは政策担当者が議論する場を提供することによって、翻訳を中心とする特許情報処理の技術開発と利用を促進することが本シンポジウムの目的である。この目的に沿って以下のようなプログラムを編成した。

午前、午後、各 2 名、合計 4 名の招待講演者に、特許庁における機械翻訳への取り組みについてのご発表、ニューラルネットを用いた自然言語処理の最新の研究動向についてのご発表、国外での機械翻訳の産業利用事例における分野適応に関するご発表、および特許翻訳における機械翻訳活用への期待に関するご発表をお願いした。これらの講演を通じて、機械翻訳の技術と特許分野における利用の現状について理解を深めていただけるものと考える。

AAMT/Japio 特許翻訳研究会からは、研究会活動の紹介を兼ねて、機械翻訳の評価に関して二件の報告をさせていただく。一つは機械翻訳の研究開発を進める上で重要な役割を果たす翻訳システム/翻訳文の自動評価手法に関する発表である。もう一つは、アジアにおける競争型機械翻訳システム評価コンテストにおける人手評価結果の分析に関する発表である。これらの報告に加えて、AAMT において、文章と翻訳の品質管理を目的として用語管理と実務日本語の規格整備に取り組む開発者に講演をお願いした。

投稿ベースの一般発表も興味深い 3 編の論文を採用することができた。特許明細書の翻訳に関する論文に加えて、特許マップの自動生成手法、特許分類の推定手法に関する論文をご発表いただく。

以上のように、特許情報処理に関するさまざまな立場からのさまざまな内容の発表からなるプログラムを編成することができた。快くお引き受けいただいた招待講演者と興味深い論文を投稿していただいた方々に感謝の意を表する次第である。この分野に関心をもつ参加者の皆様が意見を交換し、理解を深めることにより、特許情報処理がますます発展することを期待する。

2016 年 11 月第 4 回特許情報シンポジウム
実行 / プログラム委員会
委員長　　宇津呂武仁

プログラム・目次

セッション 4（特別講演＆一般講演）

16:00-16:30　特別講演「文章と翻訳の品質を改善する
—構造化用語データ UTX による用語管理と実務日本語ルール」
山本 ゆうじ（秋桜舎） …… 88

16:30-16:50　一般講演 1「課題とその対象を軸としたマトリクス型特許マップの自動生成手法の提案」
小野寺 大輝、吉岡 真治（北海道大学） …… 109

16:50-17:10　一般講演 2「特許明細書の翻訳時で注意すべきこと」
吉川 潔（翻訳業）…… 118

17:10-17:30　一般講演 3「特許文献中の重要語に着目した特許分類の推定」
綱川 隆司、佐々木 深、西田 昌史、西村 雅史（静岡大学）…… 123

17:30-17:40　閉会挨拶
守屋 敏道（日本特許情報機構）

18:00-　懇親会（グランパークカンファレンス　303 会議室、会費無料）

招待講演 1

「特許庁における機械翻訳の取り組み」

特許庁における機械翻訳の取り組み

特許情報シンポジウム
平成28年11月25日

特許庁総務部総務課特許情報室
加藤啓

目次

1. 背景
2. 機械翻訳の取り組みの現状
   - 外国文献検索への活用
   - 審査情報照会への活用
   - 外部研究機関との連携
3. 機械翻訳の課題

目次

審査実務の概要

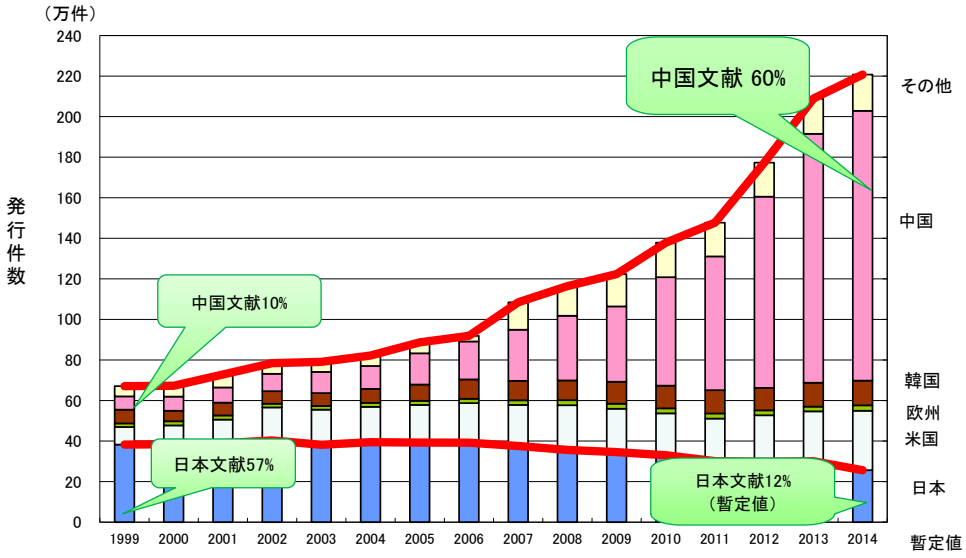
審査官による実体審査

先行技術調査とは
特許出願の出願日よりも前に公開されている文献等（外国文献含む）を調査すること

特許出願された発明が新規性、進歩性を有するかどうかは、先行技術調査の結果に基づき判断するため、**先行技術調査は非常に重要**

先行技術調査における課題

- 近年、日本語以外の言語で記載された外国特許文献の割合が急増。中国語、韓国語でしか読むことができない特許文献が、世界の特許文献の60%
- 中国語、韓国語を含む外国語特許文献を精度良く検索できる環境の整備が不可欠となっている

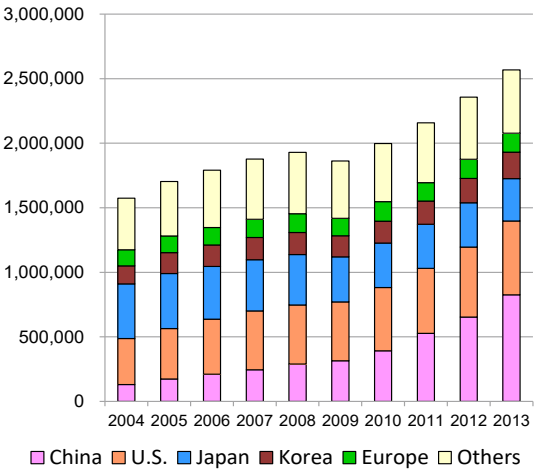


（注）世界で発行された特許文献（実用新案含む）を言語別に整理し、重複を排除したもの。複数の国に出願され、公開された同内容の特許文献について、日本語があるものは日本の特許としてカウント。日本語がない場合には、米国（英語）、欧州（英語、仏語、独語）、韓国（韓国語）、中国（中国語）の順で該当する国・地域（言語）の特許文献としてカウント。2014 年の発行件数は暫定値。

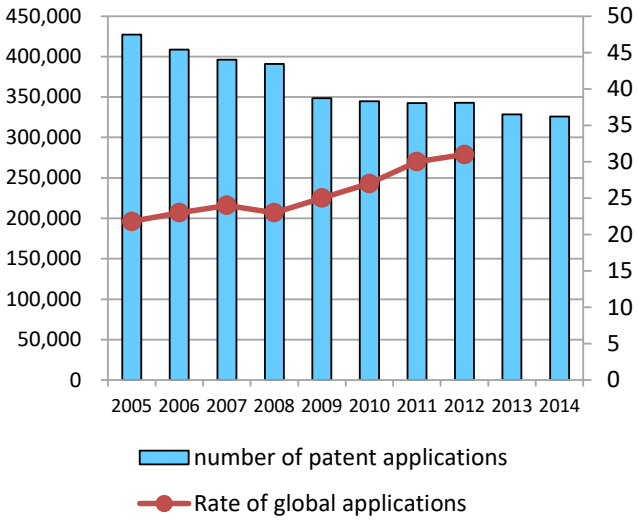
知財活動のグローバル化

- 企業活動のグローバル化が進むことで、世界における出願件数は急増。
- 日本の出願人のグローバル出願率は増加傾向。

【世界の特許出願件数】



【日本の出願人のグローバル出願率の推移】

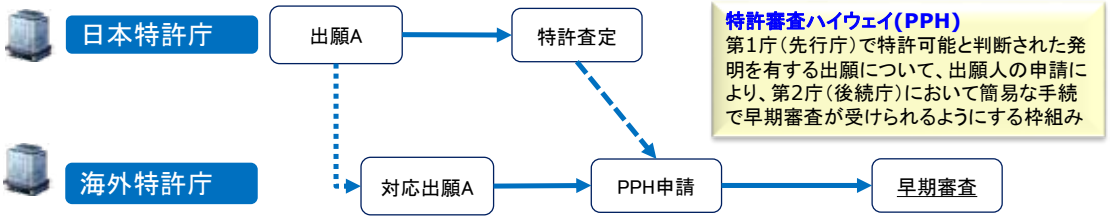


※グローバル出願率：
国内へ出願される特許出願のうち外国にも特許出願される件数の比率。外国出願した国の数は、グローバル率に影響を及ぼさない。なお、特許出願には、国内出願に基づかず直接国際出願の受理官庁としての日本国特許庁に出願されたPCT 出願を含む。

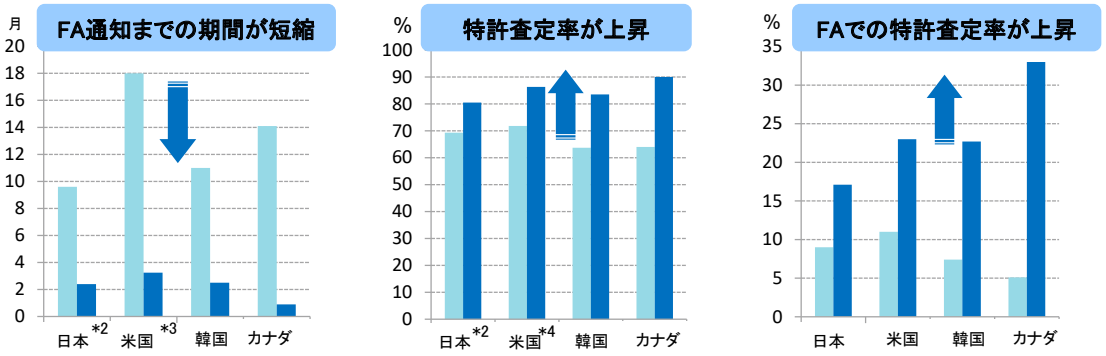
日本の審査結果の海外発信

特許庁

- 特許審査ハイウェイ（PPH: Patent Prosecution Highway）により、日本の審査結果を利用して、海外で権利を円滑に取得可能
- 日本は、米欧等先進国やアセアンを含む35の知財庁とPPHを実施



特許審査ハイウェイ（PPH）
第1庁（先行庁）で特許可能と判断された発明を有する出願について、出願人の申請により、第2庁（後続庁）において簡易な手続で早期審査が受けられるようにする枠組み



（集計期間：2015年1月～6月）

＊1 米国のPPH出願の統計にはPCT-PPHを含んでいる。 ＊2 集計期間（全出願）：2014年1月～12月
＊3 集計期間（全出願）：2013年7月～12月 ＊4 集計期間（全出願）：2014年7月～12月

審査結果の海外発信における課題

特許庁

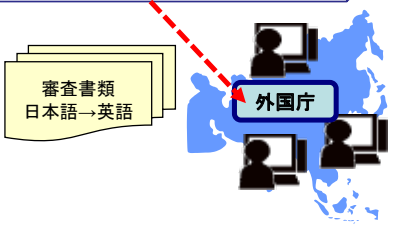
拒絶理由通知書（例）

特許出願の番号　特願2014－●●●●●●
起案日　平成26年11月11日
特許庁審査官　加藤　啓
特許出願人代理人　●●●●　様
適用条文　第29条第1項、第29条第2項

この出願は、次の理由によって拒絶をすべきものである。これについて意見が あれば、この通知書の発送の日から60日以内に意見書を提出して下さい。

理　由

1.　この出願の下記の請求項に係る発明は、その出願前に日本国内又は外国において、頒布された下記の刊行物に記載された発明又は電気通信回線を通じて公衆に利用可能となった発明であるから、特許法第29条第1項第3号に該当し、特許を受けることができない。



日本の審査結果は日本語で記載されていることから、**海外審査官等が参照する際には、翻訳をして発信する必要がある**

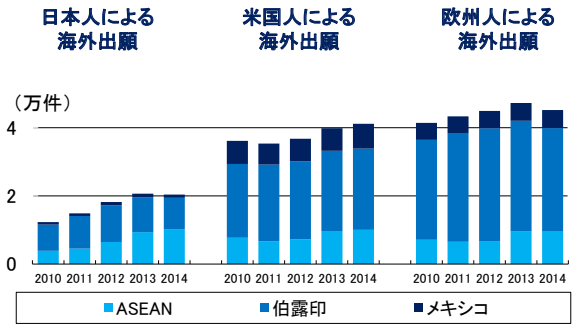
海外特許情報の確認の必要性

- 我が国企業のグローバル展開が急速に進んでいるため、進出先の出願・権利の内容を確認する必要性が急増
- ASEAN諸国など、日本語・英語以外の言語でしか内容が確認できない場合も多く、**出願・権利の内容を翻訳文により容易に確認できる環境が必要**

中期的有望事業展開先国・地域

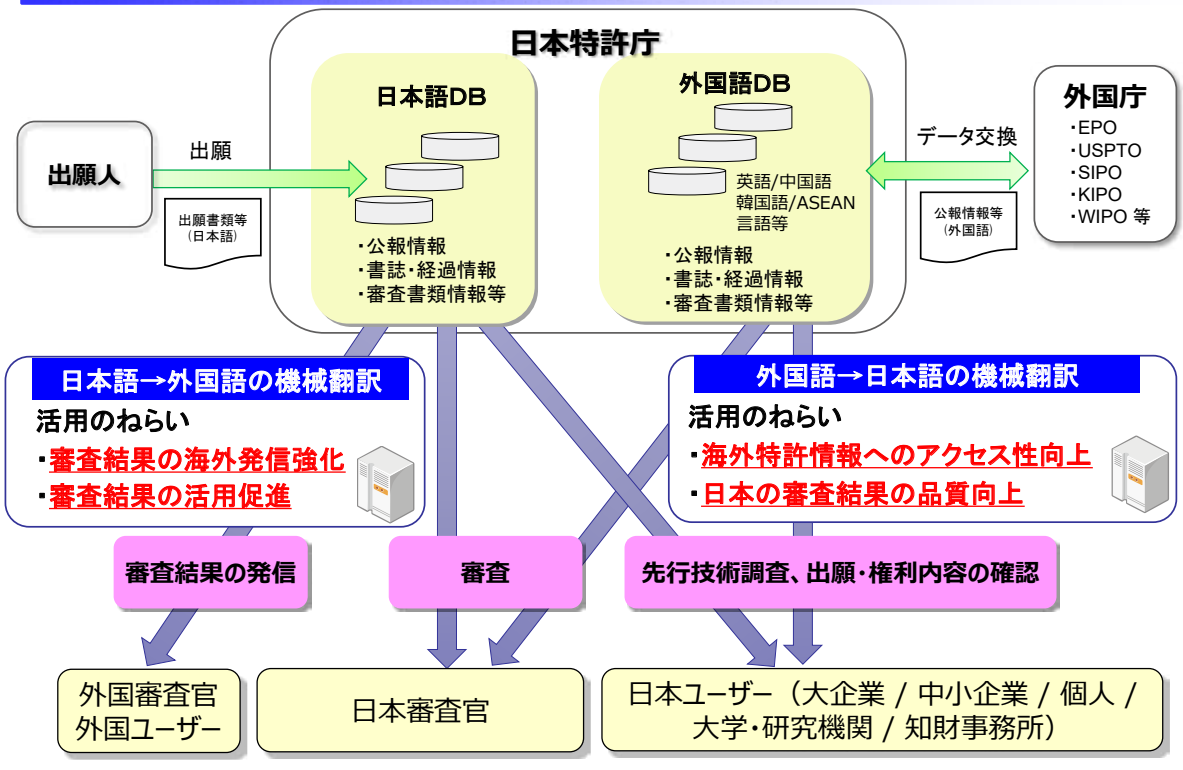
順位 2015	←	2014	国・地域名	回答社数(社) 2015 (計) 433	2014 499	得票率(%) 2015	2014
1	ー	1	インド	175	229	40.4	45.9
2	ー	2	インドネシア	168	228	38.8	45.7
2	↑	3	中国	168	218	38.8	43.7
4	ー	4	タイ	133	176	30.7	35.3
5	ー	5	ベトナム	119	155	27.5	31.1
6	ー	6	メキシコ	102	101	23.6	20.2
7	↑	8	米国	72	66	16.6	13.2
8	↑	11	フィリピン	50	50	11.5	10.0
9	↓	7	ブラジル	48	83	11.1	16.6
10	ー	10	ミャンマー	34	55	7.9	11.0
11	↑	12	マレーシア	27	46	6.2	9.2
12	↓	9	ロシア	24	60	5.5	12.0
13	↑	14	シンガポール	20	25	4.6	5.0
14	↓	13	トルコ	17	26	3.9	5.2
14	↑	15	韓国	17	20	3.9	4.0
16	↑	17	台湾	16	19	3.7	3.8
17	↓	15	カンボジア	14	20	3.2	4.0
17	↑	18	ドイツ	14	9	3.2	1.8
19	ー	19	サウジアラビア	7	7	1.6	1.4
20	↑	25	バングラデシュ	6	6	1.4	1.2
20	↑	32	ラオス	6	3	1.4	0.6
20	↑	32	英国	6	3	1.4	0.6

インド・アセアン等新興国への特許出願件数



（出典）WIPO（世界知的所有権機関）統計（2015年12月現在）
（注）欧州：EPC（欧州特許条約）加盟国。
　　欧州からの出願件数は、各年末時点のEPC加盟国の出願人による出願件数。
（注）「五庁以外」及び「その他」には、台湾への出願は含まれていない。
（注）「ASEAN」は、インドネシア、カンボジア、シンガポール、タイ、フィリピン、ブルネイ、ベトナム、マレーシアの件数に限る。

特許庁における機械翻訳活用のねらい



目次

1. 背景
2. 機械翻訳の取り組みの現状
   - 外国文献検索への活用
   - 審査情報照会への活用
   - 外部研究機関との連携
3. 機械翻訳の課題

中韓文献翻訳・検索システム①

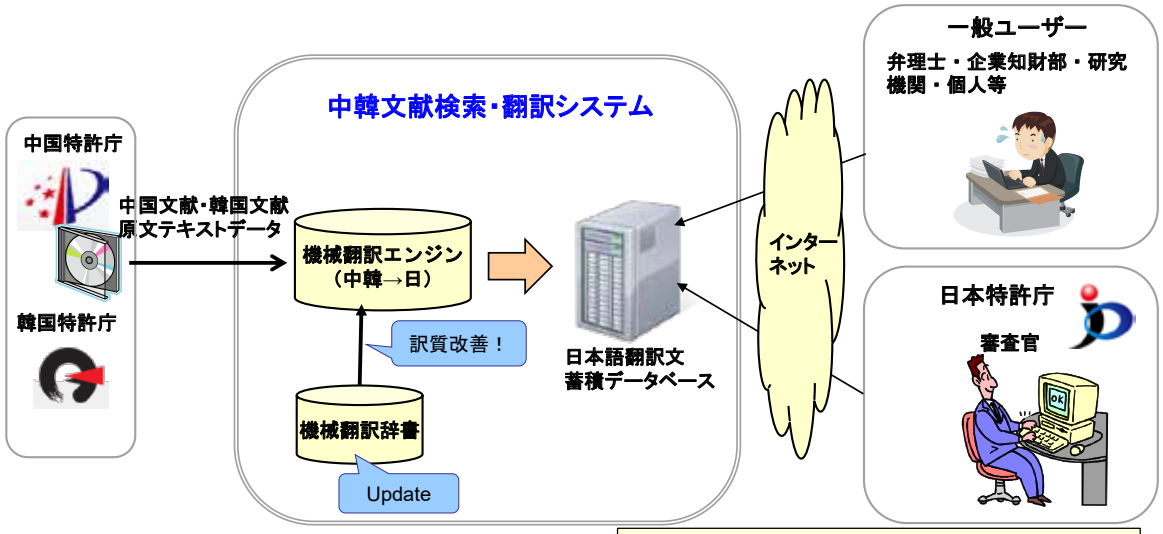
- 急増する中国・韓国語特許文献の検索の利便性を向上させるために、中国・韓国語特許文献の全文を日本語に機械翻訳し、日本語でテキスト検索及び照会を可能とするシステムを整備し、平成27年1月より本格稼働。
- 特許庁審査官に加え、一般利用者も利用が可能。
- 2003年以降に公開された中韓文献（1500万件以上）が蓄積され、今後も順次文献を蓄積。

＜中韓文献翻訳・検索システム＞
http://www.ckgs.jpo.go.jp
（サービス提供時間：開庁日の8時から22時まで）

中韓文献翻訳・検索システム②

特許庁

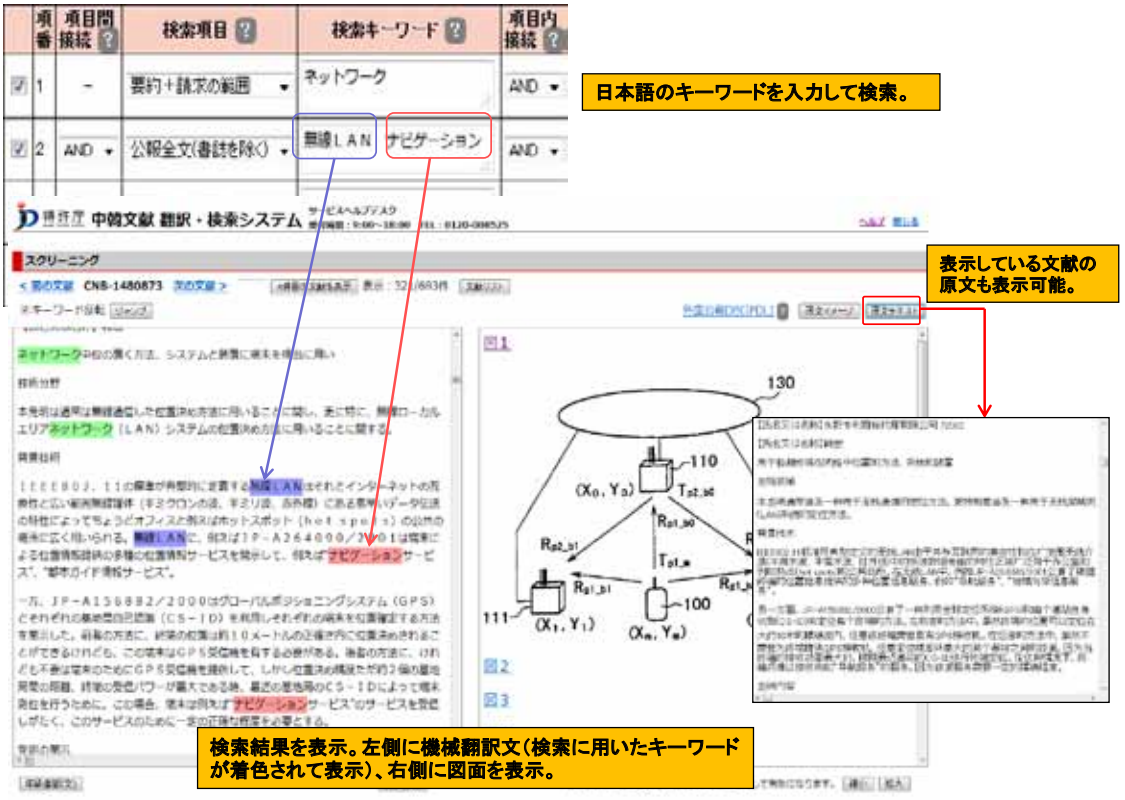
- 急増する中国・韓国語特許文献の検索の利便性を向上させるために、中国・韓国語特許文献の全文を日本語に機械翻訳し、日本語でテキスト検索及び照会を可能とするシステムを整備し、平成27年1月より本格稼働。
- 特許庁審査官に加え、一般利用者も利用が可能。
- 2003年以降に公開された中韓文献（1500万件以上）が蓄積され、今後も順次文献を蓄積。



＜中韓文献翻訳・検索システム＞
http://www.ckgs.jpo.go.jp
（サービス提供時間：開庁日の8時から22時まで）

中韓文献翻訳・検索システム③

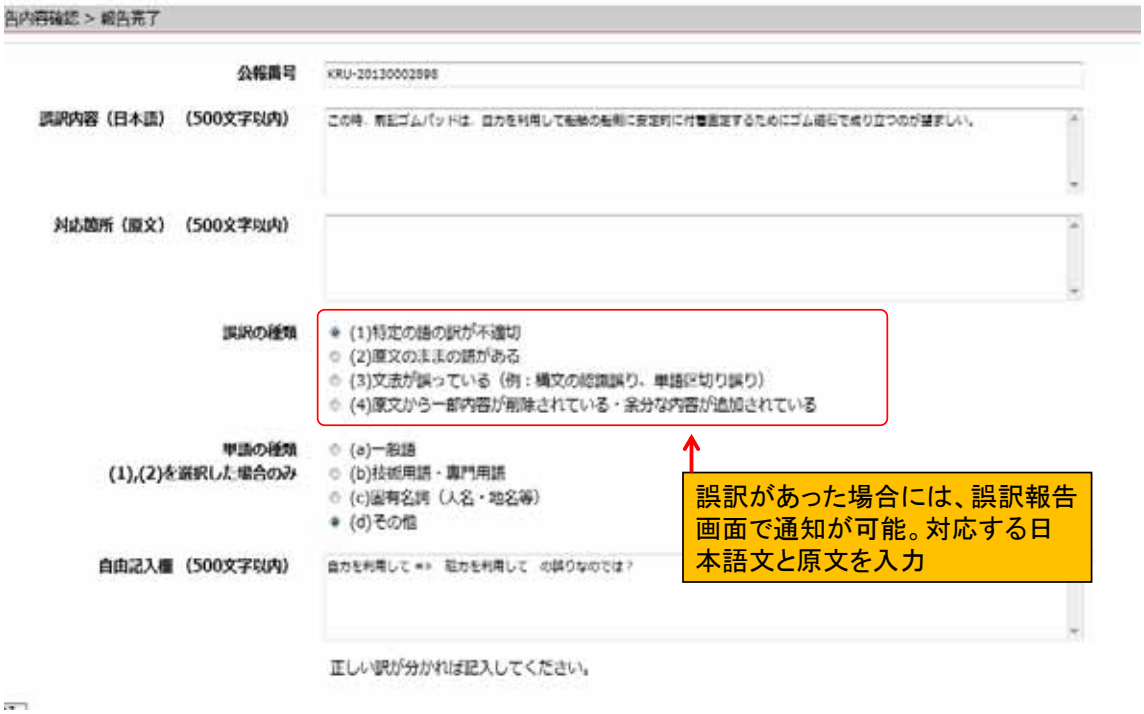
特許庁



中韓文献翻訳・検索システム④

特許庁

誤訳報告画面

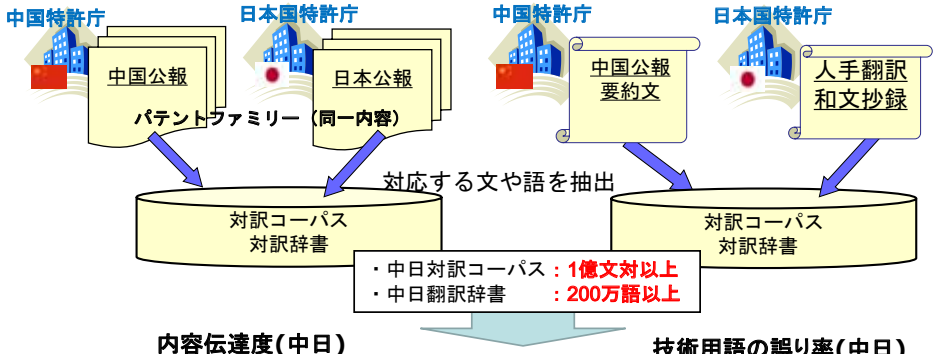


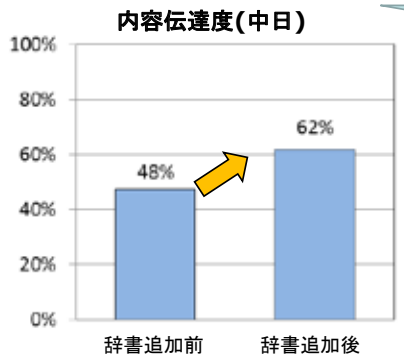
誤訳があった場合には、誤訳報告画面で通知が可能。対応する日本語文と原文を入力

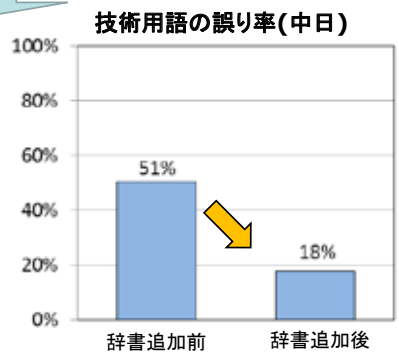
中日機械翻訳の訳質向上の取り組み

特許庁

- パテントファミリー等から、対訳コーパス・対訳辞書を継続的に作成
- 作成した辞書は、中韓文献検索・翻訳システムの翻訳精度向上に利用







※内容伝達度は中国公報の機械翻訳文700文選定、技術用語は400語抽出し、特許庁作成辞書（約200万語）の追加の前後で人手評価。

特許文献機械翻訳の品質評価手順

特許庁

- 特許庁は、機械翻訳結果の品質を適切に評価するための指針として、2014年に「特許文献機械翻訳の品質評価手順」を作成。 http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

1. 内容の伝達レベルの評価 （原文に含まれる重要な情報をどの程度正確に伝達しているかについて、5段階で評価）

5	すべての重要情報が正確に伝達されている。(100%)
4	ほとんどの重要情報は正確に伝達されている。(80%～)
3	半分以上の重要情報は正確に伝達されている。(50%～)
2	いくつかの重要情報は正確に伝達されている。(20%～)
1	文意がわからない、もしくは正確に伝達されている重要情報はほとんどない。(～20%)

2. 重要技術用語の翻訳精度の評価 （あらかじめ選定した重要技術用語について、4段階で評価）

A(適訳)	人手翻訳に照らし、技術的に同義かつ一般的に用いられる訳語である。
B(可訳)	技術用語として一般的に用いられる訳語ではないが、意味はおおむね正しい。
C(誤訳)	誤訳である。
D(不訳)	未知語、訳漏れである。

3. 流暢さの評価（機械翻訳結果の、文としての読みやすさ、理解しやすさのみを5段階で評価）

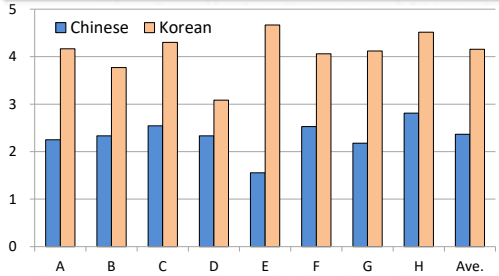
5	文意が明解で、人間が書いた日本語文に近い。
4	日本語文として不自然な箇所を若干含むが、文意は明解である。
3	日本語文として不自然な箇所があり、文意がわかりにくい。
2	日本語文法規則に反する表現をかなり含む。文意がわからない。
1	日本語文として成立していない。

中韓文献翻訳・検索システムの翻訳品質評価

特許庁

- 特許庁が策定した品質評価手順に基づき、中韓文献検索・翻訳システムの中日機械翻訳文及び韓日機械翻訳文の品質評価を実施
- 中日機械翻訳文については更なる品質向上が求められる

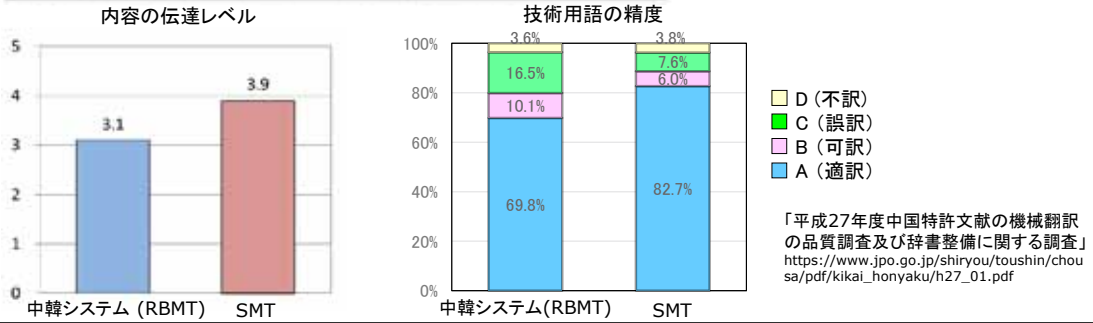
中日翻訳と韓日翻訳の比較（内容の伝達レベル）



A：生活必需品
B：処理操作、運輸
C：科学、冶金
D：繊維、紙
E：固定構造物
F：機械工学
G：物理学
H：電気

①中日翻訳と韓日翻訳の比較は、それぞれの機械翻訳100文を、5段階で人手評価
②中韓文献検索・翻訳システム(ルールベース)と統計翻訳(みんなの自動翻訳)の翻訳精度を比較。それぞれ機械翻訳700文、技術用語700語を人手評価

