第 3 回特許情報シンポジウム

論文資料集

2014 年 11 月 28 日

キャンパスイノベーションセンター東京

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：	梶　博行	静岡大学
副委員長：	横山晶一	山形大学
顧問：	辻井潤一	マイクロソフトリサーチアジア
	守屋敏道	日本特許情報機構
委員：	潮田　明	元奈良先端科学技術大学院大学
	宇津呂武仁	筑波大学
	越前谷博	北海学園大学
	江原暉将	山梨英和大学
	大塩只明	日本特許情報機構
	河合弘明	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
	須藤克仁	NTT コミュニケーション科学基礎研究所
	隅田英一郎	情報通信研究機構
	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
	塙　金治	日本特許情報機構
	早川貴之	日本特許情報機構
	宮澤信一郎	秀明大学
事務局：	野村佳代子	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」の翻訳が優先権主張や公知例調査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 年と 2012 年に開催され、今回が第 3 回目である。前 2 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。

まず、3 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、機械翻訳技術の最新の研究動向についてのご発表、および特許情報サービスにおける機械翻訳の活用事例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。

次に、AAMT/Japio 特許翻訳研究会から、研究会活動の紹介を兼ねて、2 つの重要なテーマについて報告させていただく。一つは特許文書の翻訳に不可欠な専門用語対訳辞書の構築に関する研究成果の発表、もう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。さらに、近年、関心が高まっている中国語などアジア言語の機械翻訳システムの研究開発について、それに携わっている研究会メンバーに講演をお願いした。

投稿ベースの一般発表も興味深い 6 編の論文を採用することができた。機械翻訳に関する開発サイドと利用サイドの論文に加えて、特許分類に基づくオントロジーの構築、特許文書からの情報抽出、および特許情報の分析ツール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。

2014 年 11 月　第 3 回特許情報シンポジウム
実行／プログラム委員会
委員長　梶　　博　行

プログラム

10:00-10:10　開会挨拶
辻井潤一（AAMT/Japio 特許翻訳研究会委員長，
マイクロソフトリサーチアジア 首席研究員，東京大学名誉教授）

セッション１（招待講演）　座長 : 梶博行（静岡大学）

10:10-10:40　招待講演 1「特許庁における機械翻訳への取組」
櫻井健太（特許庁）

10:40-11:10　招待講演 2「多言語機械翻訳の研究開発動向」
隅田英一郎（情報通信研究機構）

11:10-11:40　招待講演 3「特許情報検索サービスにおける機械翻訳の活用」
早川浩平（日本パテントデータサービス）

休憩（昼食）

セッション２（研究会報告・特別講演）　座長 : 江原暉将（山梨英和大学）

13:10-13:40　研究会報告 1「パテントファミリーからの専門用語対訳辞書の構築」
宇津呂武仁（筑波大学）

13:40-14:20　研究会報告 2「自動評価法を用いた機械翻訳の定量的評価」
越前谷博（北海学園大学），磯崎秀樹（岡山県立大学）

14:20-14:50　特別講演「アジア言語を中心とした機械翻訳研究」
中澤敏明（科学技術振興機構／京都大学）

休憩

セッション３（一般発表(1)）　座長 : 宇津呂武仁（筑波大学）

15:10-15:30　統計的訳語選択技術による韓日機械翻訳の高精度化
田中浩之，園尾聡，木下聡，釜谷聡史（東芝 研究開発センター）

15:30-15:50　特許事務所における機械翻訳と人手による翻訳の Mix 事例
正林真之，杉浦伸夫（正林国際特許商標事務所）

15:50-16:10　インターネットから利用できる翻訳ソフトを優れた辞書として活用する方法
吉川潔（翻訳業）

休憩

セッション４（一般発表(2)）　　座長 : 横山晶一（山形大学）

16:30-16:50　F タームに基づいたオントロジーの構築
福田悟志，難波英嗣，竹澤寿幸（広島市立大学），乾孝司（筑波大学），
岩山真（日立製作所），橋田浩一（東京大学），藤井敦（東京工業大学）

16:50-17:10　特許文書からの化学物質情報の抽出
池田紀子，田中一成（富士通研究所）

17:10-17:30　特許情報分析のためのマイニング手法と分析ツール Patent Mining eXpress
岩本圭介（NTT データ数理システム）

17:30-17:40　閉会挨拶
守屋敏道（日本特許情報機構 専務理事 特許情報研究所所長）

18:00-　懇親会（２階多目的室３、会費無料）

目 次

招待講演 1

「特許庁における機械翻訳への取組」

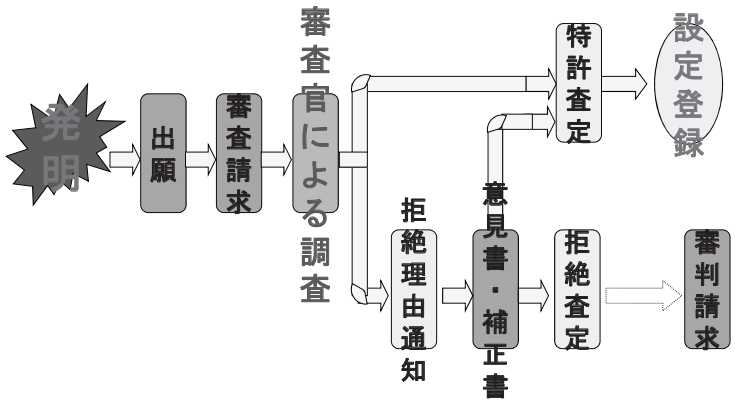
特許庁における機械翻訳への取組

特許庁特許情報企画室　櫻井健太

目次

- 特許審査の実務
- 特許審査の課題
- 機械翻訳の活用

特許審査の実務



特許審査の実務

特許になる発明とは

新しいかどうか（新規性）	×公然と知られた発明（発表、TV放映） ×公然と実施された発明（販売） ×刊行物に記載された発明（特許公報、
容易に考え出すことができないか（進歩性）	☆当業者が容易に考えつかない発明
明細書の記載は規定どおりか	☆当業者が実施可能な程度まで技術内容の記載を義務づけ
その他	☆先に出願されていないか ☆公序良俗を害しないか

先行技術調査の結果に基づき判断

特許審査の実務

- 審査官による**先行技術調査**とは

特許出願された発明が新規性、進歩性を有するかどうかを判断するため、出願日よりも前に公開されている文献を調査すること

- 特許公報に限らない
  →技術雑誌等も含む
- 国内文献に限らない
  →外国文献も含む

特許審査の実務

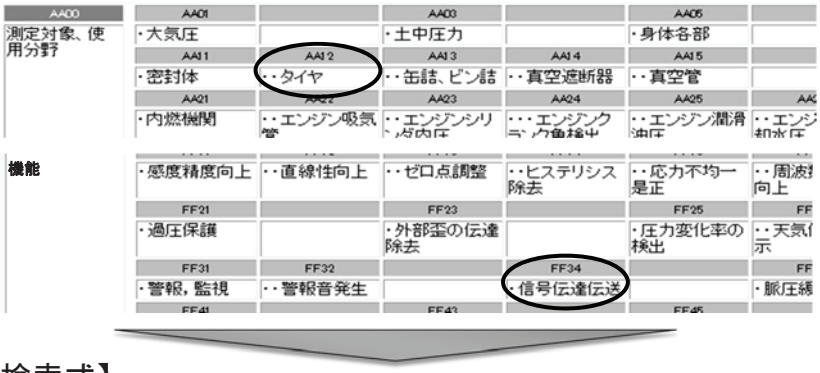
- 仮想事例

【請求項1】
車両の各タイヤに設けられタイヤの状態に関する情報を無線で送信するタイヤセンサユニットと、車体側に設けられタイヤセンサユニットから送信された情報を受信する受信装置と、受信装置で受信した情報に基づいてタイヤの状態を表示する表示装置とからなるタイヤ監視システムであって、
車体側に無線でエネルギを送出する非接触型給電部を設けるとともに、タイヤセンサユニットに非接触型給電部から送出されたエネルギに基づいて直流電源を生成する非接触型受電部を設けたことを特徴とするタイヤ監視システム。

【請求項2】
請求項1に記載されたタイヤ監視システムにおいて、タイヤセンサユニットはシート状の基板に設けられ、タイヤホイールに接着されていることを特徴とするタイヤ監視システム。

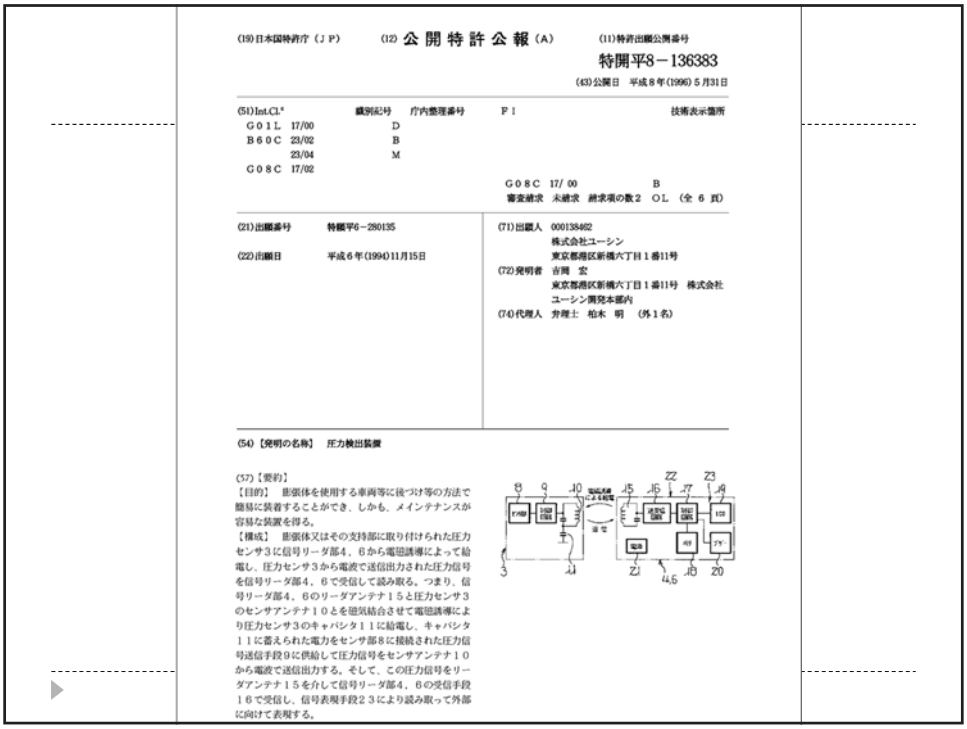
特許審査の実務

特許分類 （流体圧力測定2F055）



【検索式】

No.	テーマ	論理式	件数
¥01	2F055	AA12×FF34	272件



(19)日本国特許庁（JP）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平8－136383

(43)公開日　平成8年(1996)5月31日

(51)Int.Cl.⁶	識別記号	庁内整理番号	FI	技術表示箇所
G01L 17/00	D			
B60C 23/02	B			
23/04	M			
G08C 17/02				
			G08C 17/00 B	

審査請求　未請求　請求項の数2　OL　（全6頁）

(21)出願番号　特願平6－280135

(22)出願日　平成6年(1994)11月15日

(71)出願人　000138402
株式会社ユーシン
東京都港区新橋六丁目1番11号

(72)発明者　吉岡　宏
東京都港区新橋六丁目1番11号　株式会社ユーシン開発本部内

(74)代理人　弁理士　柏木　明　（外1名）

(54)【発明の名称】　圧力検出装置

(57)【要約】

【目的】　膨張体を使用する車両等に後づけ等の方法で簡易に装着することができ、しかも、メインテナンスが容易な装置を得る。

【構成】　膨張体又はその支持部に取り付けられた圧力センサ3に信号リーダ部4，6から電磁誘導によって給電し、圧力センサ3から電波で送信出力された圧力信号を信号リーダ部4，6で受信して読み取る。つまり、信号リーダ部4，6のリーダアンテナ15と圧力センサ3のセンサアンテナ10とを磁気結合させて電磁誘導により圧力センサ3のキャパシタ11に給電し、キャパシタ11に蓄えられた電力をセンサ部8に接続された圧力信号送信手段9に供給して圧力信号をセンサアンテナ10から電波で送信出力する。そして、この圧力信号をリーダアンテナ15を介して信号リーダ部4，6の受信手段16で受信し、信号表現手段23により読み取って外部に向けて表現する。

特許審査の実務

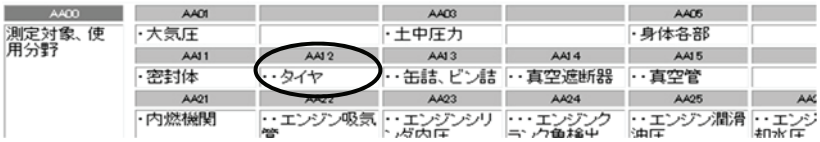
- 仮想事例

【請求項1】
車両の各タイヤに設けられタイヤの状態に関する情報を無線で送信するタイヤセンサユニットと、車体側に設けられタイヤセンサユニットから送信された情報を受信する受信装置と、受信装置で受信した情報に基づいてタイヤの状態を表示する表示装置とからなるタイヤ監視システムであって、
車体側に無線でエネルギを送出する非接触型給電部を設けるとともに、タイヤセンサユニットに非接触型給電部から送出されたエネルギに基づいて直流電源を生成する非接触型受電部を設けたことを特徴とするタイヤ監視システム。

【請求項2】
請求項1に記載されたタイヤ監視システムにおいて、タイヤセンサユニットはシート状の基板に設けられ、タイヤホイールに接着されていることを特徴とするタイヤ監視システム。

特許審査の実務

特許分類 （流体圧力測定2F055）



【追加の検索式】

No.	テーマ	論理式	件数
¥01	2F055	AA12×FF34	272件
¥02	2F055	AA12×(シート状＋フィルム状＋フレキシブル＋可撓性),IC,基板/TX－¥1	1件
¥03		タイヤ,5N,センサ/TX×(シート状＋フィルム状＋フレキシブル＋可撓性),IC,基板/TX－¥1－¥2	4件

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平9－5178

(43)公開日　平成９年(1997)１月10日

(51)Int.Cl.⁶	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０１Ｌ　3/10			Ｇ０１Ｌ　3/10	Ｃ	
Ｇ０８Ｃ　17/02			Ｇ０８Ｃ　17/00	Ｂ	

審査請求　未請求　請求項の数９　ＯＬ　（全７頁）

(21)出願番号　特願平7－156481

(22)出願日　平成７年(1995)６月22日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　射場本　正彦
茨城県ひたちなか市大字高場2520番地　株式会社日立製作所自動車機器事業部内

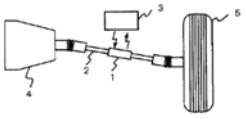
(72)発明者　鈴木　正善
茨城県ひたちなか市大字高場2520番地　株式会社日立製作所自動車機器事業部内

(72)発明者　上野　定寧
茨城県ひたちなか市大字高場2520番地　株式会社日立製作所自動車機器事業部内

(74)代理人　弁理士　平木　祐輔

(54)【発明の名称】　トルク検出システム

(57)【要約】
【目的】　信頼性及び機能性の高いトルクセンサを提供する。
【構成】　回転体２上に設けたセンサ部１に非接触で電力を送り込み、測定結果を非接触で読み取る。このため回転体２の脇の計測部３からマイクロ波を照射し、センサ部１ではそのエネルギーで電力を賄なって、計測信号を電波として計測部３に送り返す。回転体２上のセンサ部１は半導体チップと周辺部品のみで構成され、外部から独立しているので完全にモールドされる。
【効果】　接触不良等による誤動作がなく信頼性の高いセンサが得られる。センサ部は小形軽量に作れるので従来取りつけられなかった狭い部分にも取付け可能であり、トルクとともに回転数などの情報も検出できるので機能的な複合センサが得られる。



特許審査の実務

拒絶理由通知書

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　櫻井　健太
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。

特許査定

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　櫻井　健太
発明の名称　タイヤ監視システム
請求項の数　２
特許出願人　●●●●
代理人　●●●●

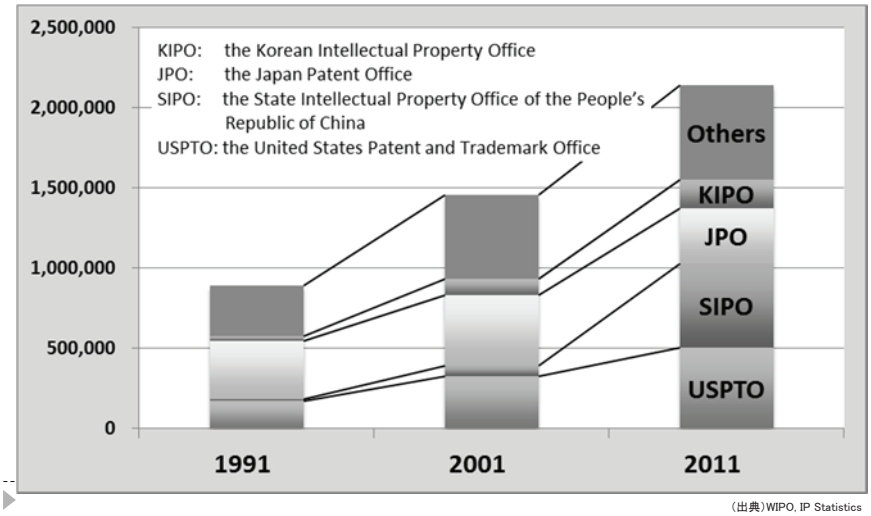
この出願については、拒絶の理由を発見しないから、特許査定をします。

目次

- 特許審査の実務
- 特許審査の課題
- 機械翻訳の活用

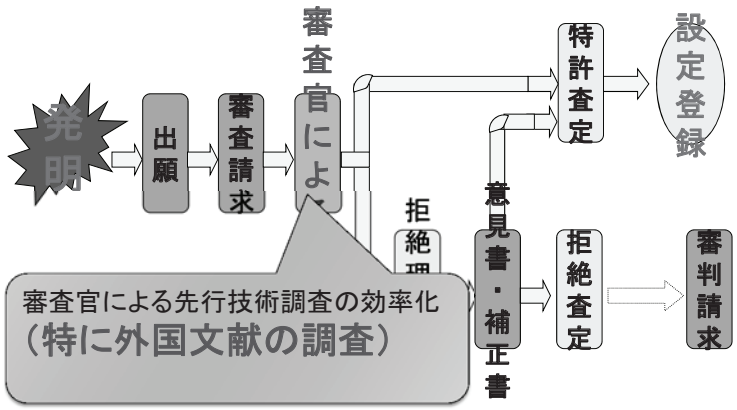
特許審査の課題

- 世界的な出願件数増



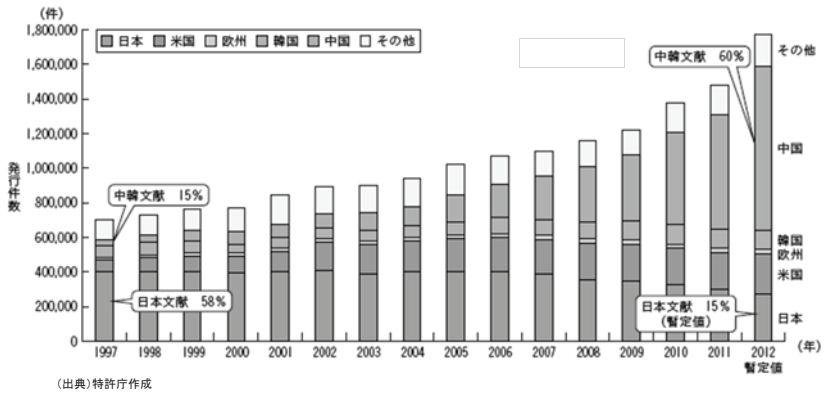
（出典）WIPO, IP Statistics

特許審査の課題



特許審査の課題

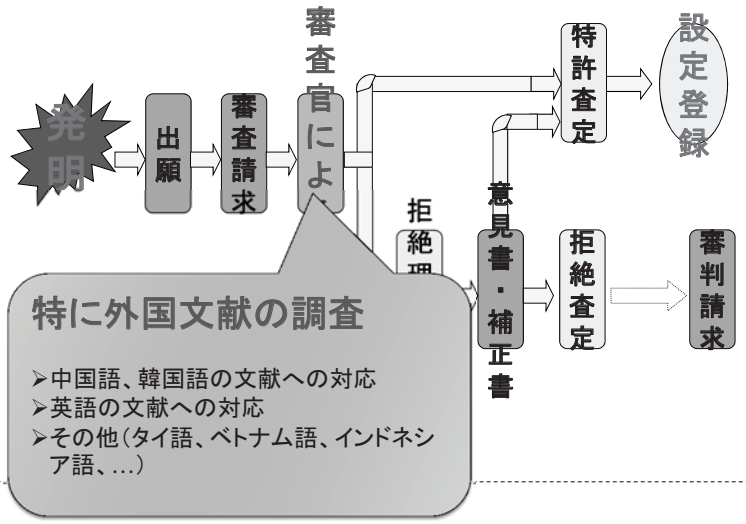
【中国、韓国の特許文献の急増】



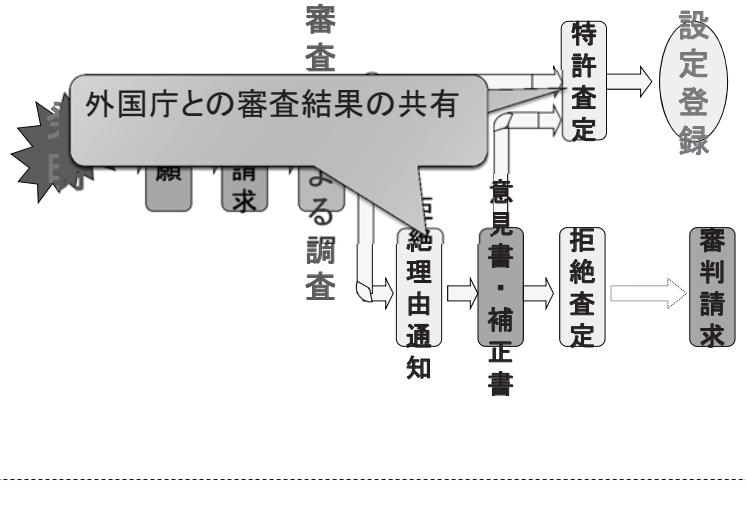
（出典）特許庁作成

（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。
2012 年の発行件数は暫定値。

特許審査の課題

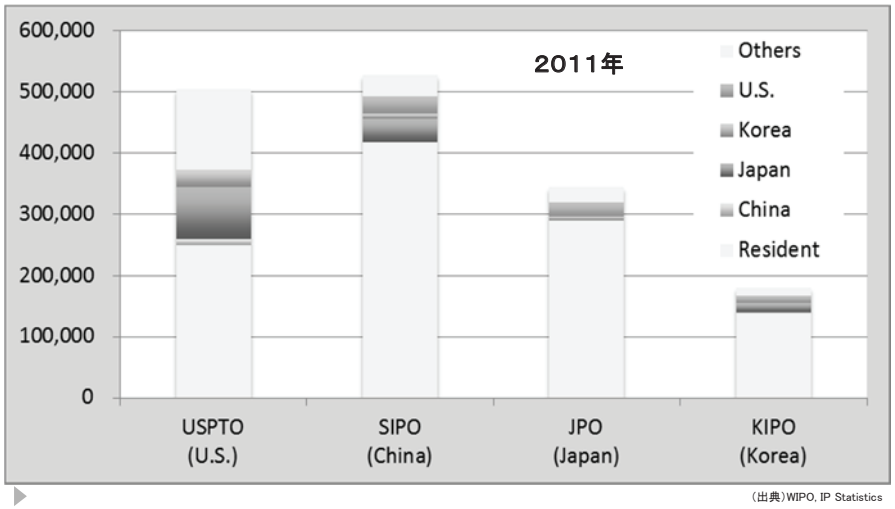


特許審査の課題



特許審査の課題

- 各庁の重複作業の削減



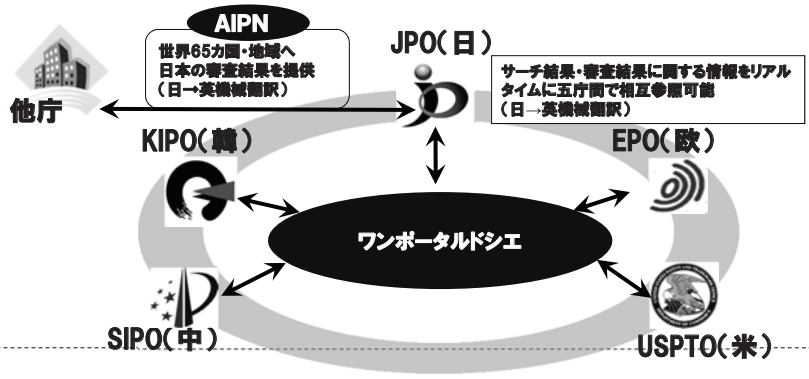
（出典）WIPO, IP Statistics

目次

- 特許審査の実務
- 特許審査の課題
- 機械翻訳の活用

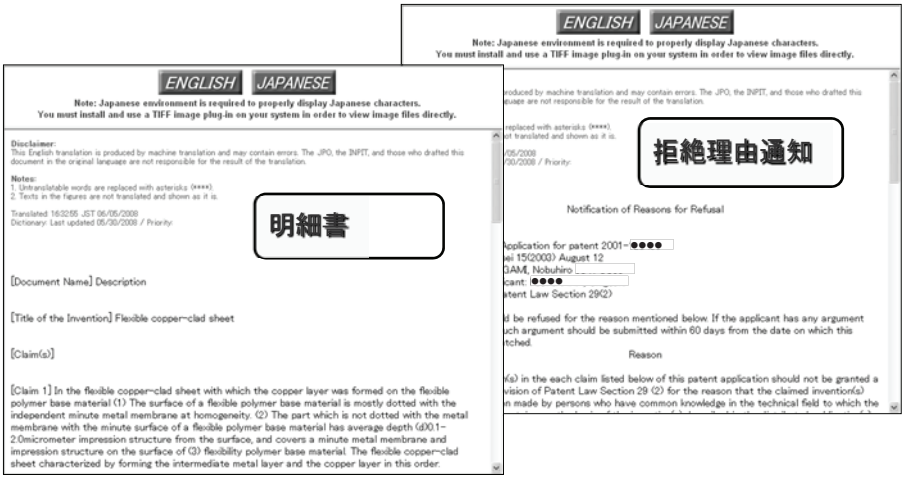
機械翻訳の活用　　（日→英機械翻訳）

- AIPN／ワンポータルドシエ
  - AIPN：Advanced Industrial Property Network （高度産業財産ネットワーク）
    途上国を含む外国特許庁に、日本の審査結果の情報を提供
  - ワンポータルドシエ
    五大特許庁各庁の包袋情報のデータベースを相互接続し、五庁審査官が各庁の包袋情報をワンストップで参照可能



機械翻訳の活用　　（日→英機械翻訳）

- AIPNにおける日英機械翻訳の例



機械翻訳の活用　　（日→英機械翻訳）

- AIPNにおける定常的な取組

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）
2014年10月時点、8万5千語を収録

海外特許庁からの誤訳フィードバック

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

機械翻訳エンジンのバージョンアップ

翻訳知識の強化や専門用語・知財用語数の増加（2011年12月）

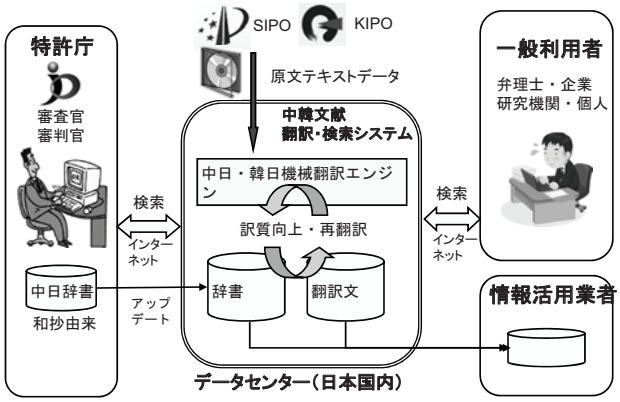
翻訳メモリの構築

拒絶理由通知に利用される定型表現を翻訳メモリに登録

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

機械翻訳の活用　　（中韓→日機械翻訳）

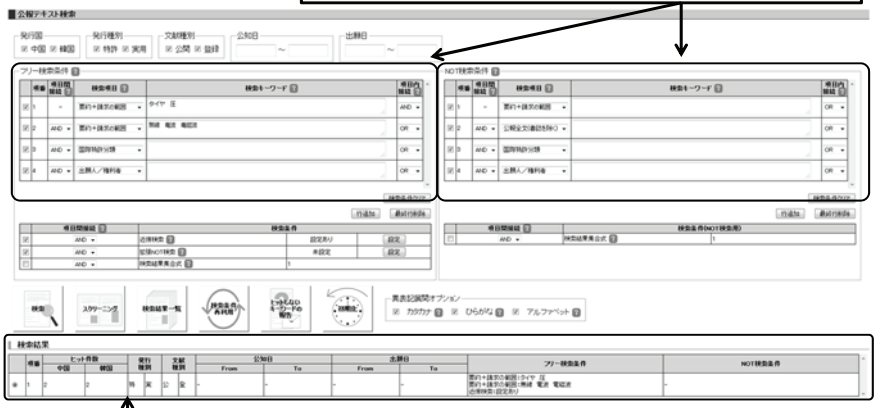
- 中韓文献翻訳・検索システム　（11月試行版リリース、1月本格リリース予定）



機械翻訳の活用　　（中韓→日機械翻訳）

- 中韓文献の公報テキスト検索

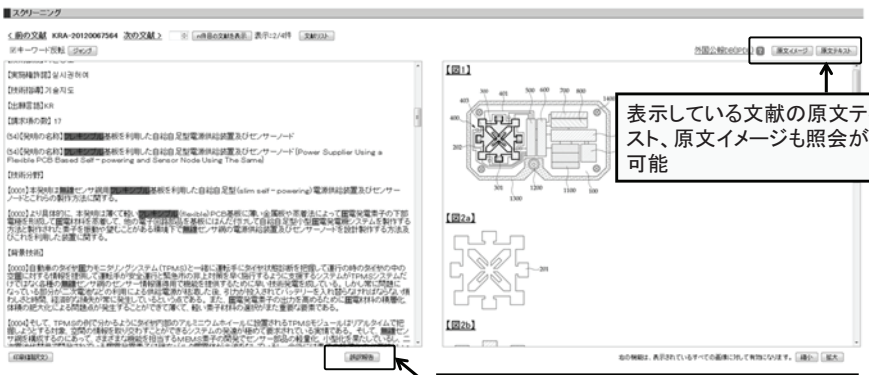
検索項目を選択し、検索キーワードを入力。フリー検索条件、ＮＯＴ検索条件により検索が可能