中韓システム(RBMT)と統計翻訳(SMT)との比較（中日翻訳）

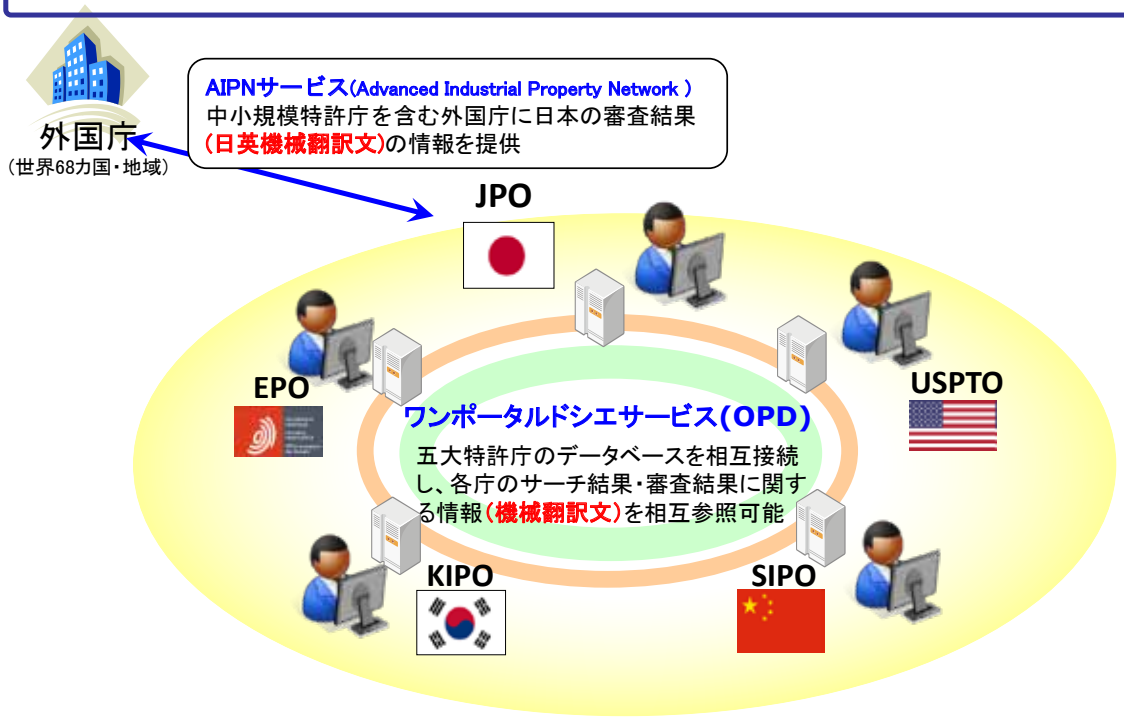


「平成27年度中国特許文献の機械翻訳の品質調査及び辞書整備に関する調査」
https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/kikai_honyaku/h27_01.pdf

目次

日英機械翻訳の活用（審査結果の海外発信）①

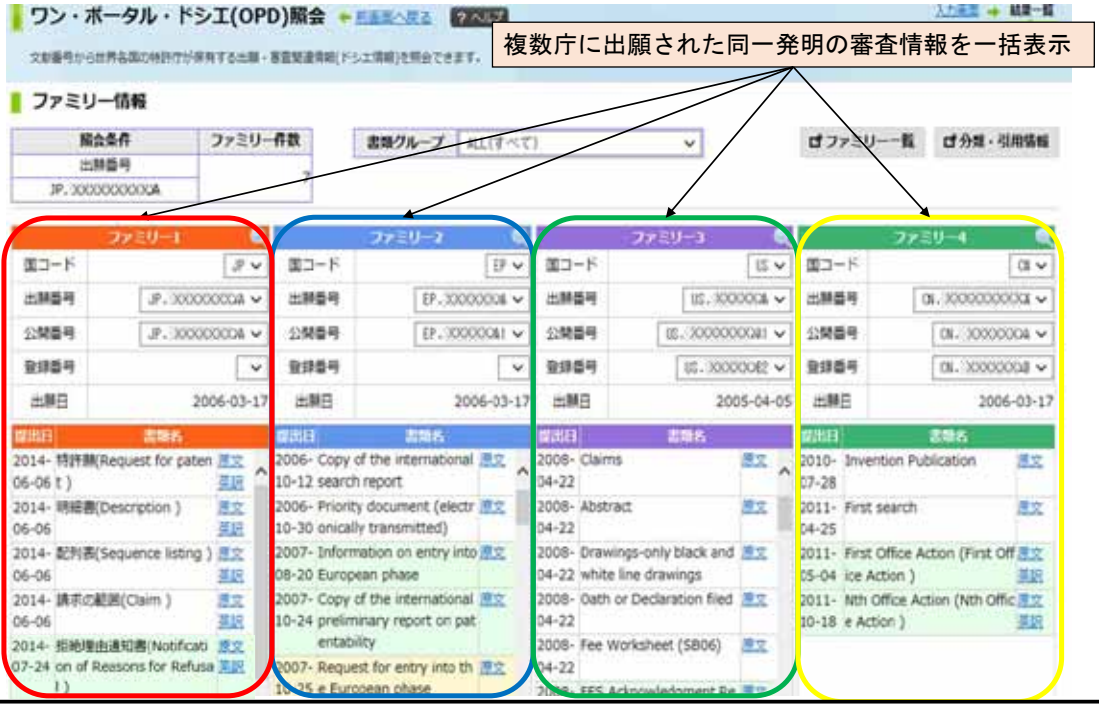
- インターネットを利用したサービスを通じて、海外の審査官、一般ユーザーに向けて、我が国の「世界最速・最高品質の特許審査」の**日英機械翻訳文**を提供。



日英機械翻訳の活用（審査結果の海外発信）②

特許庁

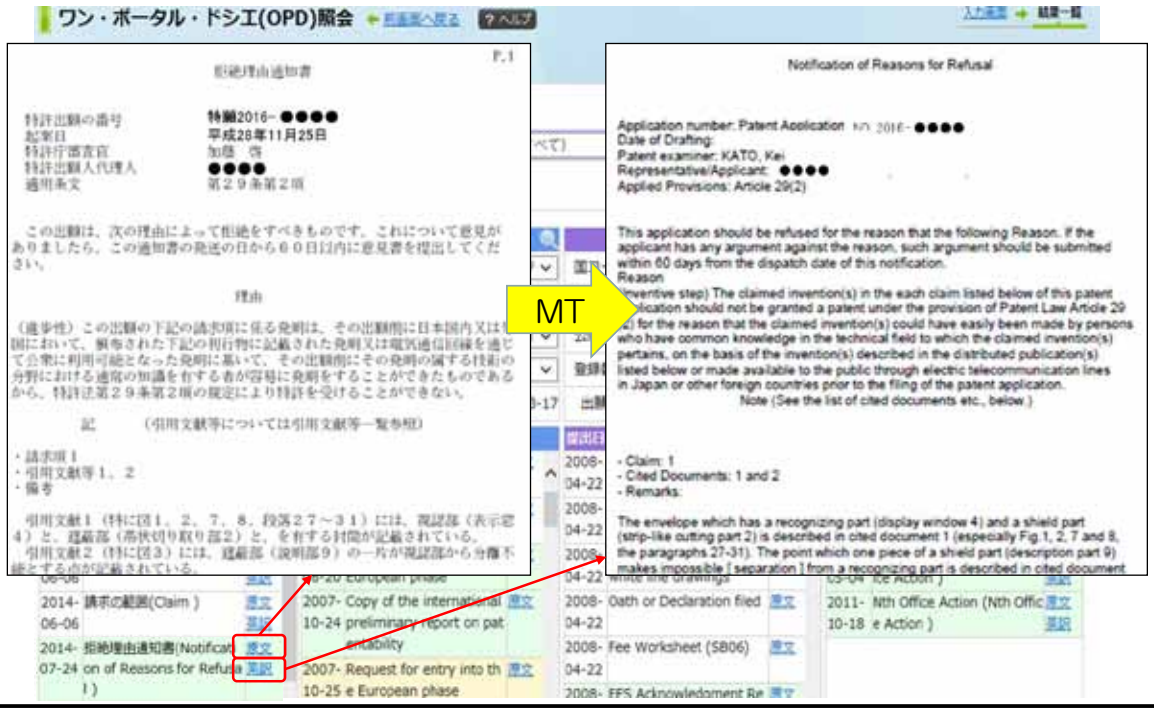
- ワンポータルドシエ（OPD）では、五庁のサーチ結果・審査結果等に関する審査情報を、一括して見やすい形式で相互参照可能（特許情報プラットフォームから提供）。



日英機械翻訳の活用（審査結果の海外発信）③

特許庁

- JPOからは原文に加えて、**英語の機械翻訳文を併せて提供している**
- 特に翻訳精度が求められているのは 、（特許審査ハイウェイの関係から）**①特許になった請求項の機械翻訳**、**②拒絶理由・拒絶査定の機械翻訳**である



審査情報（日英機械翻訳文）へのアクセス数

- 日本の審査情報（日英機械翻訳文）へのアクセス数は年々増加しており、2015年は年間で44万件以上のアクセスがあった。高精度な機械翻訳を提供する重要性がますます高まっている。

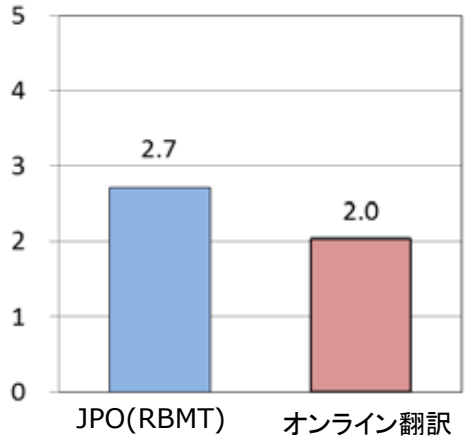
（回）ワンポータルドシエの審査情報（英日機械翻訳文）へのアクセス数

※海外からJPOのOPDで提供している書類情報(明細書等を含む)を参照した回数

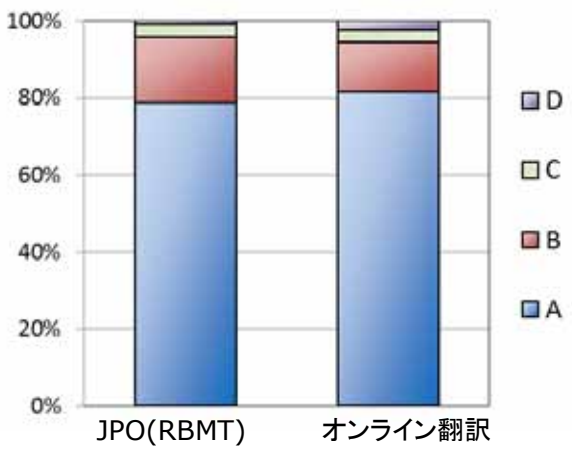
日英機械翻訳の品質評価

- JPOが策定した品質評価手順に基づき、審査書類（拒絶理由通知）の日英機械翻訳の品質評価を実施
- JPOの翻訳エンジン(RBMT)を利用した翻訳文の方が、オンライン翻訳を利用した翻訳文よりも高評価となったが、さらなる品質の改善が求められる

内容の伝達レベルの評価



重要用語の評価



更なる機械翻訳文の品質の改善が必要

「平成27年度特許審査関連情報の日英機械翻訳文の品質評価に関する調査」
https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/kikai_honyaku/h27_03.pdf

※オンライン翻訳はEspacenet(グーグル翻訳)を利用

日英機械翻訳の品質改善の取り組み

■ 日英機械翻訳システムにおける訳質向上のための定常的な取り組み

海外特許庁からの誤訳フィードバック

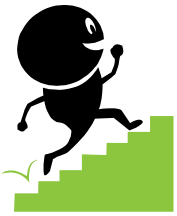
- AIPNユーザー（EPO、USPTOをはじめとする海外特許庁審査官）からの誤訳フィードバックを分析のうえ、辞書登録

未知語の収集・登録

- AIPN，IPDLにおいて、翻訳不可能な単語（未知語）のログを収集し、ユーザー辞書に追加登録（5,000語/年）
- 現在約9万語の特許用語等が登録されている

翻訳メモリの構築

- 審査官が拒絶理由通知等に利用する定型表現（汎用文例）の登録
- 現在、約1,570 文/語が登録されている



目次

機械翻訳に関する外部機関との連携・協力①

特許庁

- 平成26年7月、特許庁と国立研究開発法人情報通信研究機構（NICT）は、特許文献の機械翻訳に関する連携・協力を進めることに合意
- 合意後、継続的に協議を重ねつつ、対訳コーパスや翻訳エンジンの作成などを実施

機械翻訳に関する外部機関との連携・協力②

特許庁

協力のイメージ

NICTとJPOの協力

自動翻訳の構築

NICT ⇄ JPO

リソースの提供・品質評価など
フィードバック

↓ 多言語の自動翻訳技術の民間提供

日本の特許検索・調査の高度化、効率化に貢献

特許文献をもとにした対訳コーパスの作成・外部提供

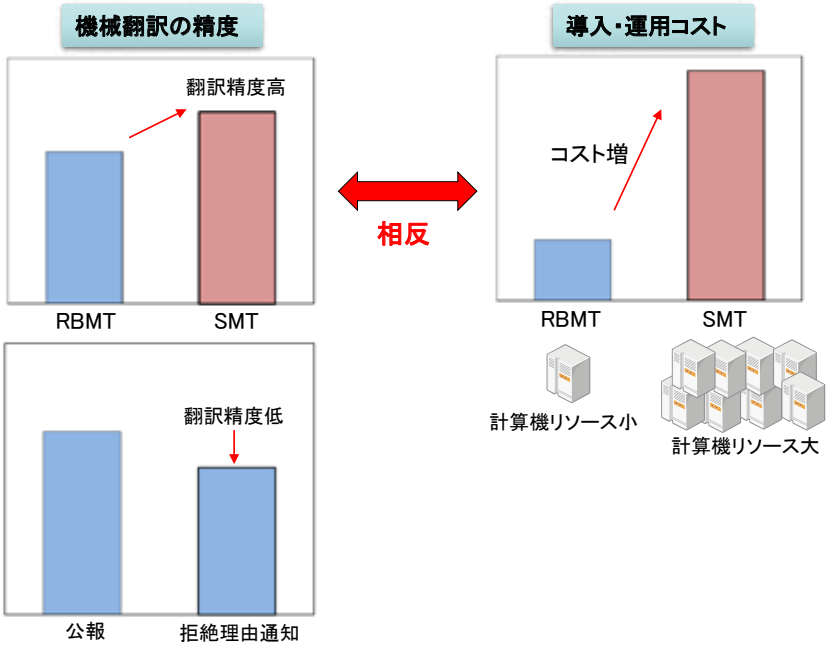
翻訳エンジンの開発・検索環境の検証

目次

1. 背景
2. 機械翻訳の取り組みの現状
   - 外国文献検索への活用
   - 審査情報照会への活用
   - 外部研究機関との連携
3. **機械翻訳の課題**

機械翻訳における課題①

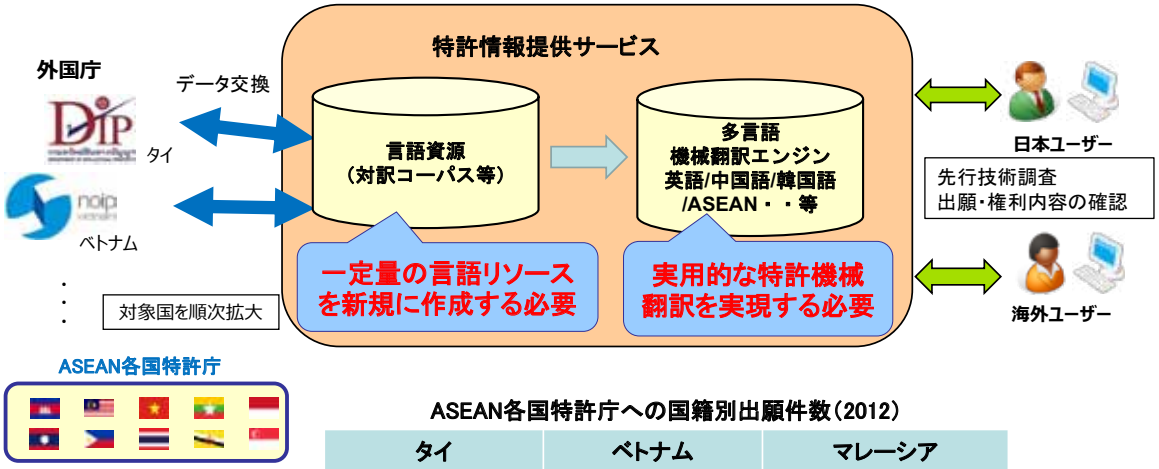
- 機械翻訳の更なる活用のためには翻訳精度の向上が必要。統計翻訳システムの導入も考えられるが、多大な計算機リソースが必要なため、コスト増になるなどの課題も。
- 審査書類（拒絶理由通知等）の翻訳精度が低いため、審査書類の翻訳精度向上に向けた取り組みを実施する予定。



機械翻訳における課題②

特許庁

➢ 近年日本からの出願が増加しているASEAN諸国等へも対応するため、**特許機械翻訳の多言語化**へ向けて研究を進める予定。



ASEAN各国特許庁への国籍別出願件数（2012）

タイ	ベトナム	マレーシア
日本　2595件	日本　1212件	米国　1695件
米国　1498件	米国　676件	日本　1248件
タイ　604件	ベトナム　382件	マレーシア　1114件

最後に

特許庁

特許庁は、機械翻訳の研究の進展を
大きく期待しています！！！

特許庁
JAPAN PATENT OFFICE

ご静聴ありがとうございました

招待講演 2

「ニューラルネットワークを用いた自然言語処理の最先端」

ニューラルネットワークを用いた自然言語処理の最先端

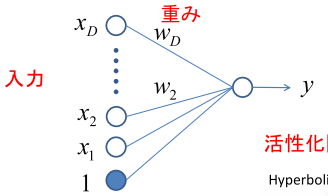
鶴岡慶雅（東京大学）

概要

- ニューラルネットワーク
  - 多層ニューラルネットワーク
  - リカレントニューラルネットワーク
  - 畳み込みニューラルネットワーク
- 自然言語処理
  - 機械翻訳、対話
  - 画像キャプション生成
  - 文分類
  - 質問応答
  - 構文解析

ニューラルネットワーク

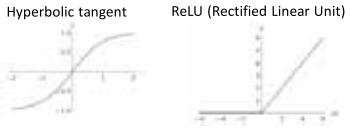
- ニューロン



入力の線形和に非線形な活性化関数を適用

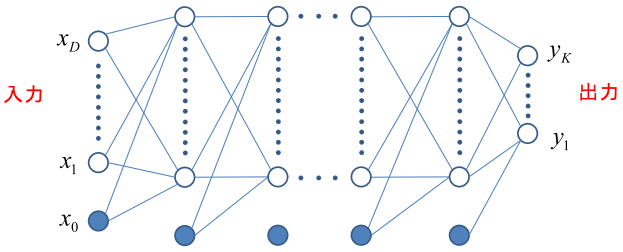
$$y = f\left(\sum_{i=0}^{D} w_i x_i\right)$$

活性化関数



多層ニューラルネットワーク

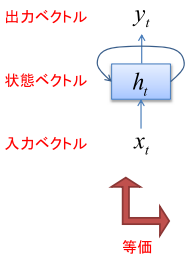
- 多数の入出力のペアから入出力関係を学習



入出力の次元は固定
→ 不定形な構造を持つ入出力は扱いにくい

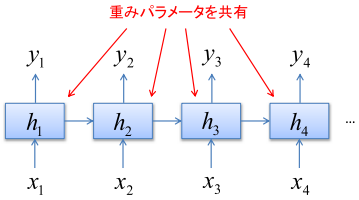
リカレントニューラルネットワーク（Recurrent Neural Network, RNN）

- 任意の長さの系列を扱うことができる



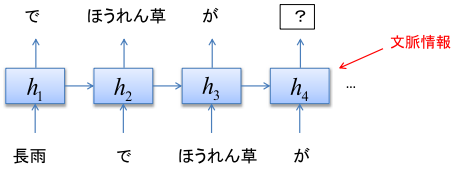
$$h_t = sigmoid\left(W^{hx} x_t + W^{hh} h_{t-1}\right)$$

$$y_t = W^{yh} h_t$$



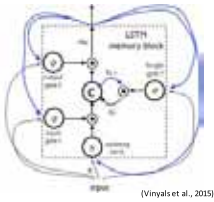
RNNと自然言語処理

- 自然言語処理では文字や単語の系列を扱う
  - 言語モデル 、品詞タグ付け、固有表現認識、機械翻訳 、etc.
- 例）言語モデル
  - 次の単語を予測



LSTM (Long Short-Term Memory)

- 単純な RNN の問題点
  - 勾配消失問題
  - 長距離の依存関係をとらえられない
- Long Short-Term Memory (LSTM)



$$i_t = \sigma\left(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}\right)$$
$$f_t = \sigma\left(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}\right)$$
$$o_t = \sigma\left(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}\right)$$
$$\tilde{c}_t = \tanh\left(W^{(\tilde{c})}x_t + U^{(\tilde{c})}h_{t-1} + b^{(\tilde{c})}\right)$$
$$c_t = i_t \circ \tilde{c}_t + f_t \circ c_{t-1}$$
$$h_t = o_t \circ \tanh(c_t)$$

(Vinyals et al., 2015)

ニューラル機械翻訳

- ある言語の文を他の言語に変換

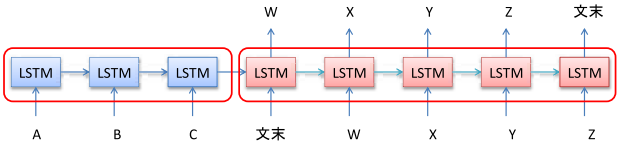
I'm here on vacation

Je suis là pour les vacances

- 多数の翻訳例から翻訳モデルを学習
  - 例. WMT'14 English-to-French データセット
    - 1200万文
    - 約3億語（英）
    - 約3億5千万語（仏）

ニューラル機械翻訳

- エンコーダー・デコーダーモデル (Sutskever et al., 2014)
  - Encoder RNN
    - 翻訳元の文を読み込み、実数値ベクトルに変換
  - Decoder RNN
    - 実数値ベクトルから翻訳先言語の文を生成



出力例

モデルの出力

Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .

正解の翻訳

Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .

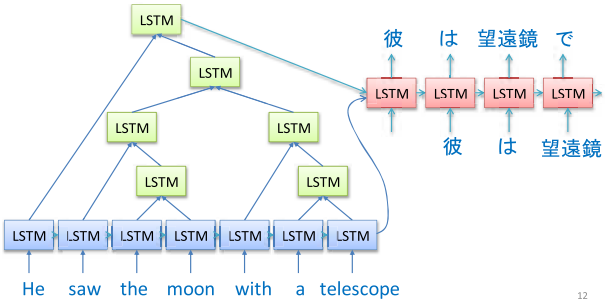
Sutskever et al., Sequence to Sequence Learning with Neural Networks, NIPS 2014

入力文のベクトル表現

Sutskever et al., Sequence to Sequence Learning with Neural Networks, NIPS 2014

Tree-to-sequence 機械翻訳

- 入力文の構文構造を利用（Eriguchi et al. 2016）



Tree-to-sequence 機械翻訳

- 学習データ
  - WAT'15 English-to-Japanese データセット
  - 135万文ペア
- 翻訳精度

	BLEU	RIBES
Tree-to-string statistical MT (Neubig, 2014)	36.6	79.6
Neural reranking (Neubig et al., 2015)	38.2	81.4
Sequence-to-sequence LSTM (Zhu, 2015)	36.2	80.9
Tree-to-sequence モデル	36.9	82.4

13

翻訳例

In information technology and electron field, the application of nanotechnology to next generation semiconductors, high-density information record technology, miniature integrated circuit elements, electric power saving displays using carbon nano-tube, etc. can be expected.

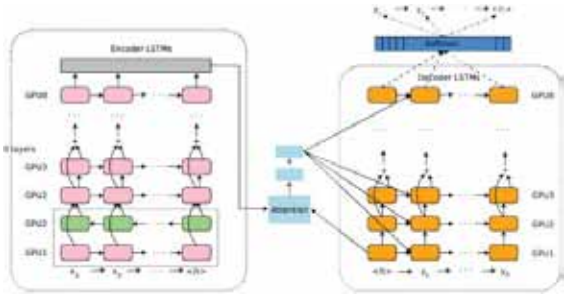
情報 技術 と 電子 分野 で は ， 次 世代 半 導体 へ の ナノテクノロジー の 応用 ， 高 密度 情報 記録 技術 ， 超 小型 集積 回路 要素 ， カーボン ナノ チューブ を 用い た 省 電力 表示 など が 期待 でき る 。

14

Google Neural Machine Translation system (GNMT) (Wu et al., 2016)

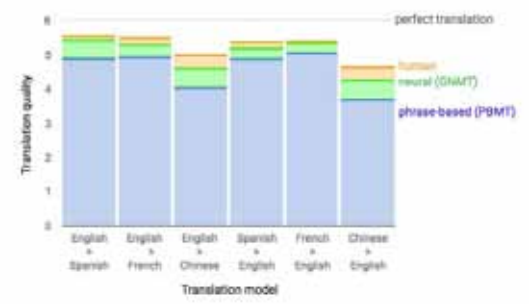
- モデル
  - エンコーダ： 8層LSTM（最下層はbidirectional）
  - デコーダ：8層LSTM
  - アテンション： デコーダの最下層からエンコーダの最上層
  - 翻訳単位：Wordpiece モデル
- 学習データ
  - 評価用
    - WMTコーパス： 英仏（3600万文）、英独（500万文）、etc
  - Production用
    - WMTコーパスの百倍から千倍

GNMT system



（Wu et al., 2016）

GNMTの翻訳性能



https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

ニューラル会話モデル (Vinyals & Le, 2015)

- エンコーダー・デコーダーモデルで会話文を学習
  - 学習データ
    - 映画の中での会話　6200万文
- 学習したモデルと人間との会話の例

Human: who is skywalker ?
Machine: he is a hero .
Human: who is bill clinton ?
Machine: he 's a billionaire .
Human: is sky blue or black ?
Machine: blue .
Human: does a cat have a tail ?
Machine: yes .

Human: does a cat have a wing ?
Machine: no
Human: can a cat fly ?
Machine: no .
Human: how many legs does a cat have ?
Machine: four , i think .
Human: how many legs does a spider have ?
Machine: three , i think .

18

質問応答（QA）

文書

Mary got the football there.
John moved to the bedroom.
Sandra went back to the kitchen.
Mary travelled to the hallway.
John got the football there.
John went to the hallway.
John put down the football.
Mary went to the garden.

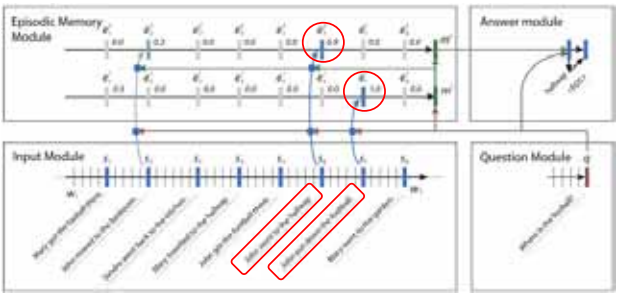



質問

Where is the football?

Dynamic Memory Networks (Kumar et al., 2016)

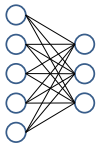
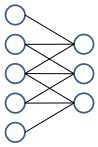
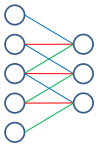
- 答えを導出するために必要な文を順次推定



20

畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）

- 全結合
- 局所的結合
- パラメータ共有

パラメータ数
5 x 3 = 15

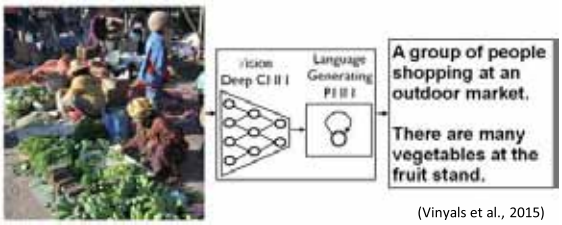
パラメータ数
3 x 3 = 9

パラメータ数
3

パラメータ数を減らすことにより過学習を回避
画像認識、テキスト分類などに有効

21

画像の説明文の生成

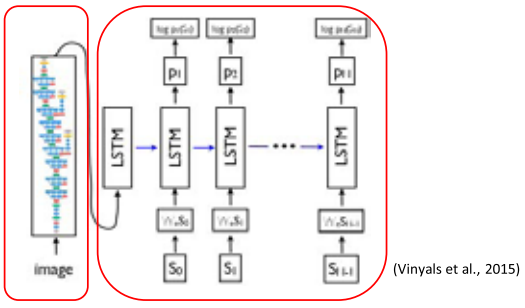


(Vinyals et al., 2015)

1. 大量のラベル付き画像で画像認識CNNを学習
2. 説明文付きの画像で言語生成RNNを学習

22

画像の説明文の生成



(Vinyals et al., 2015)

CNN　　RNN

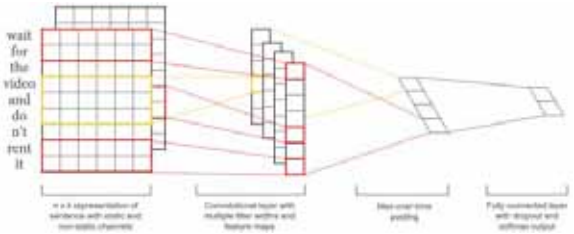
23

説明文生成例 (Vinyals et al., 2015)

24

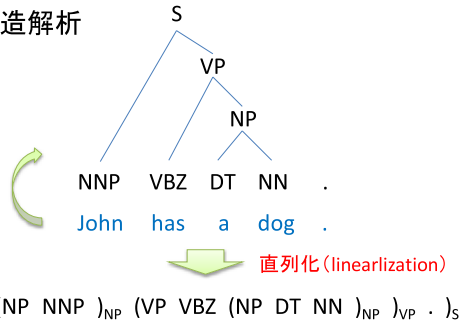
CNN による文分類

- Kim (2014)



構文解析

- 句構造解析

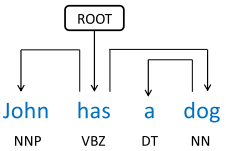


エンコーダ・デコーダモデルによる句構造解析（Vinyals et al., 2015）

- モデル
  - エンコーダ・デコーダ：3層LSTM＋アテンション
- 学習データ
  - 複数のツリーバンク
  - Webテキスト
    - Berkeley Parser と ZPar の出力が一致したもの
- 性能
  - 精度：92.1 (f-score) on WSJ23
  - 速度：100文/秒 以上

構文解析

- 依存構造解析
  - 単語間の係り受け



SyntaxNet (Andor et al., 2016)

- 遷移型モデル
  - アクションの系列で構文構造を決定
  - Feedforward NN でアクション選択
  - ビーム探索と early update による全体最適化
- 精度
  - Unlabeled attachment score (UAS): 94.61%
  - Labeled attachment score (LAS): 92.79%

まとめ

- ニューラルネットワーク
  - リカレントニューラルネットワーク
  - 畳み込みニューラルネットワーク
- エンコーダー・デコーダーモデル
  - 機械翻訳、対話システム、構文解析、etc
- 複雑なタスクを簡単なアーキテクチャで実現
  - End-to-end 学習

研究会報告 1

「翻訳自動評価法～翻訳の質を推定する技術の進化」

翻訳自動評価法
翻訳の質を推定する技術の進化

磯崎　秀樹
岡山県立大学

2016 年 11 月 25 日 第 4 回特許情報シンポジウム

翻訳ソフトの改良と人手評価 1

翻訳ソフトの作成では、出力を見て改善する作業をくりかえす。

しかし、よかれと思ってやった変更に副作用があり、全体的にはむしろ悪くなっていることがある。

そこで、新しい訳が古い訳よりも本当にいいか確認する必要がある。

色んな文を訳して、その訳を人間が見て採点すればよい。

これを「**人手評価**」といい、以下の 2 つの評価尺度が有名。

- **妥当性 (adequacy)**: 訳が原文にどれくらい忠実かの評価。
- **流暢性 (fluency)**: 訳がどれくらい流暢かの評価。

(これらは信頼性が低く、近年の人手評価では好まれない。)

自動評価の必要性 2

翻訳ソフトの出力する何千という文を、人手で評価するのは大変。
人件費も時間もかかる。

そこで、「**自動評価**」が考案された。

あらかじめ、人間が理想的な訳「**参照訳**」を作っておく。

翻訳ソフトの出力した訳「**機械訳**」と参照訳の**類似度を計算**。

世界で標準的に用いられている類似度は、IBM が提案した $\mathbf{BLUE}$ (BiLingual Evaluation Understudy))。

初期の自動評価法 3

BLEU より前には、音声認識分野で使われいる $\mathbf{WER}$ (Word Error Rate) が使われていた。

WER は、機械訳を書き換えて参照訳にするのに必要な、
単語の追加・削除・置換の操作の回数に基づく尺度。

機械訳が参照訳と同じなら WER $= 0.0$ で、
違うほど WER が大きいので、$1 -$ WER を類似度とする。

しかし、語順の近い欧米言語間では、逐語訳でもかなりわかり、
WER は語順の間違いに厳しすぎると批判された。

そこで、語順の間違いを大目に見る尺度が求められた。

PER と TER 4

これらは、WER をベースにして、語順の違いに甘くしたもの。

- **PER** (Position-independent word Error Rate):
  文を単語の集合とみなすことで、語順を完全に無視。
- **TER** (Translation Edit Rate): (Sover et al. 2006)
  複数語からなるフレーズ一度に動かすのを 1 操作とみなし、
  語順の違いを大目に見る。

磯崎が語順に関する論文を国際会議に投稿したとき、
「語順など重要な問題ではない」と書いた査読者がいる。

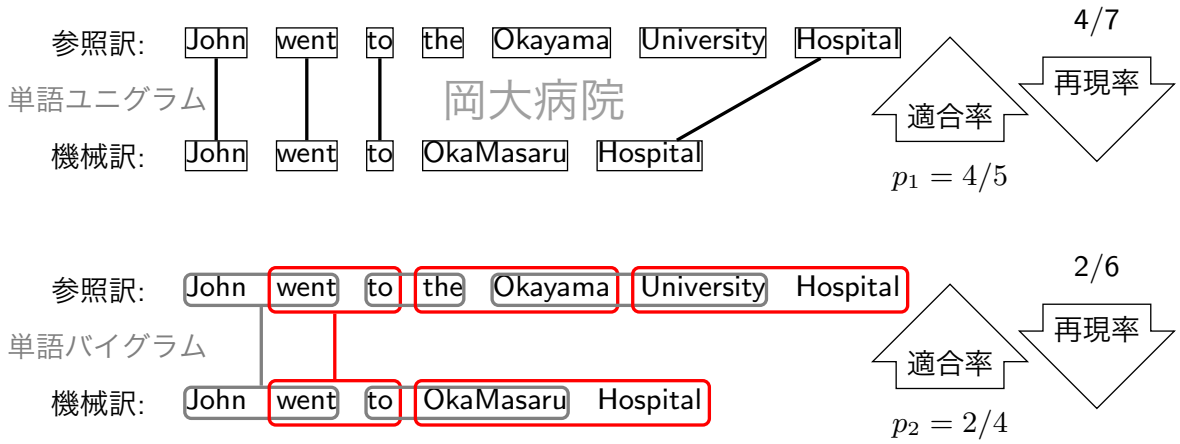
英語の語順で苦労している日本人には信じがたいが、
これが欧米の一部の研究者の認識らしい。

BLEU (単語 n グラム適合率) 5

Papineni et al. 2002

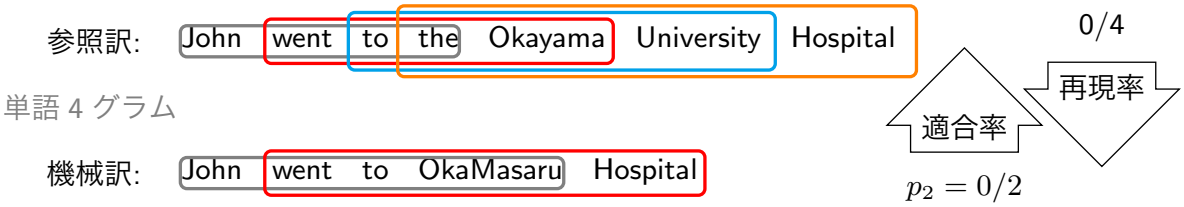
機械訳と参照訳の間で、共通している単語やフレーズが多いほど、
よい訳である、という考え方にもとづく。

単語 n グラムの適合率を p_n で表し、p_1 から p_4 の相乗平均
$\sqrt[4]{p_1 p_2 p_3 p_4}$ をベースにしている。



BLEU (複数参照訳) 6

p_4 は、共通している単語 4 グラムが存在しないと 0 点なので、
BLEU $= \sqrt[4]{p_1 p_2 p_3 p_4}$ も 0 点になってしまう。



参照訳が一つだけだと、BLEU が 0 点の文が多くなる。

そこで BLEU を使うときは、なるべく 0 点にならないように、
参照訳をいくつも用意しなければならない。

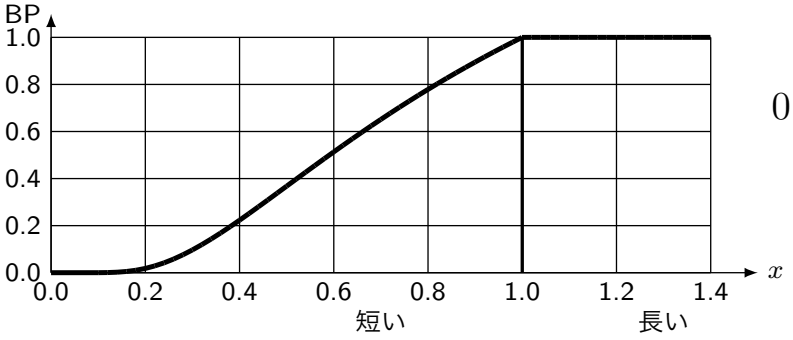
BLEU (Brevity Penalty) 7

複数参照訳のせいで、BLEU は再現率を使ってないので、
自信のない部分を出力しなければ適合率が上がる。

これを防ぐため、BLEU では、短い機械訳にペナルティを与える。

文の長さの比「機械訳の長さ/参照訳の長さ」を x とすると、以下の
BP (Brevity Penalty) を掛けることで、短すぎる訳の点数を下げる。

$$\mathrm{BP}(x) \stackrel{\mathrm{def}}{=} \min\left(1, \exp\left(1 - 1/x\right)\right), \qquad \mathrm{BLEU} = \mathrm{BP} \times \sqrt[4]{p_1 p_2 p_3 p_4}$$



$$0.0 \leq \mathrm{BLEU} \leq 1.0$$

BLEU (人手評価との相関の低さ) 8

NTCIR 特許翻訳タスクの実験結果によると、
英日翻訳や日英翻訳では、BLEU と人手評価の相関が低い。
NTCIR-7 の日英翻訳タスクの場合、
BLEU と人手評価の順位相関 (Spearman's ρ) は 0.5 程度しかない。

BLEU が大局的な語順を考慮していないのが原因。

BLEU は、因果関係が逆の訳などに高い点数を与えることがある。

原文	彼は雨に濡れたので、風邪を引いた。	妥当性	BLEU
参照訳	He caught a cold because he got soaked in the rain.	○	1.00
機械訳1	He caught a cold because he had gotten wet in the rain.	○	**0.53**
機械訳2	He got soaked in the rain because he caught a cold.	×	**0.74**

RIBES(人手評価との相関の高さ) 9

Isozaki et al. 2010, 平尾ら 2011

そこで磯崎は、大局的な語順を考慮した **RIBES** を提案。
RIBES は人手評価と相関が高い。

NTCIR-7 日英翻訳における人手評価 (妥当性と流暢性の平均) と
自動評価の順位相関 (Spearman's ρ)

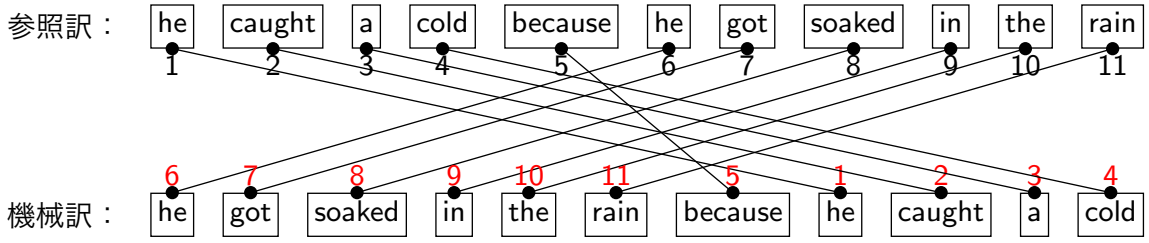
自動評価法	妥当性	流暢性
RIBES	**0.947**	**0.879**
BLEU	0.515	0.500

原文	彼は雨に濡れたので、風邪を引いた。	妥当性	BLEU	RIBEU
参照訳	He caught a cold because he got soaked in the rain.	○	1.00	1.00
機械訳1	He caught a cold because he had gotten wet in the rain.	○	0.53	**0.93**
機械訳2	He got soaked in the rain because he caught a cold.	×	0.74	**0.38**

RIBES(大局的語順の考慮) 10

RIBES は語順の近さをベースとした類似度。

語順の近さは、Kendall's τ という順位相関係数で測定。

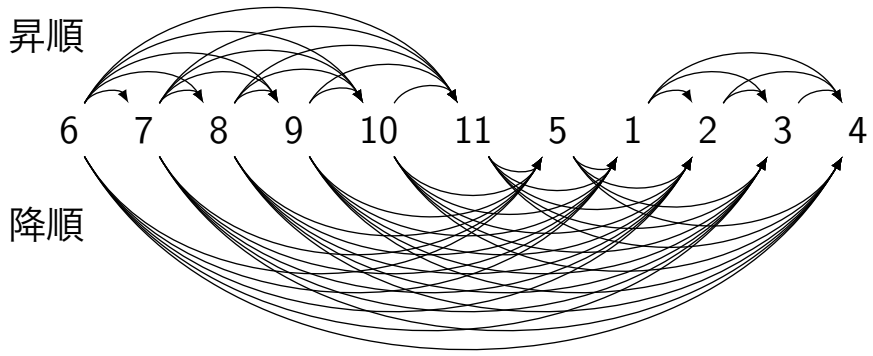


この機械訳の語順は、[6,7,8,9,10,11,5,1,2,3,4] という整数のリストで表せる。

RIBES(NKT) 11

整数リストから、整数を 2 つ取り出してできるペアのうち、昇順ペアの割合を NKT (Normalized Kendall's τ) と呼ぶ。

この場合、要素が 11 個あるので、${}_{11}C_2 = 55$ ペアあるはず。



昇順なのは、6〜11 の部分の ${}_6C_2 = 15$ ペアと、1〜4 の部分の ${}_4C_2 = 6$ ペアの合計 21 ペアなので、NKT $= 21/55 = 0.38$。

NKT を日英翻訳の自動評価に使ってみたところ、人手評価と高い相関があることが判明。

RIBES(ペナルティー) 12

平尾ら 2014

しかし、NKT には、共通する単語が少ないとき、NKT は過大 (過少) 評価する、という弱点がある。

そこで、単語適合率 p_1 の α 乗をペナルティとして掛ける。

しかし、適合率は自信のあるところだけを出すという方法で上げられる。そこで、BLEU の BP の β 乗を掛ける。

以上により、RIBES は以下の式により定義される。

$$\text{RIBES} \stackrel{\text{def}}{=} \text{NKT} \times P^{\alpha} \times \text{BP}^{\beta}$$

RIBES(スクランブリング) 13

Isozaki and Kouchi 2015

RIBES は語順の類似度によって訳を評価する。

しかし、日本語の語順は、比較的自由と言われている。
たとえば、以下の 2 つの文は同じ意味であり、どちらでもよい。

❶ 太郎は水族館でイルカを見た。
❷ 水族館で太郎はイルカを見た。

しかし、次の語順は意味が変わってしまうので、許容できない。

❸ イルカは水族館で太郎を見た。

RIBES は語順を重視して評価するので、❶を参照訳、❷を機械訳として採点すると、悪い点になり、人手評価とずれる。

これは係り受け解析で解決できそうだが、機械訳が係り受け解析できるほど質が高いとは限らない。

Isozaki and Kouchi 2015 は、参照訳の係り受け解析だけで、
この問題の解決する方法を提案した。

欧米の最近の翻訳自動評価 14

ここまで、英日・日英翻訳の翻訳自動評価の話をしてきた。

欧米にも BLEU に対する不満の声はあり、国際会議

WMT $\left\|\begin{array}{l}\text{Workshop on Statistical Machine Translation}\\ \longrightarrow \text{Conference on Machine Translation}\end{array}\right\|$

などで、新しい翻訳自動評価の手法が活発に研究されている。

RIBES と同じ 2010 年に、イギリスで LRscore という Kendall's τ をベースにした自動評価法が独立に提案されている。
しかし、語順の近い中国語と英語の翻訳結果で実験しているので、RIBES ほど大きな差は出ていない。

2010〜2012 年の欧米での翻訳自動評価法については、
以下の資料にまとめてあるので参照されたい。

磯崎：最近の自動評価法の研究動向と RIBES、AAMT/Japio 特許翻訳研究会、特許文書の機械翻訳結果評価方法検討会資料集, 2012.

まとめと今後の課題 15

- 欧米で標準的に用いられている BLEU という翻訳自動評価法は、日英・英日翻訳では人手評価と相関が低い。
- そこで、語順の近さに注目した自動評価法 RIBES を提案した。RIBES は人手評価と相関が高い。
- 参照訳の係り受け木を用いて参照訳を増やすことで日本語のスクランブリングに対応させた。

RIBES は日英・英日翻訳にかかわる多くの翻訳研究者に利用されているが、以下の問題点が指摘されている。

- RIBES を目的関数としてチューニングすることが難しい。
- ニューラルネットを用いた NMT (Neural Machine Translation) は、間違いの傾向が異なり、NMT に合わせた改良が必要。

研究会報告 2

「第 3 回アジア翻訳ワークショップの人手評価結果の分析」

第3回アジア翻訳ワークショップの人手評価結果の分析

中澤 敏明
科学技術振興機構（JST）

2016年11月25日 第4回特許情報シンポジウム

アジア翻訳ワークショップ (WAT)

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

- Workshop on Asian Translation
  - アジア言語を対象とした機械翻訳評価ワークショップ
  - 日本語、中国語、韓国語、インドネシア語、ヒンディー語、英語
  - 2014年より開催、今年で3回目
- WAT2016
  - 12月12日に大阪で開催（Coling2016併設）
  - 今回から研究論文も募集
  - 招待講演はGoogleの賀沢氏

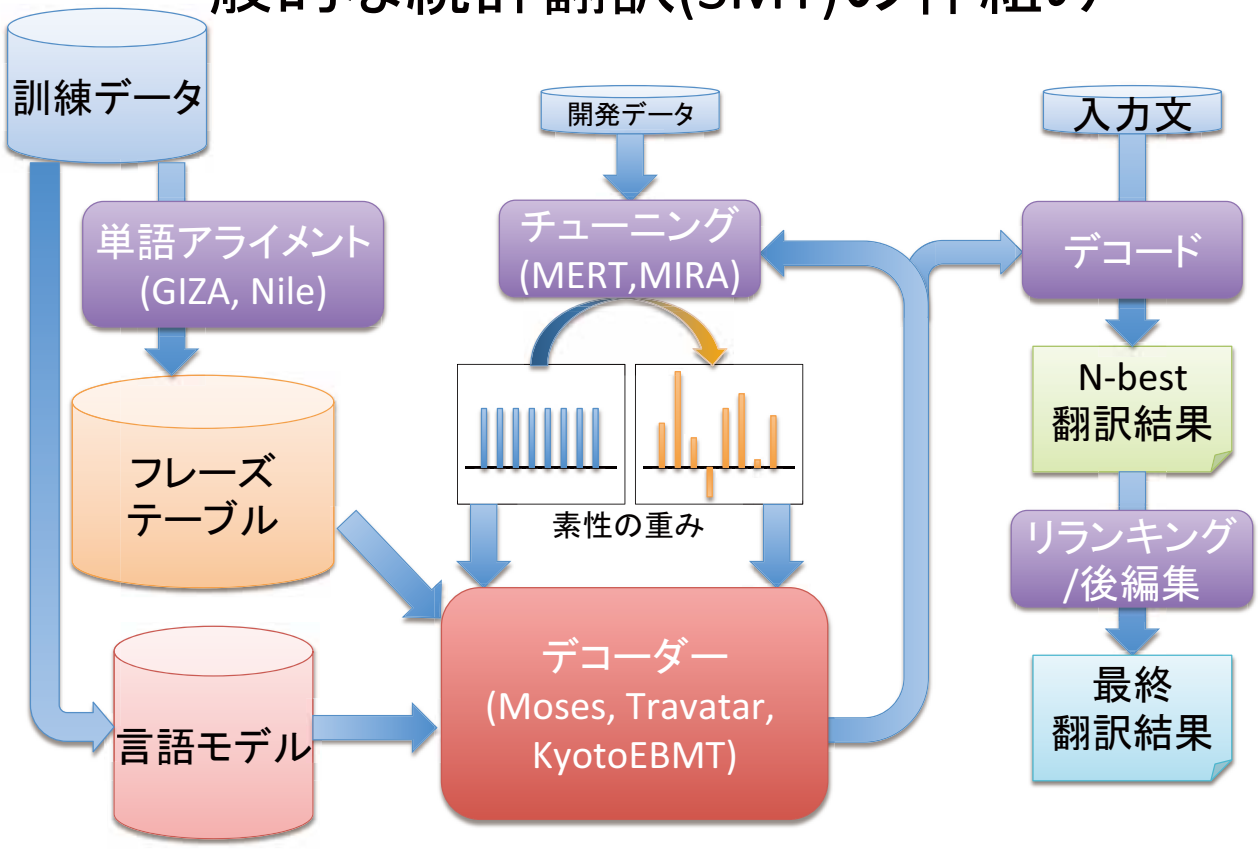
WAT2016翻訳タスク

- 科学技術論文（ASPEC）
  - 日⇔英、日⇔中
- 特許（JPC）
  - 日⇔英、日⇔中、日⇔韓
- 新聞記事（BPPT）
  - インドネシア⇔英
- 混合ドメイン
  - ヒンディー⇔英、ヒンディー⇔日

翻訳タスク参加チーム一覧

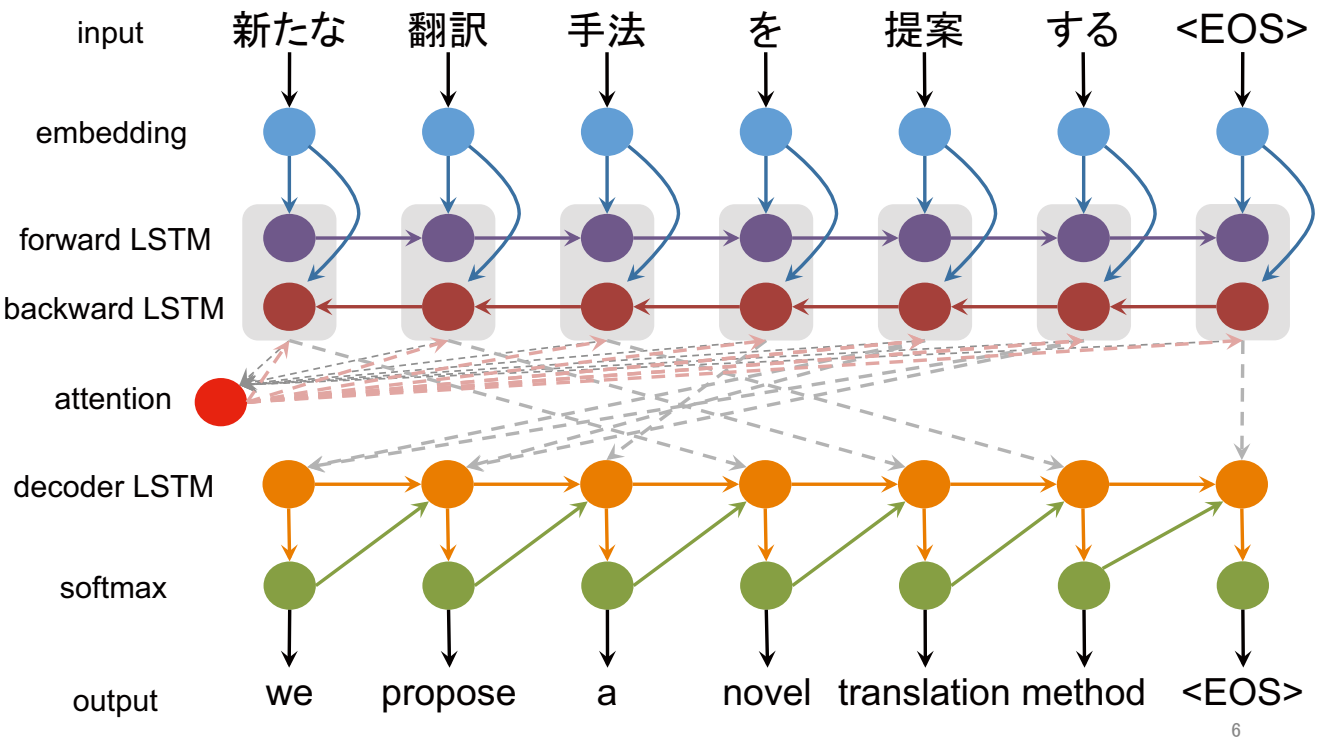
NMT Team ID	Organization (company / outside Japan)	ASPEC JE	ASPEC EJ	ASPEC JC	ASPEC CJ	JPC JE	JPC EJ	JPC JC	JPC CJ	JPC JK	JPC KJ	BPPT EI	BPPT IE	IITBC HE	IITBC EH	pivot HJ	pivot JH
NAIST	Nara Institute of Science and Technology	✓															
Kyoto-U	Kyoto University	✓	✓	✓	✓												
TMU	Tokyo Metropolitan University	✓															
bjtu_nlp	Beijing Jiaotong University	✓	✓	✓	✓	✓	✓	✓	✓								
Sense	Saarland University											✓	✓				
NICT-2	National Institute of Information and Communication Technology	✓	✓	✓	✓	✓	✓	✓	✓								
WASUIPS	Waseda University								✓								
EHR	Ehara NLP Research Laboratory		✓		✓				✓		✓				✓	✓	
ntt	NTT Communication Science Laboratories								✓								
TOKYOMT	Weblio, Inc.		✓														
IITB-EN-ID	Indian Institute of Technology Bombay											✓	✓				
JAPIO	Japan Patent Information Organization		✓		✓		✓		✓		✓						
IITP-MT	Indian Institute of Technology Patna														✓		
UT-KAY	University of Tokyo				✓												
UT-AKY	University of Tokyo		✓														
	# of participants	5	7	3	6	2	3	2	6	0	2	2	2	0	2	1	0

一般的な統計翻訳(SMT)の枠組み



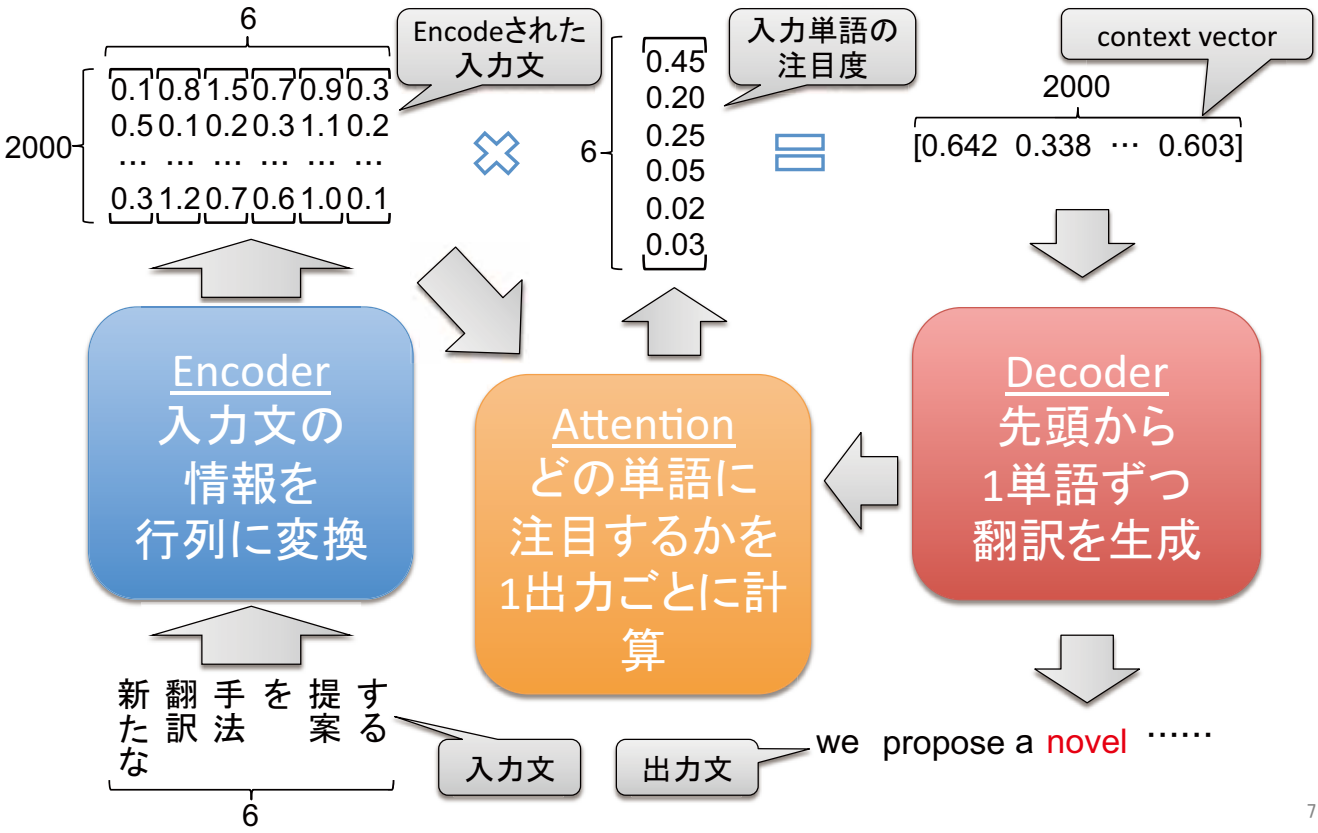
5

Attention-based Neural Machine Translation

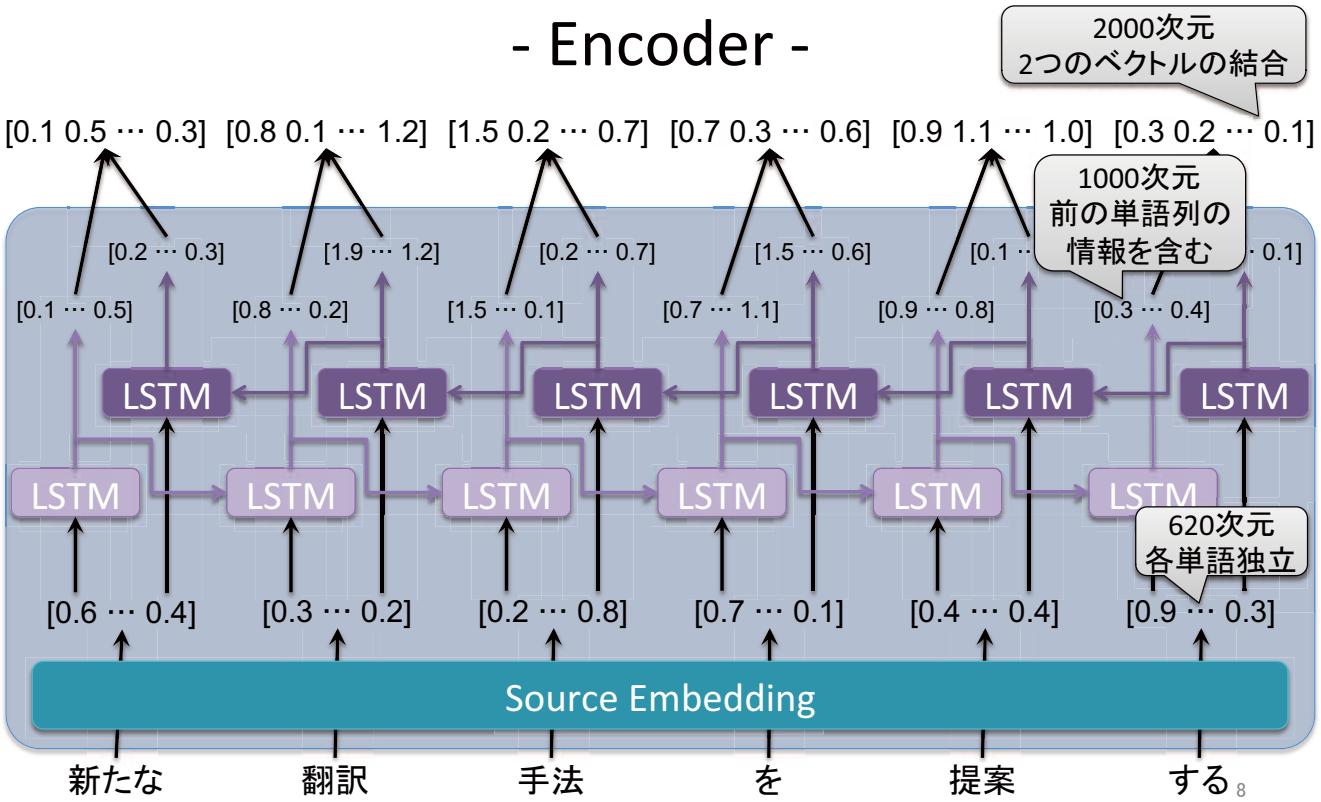


6

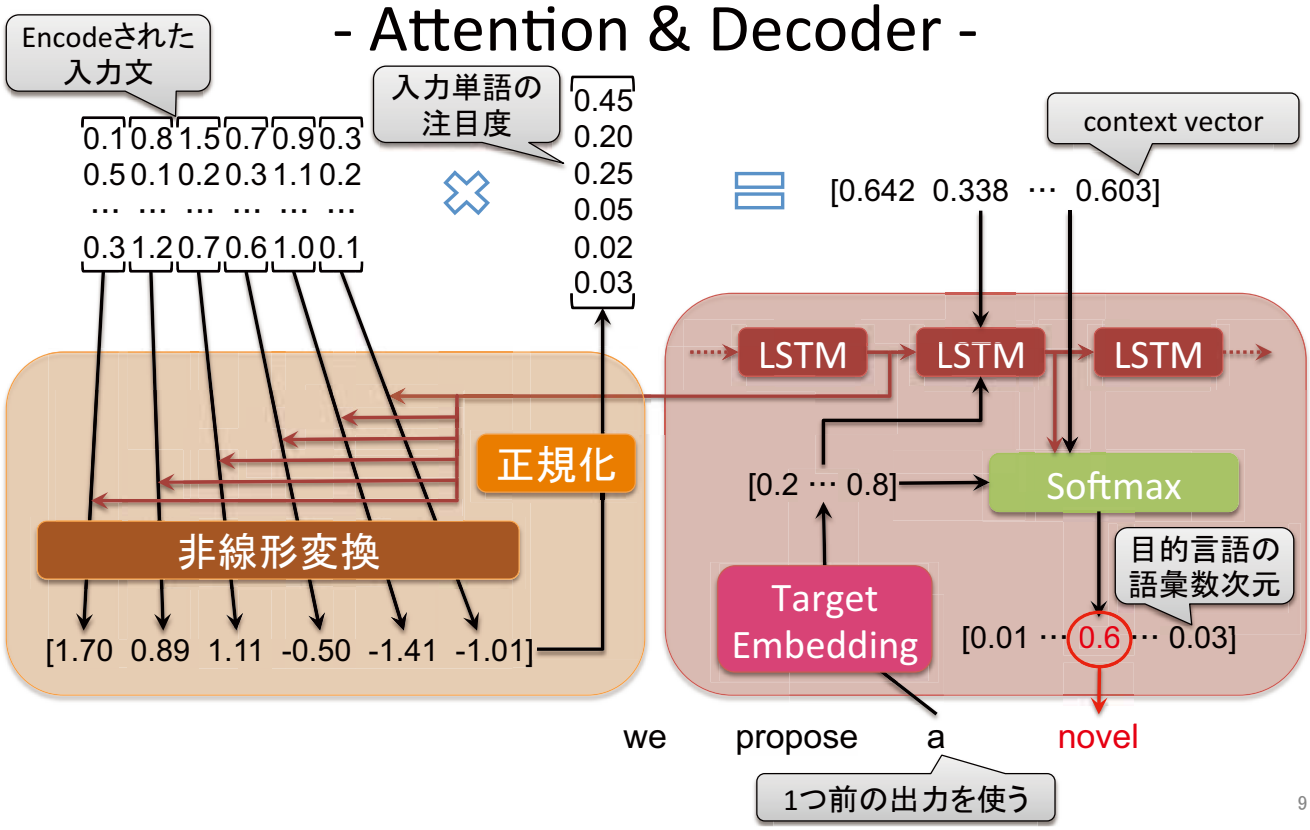
Attention-based Neural Machine Translation