検索結果を表示

機械翻訳の活用　　（中韓→日機械翻訳）

- 中韓文献のスクリーニング画面



表示している文献の原文テキスト、原文イメージも照会が可能

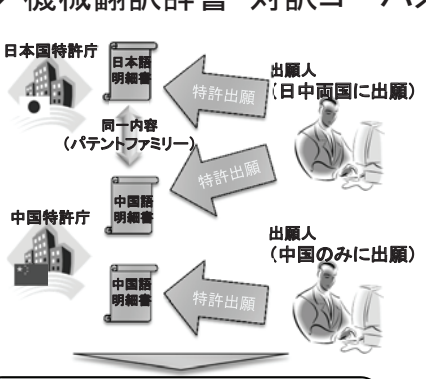
誤訳があった場合には、誤訳報告が可能

機械翻訳の活用　　（中韓→日機械翻訳）

- 中韓機械翻訳の誤訳報告画面

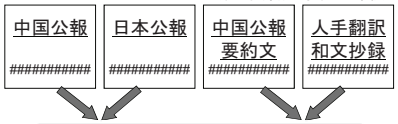
機械翻訳の活用　　（中→日機械翻訳）

- 機械翻訳辞書・対訳コーパスの作成



中国のみに出願された特許（パテントファミリーの存在しない特許）については、日本の審査官による調査を可能とするため要約の人手翻訳（和文抄録）を作成

（和文抄録作成事業の成果物）



対訳コーパス（文と文の対）の作成

翻訳辞書（単語と単語の対）の作成

No.	中国語	日本語
1	药物化学	医薬品化学
2	…	…

中韓文献翻訳・検索システム等の機械翻訳エンジンの翻訳精度向上に利用

以下のデータを作成済み
・対訳コーパス：　約１億文対
・翻訳辞書：　　約２００万語対

機械翻訳の活用　　（機械翻訳品質評価）

▸ 特許文献機械翻訳の品質評価手順

（http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm）

・相対評価

複数の機械翻訳システム間の翻訳精度比較や、特定の機械翻訳システムの翻訳品質の過去と現在の比較のため

・絶対評価

機械翻訳結果の特定の用途への有用性の判断等のため

（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段階で評価する。
（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、適訳、可訳、誤訳、不訳（未知語や訳漏れ）の４段階で評価する。
（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段階で評価する。

・フィードバックのための評価

機械翻訳システムの具体的な弱点を把握し、品質向上につなげていくため

【必須項目】【任意項目】【翻訳言語別任意項目】からなるチェックリストによる評価

機械翻訳の活用　　（機械翻訳品質評価）

▸ 特許文献機械翻訳の品質評価手順

相対評価、絶対評価の観点

(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段階で評価する。

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段階で評価する。

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。

機械翻訳の活用　　（機械翻訳品質評価）

▸ 特許文献機械翻訳の品質評価手順

(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段階で評価する。

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。

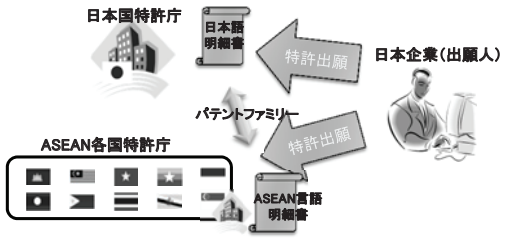
フィードバックのための評価の観点

機械翻訳によく見られるミスを収集したチェックリストを用いて評価する。

必須項目	「語の欠落」,「不要語の混入」, …
任意項目	「訳語のゆれ」,「不適切な文の切断」, …
翻訳言語別任意項目 （英日翻訳の場合）	文中のピリオドを適切に処理していない, 「:」、「;」を用いた文を適切に処理していない, :

機械翻訳の活用　　（ASEAN言語等）

▸ ASEAN言語等の機械翻訳



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

パテントファミリーをもとに
日本語とASEAN言語の文献対を作成

↓

日本語・ASEAN言語間の
対訳コーパス・辞書の作成

↓

日本語・ASEAN言語間の機械翻訳の精度向上

- ✓ ASEAN特許文献の日本語への機械翻訳
- ✓ 日本語特許文献のASEAN言語への機械翻訳
- ✓ 日本語・ASEAN言語間の機械翻訳ツール

招待講演 2

「多言語機械翻訳の研究開発動向」

第3回特許情報シンポジウム

「多言語機械翻訳の研究開発動向」

～特許翻訳と東京五輪向けの音声翻訳を絡めて～

情報通信研究機構(NICT)
隅田　英一郎
eiichiro.sumita@nict.go.jp

言葉の壁
のない
世界

特許訴訟の大半は翻訳！

⇒高速化・低コスト化が必須

武田薬品、米国アクトス特許訴訟控訴審で弁護士費用も勝ち取る

武田薬品工業は12月9日、同社と米国子会社の武田テバ・ファーマシューティカルズ・ノースアメリカ(TPNA)社が、ミラン社およびアルファーム社に対して提起していた糖尿病治療薬「アクトス」に関する特許侵害訴訟 で、米連邦巡回控訴裁が8日、同社とTPNA社が負担した弁護士費用の支払いをミラン社およびアルファーム社に命じたニューヨーク南部連邦地裁の判決を全 面的に支持する判決を下したと発表した。

この判決により、ミラン社が1140万ドル、アルファーム社が50万ドルに、それぞれ利子を加えた金額を支払うことになるという。

同社は、TPNA社とともに、アクトスの後発品を販売するため米国食品医薬品局(FDA)に簡略承認申請を行ったミラン社およびアルファーム社に対する特許侵害訴訟を提起。この訴訟で被告の両社は、アクトスの活性成分ピオグリタゾンに関する米国特許(特許番号6,087,777)の有効性と権利行使 可能性に対する主張を試みた。しかし、ニューヨーク南部連邦地裁は、両社の主張はいずれも法律(Hatch-Waxman法)が要求する誠実性に裏付けられず、さらに訴訟中にも違反行為を続けたと判断し、両社に弁護士費用の全額償還を命ずる判決を下し、さらに今回、米連邦巡回控訴裁の判決も、このニューヨーク南部連邦地裁の判決を支持したものとしている。

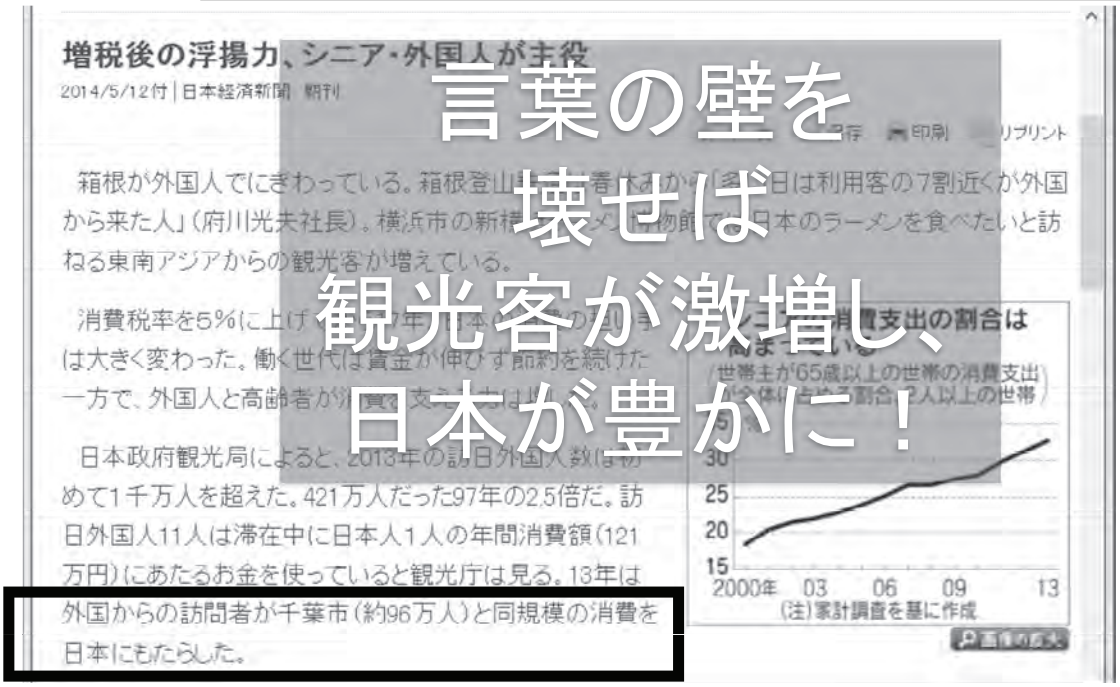
英語という言葉の壁が出現！

http://news.braina.com/2008/1209/judge_20081209_001____.html

観光は国内経済浮揚の二本柱の一つ

訪日外国人1000万人が1.4兆円を消費

（観光庁調査http://www.mlit.go.jp/common/001037717.pdf ）



増税後の浮揚力、シニア・外国人が主役

2014/5/12付 | 日本経済新聞 朝刊

箱根が外国人でにぎわっている。箱根登山[illegible]は利用客の7割近くが外国から来た人」（府川光夫社長）。横浜市の新横[illegible]で日本のラーメンを食べたいと訪ねる東南アジアからの観光客が増えている。

消費税率を5%に上げ[illegible]は大きく変わった。働く世代は賃金が伸びず節約を続けた一方で、外国人と高齢者が[illegible]

日本政府観光局によると、2013年の訪日外国人数は初めて1千万人を超えた。421万人だった97年の2.5倍だ。訪日外国人11人は滞在中に日本人1人の年間消費額（121万円）にあたるお金を使っていると観光庁は見る。13年は外国からの訪問者が千葉市（約96万人）と同規模の消費を日本にもたらした。

[illegible]支出の割合は

（世帯主が65歳以上の世帯の消費支出）

30
25
20
15
2000年 03 06 09 13

（注）家計調査を基に作成

言葉の壁を
壊せば
観光客が激増し、
日本が豊かに！

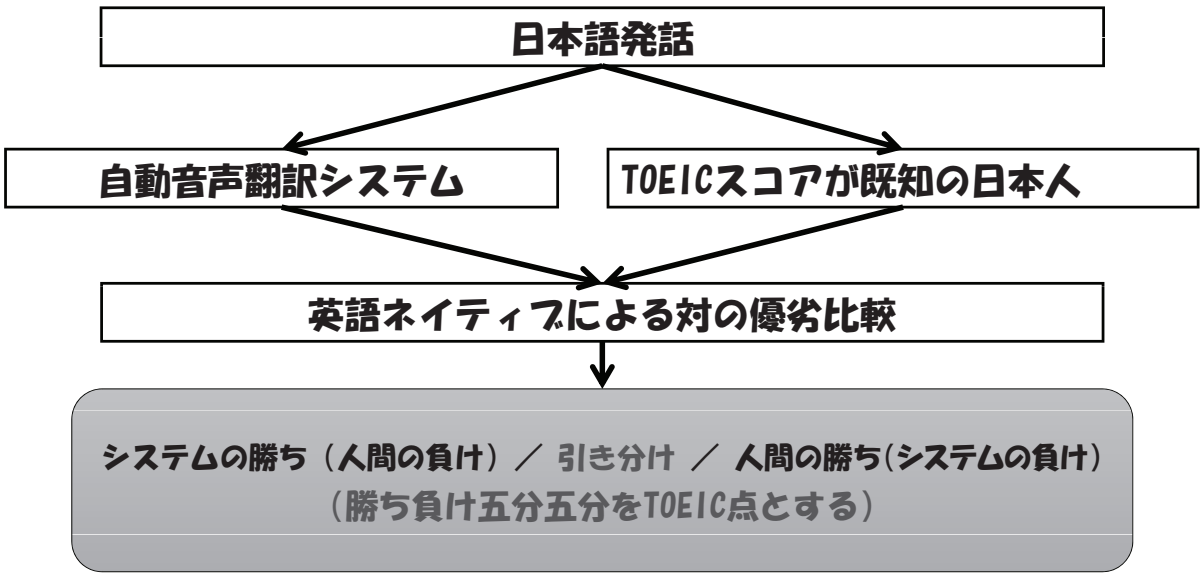
①今、動いている音声翻訳

観光/VoiceTra@ベトナム

- 2014年7月17日　ベトナムハノイの雑貨店に行きました。
- 良いお土産を見つけました。(ん？「アジアでは値引き交渉があたりまえだと聞いた」なぁ‥)
- しかし、ベトナム語はわからない。店員さんは英語は通じない‥
- そこで、iPhoneを取り出し‥‥

VoiceTraの翻訳性能はTOEIC600点の日本人に相当

いろいろなTOEICスコアの人間の音声翻訳能力と比較し、音声翻訳システムの能力がTOEICスコアでどのくらいの人間に相当するかを評価。



②音声翻訳の社会実装

グローバルコミュニケーション（GC）計画

グローバルコミュニケーション計画（GC計画）

http://www.soumu.go.jp/main_content/000285578.pdf

- 平成26年4月11日
  - **総務大臣が記者会見にて発表**
- Mission
  世界の「言葉の壁」をなくす
- Vision
  （1）グローバルで自由な交流の実現
  （2）日本のプレゼンス向上
  （3）東京オリンピック・パラリンピックでの「おもてなし」
- Action
  - 関係する企業や関係省庁等と連携、協力しながら、まずは6年間のロードマップを共有して取り組む
    - （プロジェクト1）病院、商業施設、観光地等における社会実証
    - （プロジェクト2）多言語音声翻訳の対応領域、対応言語を拡大するための集中的な研究開発投資
    - （プロジェクト3）2020年東京オリンピックにおける **社会実装**

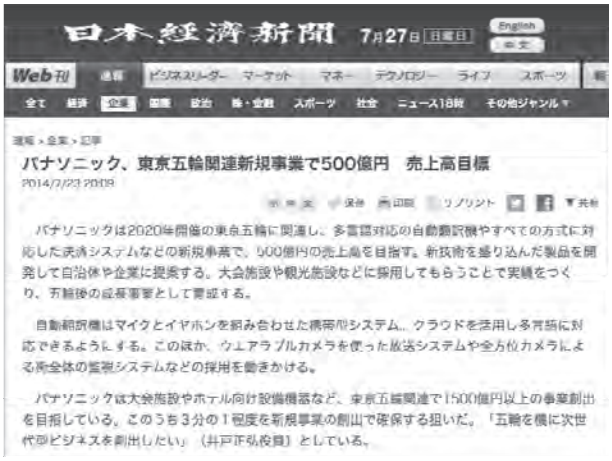
目標：ショッピング、交通、医療、ホテルなどで**「見慣れた普通の」のICT機器**として活用される

企業の動き

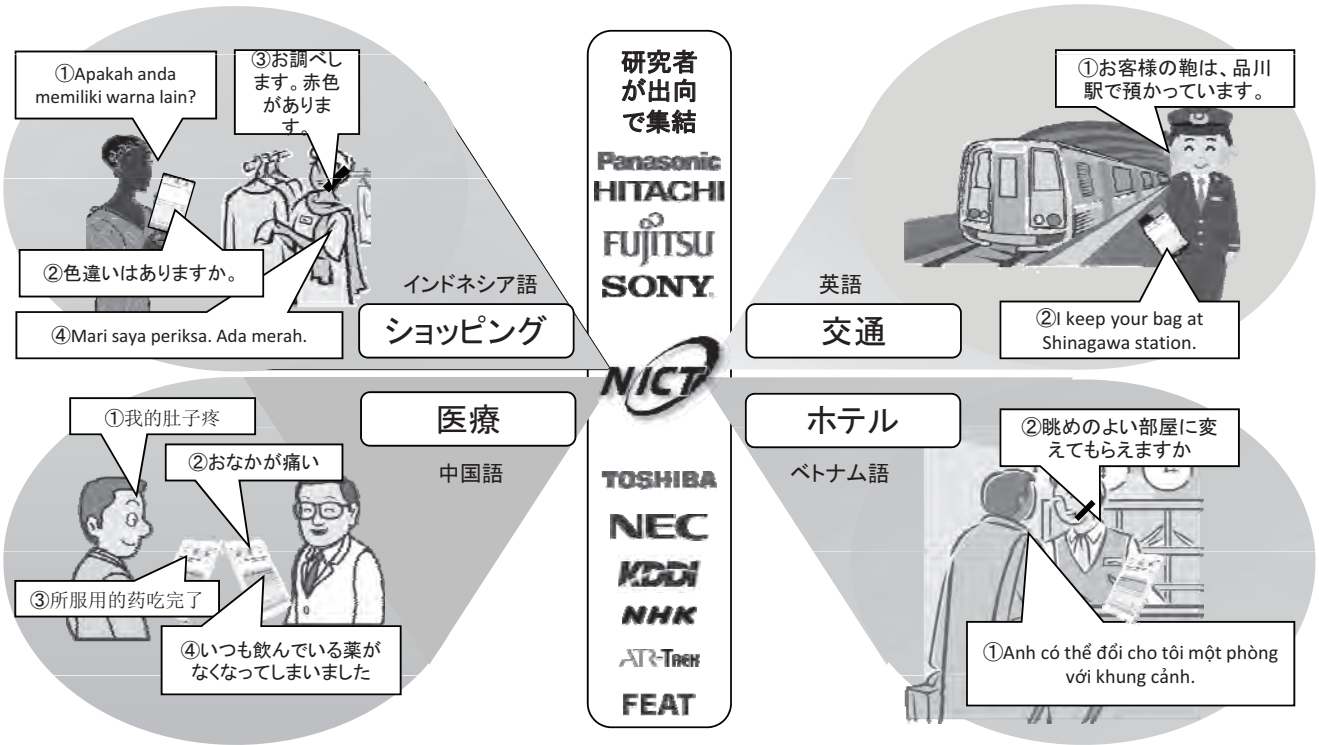
NTTドコモの機械翻訳の合弁会社設立（H26.9.29）

京急での活用 朝日新聞（H26.7.16）

パナソニックの計画 日経新聞（H26.7.23）



2020年に向けて10月より研究開始！



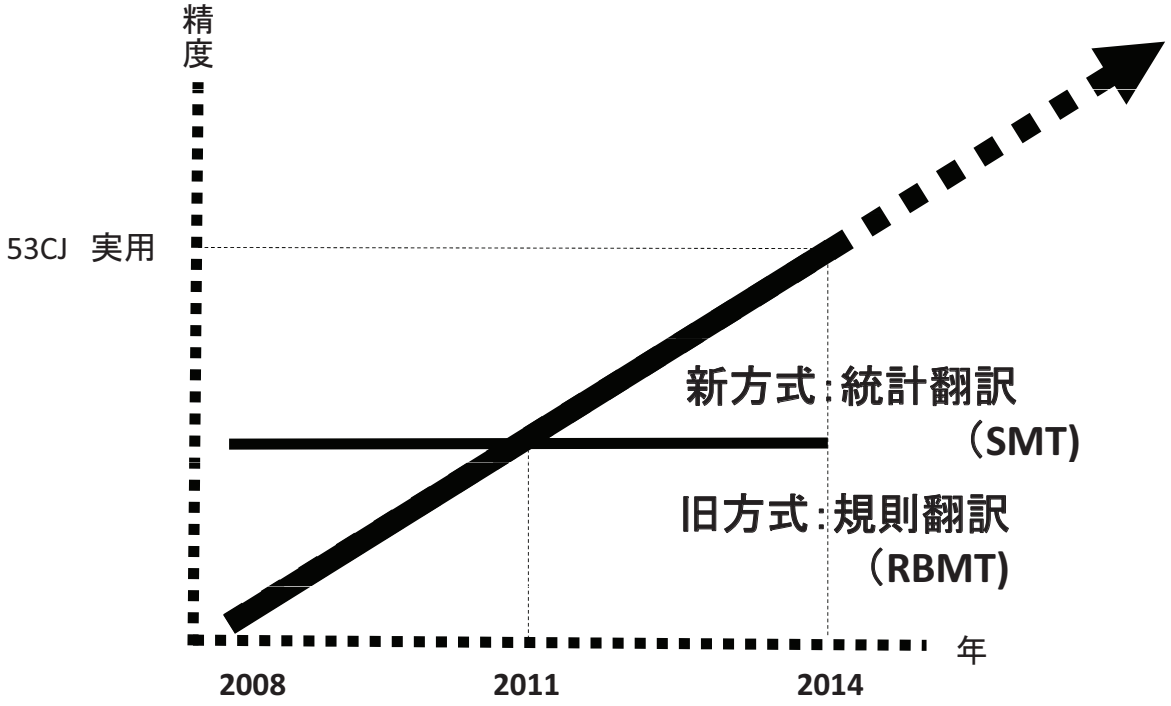
(A) 自動翻訳技術の現在

旧⇒新
規則翻訳(RBMT)⇒統計翻訳(SMT)

パラダイムシフトがおこっている

	規則翻訳(RBMT)	統計翻訳(SMT)
翻訳品質	中程度	• 専用システムは高品質 • 汎用システムは中程度
処理速度	速い	速くはない
多言語化	向かない	向いている
VENDER	クロスランゲージ、高電社、SYSTRAN	NICT、Google
ポイント	規則や辞書を作成する人間の専門家	大規模な対訳コーパスと自動学習

新翻訳方式は高精度



従来方式：規則翻訳（RBMT)

1. I bought a car with four **doors**
2. I bought a car with four **dollars**

文法の解析

1. (I (bought (a car (with four doors))))
2. (I (bought a car) (with four dollars)))

単語や語順の変換

必要な規則や辞書を
人間が作成

1. (私は ((4ドアの) 車を) 買った))
2. (私は (4ドルで) (車を買った))

文の生成

1. 私は4ドアの車を買った
2. 私は4ドルで車を買った

新方式：統計翻訳（SMT）

1.	京都駅はどこですか	Could you direct me to Kyoto station?
2.	駅はどこですか	Where is the station?
3.	トイレはどこですか	Where is the rest room?
4.	タクシー乗場はどこですか	Where is the taxi stand?
5.	ここはどこですか	Where am I?

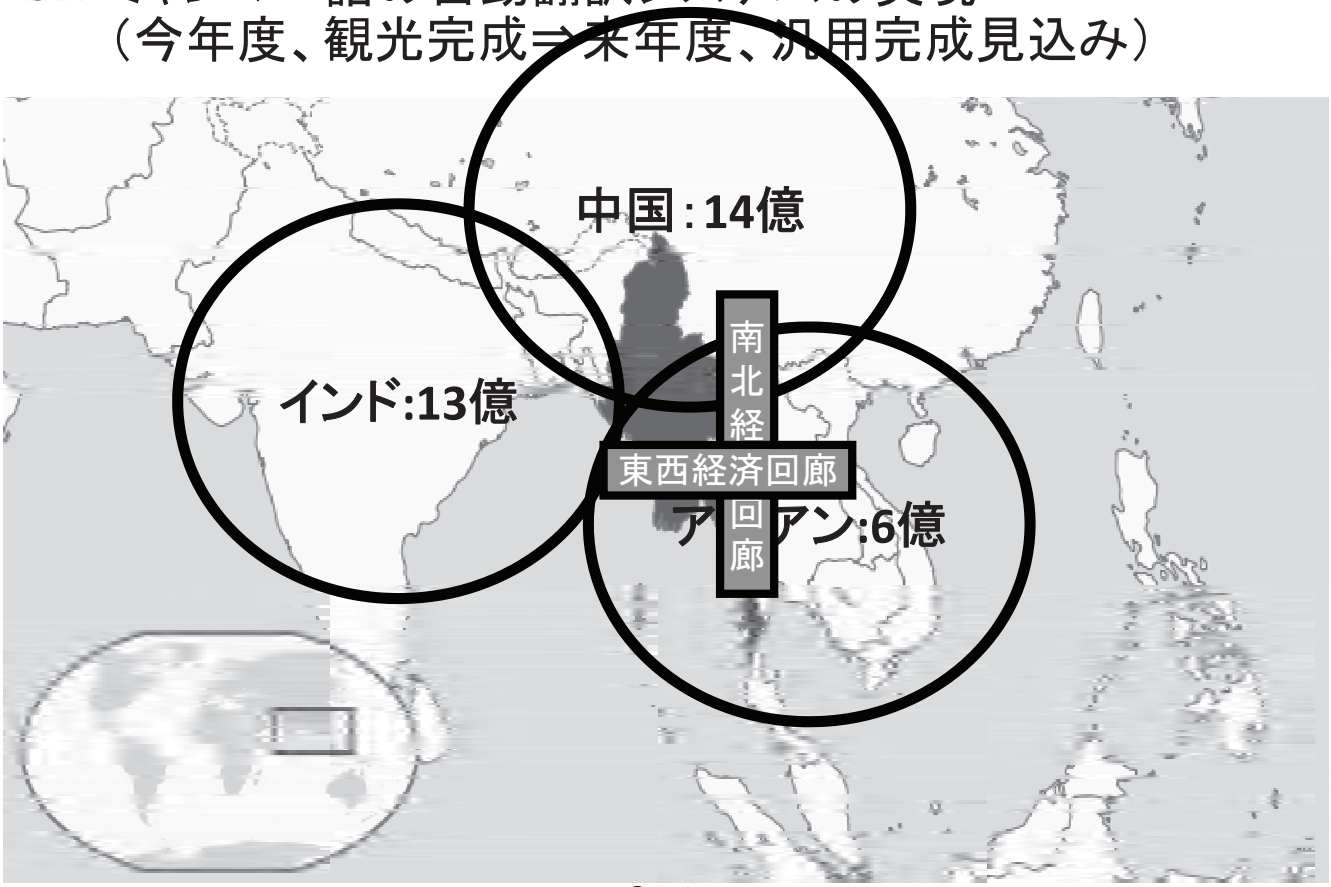
↓

確率付対訳辞書の自動作成

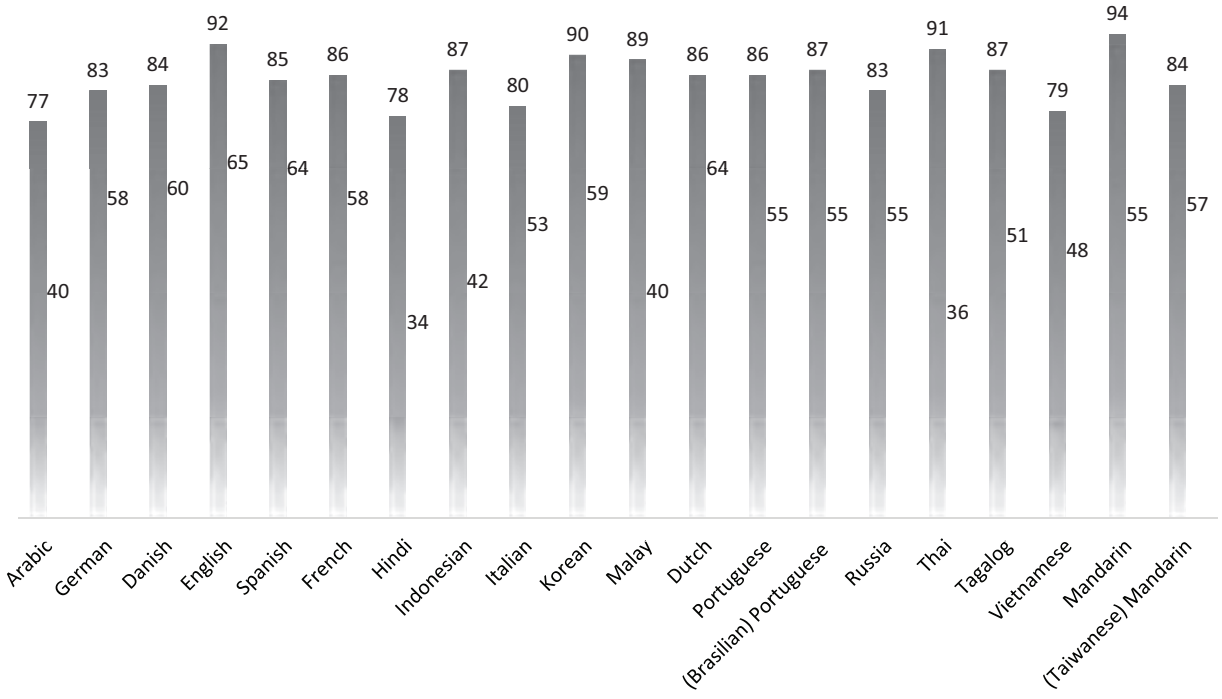
↓

どこですか	→	Where is	3/5=60%
どこですか	→	Could you direct me to	1/5=20%
どこですか	→	Where am	1/5=20%

ミャンマー語の自動翻訳システムの実現
（今年度、観光完成⇒来年度、汎用完成見込み）



統計翻訳（SMT）は多言語化を高精度で実現
（旅行分野で20言語から日本語へ翻訳して評価）



統計翻訳（SMT)は売れています。

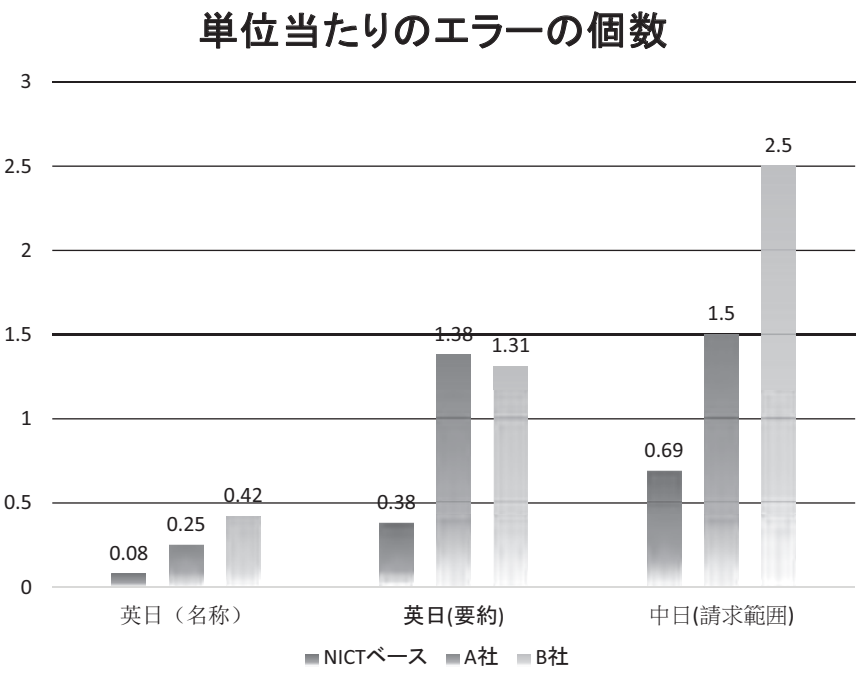
↓

（買う前に各組織は評価するので、）
第三者によってSMTが良いことが証明されている。

統計翻訳（SMT）は特許も高精度

～NICTの自動翻訳を導入した(株)日本発明資料による評価～

※AAMTジャーナル56号（2014年6月）からの引用・編集

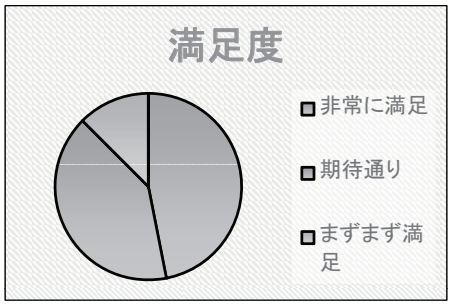


自分でやるには

講習会・教科書

Alagin自然言語処理セミナー

- 2014年3月開催し、企業や大学から57名参加し好評だった。
- 2015年3月も新宿で開催の準備中。



コロナ社

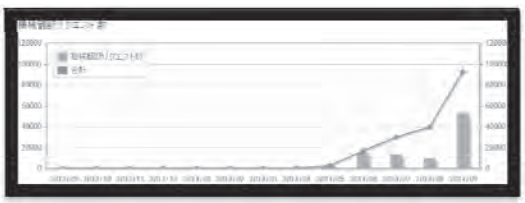
日本発のツール・言語資源

種類	詳細	公開場所・時期
プログラム	フレーズアライメントツール pialign - Phrasal ITG Aligner	http://www.phontron.com/pialign/
	教師なし識別構文解析ツール lader - Latent Derivation Reorderer	http://www.phontron.com/lader/
	翻訳デコーダ cicada	http://www2.nict.go.jp/univ-com/multi_trans/cicada/
	言語モデルの作成ツール expgram	http://www2.nict.go.jp/univ-com/multi_trans/expgram/
対訳コーパス・辞書	日英特許対訳コーパス（約300万文）	NTCIR
	日英新聞記事対訳コーパス（JENAAD）	http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/jea/index-ja.html
	Wikipedia日英京都関連文書対訳コーパス	http://alaginrc.nict.go.jp/WikiCorpus/
	日英中（対訳）基本文コーパス	http://nlp.ist.i.kyoto-u.ac.jp/index.php
	IWSLTの訓練・試験コーパス	https://alaginrc.nict.go.jp/resources/nictmastar/resource-info/abstract.html#B-1

もっと手軽にやるには

「みんなの自動翻訳@TexTra®」
https://mt-auto-minhon-mlt.ucri.jgn-x.jp/

自動翻訳をみんなで育てるサイトです。



利用が伸びています。

（B）自動翻訳技術のちょっと未来

性能向上の要点は2つ

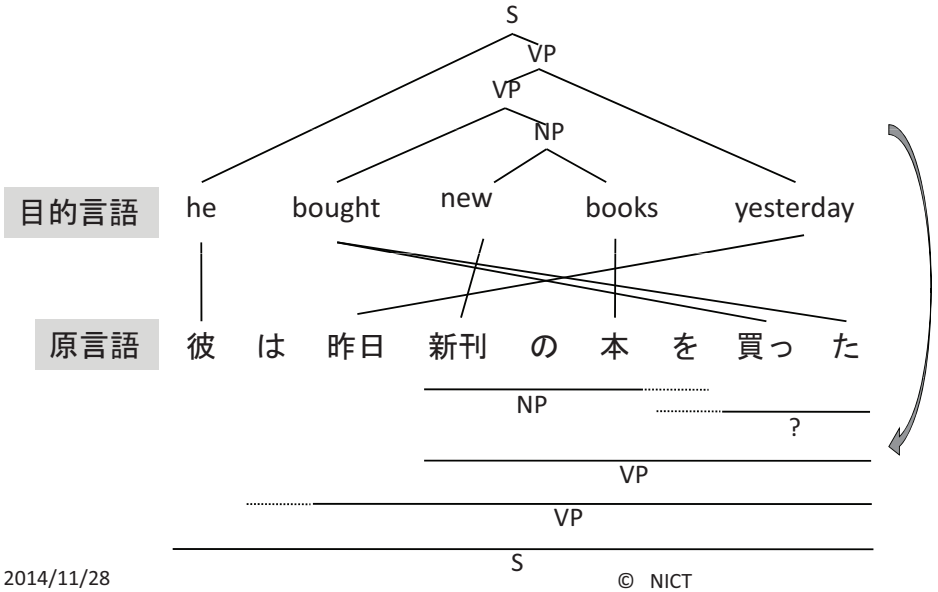
精度 = SMT(アルゴリズム, 対訳量)

毎年、進化中

The more the better!

アルゴリズムの話☺
文法を自動獲得する多言語翻訳

- 従来法
  - 原言語･目的言語の構文解析要
- 提案法
  - 目的言語の構文解析のみでよい。 ∴ 日本語(英語)構文解析はある ∴ 日英⇔多言語翻訳可能



- 部分構造をプロジェクション
- 統計的に不足部を補完＆曖昧性解消
- Pitman-Yor プロセスで原言語CFGモデルを作成

従来法	34.28% BLEU
提案法	36.22% BLEU

対訳量の話☺
日本全体の対訳データ

DISKに格納されて忘れられている5億文/年の超大規模対訳データ

（C）自動翻訳技術のさらに未来

統計翻訳の進化

ありがとうございました ☺

招待講演 3

「特許情報検索サービスにおける機械翻訳の活用」

AAMT/Japio特許翻訳研究会
第3回特許情報シンポジウム（11/28）

特許情報検索サービスにおける機械翻訳の活用

JPDS　日本パテントデータサービス㈱
企画室　早川浩平

JPDS 日本パテントデータサービス株式会社

目次

JPDS 日本パテントデータサービス株式会社

会社概要

会社名　日本パテントデータサービス株式会社
Japan Patent Data Service Co., Ltd.
代表者　代表取締役 仲田 正利

設 立	昭和63年10月
資本金	2,000万円
本 社	東京都港区西新橋2-8-6 住友不動産日比谷ビル
拠 点	東京本社・名古屋営業所・大阪営業所・九州営業所
社員数	43名
事業内容	特許情報検索サービス・特許管理システム 特許調査サービス・知財研修セミナー 他、知的財産に関連する事業

3

JPDS 日本パテントデータサービス株式会社

沿 革

1988年10月	特許情報の国内外提供を目的として設立
1992年12月	JP-ROM電子特許公報検索システム発表
1993年5月	日本最初の電子特許データの編集加工販売を開始
1997年11月	JP-ROMクライアントサーバーモデル発表
2000年11月	JP-NET 500万件の全文明細書検索サービス開始
2003年6月	New Client Server System（NewCSS）販売開始
2005年12月	ぱっとマイニングソフト発売（ワイズシステム社契約）
2006年4月	知的財産研修事業部 発足
2008年7月	知財ワークス株式会社設立
2009年11月	海外DB翻訳機能サービス開始
2012年1月	中国知識産権出版社と代理店契約締結
2012年2月	インドClairvolex社と代理店契約締結
2013年4月	知財システム開発株式会社設立　PATAS特許事務管理システム販売開始

4

JPDS 日本パテントデータサービス株式会社

商品・サービス紹介

- 特許情報検索サービス
  「JP-NET/NewCSS」

- 特許事務管理システム
  「PATAS」

- 特許調査サービス

- 知的財産研修セミナー
  「知財寺子屋」

5

JPDS 日本パテントデータサービス株式会社

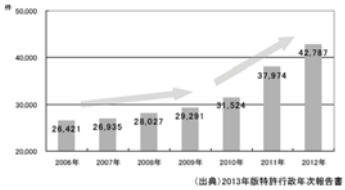
海外特許を取り巻く状況

海外特許の重要性・調査の必要性が上昇！

- 生産・市場はグローバル化
- 特許権は属地主義を採択
- 事業展開を行う国の権利の重要性向上
- 国際出願の比率の向上

→出願時・調査時に海外特許を無視
出来ない状況になっている

■ 日本からの国際特許（PCT）出願件数の推移



（出典）2013年版特許行政年次報告書

6

JPDS 日本パテントデータサービス株式会社

特許情報と機械翻訳

各国特許庁が機械翻訳に対して積極的に取組みを実施

- EPO→Patent Translate(Google)
  （英語⇔31言語　独・仏語⇔27言語）
  ※EMWでは2012年、2013年に機械翻訳をテーマにピックアップ

- WIPO→Google Translate(Google)
  Microsoft translate(Microsoft)
  （世界10ヶ国の出願 各対応言語へ翻訳）
- WIPO→TAPTA（Moses） タイトル・要約の技術用語辞書を作成し、翻訳
  →CLIR 各国語の翻訳+同義語辞書（12言語）
  →WIPO Pearl 同じ技術分野に関連する技術用語をアシスト
- JPO→中国・韓国特許の機械翻訳

7

JPDS 日本パテントデータサービス株式会社

特許情報と機械翻訳

- **各国特許庁が機械翻訳の研究に注力**
  →国際出願の拡大背景
  →一般ユーザーの調査環境の改善（無駄な出願・訴訟の抑止）
  →審査官の審査の品質向上

- **GoogleやMicrosoftが特許分野に注目**
  →テキスト量が膨大
  →技術用語が多く、新しい言葉がでる
  →対応する完全な人手翻訳が存在する
  →統計データ翻訳にとって魅力的な分野？

特許情報と機械翻訳は切り離せない関係になっている

8

JPDS 日本パテントデータサービス株式会社

JPDSの特許情報検索サービス概要

サービス名称 「JP-NET」「NewCSS」

- 専用ブラウザの高速検索・高速表示
- 検索結果を豊富な形式で出力
- 安価な定額制
- 信頼できる導入実績
- 安心のサポート

9

JPDS 日本パテントデータサービス株式会社

JPDSの特許情報検索サービス概要

- 日本の四法（特許・実案S46～、意匠S39～、商標 登録1号～）
- 海外80ヶ国の情報提供（収録範囲は下図参照）

収録国	収録年度	収録言語	備考
アメリカ(US)	1985～	英語	
ヨーロッパ(EP)	1978～	英・仏・独語	仏語・独語MT
国際公開(WO)	1978～	10ヶ国語	
中国(CN)	1985～	英語	要約マニュアル翻訳 全文MT
ドイツ(DE)	2007～	英・独語	独語MT 原文の表示可
イギリス(GB)	1990～	英語（要約）	
韓国(KR)	1978～	英・韓語	要約（英語） 全文（韓国語）
その他 73カ国 (DE,GB等含む)	1978～	英語（要約）	DOCDB

機械翻訳機能の対応国
串刺し検索対応国

10

JPDS 日本パテントデータサービス株式会社

海外特許検索サービスの歴史

2008年　US特許の提供からUS/EP/WOの提供へ拡大し、
　　海外DBサービスを開始
2009年　機械翻訳サービスを導入・提供開始
　　リーガルステータスデータ提供開始
2010年　和訳印刷オプション　提供開始
　　中国特許の英語全文データの提供開始
　　34ヶ国のDOCDBデータの収録
2011年　ドイツデータ、イギリスデータ提供開始
2012年　80ヶ国のDOCDBデータの収録
　　EP特許全文検索データ提供開始
2013年　インド特許のジャーナルデータ・台湾のガゼットデータ収録開始
2014年　ファミリーベース検索機能　提供開始

11

JPDS 日本パテントデータサービス株式会社

海外検索サービスの目標

【課題】

権利調査は各企業の事業展開に対応する必要性があり、
検索サービスの収録国拡大は必須。
データを入手できても、様々な言語を扱える人は多くはない。

- 全ての収録国の情報を英語に翻訳し、串刺しで検索が行える環境の整備を行う
- 検索した結果の英語は日本語に翻訳し、閲覧可能にすることで調査効率の向上を目指す

➡ **機械翻訳の活用は必要不可欠！**

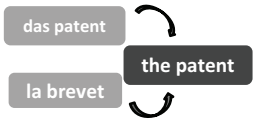
12

JPDS 日本パテントデータサービス株式会社

機械翻訳の活用範囲

- **機械翻訳したデータの収録・検索（EP・ドイツ）**
  →データ更新段階でドイツ語・フランス語の案件の全件を英語に翻訳してから収録。

- **公報の表示時に機械翻訳**
  →英語のデータを日本語に翻訳



- **公報の出力で機械翻訳したデータを出力**
  →Excel、PDF、紙出力の活用

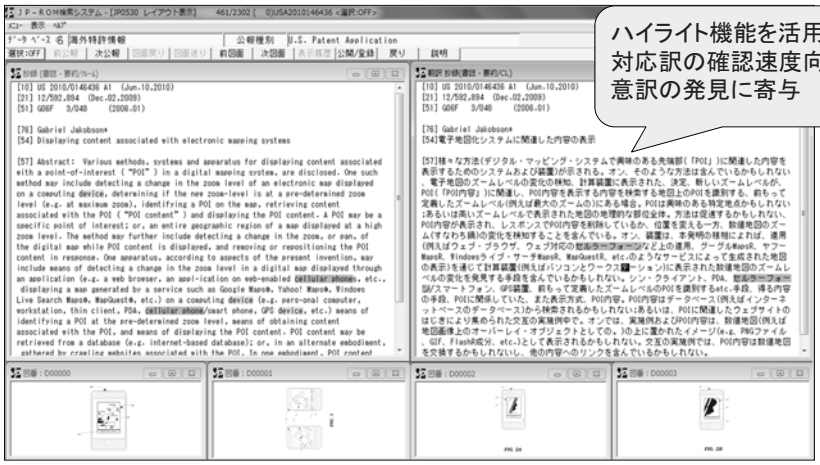
独仏語→英語と、英語→日本語の翻訳ソフトは別システム

13

JPDS 日本パテントデータサービス株式会社

実際の活用方法

- 公報内容の翻訳（レイアウト表示機能での対訳比較表示）

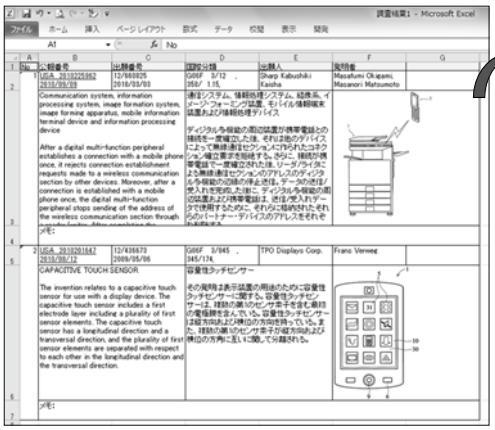


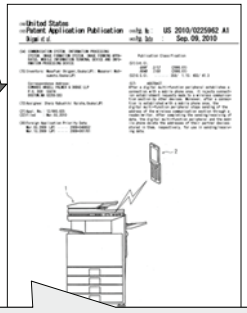
14

JPDS 日本パテントデータサービス株式会社

実際の活用方法

- Excelで和訳と原文の並列ダウンロード（調査結果の共有に活用）





気になった案件はPDFリンクで公報全文を表示することで、即確認が可能。

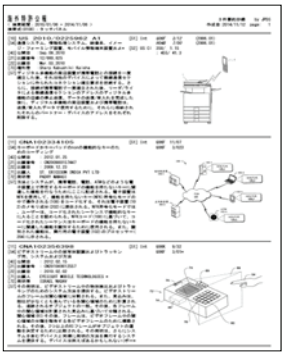
15

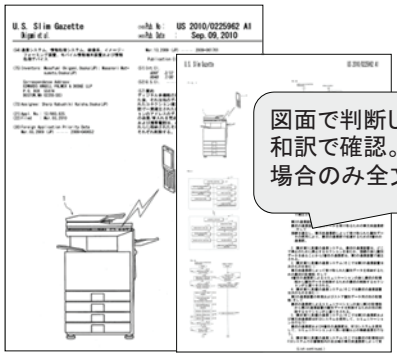
JPDS 日本パテントデータサービス株式会社

実際の活用方法

- 和訳出力物のSDI回覧（様々な抄録形式）

※SDI･･･Selective Dissemination of Informationの略。定期的に選択した情報（分類や出願人等）を普及するために関係者に発信（回覧）する業務。





図面で判断し、必要な案件を和訳で確認。さらに必要な場合のみ全文を取得依頼。

抄録形式例＞3件抄録、1件抄録、ガゼット（全図面抄録）、スリムガゼット、レイアウト･･･他

16

JPDS 日本パテントデータサービス株式会社

利用者の声

- **メリット**
  - 日本語なので見やすい（閲覧効率向上）
  - 比較表示での辞書のような感覚の利用が可能
  - 開発者への利用もさせやすい
  - 翻訳依頼に出す前の確認が可能

  1つのシステムで完結出来る事が大きなメリット

- **デメリット**
  - 精度が低い ＝ 無料の翻訳サービスより悪い場合も多い
  - スピードが遅い

- **要望**
  - 自分の関連技術は辞書を独自作成したい

17

JPDS 日本パテントデータサービス株式会社

提供者側の声

- **メリット**
  - エンドユーザーへの評価向上に直結する
  - サーバーに導入する事で、全ユーザーに 同条件で提供出来る
  - ユーザーの生の評価を得て、
    今後の辞書改善の情報が集まりやすい。

- **デメリット**
  - スピードが遅い
  - 精度の改善がしづらい
  - 誤訳の対応が難しい
  - 新規出現ワードの対応が難しい
  - ソフト自体の改善に自由度がない
  - メンテナンスの手間

提供側の工夫

翻訳する箇所を個別に設定可能にすることで、サーバー負荷とユーザー待機時間の削減を実現

18

JPDS 日本パテントデータサービス株式会社

特許翻訳　誤訳例

1．特許用語独特の言い回しで翻訳されるべき語で、一般用語の翻訳を返す語がある。

例＞said

→正訳「該」、「前記（の）」　　誤訳「言った」

※通例、”the said”で使われることが多くその場合は正訳を当てるべき。

2．動詞の過去分詞系が動詞扱いになる、英語にはない言葉が追加される

例＞Hand guided roller（EP1096072）

→正訳「ハンドガイドローラ」、「手でガイドする（した）ローラ」

誤訳「手はインキを移しローラをガイドした」

19

JPDS 日本パテントデータサービス株式会社

特許翻訳　誤訳例

3．同じ綴りで複数の意味を持つ多義語において文脈に沿わない翻訳を返す語がある。

例）品詞が同じグループ

art（技術vs.芸術）、critical（危機的vs.批判的）、amount（合計vs.金額）等

4．品詞が異なる単語については文脈に関係なく正しく訳されるべき。

例）品詞が異なるグループ

patient（患者vs.忍耐強い）

close（近いvs.終了、閉じる）※to closeやclosedの場合は正しく訳せる傾向あり

light（光vs.軽い）※但し、同じ文献内でも訳語が違い、正しく訳せている箇所もある

20

JPDS 日本パテントデータサービス株式会社

特許翻訳　誤訳例

5. 訳出されず英単語のまま返ってくる語がある。
例）endothermically（吸熱）

※辞書の強化を図るべき。
明細書上のテキストの綴りが誤っている場合は致し方ないと思われるが、一部Googleでは綴りが誤っていてもそれと近い綴りの単語と認識して翻訳を返す事例がある。

21

JPDS 日本パテントデータサービス株式会社

検索サービスが機械翻訳に求めるもの

- 翻訳のスピードアップ
- 辞書の充実による精度向上

- 分類毎による使用辞書の自動変更
- 個々のユーザーよる辞書構築
- 多義語の変換（選択機能）

22

JPDS 日本パテントデータサービス株式会社

アメリカ英語とイギリス英語

- 機械翻訳/人手翻訳関わらず各国言語⇔英語の翻訳を行う場合、英語の種類に注意が必要。どちらが利用されているか確認した。

検索期間　：1978/01/01 ～ 2014/10/16　　作成日：2014/10/28　page:1
公報種別　：WO, EA, EB, US(A, B, P, D, RE, H, T), CN(A, C, B, U, Y), KR(A, B, U, Y), GB(A, B), DE(A, B, Y), DD(A, B)

式No	検索コマンド	特許	実案
#001	//アメリカ英語とイギリス英語の綴り比較	0	0
#002	//上段がアメリカ英語、下段がイギリス英語	0	0
#003	//	0	0
#004	HTX=FIBER	2050701	141518
#005	HTX=FIBRE	681464	102172
#006	T=#4 + #5	2417706	182641
#007	//	0	0
#008	HTX=CENTER	4658251	422491
#009	HTX=CENTRE	1533581	461123
#010	T=#8 + #9	5702383	700954
#011	//	0	0
#012	HTX=COLOR	2863105	154199
#013	HTX=COLOUR	605032	47795
#014	T=#12 + #13	3167748	173089

機械翻訳の場合・・・言語選択時に設定が可能。しかし、構文の記載や辞書構築の都合等でアメリカ英語に設定していてもイギリス英語が翻訳される場合もあった。

23

JPDS 日本パテントデータサービス株式会社

アメリカ英語とイギリス英語

- アメリカ特許（全て人手翻訳）でも1割程度のイギリス英語の利用が見られる。→これらの検索もOR検索する必要性がある。

検索期間　：1978/01/01 ～ 2014/10/16　　作成日：2014/10/28　page:1
公報種別　：US(A, B, P, D, RE, H, T)
検索式(0023)：米語vs英語

式No	検索コマンド	特許	実案
#001	//アメリカ英語とイギリス英語の綴り比較	0	
#002	//上段がアメリカ英語、下段がイギリス英語	0	
#003	//	0	
#004	HTX=FIBER	924674	
#005	HTX=FIBRE	108219	
#006	T=#4 + #5	1001554	
#007	//	0	
#008	HTX=CENTER	2380756	
#009	HTX=CENTRE	175629	
#010	T=#8 + #9	2515513	
#011	//	0	
#012	HTX=COLOR	1341297	
#013	HTX=COLOUR	121288	
#014	T=#12 + #13	1414604	

24

JPDS 日本パテントデータサービス株式会社

まとめ

- 特許情報と機械翻訳は密接な関係が続き、今後も必要性は高くなる
- 機械翻訳は一層の精度向上が求められる

- 検索サービスでは、翻訳＋αが必須の時代に！ユーザーニーズとのマッチの実現がカギ

様々な情報を収集し、常にニーズキャッチが必要

25

JPDS 日本パテントデータサービス株式会社

ご静聴ありがとうございました

知財戦略の総合サポート

JPDS 日本パテントデータサービス株式会社
ホームページ http://www.jpds.co.jp

〒105-0003 東京都港区西新橋２－８－６住友不動産日比谷ビル tel:03(3580)8021 fax:03(5512)7810
〒460-0008 名古屋市中区栄２－１０－１９ 名古屋商工会議所ビル tel:052(219)4561 fax:052(219)4581
〒550-0004 大阪市西区靭本町１－７－１８ ビーイングビル tel:06(6448)7401 fax:06(6459)4588
〒812-0013 福岡市博多区博多駅東２－６－２３ 博多駅前第2ビル tel:092(405)2341 fax:092(405)2342

E-mail 東京 tokyo-sales@jpds.co.jp 名古屋 nagoya-sales@jpds.co.jp
大阪 osaka-sales@jpds.co.jp 九 州 kyushu-sales@jpds.co.jp

26

研究会報告 1

「パテントファミリーからの専門用語対訳辞書の構築」

パテントファミリーからの専門用語対訳辞書の構築

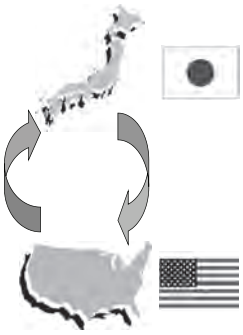
宇津呂武仁　董麗娟　龍梓　山本幹雄

筑波大学 システム情報系
筑波大学大学院 システム情報工学研究科

2014/11/28　第3回特許情報シンポジウム
＠キャンパス・イノベーションセンター東京

研究の背景(1)

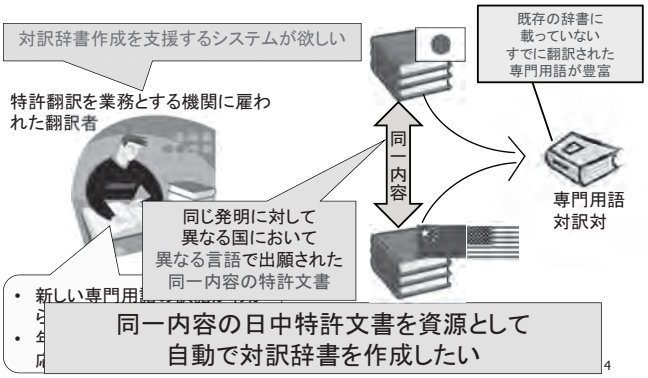
- 特許文書の翻訳は，他国への特許申請や特許文書の言語横断検索などといったサービスにおいて，不可欠である．
- 特許文書翻訳の過程において，専門用語の対訳辞書は重要な情報源である．



研究の背景(1)

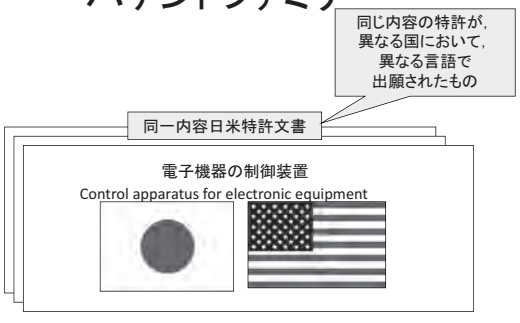
- 特許文書の翻訳は，他国への特許申請や特許文書の言語横断検索などといったサービスにおいて，不可欠である．
- 特許文書翻訳の過程において，専門用語の対訳辞書は重要な情報源である．