Attention-based Neural Machine Translation - Encoder -



Attention-based Neural Machine Translation - Attention & Decoder -



9

SMTとNMTの違い

- NMTではフレーズテーブルという概念は存在しない
  - 単語アライメントは必要ない
- NMTは、SMTのように入力文を単語（フレーズ）単位で「置き換える」ことで翻訳しているわけではなく、入力文を見ながら翻訳文を作り出している
  - 入力文を過不足なくカバーして翻訳することができない
- NMTでは入口（入力文）と出口（翻訳文）以外は全て数値計算（行列の積など）だけで動いている
- 現状のNMTの枠組みでは、指定部分を指定訳にするなどの前処理・後処理ができない

10

SMTとNMTの違い

- モデルのトレーニングに時間がかかる
  - 100万文の対訳コーパスで3日ぐらい
  - GPUが必須、CPUでは10倍以上時間がかかる
- 使う対訳コーパスによって最高のパフォーマンスが出る設定が全然ちがう
  - この設定を自動で決められるようにはなっていない
  - どの設定がいいかは経験や直感に頼る部分が大きい
  - 初期パラメータをランダムに設定するため、同じコーパス、同じ設定でトレーニングしても、最終的な精度が全然違うことがある

翻訳結果評価方法

- 自動評価
  - BLEU, RIBES, AM-FM [Banchs+, 2015]
  - 自動評価サーバーを用意し、参加者が翻訳結果を提出すると自動的に評価され、都度Webに結果が公開される
- 人手評価: 二段階評価
  1. 一対比較評価
     - 各チーム各言語対2つまで翻訳結果を提出可能
  2. 特許庁が公開している特許文献機械翻訳の品質評価手順のうち、「内容の伝達レベルの評価」
     - 一対比較評価の結果、各言語対上位3システムを対象に評価

一対比較評価

- 各システムの翻訳結果をベースライン（フレーズベースSMT）と文ごとに比較
- 評価対象文はテストセットのうちの400文
- 5人の評価者が文ごとにベースラインより良いか（+1）、悪いか（-1）、同程度か（0）を評価
- 5人の評価を足し合わせて、+2以上ならばその文はベースラインより良いと、-2以下ならば悪いと、それ以外ならば同程度と判定
- 最終的にスコアを以下の式で計算
  - Pairwise Score = 100 × (勝ち数 - 負け数) / 400

13

内容の伝達レベルの評価

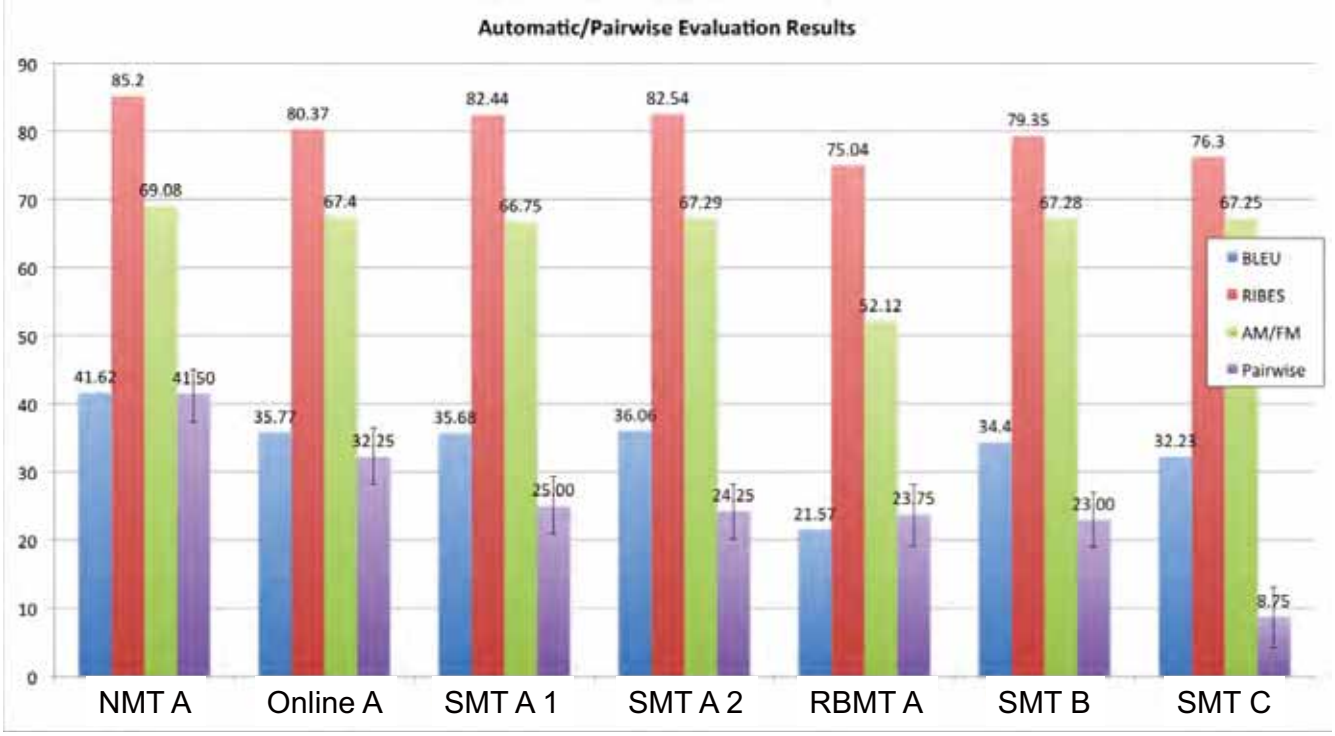
- 一対評価対象文のうちの200文を評価
- 2人の評価者が文ごとに特許庁の基準により評価

評価値	評価基準
5	すべての重要情報が正確に伝達されている。 (100%)
4	ほとんどの重要情報は正確に伝達されている。 (80%～)
3	半分以上の重要情報は正確に伝達されている。 (50%～)
2	いくつかの重要情報は正確に伝達されている。 (20%～)
1	文意がわからない、もしくは正確に伝達されている重要情報はほとんどない。 (～20%)

https://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

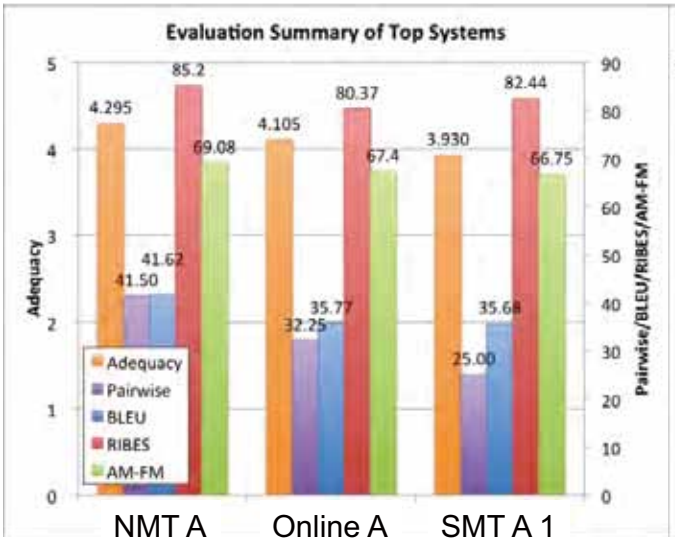
14

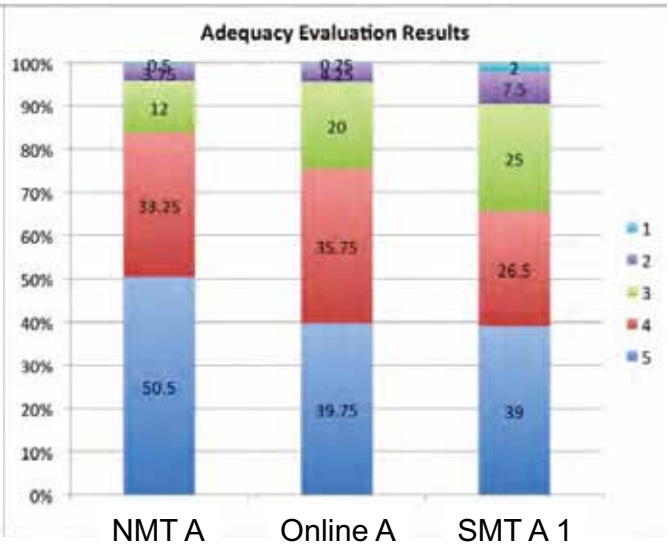
日→英 評価結果



15

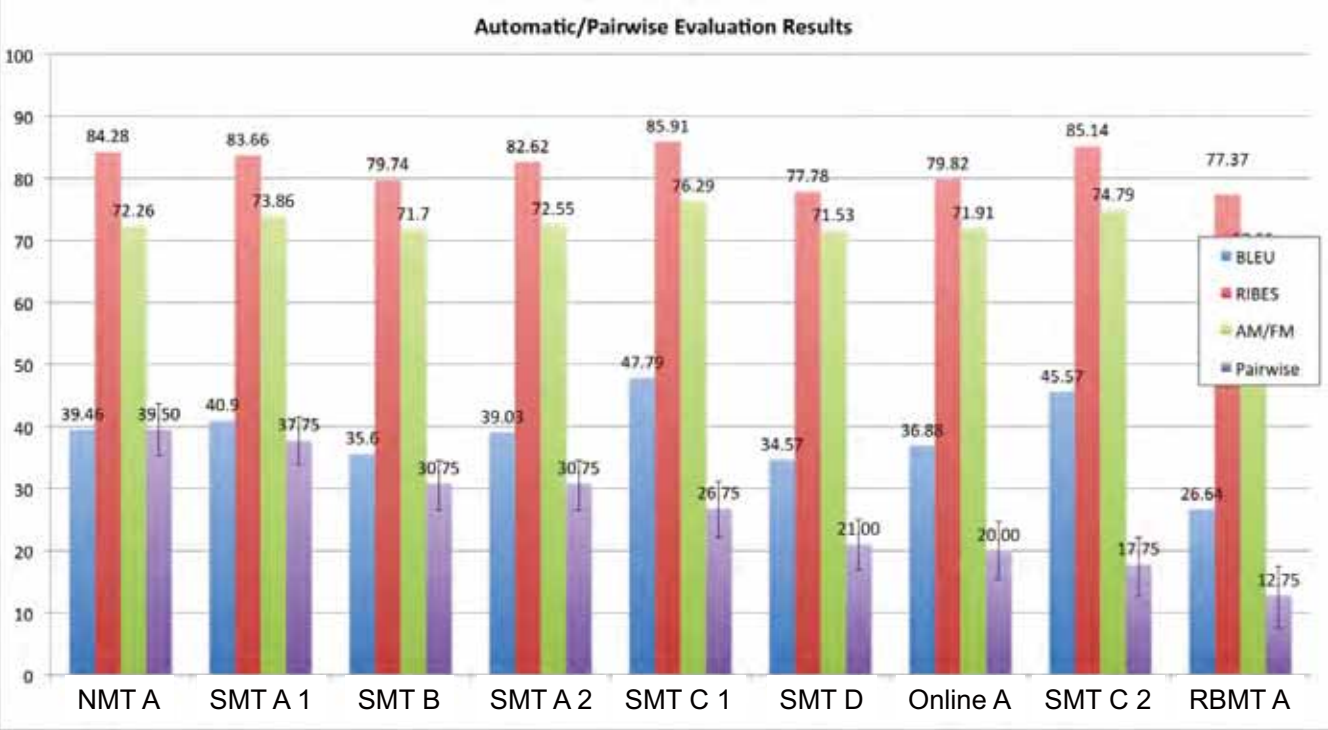
日→英 評価結果





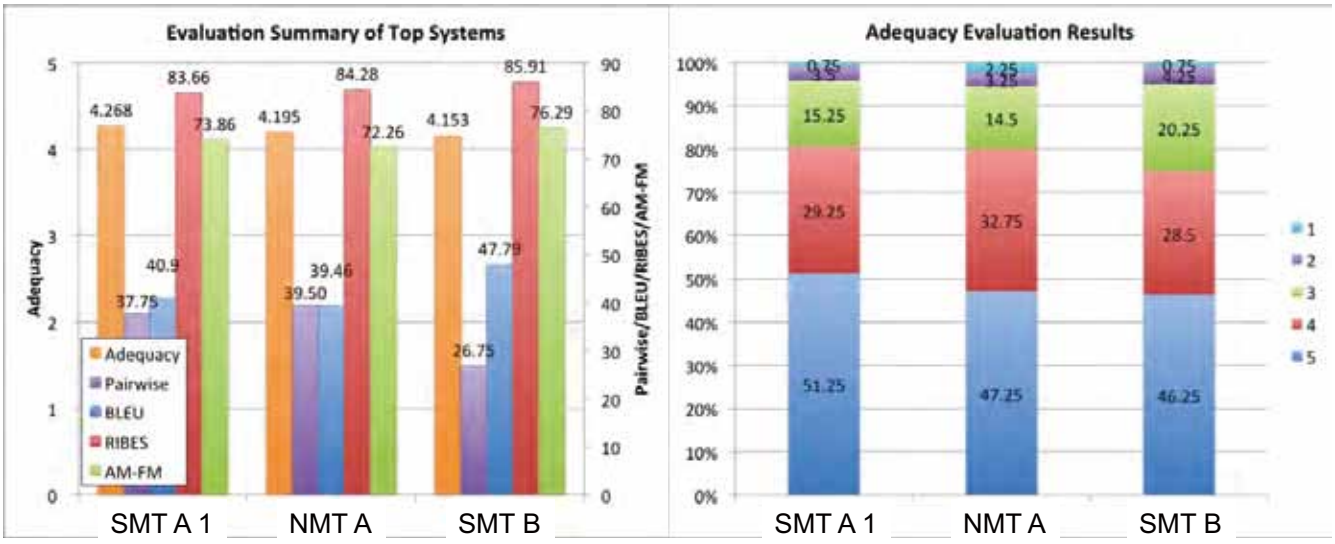
16

英→日 評価結果



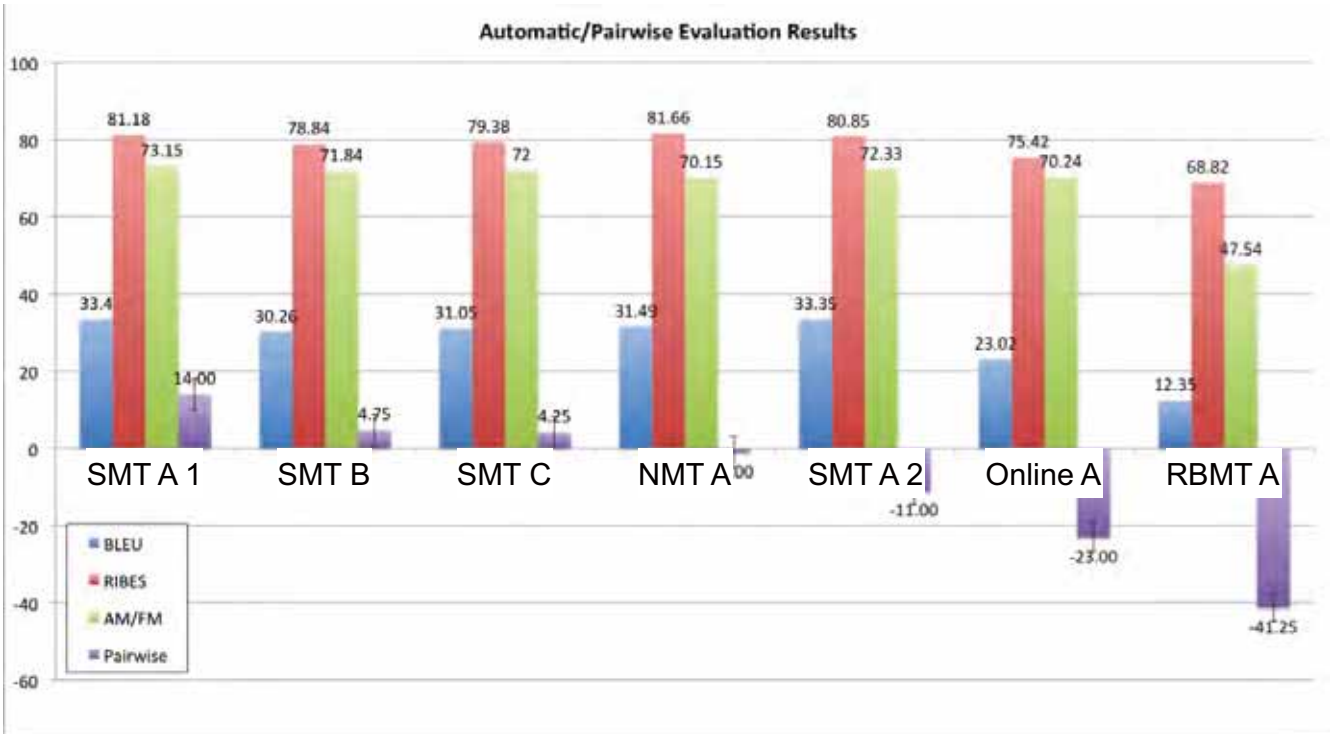
17

英→日 評価結果



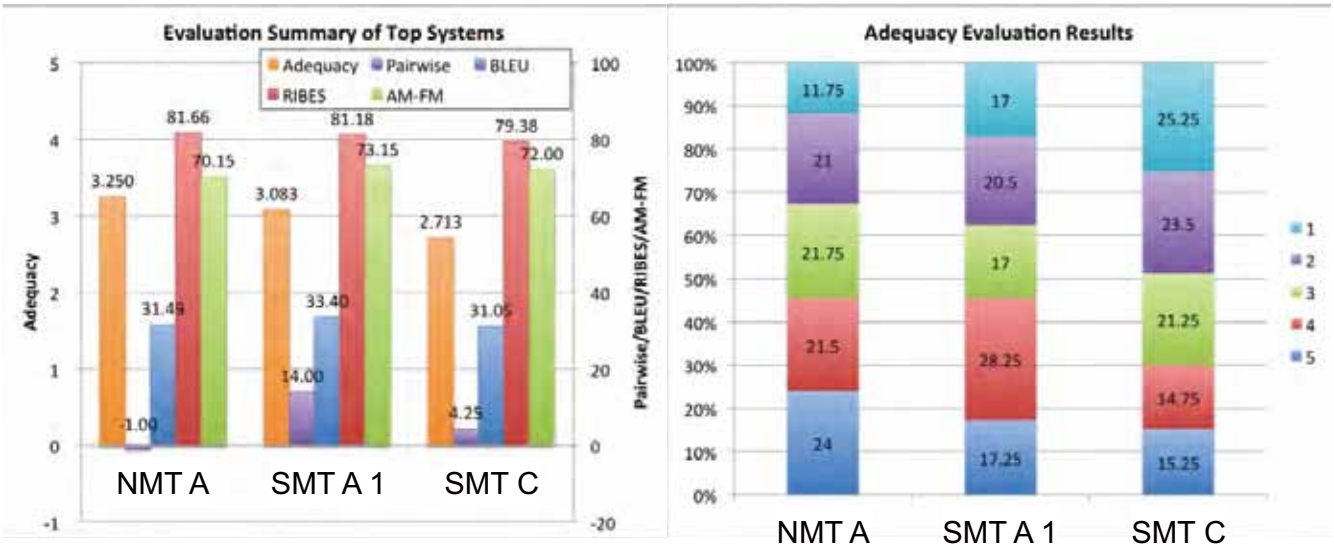
18

日→中 評価結果



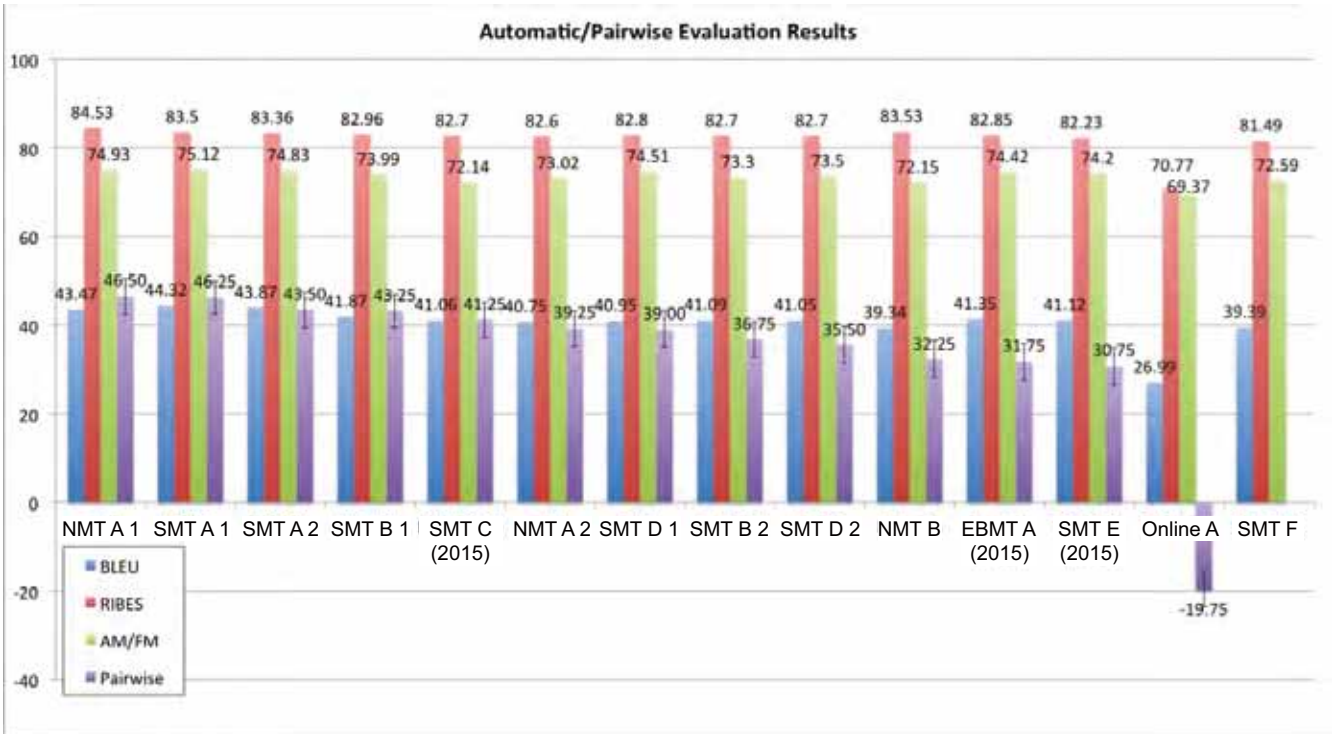
19

日→中 評価結果

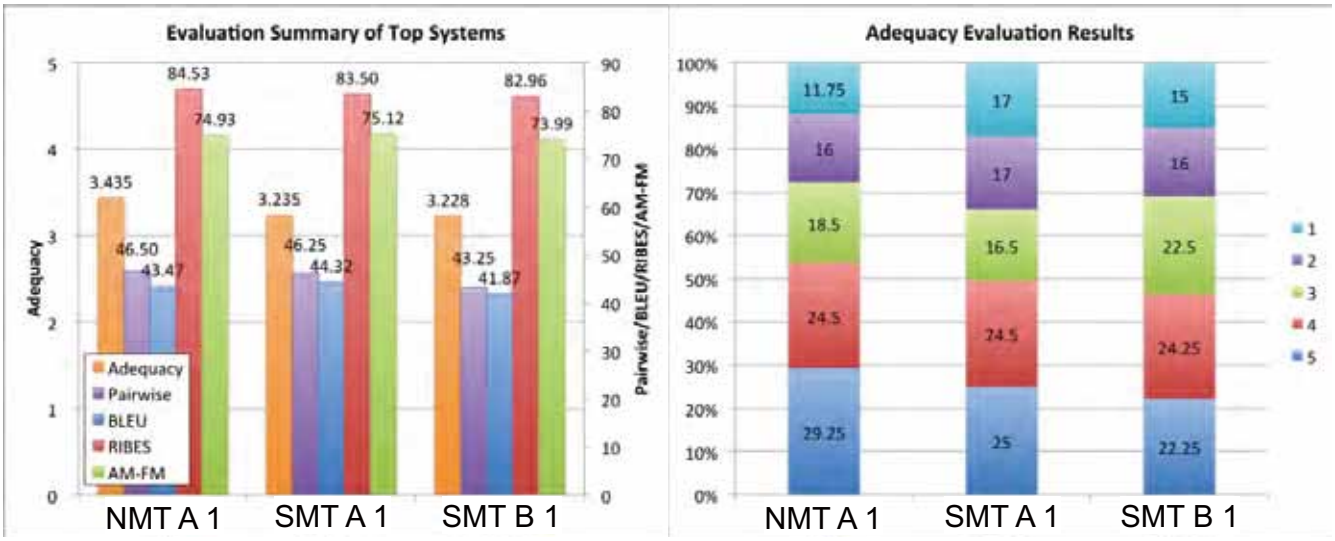


20

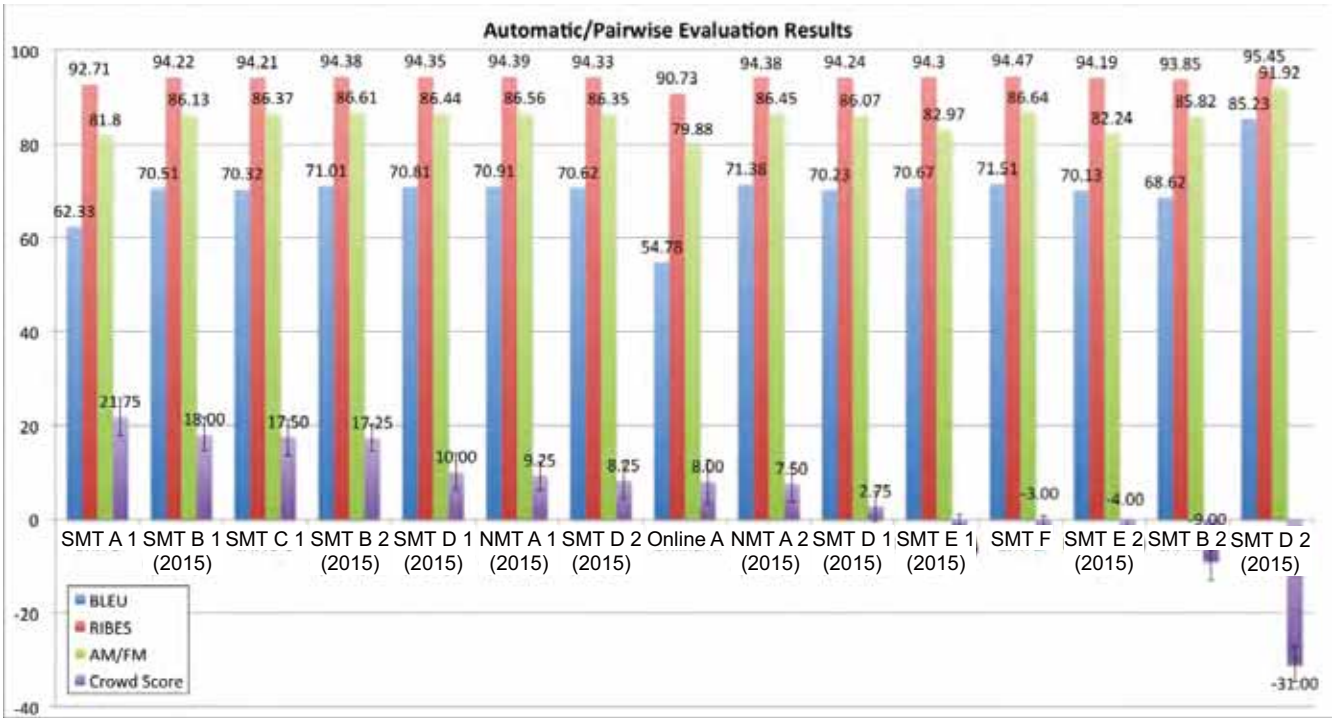
中→日 評価結果



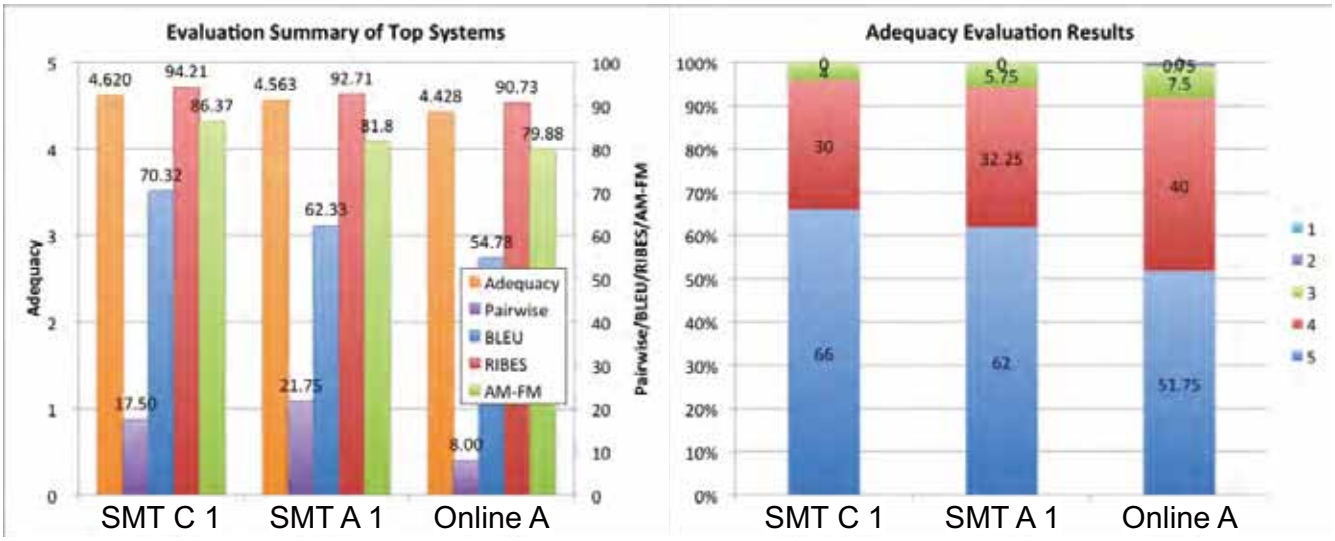
中→日 評価結果



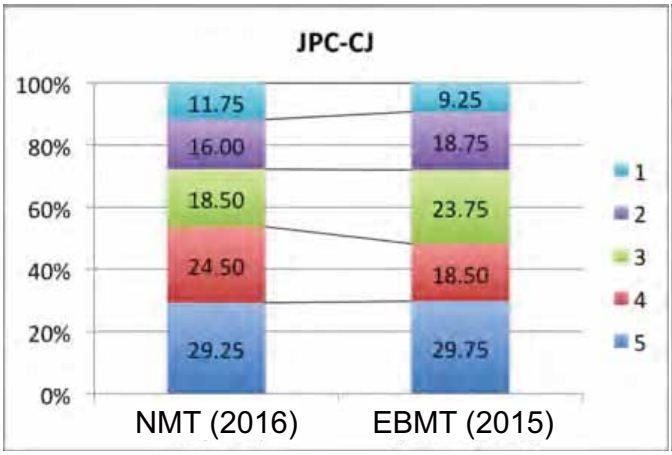
韓→日 評価結果



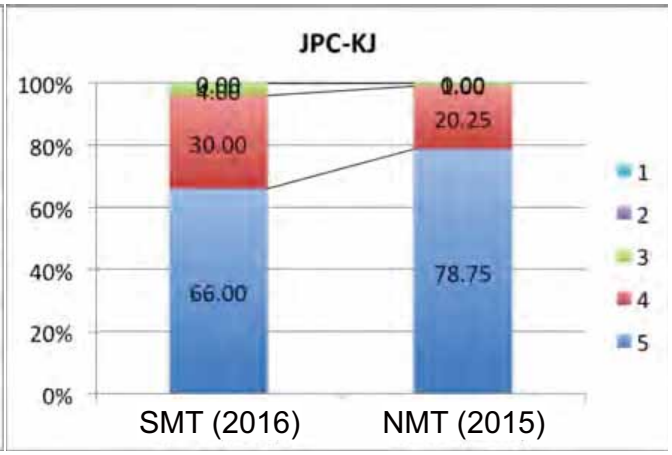
韓→日 評価結果



特許翻訳 経年評価



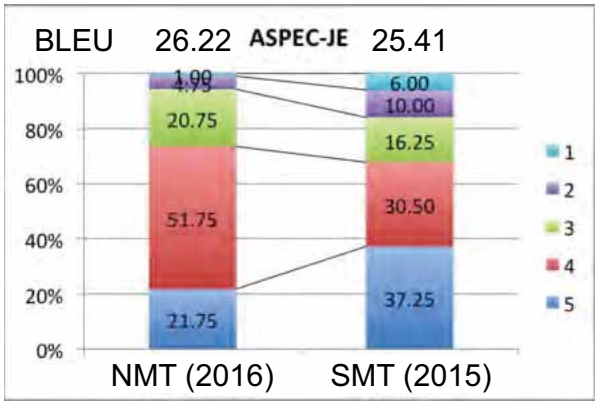
BLEU 44.99 41.35

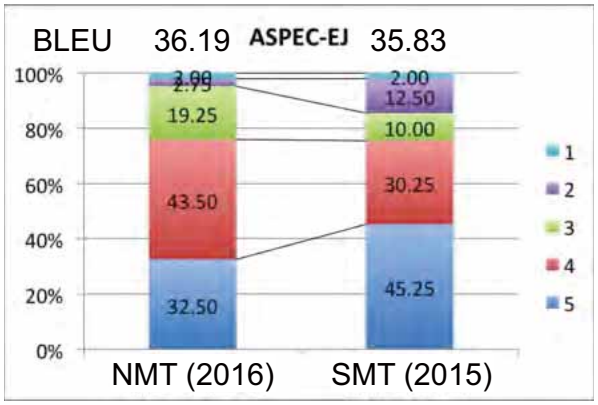


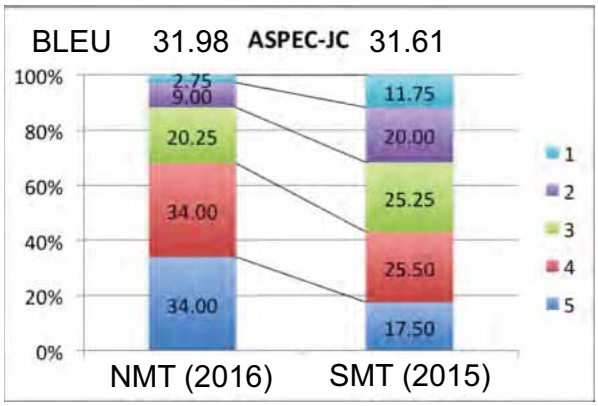
BLEU 71.51 71.38

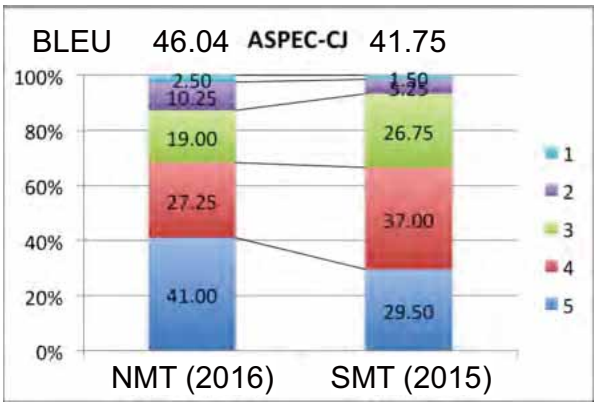
25

科学技術論文翻訳 経年評価









26

翻訳結果の分析

翻訳例

1. SMTにおいて(5, 5)の評価だったものが、NMTにおいてどちらも5未満の評価になったもの
2. NMTにおいてどちらも2以下の評価になったもの

- 中→日 特許文、日→英・英→日 科学技術論文

特許 中→日 1

IN	“聚合酶链反应”(“PCR”)是本领域技术人员熟知的。
REF	「ポリメラーゼ連鎖反応」(「PCR」)は、当業者によく知られている。
SMT	「ポリメラーゼ連鎖反応(PCR)」は、当業者に周知である。(5, 5)
NMT	「重反応」(「PCR」)は当業者に周知である。(4, 2)