研究の背景（2）



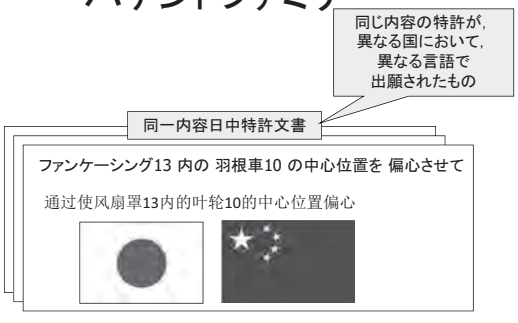
同一内容の日中特許文書を資源として自動で対訳辞書を作成したい

パテントファミリー



同一内容日米特許文書には専門用語が豊富
→これを資源として自動で対訳辞書を作成したい

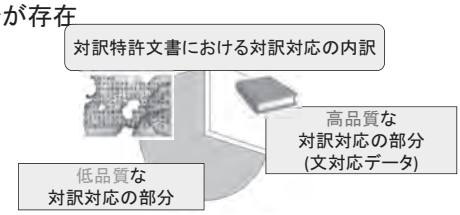
パテントファミリー



同一内容日中特許文書には専門用語が豊富
→これを資源として自動で対訳辞書を作成したい

対訳対応の内訳

- 対訳辞書作成の資源
- パテントファミリーの対訳特許文書
- 高品質な対訳対応の部分と低品質な対訳対応の部分が存在



高品質・低品質な対訳対応の部分の特徴

- 高品質な対訳対応の部分
  - 文対応が1対1
  - 文の内容が対応
- 高品質な部分
  - 対訳文対の抽出が相対的に容易
  - 対訳辞書の作成が相対的に容易
- 低品質な部分
  - コンパラブルコーパスとしては高品質
  - 言語資源としては豊富

パテントファミリーからの専門用語対訳辞書の構築 ――― タスクの内訳 ―――

言語資源の品質	タスク	日米	日中
高品質（対訳文対応有）	対訳対抽出	［森下他 2010］（信学会論文誌）	［董・龍他 2014］（言語処理学会年次大会）
	同義対訳対抽出	［梁他 2012］（言語処理学会年次大会）	［龍・董他 2014］（言語処理学会年次大会）
低品質（対訳文対応無）	対訳対抽出	［豊田他 2013］（NL研・言語処理学会年次大会）	

パテントファミリーからの専門用語対訳辞書の構築 ――― タスクの内訳 ―――

言語資源の品質	タスク	日米	日中
高品質（対訳文対応有）	対訳対抽出	［森下他 2010］（信学会論文誌）	［董・龍他 2014］（言語処理学会年次大会）
	同義対訳対抽出	［梁他 2012］（言語処理学会年次大会）	［龍・董他 2014］（言語処理学会年次大会）
低品質（対訳文対応無）	対訳対抽出	［豊田他 2013］（NL研・言語処理学会年次大会）	

パテントファミリーからの専門用語対訳辞書の構築 ――― タスクの内訳 ―――

言語資源の品質	タスク	日米	日中
高品質（対訳文対応有）	対訳対抽出	［森下他 2010］（信学会論文誌）	［董・龍他 2014］（言語処理学会年次大会）
	同義対訳対抽出	［梁他 2012］（言語処理学会年次大会）	［龍・董他 2014］（言語処理学会年次大会）
低品質（対訳文対応無）	対訳対抽出	［豊田他 2013］（NL研・言語処理学会年次大会）	

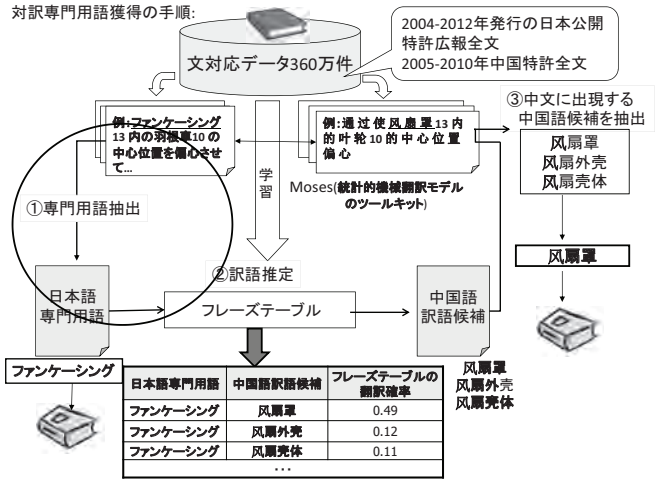
日中パテントファミリーから抽出した対訳文を用いた専門用語の訳語推定

†董麗娟 † 龍梓 †豊田樹生
‡宇津呂武仁 ‡山本幹雄 § 三橋朋晴

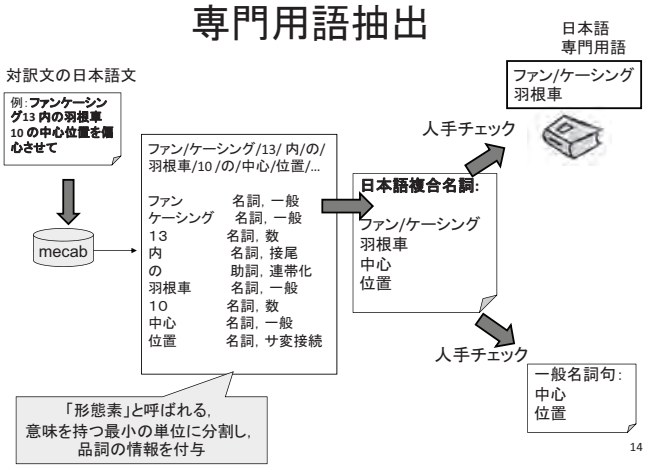
†筑波大学大学院 システム情報工学研究科
‡筑波大学 システム情報系
§ 日本特許情報機構

2014/03/18 言語処理学会 第20回年次大会

対訳専門用語獲得の手順:

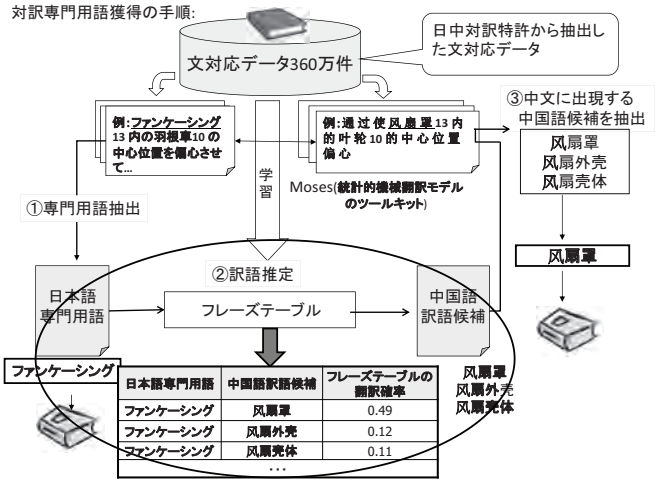


専門用語抽出

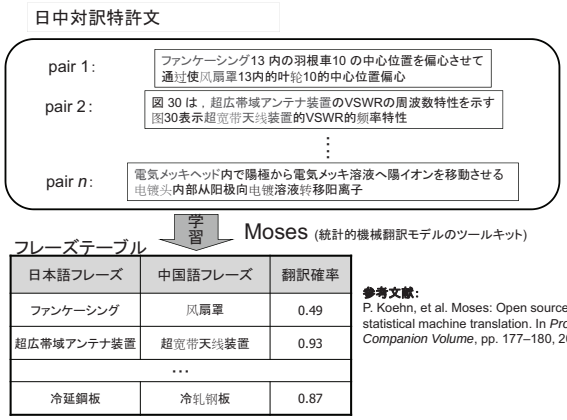


14

対訳専門用語獲得の手順:



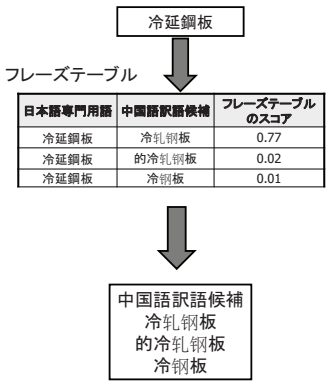
翻訳モデル学習



参考文献:
P. Koehn, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

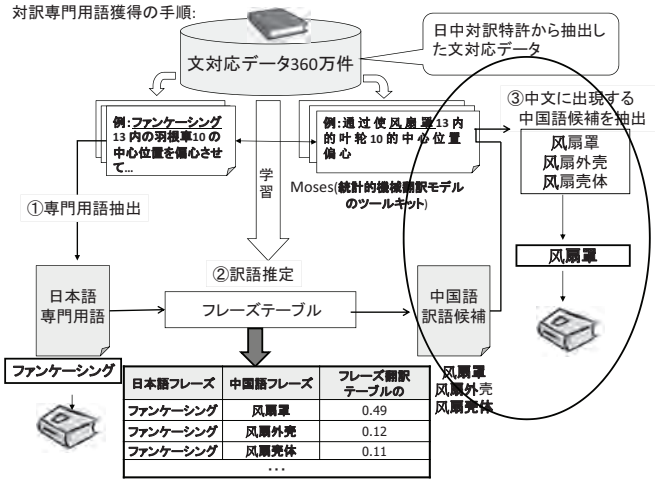
16

訳語推定



17

対訳専門用語獲得の手順:



中国語訳語候補抽出

日中対訳文

高/成形/性/冷/延/鋼板/と/その/製造/方法/…

具有/优良/可/成形性/和/高屈强/比/的/冷轧/钢板/及/其/制造/方法/…

フレーズテーブル

日本語専門用語	中国語訳語候補	翻訳確率
冷延鋼板	冷轧钢板	0.77
冷延鋼板	的冷轧钢板	0.02
冷延鋼板	冷钢板	0.01

中国語訳語候補	翻訳確率
冷轧钢板	0.77
的冷轧钢板	0.02

評価実験

日中対訳特許文

…三原色を用いてフルカラーの有機ELパネルを作製…
…**使用三原色制造全彩色有机EL面板**…

形態素単位の日本語文

…/三原色/を/用い/て/フルカラー/の/有機/EL/パネル/を/作製/…

形態素単位の中国語文

…/**使用**/**三原色**/**制造**/**全**/**彩色**/**有机**/**EL**/**面板**/…

文字単位の中国語文

…/**使**/**用**/**三**/**原**/**色**/**制**/**造**/**全**/**彩**/**色**/**有**/**机**/**EL**/**面**/**板**/…

評価結果

翻訳確率1位の訳語候補の再現率・適合率・F値(%)

	形態素単位	文字単位
再現率	96.3 (497/516)	95.9 (495/516)
適合率	97.8 (497/508)	96.9 (495/511)
F値	97.0	96.4

- 両者はほぼ97%程度の適合率およびF値を達成した
- 両者の誤りの傾向は異なっている

パテントファミリーからの専門用語対訳辞書の構築
――― タスクの内訳 ―――

言語資源の品質	タスク	日米	日中
高品質（対訳文対応有）	対訳対抽出	［森下他 2010］（信学会論文誌）	［董・龍他 2014］（言語処理学会年次大会）
	同義対訳対抽出	［梁他 2012］（言語処理学会年次大会）	［龍・董他 2014］（言語処理学会年次大会）
低品質（対訳文対応無）	対訳対抽出	［豊田他 2013］（NL研・言語処理学会年次大会）	

日中パテントファミリーから抽出した対訳文を用いた同義対訳専門用語の同定手法

†龍梓　†董麗娟　†豊田樹生
†宇津呂武仁　‡三橋朋晴　†山本幹雄

†筑波大学大学院 システム情報工学研究科
‡日本特許情報機構

2014/03/20　言語処理学会　第20回年次大会

同義語同定の動機

- 先行研究では，複数の対訳専門用語間の同義・異義関係を同定することができない

中日対訳特許文：

…エアクリーナ1内…
される。
…空气滤清器1内…

…二次空気導入管…
されている。
…补充空气引入管60的上游端被连接到空气清洁器11。

…エアクリーナ30…
止し、…。
…防止通过空气冷…

同義

異義

対訳専門用語：

エアクリーナ
空気濾清器

エアクリーナ
空気清浄器

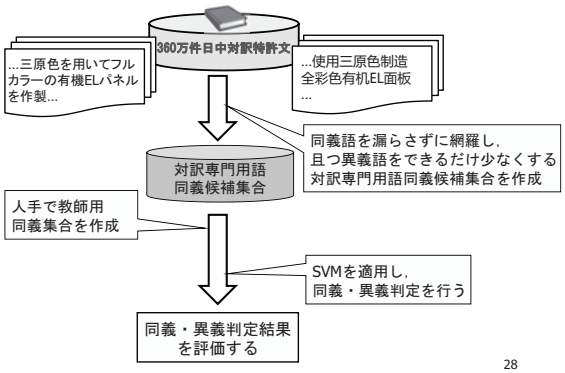
エアクリーナ
空気冷却器

同義語同定の動機

- ❖<u>専門用語対訳辞書作成者のため</u>:
  - ❖ 同義語辞書を整備し，翻訳者の理解を助ける.

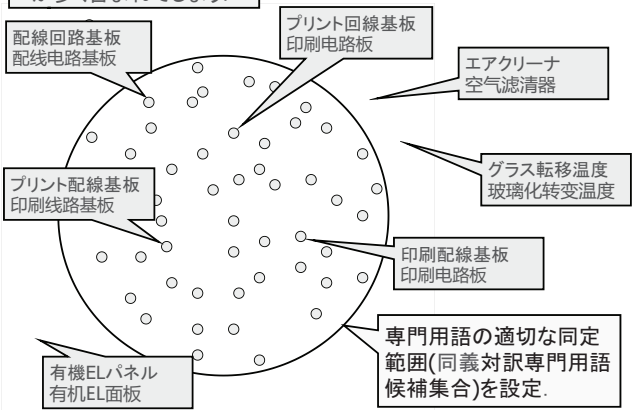
- ❖<u>機械翻訳のため</u>:
  - ❖ 機械翻訳において用いる翻訳辞書において，同義語情報を整備することにより，カバレージが向上する.

提案手法の流れ

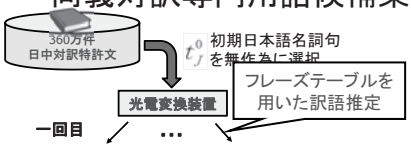


専門用語候補の探索範囲

探索範囲が広すぎて，同義でない対訳専門用語が多く含まれてしまう.



フレーズテーブルを用いた同義対訳専門用語候補集合の作成方法



フレーズテーブルの作成

日中対訳特許文(360万件)

pair 1: ファン/ケーシング/13/ 内/の/羽根車/10/ の/中心/位置/を/偏心/させて
通过/使/风扇/罩/13/内/的/叶轮/10/的/中心/位置/偏心

pair 2: 図 /30 /は/ , /超/広帯域/アンテナ/装置/の/VSWR/の/周波数/特性/を/示す
图/30/表示/超宽带/天线/装置/的/VSWR/的/频率/特性

pair *n*: 高/成形/性/冷/延/鋼板/と/その/製造/方法/...
高屈强/比/的/冷轧/钢板/及/其/制造/方法/...

学習 Moses (統計的機械翻訳モデルのツールキット)

EMアルゴリズムを用いて日中で単語対応付け

フレーズテーブル

日本語フレーズ	中国語フレーズ	翻訳確率
ファン/ケーシング	风扇/罩	0.49
超/広帯域/アンテナ/装置	超宽带/天线/装置	0.93
	...	
冷/延/鋼板	冷轧/钢板	0.87

参考文献:
P. Koehn, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

対訳専門用語の自動収集手法 [董 14]

対訳特許文(1文):

.....在/使用/三原色/制造/全彩色/有机/EL/面板/时.....

.....

三原色/を/用い/て/フル カラー/の/有機/EL/パネル/を/作製/しよ/う/と/する/場合/... ...

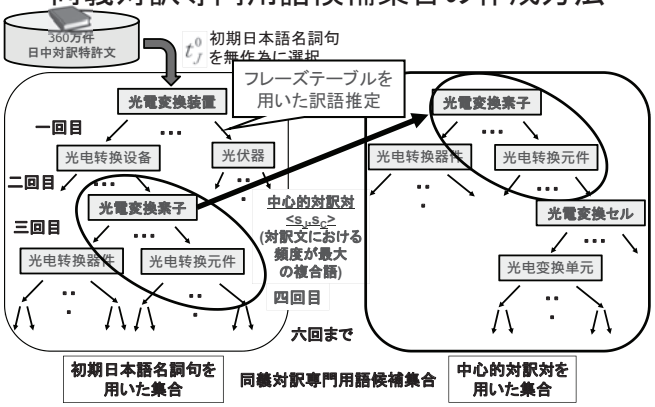
フレーズテーブル

中国語フレーズ	日本語フレーズ	翻訳確率
有机EL面板	有機ELパネル	0.73
~~有机EL面板~~	~~有機 EL ディスプレイ~~	~~0.1~~
有机EL面板	有機ELパネルを	0.05
	...	

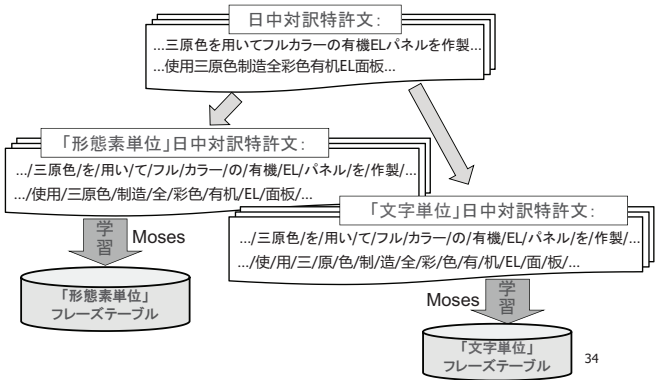
対訳専門用語

有机EL面板
有機ELパネル

フレーズテーブルを用いた同義対訳専門用語候補集合の作成方法



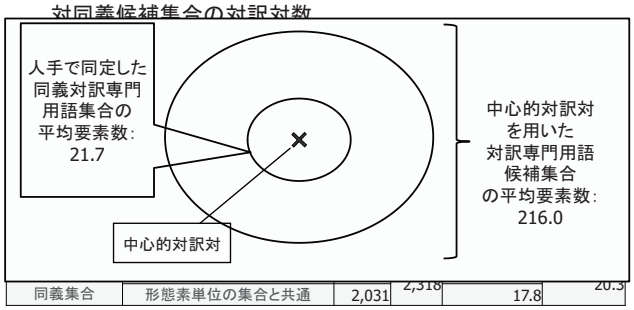
「形態素単位」フレーズテーブルと「文字単位」フレーズテーブル



34

中心的対訳対を用いた同義対訳専門用語候補集合の生成結果

❖ 形態素単位フレーズテーブルを用いた場合の専門用語対訳対同義候補集合の対訳対数

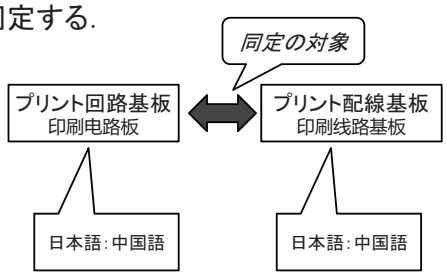


同義集合	形態素単位の集合と共通	2,031	2,318	17.8	20.3

36

同義・異義同定の対象

❖中心的対訳対と対訳専門用語間の関係を同定する.



37

対訳専門用語の同義・異義集合を同定するための素性

分類	素性名
対訳対 (t_i,t_c) の特性を規定する	f_2:共起頻度
	f_3:中国語訳語の順位
	f_4:日本語訳語の順位
	f_5:日本語文字数
	f_6:中国語文字数
	f_7:訳語推定における繰り返しの回数
対訳対 (t_i,t_c) と中心的対訳対 (s_j,s_c) の関の関係を規定する	f_8:日本語用語が同一
	f_9:中国語用語が同一
	f_{10}:編集距離類似度
	f_{11}:バイグラム類似度
	f_{12}:日本語用語の同一形態素の割合
	f_{13}:中国語用語の同一文字の割合
	f_{14}:日本語用語の文字列の包含関係もしくは異表記
	f_{15}:中国語用語の文字列の包含関係
	f_{16}:フレーズテーブルの共通訳の割合
	f_{17}:全非共有箇所に対しフレーズテーブルにおける共通訳の割合
	f_{18}:フレーズテーブルの訳語関係が存在

38

「形態素単位」∩「文字単位」

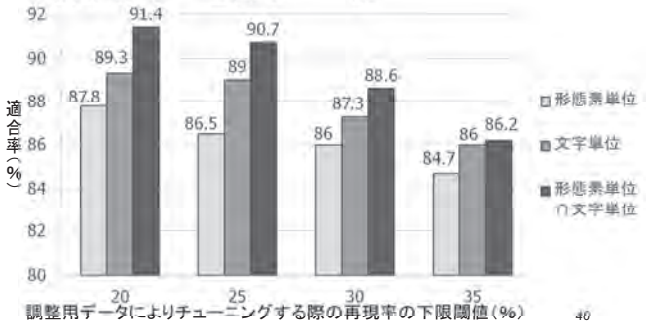
❖「形態素単位」・「文字単位」の両方が一致して同義と判定する場合のみ同義と判定する

対訳対 <t_i,t_c>	中心的対訳対 <s_j,s_c>	SVMの同義・異義判定 「形態素単位」での同義・異義判定	SVMの同義・異義判定 「文字単位」での同義・異義判定	「形態素単位∩文字単位」による同義・異義判定	人手による同義・異義判定
〈印刷/回路/基板, 印刷/电路/基板 or 印/刷/电/路/基/板〉	〈プリント/回路/基板, 印刷/电路/板 or 印/刷/电/路/板〉	同義	同義	➡ 同義	同義
〈電磁/駆動/装置, 电池/驱动/装置 or 电/池/驱/动/装/置〉	〈電磁/駆動/装置, 电磁/驱动/装置 or 电/磁/驱/动/装/置〉	異義	同義	➡ 異義	異義
〈磁気/記録/媒体, 使磁/记录/媒体〉	〈磁気/記録/媒体, 磁性/记录/介质 or 磁/性/记/录/介/质〉	同義	<t_i,t_c>は候補集合に含まない	➡ 異義	異義
… …					

39

「形態素単位」∩「文字単位」の評価結果

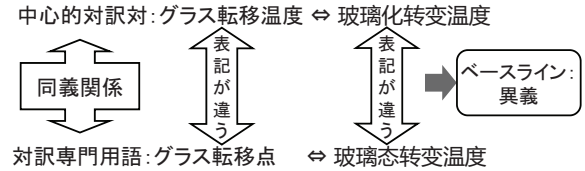
❖評価結果の適合率の比較



40

提案手法による改善例（同義）

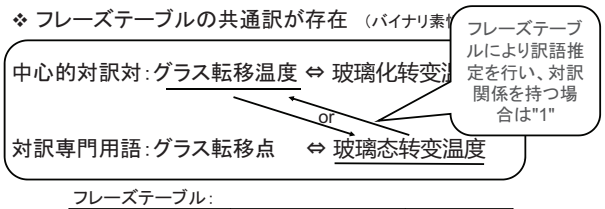
❖提案手法のみで同義と判定し正解



提案手法では、「フレーズテーブルの同一訳の割合」や「フレーズテーブルの共通訳が存在」などの素性の効果によって同義と判定し正解

42

素性例



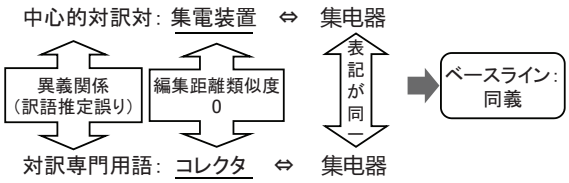
フレーズテーブル：

日本語フレーズ	中国語フレーズ	翻訳確率
グラス転移温度	玻璃化转变温度	0.27
グラス転移温度	玻璃转移温度	0.26
…		
グラス転移温度	玻璃态转变温度	0.08
…		

43

提案手法による改善例（異義）

❖提案手法のみで異義と判定し正解



提案手法では、「日本語編集距離類似度」や「日本語バイグラム類似度」などの素性の効果によって異義と判定し正解

44

パテントファミリーからの専門用語対訳辞書の構築
――― タスクの内訳 ―――

言語資源の品質	タスク	日米	日中
高品質（対訳文対応有）	対訳対抽出	［森下他 2010］（信学会論文誌）	［董・龍他 2014］（言語処理学会年次大会）
	同義対訳対抽出	［梁他 2012］（言語処理学会年次大会）	［龍・董他 2014］（言語処理学会年次大会）
低品質（対訳文対応無）	対訳対抽出	［豊田他 2013］（NL研・言語処理学会年次大会）	

45

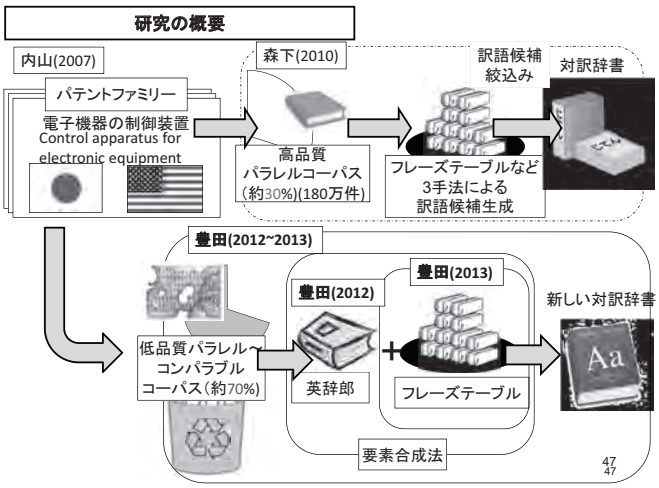
パテントファミリーを用いた
専門用語訳語獲得における
対訳文対非抽出部分
およびフレーズテーブルの利用

†豊田樹生　†龍梓　†董麗娟
‡宇津呂武仁　‡山本幹雄

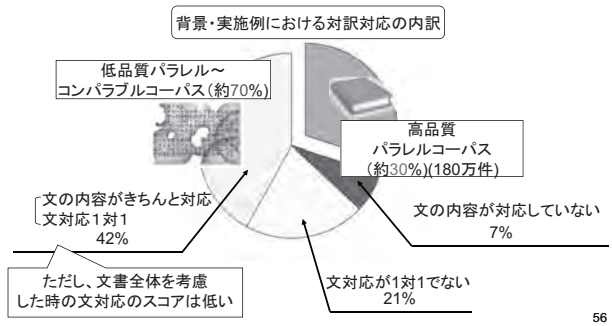
†筑波大学大学院 システム情報工学研究科
‡筑波大学 システム情報系

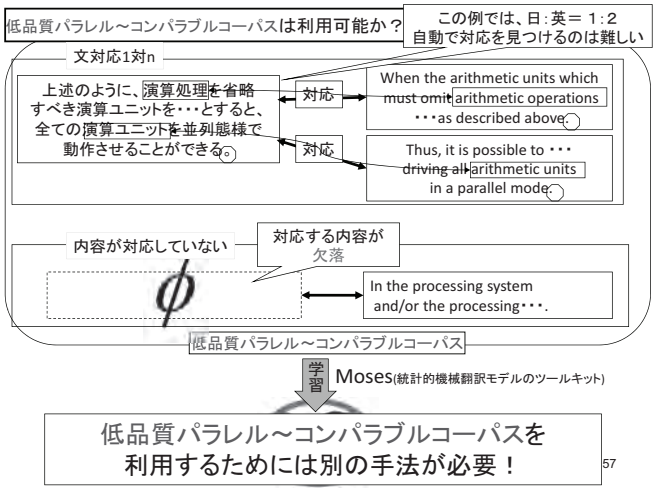
2013/07/19 情報処理学会 第212回自然言語処理研究会

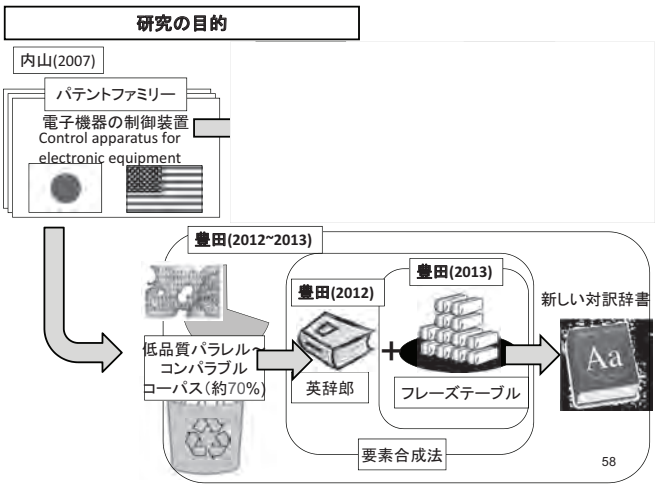
46

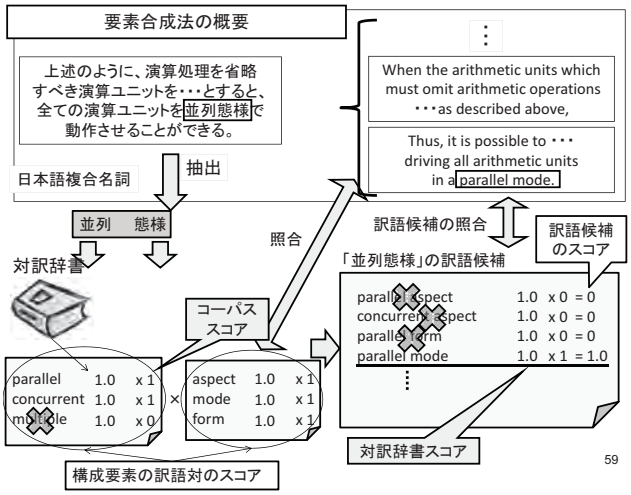


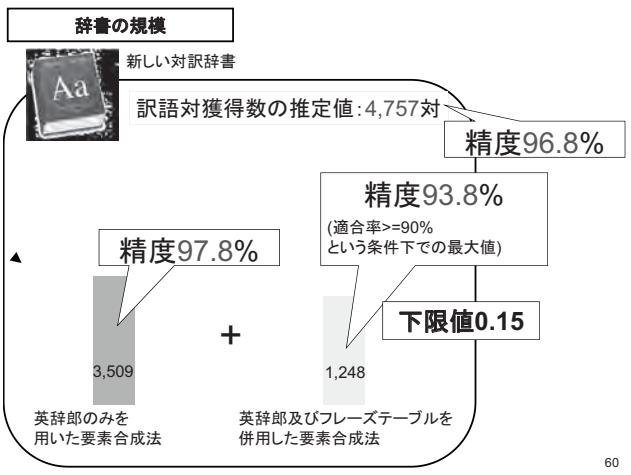
低品質パラレル～コンパラブルコーパスの内訳











まとめと今後の課題

言語資源の品質	タスク	日米	日中
高品質（対訳文対応有）	対訳対抽出	［森下他 2010］（信学会論文誌）	［董・龍他 2014］（言語処理学会年次大会）
	同義対訳対抽出	［梁他 2012］（言語処理学会年次大会）	［龍・董他 2014］（言語処理学会年次大会）
低品質（対訳文対応無）	対訳対抽出	［豊田他 2013］（NL研・言語処理学会年次大会）	

61

まとめと今後の課題

言語資源の品質	タスク	日米	日中
高品質（対訳文対応有）	対訳対抽出	［森下他 2010］（信学会論文誌）	［董・龍他 2014］（言語処理学会年次大会）
	同義対訳対抽出	［梁他 2012］（言語処理学会年次大会）	［龍・董他 2014］（言語処理学会年次大会）
低品質（対訳文対応無）	対訳対抽出	［豊田他 2013］（NL研・言語処理学会年次大会）	

62

まとめ

- 言語資源：日中パテントファミリーから抽出した
  360万件の日中対訳特許文
- 手法：句に基づく統計的機械翻訳モデルによって学習されたフレーズテーブルおよび日中対訳文を知識源とすることによって，
  専門用語の訳語推定・同義対訳専門用語抽出
  を行った

63

今後の課題

- Support Vector Machines (SVM)の導入による高精度化

64

研究会報告 2

「自動評価法を用いた機械翻訳の定量的評価」

研究報告２

自動評価法を用いた機械翻訳の定量的評価

越前谷博（北海学園大学）・磯崎秀樹（岡山県立大学）

目次

自動評価法とは

- なぜ必要なのか
  - 人間による評価は精度は高いが、時間やコストがかかり、再現性の点で問題がある
  - 機械翻訳システムの開発サイクルのスピードアップに有効

自動評価法とは

- 機械翻訳システムの訳文に対し、定量的な評価を完全自動で行うための技術
  - 入力：機械翻訳システムの訳文（システム訳）、人手による正しい訳文（参照訳）
  - 出力：スコア（例：0.0～1.0）
- システム訳に対する評価単位：セグメントレベル（1文）、システムレベル（複数文）
- 自動評価法に対する評価（メタ評価）：自動評価法によるスコアと人手評価によるスコアと間の相関を求める（例：スピアンマンの相関係数）

自動評価法とは

- どんな自動評価法が求められているのか
  - 人間による評価との相関が高い
  - 処理速度が速い
  - 機械翻訳システムへのフィードバックに利用できる（どこが悪いのかを示してくれる）

自動評価法における動向
～Workshop on Statistical Machine Translationに参加して

自動評価法における動向：Workshop on Statistical Machine Translation（WMT）

- 2006年よりACL主催の国際会議のワークショップとして毎年開催されている。
- 機械翻訳に関するいくつかのタスクを選定し、タスクごとに評価ワークショップを実施
- EU言語を対象とした機械翻訳技術の進展を目的とするThe EuroMatrix（Statistical and Hybrid Machine Translation Between All European Languages）Projectの活動の一つとして始まった。

自動評価法における動向：WMT2014

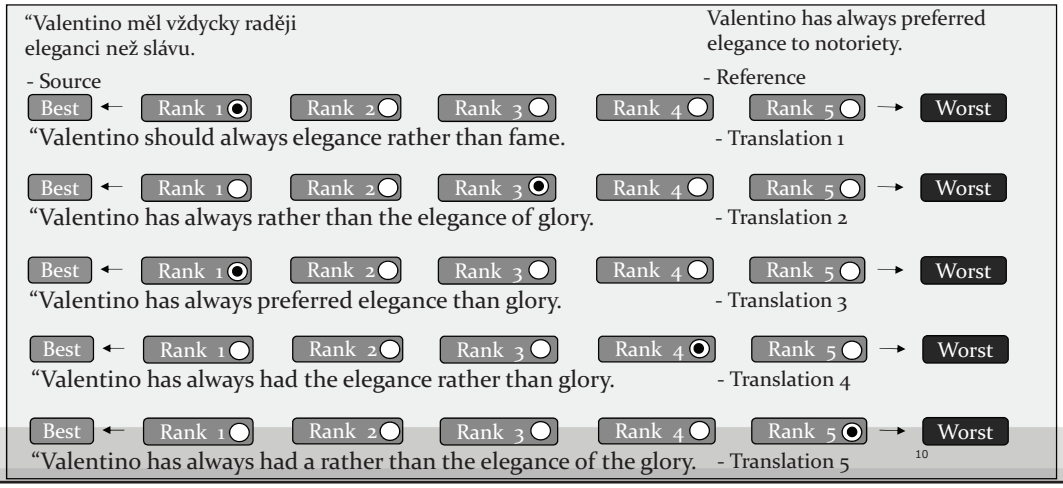
- WMT2014の概要
  - 2014年6月26日～27日、ACL2014のワークショップとしてボルチモアにて開催
  - 対象タスク
    - 翻訳タスク（Translation task）
    - 自動評価タスク（Metrics task）
    - 品質推定タスク（Quality Estimation task）
    - 医療翻訳タスク（Medical translation task）
    - その他：Data and Adaptation、Translation Models

自動評価法における動向：WMT2014

- 自動評価タスクにおけるテストコレクション
  - システム訳
    - 分野：オンラインニュース記事
    - 翻訳タスクに提出された110の機械翻訳システムのシステム訳を使用
    - 言語ペアとテストセット：French-English：3,003文, Hindi-English：2,507文, German-English：3,003文, Czech-English：3,003文, Russian-English：3,003文
    - 機械翻訳システム：cs-en:**5**システム、de-en:**13**システム、en-cs:**10**システム、en-de:**18**システム、en-fr:**13**システム、en-hi:**12**システム、en-ru:**9**システム、fr-en:**8**システム、hi-en:**9**システム、ru-en:**13**システム (en: English, cs: Czech, de: German, fr: French, hi: Hindi, ru: Russian)
    - セグメント数：cs-en:**15,015**文、de-en:**339,039**文、en-cs:**30,030**文、en-de:**49,266**文、en-fr:**39,039**文、en-hi:**30,084**文、en-ru:**27,027**文、fr-en:**24,024**文、hi-en:**22,563**文、ru-en:**39,039**文　　トータル：**315,126**文
- データの提出
  - システム訳と参照訳を用いて、開発した自動評価法よりスコアを求める
  - システムレベル：110スコア、セグメントレベル：315,126スコア

自動評価法における動向：WMT2014

- 自動評価タスクにおけるテストコレクション
  - 人手評価



自動評価法における動向：WMT2014

- 自動評価タスクにおける参加チーム
  - 12のグループより23の自動評価法が参加

Metrics	Sys	Seg	Authors
APAC	●	●	Hokkai-Gakuen University（Echizen'ya, 2014）
BEER		●	University of Amsterdam（Stanojevic and Sima'an,2014）
RED-＊	●	●	Dublin City University（Wu and Yu,2014）
DISCO TK-＊	●	●	Qatar Computing Research Institute（Guzman et al.,2014）
ELEXR	●		University of Tehran（Mahmoudi et al.,2014）
LAYERED	●		Indian Institute of Tech.（Gautam and Bhattacharyya,2014）
METEOR	●	●	Carnegie Mellon University（Denkowski and Lavie,2014）
AMBER	●	●	National Research Council of Canada（Chen and Cherry, 2014）
BLEU-NRC	●	●	National Research Council of Canada（Chen and Cherry,2014）
PARMESAN	●		Charles University in Prague（Barancikova,2014）
TBLEU	●		Charles University in Prague（Libovicky and Pecina,2014）
UPC-＊	●	●	Technical University of Catalunya（Gonzalez et al.,2014）
VERTA-＊	●	●	University of Barcelona（Comelles and Atserias,2014）

11

自動評価法における動向：WMT2014

- システムレベルのメタ評価
  - ピアソンの相関係数

$$r = \frac{\sum_{i=1}^{n}(H_i - \bar{H})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \bar{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \bar{M})^2}}$$

MTシステムS_iに対する人手評価のスコア： H_i　　人手評価のスコアの平均： $\overline{H}$

MTシステムS_iに対する自動評価法のスコア： M_i　　自動評価法のスコアの平均： $\overline{M}$

  - 人手評価
    - TrueSkillを使用・・・ベイズ理論に基づくランキングアルゴリズム

自動評価法における動向：WMT2014

- システムレベルのメタ評価
  - 訳文：into English

From	fr	de	hi	cs	ru	Avg
DISCOTK-PARTY-TUNED	.98	**.94**	.96	.97	**.87**	**.94**
LAYERED	.97	.89	**.98**	.94	.85	.93
DISCOTK-PARTY	.97	.92	.86	.98	.86	.92
UPC-STOUT	.97	.91	.90	.95	.84	.91
VERTA-W	.96	.87	.92	.93	.85	.91
VERTA-EQ	.96	.85	.93	.94	.84	.90
TBLEU	.95	.83	.95	.96	.80	.90
BLEU-NRC	.95	.82	.96	.95	.79	.89
BLEU	.95	.83	.96	.91	.79	.89
UPC-IPA	.97	.89	.91	.82	.81	.88
CDER	.95	.82	.83	.97	.80	.87
APAC	.96	.82	.79	.98	.82	.87
REDSYS	**.98**	.90	.68	.99	.81	.87
REDSYSSENT	.98	.91	.64	**.99**	.81	.87
NIST	.96	.81	.78	.98	.80	.87
DISCOTK-LIGHT	.96	.93	.56	.95	.79	.84
METEOR	.98	.93	.46	.98	.81	.83
WER	.95	.76	.61	.97	.81	.82
AMBER	.95	.91	.51	.74	.80	.78
ELEXR	.97	.86	.54	.94	-.40	.58

自動評価法における動向：WMT2014

- システムレベルのメタ評価
  - 訳文：out of English

Into	fr	hi	cs	ru	Avg	de
NIST	.94	.98	.98	.93	**.96**	.20
CDER	.95	.95	.98	.94	.95	.28
AMBER	.93	**.99**	.97	.93	.95	.24
METEOR	.94	.98	.98	.92	.95	.26
BELU	.94	.97	.98	.91	.95	.22
PER	.94	.93	**.99**	**.94**	.95	.19
APAC	.95	.94	.97	.93	.95	.35
TBLEU	.93	.97	.97	.91	.95	.24
BLEU-NRC	.93	.97	.97	.90	.95	.20
ELEXR	.89	.96	.98	.94	.94	.26
TER	.95	.83	.98	.93	.92	.32
WER	**.96**	.52	.98	.93	.85	**.36**
PARMESAN	-	-	.96	-	.96	-
UPC-IPA	.94	-	.97	.92	.94	.28
REDSYSSENT	.94	-	-	-	.94	.21
REDSYS	.94	-	-	-	.94	.21
UPC-STOUT	.94	-	.94	.92	.93	.30

自動評価法における動向：WMT2014

- セグメントレベルのメタ評価
  - ケンドールの順位相関係数

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}$$

自動評価法のスコアと人手評価のスコアが一致　：*Concordant*

自動評価法のスコアと人手評価のスコアが不一致：*Discordant*

$\tau \in [-1, 1]$

  - 人手評価

Best ← Rank 1 ◉ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

"Valentino should always elegance rather than fame. - Translation 1

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ◉ Rank 4 ○ Rank 5 ○ → Worst

"Valentino has always rather than the elegance of glory. - Translation 2

自動評価法における動向：WMT2014

- セグメントレベルのメタ評価
  - ケンドールの順位相関係数
  - 例：

Human	Metric	結果
A<B	A<B	一致:1
C>A	C>A	一致:1
C>B	C<B	不一致:-1

$$\tau = \frac{2 \cdot 1 + 1 \cdot (-1)}{2 + 1} = \frac{1}{3}$$

  - WMT2014 variant
    - 自動評価法の結果のみが"="の場合は0とする
    - その場合、分母のみが増加

		Metric		
		<	=	>
Human	<	1	0	-1
	=	X	X	X
	>	-1	0	1

自動評価法における動向：WMT2014

- セグメントレベルのメタ評価
  - 訳文：into English
  - ペア数：fr-en：26,090
    de-en：25,260
    hi-en：20,900
    cs-en：21,130
    ru-en：34,460

From	fr	de	hi	cs	ru	Avg
DISCOTK-PARTY-TUNED	**.43**	**.38**	.43	**.33**	**.35**	**.39**
BEER	.42	.34	**.44**	.28	.33	.36
REDCOMBSENT	.41	.34	.42	.28	.34	.36
REDCOMBSYSSENT	.41	.34	.42	.28	.34	.36
METEOR	.41	.33	.42	.28	.33	.35
REDSYSSENT	.40	.34	.39	.28	.32	.35
REDSENT	.40	.34	.38	.28	.32	.35
UPC-IPA	.41	.34	.37	.27	.32	.34
UPC-STOUT	.40	.34	.35	.28	.32	.34
VERTA-W	.40	.32	.39	.26	.31	.34
VERTA-EQ	.41	.31	.38	.26	.31	.34
DISCOTK-PARTY	.39	.33	.36	.26	.31	.33
AMBER	.37	.31	.36	.25	.29	.32
BLEU-NRC	.38	.27	.32	.23	.27	.29
SENTBLEU	.38	.27	.30	.21	.26	.29
APAC	.36	.27	.29	.20	.28	.28
DISCOTK-LIGHT	.31	.22	.24	.19	.21	.23
DISCOTK-LIGHT-KOOL	.00	.00	.00	.00	.00	.00

自動評価法における動向：WMT2014

- システムレベルのメタ評価
  - 訳文：out of English
  - ペア数：en-fr：33,350
    en-de：54,660
    en-hi：28,120
    en-cs：55,900
    en-ru：28,960

Into	fr	de	hi	cs	ru	Avg
BEER	.29	**.27**	.25	**.34**	**.44**	**.32**
METEOR	.28	.24	.26	.32	.43	.31
AMBER	.26	.23	**.29**	.30	.40	.30
BLEU-NRC	.26	.20	.23	.30	.39	.28
APAC	.25	.21	.20	.29	.39	.27
SENTBLEU	.26	.19	.23	.29	.38	.27
UPC-STOUT	.28	.23	-	.28	.42	.30
UPC-IPA	.26	.23	-	.30	.43	.30
REDSENT	**.29**	.24	-	-	-	.27
REDCOMBSYSSENT	.29	.24	-	-	-	.27
REDCOMBSENT	.29	.24	-	-	-	.27
REDSYSSENT	.29	.24	-	-	-	.26

自動評価法における動向：WMT2014

- システムレベルの総評
  - 相関係数が0.8～1.0の範囲であり、全体的に高い相関である
  - out of Englishにおいてベースライン（NIST, CDER, BLEU, PER）が高順位である
    - English-Hindiを除くとWERも高順位である
  - into Germanの相関係数が非常に低い
    - 機械翻訳システムの数（18）が他の言語間より多かった。
    - 自動評価法において、似たような性能のシステムを差別化することは難しい。
  - METEORではnon-Latin scriptから英語の順位が低い
- セグメントレベルの総評
  - 相関係数は約0.4であり、まだまだ不十分

⇩

自動評価タスクは変わらず興味深いタスクである
（12チームが参加）

自動評価法における動向：WMT2014

- WMT2014に参加しての感想
  - 提案手法（APAC）の位置づけの把握に有効
    - 参加前：システムレベルではそれほど有効ではないが、セグメントレベルでは有効
    - 結果：システムレベルはまあまあの順位だが、セグメントレベルの順位は低い
  - 似たような性能のシステムであっても正しく評価できなければならない

Into	fr	de	hi	cs	ru	Avg
APAC	.95	.35	.94	.97	.93	**.83**
CDER	.95	.28	.95	.98	.94	.82
METEOR	.94	.26	.98	.98	.92	.82
AMBER	.93	.24	**.99**	.97	.93	.81
NIST	.94	.20	.98	.98	.93	.81
ELEXR	.89	.26	.96	.98	.94	.81
BELU	.94	.22	.97	.98	.91	.80
TBLEU	.93	.24	.97	.97	.91	.80
TER	.95	.32	.83	.98	.93	.80
PER	.94	.19	.93	**.99**	**.94**	.80
BLEU-NRC	.93	.20	.97	.97	.90	.80
WER	**.96**	**.36**	.52	.98	.93	.75
PARMESAN	-		-	.96	-	.96
UPC-IPA	.94	.28	-	.97	.92	.78
UPC-STOUT	.94	.30	-	.94	.92	.78
REDSYSSENT	.94	.21	-	-	-	.58
REDSYS	.94	.21	-	-	-	.58

参考文献：

[1] M. Macháček and O. Bojar: Results of the WMT14 Metrics Shared Task, Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.293-301 (2014).

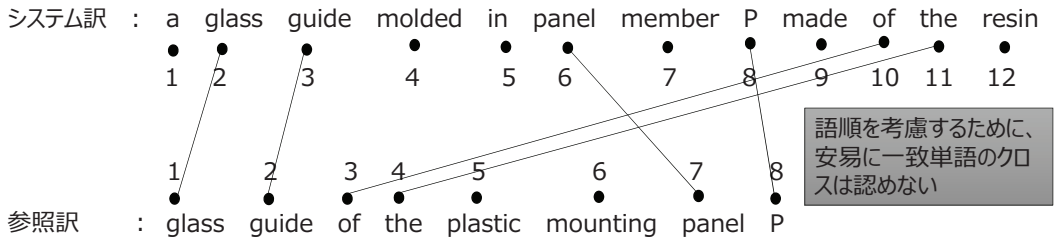
[2] O. Bojar, C. Buck, C. Federman, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia and A. Tamchyna: Findings of the 2014 Workshop on Statistical Machine Translation, Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.12-58 (2014).

自動評価法：APAC

自動評価法：APAC

- 特徴
  - 多義性のある一致単語列（チャンク）を大局的な観点から一意に決定：正しいチャンクを決定
  - 一致単語の語順の違いに柔軟に対応：パラメータの使用

- チャンクの決定方法



自動評価法：APAC

$$score = \sum_{c \in c\_num} \left(length(c)^{\beta} \times pos \right)$$

$$pos = \left(1.0 - \left| \frac{c_i}{m} - \frac{c_j}{n} \right| \right)$$

- チャンクの決定方法

候補1：

システム訳　：a glass guide molded in panel member P made of the resin

参照訳　：glass guide of the plastic mounting panel P

score = 3.499

候補2：

システム訳　：a glass guide molded in panel member P made of the resin

参照訳　：glass guide of the plastic mounting panel P

score = 3.446

パラメータβ：デフォルト値は1.2

自動評価法：APAC

- スコアの算出方法[1]

システム訳　：a glass guide molded in panel member P made of the resin

参照訳　：glass guide of the plastic mounting panel P

⇩ チャンクを再帰的に決定

システム訳　：a glass guide molded in panel member P made of the resin

参照訳　：glass guide of the plastic mounting panel P

$$R = \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \sum_{c \in c\_num} length(c)^{\beta} \right)}{m^{\beta}} \right)^{\frac{1}{\beta}}$$

$$P = \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \sum_{c \in c\_num} length(c)^{\beta} \right)}{n^{\beta}} \right)^{\frac{1}{\beta}}$$

$$AE\ score = \frac{(1+\gamma^2)RP}{R+\gamma^2 P}$$

パラメータα：デフォルト値は0.1　パラメータβ：デフォルト値は1.2　AE score = 0.3268

[1] H. Echizen-ya and K. Araki:Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, Proceedings of the Eleventh Machine Translation Summit (MT SUMMIT XI), pp.151-158 (2007).

自動評価法：APAC

- 改良[2]
  - 問題点：短い文のスコアが過度に小さくなる

⇩

短い文ほど不一致単語の重みが大きくなる

システム訳　：the doctor treated a patient
参照訳　：the doctor cured a patient

$$P = \left\{ \left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \times Ch\_score)}{m^\beta} \right)^{\frac{1}{\beta}} + 0.5 \times Prize\_m \right\} / 2.0$$

$$Prize\_m = \frac{1}{\log(m) + 1}$$

$$R = \left\{ \left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \times Ch\_score)}{n^\beta} \right)^{\frac{1}{\beta}} + 0.5 \times Prize\_n \right\} / 2.0$$

$$Prize\_n = \frac{1}{\log(n) + 1}$$

$$Ch\_score = \sum_{c \in c\_num} length(ch)^\beta$$

[2] H. Echizen-ya, K. Araki and E. Hovy: Application of Prize based on Sentence Length in Chunk-based Automatic Evaluation of Machine Translation, Results of the WMT14 Metrics Shared Task, Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.381-386 (2014).

自動評価法：APAC

- 性能評価(JE)
  - WMT2012におけるシステムレベルの相関係数（Spearman's rank）

Metrics	cs-en(6)	de-en(16)	es-en(12)	fr-en(15)	Avg.	Rank
APAC	0.886	0.650	0.958	0.811	0.826	6
IMPACT	0.886	0.676	0.958	0.807	0.832	4
RIBES	0.943	0.732	0.944	0.814	0.858	2
METEOR	0.943	0.841	0.979	0.818	0.895	1
BLEU	0.886	0.674	0.958	0.796	0.828	5
NIST	0.943	0.700	0.944	0.779	0.841	3

  - WMT2012におけるセグメントレベルの相関係数（Kendall tau rank）

Metrics	cs-en (11,155)	de-en (12,042)	es-en (9,880)	fr-en (11,682)	Avg.	Rank
APAC	0.185	0.204	0.209	0.226	0.206	3
IMPACT	0.189	0.207	0.208	0.226	0.207	2
RIBES	0.055	0.125	0.114	0.115	0.102	4
METEOR	0.223	0.279	0.248	0.243	0.248	1

自動評価法：APAC

- 性能評価(JE)
  - WMT2013におけるシステムレベルの相関係数（Spearman's rank）

Metrics	cs-en(11)	de-en(17)	es-en(12)	fr-en(13)	ru-en(19)	Avg.	Rank
APAC	0.900	0.904	0.916	0.934	0.709	0.873	4
IMPACT	0.909	0.909	0.937	0.934	0.721	0.882	2
RIBES	0.900	0.912	0.930	0.978	0.670	0.878	3
METEOR	0.982	0.946	0.923	0.967	0.889	0.941	1
BLEU	0.945	0.897	0.853	0.951	0.614	0.852	5
NIST	0.900	0.828	0.804	0.786	0.465	0.757	6

  - WMT2013におけるセグメントレベルの相関係数（Kendall tau rank）

Metrics	cs-en (85,469)	de-en (128,668)	es-en (67,832)	fr-en (80,741)	ru-en (151,422)	Avg.	Rank
APAC	0.144	0.163	0.169	0.139	0.121	0.147	3
IMPACT	0.148	0.167	0.176	0.142	0.123	0.151	2
RIBES	0.044	0.063	0.056	0.018	0.003	0.037	4
METEOR	0.222	0.236	0.241	0.194	0.226	0.224	1

自動評価法：APAC

- 性能評価(JE)
  - NTCIR-7におけるシステムレベルの相関係数（Spearman's rank）

Metrics	Adequacy(15)	Fluency(15)	Avg.	Rank
APAC	0.872	0.805	0.839	2
IMPACT	0.872	0.805	0.839	2
RIBES	0.963	0.918	0.941	1
METEOR	0.424	0.380	0.402	6
BLEU	0.582	0.586	0.584	4
NIST	0.578	0.568	0.573	5

  - NTCIR-7におけるセグメントレベルの相関係数（Kendall tau rank）

Metrics	Adequacy(1,500)	Fluency(1,500)	Avg.	Rank
APAC	0.494	0.489	0.491	1
IMPACT	0.482	0.476	0.479	2
RIBES	0.370	0.341	0.356	3
METEOR	0.366	0.383	0.375	4

自動評価法：APAC

- 性能評価(JE)
  - NTCIR-9におけるシステムレベルの相関係数（Spearman's rank）

Metrics	Adequacy(19)	Acceptance(14)	Avg.	Rank
APAC	0.182	0.298	0.240	2
IMPACT	0.182	0.298	0.240	2
RIBES	0.660	0.640	0.650	1
METEOR	-0.081	0.015	-0.033	5
BLEU	-0.123	0.059	-0.032	4
NIST	-0.344	-0.275	-0.309	6

  - NTCIR-9におけるセグメントレベルの相関係数（Kendall tau rank）

Metrics	Adequacy(5,700)	Acceptance(5,700)	Avg.	Rank
APAC	0.250	0.261	0.256	2
IMPACT	0.242	0.250	0.246	3
RIBES	0.281	0.339	0.310	1
METEOR	0.167	0.217	0.192	4

自動評価法：APAC

- APACの特徴
  - Chef's tips for evaluation

	データ	優劣
WMT	システムレベル	METEOR > RIBES > APAC
	セグメントレベル	METEOR > APAC > RIBES
NTCIR	システムレベル	RIBES > APAC > METEOR
	セグメントレベル	APAC > RIBES > METEOR (NTCIR-7)
	セグメントレベル	RIBES > APAC > MEIEOR (NETCIR-9)

  - 相対的には安定した性能を示している。

自動評価法：RIBES

自動評価法：RIBES

- システム訳と参照訳の間の**語順の近さ**を測定
- 日英・英日の翻訳において**人手評価と強い相関**がある

NTCIR-7 日英翻訳でのメタ評価

妥当性とのシステムレベルの相関、単一参照訳、スピアマンの相関係数

BLEU	METEOR	ROUGE-L	IMPACT	RIBES
0.515	0.490	0.903	0.826	**0.947**

自動評価法：RIBES

- EMNLP版[1]のRIBESは以下の式で定義される

$$\mathrm{RIBES} \overset{\mathrm{def}}{=} \mathrm{NKT} \times P^{\alpha}$$

- NKT $\overset{\mathrm{def}}{=} \frac{\tau + 1}{2}$ は正規化したKendall's τ
  - システム訳と参照訳で共通する単語の**語順の近さ**を表す。
- *P*は単語の適合率
  - α（$0 \leq \alpha \leq 1$）は*P*の影響を制御するパラメータ
  - デフォルト値は0.2
- （低評価）$0.0 \leq$ RIBES ≤ 1.0（高評価）

[1] H. Isozaki, T. Hirao, K. Duh, K. Sudoh and H. Tsukada: Automatic Evaluation of Translation Quality for Distant Language Pairs, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP2010), pp.944-952 (2010).

自動評価法：RIBES

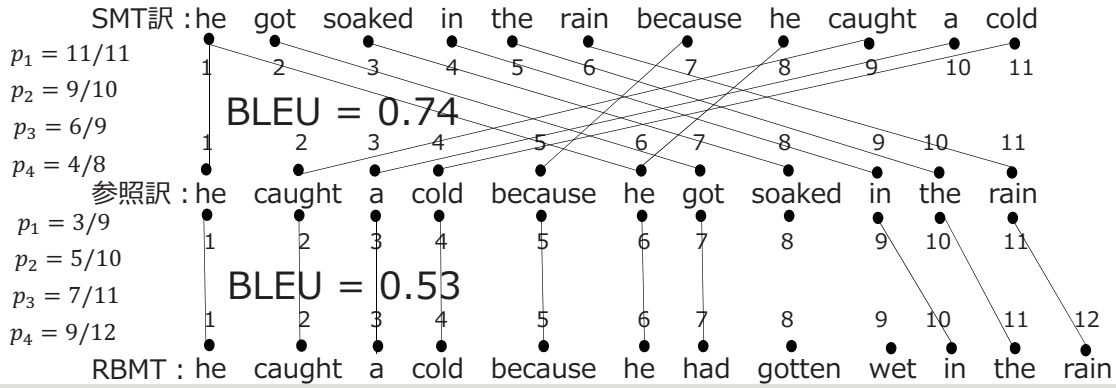
- BLEUの問題点
  - SMTの語順が大きく誤っていても高いスコアとなる。
  - 因果関係が逆の例

参照訳：
He caught a cold because he got soaked in the rain.