IN	具体来讲，如图2中所示地执行晶锭生长步骤(S110)。
REF	具体的には、図2に示すように、まずインゴット成長工程(S110)を実施する。
SMT	具体的には、図2に示すように、インゴット成長工程(S110)が行われる。(5, 5)
NMT	具体的には、図2に示すように結晶成長工程を行う(S110)。(4, 4)

IN	关于施加与最佳值错开的电压时获得的电流量受到血液试样内的血细胞比容值(Hct值)的影响。
REF	最適値から外れた電圧を印加した時に取得される電流量は、血液試料内のヘマトクリット値(Hct値)によって左右され得る。
SMT	最適値からずれる電圧を印加した際に得られる電流量は、血液試料内のヘマトクリット値(Hct値)の影響を受ける。(5, 5)
NMT	最適な値の電圧を印加する際に得られる電流量は、血液試料内のヘマトクリット値(Hct値)の影響を受ける。(4, 4)

特許 中→日 2

IN	本发明片剂还可以包含包衣层，它可占片剂组合物的约0-15%(重量)。
REF	また錠剤は、錠剤組成全量中0～約15重量%を含むコーティング層を有することができる。
SMT	本発明の錠剤は、コーティング層をさらに含むことができるが、それは、錠剤組成物の約0～15重量%を占めることができる。(5, 5)
NMT	本発明の錠剤はまた、錠剤組成物の約0～15重量%(重量)を含有していてもよい。(1, 2)

IN	使用由高取向热解石墨(HOPG)制成的单色器702作为X射线的单色化单元以便提取Mo的特征X射线。
REF	X線の単色化手段としては高配向性熱分解黒鉛(HOPG)のモノクロメータ702を用いMoの特性X線部分を抽出する。
SMT	X線として高配向熱分解黒鉛(HOPG)からなるモノクロメータ702の単色化ユニットを用いてMoの特性X線を抽出することができる。(3, 3)
NMT	高配向熱分解黒鉛(HOPG)からなるの702は、X線を用いてX線を抽出するために、X線を使用する。(1, 2)

論文 日→英 1

IN	Hartree–Fock–Bogoliubov法を用いて変形した原子核の対密度分布を調べた。
REF	Using Hartree–Fock–Bogoliubov method, pairing density distribution of the deformed nuclei was investigated.
SMT	The pair density distribution of deformed nuclei was investigated using Hartree–Fock–Bogoliubov method. (5, 5)
NMT	The density distribution of deformed nuclei is investigated by the Hartree – Fock – Bogoliubov method. (4, 4)

IN	軸対称な四重極変形した原子核を対象にした。
REF	Axial symmetric quadrupole deformed nuclei were examined.
SMT	The axisymmetric quadrupole deformation nuclei were made to be an object. (5, 5)
NMT	An axially symmetric quadrupole nucleus is studied. (4, 3)

IN	表面型対相関を用いて密度分布と対密度分布の角度依存性を求めた。
REF	Angular dependences of the density distribution and pairing density distribution were obtained with the surface type pairing correlation.
SMT	The angular dependence of density distribution and pair density distribution was obtained using a surface–type pair correlation. (5, 5)
NMT	The surface – pair correlation is used to obtain the angular dependence of the density distribution and the density distribution. (4, 4)

論文 日→英 2

IN	食事については，全介助から半介助となった。
REF	On the meal, changing from all support to half support was realized.
SMT	The meal became half assistance from the assistance. (1, 4)
NMT	On the meal, it became the assistance from all assistance assistance. (1, 2)

IN	排せつ，排尿はまだ時間を要する
REF	It takes much more times in excretion and urination.
SMT	The excretion, urination, still needs the time. (5, 5)
NMT	Excretion and voidance are necessary for a long time. (1, 2)

IN	スキー滑走中に転倒受傷し右膝変形ととう痛にて受診。
REF	A man tumbled over in a ski slide and was injured and then had right knee deformation and pain and so received medical examination.
SMT	He had medical examination by pain and right knee deformation injured overturning during ski. (3, 4)
NMT	The ski was injured in ski ski sliding, and it had medical examination by right knee deformation and pain. (2, 2)

論文 英→日 1

IN	Using Hartree–Fock–Bogoliubov method, pairing density distribution of the deformed nuclei was investigated.
REF	Hartree–Fock–Bogoliubov法を用いて変形した原子核の対密度分布を調べた。
SMT	Hartree - Fock - Bogoliubov法を用いて，変形した核の対密度分布を調べた。(5, 5)
NMT	標記の方法を用いて，変形核の対密度分布を調べた。(4, 3)

IN	By this approach, monodispersed swollen styrene particles were obtained and the fine particles in the title were synthesized.
REF	この方法で，単分散なスチレン膨潤粒子を得，標記微粒子を合成した。
SMT	このアプローチにより，単分散膨潤スチレン粒子が得られ，標記微粒子を合成した。(5, 5)
NMT	この方法により，単分散のスチレン粒子を得，標記微粒子を合成した。(4, 4)

IN	No side effect was noted during treatment.
REF	治療中，副作用は認めなかった。
SMT	治療中に副作用はなかった。(5, 5)
NMT	副作用は認めなかった。(4, 3)

論文 英→日 2

IN	The liver and intestine were extracorporeally observed through the central abdominal wall in the gestational age.
REF	在胎中，肝と腸が中心腹壁を介し体外的に観察された。
SMT	肝臓や小腸は妊娠期間における中枢腹壁を通して体外に観察された。(5, 4)
NMT	妊娠時の腹部CTでは，肝，腸管はほぼ全周性に描出された。(1, 1)

IN	Pilot plant tests showed the system gives clean water reusable as feed water to deionizers.
REF	パイロットプラントによる結果は，純水装置の原水として再利用可能レベル迄処理できた。
SMT	パイロットプラント試験はdeionizersに水として清浄水再使用を与えるシステムを示した。(3, 4)
NMT	パイロットプラント試験により，本システムの有効性を確認した。(2, 1)

IN	Polonium 210 is found, in an extremely low amount though, in the cigarette smoke.
REF	またポロニウム210はごく微量であるが，タバコの煙の中にも含まれている。
SMT	たばこ煙が極めて低い量で210ポロニウムが見られる。(3, 5)
NMT	シガレット210は極低濃度ではあるが，たばこ煙では検出されていない。(1, 2)

結果からわかること

- NMTはアジア言語の翻訳においても有効
- NMTでは翻訳文の流暢さは完璧に近い
- ただし正確性は完璧ではない
  - 平均的な翻訳精度は格段に向上するが、訳抜けが起こりやすく、完璧な翻訳になる割合が低い
  - 低頻度語の翻訳に弱い
  - たまにおかしくなったように同じ単語を繰り返し出力する
- 上記問題の解決方法はすでに提案され始めている

35

結果からわかること

- 日⇔中翻訳はかなり向上した
  - 同様の結果は中英翻訳でも確認されている
    - https://arxiv.org/abs/1610.01108
  - おそらく、中国語の単語分割誤りの影響がNMTでは低減されることが原因
- そもそも翻訳精度が高かった日⇔英翻訳では前期の問題により日⇔中ほど大きな向上は見られない

36

今後の展望

- ニューラル翻訳(NMT)はここ数年で急激に発展し、統計翻訳の精度を追い越している
- NMTの研究はまだ発展する可能性が高い
- ただし、過去の遺産(SMT)が活躍する場もまだ残っていることは確かで、うまく組み合わせられると良い
  - 特に対訳コーパスが少量の場合NMTはSMTよりも弱い
  - NMTの発展次第では本当に遺産になる可能性もあるが

今後の展望

- NMTをプロダクト化したところも出ている
  - Baidu (http://www.aclweb.org/anthology/W15-4110)
    - 中⇔英 (2015/05/20)
  - SYSTRAN (http://blog.systransoft.com)
    - 12言語 (2016/08/31, 日、韓、英、仏など)
  - WIPO (http://www.wipo.int/pressroom/en/articles/2016/article_0014.html)
    - 特許 中、日、韓→英 (2016/10/31)
  - Google (https://research.googleblog.com/2016/09/a-neural-network-for-machine.html)
    - 中⇔英(2016/09/27) 、日⇔英 (2016/11/11)
- これらとどう渡り合うか？

招待講演 3

“Domain Adaptation for Machine Translation at NAVER LABS.”

Domain Adaptation for Machine Translation at NAVER LABS

Hyoung-Gyu Lee

NAVER | L | A | B | S |

Outline

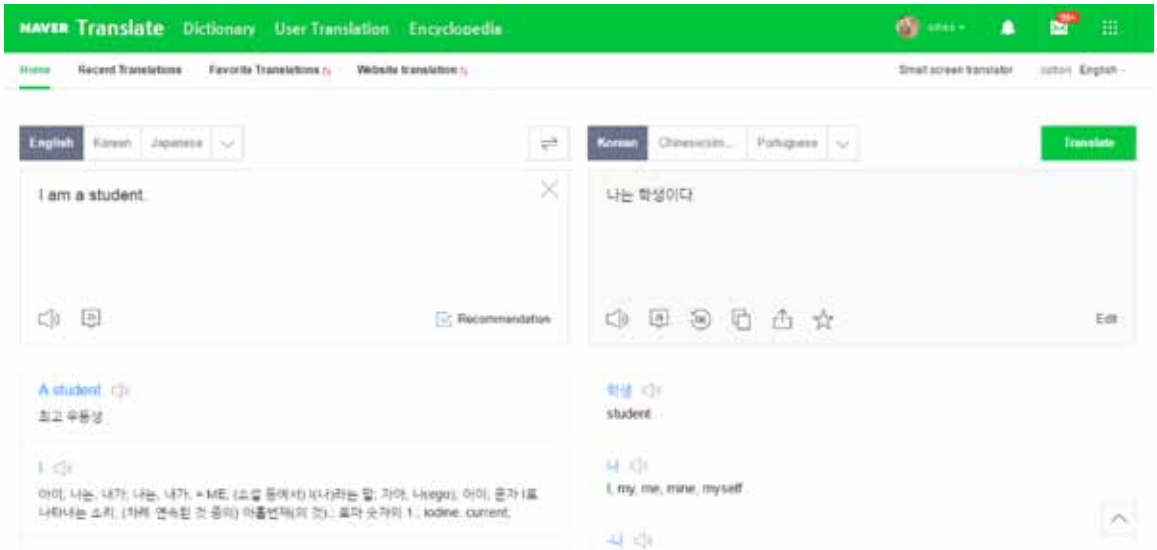
대외비

- Introduction
  - NAVER MT Services
  - MT work at NAVER LABS
- Intro to Domain Adaptation for MT
- Domain Adaptation for MT at NAVER LABS
  - IT Manual Domain
  - Experimental Results
- Summary Future work

NAVER MT Services

NAVER MT Service

- http://translate.naver.com



NAVER MT Service

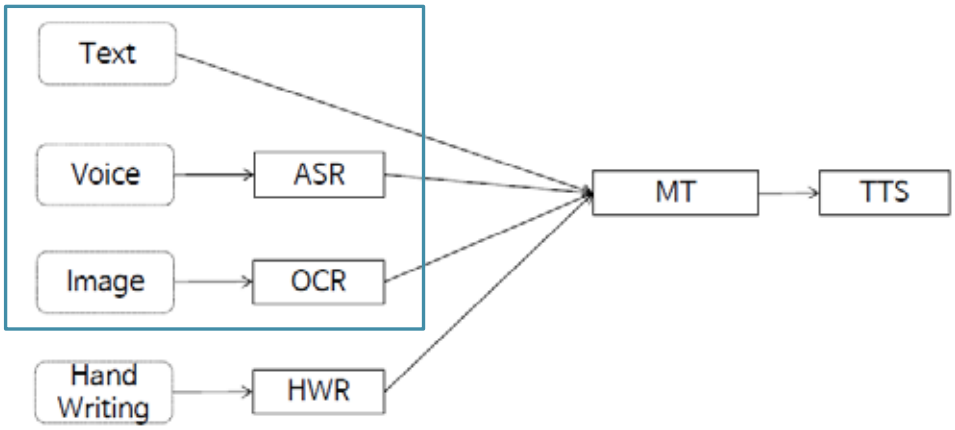
- LINE Translation bot

NAVER MT Service

- NAVER Open API
  - https://developers.naver.com/products/translator
  - Korean <-> English
  - Korean <-> Chinese
  - Korean <-> Japanese
  - 10,000 characters/day

NAVER MT Service

- Papago Translator App
  - Mobile app service of NAVER MT
  - Voice, image, text as input
  - Using in-house engines for ASR, OCR, TTS, and MT



NAVER MT Service

- Papago Translator App
  - Android, iOS

papago

Text input

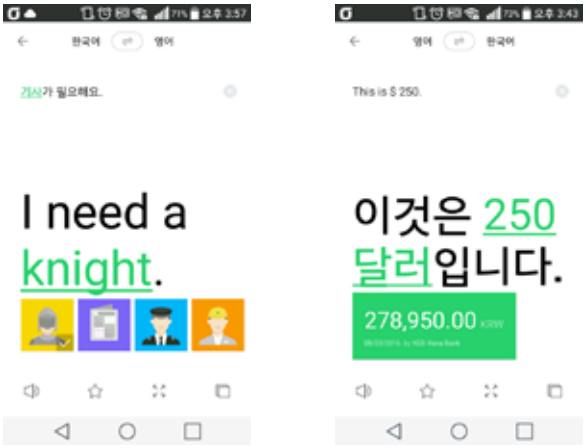
Voice input

1-to-1 conversation

Image input

NAVER MT Service

- Papago Translator App
  - "papago: A Machine Translation Service with Word Sense Disambiguation and Currency Conversion" (Coling 2016 Demonstration)
    - 1) WSD based on User Feedback
    - 2) Instant Currency Conversion



MT Work at NAVER LABS

MT Work at NAVER LABS

- MT Training Corpus
  - Most important work for MT!
  - Need much money
    - Little free available and high quality data
    - Partnerships with several translation companies

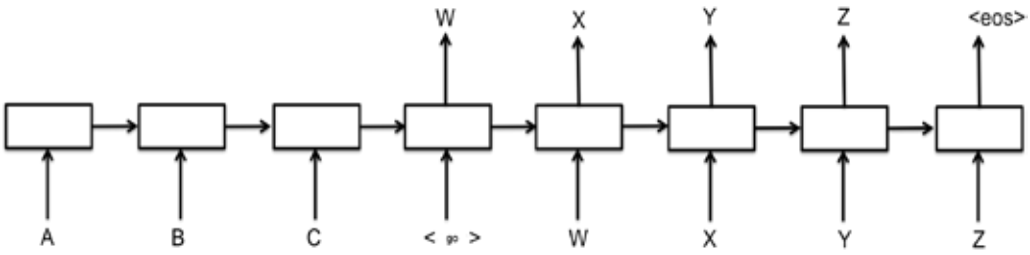
MT Work at NAVER LABS

- Traditional SMT
  - Phrase-based
    - Japanese <-> Korean, …
    - Using pre-reordering method for several languages
  - Hierarchical Phrase-based
    - Chinese <-> English,
    - Korean <-> English, …

대외비

MT Work at NAVER LABS

- NMT
  - RNN sequence-to-sequence model
  - LSTM, GRU
  - Korean <-> English, Korean <-> Chinese, …
  - On-going work
  - We have opened the NMT service for Korea <-> English !
    - NAVER Translate, Papago app, NAVER labspace



NAVER LABS

대외비

MT Work at NAVER LABS

- MT Competition
  - WAT 2015
    - Participate in two tasks
      - Korean→Japanese: Phrase-based SMT
      - English→Japanese: Syntax-based SMT + NMT Rescoring

Team	HUMAN	BLEU
NAVER	4.77	94.38
NICT	4.75	94.35
Online	4.52	90.92
SENSE	4.32	95.45

[K→J]

Team	HUMAN	BLEU
NAIST	4.04	35.83
NAVER	4.00	34.60
WEBLIO	3.81	33.23

[E→J]

NAVER LABS

Intro to Domain Adaptation for MT

Intro to Domain Adaptation for MT

- Why Domain Adaptation?
  - Mismatch between
    - Domain for which training data are available
    - Target domain of a machine translation system
  - Different domains may vary by topic or text style

Intro to Domain Adaptation for MT

대외비

(1) If we have in-domain parallel data
- Hard to acquire the data
- Just training

(2) If we have in-domain monolingual data
- Relatively easy to acquire the data
- LM adaptation
- Automatically generate in-domain parallel data using translator

(3) If we have in-domain monolingual data & general-domain parallel data
- Select In-domain sentences with sentence classifier

Intro to Domain Adaptation for MT

대외비

Acquire as much in-domain data as you can!

Previous Work of Domain Adaptation

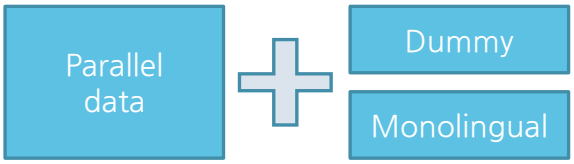
대외비

- Domain Adaptation via Pseudo In-Domain Data Selection (Axelrod et al., EMNLP 11)
  - Cross-entropy based method
  - To score each sentence,
    - Use domain-specific LM trained with in-domain monolingual data
  - Show that relatively tiny amounts of in-domain data can prove more useful than the entire general-domain corpus

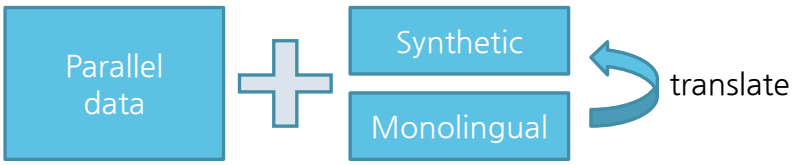
Previous Work of Domain Adaptation

대외비

- Improving Neural Machine Translation Models with Monolingual Data (Rico Sennrich et al. ACL 2016)
  - Dummy source sentences



  - Synthetic source sentences



Previous Work of Domain Adaptation

- If target domain is not determined
  - Mixture model approach
    - (Foster and Kuhn, 2007), (Schwenk and Koehn, 2008)
    - Combination of multiple domain-specific models (TM or LM)

    - MT Process
      - Input sentence
        → Domain classification → Domain proportions
        → Weighting domain-specific models
        → Translation

Domain Adaptation at NAVER LABS

Domain Adaptation at NAVER LABS

대외비

- Task: IT Manual Domain Adaptation
  - Task scenario
    - Domain-targeted translation task: IT manual
    - We have both large in-domain parallel data and large general-domain parallel data
  - IT manual data
    - Provided by hansemEUG
      - A translation company in Korea
      - www.ezuserguide.com
    - About 2 million sentence pairs

Domain Adaptation at NAVER LABS

대외비

- Method
  - Utilize both in-domain data and general-domain data
  - Select In-domain data from general-domain data
    - Adopt method proposed by (Axelrod et al., EMNLP 11)
  - Comparative study
    - General-domain
    - In-domain
    - In-domain + General-domain
    - In-domain + In-domain selected from general-domain

Examples of In-domain Selection

Settings
- 4-gram English LM and 4-gram Korean LM
- 2 million training sentences for each LM
- Rare words as UNK
- Top-15

Warning: This computer program is protected by copyright law and international treaties.	경고: 이 컴퓨터 프로그램은 저작권법과 국제 협약의 보호를 받습니다.
Error! Can't edit a pilot profile with a name that has empty spaces at the beginning or the end.	오류! 처음이나 끝에 공백이 있는 이름으로 조종사 프로필을 편집할 수 없습니다.
For more information, contact your system administrator.	자세한 내용은 시스템 관리자에게 문의하십시오.
Error! Can't create a pilot profile with a name that has empty spaces at the beginning or the end.	오류! 처음이나 끝에 공백이 있는 이름으로 조종사 프로필을 만들 수 없습니다.
Type your user name and password, and then click OK.	사용자 이름과 암호를 입력한 다음 확인을 클릭합니다.
You can do the following:	다음 작업을 수행할 수 있습니다.
You have successfully completed the Initialize and Convert Disk Wizard.	디스크 초기화 및 변환 마법사를 완료했습니다.
This will delete all headers and message bodies and will reset the folder so that headers will be re-downloaded.	모든 머리글과 메시지 본문을 삭제하고 폴더를 재설정하여 머리글을 다시 다운로드할 수 있도록 합니다.
You might need to restart your computer for the changes to take effect.	변경 사항을 적용하려면 컴퓨터를 다시 시작해야 합니다.
For more information, see the documentation that came with your computer.	자세한 내용은 컴퓨터 구입 시 함께 제공된 설명서를 참조하십시오.
The time between the most recent refresh of a record timestamp and the moment when the timestamp may be refreshed again.	레코드 타임스탬프의 최신 새로 고침과 타임스탬프가 다시 새로 고쳐지는 순간 사이의 시간입니다.
Type the user name, password, and domain of an account with administrative rights.	관리자 권한을 가진 계정의 사용자 이름, 암호, 도메인을 입력하십시오.
You can also do the following:	또한 다음 작업을 수행할 수 있습니다.
Site link bridge objects must link at least two site link objects.	사이트 링크 브리지 개체는 적어도 두 개의 사이트 링크 개체를 연결해야 합니다.
Choose one of the following:	다음 중 하나를 선택합니다.

Experimental Results

- Experimental environments
  - Evaluation set
    - IT manual 1,000 sentence pairs for BLEU eval.
    - IT manual 100 sentences for human eval.
  - Human evaluation
    - Count of error words
    - Lower is better
    - Evaluated by hansem
  - All systems are implemented by using in-house engines
    - Tokenizer, SMT decoder, NMT seq2seq, Beam search decoder, …

NAVER LABS

Experimental Results

- In SMT,
  - HPB model
  - Large general-domain vs. Large In-domain data
  - Effect of In-domain data selection from general-domain data

SYS	KoEn		EnKo	
	BLEU	Human (#Errors)	BLEU	Human (#Errors)
General-domain	15.78	-	20.41	-
In-domain 2M	23.64	238	33.17	236
In-domain 2M + General-domain	23.90	221	33.08	223
In-domain 2M + In-domain selection 200k	24.32	197	34.80	238

Experimental Results

- In NMT,
  - SMT vs. NMT
  - Effect of In-domain data selection from general-domain data

SYS	KoEn		EnKo	
	BLEU	Human (#Errors)	BLEU	Human (#Errors)
HPB, In-domain 2M + In-domain selection 200k	24.32	197	34.80	238
NMT, In-domain 2M + In-domain selection 200k	27.39	119	39.00	201

Summary and Future Work

Summary

- Domain Adaptation
  - Required to develop translator for target domain
- Our experiments shows
  - Large in-domain parallel data is powerful
  - In-domain data selection from general-domain data
    - Effective for domain adaptation
    - SMT model size reduction
    - NMT training time reduction
- Need to collect as much in-domain data as possible
- Our domain-adapted MT is on-going work

Future Work

- Adaptation to Additional Domains
  - Shopping, e-commerce
  - Shop name
- Further Study in NMT
  - Domain adaptation by Integrating In-domain data selection into NMT
  - Utilize in-domain monolingual data

Thank You

招待講演 4

「日本語の素晴らしさとユーザーの機械翻訳への大きな期待」

日本語の素晴らしさと
ユーザーの機械翻訳への大きな期待

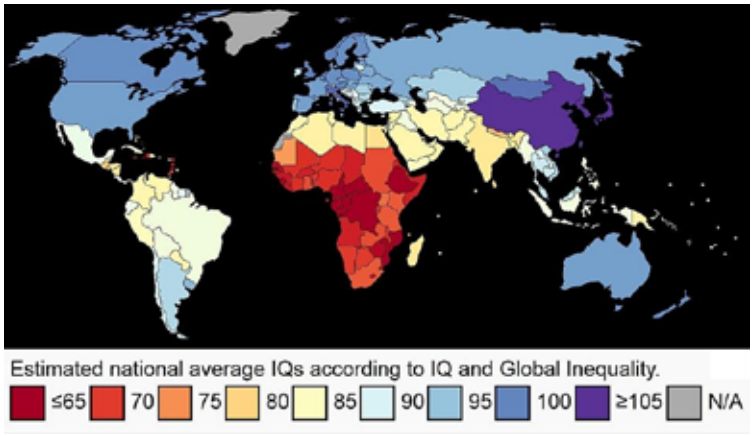
奥山尚一
久遠特許事務所

2016年11月25日
第4回特許情報シンポジウム

日本語かっけー

- みんな仮名は簡単に分かる（識字率99%）
  - これで一応なんとかなる
- その後、少しずつ漢字を覚える
  - 常用漢字　2136字／4388音訓！
- 右脳と左脳の両方を使う
- 平均IQは１１０！
  - Nature 297, 222 - 223 (20 May 1982)　Richard Lynn, The New University of Ulster,（共同ヨーロッパ総局発の記事）

最近のIQは105



講演者について

- 理工学部電気工学科卒
- アメリカで理論化学のPh.D.
- 弁理士(1990年)　弁理士会会長(2011年から13年)

- 特許翻訳学校の創設(17年前)

- 株式会社クールデザイン設立(17年前)(オンライン署名個人認証)
- 内閣府知的財産戦略本部有識者本部員(もうすぐ任期切れ)

日本の翻訳の歴史 - 外との大衝突

- 5世紀には中国から漢字の書物を輸入、普及
  - 万葉仮名という綴り方を発明
- 9世紀に入ると、そこから平仮名と片仮名
  - 梵語（サンスクリット語）に起源のある仏教用語

- 明治初期の「大翻訳時代」
- まず「概念」を導入しなければならなかった
- そこで和製漢語
  - 「社会」(society)、「文化」(culture)、「文明」(civilization)、「時間」(time)、「美術」(art)、「空間」(space)、「科学」、「彼、彼女」
  - 「恋愛」、ロマン派詩人北村透谷により「love」の訳として作られた
    - それまでは「恋」とか「色事」とか
  - 元来の意味とは全く違うものとして生まれ変わらせた
    - consciousnessは「意識」。「意識」は元々は「分別の心」という意味
    - feeling/moodは「気分」に。「その人の本来の性質」という意味

自由

- Freedom?　Liberty?

- 「自由」は古典中国語では「後漢書」、日本では「続日本紀」まで遡ることができる。
- 我儘放蕩(わがままほうとう)の意味であった。徒然草に「よろづ自由にして、大方、人に従うといふことなし」(60段)

経済

- Economy

- 古代中国の「経国済民」もしくは「経世済民」の略
- 「経国済民」「経世済民」は、国（世）を治め民を救済することを意味し、現代でいう「政治」の意味に近い
- 日本では、江戸時代の学者用語に現れ、理念的な政治政策の意味
- しだいに経済運営の意味で使われる
- いまの辞書(goo辞書)：《名》人間の生活に必要な財貨・サービスを生産・分配・消費する活動。また、それらを通じて形成される社会関係。

科学

- 明治初期に“science”に対して日本でつくられた造語
- この「科学」というのは，文字どおり「科」の「学」。外「科」，内「科」，小児「科」，産婦人「科」の「科」。つまり個別の「科」に分れた、専門分化した学問という意味
- 一方，1600年代のアマチュア的「知的活動」は、専門分化した「科」学ではなかった
  - ロバート＝フック（1635～1703）は、現在でいう物理学、化学、生物学、地学などのいろんなことを研究
  - もちろん，当時はphysics（物理学），chemistry（化学），biology（生物学），geology（地質学）という言葉もなかった
- 1800年後半には“science”は，すでに専門分化した「科」学になっていた
  - 明治の初期に日本人が「科学」という訳語を作った

- そもそも“science”はいつから使われている？
  - ラテン語の「知る」（scere）に基づいている　たんなる「知識」（scientia）というような意味
    - 1600年のはじめころ
  - 1660年に成立した王認学会（ロイヤル・ソサエティ）の正式名称は「The Royal Society of London for Improving Natural Knowledge」（自然の知識の増進のためのロンドンの王認学会）
  - 王認学会などを中心に行われていた1600年代の「知的活動」は“science”ではなく、“natural philosophy”と呼ばれていた

化学

- 最初は、「舎密（セイミ）」　(1837頃)
- 幕末の蘭学者宇田川榕菴が、オランダ語のChemie （科学）の音をそのまま活かして、日本語化しようとした
  - Wikipediaより
- 蘭学者川本幸民（1810-1871）の著書「化学新書」(1860)が最初
  - ドイツのステックハルトの“Die Schule der Chemie”のオランダ語版を和訳

例えば、宇田川榕庵の場合

- 「舎密開宗」1837～1847年に出版
- 原著はイギリスの化学者ウィリアム・ヘンリーの“Elements of Experimental Chemistry”（1799年）で、ヨーロッパでドイツ語からさらにオランダ語に重訳され、それを日本語に
- ヨーロッパとの出版の時差が38～48年。単なる翻訳ではなかった
- 酸素、水素、窒素、炭素、白金等の元素名や元素、酸化、還元、溶解、分析等の化学用語、細胞、属等を訳出した
- なお、coffeeを「珈琲」と標記した
  - Wikipediaより

これだけイメージが広がる言語があるか

- 万葉集からイメージが広がるか
  - 英語だって、せいぜいシェークスピア(1556-1616)から
  - まあ、ラテン語もあるか
- 言語をサポートする力
  - 文学、科学、技術
- 大学院まで日本語で
  - そんな国数えてみて？
  - 普通、科学技術は英語

日本語という言語

- 孤立言語
  - どの言語とも類縁でない　韓国語とも語彙が違う
- 母音を主体としている
- 表意文字と表音文字の混在
  - ひらがな、カタカナ、漢字、アルファベット（ほぼ完全無欠）
    - ロシア語、タイ語、アラビア語、ハングルは無理か
- 2000年前からの伝承と積み重ね
    - 古事記は712年、日本書紀720年、万葉集759年以降（630年ころの歌から？）
- 日常語彙が豊富

母音を主体としている

- 雨がザーザー降る　　雪がシンシンと積もる
- 雲がふんわり浮かんでいる
- ぽかぽかした小春日和　　ぽかぽかなぐられる

- 人間界と自然の近さを生んだ
- 工業製品の優秀さにつながる？

漢字かな混じり文

- 表意文字と表音文字の混在

- 漢字一つ一つに意味がある
  - 視覚による処理
- 一目で全体を理解しやすい（右脳の世界）
- 漢字があるので情報量が多く、仮名があるので表現が自由

膠着言語

- 最後まで聞かないと読まないと分からない
  - 僕は、明日仕事に行く
    - かもしれない
    - 気が無い
    - べきでしょうか
- 表現のニュアンスが豊かになる

- 英語とは違いますよね

機械翻訳がやってくる！

機械翻訳ができると

- 言語は、コミュニケーションのツールである必要がなくなる
- 言語は、思考と認識のツールになる

- そうであるなら、日本語は圧倒的に有利
  - この優位性を維持する努力が必要

- 英語の呪縛から逃れられるか？
  - 英語がコミュニケーションのツールである必要がなくなる

機械翻訳への期待

- 例えば、日本の知財判例の世界への発信が容易になる
  - その他の情報も本質的に同じ
- 特許情報も自由に流れるようになる
  - グローバル・ドシエ（昨年１１月から）

- とりあえず、翻訳者にとっては下訳作成のツールになる

- そして、その後の世界は！

日本語が世界の中心へ！

コメントは、okuyama@quon-ip.jpへ

20

特別講演

「文章と翻訳の品質を改善する―構造化用語データ UTX による用語管理と実務日本語ルール」

http://www.aamt.info/japanese/utx/

AAMT（アジア太平洋機械翻訳協会）

シンプルな用語集形式 UTX

UTX用語集形式とは?