SMT訳：
He got soaked in the rain because he caught a cold.

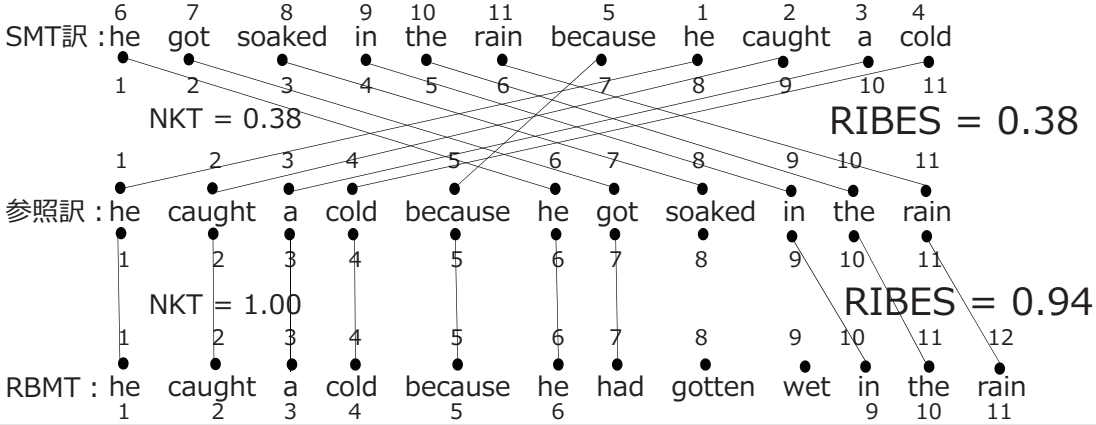
自動評価法：RIBES

- BLEUの問題点
  - SMTの語順が大きく誤っていても高いスコアとなる。
  - 因果関係が逆の例



自動評価法：RIBES

- RIBESの評価
  - SMT訳よりもRBMTを高く評価



自動評価法：RIBES

- RIBESの改良
  - EMNLP版のRIBESに対して、BLEUのBrevity Penaltyを導入

参照訳： John went to a restaurant yesterday

システム訳：to a

語順（NKT）もユニグラム適合率（P）も完全一致なので、従来だと1.0となってしまう。

  - 以下の式で定義[2]

$$\mathrm{RIBES} \overset{\mathrm{def}}{=} \mathrm{NKT} \times P^{\alpha} \times \mathbf{BP}^{\beta}$$

    - デフォルト値はα=0.25、β=0.10

http://www.kecl.ntt.co.jp/icl/lirg/ribes

[2] 平尾、磯崎、須藤、Duh、塚田、永田： 語順の相関に基づく機械翻訳の自動評価法、自然言語処理、Vol. 21、No. 3, pp.421-444 (2014).

自動評価法：RIBES

- 性能評価
  - NTCIR-9, 10 Patent MTがRIBESを標準的な自動評価法として採用

NTCIR-9, 10 Patent MTでのメタ評価

妥当性とのシステムレベルの相関、単一参照訳、スピアマンの相関係数

		BLEU	NIST	RIBES
NTCIR-9	JE	**-0.042**	-0.114	**0.632**
NTCIR-9	EJ	**-0.029**	-0.074	**0.716**
NTCIR-10	JE	**0.31**	0.36	**0.88**
NTCIR-10	EJ	**0.36**	0.22	**0.79**

  - 現在、日英・英日翻訳のほとんどの論文がRIBESを使用
  - 言語処理学会第20回年次大会（NLP2014）にて18本の機械翻訳の論文がRIBESを使用

自動評価法：RIBES

- RIBESのさらなる改良

  日本語は語順が比較的自由（スクランブリング）。

  太郎はイタリアでピザを食べた。

  イタリアで太郎はピザを食べた。

  日本語訳の評価をする場合に、この点を考慮すべき。

  与えられた参照文の係り受け木から、他の語順を自動生成して参照訳に追加

- RIBESの文レベルの相関係数が若干改善された。

  NTCIR-7 Mosesベースラインで Spearman's ρ が 0.607から 0.670 に向上など。

H. Isozaki, N. Kouchi, T. Hirao:
Dependency-based Automatic Enumeration of Semantically Equivalent Word Orders for Evaluating Japanese Translations, WMT-2014.

まとめ

- 現時点での最適な自動評価法は何か
  - 求めるものによって変わる
  - 一般的な翻訳データ(WMT)、特許翻訳データ(NTCIR)、対象言語、システムレベル、セグメントレベル
- 今後の課題
  - セグメントレベルの評価精度（相関係数）の向上

特別講演

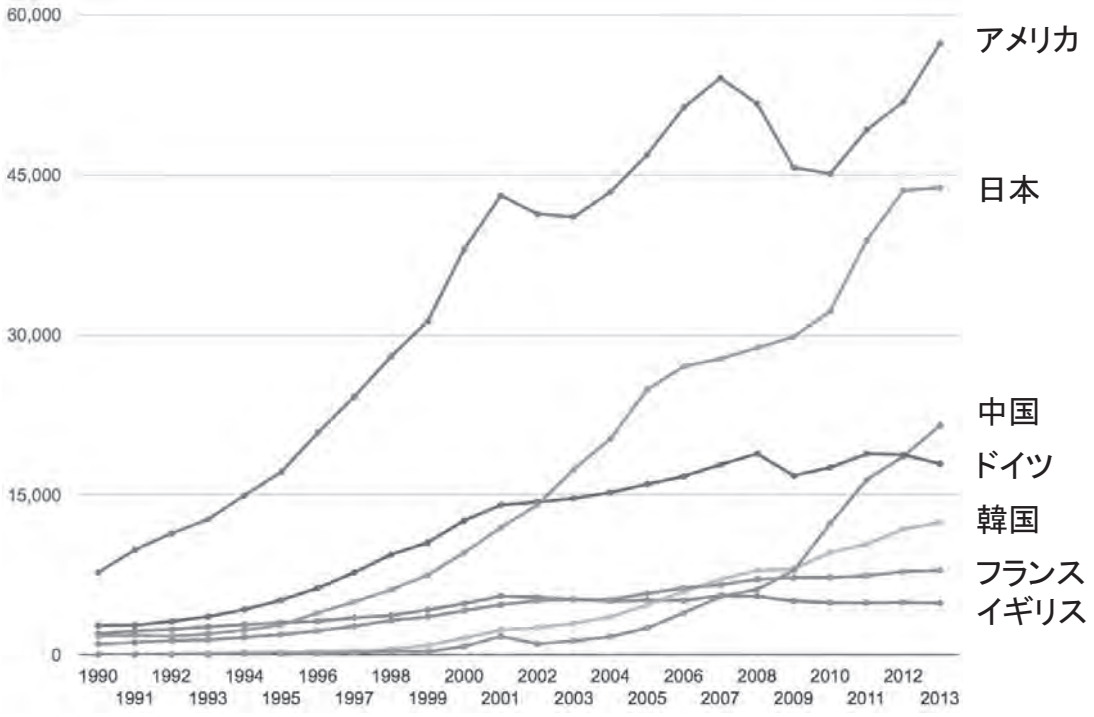
「アジア言語を中心とした機械翻訳研究」

アジア言語を中心とした機械翻訳研究

中澤 敏明
科学技術振興機構（JST）/京都大学

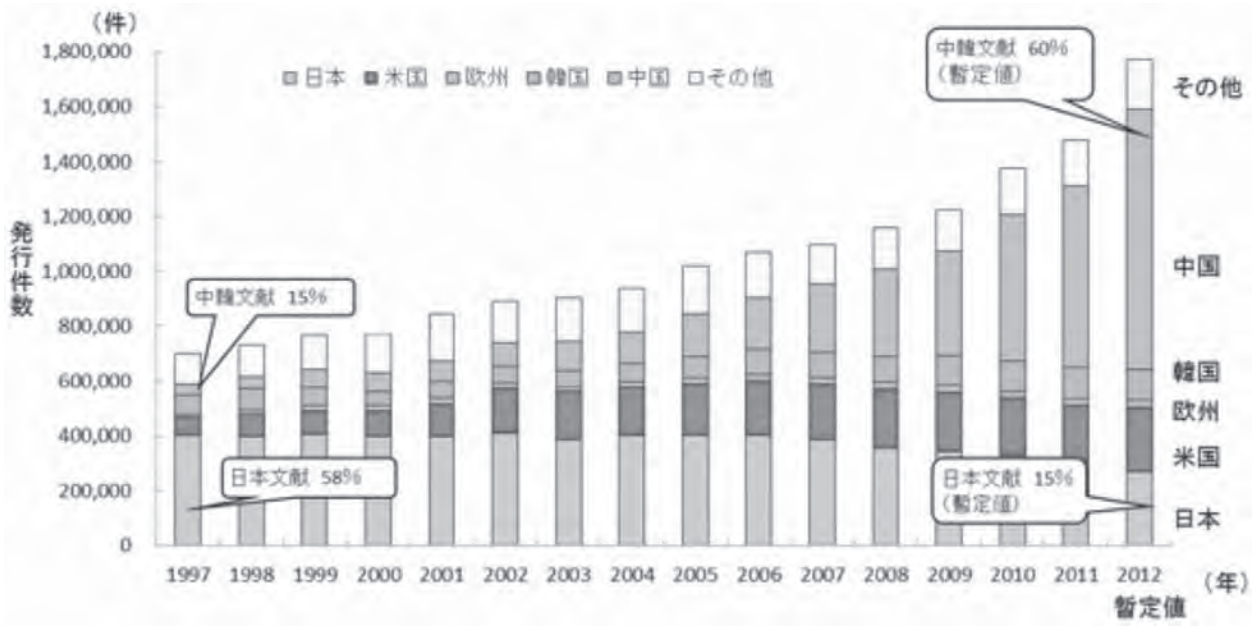
2014年11月28日 第3回特許情報シンポジウム

国際特許出願件数



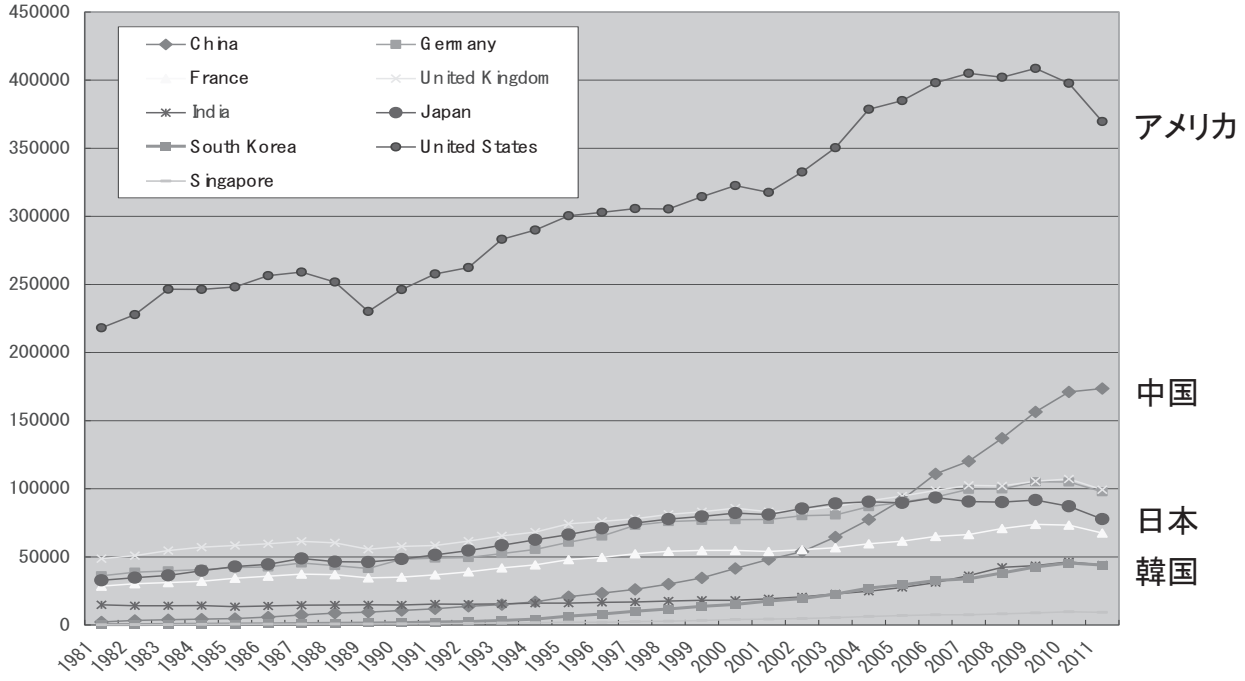
http://www.globalnote.jp/post-5380.html

世界の特許文献



http://www.meti.go.jp/press/2014/11/20141112003/20141112003.html

世界の科学技術論文数



※トムソンロイターのWeb of Scienceのデータを元にJSTが集計

Frontrunner 5000

http://f5000.istic.ac.cn

- 中国科学技术信息研究所(ISTIC)が発表
- およそ4600ある中国の科学技術論文誌から、優れた315論文誌を選出
- 中でも各分野で最も優れた論文(およそ)5000論文を集めた
- 論文概要は英語1000語以内で書かれているが、本文は中国語
  - 国外からのアクセスが期待できない

5

情報アクセスの促進

- 英語以外の言語で書かれた文書量の増大
- その中にも重要な情報は含まれている
- 他言語の重要な情報への容易なアクセスには機械翻訳技術は必要不可欠
  - JPOによる中・韓特許文献翻訳・検索システム
  - JSTによる日中・中日機械翻訳実用化プロジェクト

6

目次

- 日中・中日機械翻訳実用化プロジェクト
  - 言語資源の構築
  - 機械翻訳エンジンの開発
- 1st Workshop on Asian Translation (WAT2014)
  - 概要説明
  - 評価手法
  - 評価結果
- まとめ

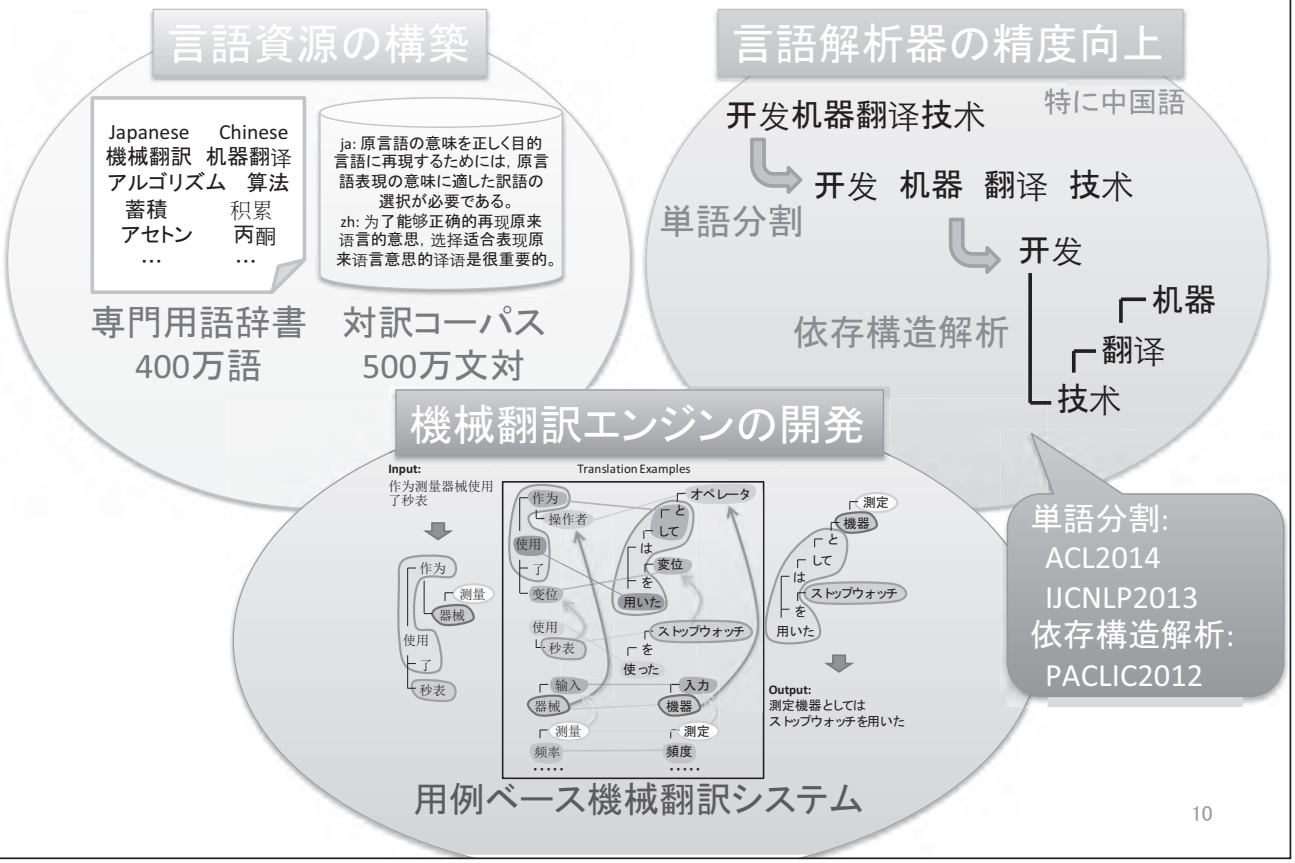
日中・中日機械翻訳実用化プロジェクト

プロジェクト概要

- 期間: 2013年から5年間
- 参加機関
  - 日本: JST, 京大（協力機関: 筑波大, NICT）
  - 中国: ISTIC, CAS, BJTU, HIT
- 機械翻訳技術により日中間の言語障壁を取り除き、科学技術交流の促進を目指す

http://foresight.jst.go.jp/jazh_zhja_mt/

プロジェクトの目標



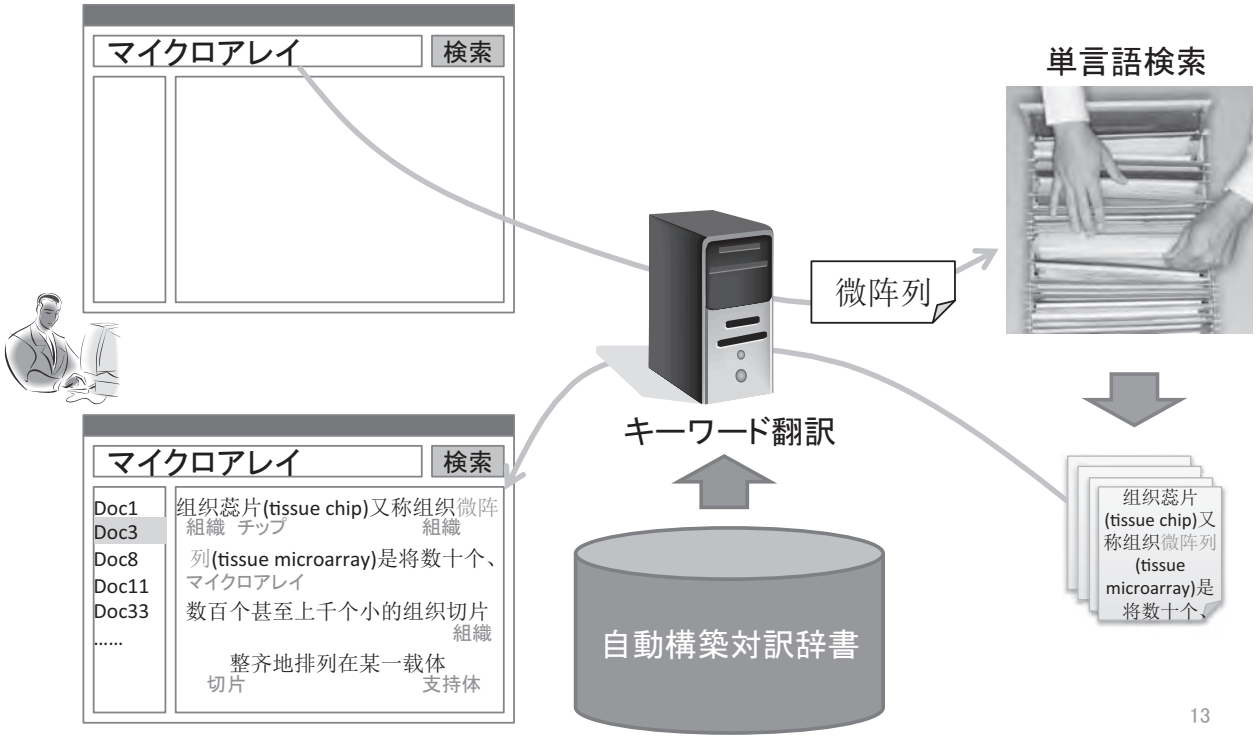
10

日中言語資源の構築

専門用語対訳辞書の構築

- 中英、英日の既存の言語資源から、英語を介して構築
  - 中英 : 論文抄録6M、論文タイトル1M
  - 英日 : 論文抄録23.4M、論文タイトル22.6M
- 現時点での辞書サイズ
  - 中英 : 自動獲得 823,356ペア
    中国側から提供された辞書 3Mペア
  - 英日 : 自動獲得 8,079,137ペア
  - 中英日 : 自動獲得 1,843,959ペア

言語横断文書検索



13

言語横断文書検索

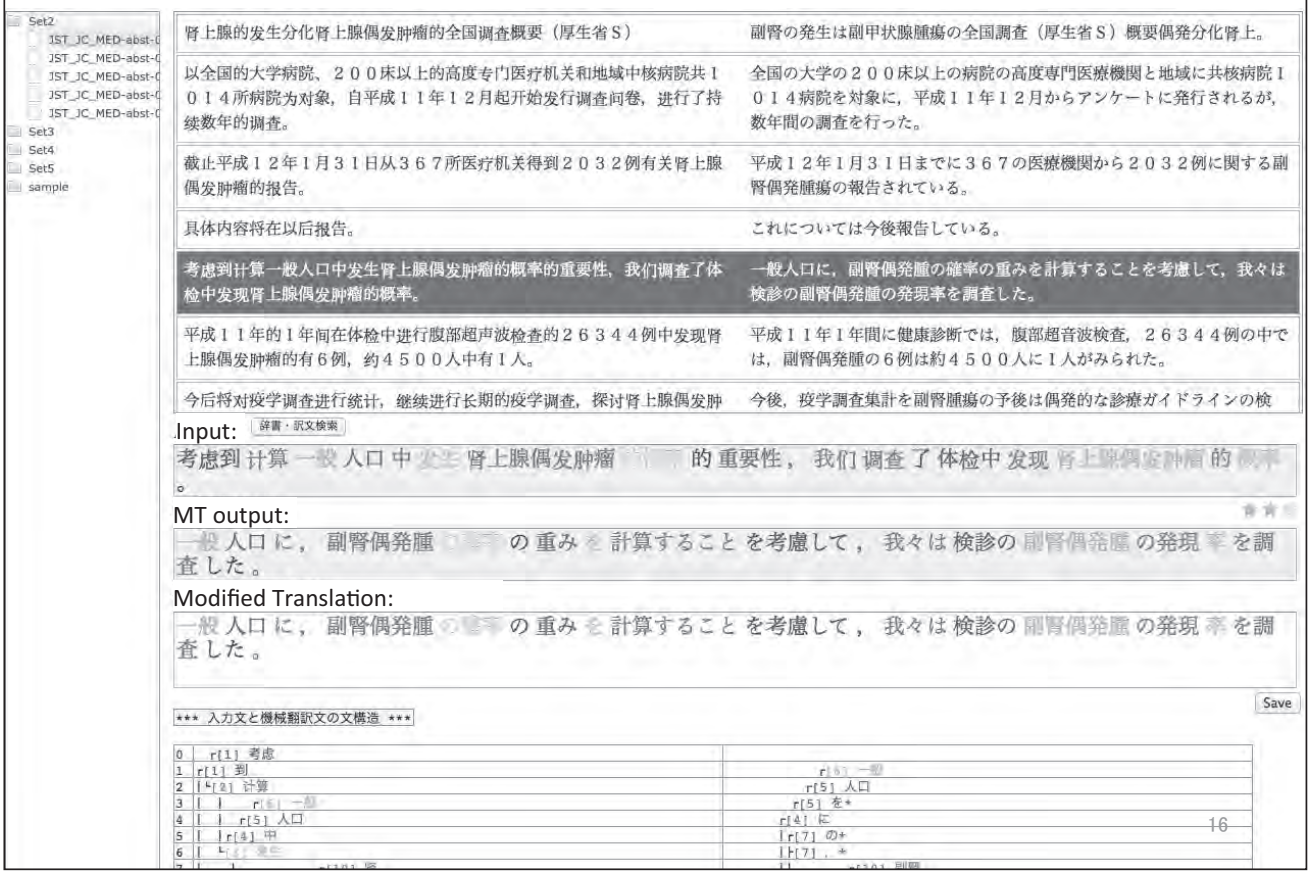
14

日中対訳コーパス構築

- 日本側
  - 既存の対訳抄録からの対訳文の自動獲得
    - 現状2.4M文ペア
  - 中国語文献の人手による日本語への翻訳
    - 現状7,000記事、さらに7,000記事を今年度中に翻訳
- 中国側
  - 翻訳支援ツールを使用した（Computer-assisted Translation: CAT）日英コーパスの中国語への翻訳

15

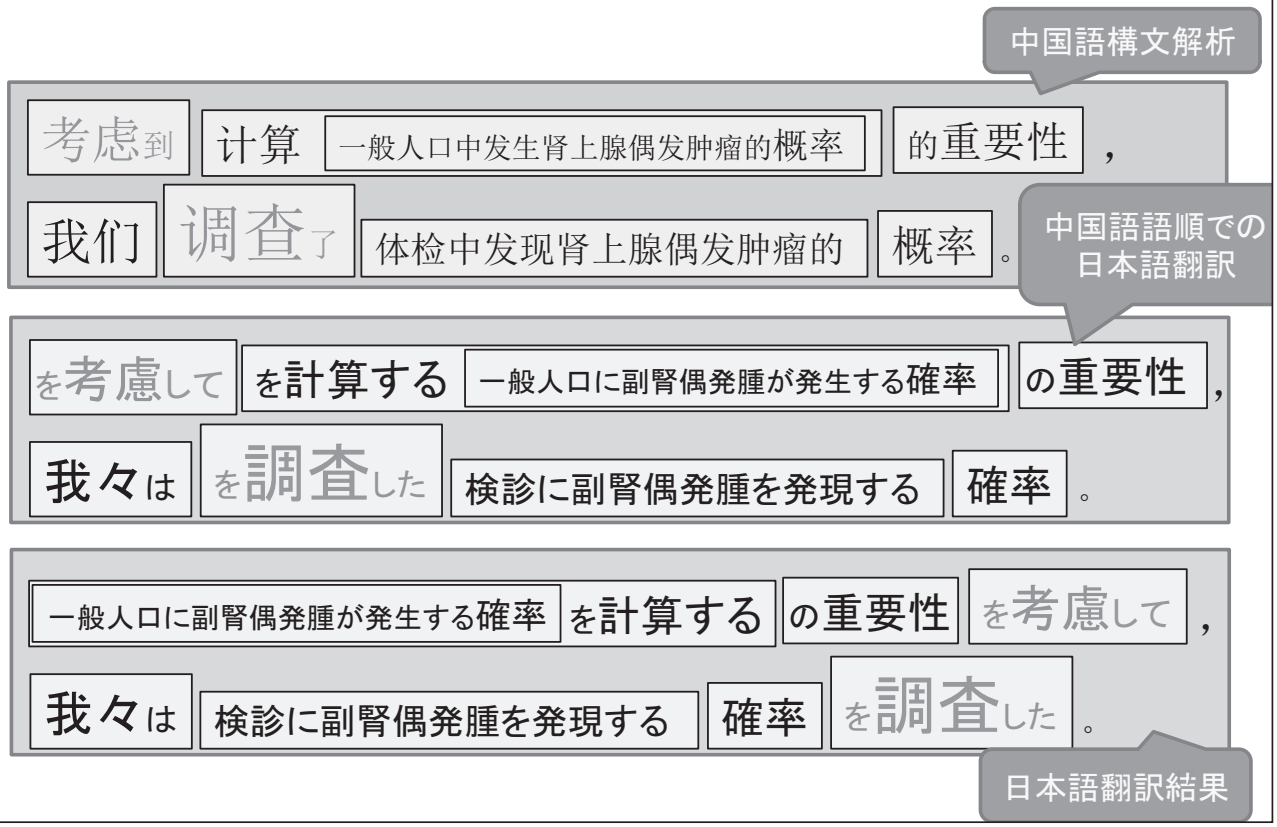
機械翻訳の後編集インターフェース



16

文構造の可視化

[Kishimoto et. al, 2014 WPTP3]