AAMTが策定したシンプルな用語集形式

専門用語を管理する用語集の作り方のルール

Excelなどでも編集できるタブ区切り形式

UTXの利点

専門用語を含む用語集を簡単に作成・管理できる

共通の形式なので用語データを共有・再利用しやすい

人間翻訳と機械翻訳を効率よく正確にできる

UTX 用語集形式サンプル

用語集全体についての情報（作成日、使用許諾など）

#UTX 1.11; en-US/ja; 2016-04-01T19:00:00+09:00; copyright: AAMT (2014); license: CC BY 4.0

#src	tgt	src:pos	term status
Asia-Pacific Association for Machine Translation	アジア太平洋機械翻訳協会	properNoun	approved
contributor	用語提出者	noun	provisional
optional	省略可能	adjective	approved
optional	オプショナル	adjective	forbidden
merge	統合する	verb	approved
merge	マージする	verb	forbidden
unidirectional	一方向	adjective	approved
monodirectional	一方向	adjective	non-standard

原語　訳語　品詞　用語ステータス（省略可）

安心して使える用語か、禁止用語か区別できる

概要

UTX（Universal Terminology eXchange）とは、AAMT（アジア太平洋機械翻訳協会）が策定した、用語集形式です。UTX用語集は、翻訳者のための用語集として使えることに加え、各種形式に変換して、さまざまな翻訳支援ツールで使えます。特に翻訳ソフトの用語データ（ユーザー辞書）として使うことで、翻訳精度を大きく向上できます。ユーザー視点から、**シンプルで作りやすく、使いやすい**ことを目指しています。

AAMTは、機械翻訳の研究開発者、製造販売者、利用者から構成される団体で、メンバーはボランティアです。

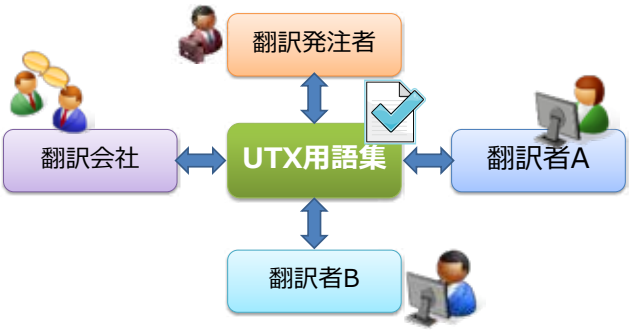
なぜ翻訳で用語集が必要か

企業・大学・組織での専門的な翻訳では、用語集がないと、専門用語を正しく訳せません。初期段階で用語集をしっかり作れば、同じ言葉を何度も調べずにすみ、翻訳の費用と労力を減らせます。用語集は、訳語統一や誤訳チェックにも役立ちます。高品質な翻訳を正確に行うには、用語集は必須です。

なぜUTXを使うといいのか

UTXを使えば、用語集を簡単に作成・共有・再利用して、翻訳の質を向上できます。「翻訳ソフトは変な訳ばかり出す」と思われていませんか？　その理由は、語句をどう訳すべきかという**翻訳**

共通規格なのでユーザー間で用語集を共有・再利用しやすい



知識が不足しているからです。まず、翻訳知識をUTX用語集として蓄積し、それをユーザー辞書に変換することで、翻訳ソフトの翻訳精度を大きく改善できます。

Excelやテキスト ファイルでも、各項目の形式が共通化されていないと、共有や再利用は困難です。さまざまな用語集がインターネットで公開されていますが、実際にはすぐに活用できず、手間のかかる修正と調整が必要です。しかし、UTXのような標準規格に沿った形式であれば、さまざまなツールで用語集を広く共有し、すばやく再利用できるようになります。UTXを経由することで、異なる用語集形式の変換を橋渡しできます。

だれが作り、使うのか

主に翻訳者や翻訳ソフトのユーザーが、作り、使うことを想定しています。UTX用語集は、原語、訳語、品詞など、最低限の情報のみでも作れます。必要な情報があれば追加できます。

どのような分野で使うのか

IT、医療、法律、工学など、専門用語が多い翻訳であれば、どのような分野でも使えます。

どんな語を含むのか

UTX用語集は、製品・部品名、病名、薬品名、法律名など、**特定分野の専門用語**や、人名、地名、施設名などの**固有名詞**を含みます。一般的な語は含めません。

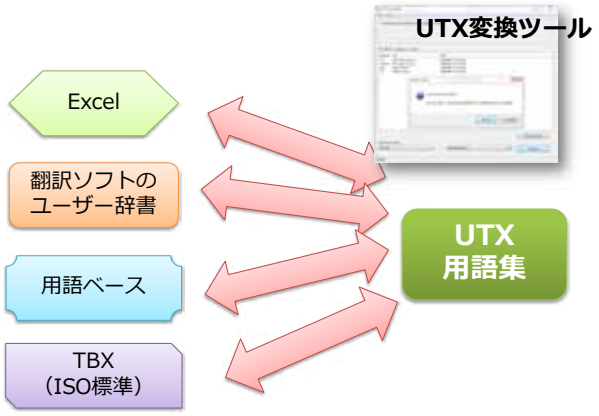
どのような状況で作るのか

1. 用語集をゼロから作成する場合
2. 多人数による翻訳で発生する訳語を収集しまとめる場合
3. 各種用語データ間変換で、中間変換形式として使う場合

どうやって作るのか・使うのか

UTX用語集は、Excelなどの**表計算ソフトやテキスト エディターで簡単に作成・編集・表示**できます。翻訳メモリー ツールOmegaT、用語ツールApSIC Xbenchなどのツールでは、UTXをほぼそのまま使えます。

UTXと各種形式の相互変換



また、各種形式とUTX形式の相互変換を行うことで、各種ツールで使用できます。 変換ツールとして、公式UTX変換ツールやGlossary Converterがあります。公式UTX変換ツールでは、UTXと、ATLAS、The翻訳、PC-TranserなどTranserシリーズの間で相互変換ができます。

詳細:〈http://www.aamt.info/japanese/utx/tools.htm〉

どこで使われているのか

UTXは、企業で使われている他、特許庁でも機械翻訳辞書(220万語)が作成されています。企業事例も公開予定。

費用はかかる？　より詳しく知りたい

UTX仕様書、サンプル用語集、変換ツールは、無料でダウンロードして使用できます。UTX仕様に基づいて、どなたでもご自由にUTX用語集を作成・公開・共有できます。「用語集を作りたい、用語データを活用したい」とお考えの組織や企業の方は、ご連絡いただければアドバイスを差し上げられます。

AAMT(アジア太平洋機械翻訳協会)機械翻訳課題調査委員会 共有化・標準化ワーキンググループ

http://www.aamt.info/japanese/utx/

問い合わせ先：aamt-info@aamt.info

メンバー(順不同)

山本 ゆうじ（リーダー）	秋桜舎
村田 稔樹	沖電気工業株式会社
Francis Bond	南洋理工大学（シンガポール）
島津 美和子	東芝ソリューション株式会社
大倉 清司	株式会社富士通研究所
加藤マイケル孝仁	ジャパニーズ・グレイツ株式会社
秋元 圭	合同会社ことばや
高橋 博之	株式会社クロスランゲージ

2016年8月版

UTX 用語管理で翻訳と文書の効率・品質を高める

http://www.aamt.info/japanese/utx/

アメリカ英語／日本語／中国語（簡体字）

#UTX 1.20; lang: en-US/ja/zh-CH; copyright: AAMT (2016); license: CC BY 4.0					
#term:en-US	tgt:ja	term status:en-US	term status:ja	concept ID	term:zh-CH
#用語（アメリカ英語）	**用語（日本語）**	**用語ステータス（アメリカ英語）**	**用語ステータス（日本語）**	**概念ID**	**用語（中国語簡体字）**
Asia-Pacific Association for Machine Translation	アジア太平洋機械翻訳協会	approved	approved		亚洲太平洋机器翻译协会
contributor	用語提出者	approved	provisional		用语提交者
entry	項目	approved	approved	1	条目
entry	エントリー	approved	forbidden	1	
unidirectional	一方向	approved	approved	2	单向
monodirectional	一方向	non-standard	approved	2	

用語集を「見える化」して社内の専門知識を集約・共有・再利用

UTX 用語集の仕様は、用語の専門家により考えぬかれた用語集作成のルールです。シンプルな表形式の一覧で表されるため、**埋もれていた社内の専門知識を集約**し、文書コンテンツとして共有・再利用しやすくできます。用語ステータスに基づいて自動色分けすると、どの用語を使い、どの用語を禁止すべきか見やすくできます（上図参照）。

さまざまなツールで活用できるシンプル・軽量な標準化形式

社内で作った Excel 用語集が活用されずに放置されていませんか？　UTX 用語データの構造は、仕様に基づいて標準化されているため、機械翻訳、翻訳支援ツール（翻訳メモリー）など**さまざまなツールで活用**できます。UTX 用語データは、公式変換ツールなどで、各社翻訳ソフトのユーザー辞書や、ISO 規格の用語集形式 TBX にも変換可能。

難しい用語を書き換えて読みやすく翻訳しやすい日本語にする

UTX では日本語書き換えと対訳を 1 つの用語集で管理可能。用語データに基づき、文書内に難しい・あいまいな用語、その他差別表現など不適切な用語があれば、適切な用語に書き換えることができます。日本語文書を読みやすくすると、海外展開するときに翻訳しやすくなります。機械翻訳の前処理（プレ エディット）にも使用できます。

機械翻訳での用語の適用およびチェック（ポスト エディット）

統計機械翻訳では用語レベルの正確性は保証されません。しかし、後編集（ポスト エディット）で UTX 用語データを使えば、**訳された用語をチェック**できます。ルールベース機械翻訳（市販の翻訳ソフトなど）では、最初から UTX 用語データに基づいた正確な用語で翻訳できます。

CMS、DITA、各種 XML での用語および表記の統一

共有・再利用される構造化文書、文書コンテンツ、マニュアルでは、多数の書き手が関わるため用語と表記が一貫している必要があります。シンプルな UTX 形式に基づけば、**必要最小限の手間で用語と表記の統一**が可能です。

なぜ Excel より UTX 用語集が便利？

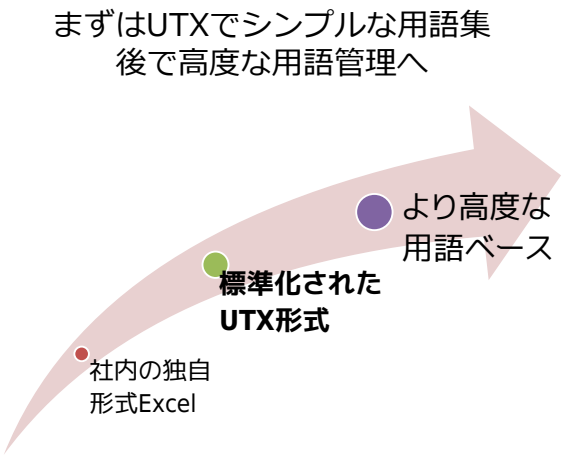
用語集を文書作成や翻訳に活用されていますか？

「文書の専門用語が難解・意味不明で困る」

「ねじ、ネジ、スクリューなど、用語がバラついているので統一したい」

そういうお悩みが何年も前からずっとありませんか？　それなら用語集さえ作れば解決できるはず。

しかし、用語集作成は、最初の一歩を踏み出さないと永遠に悩み続けるだけです。

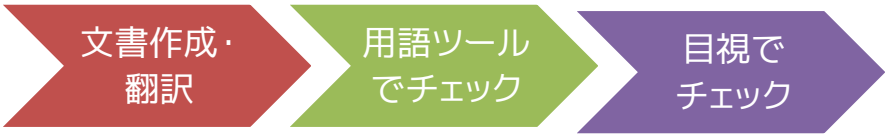


用語集の「最初の一歩」は Excel 用語集でいいのです。そのほうが楽です。しかし、**Excel 用語集は、用語の問題に気づいたときに参照されるだけです。**見過ごした問題はそのまま放置されます。しかも文書と用語集をつきあわせるのが面倒で、やがてだれも用語集を使わなくなります。用語集を確実に統一する**唯一の解決策は、文書作成・翻訳工程にアクティブな自動用語チェックを組み込むことです。**

また目視の用語チェックは重要ですが、目視チェックだけでは効率が悪く、必ず見落としが発生します。

目視に加えて、用語ツールで用語チェックをしましょう。

ツールでチェックすると今までどれだけ目視のみで見落としをしてきたか驚くはずです。



用語チェックするのに複雑な用語データはいりません。

「用語の定義」、「対応する訳語（翻訳の場合）」をしっかり決めてさえいれば十分です。SDL Trados のような翻訳支援ツールには用語検証機能があります。また ApSIC Xbench のような用語ツールでもチェックできます（方法説明：http://cosmoshouse.com/utx/tools.htm#xbench）。

用語ツールを使うには、**Excel 用語集を「用語データ」として整理する**必要があります。そこで

UTX を使えば、シンプルな用語データをきっちり作りこめます。

同じ意味でも、どの用語が正しく、どの用語が許容範囲または禁止かしっかり指定できます。そうすると、その用語データを用語ツールで活用し、再利用できるのです。**UTX は、標準化されたデータ構造**なので XML 形式にもスムーズに移行・変換できます。

UTX の詳細・作成法は http://www.aamt.info/japanese/utx/まで

用語集形式 UTX の新バージョン 1.20 が登場！

「用語・訳語がバラバラで本当に困っている……」

UTX なら用語データを用語ツールで
活用して用語統一を簡単に実現。
翻訳コストを下げ、より正確な翻訳を！

標準規格だから

同じ用語データを共有・再利用！
簡単に作成、高度に活用！

新バージョン UTX 1.20 の特徴

よりシンプルに	・最低限の情報だけで用語集を作成できる ・Excelで編集できる扱いやすさはそのまま
より柔軟に	・複数のサブ用語集を一つの用語集にまとめて管理
多言語用語集に対応	・日英中韓など複数言語を一つの用語集で集中管理
移行も簡単	・UTX 1.11の冒頭数行を変更するだけ
より分かりやすく	・仕様書の具体的な実例が充実

UTX 1.20 仕様書で、いますぐ 100 語の基本用語集を作成しませんか？

仕様書・用語データのダウンロード：http://www.aamt.info/japanese/utx/

UTX 変換ツール（UTX Converter）

2015/06/07

UTX 変換ツールとは？

UTX 変換ツール（UTX Converter）は**用語集・翻訳ソフト辞書変換ツール**です。

AAMT（アジア太平洋機械翻訳協会）が仕様を策定した汎用性の高い用語集形式である UTX と、翻訳ソフトなどのユーザー辞書の間で、ファイル形式の変換ができます。

UTX 変換ツールは AAMT が開発を推進しているツールで、オープンソースで公開されており、無償で利用できます。

対応辞書データ形式

UTX	⇔	富士通	ATLAS
	⇔	東芝	The 翻訳
	⇔	クロスランゲージ	PC-Transer/MED-Transer/PAT-Transer/Legal Transer
	→	SDL	MultiTerm（インポート用テキスト）

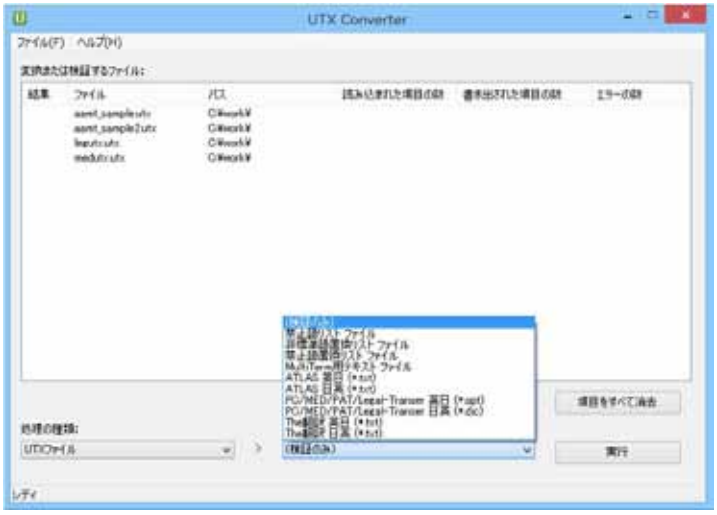
UTX 形式に関する機能

- ◆ **UTX 形式が適切な構造かのチェック**
- ◆ **禁止語の抽出**

  禁止語を一括検索するためのリスト作成
- ◆ **禁止語と承認語のペアの抽出**

  禁止語を承認語に置換するためのリスト作成
- ◆ **非標準語と承認語のペアの抽出**

  非標準語を承認語に置換するためのリスト作成



動作環境: Windows 7 / Windows 8.1 （各 OS 32bit/64bit 版）

==========

■UTX 変換ツール公式サイト

http://utxconv.sourceforge.net/ja/

■UTX 変換ツール（UTX Converter）ダウンロードページ

http://sourceforge.net/projects/utxconv/files/

■用語集形式 UTX の仕様

http://www.aamt.info/japanese/utx/

実務日本語

共有・再利用を意識した電子文書のベスト プラクティス

良いモノづくりはできても文章がだらだらして要領を得ないのはなぜ？

簡潔に要点を絞り込み、会席料理のようにすっきりまとめた日本語文章を書きませんか？

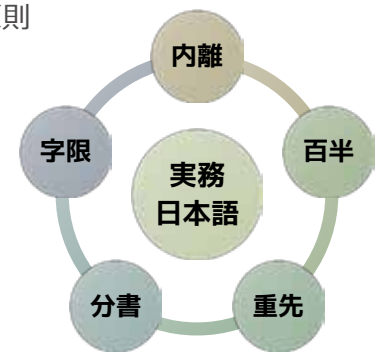
実務日本語とは

■実務文書向けの表記原則

■翻訳用途では

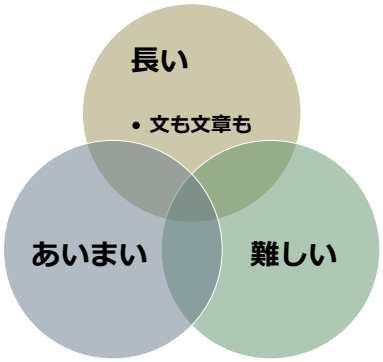
●訳文日本語の改善

●原文日本語の改善



人間にも機械にも読みやすく！

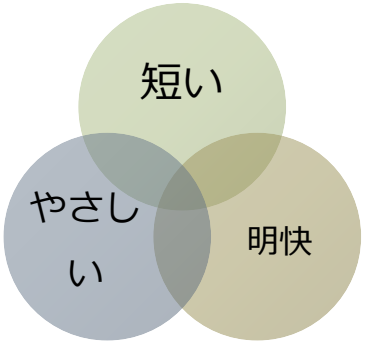
悪文の問題点



相互に関連している

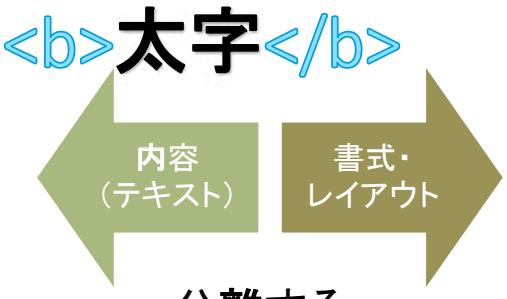
良文の3要素

余分がなく重要な情報が伝わる



1度読めばすぐに理解できる　解釈のブレがない

内離ルール



分離する

実務日本語の詳細はhttp://cosmoshouse.com/jitsumu-nihongo/

複雑な作文ルールを現場で使えますか？

■文章の書き直しには時間とコストが発生する
■時間とコストを抑えるには？

1. 作文ルールをシンプルにする
2. 目視のみでなく必ずツールでもチェックする

どうすればいいのか

■最も実用的で現実的な解決

百半ルールを使う

■百半ルール：
「1文が100字を超えたら半分にする」

■字数を削って100字以下にするのではない

■実質的には**「長い複文を単文にする」**だが、
そのままではルールとして運用しづらい

100字以上の文例

■非特許文献１では、トピック依存性は、デコード処理開始前にデータを集合に分け、その後、前処理パスでソース文の全てによって学習を済ませた分類器により、ソース文のクラスを予測し、予測されたクラスに特定の別々のモデルを用いて、これらの集合を独立してデコードすることによって実現される。

百半ルールを使うと…

■非特許文献１では、トピック依存性は、デコード処理開始前にデータを集合に分ける。（←39字）その後、前処理パスでソース文の全てによって学習を済ませた分類器により、ソース文のクラスを予測する。（←49字）予測されたクラスに特定の別々のモデルを用いて、これらの集合を独立してデコードすることによって実現される。（←52字）

（特開　統計的機械翻訳装置　JPA_2009294747）

文章ルールは、ここまで単純化しないと徹底した実施は難しいかもしれません。低コストで確実に実施できるルールは少ないからです。極論かもしれませんが、100字を超えた文は、なにがあっても一切、文として認めず、金輪際受け付けないようにすれば、良いのではないでしょうか？

実務日本語の詳細はhttp://cosmoshouse.com/jitsumu-nihongo/

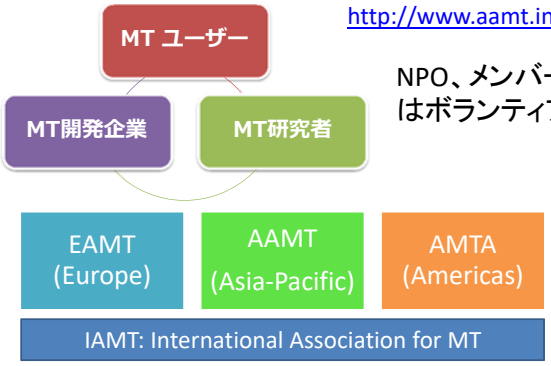
文章と翻訳の品質を改善する―構造化用語データUTXによる用語管理と実務日本語ルール

2016特許情報シンポジウム（AAMT-Japio）
秋桜舎 代表：山本ゆうじ

秋桜舎代表・山本ゆうじ

■言語・翻訳コンサルタント
■AAMT・UTXチーム リーダー
■ISO/TC37（専門用語・翻訳・通訳）エキスパート
■産業日本語研究会委員

AAMT (Asia-Pacific Association for MT)
アジア太平洋機械翻訳協会



http://www.aamt.info/

NPO、メンバーはボランティア

UTXチーム メンバー（順不同・敬称略）

- 山本ゆうじ（リーダー） 秋桜舎
- 秋元圭　合同会社ことばや
- 大倉清司　株式会社富士通研究所
- 加藤マイケル孝仁　ジャパニーズ・グレイツ株式会社
- 島津美和子　東芝ソリューション株式会社
- 村田稔樹　沖電気工業株式会社
- Francis Bond　南洋理工大学（シンガポール）
- 高橋 博之　株式会社クロスランゲージ

新メンバー募集中！

本日の内容

■実務日本語
■用語集形式UTX

実務日本語とは

■実務文書を分かりやすく書くための表記原則

■翻訳用途では
- ●訳文日本語の改善
- ●原文日本語の改善

人間にも機械にも読みやすく！

「悪文」文書の問題点

■正しく検索できない
- ●ウェブでも、イントラネットでも、データベースでも

■翻訳が困難になる
- ●費用が増加する
- ●時間が掛かる
- ●人間にも機械的処理にとっても

複雑な作文ルールを現場で使えますか？

■文章の書き直しには時間とコストが発生する

■時間とコストを抑えるには？

1. 作文ルールをシンプルにする
2. 目視のみでなく必ずツールでもチェックする

文を短く書くにはどうすればいいのか

■最も実用的で現実的な解決

百半ルールを使う

■百半ルール：
「1文が100字を超えたら半分にする」

■**字数を削って100字以下にするのではない**

■実質的には**「長い複文を単文にする」**だが、
そのままではルールとして運用しづらい

内離ルール

太字

内容（テキスト）

書式・レイアウト

分離する

ハード リターンとソフト リターンの問題

ハード リターン

- 実際の改行【これ→】↵
- WordなどでのEnter

ソフト リターン

- 見かけ上の改行【これ→】↓
  WordなどでのShift+Enter

特許文書の例

「前記翻訳装置は、¶

定期的に、ネットワーク上の複数のウェブページの翻訳を、少なくとも予め記憶している辞書と文法規則を参照して行なう翻訳部と、¶

前記翻訳部が翻訳した翻訳結果を格納しておく格納部と、¶

前記クライアントが行なう、所定のウェブページの翻訳結果の要求を受信する要求受信部と、¶」

【公開番号】特開2007-334905
https://www7.j-platpat.inpit.go.jp/tkk/tokujitsu/tkkt/TKKT_GM301_Detailed.action

特許文書では

ハード リターンでは困る

- ーセンテンスが分かれてしまう

ソフト リターンでも困る

- 文が長いまま

根本的には文を短くするしかない

今後の問題提起

1. 原文をどうしたらよくできるか
2. 作文講習、規則施行をどう行うのか？
3. MTのベースになる人間翻訳の品質をどう上げるか

構造化用語データ UTX用語集形式

体系的翻訳と用語集

体系的翻訳

■企業では大量の多言語文書をすばやく高品質に翻訳する必要がある（精訳）

■**翻訳ツール**や**用語管理**で翻訳工程を改善する

■日本では体系的でないことも多い

用語ベース=用語データベース

ヨーロッパの用語集の例：IATE

■IATE (InterActive Terminology for Europe)（EU）

■**1300万項目、800万語、24か国語の用語集**

■商業用途を含め無償

■TBX形式

アメリカの言語資産の例：MeSH

■Medical Subject Headings (MeSH)

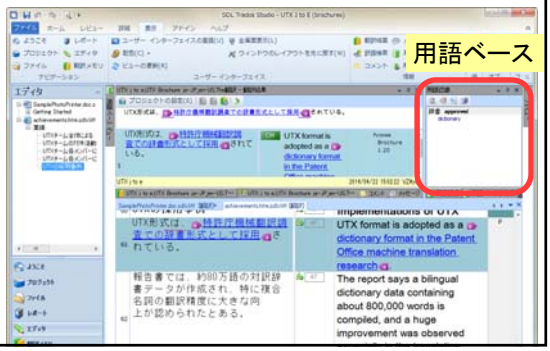
■U.S. National Library of Medicine

■**27,883 の件名見出し**

■**87,000 entry terms**

翻訳支援ツール（翻訳メモリー）の実例

■手作業の翻訳では効率が悪く、間違いに気づかない



構造化用語データ UTX用語集形式とはなにか

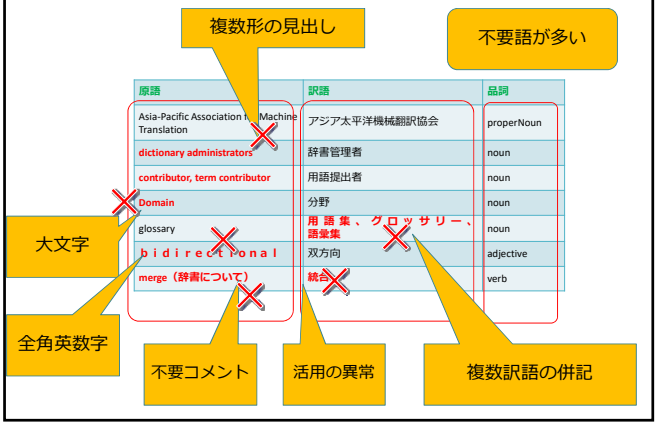
問題：用語集は適切に活用されていない

1. 用語集の管理者や管理方法が不明確
2. 用語集の構造が複雑
3. 用語集の形式がバラバラ
4. 不適切な語が多数混入

文書や翻訳を読みやすくするには用語集は不可欠

分かりにくい用語を禁止する	分かりやすい用語を一貫して使う
• 褥瘡 • 増悪させる	• 床ずれ • 悪化させる

構造化・標準化されていないExcel用語データには問題が多い



構造化されていない用語集

1. Excelのみ、または特定のツールでしか使えない
2. データを転用・変換・共有・再利用できない
3. 翻訳言語方向を反転できない

UTX用語集形式とは?

AAMTが策定したシンプルな用語集形式

専門用語を管理する用語集の作り方のルール

Excelなどでも編集できるタブ区切り形式

UTXは「構造化用語データの形式」

1. 明確に定義された構造を持つ
2. さまざまなシナリオに対応できる
3. データとしてすぐに利用できる
4. 他の形式に変換できる
5. さまざまなツールで使える
6. 翻訳方向を反転しても使える

誤解にご注意

■UTXは用語データそのものではない

■UTXは「用語作成のルール」、「専門用語データの標準規格」

UTX用語集形式の特徴

シンプル	• 表計算形式で作りやすく管理しやすい
信頼できる	• 用語ステータスで品質管理
複数形式と相互変換	• 翻訳ソフトや用語ベースなど
双方向・多方向	• たとえば英日と日英、また多言語を1つの用語集で管理できる
共有・再利用しやすい	• 構造化・標準化されており無料で使用できる

文書に埋もれた用語知識を「見える化」する

一覧できる**表形式**と定義された**用語ステータス**で
データとしての形式を整え、さまざまな環境で共有・再利用できるようにする

#UTX 1.11; en-US/ja; 2016-04-01T19:00:00+09:00; copyright: AAMT (2014); license: CC BY 4.0			
#原語	訳語	品詞	用語ステータス（省略可）
#src	tgt	src:pos	term status
Asia-Pacific Association for Machine Translation	アジア太平洋機械翻訳協会	properNoun	approved
contributor	用語提出者	noun	provisional
optional	省略可能	adjective	approved
optional	オプショナル	adjective	forbidden
merge	統合する	verb	approved
merge	マージする	verb	forbidden
unidirectional	一方向	adjective	approved
monodirectional	一方向	adjective	non-standard

拡張可能

安心して使える用語か、禁止用語か区別できる

二言語双方向のUTX用語管理

■日英・英日用語データを一つの用語データで管理

■用語ステータスと概念IDで
1:n、n:1、n:nの対応関係を記述できる

多言語多方向の用語管理
多言語UTX用語データの例

アメリカ英語／日本語／中国語（簡体字）

#UTX 1.20; lang: en-US/ja/zh-CH; copyright: AAMT (2016); license: CC BY 4.0					
#term:en-US	tgt:ja	term status:en-US	term status:ja	concept ID	term:zh-CH
#用語（アメリカ英語）	用語（日本語）	用語ステータス（アメリカ英語）	用語ステータス（日本語）	概念ID	用語（中国語簡体字）
Asia-Pacific Association for Machine Translation	アジア太平洋機械翻訳協会	approved	approved		亚洲太平洋机器翻译协会
contributor	用語提出者	approved	provisional		用语提交者
entry	項目	approved	approved	1	条目
entry	エントリー	approved	forbidden	1	
unidirectional	一方向	approved	approved	2	单向
monodirectional	一方向	non-standard	approved	2	

UTXの用途の例

- 書き換え置換用の日本語用語集として使う
- 類似概念の用語をまとめて検索しやすくする
- 機械翻訳の精度を向上する
- 人間・機械翻訳のチェックに使う

UTXの実例：デンソー様の技術開発センターDP-EDA改革室

■DP-EDA改革室（DENSO Project - Electronic Design Automation）

■用語整理を進めるために、1093語の用語集をUTX形式に変換

■SDL Trados用語ベースなどに変換して活用

UTXの実例：特許庁が220万語中日辞書をUTX形式で作成

「『中日対訳辞書データ』を機械翻訳辞書に追加することにより、用語（名詞）の翻訳精度に関して一定の向上効果が得られることが確認できた。」

特許庁「平成24 年度 中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査 調査報告書 概要版」

https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/kikai_honyaku/h24.pdf

UTX用語データで機械翻訳を改善する

ニューラル機械翻訳の課題：用語レベルで不正確（用語の欠落）

■原文：”AAMT created its first version of the specification, UPF (Universal PlatForm), with support from IPA (Information-technology Promotion Agency, an institute in Japan) in 1995.”

■訳文「AAMTは、1995年にIPA（日本の研究所）の支援を受けて、UPF（Universal PlatForm）という仕様の最初のバージョンを作成しました。」

Googleニューラル機械翻訳で英日翻訳

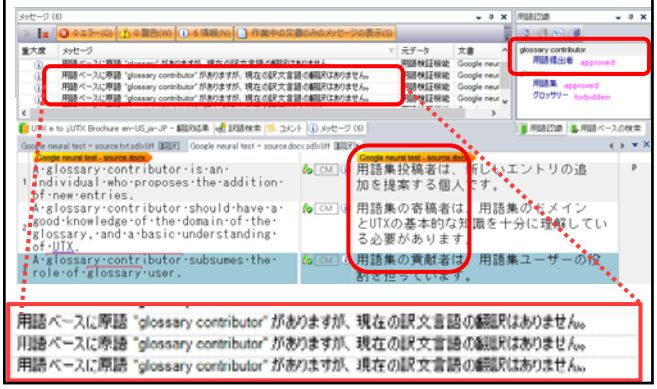
ニューラル機械翻訳の課題：用語レベルで不正確（一貫性がない）

■原文：”A glossary contributor is an individual who proposes the addition of new entries. A glossary contributor should have a good knowledge of the domain of the glossary, and a basic understanding of UTX. A glossary contributor subsumes the role of glossary user.”

■訳文「用語集投稿者は、新しいエントリの追加を提案する個人です。 用語集の寄稿者は、用語集のドメインとUTXの基本的な知識を十分に理解している必要があります。用語集の貢献者は、用語集ユーザーの役割を担っています。」

Googleニューラル機械翻訳で英日翻訳

翻訳支援ツールでの用語チェック



機械翻訳と人間の精訳の差はなにか

■技術翻訳では自然さより用語が重要

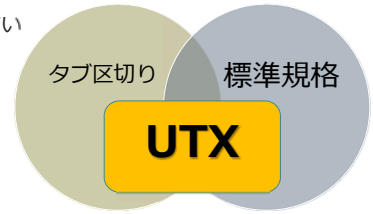
■文書と用語が不一致

「専門用語」や固有名詞を正しく訳すには
管理された構造化用語データが必要

「標準化されたタブ区切りデータ」の利点

■タブ区切りの利点
- ●Excelで一覧編集できる
- ●XMLのようなタグがないのでデータが軽量

■標準化形式の利点
- ●共有・再利用しやすい
- ●ツールを活用できる
- ●再調整が不要



ニューラル機械翻訳での用語データ活用

■用語レベルでの正確性が保証されない

■UTXの用語データで精度を向上できる

UTXデータで日本語を読みやすくする

文書を読みやすくするには用語集は不可欠

分かりにくい用語を禁止する	分かりやすい用語を一貫して使う
• 褥瘡 • 増悪	• 床ずれ • 悪化

コーパス表記のゆれ⇔用語の一貫性

インターフェイス	interface
インターフェース	interface
インタフェース	interface
インタフェイス	interface

同じ概念

検索や翻訳チェックに支障

統一するには用語集が必要！

cf. Fターム

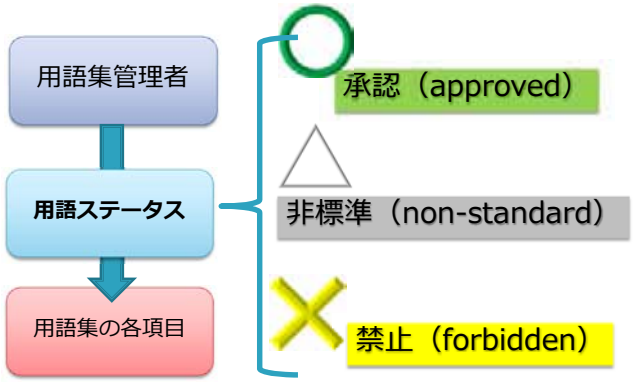
概念IDで類義語をまとめる

■既定の原語および訳語

■原語および訳語のバリエーション（同義語≒別訳語）

英語	日本語	用語ステータス	概念ID
term:en	term:ja	term status:ja	concept ID
bedsore	褥瘡	forbidden	1
bedsore	床ずれ	approved	1
aggravate	増悪させる	forbidden	2
aggravate	悪化させる	approved	2

用語ステータスで「信頼」を共有する



UTXの「既定訳語」（承認語）の意義

既定の訳語がないTBXでの同義語の組み合わせ
3x6=18ペア

English	French
artificial language	langue artificielle
artificial language	langage artificiel
artificial language	langue formelle
artificial language	langage formel
artificial language	langage formalisé
artificial language	langue formalisée
formal language	langue artificielle
formal language	langage artificiel
formal language	langue formelle
formal language	langage formel
formal language	langage formalisé
formal language	langue formalisée
formalized language	langue artificielle
formalized language	langage artificiel
formalized language	langue formelle
formalized language	langage formel
formalized language	langage formalisé
formalized language	langue formalisée

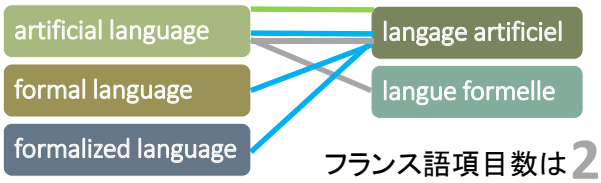
UTXでapprovedを指定した場合
3+6-1=8ペア

English	term status:en	French	term status:fr
artificial language	approved	langue artificielle	approved
artificial language		langage artificiel	non-standard
artificial language		langue formelle	non-standard
artificial language		langage formel	non-standard
artificial language		langage formalisé	non-standard
artificial language		langue formalisée	non-standard
formal language	non-standard	langue artificielle	
formalized language	non-standard	langue artificielle	

同義語が18→8ペアに減少

表形式でもn+n-1項目

既定訳語ペア

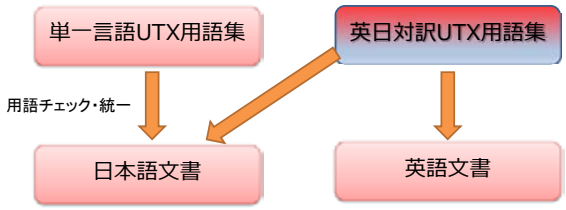


フランス語項目数は 2

英語項目数は 3

UTXは 日本語文書も翻訳も改善できる

- 「同じ概念（別訳語）」を概念IDでまとめられる



同義語の中の禁止語と承認語を区別できる

UTX用語集で原語が同じで訳語が異なる場合

term:en	term:ja	term status:ja	yomi:ja
bedsore	褥瘡	forbidden	じょくそう
bedsore	床ずれ	approved	とこずれ
aggravate	増悪させる	forbidden	ぞうあくさせる
aggravate	悪化させる	approved	あっかさせる

置換リストとして抽出できる

置換前	置換後
褥瘡	床ずれ
増悪させる	悪化させる

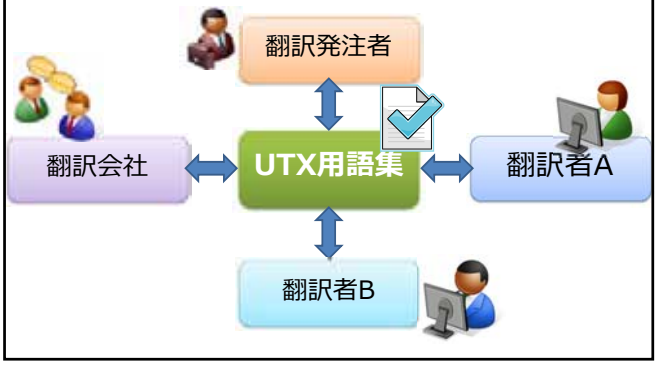
置換して
分かりやすい用語に統一できる

用語ツール＋UTXの活用シナリオ

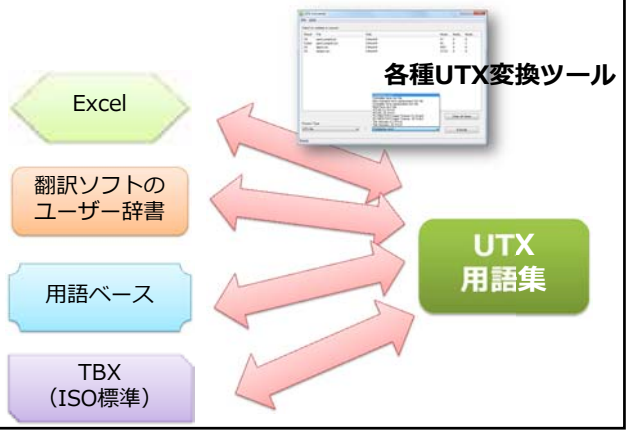
1. 承認語（分かりやすく適切な語）が使われているかチェックする
2. 同じ語が一貫して使われているかチェックする
3. 禁止語を検出し承認語に修正する

UTXデータを様々な形式で共有・再利用する

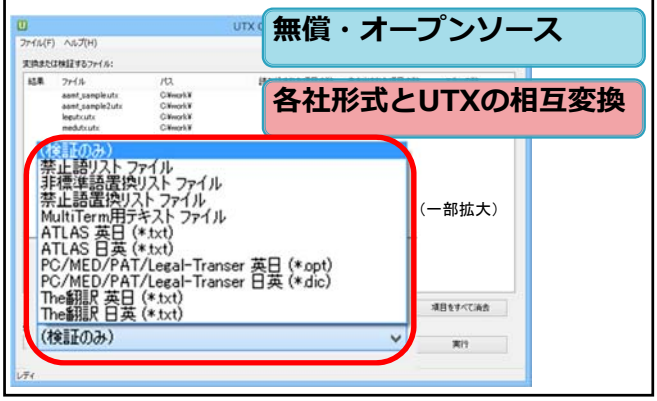
標準規格なのでユーザー間で用語集を共有・再利用しやすい



UTXと各種形式の相互変換

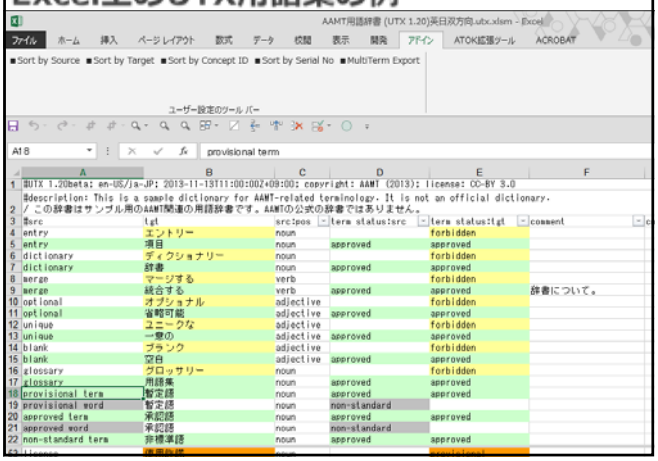


公式UTX変換ツール

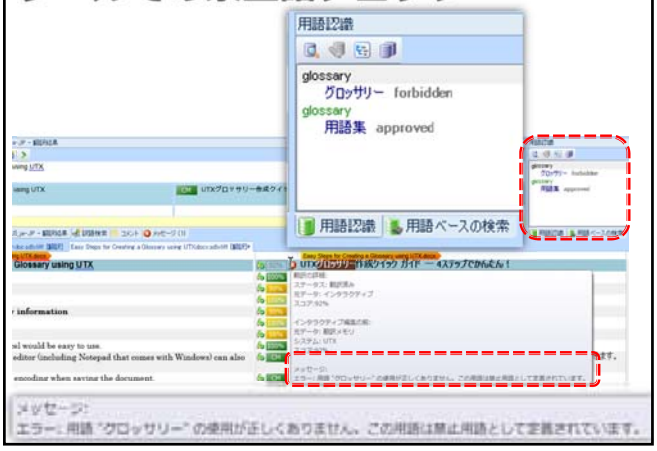


各社の市販翻訳ソフトで使用できる

Excel上のUTX用語集の例



ツールでの禁止語チェック



まとめ・用語集形式UTXの今後

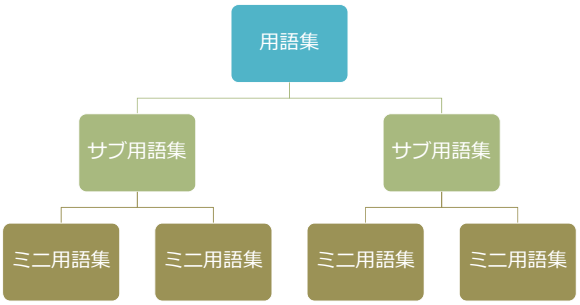
ISOとの連係

1. UTXはISOのTBX形式のグループと連係
2. TBXは高機能だが複雑
   1. →シンプルなUTXとすみわけできる
3. 変換ツールで相互に変換可能
4. 新バージョンUTX 1.20ではTBXとの互換性が向上

UTX ⇔ TBX

断片化した用語データを統合して オープン データ言語資源として活用する

データ構造が標準化されているのですぐにまとめられる



提言

1. 専門用語データを共有・再利用する仕組みが必要
2. 「用語ファーム」でオープン データの構造化用語資産を作り育てる（UTX形式）

詳細は

UTX 用語で検索

http://www.aamt.info/japanese/utx/

■無料用語集や、用語集を作成するための仕様がダウンロードできる

ご清聴ありがとうございました

一般講演 1

「課題とその対象を軸としたマトリクス型特許マップの自動生成手法の提案」

課題とその対象を軸としたマトリクス型特許マップの自動生成手法の提案

小野寺 大輝　吉岡 真治

北海道大学 情報科学研究科
onodera@kb.ist.hokudai.ac.jp yoshioka@ist.hokudai.ac.jp

概要

特許マップとは、特定の分野の特許の出願状況を整理したものであり、その分野における特許の現状を把握するために活用される。この特許マップには、様々なタイプのものがあるが、本研究では、特許の効能を示す表現（特長表現と呼ぶ）が、技術的な課題 (生産性や使い勝手) と、その課題を解決するために注目する対象 (エンジンや操作パネル) の組み合わせを用いて表現される（操作パネルの使い勝手を向上する）ことが多いことに注目し、課題と対象を軸としたマトリクス型特許マップの自動生成手法を提案する。そのために、文書を複数のトピックの組み合わせとして表現可能なトピックモデル (LDA) を拡張して、特定の観点（facet）の単語群に関するトピックという考え方を導入した facet-based LDA(fbLDA) を提案し、特許マップ作成に利用する。具体的には、特許文書が、技術的課題に関する観点の語句に関するトピック、構成要素などの対象に関する観点の語句に関するトピック、その他一般の語句に関するトピックの組み合わせから構成されると考えたトピックモデルを作成し、各々の特許と関連の深い技術的課題と対象のトピックの情報を用いて、各特許を技術的課題と対象のマトリクス上に配置することにより、特許マップを作成する。本システムの有用性を検証するために、2001,2002 年の公開特許公報に基づいて、主に IC タグの分野の特許を分析した結果について紹介する。

1　はじめに

特許には、様々な技術課題とその解決方法などが示されており、それぞれの企業や大学等の研究開発における知的財産の保護を目的に毎年多数の特許が出願されている。西山ら [5] は、個々の特許が「操作パネルの使い勝手を改善する」や「エンジンのノイズを抑える」といった一般的な技術課題の改善あるいは低減といった形で効能が示されることが多いことに注目し、これらを特長表現と呼び、文パターンを用いることにより、自動抽出を行う手法を提案している。我々は、この特長表現の多くが、技術的な課題 (使い勝手やノイズ) と、その課題を解決するために注目する対象 (操作パネルやエンジン) の組み合わせを用いて表現されていることが多いこと、また、一つの対象が複数の技術的な課題に関連するとともに、一つの技術課題に対する解決方法が、注目する対象によって特徴づけられることに注目した、課題とその対象に注目したマトリクス型特許マップを作成するプロトタイプ手法を既に提案している。[4] しかし、上記の手法においては課題語と対象語のクラスタリング手法の検討が十分ではなかった。そこで、本研究では多数存在する課題と対象のクラスタリングの手法として、文書を複数のトピックの組み合わせとして表現可能なトピックモデル (LDA) を利用し、特許文書が課題とその関連語のトピック、対象とその関連語のトピック、一般の語のトピックの組み合わせから成っていると仮定

し、facet-based LDA(fbLDA) を提案する。本モデルを用いることにより、課題と対象に関する語が同時にクラスタリングされ、共起語の情報も利用することで、まとまりのあるクラスタを作成することが可能になる。また、本システムを用いて 2001,2002 年の公開特許公報を分析した結果について実際に作成した特許マップと作成されたクラスタの評価を行った。

2 課題と対象に注目した特許マップと特許における特有の表現

2.1 技術動向把握のための特許マップ

特許マップは目的に応じて様々な可視化の方法が提案されており、各企業がどんな対象に対しての特許をどれくらい持っているかという特許マップや、マトリクス型特許マップとして軸要素を技術課題と解決方法としたものも存在する。特許マップの形態はこれらの特許マップ以外にも多く存在し、特許情報から読み取りたい情報によって様々な形となる。我々は既に特長表現 [5] の多くに、評価の基準である観点（例えば安定動作、操作性）と、それをどの部分により実現するかを示した対象（例えば液晶パネル、LED 照明）の 2 つを含むことから、この組み合わせによる特許マップの生成手法を提案してきた [4]。この手法では、文書中からパターンを用いて対象語と観点語を収集し、それらの単語のクラスタリングを別々に行うことで、特許マップの作成を行う。このクラスタリングを行う方法として、対象語と観点語の各々について、それらの語を含む文を集めた代表文書に対して LDA によるクラスタリングを行っていた。しかし、クラスタリングの手法の検討が不十分であり、クラスタリングの対象となる文書の作成方法も恣意的であったため、本研究ではクラスタリング手法として fbLDA を提案し、対象語と観点語に関する文書を別々に新たに作成して LDA を適用するのではなく、対象語と対象語に関連する語からなるトピックと観点語と観点語に関連する語からなるトピック、それ以外の一般トピックを作成し、対象語や観点語ではない一般語の共起情報も用いたクラスタリングを行う。なお、観点語については技術課題であることがほとんどであるために、課題語という表現に変更した。このような技術課題と対象を軸とした特許マップによって集中的に開発されている対象とその技術、逆に注目されていない技術課題やその対象を発見できるようになることが望まれる。

2.2 特長表現抽出のための記述パターン

特許マップを作成するために特許から特長表現 [5] と呼ばれる表現を見つけ出すために特許の請求項から明細書に至るまで全ての情報を用いると周辺技術の情報など他の情報の存在によって特定の特許の固有性が減少し、上手く特徴を捉えられなくなる可能性がある。そこで [4] においても研究の全体像が示される明細書中の、従来技術との差異が強調して書かれるであろう発明の効果の箇所を解析の対象としている。発明の効果から向上・改善された当該技術の新たな長所を抽出することで、特許特有の長所がまとまった特許マップになると考えられる。上で述べたように特長表現は当該技術の新たな長所を表した表現であり、例えば特許中に携帯電話の操作性を向上させたという表現や製造時の変形を防止することが可能となったなどの表現がある場合は特長表現として抽出されることになる。これらの表現の特長として～を向上させる、～を高める、～を抑制するなどのパターンを用いて記述されることが多いため、本研究ではこれらに当てはまるパターンを抽出していた。具体的には以下のパターンを用いている。

- ・[対象] の [課題] を向上
- ・[対象] の [課題] を高める
- ・[対象] の [課題] を抑制
- ・[対象] の [課題] を防止
- ・[対象] の [課題] を低減

これらのパターンで抽出される課題と対象をマトリクスマップとするためにトピックを考慮したクラスタリングを行うことで、似た意味を持つ単語がまとまったクラスタとなると考えている。

3　facet-based LDA による発明効果の分類

前節におけるパターン検索の結果として多くの課題と対象が発見できるが、そのままマトリクスマップを作成するとほとんどの値が 0 となってしまい、視認性の悪い特許マップになる。そこで本研究で提案する facet-based LDA を用いて単語をクラスタリングすることで似た意味をもつ単語が一つのクラスタを形成し、それらのクラスタによってマトリクスマップを構成することができれば視認性の良い特許マップが作成できると考えた。

3.1　Latent Dirichlet Allocation における分類

LDA(Latent Dirichlet Allocation)[2] は文書の確率的生成モデルの一つであり、図 1に表されるような階層ベイズモデルとしてモデリングされている。LDA では一般的に文書が複数のトピックの組合わせにより構成されていると考え、更にそのそれぞれのトピックから一定の確率分布に従って単語が生成されると考えている。一般的に LDA においては各文書におけるトピック分布、トピックごとの単語の分布についても更にディリクレ分布を仮定する。つまり、トピック分布や単語分布が解析的に求められない、正しいパラメータを実験的に求める必要がなく、仮定したディリクレ分布のパラメータを尤度に基づいて計算することで求めることができ、更にこのディリクレ分布で得られたパラメータで定まるトピック分布、単語分布が得られる。更にこのトピック分布、単語分布から各単語がサンプリングされると考えられ、LDA ではトピック分布と単語分布として多項分布を考えているために、分布の共役性によって MCMC や変分ベイズ法による近似計算が LDA において行われている。このようにディリクレ分布を仮定することによって幾何学な解釈として文書を各トピックと各単語をノードとする単体上の空間の一点として表すことができるので、Latent Driichlet Allocation と呼ばれている。一般に似た文書はこの空間上で近い位置に配置されることになるため、LDA を文書群に適用すると、単語の共起性に基づいた単語のクラスタリングを行うことが可能となる。図 2はオリンピックの文書に LDA を適用することを考えたときのイメージ図である。複数のトピックの組合わせとして記述されると考えられる文書においては LDA はそのトピックごとに単語の確率分布を適切に定めると考えられている。

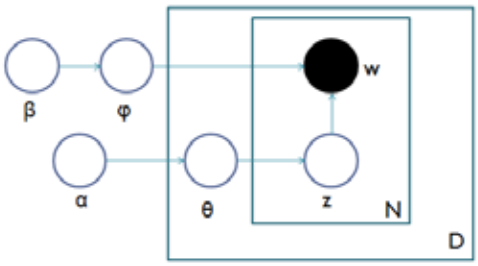


図1　LDA のグラフィカルイメージ

図2　オリンピックの文書について LDA を適用した例

3.2　facet-based LDA の提案と特許文書群への応用

本研究では、前節で述べた LDA における文書がトピックの組み合わせにより構成されているという考え方を更に拡張し、特許文書を対象・課題という二つの軸から分析するために、主に、対象に関連するトピック、課題に関連するトピックといったトピックのタイプを考え、その組み合わせにより、特許文書が作成されると考えたトピックモデルを考える。このように、分類したい軸に応じたトピックを考えることにより、一つの特許文書に対応する主な対象に関連するトピック、課題に関連するトピックの情報を用いることにより、特許がマトリクス上に配置可能となり、 1 回のトピックモデルの構築で、対象と課題について、同時にクラスタリングが行われることになる。特許におけるこのような対象や課題を分析のための観点 (facet) であると考え、提案

するトピックモデルを facet−based LDA(fbLDA) と呼ぶ。fbLDA では、通常の LDA と異なり、トピックを各観点に関連するトピック群、一般のトピック群といった形に分けて、トピックを作成する。具体的には、各観点に対応する語群を情報として与え、特定の観点に属する語は、その観点を表すトピック群の中のトピックからのみ生成され、他のトピック群からは生成されないという制約を導入する。この制約を導入することにより、各々の観点に属する語を持つ文書は、必ず、対応する観点に関連するトピック群の中からトピックを選択する必要が生じるため、結果として、各々の観点に属する単語を全て持つような文書は、各観点を代表するトピックの混合として表されることとなる。観点に関連するトピックを作る方法としては、観点に属する語のみで生成する方法も考えられるが、観点に関する語が少ない場合に、それらの語の間の共起情報だけでは、まとまりのあるトピックを作成するのが困難なため、観点に関する語と特徴的に共起するような語の情報も用いてトピックを生成することとした。表 1に二つの観点を持つ場合のｆｂ LDA における各トピックのタイプと、語のタイプにおける単語の発生確率に関する制約を示す。 なお、表中の 0 については 0 に

表1　発生確率に関する制約

	課題語	対象語	一般語
課題トピック	推定値	0	推定値
対象トピック	0	推定値	推定値
一般トピック	0	0	推定値

非常に近い値を代入し、推定値については変分ベイズ法によって求まる値である。つまり、課題語は課題トピックからしか生成されず、対象語は対象トピックからしか生成されない、一般語についてはどのトピックからでも生成されるという制約になっている。この表からわかるように、fbLDA では、一つの観点に関するトピックを作成する際には、他の観点に属する語の共起情報を用いないため、fbLDA では、異なる観点の語に関するクラスタリングは、基本的には、独立に分類が行われることになる。特許マップの事例では、対象と課題のトピックは、各々独立に作成され、両者の関係は、マトリクスマップ上で確認されることとなる。一般的な LDA を用いると、この独立性が考慮されないために、数が多い交点に対応するような代表的な特許群に対応するトピックが生成されることが想定されるため、その結果に基づいて、後処理で対象や観点のクラスタリングを行うよりも、有用なクラスタリングが行われることが期待される。

本研究では特許文書群が課題トピックと対象トピック、それ以外の一般トピックから構成されると考え、LDA を目的に合うように改良した形で特許文書群に適用することを考える。上記で述べたように特許群に対しても LDA を適用することで特許文書群中の単語をトピックに分割し、特許マップ作成に役立てるために、対象を中心としたトピック、課題を中心としたトピック、その他一般トピックを作成し、作成されたトピックをマトリクスマップの軸要素として配置したい。更にこのトピックのトピックという概念、つまり今回で言えば対象と課題がファセットと呼ばれ、このファセットに基づいてトピックが形成されるために facet-based LDA とした。しかし、LDA においてはトピックの種類の情報は与えられず、偶然適切なトピックが作成されることも考えられるが、本研究では制約を導入することによって確実に対象トピック、課題トピックが作られるようにした。図 3に本研究で提案する facet-based LDA のイメージを表した。はじめにパターン検索によって得られた課題語と対象語のリストを作成し、特許文書群を課題トピック、対象トピック、その他一般トピックから構成されると仮定して制約を導入することで、課題トピックには課題語と課題語と関連のある語が分布の上位に、対象トピックには対象語と対象語と関連のある語が分布の上位に来るようにしており、それらの作成された課題トピック、対象トピックのラベル付けを手動で行うことによって、それらをマトリクスマップの軸要素とし、実際の要素としては課題と対象が現れた回数が記録されることとなる。各作業の詳細を以下に示す。

1. 抽出された課題と対象語のリストをそれぞれ作

成する

2. 制約を導入するために課題トピックからは対象語が生成されないように、課題トピックの対象語生成確率に 0 に近い値を代入し、対象トピックからは課題語が生成されないように、対象トピックの課題語の生成確率に 0 に近い値を代入し、一般トピックには課題語、対象語がともに生成されないように 0 に近い値を設定する。
3. 得られたトピックの上位語に特徴的な語が現れるので、上位 7 8 件程度を中心としたトピックと考え、それらのトピックのラベルを要素を見て手動でつける
4. ラベルをつけたトピックをマトリクスマップの軸に配置し、要素としてトピックに含まれる課題語と対象語のペアの出現回数を設定し、マトリクスマップとする。

4 実験結果と考察

4.1 実験データについて

使用する特許データは，国立情報学研究所で作成された NTCIR-5 PATENT[3] の公開特許公報全文データ中から，国際特許分類 (IPC)「G06K 19/07」(主に IC タグ) 分野の特許 1972 件 (2001 年〜 2002 年) を用いた。facet-biased トピックモデルの実装には [2] の著者である blei 氏が公開している LDA の implement である lda-c を利用させていただいた。パターンによって獲得できた課題語、対象語、その他一般語の数は表 2のようになった。この出現回

表2 それぞれのファセット語と一般語の出現回数

	語彙数	出現回数
対象語	80	30,312
課題語	140	215,337
一般語	20117	428,300

数の比率を考慮し、トピック数をそれぞれ課題トピックを 10, 対象トピックを 10, その他一般トピックを 80 トピック、全てで 100 トピックとして fbLDA の計算を行う。実際に得られた対象語と課題語のリストの一部を表 3に示す。

表3 対象語と課題語のリストの一部

対象語	課題語
IC チップ	強度
IC タグ	再利用
非接触型 IC カード	セキュリティー
万引き防止システム	信頼性
情報記録担体	利便性
製造	歩留まり

4.2 実験結果

対象語と課題語のリストを用いて対象トピック、課題トピックを作成した結果を表 4と表 5に示す。

表 4は対象トピックについてのクラスタリングの結果であり、最後のトピックについては不明トピック、つまり名前をつけることが難しいトピックが作成されてしまったが、他のトピックについては概ね似た意味の単語が集まったクラスタを作成することができたと考えられる。ラベルは手動でつけたが、ラベルの粒度として非接触カードと IC を使ったサービスなど情報の粒度が異なる分類名になってしまっており、当初として想定していたラベル群とは少し異なる結果となった。想定していたラベル群と異なる結果となった原因については特許の特性に理由があると考えており、特許は基本的に基礎技術の特許とその周辺の応用技術の特許が存在する。当然基礎技術においては例えば根本的な IC カードの設計であったり、製造技術の新たな手法であったりするが、応用技術の特許については IC を使ったサービスに書かれているようにこれまでの既存の技術を車内精算のシステムに応用したり、クレジットサービスなどに応用した事例も特許として現れる。その結果として基礎技術と応用技術の特許がそれぞれ混ぜ合わさっている。更に似たような単語が異なるクラスタに含まれてしまった問題については、共起が似た意味を持つ特許同士で起こっているのではなく、特許を記述する弁理士などに依存しているようにも見受けられた。特に対象については広めに請求項を獲得するために抽象的な文言を用いることが非常に