機械翻訳エンジンの開発

18

動機

- 近年のコーパスベース機械翻訳の成功
  - 特に英仏など言語構造の似た言語対
  - ルールベース翻訳よりも高精度なことも
- 言語構造や語順の大きく異なる言語対で高精度な翻訳精度を達成するには構造情報の利用が必須
  - 日英翻訳や日中翻訳など

19

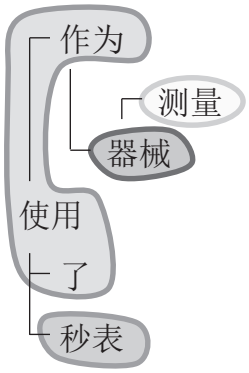
アプローチ

- 依存構造木上での単語アライメント (⇔ GIZA++)
  - [Nakazawa+, COLING2012], [Nakazawa+, IJCNLP2011]
- 依存構造木同士の翻訳 (⇔ Phrase-based SMT)
  - [John+, ACL2014]
- 高速なオンライン用例検索
  - [Cromieres, EMNLP2011]
- ラティス構造を利用した効率的なデコード
  - [Cromieres+, EMNLP2014]

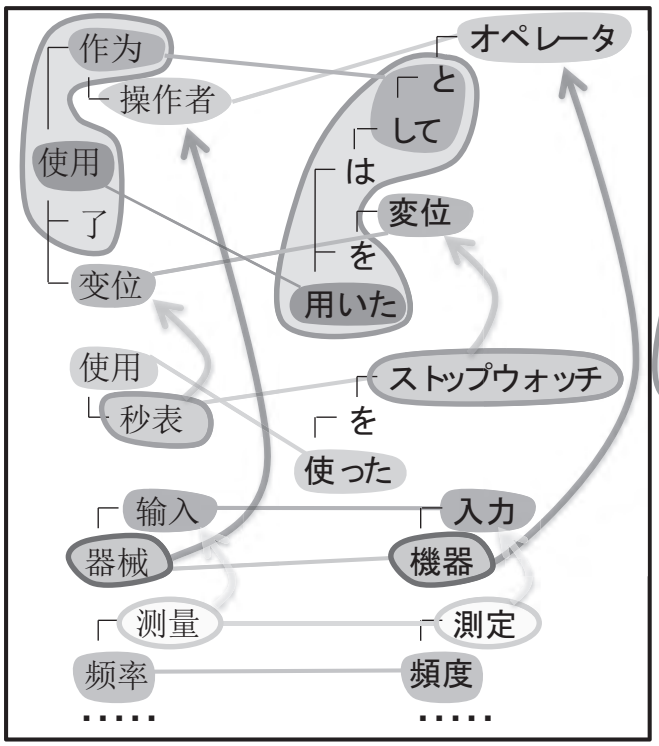
20

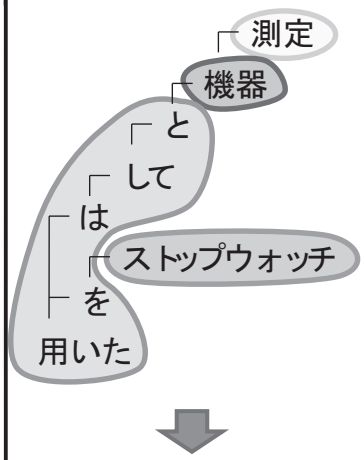
KyotoEBMTの概要

Input:
作为测量器械使用了秒表



Translation Examples



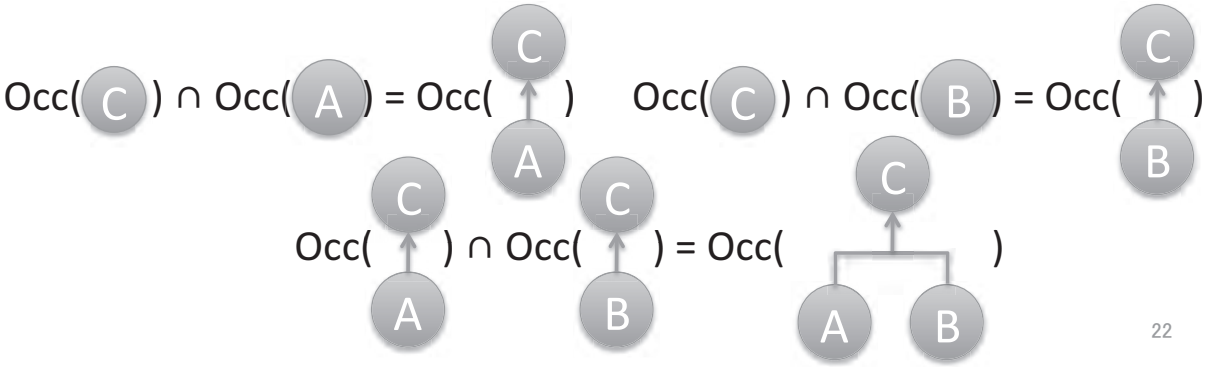


Output:
测定机器としては
ストップウォッチを用いた

21

高速なオンライン用例検索

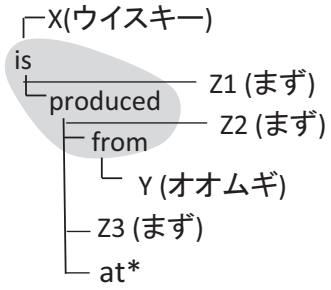
- 対訳コーパス全体から、入力文の翻訳に使える部分（部分木）を高速に検索
  - 事前に全ての翻訳知識を作り出す必要がない
- 転置インデックスを使い、小さな部分木の出現の積集合を繰り返し計算

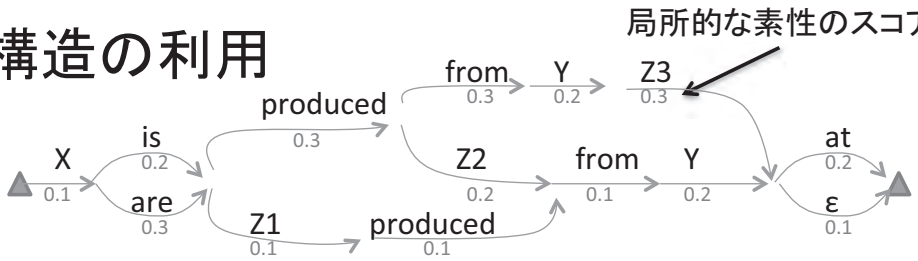


22

ラティス構造によるデコード

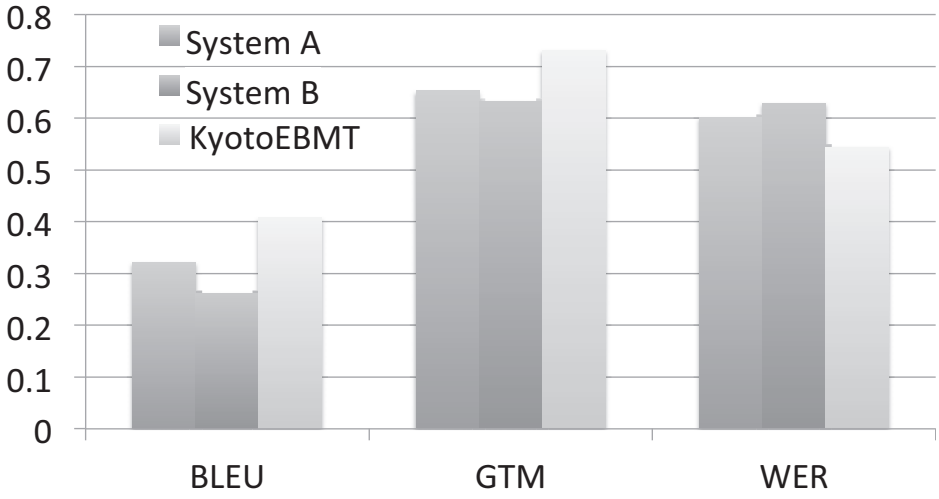
- KyotoEBMTでの翻訳の難しさ
  - アライメント時に対応のない語の扱い（図中の*）
  - 用例の組み合わせ方の曖昧性（図中のZ）
  - 非局所的な素性（言語モデルなど）の利用
- ラティス構造の利用





翻訳精度

- ISTICによる日→中翻訳の自動評価
- 中国国内の翻訳サービスと比較



24

Next Step

- 入力文の構文”森”化
  - 構文解析誤りの影響を低減
- 木構造言語モデルの利用
  - 出力木構造の情報を利用
- Deep Learning技術の利用
  - ~~最近いたるところで流行っている~~
  - 単語、文構造の単なる置き換え作業から意味の翻訳へ

25

1ST WORKSHOP ON ASIAN TRANSLATION (WAT2014)

26

独立行政法人 科学技術振興機構 NICT 京都大学 KYOTO UNIVERSITY http://lotus.kuee.kyoto-u.ac.jp/WAT/

WAT 2014
The 1st 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 (2014年は日本語、中国語、英語のみ)
- 科学技術論文を翻訳対象として採用
- 日⇔中翻訳を言語対として採用
- テストセットが段落単位になっており、文脈を考慮した機械翻訳の可能性を検討可能
- テストセットを含む全てのデータを一般公開
  - ASPECを利用
  - 機械翻訳研究の継続的な発展に貢献

27

独立行政法人 科学技術振興機構 NICT http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

ASPEC
(Asian Scientific Paper Excerpt Corpus)

- 2006年度から2010年度に日本で実施された、科学技術振興調整費による重点課題解決型研究「日中・中日言語処理技術の開発研究」の成果の一部
- 日英科学技術論文抄録コーパス（ASPEC-JE）
  - JSTが所有する約200万件の学術論文日英抄録から抽出された300万文対
- 日中科学技術論文抜粋コーパス（ASPEC-JC）
  - JSTの運営する電子ジャーナルサイトJ-STAGE登載の和文論文を、出版学会の許諾を得て中国語に翻訳して作成した68万文対

28

機械翻訳タスクの参加チーム

Team ID	J->E	E->J	J->C	C->J
NAIST	✓	✓	✓	✓
EIWA	✓			✓
Kyoto-U	✓	✓	✓	✓
WEBLIO-EJ1		✓		
TMU	✓			
BJTUNLP			✓	

Team ID	J->E	E->J	J->C	C->J
NII	✓			
SAS_MT		✓		✓
Sense	✓	✓	✓	✓
NICT			✓	
TOSHIBA	✓		✓	
WASUIPS			✓*	✓*

会社

国外

* 自動評価にのみデータを提出

29

当日の参加者は50名以上！

2014年10月4日撮影

30

WAT2014での自動評価

- 自動評価サーバーを用意
  - 複数の単語分割ツール、BLEUとRIBESで評価
- 現在も稼働中
  - いつでも最新の翻訳結果を継続評価可能

評価結果の閲覧:

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

翻訳結果の提出(無料の利用登録が必要):

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

31

機械翻訳の人手評価

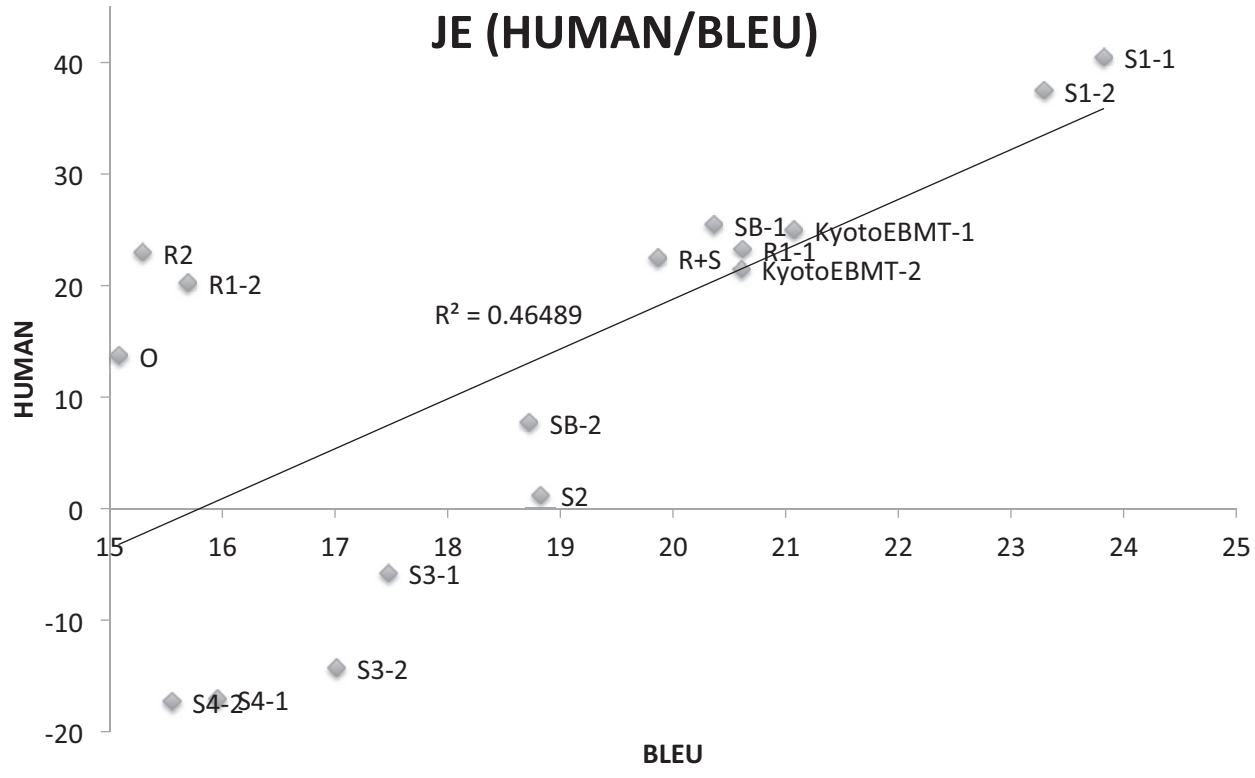
- お金も時間もかかる
- 評価者ごとに基準が異なり、結果が不安定
- 様々な方法が存在
  - Adequacy/Fluency (IWSLT)
  - Ranking (WMT, IWSLT)
  - Acceptability (NTCIR)
  - 特許審査評価 (NTCIR)
  - 特許文献機械翻訳の品質評価手順 (JPO)

32

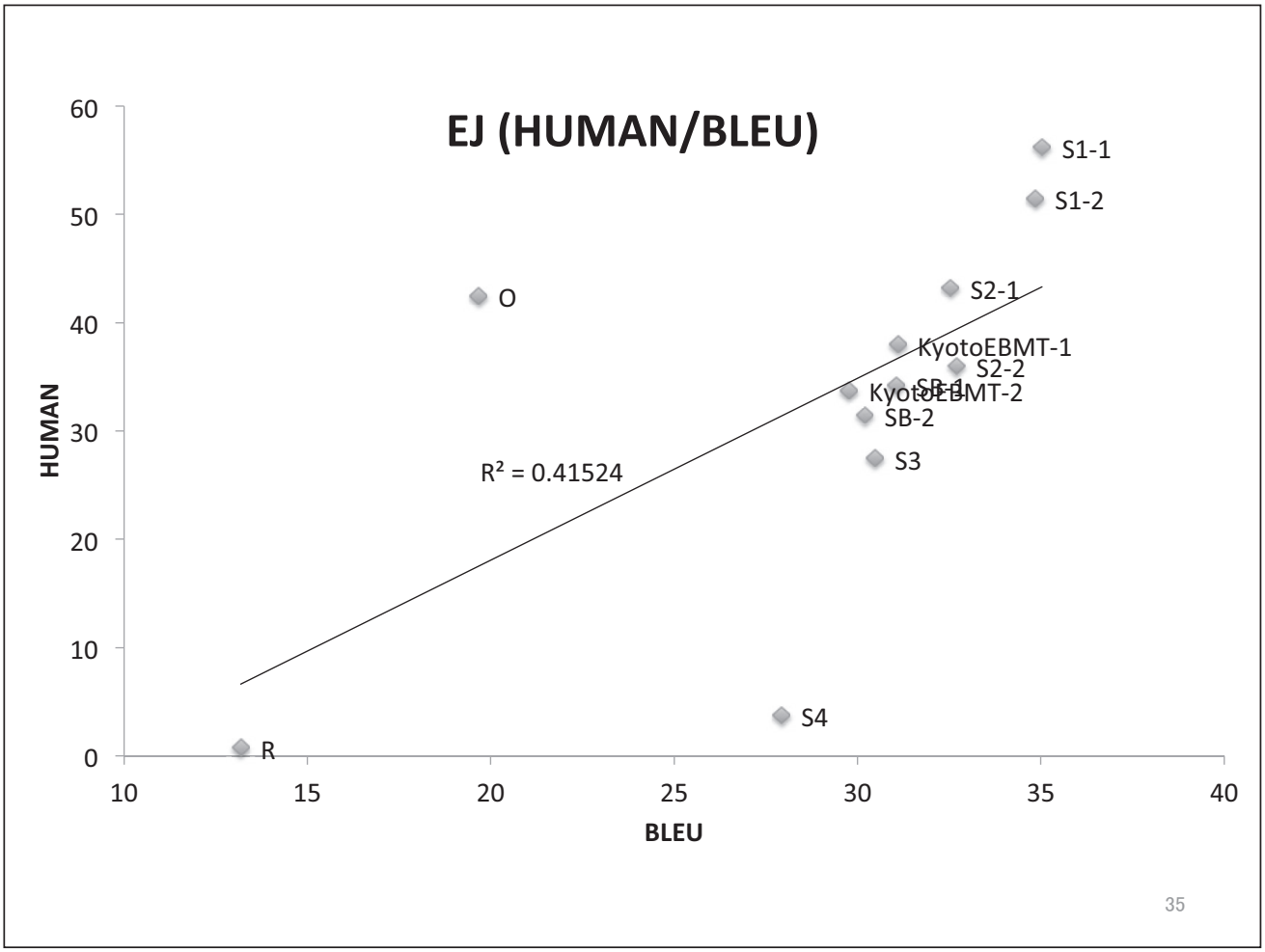
WAT2014での人手評価

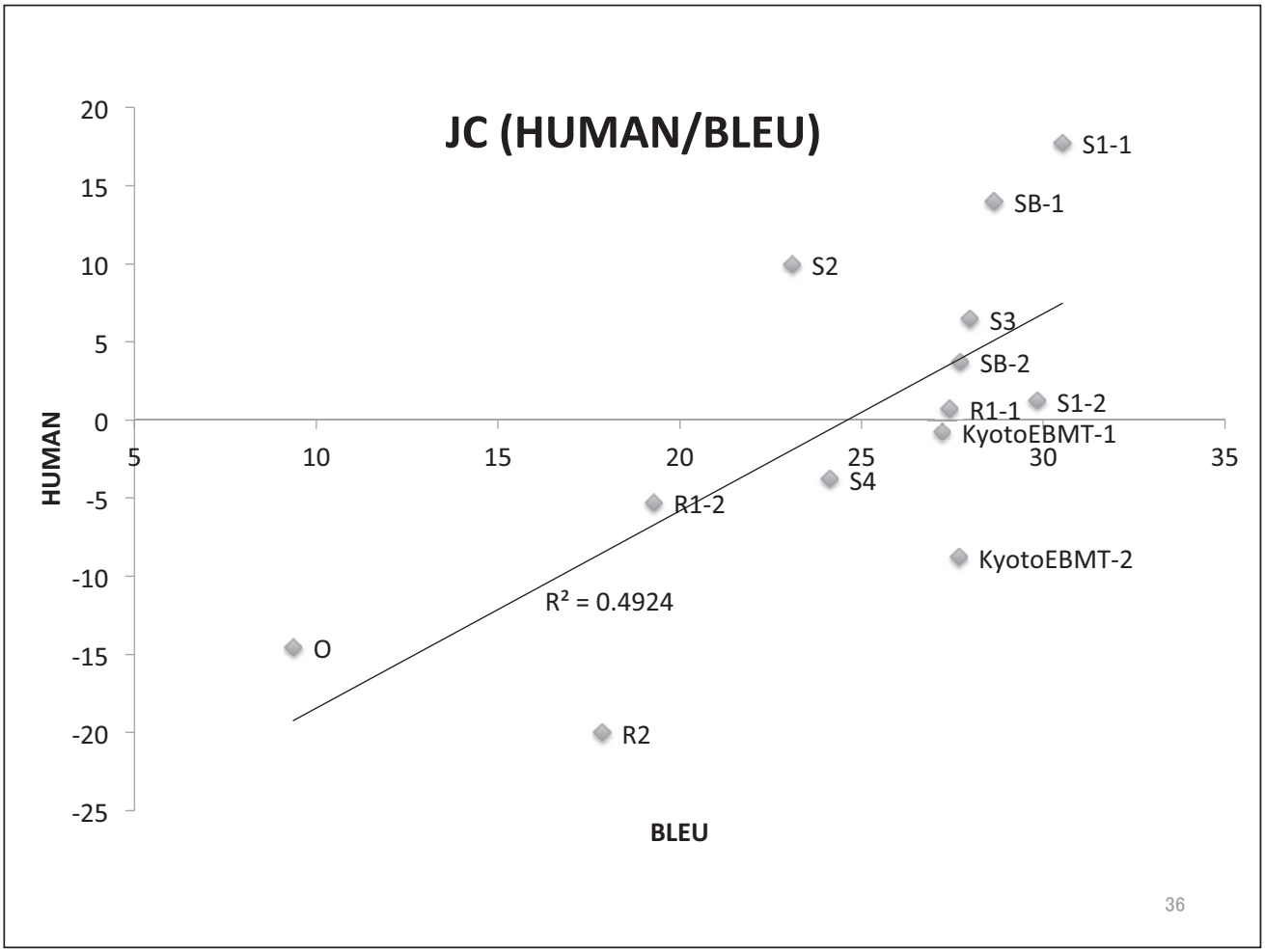
- お金も時間もかかる
  - クラウドソーシングを利用することで低減
- 評価者ごとに基準が異なり、結果が不安定
  - 複数人の評価を用いて総合判断
- 様々な方法が存在
  - HUMANスコアを利用

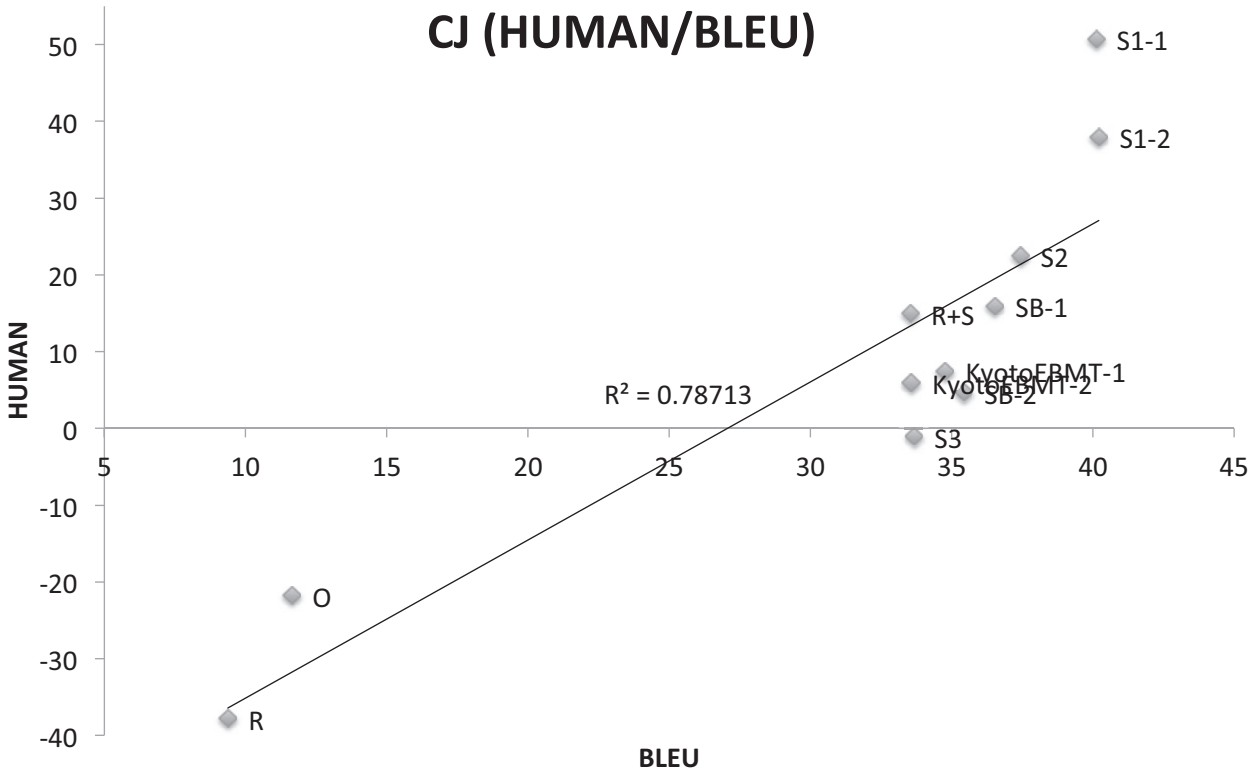
33



34


EJ (HUMAN/BLEU)
HUMAN
BLEU
S1-1
S1-2
S2-1
O
KyotoEBMT-1
S2-2
SB-1
KyotoEBMT-2
SB-2
S3
R² = 0.41524
S4
R
35


JC (HUMAN/BLEU)
HUMAN
BLEU
S1-1
SB-1
S2
S3
SB-2
R1-1
S1-2
KyotoEBMT-1
S4
R1-2
KyotoEBMT-2
R² = 0.4924
O
R2
36



37

Next Step

- WATは毎年開催の予定
  - より多くの言語、ドメインを含める
  - WAT2015で検討中
    - インドネシア語-英語の新聞記事の翻訳
    - 日本語-中国語の特許文献の翻訳
- 言語資源の共有
  - 単言語/対訳コーパス、辞書など
- 文脈を利用した機械翻訳の重要性を検討

38

まとめ

- 英語以外の言語で書かれた文書数の増大
  - 他言語の情報への容易なアクセス方法が必要
- 日中・中日機械翻訳実用化プロジェクト
  - 両国間の科学技術交流を促進
- 近年のアジア諸国の発展
  - 日中韓とASEANなどが一体となり、さらに発展
- 日本での機械翻訳技術の活用は遅れている
  - 研究者と利用者（翻訳者）との歩み寄り
  - 効率的な人手翻訳、言語資源の蓄積

39

一般発表

「統計的訳語選択技術による韓日機械翻訳の高精度化」

統計的訳語選択技術による韓日機械翻訳の高精度化

田中　浩之、園尾　聡、木下　聡、釜谷　聡史

（株）東芝研究開発センター

概要　我々は、当社がこれまで開発してきた規則ベース翻訳技術を基礎として、韓日翻訳エンジンを開発した。規則ベース翻訳技術は、統計翻訳技術と比べて、きめ細かい訳出制御ができることが特徴であるが、高精度な翻訳を実現するには、訳し分け規則等を充実させる必要がある。人手による規則の開発には、長い時間と高い費用を要する為、規則構築を可能な限り自動化することで、コストの低減を図りたい。その解決方法として、対訳コーパス等の知識源から自動的に訳し分け規則を学習することが考えられる。しかし、学習には質の良い対訳コーパスが相当量必要であり、その確保は容易ではない。そこで我々は、単言語の明細書データから作成した大規模 N-gram 言語モデルと、パテントファミリーから対訳関係を推定した対訳文書対で学習した訳語優先度を用い、翻訳辞書中にある訳語候補から適切な訳語を決定する自動訳語選択技術を開発した。韓国語特許明細書の韓日翻訳精度について評価した結果、自動訳語選択をしなかった場合に対し、ＢＬＥＵスコアで８．５ポイントの精度改善効果があった。

1. はじめに

近年、韓国語で出願された特許文書の内容を迅速に把握したいというニーズが高まっている。そこで我々は日・中・英の３言語サポートに加え、新たに韓日・日韓機械翻訳システムを開発した[1]。

これまで当社で開発してきた規則翻訳システムでは、原言語側の構文木を作成し、それを対象言語の構文木に変換する規則を作成することで翻訳を実行する。しかしながら、韓国語と日本語との間の変換に関しては、互いの言語が似通っており、辞書を参照して単語を置換するといった処理である程度までは対処することが可能であるため、構文木ではなく、形態素解析結果の一次元構造を使って翻訳している。

図１は韓日翻訳システムの概要である。文が入力されると、形態素解析器が辞書を参照し、形態素毎に分け、一次元の系列を作る。次に、一次元の形態素系列に対し、可能な訳語を付与し、訳語のラティスを作成する。この時、解析誤りによる誤訳を低減させるため、ある種の品詞を持つ語に対して、同じ表記で異なる品詞を持つ語の訳語を候補として追加する（例えば普通名詞と固有名詞）。また、訳がより自然になるように、一部の助詞に対しては動的に訳語を補完する処理等を行っている（図中の“の”）。その後、訳語候補のラティスから最適な訳語系列を選択し、最後に時制等