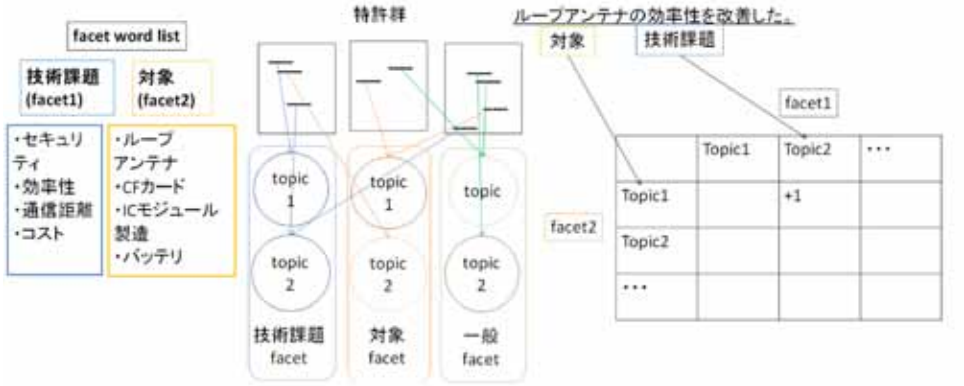


図3　facet-based LDA のイメージ

多く、審査の結果が不合格であれば請求項を修正して狭めの請求範囲に変更するなどといった出願手法も考えられるため、特許の抽象度は特許ごとに当然異なる。更に抽象的な単語レベルで見ると同じ意味の単語が多く出現する。特に非接触型の IC カードについては非接触 IC カード、非接触情報担体、非接触情報メディア、非接触型情報メディアなど、表記揺れだけでなく、非常に多くのパターンが存在し、これらを統一的に扱うためには表記揺れ獲得手法の他に異なる弁理士などによって記述された同じ意味の単語を認識する方法が新たに必要となる。更に正規表現によるパターン検索においては、パターンのみである程度の抽出精度はあるものの、表を参照すると接続部など単語だけでは意味を認識できない語も含まれてしまっている。表 5は観点トピックについてのクラスタリングの結果である。表 2に見られるように対象語と課題語の語彙数は 2 倍程度、課題語の方が多く、出現回数に至っては約 7 倍の差で課題語の方が多い結果となっている。今回の抽出結果としては一つの対象語につき 7 個程度の課題語のセットが存在したと考えられるが、パターン検索においては対象語に比べて課題語の検索は非常に難しいと考えられ、対象語は複合名詞などで品詞で絞り込むことが可能であるが、課題語は収集したデータをログデータとして蓄積できるようになった、などの表現は単なるパターン検索では獲得が難しく、今回のパターン検索で抽出できた課題語は語尾が～性やなどで表されるような簡単な語句しか獲得できなかった。表 2を参照すると、特に信頼性やセキュリティ、生産性などのクラスタにおいては非常に良いまとまりの良いクラスタを作成することができたが、物理的強度のように同じようなラベルを持つクラスタも作成されてしまった。対象語、課題語共に共起する語が似ているが故にうまく形成することができたクラスタもあれば、単語によっては本来の意味とは異なった書き手による単語の類似性で共起性が判断され、うまく分類できなかった事例もあると思われる。総じてトピック数の設定を特許マップの視認性を考え、10×10 で作成したが、特許の分野ごとにキーワードの数が変わることが想定されるために、トピック数のパラメータ調整を行う必要があった。図 4に本研究で作成した特許マップを示す。表 4と表 5においてラベル付けをしたトピックが軸要素に配置され、マトリクスの要素としてはそのクラスタに含まれる課題語と対象語のペアの出現回数となっている。図 3にあるようにループアンテナの効率性を改善したという特長表現でループアンテナが対象語で効率性が課題語として獲得できていた場合を考える。このときループアンテナが対象トピックのトピック 1 に、効率性が課題トピックのトピック

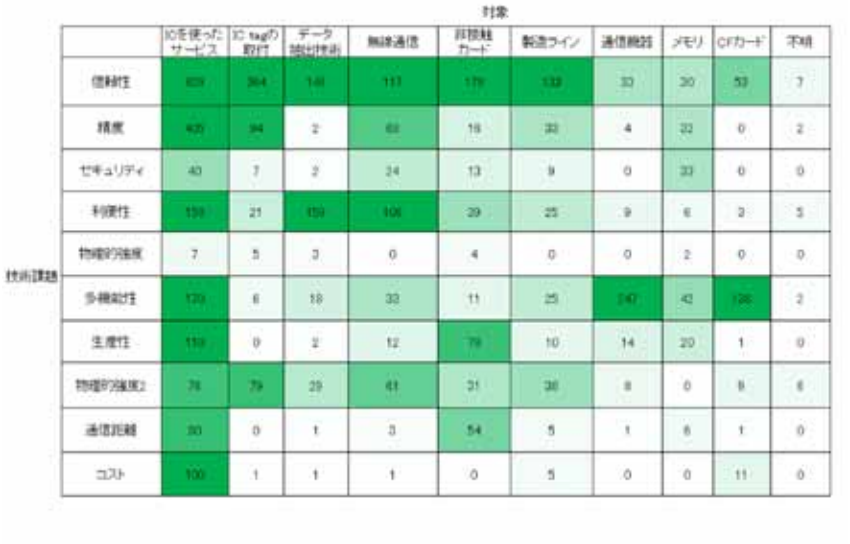


図4　本研究で作成した特許マップ

2 に fbLDA によって割り当てられたとすると、マトリクスの (1,2) 要素に +1 が加算されることになる。この特許マップ上では特に信頼性に関する発明が 2001 年〜 2002 年で多くされていたことがわかる。そして IC を使った応用についての発明も盛んなことから、2001 年〜 2002 年時点では IC の基礎技術に関する発明は既に多くなされて特許出願されており、この年代ではデータや通信の信頼性の向上及び、IC 技術を様々な場面に応用する試みが多くなされていたのかもしれないという推測も可能になる。このような技術動向の把握は分野の把握や先ほど述べたような集中的に発明されている箇所や見落としがちだが重要な箇所を発見できる可能性があると考えられる。しかし、技術動向を把握するためには時系列で特定の技術に対する発明がどのように変遷しているのかなどといった情報も重要になるため、時系列での把握が可能な特許マップが必要となってくる。

5　まとめと今後の課題

本研究ではマトリクス型特許マップとして当該技術の新たな長所としての特長表現と更にそれを分割した課題と対象によって構成される特許マップの自動生成手法を提案した。今後の課題としてはまずデータについては、発明の効果文から適切な共起情報が得られていないと考えられる可能性もあるために請求項を用いて分析をすることも考えられる。更にこれまでパターン検索で行なっていた課題語と対象語の抽出も、名詞や助詞の連接になっている課題語の抽出や、その抽出語の類似度計算など、fbLDA などで処理をする前の段階で改善できる部分も多く存在する。更に fbLDA においては LDA の拡張として時系列トピックモデルである Dynamic Topic Model[1] が提案されており、fbLDA においても時系列の拡張を行うことで時系列に技術動向を把握することのできる特許マップが作成できる可能性がある。

謝辞

本研究の一部は、科研費 25280035,26280111 の支援を受けて実施した。

参考文献

[1] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*,

表4　対象トピック

クラスタ	単語
IC を使ったサービス	車内精算システム, 非接触型 IC メディア, IC メモリーカード, クレジットサービス, 暗号表示カード
IC タグの取り付け	IC タグ, 製造工程, 半導体装置, RFID タグ, 半導体チップ
データ抽出技術	データ, データ抽出, IC メモリカード, システム, 情報
無線通信	無線通信カード, 情報処理システム, 非接触データ送受信体, 物体, 情報
非接触カード	IC モジュール, 非接触 IC カード, 利用者,　万引き防止システム, 非接触型記録メディア
製造ライン	工程, 電子部品, パッケージ, 製造工程, 接続部
通信機器	アンテナ, 無線通信カード, 非接触 IC カード, 非接触データ送受信体, アンテナコイル
メモリ	データ, メモリカード, IC メモリカード, システム, 情報
CF カード	認証、CF カード、情報記憶装置、接続部、無線通信カード
不明	カード表面、荷物、システム、利用者、サービス提供

表5　課題トピック

クラスタ	単語
信頼性	平滑性, 耐障害性, セキュリティ, 信頼性, 通信可能距離
精度	精度、薄型化、分解、安全性、流通、耐久性
セキュリティ	強度、書き換え、耐久性、破損、セキュリティー、損傷
利便性	短縮、伝送、利便性、凹凸、再利用、簡素化、改竄
物理的強度 1	形状, 増加, 短縮, 耐熱性, 強度
多機能性	多機能, 拡張, ばらつき, 実用性, 情報
生産性	作業負荷, 生産性, 歩留まり, 強度, 設計の自由度, 競争力
物理的強度 2	強度, 表面平滑性, 損傷, 耐熱性, セキュリティー
通信距離	簡素化, 通信距離、通信特性、偽造、破壊
コスト	コスト、品質、性能、生産性、断線、反り、歩留まり

pages 113–120. ACM, 2006.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] Atsushi Fujii, Makoto Iwayama, and Noriko K. Overview of patent retrieval task at ntcir-5. In *In Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 269–277, 2005.

[4] 岸桂太 and 吉岡真治. 特長表現に注目した対象-観点型特許マップの自動生成. 研究報告デジタルドキュメント（*DD*）, 2014(9):1–5, mar 2014.

[5] 西山莉紗. 技術文書の情報編纂: 課題・特長・手段を表す表現の抽出と利用 (人工知能学会全国大会 (第 26 回) 文化, 科学技術と未来) – (近未来チャレンジセッション「nfc-4 (卒業セッション) 情報編纂の基盤技術」). 人工知能学会全国大会論文集, 26:1–4, 2012.

一般講演 2

「特許明細書の翻訳時で注意すべきこと」

特許明細書の翻訳で注意すべきこと（翻訳者からのノウハウ）

Attention Item in Translation of Patent Specification (Know-how from Translator)

吉川潔（Kiyoshi Kikkawa）　新潟県に居住する特許翻訳者

はじめに

新潟の田舎で東京の特許事務所から特許明細書の原稿をメールで受信し、翻訳後に返信するという仕事をテレワーク（SOHO）で、３０数年行ってきた。

その間に国際出願で高名な弁理士から指導を受けた。それをＭＴの開発に役立ててほしいので記す。

その多くは学校の教科書や翻訳用の参考書に未記載である。本年のＡＡＭＴ総会で、翻訳者のノウハウをコンピュータに取り込むこと、翻訳者の知識をＡＩ手法で収集し整理し自己増強することが重要と言われた。

下記が、その趣旨に沿うと確信する。特に⑦の数量表現を、ＡＩ（人口知能）や深層学習（deep learning）や用例翻訳に役立ててほしい。

最近、数社からＡＩや深層学習に基づくＭＴが発表されているが、本原稿の出稿日で試訳できる状態でない。従って、それらを対象外にして、本原稿を作成している。

① 単数と複数の区別を正確に

日本人は単数と複数の区別に鈍感だが、アングロサクソンは敏感である。区別を正確に要求する。日本語の原稿が曖昧でも、技術的な背景を考慮して、単数と複数を正確に英訳することが要求される。

1-1 （最近の自動車はモータを備えている）

A recent car includes a motor. →

「一台の自動車が、１個のモータを有する」の意味 ？

最近の自動車は数十個のモータを具備するので、日本語原稿に（複数）と明記してなくとも、複数で訳す。

A recent car includes motors.

1-2 （カメラとタブレットは電池を備えている）を、下記のように英訳すると、

A camera and a tablet contain a battery.

「一つのカメラと一つのタブレットが１個の電池を共有」の意味 ？

A camera and a tablet contain batteries.

[一つのカメラと一つのタブレットが幾つかの電池を共有]の意味 ？

Each of a camera and a tablet contains a battery.

「一つのカメラと一つのタブレットの各々が１個の電池を持つ」の意味。

３段目が日本語原稿を正確に訳すと想定すると、原稿にない「each of」を追記する。

1-3 （プログラムは、コマンドを有する複数のメニューを含んでいる）を、下記のように英訳すると、

A program includes a plurality of menus having a command.

上記は、一つのプログラムは、一つのコマンドを有する複数のメニューを含んでいる
即ち、複数のメニューが一つの
　コマンドを共有するという意味になる？
A program includes a plurality of menus
each having a command.

　一つのプログラムは複数のメニューを含み、各メニューが一つのコマンドを有する。
即ち、一つのメニューが一つのコマンドを有する、複数のメニューを一つのプログラムが有する。

下段の英文が日本語原稿の意味を正確に表すと想定すると、原稿にない「each」を追記して英訳

1-4　（これは、異なる波長を有する幾つかの信号を送る）の文意は、
（これは、一つ一つの信号の波長が異なる、幾つかの信号を送る）。
原稿にない each of which を補って英訳。
This sends several signals
each of which has a different wavelength.

1-5　この物質は、一次粒子が凝集した
二次粒子を含む

この物質は単数としても、
　二次粒子は単数でなく複数なことは常識。
　一次粒子が凝集ということは、１個の二次粒子に複数の一次粒子が凝集を意味する。

従って、この物質は複数の二次粒子を含み、一つの二次粒子に複数の一次粒子が凝集の趣旨で英訳する。

②　日本語原稿の表現が曖昧な場合、技術的な背景を考慮して、補いながら英訳する
2-1　サンプルＡとサンプルＢの直径は、
サンプルＣの直径と同じ
サンプルＡと、サンプルＢと、
サンプルＣの直径が、全て同じ？
サンプルＡとサンプルＢの二つの合計の
直径が、サンプルＣの直径と同じ？
この場合、サンプルＡとサンプルＢの
直径は等しい　？

2-2　サンプルＡとＢが
所定値であるかどうか調べる。
サンプルＡとＢの合計が、
所定値以下であるかどうか調べる？
サンプルＡとＢが、それぞれ単独で、
所定値であるかどうか調べる？
この場合、所定値は、
サンプルＡとＢとで異なる　？

2-3　サンプル１とサンプル２の
長さは等しい。
A sample 1 and a sample 2 are equal.
A sample 1 and a sample 2 are
equal to each other.
（お互いに等しいという語句を加える）
サンプルＡとサンプルＢが
他の何かと等しい？

③　不鮮明な修飾と被修飾の関係を
見抜いて英訳
3-1（増幅器が増幅できる周波数帯が、
この分野で用いられる）を英訳する際に、
電気知識のある翻訳者は（増幅器が ---
を増幅できる周波数帯）と意訳。
周波数帯そのものが増幅器で
増減することは考えられない！

3-2　（スイッチS1、S2が、ノード１と抵抗A､B間に接続している）を英訳する際に、
スイッチＳ１が、ノード１と抵抗Ａの間に接続し、スイッチＳ２が、ノード１と抵抗Ｂの間に接続している。
電気配線図を見て、上記のように区分けして意訳しなければならない　！

3-3 これは、炭素繊維やガラス繊維から
作られたシートである。
This is a sheet made of carbon fiber
and glass fiber.
炭素繊維とガラス繊維の両方から
作られたシートの意味
This is a sheet made of carbon fiber
or glass fiber.
炭素繊維またはガラス繊維の一方
から作られたシートの意味
This is a sheet made of carbon fiber
and/or glass fiber.
炭素繊維及び又はガラス繊維から
作られたシート（両方でも一方でも可）
日本語原稿だけでは、上記の三つのどれか不明なので、前後の文脈から推定して訳す

3-4 それは、領域に生成された電荷が転送される転送路（特許庁の出版物から抜粋）
（特許庁の英訳　×）
It is a transfer path through which charges defined in a region is transferred.
特許庁の図面から　「それは、領域に生成され且つ電荷が転送される転送路」の意味

正訳　→　It is a transfer path
which is defined in a region and
through which charges are transferred.

3-5　　主語や目的語や補語を補充
サンプルを､壁と垂直な位置に配置したり、
それと平行な位置に配置する
A sample is arranged at a position vertical to a wall, or arranged at a position parallel to the wall.

④　日本語原稿には、単純な誤字脱字だけでなく、下記のような気付きにくい
誤記もあるので注意が必須

4-1
事故は自動車の加速度のために発生　×
→　事故は、自動車の過速度の誤記

事故は、自動車の加速のために発生　○
→ 事故は、自動車の急加速のために発生

4-2　サンプルを、マット上の表面に置く
A sample is placed on a surface
on a mat.
サンプルを、マット状の表面に置く
A sample is placed on
a mat-shaped surface.

4-3　　ワープロ機能の単純な変換ミス
充電対称の電池　→　充電対象の電池

⑤　分詞構文の主語に注意（日本語は、
主語を省くことが多いので間違う）
北海道に旅行したら、雪が残っていた。
下記は×
When travelling in Hokkaido,
snow still remained.
雪が北海道に旅行したら、雪が残っていた。
When I travelled in Hokkaido,
snow still remained.
私が北海道に旅行したら、雪が残っていた。

分詞構文の主語が省略してある場合、
本文の主語が、分詞構文の主語になるのが英文法！

⑥　英文和訳の例
２０年前、米国のA社が某案を日本特許庁に申請したらパスした。内定後３０日以内に異議が無いと、当時は登録になった。
しかし、明細書に「真の半導体」という語句があるが、これは半導体の物理的な状態を表現していない。不的確な表現という異議申立があった。反論が認められ、最後の土壇場で、A社の申請は却下！
A社の英文原稿に
「intrinsic semiconductor」があり、
「真性半導体」が正訳。「真の半導体」と不的確に訳したので、却下になったと推測。

⑦　MT は数量関連の表現の翻訳が苦手。
類似の例を集めて、用例翻訳として活用
7-1　例えば、分数を正確に英和訳できない。
An aperture is reduced to
several one –tenths.
「口径を数十分の一に減らす」が正訳

7-2　日本語と英語の違いとして、
主語と述語の語順の違いがあるが、
形容詞的な表現の語順にも違いがある
It is a pack one pack earlier.
「それは一つ前のパックである」が正訳

7-3　動詞句の語尾の変化に追従できない
「本はいくらですか」を正訳しても、
「本はいくら」、「本はいくらだ」
「本はいくらだった」
「本はいくらでしょうか」のように動詞句の語尾が変わると、誤訳する MT が多い。動詞句の語尾変動に正確に追従して正訳すべきである。

7-4　「比較に関連した表現」：
下記の英訳で、上段は全ての MT が正訳するが、中段は半々、下段は全滅。
A店の値段は、B店より高い。
A店の値段は、B店より５％高い。
A店の値段は、B店より５％以上高い。
A店の値段は、B店より１０円高い。

7-5「at least に関連した表現」
少なくとも一部の～
少なくとも１個から５個多い
少なくとも～と同じ長さ
～を少なくとも制御する

7-6　「できるだけ～の表現」
できるだけ～を短く
できるだけ短周期で反転させる
～から、できるだけ下げる
できるだけ～の近くに配置する

7-7　「～程度の表現」
ベースに接しない程度で右側に配置する
子供が読める程度に短文にする
湯温が４１℃になる程度に加熱する
ピンが容易に動かない程度の深さ

⑧　今後の方針：上記の⑦の数量や形容句に関係した語句や文例を集めて体系化し、用例翻訳に役立つようにする。
動詞句や形容詞句の語尾変動の誤訳事例を集めて体系化し、AI や深層学習に役立つようにする。

一般講演 3

「特許文献中の重要語に着目した特許分類の推定」

特許文献中の重要語に着目した特許分類の推定

綱川 隆司[1,2]　佐々木 深[2]　西田 昌史[1,2]　西村 雅史[1,2]

[1]静岡大学学術院情報学領域

[2]静岡大学大学院総合科学技術研究科情報学専攻

{tuna@inf, gs15021@s.inf, nishida@inf, nisimura@inf}.shizuoka.ac.jp

1　はじめに

特許審査等のための先行技術調査を効率的に行うため，特許文献の検索においては，IPC（国際特許分類），FI，F タームといった特許分類が検索インデックスとして重要な役割を果たしている [1]. これらの分類を用いた検索はキーワード検索と比べ，ある観点に対して網羅的な検索結果が得られるという特長がある．より正確で十分な粒度の分類を行うことが検索の効率化につながるため，分類付与作業や分類の改正への対応に人手による作業が必要となる．

分類付与作業は，効率化のために最初に粗い分類を自動的に行っており [2] ，より細かい分類へ対応した高精度な自動分類システムが望まれる．従来，既存の特許文献に付与された分類を訓練データとして用いる教師あり機械学習による分類方法が提案されている [3-6]. しかし，分類によってはその分類が付与されている特許文献の数が訓練には不十分である，分類改正による新設の分類には訓練データが存在しない，といった課題があり，分類精度も向上の余地が残されている．

本研究では，従来の機械学習手法に加え，特許文献中の重要語に着目して F ターム自動推定の精度向上を図る．また，訓練データが不足する分類に対し，抽出された重要語と比較して自動分類する方法について検討する．

2　F ターム

日本で用いられている特許分類は, 分類の粗い順に IPC（国際特許分類），FI，F タームがある [1]．IPC は 1975 年から用いられている国際的に定められた世界共通の分類体系である．IPC では，発明に関する全技術分野を 8 分野のセクションに分け，以下クラス，サブクラス，メイングループ，サブグループの順に段階的に細分化されており，およそ 7 万の分類が存在する．

IPC は国際的に通用する分類であるため，各国の特許を効率的に検索できる一方で，国ごとに開発が活発な分野が異なる等の理由で，一部の分野に文献が集中することがあり得る．このことから，日本では IPC をさらに展開した索引として FI および F タームが用いられている．

FI は IPC に付加する形でサブグループの下位分類として展開記号・分冊識別記号を付与したものである. 一方，F タームは種々の技術的観点（目的，用途，構造，材料等）から IPC を分野ごとに再区分したものであり，複数観点を組み合わせることで関連特許をより効率的に絞り込むことを目的に定められている．

F タームは，FI の全技術分野を約 2600 の“テーマ”に区分し，一部のテーマについて，テーマごとに定義された観点に対して割り当てられている．F タームはテーマを表すテーマコード（英数字 5 桁），観点（英字 2 桁），数字（2 桁）で構成される．例えば，テーマ“ハードウェアの冗長性”のテーマコードは 5B034 であり，4 つの観点“受動的冗長”，“能動的冗長”，“冗長回路”，“機能・構成”を持つ．F タームは表 1 に示すように観点またはその下位分類に対応している．各 F タームの先頭のドット（・）は階層の深さを示し，ドットの数が多いほど下位の階層を表す．表 1 で形成される階層構造を図 1 に示す．F タームは組合せで検索されることが想定されており，1 つの特許に対して同一テーマコード内の多数の F タームが付与されている．

3　関連研究

NTCIR-5 および NTCIR-6 の特許検索タスクにおいて，F タームの分類に焦点を当てた特許文献の自動分類タスクが設けられ，テストコレクションが公開されている [7,8]．NTCIR-6 分類タスクは，1993～1997 年に公開された日本公開特許公報の特許文献全文を訓練データとし，1998～1999 年の特許文献 21606 件に対して F タームを付与する課題であり，6 グループがタスクに参加した．Li et al. [3] は特許文献の bag-of-words を素性として用いた SVM による分類器を用い，完全一致による評価で F 値 0.4125 を達成し最高性能を得た．Fujino and Isozaki [4] は，特許文献の各要素（発明の名称，出願人・発明者，要約

表 1　F タームリストの例

5B034	ハードウェアの冗長性						
観点	F ターム						
AA	AA00	AA01	AA02	AA03	AA04	AA05	…
	受動的冗長	・二重化	・・照合	・・・圧縮照合	・多重化	・・多数決	…
BB	BB00	BB01	BB02	BB03	BB04	BB05	…
	受動的冗長	・切替	・・予備切替	・・・共通予備	・・選択	・・・信頼度	…
		BB11	BB12	BB13		BB15	…
		・再構成	・・緊急制御回路	・・機番変更		・切離し	…
…	…	…	…	…	…	…	…

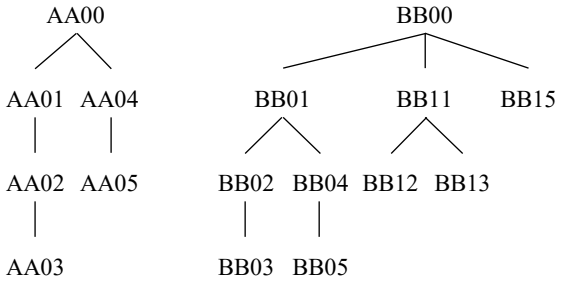


図 1　F タームリスト階層の例

書，特許範囲，明細書）について bag-of-words によるナイーブベイズの二値分類器を作り，これらを最大エントロピー法に基づき組み合わせる方法を提案した．Murata et al. [5] は k-NN 法に基づく方法を提案し，特許文献間の類似度として SMART [9] や BM25 [10] を用いて評価を行った．

小林 [6] は，F タームの付与根拠データ[1]を用いた結果をもとに，付与根拠データが少ない分野に対して機械学習に基づき精度を向上させるため，tf-idf および分類の階層構造を利用した分類推定方法を提案した．

4　提案方法

本研究では，特許文献に付与されるべき F タームを自動推定するため，既存の特許文献に人手で付与された F タームの分類情報を学習データとする機械学習手法を用いる．特許文献分類のための機械学習手法として有効である SVM（サポートベクタマシン）を用い，特許文献の bag-of-words に加え，特許文献に出現する重要語を素性として用いる．

4.1　用いる素性

(1) 出現語の bag-of-words（正規化 tf-idf）
形態素解析ソフト MeCab[2] を使用し特許文献から名詞（非自立名詞，固有名詞，数は除く），自立動詞，自立形容詞，未知語を抽出する．これらに対し tf-idf で重みづけしたベクトルを求め，正規化して各出現語に対する素性値とする．

(2) 特許文献の重要語の出現有無
特許文献中には属する技術分野の専門用語が多く含まれており，形態素解析によってこれらの単語を分割すると本来の意味を反映されないケースが想定される．そこで，重要語自動抽出モジュール TermExtract[3] [11] を使用し，特許文献を特徴づける重要語の抽出を行う．このとき，TermExtract によって計算される重要度が 20 以上のものを重要語として抽出し，それらの重要語が特許文献に出現するかどうかを素性として加える．ただし，(1) の素性値とのバランスをとるため，(1) で求めた正規化ベクトルの各要素の値の平均値を算出し，出現した重要語の素性値とする．

4.2　学習アルゴリズム

1 つの特許文献に対し複数の F タームが付与されるため，学習データに付与されているテーマコードに対し，展開される F タームそれぞれに対する分類器をテーマコード別に学習する．すなわち，ある F タームの分類器を学習する際，対象 F タームが付与された文献を正例，展開元のテーマコードが付与され，かつ対象 F タームが付与されていない文献を負例とする．このとき，負例が圧倒的に多くアンバランスな学習データになるため，正例

[1] 特許文献への分類付与者が，付与することとなった根拠箇所を明細書等から抽出したもの．

[2] http://taku910.github.io/mecab/
[3] http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

はすべて採用し，負例は最大で正例数の 2 倍の数だけランダムに選択する[3,12].

また，F タームは階層構造を成しており，親子関係や祖孫関係にあたるような F タームが同時に付与されることはあまりない．上記のような場合，下位のものは上位の特徴を継承しさらに具体的な分類を示すため，本研究においては階層関係を考慮し，下位のものを選択しそれ以外は除外する．

5 評価実験

5.1 実験データ

NTCIR-6 のデータコレクションを使用し，1993～1997 年に発行された特許文献を訓練データ，1998～1999 年に発行されたものをテストデータとする．テストデータに付与されるテーマコードは 108 種類あり，本実験では，この中から 20 個ランダムに選択し，そのテーマコードが付与された特許文献を対象とする．本実験に用いた 20 のテーマコードに含まれる文献数および 1 文献に付与された F ターム数を表 2 に示す．

5.2 評価方法

テストデータにおいて，提案手法によって付与された F タームと本来付与されるべき正解 F タームに対し適合率, 再現率, F 値を計算し F ターム推定性能を評価する．それぞれの指標は以下の式で定義される．

$$\text{適合率} = \frac{\text{（推定した F タームで正解のものの数）}}{\text{（推定した F タームの数）}},$$

$$\text{再現率} = \frac{\text{（推定した F タームで正解のものの数）}}{\text{（正解の F タームの数）}},$$

$$\text{F 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}.$$

また，F タームの階層構造を考慮する場合としない場合の精度についても比較を行う．

5.3 結果

表 3 に，20 のテーマコードのいずれかを持つ特許文献全体に対する適合率，再現率，F 値を示す．素性として bag-of-words の tf-idf のみを用いた場合と比べ，重要語を素性として加えることで付与する F ターム数が増加し，再現率の改善がみられた．一方，適合率の低下により，F 値はほぼ同水準にとどまった．

テーマコードによって文献数に偏りがあり，たとえば，テーマコード 3H045（容積形ポンプの制御）とテーマコード 2C088（弾球遊技機（パチンコ等））の文献数はそれぞれ 1222，7794 であり大きく異なる．また，表 4，表 5 にそれぞれの適合率，再現率，F 値を示す．いずれも再現率は改善したが，テーマコード 2C088 では正解の F ターム数と比べ付与した F ターム数が比較的多いために低い適合率にとどまり，F 値の改善にいたらなかった．20 のテーマコードのうち，提案方法の F 値が最良だったものは 12 あった．さらに，テーマコードの付与数だけでなく，そこから展開される F タームにも付与数にも差があり，学習データにおいて正例が極端に少ない F タームが存在する．この場合，テストデータを用いた推定で，当該 F タームは一回も付与されないという結果がみられた．

表 2　実験に用いたデータ

訓練データの文献数		45631
テストデータの文献数		4012
1 文献に付与された F ターム数	平均	8.37
	最大	43
	最小	1

表 3　F ターム推定結果

素性	階層考慮	適合率	再現率	F 値
(1)のみ	なし	0.227	0.394	0.288
	あり	**0.242**	0.379	0.296
全て（提案手法）	なし	0.204	**0.490**	0.288
	あり	0.219	0.465	**0.298**

表 4　テーマコード 3H045
（容積形ポンプの制御）の F ターム推定結果

素性	階層考慮	適合率	再現率	F 値
(1)のみ	なし	0.353	0.353	0.353
	あり	**0.373**	0.341	0.356
全て（提案手法）	なし	0.288	**0.461**	0.355
	あり	0.312	0.431	**0.362**

表 5　テーマコード 2C088
（弾球遊技機（パチンコ等））　の F ターム推定結果

素性	階層考慮	適合率	再現率	F 値
(1)のみ	なし	0.070	0.496	0.123
	あり	**0.082**	0.477	**0.140**
全て（提案手法）	なし	0.060	**0.577**	0.109
	あり	0.073	0.541	0.129

5.4 訓練データが少ない F タームの分類推定

テーマ“旋削加工”(3B045) には F タームが 197 個あり，そのうち 3B045DA20（旋削加工の形態＞・非円形断

面＞・・多角形）が付与された特許文献は訓練データに 3 件，テストデータに 1 件のみ存在した．F ターム推定の結果，テストデータに含まれるテーマコード 3B045 を持つ特許 193 件すべてについて当該 F タームは付与されなかった．訓練データ 3 件には，多角形を示す“ポリゴン”という語が出現するが，テストデータの 1 件では“非円形”・“突起”の 2 語で多角形であることが示唆されており，この観点において語彙が一致しなかった．また，訓練データの文献数が少ないため，他の出現語彙の類似性からこれらの文献間の共通性を見出して当該 F タームが付与されることも期待できない．このケースでは，F タームの“多角形”という語自身を利用し，多角形と関連のある語句を用いて比較するといった手法が必要になると考えられる．

また，F ターム 3G023AF04（内燃機関燃焼法＞対象とする機関＞・側弁式機関）についても，訓練データ 5 件，テストデータ 2 件にのみ付与されており，テストデータのいずれにも当該 F タームは付与されなかった．訓練データの文献のいくつかには“側弁”，“サイドバルブ”の語句が含まれているが，テストデータの文献には含まれておらず，形状は文章の説明および図によって示されている．このケースも訓練データが不十分であり，この F タームに特徴的な重要語の共通性もみられないため本手法では正しい分類推定が困難な例である．

6 おわりに

本研究では，特許文献に対して特許分類 F タームを自動的に付与するため，特許文献中の重要語を用いた機械学習手法を提案した．評価実験の結果，付与する F ターム数を増加させ再現率を改善したが，F 値は重要語の追加前と比べ同水準の性能にとどまった．また，訓練データの少ない F タームの推定結果について調査した．

今後の課題として，訓練データの少ない F タームの推定に対応するため，F ターム自身および F タームの説明を利用して当該 F タームに関連する語句を集め，特許文献と比較する方法を検討する．また，テーマコードごとの F タームの分布を分類推定に利用する方法も考えられる．

参考文献

[1] 独立行政法人工業所有権情報・研修館. (2016). 特許分類の概要とそれらを用いた先行技術調査～IPC, FI，F ターム編～（平成 28 年度版）. http://www.inpit.go.jp/content/100798564.pdf.

[2] 古屋野 浩史. (2007). 特許分類等の付与精度向上への取り組み. *Japio 2007 YEAR BOOK*, pp.118-119.

[3] Li, Y., Bontcheva, K., and Cunningham, H. (2007). SVM based learning system for F-term patent classification. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 396-402.

[4] Fujino, A. and Isozaki, H. (2007). Multi-label patent classification at NTT Communication Science Laboratories. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 381-384.

[5] Murata, M., Kanamaru, T., Shirado, T., and Isahara, H. (2007). Using the k-nearest neighbor method and SMART weighting in the patent document categorization subtask at NTCIR-6. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 407-413.

[6] 小林 英司. (2015). 特許分類の自動推定に向けた取り組み－機械学習による自動分類推定の課題と今後の展開－. *Japio YEAR BOOK 2015*, pp. 272-275.

[7] Iwayama, M., Fujii, A., and Kando, N. (2005). Overview of classification subtask at NTCIR-5 patent retrieval task. In *Proc. of NTCIR-5 Workshop Meeting*,

[8] Iwayama, M., Fujii, A., and Kando, N. (2007). Overview of classification subtask at NTCIR-6 patent retrieval task. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 366-372.

[9] Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96)*, pp. 21–29.

[10] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1994). Okapi at TREC-3. In *Proc. of the 3rd Text REtrieval Conference (TREC-3)*, pp. 109–126.

[11] 中川 浩志, 湯本 紘彰, 森 辰則. (2003). 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, 10(1), 27-45.

[12] Li, Y. and Shawe-Taylor, J. (2003). The SVM with uneven margins and Chinese document categorisation. In *Proc. of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC 17)*, pp. 216-227.

平成28年度AAMT/Japio特許翻訳研究会
第4回特許情報シンポジウム論文資料集

発 行 日　平成28年11月

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

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

印　　刷　株式会社 インターグループ