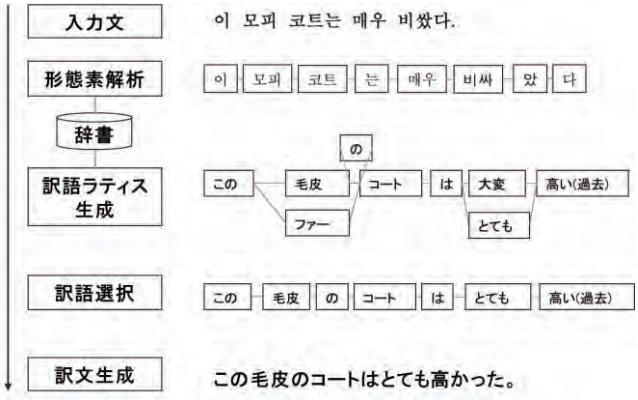


図 1　韓日翻訳システムの概要

を考慮して訳文を生成する。日韓翻訳も基本的には同じ仕組みである。

一方で、訳語を決定するための規則あるいは知識が、高精度な翻訳を行うためには必要である。特に、韓国語特有の問題として同音異義語がある。韓国語は同じ表記で違う意味を持つものが多く、文脈を考慮しないと訳語が決定できないものが多い。例えば、「3 연패」という単語は「３連覇」、もしくは「３連敗」という訳を取りうるため、文脈によって全く正反対の意味となる。よって、とりわけ韓日翻訳においては適切な訳語を選択することが精度に大きく関係する。

このような、文脈に応じた訳語の選択を行うためには、非常に多くの規則を作成する必要がある。しかし、人手による規則の開発は長い時間と高い費用を要す

る。その為、訳語側単言語データで作成可能な言語モデルと、コンパラブルな対訳データを用いて学習する訳語優先度を使って、比較的容易に学習データを用意出来、短期間で高精度な訳語選択を行う手法を開発した。

次の第２章で提案手法と実装上の工夫について詳しく説明する。第３章では韓日翻訳における統計的訳語選択の効果を特許コーパスを用いて検証した結果を紹介する。第４章で実際の処理例を用いて考察を行い、第５章でまとめる。

2. 提案手法

2-1. 言語モデル

言語モデルの構築方法は、例えば Modified Kneser-Ney Smoothing[2,3]や近年ではニューラルネットワークを用いた RNNLM[4]等があるが、本手法では N-gram 確率を線形補間するモデルを用いる:

$$\ln P\left(w_i \middle| \mathrm{w}_{\mathrm{i}-(\mathrm{n}-1)}, \dots, \mathrm{w}_{(\mathrm{i}-1)}\right) = \sum_n \alpha_{\mathrm{n}} \ln p\left(w_i \middle| w_{i-(n-1)}, \dots, w_{i-1}\right)$$

ここで、w_iは i 番目の単語であり、$p(\cdot)$は最尤推定した各次数の N-gram 確率である。

線形補間モデルを採用した理由は、各 N-gram の次数に対するパラメータを後で紹介する他のパラメータと同時最適化出来るようにするためである。また、今回の方式においては、低次の N-gram に関し、その寄与を小さくする方が良い結果となる事が経験的に確かめられた。通常のスムージング手法では低次の項の影響により訳語選択が過度に行われ、流暢だが正確でない翻訳結果となってしまう例が多く見られた。これは、同音異義語が多い韓日翻訳特有の現象である可能性もある。

また、言語モデルの構築を行う際には、分かち書きを行う必要がある。通常、この為には形態素解析器等を用いて文章を形態素系列に変換する。しかし、訳語を選択する場合、辞書中で記述された訳語は、区切りが原言語側の形態素解析結果の形態素区切りに依存する為、必ずしも対象言語の形態素解析結果の区切りと一致しているとは限らない。そこで、辞書中の訳語を形態素の辞書として参照し、全ての形態素の区切れる可能性を展開して N-gram カウントを計算した。

2-2. 訳語優先度

外国出願された特許明細書は、対訳文書対であると見做す事が出来る。そこで今回、韓国から日本に外国出願された日本語特許明細書と、オリジナルとなる韓国特許庁に出願された韓国語特許明細書とを用いて、構造アライメントと文アライメントを行った。

具体的には、１９９６年から２０１２年までの日本の登録特許から優先権主張国が韓国である特許を抽出する。そして、優先権主張番号からオリジナルの韓国特許を検索し、対訳特許文書ペアとした。次に、明細書中の以下の項目がそれぞれ対訳として対応すると仮定し対応付け可能な以下の項目に含まれる文を抽出し、対訳パラグラフ候補を抽出した。

- 請求項
- 技術分野
- 背景技術
- 課題
- 解決手段
- 図の説明
- 発明の形態

そして､各パラグラフから文アライメント器[5]を用いて文アライメントを行った結果、６６１００文の対訳文候補を抽出した。

次に、得られた対訳文候補を用いて訳語の優先度を以下の様に計算する。まず、原文中の一つの語に対する可能な訳語リストを辞書中から取得する。

次に訳文候補中に訳語リスト中の語があるかどうかを判定する。その際、今の単語位置の相対的な位置関係を考慮して探索範囲を限定する。例えば、原文の文字数が２０文字、対訳文候補の文字数が２５文字、注目している原文中の語の先頭位置が１０文字目の場合、対訳文候補の２５×１０／２０の位置を中心に探索を行う（小数点は四捨五入する）。今回、探索範囲の大きさは対訳文候補文字数の 0.5 倍とした。最終的に、最尤推定結果を訳語の訳語優先度とする。

2-3. デフォルト訳語

翻訳辞書中には、予めその語にとって多くの場合用いられる意味に対応する訳語がデフォルト訳語(第一訳語)

	一般ドメイン	特許ドメイン
言語モデル	日本語 Web テキスト ３４億文	明細書テキスト ３億７千万文
訳語優先度モデル	ニュース対訳コーパス（自動アライメント）１７万文	国際出願特許ペア（自動アライメント）６６１００文
パラメータ最適化	ニュース対訳１０００文	国際出願特許ペア１０００文

表 1　使用したコーパスの内訳

	devset	評価用
化学	250	100
機械	250	100
情報通信	250	100
電気	250	100
その他	-	100

表 2　特許ドメインにおけるの最適化、評価コーパスそれぞれの内訳

として格納されている。このデフォルト訳語はある種の人出による訳語優先情報と見做すことが出来る。よって、辞書中の訳語リストに対し、デフォルト訳語には一定の優先度を付与し、他の訳語よりも優先されるように調整した。

2-4 モデル

上記言語モデルと訳語優先度を用いて、対数線形モデルを仮定した：

$$\bar{w} = \mathrm{argmax}_{\mathrm{w}} \sum_i \left(\sum_n \alpha_n \ln P_n(w_i) + \beta \ln P(w_i) + \gamma R(w_i) \right)$$

$P_n(w_i)$は各次数の最尤 N-gram 確率、$P(w_i)$は訳語の訳語優先度、$R(w_i)$はデフォルト訳語フラグであり、デフォルト訳語なら１、そうでないなら０である。$\bar{w}$は出力する訳語系列である。α_n、β、γは最適化パラメータである。

3. 実験結果

実験を行うにあたり、一般ドメインのコーパスと、特許ドメインのコーパスの二種類を準備した（表１）。

一般ドメインのコーパスとして以下のものを用意した。まず、言語モデル構築用に、ウェブ上のテキストをクローリングして収集した、およそ３４億文から

	BLEU	NIST
baseline	0.425	8.65
general	0.502	9.66
patent	0.510	9.78

表 3　各システムの精度評価値

なる雑多な日本語テキストデータを作成した。また、訳語優先度学習用に、ニュース対訳記事から自動で作成した対訳候補文 17 万文を用いた。パラメータ最適化用には、予め作成した１０００文の対訳文を使用した。

特許ドメインのコーパスとしては、言語モデル用に日本語特許公報４年分（２００５―２００８）から作成した日本語コーパス、およそ３億７千万文を用意した。また、2-2. で紹介した対訳文候補の内、１０００文を特許ドメインにおけるパラメータ最適化用の devset、５００文を本実験の評価に用いるコーパスとした。これら１５００文の選定は、国際特許分類（IPC）を用いて表２のように行った。さらに、これら１５００文に対して人出によるチェックを行い、誤りを修正した。残りの６４６００文を訳語優先度学習に用いた。

言語モデルの次数は 4-gram までを用い、モデルパラメータは BLEU[6]を目的関数にして最適化を行った。また、その際の最適化アルゴリズムは Powell 法[7]を用いた。

実験の結果を表３に示す。baseline は訳語選択導入前の、デフォルト訳語を使った訳出の場合のスコアである。general はコーパスに一般ドメインのものを用いた場合のスコア、patent はコーパスに特許ドメインのものを用いた場合のスコアである。訳語選択導入前の BLEU スコアと比較すると、一般ドメインコーパスを用いた場合に７．８ポイント、特許ドメインコーパスを用いた場合には８．５ポイントの精度向上が得られた。

4. 考察

文例１は、baseline、general、patent の各システムで翻訳した文章の例である：

原文： ITO 타겟재를 이용하는 제조 공정시 발생하는 소량의 주석을 포함하는 폐용액들도 원료로 이용할 수 있는 장점이 있다.

baseline: ＩＴＯターゲット材を利用する製造工程市発

生する少量の注釈を含む廃溶液等も原料で利用する長所がある。
general：ＩＴＯターゲット材を利用する製造工程時発生する少量の注を含む廃溶液等も原料と付け入る長所がある。
patent：ＩＴＯターゲット材を利用する製造工程時発生する少量の錫を含む廃溶液等も原料として利用することができる長所がある。

baseline の訳出に対し、general は原文中“공정시”に対応する“工程時”の部分（市→時）が正しく選択されている。一方、patent では“주석”に対応する“錫“という訳語や、”として利用することが出来る“という部分がより良い訳になっている。

その他の例としては、例えば、“同じ層に：同じ階で：同一層に”（baseline:general:patent）、“画素正義幕：画素の定義幕：画素定義膜”、“シートの外見にコーティング：シートの界面にコーティング：シートの外面にコーティング”といった改善箇所が見られた。

本手法における課題としては、文脈に応じて訳語が動的に変化するため、時に訳語の統一感が失われる場合がある。この問題に対しては文書全体に対する文脈処理や、ユーザー辞書中の訳をより強く優先する、といった事が今後必要になると考えられる。また、辞書中に無い訳語は選択できない、といった問題もある。これに関しては、効率よく辞書を拡充していく必要がある。

5. おわりに

本稿では、韓日翻訳システムにおける統計的訳語選択技術について紹介した。本手法は辞書ベースで動作する翻訳エンジンと連携動作するため、対訳コーパスが少量しか用意できない場合でも単言語のコーパスを一定量用意すればよい事が特徴である。これは、対訳コーパスが大量に必要な統計翻訳システムに対する大きな利点であると考えられる。

参考文献

1) The 翻訳®シリーズ http://pf.toshiba-sol.co.jp/prod/hon_yaku/
2) Reinhard Kneser and Hermann Ney, “Improved backing-off for m-gram language modeling”, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 181–184, 1995.
3) Stanley Chen and Joshua Goodman, “An empirical study of smoothing techniques for language modeling”. Technical Report TR-10-98, Harvard University, August. 1998.
4) Mikolov Tomáš et al., “Recurrent neural network based language model”, In Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010.
5) Robert C. Moore, “Fast and accurate sentence alignment of bilingual corpora”, In Proceedings of AMTA, 2002.
6) K. Papineni et al., “Bleu: a method for automatic evaluation of machine translation”, In Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL), July. 2002.
7) Powell, Michael JD. "An efficient method for finding the minimum of a function of several variables without calculating derivatives." The computer journal 7.2, pp. 155-162, 1964.

一般発表

「特許事務所における機械翻訳と人手による翻訳の Mix 事例」

特許事務所における機械翻訳と人手による翻訳の Mix 事例

正林真之　　杉浦伸夫
正林国際特許商標事務所
東京都千代田区丸の内１－７－１２　サピアタワー
http://www.sho-pat.com

国際特許商標事務所における平素の業務で、翻訳ソフトウェアやそれと連携するツール（以下、 機械翻訳という）がどのように活用されているかについて紹介する。

機械翻訳を利用するメリットとして、大意を迅速に把握することができ、特に IT や化学分野で訳語が統一されることが挙げられる一方、訳文検索時にユーザーが直感的に用いると予想される

キーワードとの一致性、もしくは近似性が懸念される。

翻って、プロの翻訳家もしくは企業に大量の翻訳を依頼した場合に、複数の翻訳家が手分けして作業を行ったことにより、訳語が統一されていないケースが散見されることには疑いがない。

また、機械翻訳による前処理がなされていると客観的に感じられることがありながらも、翻訳家は独自に作成した用語集や翻訳メモリー等のノウハウを公開することには慎重である。

このような観点から、ユーザーは機械翻訳とプロの翻訳家を組み合わせ、いかに費用を抑えつつも高い翻訳品質を確保するかが当面の課題となる。

当所では、まずは機械翻訳による前処理を行うことで大意を把握し、重点的に読むべき箇所や人手による翻訳を必要とする箇所を選別し、翻訳品質とコストの配分を図っている。

さらに、特許や医療分野などに特化した機械翻訳が出ている点にも着目し、分野に応じて製品を使い分けることも励行している。

今回は機械翻訳についての当所での具体的な取り組みについて紹介する。

当所では、「翻訳メモリ機能」と「翻訳辞書による機械翻訳」との組み合わせを「Hybrid 翻訳」と位置付け、一定の効果を実感している。それぞれの利用方法とメリットとしては下記が挙げられる。

１）翻訳メモリ機能

【具体的な利用方法】

翻訳作業中、日英の対訳データが、翻訳メモリファイルに自動的に蓄積（フレーズ単位）。蓄積された翻訳メモリからは、完全一致文のみならず、部分一致も検索され、以後の訳文に反映可能。翻訳メモリは、単

語単位で検索可能。
【メリット】
１－１）単独案件内の整合性向上（文単位の訳ぶれ防止）⇒ 単独案件の品質向上
１－２）関連案件間の整合性向上（文単位の訳ぶれ防止）⇒ シリーズ案件の品質向上
１－３）既訳の過去案件の再利用（メモリ化）⇒ 翻訳文再利用による時間短縮
１－４）共通箇所（従来技術や定型文等）の自動反映 ⇒ 繰り返しの翻訳が不要になることによる時間短縮

２）翻訳辞書による機械翻訳
【具体的な利用方法】
　当所では、明細書に頻出する化合物名の羅列などに対し、タイプミスをはじめとするヒューマンエラーを防止することを主目的として翻訳辞書を利用。
【メリット】
２－１）訳語指定の反映、機械翻訳⇒品質向上、時間短縮
２－２）単独案件内の整合性向上（単語単位の訳ぶれ防止）⇒ 単独案件の品質向上
２－３）関連案件間の整合性向上（単語単位の訳ぶれ防止）⇒ シリーズ案件の品質向上

　当所では上記の「Hybrid 翻訳」により、機械翻訳の品質向上と時間短縮の両立が可能であると判断し、日々の業務において実践している。

第３回特許情報シンポジウム

特許事務所における機械翻訳と人手による翻訳のMix事例

2014.11.28

正林国際特許商標事務所
東京都千代田区丸の内１－７－１２ サピアタワー
TEL: 03-6895-4500
FAX: 03-6895-4511
URL: http://www.sho-pat.com

機械翻訳と人手による翻訳の相違点

機械翻訳を利用する最大のメリットとも言えるスピード、人の手による翻訳の特長である正確性など、6項目に分けて比較を行った。

	機械翻訳	人手による翻訳
スピード	速い	遅い～中
正確性	低～高	中～高
訳語の統一性	高	低～高
想像しやすい検索キーワードとの一致性・近似性	低～中	中～高
費用	中 × 1回の支払い（または月額制）	中～高 × 毎回
学習能力	高	中～高

特許庁に提出された日英翻訳における訳語の相違

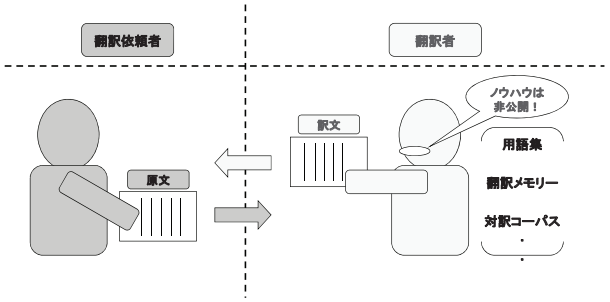
「熱圧着」という日本語の技術用語が、どのように英訳されて特許庁に申請されているかを比較する。

多数派の英訳	少数派の英訳	
Thermal compression bonding	Thermal crimping	Heat application
Thermocompression	Thermal pressure bonding	Heat press bond
Thermocompression bonding	Thermal welding	Heat pressure bond
	Thermally contact-bonded	Hot-press
	Thermally pressure-bonding	Hot press bonding
		Hot pressing

出典 Weblio 英語例文

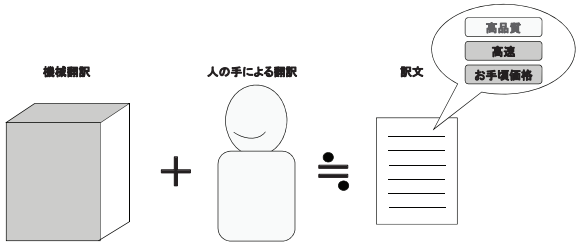
翻訳作業時に発生、蓄積されたノウハウの所有と開示

翻訳家は、独自に作成した用語集や翻訳メモリーを公開することには慎重。



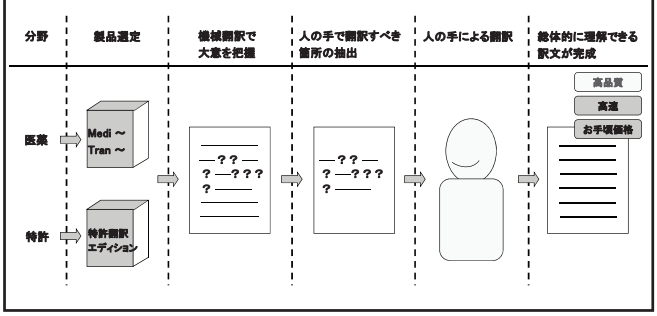
異なる翻訳手法の効果的な活用方法と課題

機械翻訳と人の手による翻訳のそれぞれの長所を組み合わせ、費用を抑えて高い翻訳品質を確保することが当面の課題。



当事務所における機械翻訳と人手による翻訳のMix 事例

機械翻訳による一次翻訳と、それに続けてプロの翻訳家に依頼する翻訳量と費用を抑えつつ、短時間で効率的な翻訳結果を得るために行っているプロセスを紹介する。



Shobayashi International Patent & Trademark Office

Hybrid翻訳のご紹介(1/2)

7

以下の2機能を活用した翻訳

■ 翻訳メモリ機能

翻訳作業中、日英の対訳データが、翻訳メモリファイルに自動的に蓄積。(フレーズ単位)。
蓄積された翻訳メモリからは、完全一致文のみならず、部分一致も検索され、以後の訳文に反映可能。翻訳メモリは、単語単位で検索可能。

【メリット】

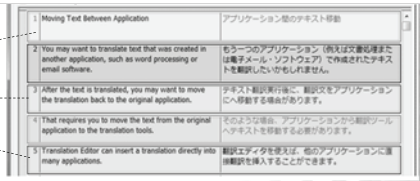
単独案件内の整合性向上(文単位の訳ぶれ防止)⇒単独案件の品質向上
関連案件間の整合性向上(文単位の訳ぶれ防止)⇒シリーズ案件の品質向上
既訳の過去案件の再利用(メモリ化)⇒翻訳文再利用による時間短縮
共通箇所(従来技術や定型文等)の自動反映⇒繰り返しの翻訳が不要、時間短縮

翻訳メモリの検索結果(例)

完全一致

類似文

文型一致



1	Moving Text Between Application	アプリケーション間のテキスト移動
2	You may want to translate text that was created in another application, such as word processing or email software.	もう一つのアプリケーション（例えば文書処理または電子メール・ソフトウェア）で作成されたテキストを翻訳したいかもしれません。
3	After the text is translated, you may want to move the translation back to the original application.	テキスト翻訳実行後に、翻訳文をアプリケーションへ移動する場合があります。
4	That requires you to move the text from the original application to the translation tools.	そのような場合、アプリケーションから翻訳ツールへテキストを移動する必要があります。
5	Translation Editor can insert a translation directly into many applications.	翻訳エディタを使えば、他のアプリケーションに直接翻訳を挿入することができます。

Shobayashi International Patent & Trademark Office

Hybrid翻訳のご紹介(2/2)

8

■ 翻訳辞書による機械翻訳

当所では、明細書全体を機械翻訳することはしていませんが、化合物名の羅列などには、タイプミスをはじめとするヒューマンエラーを防止する上でも有効な機能。

【メリット】

訳語指定の反映、機械翻訳⇒品質向上、時間短縮
単独案件内の整合性向上(単語単位の訳ぶれ防止)⇒単独案件の品質向上
関連案件間の整合性向上(単語単位の訳ぶれ防止)⇒シリーズ案件の品質向上

★ 翻訳における品質向上、時間短縮が両立可能

一般発表

「インターネットから利用できる翻訳ソフトを

優れた辞書として活用する方法」

インターネットから利用できる翻訳ソフトを優れた辞書として活用する方法

Method of Using Translation Software Available from Internet as Excellent Dictionary

吉川潔（Kiyoshi Kikkawa）　新潟県に居住する特許翻訳者

今回の発表の概要

① はじめに
② インターネットから利用できる翻訳ソフトを優れた辞書として活用する方法
③ 名詞句から形容詞・副詞句に拡充
④ 動詞句の語尾変化に追従
⑤ 今後の課題

① 概要

1-1　発表者の翻訳経験

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電子メールで受信し、翻訳後に返信するという翻訳の仕事を、いわゆる、テレワーク（ＳＯＨＯ、在宅作業）という勤労形態で３０年間行ってきた。

特許事務所は、情報処理技術（ＩＣＴ）を用いた特許関連の文献調査に関心がある。そこで、私の翻訳体験から得た特許事務所の意向を、翻訳ソフトの研究や開発の関係者に反映できたらと願っている。

1-2　翻訳ソフトの試訳

翻訳作業中に基本語句とそこから逸脱する語句に出会うと、それらをパソコンに保存していたら約３千に達した。１４年前から、８社が市販する翻訳ソフトを用いて、上記の語句を試訳していた。

そのなかから、主な４社のソフトを試訳対象にしている。各社の廉価版と高価版の翻訳ソフトを購入し差異も調べた。それらは７年前に発売のモデルであるが、最新版の翻訳ソフトといえでも、ＯＳの改訂版に対応するために調整しただけで、翻訳精度の実態は同じという情報を得ている。

1-3　試訳結果からの提案

４年前の第一回特許情報シンポジウムで「翻訳ソフト実用化の提案」を発表し、様々な誤訳の実例を示した。長文の翻訳は誤訳しやすい。というよりも、３０字以上になると（和文英訳の場合、英訳後の単語数）原稿の何処かに「てにをは」を含めた文法ミスが存在しがちである。特許明細書では、語数が１００字以上になることが多いので、人間翻訳者でも、図面と原稿を繰り返し読んで訳している。この翻訳現場の実態を考慮すると、原文が正しい文法のもとで記してあることを条件にする翻訳ソフトに過大な期待は無理である。そこで、翻訳対象の単語数を制限し、短文で数量表現に特化した翻訳ソフトの開発を提案した。

２年前の第二回特許シンポジウムでは、「コンマ（,）と（ａｎｄ）と（ｔｏ）が入り混じる語句の誤訳の実例と対策」を述べた。誤訳と正訳のあいだに規則性がない。自然現象に論理的な原理原則があり数式で表現できるが、言語の規則（文法）には、数式で表現できない難しさがあると言える。

対策として、類似文を最大限にインプットして、共通部分と規則性を見出し、フロ

ーチャート化するか用例翻訳として、翻訳ソフトにプログラミングすることを、誤訳解決の一つとして提案した。

1-4　インターネット上の翻訳ソフト

翻訳ソフトの会社は、インターネットから無料で利用できる幾つかの翻訳ソフトの影響を受けていた（市販し且つインターネット上で無料で提供する会社もある)。

そこで、インターネットから無料で利用できるG社の翻訳ソフトを、前述の主な４社と同様に、英文和訳を重点に試訳した。

それは、G社を含めた５社の比較結果として、本年６月に関係者に配布している。それに加えて、G社の英訳も調べてみた。

更に、最近、M社もインターネット上で無料利用可能な翻訳ソフト提供していることを知り、M社の英訳も調べた。

私が単独で試訳できる範囲は、翻訳分野の全体からみると、ごく一部にすぎない。従って、その試訳結果から、全体の傾向や各社の優劣を指摘することは差し控える。

その調査の過程で、幾つかの翻訳ソフトが名詞句を正確に訳していることに気付いた。今回は、それについて説明する。

②　インターネットから利用できる翻訳ソフトを優れた辞書として活用する方法

特許は発明であり、新しい技術を意図しているので、特許明細書には聞き慣れない単語や用語や造語が続出する。

・翻訳作業で最も困ることは、単語（用語）を辞書で調べても適訳が見つからないことである。

・最も助かることは、どんな単語（用語）でも見つけてくれる「お助けマン」のような辞書の存在である。

2-1　活用方法

ある時、特許明細書の原稿で（双船尾船）という用語に出会った。書籍辞書やCD-ROM辞書に非記載。インターネットから無料で利用できるWeblio（JSM科学技術用語日英対訳辞書など）も非記載。同様に利用可能な翻訳ソフトで訳してみた。

例えば、「**それは（～）です**」という１文のなかに単語を入力して訳してみる。

2-2　事例１（双船尾船）

「それは（双船尾船）です」

と入力すると、

G社　There is a bi-stern ship.
M社　It is the double-stern ship.
E社　It is 双船尾船.
Y社　It is 双船尾船.
N社　It is [arega] [**] stern ship.

船舶用用語集の定義から、G社とM社が適訳らしい。インターネット上で使用実例数を確認して、英訳文に用いた。

2-3　事例２（ラシュバ場）

（ラシュバ場）は磁界に関係する用語なことは分かっていたが、手持ちの辞書をひいても非記載である。上記と同様に、

「それは（ラシュバ場）です」

と入力すると、

Ｇ社　It is Rashba field.
Ｍ社　It's fields.
Ｅ社　It is a ラシュバ place.
Ｙ社　It is ラシュバ ground.
Ｎ社　It is [rashuba] place.

Ｇ社の訳語は、電磁気学の書籍をみると正訳である。そして、インターネットから無料で利用できる翻訳ソフトが、名詞句ならば、**正訳するらしい**ことに気付いた。
そこで、Ｇ社を中心にして調べてみた。

2-4 事例３（ヤドリバエ型マイクロフォン）
上記と同様に入力すると、
[It is a **tachinids microphone**]
という訳文が現れる。
そのなかから、
[tachinids microphone] を選択。

（ヤドリバエ型マイクロフォン）は、
（ヤドリバエ）と（マイクロフォン）の２語を辞書から調べて連結すれば訳出可能。しかし、１回で済めば非常に楽である。

2-5　事例４（ギア列）
「それは（ギア列）です」と入力。
「It is is a gear train」という訳文が現れるので、「**gear train**」を選択する。

並の辞書の場合、(ギア) は「ｇｅａｒ」。(列) は「ｌｉｎｅ」や「ｒｏｗ」。
二つを単純に連結して「gear line」や「gear row」と早合点するのは、？？？

G 社は、学術論文などを含んでいるビッグデータから、統計的な組み合わせに準じて「gear train」を適訳と瞬時に選定するらしい（但し、あくまでも発表者の推定）。

2-6　事例５（メルカプト系シラン）
上記と同様に入力すると、
「It is **mercapto-based silane**」が現れる。

（～系～）は、（～系の会社）や（体育会系の～）は、[~ system], [~ group], [~ series] の３通りある。

化学関連の（～系～）は「~ based ~」が多いことは、インターネットの検索機能から出現数を調べると分かる。複数の候補がある場合、出現数の比較調査も重要。

2-7　事例６
化学系や医薬系の新語には、長々と続くカタカナ表記が多く、辞書に非記載の場合が多い。
例えば、「それらは、(ポリテトラフルオロエチレン、パーフルオロアルコキシエチレン共重合体、テトラフルオロエチレン、ヘキサフルオロプロピレン共重合体、ポリクロロトリフルオロエチレン）です」
と入力すると、
(Pplytetrafluoroethylene, perfluoroalkoxy ethylene copolymer, tetrafluoroethylene, hexafluoropropylene copolymer, poly chlorotrifluoroethylene) と現れる。

各々、長いカタカナが続く単語なのに、
５語を同時に訳してくれる。

辞書といえば、単語を一つ一つ調べるのが

常識なのに、インターネットから無料で
利用できる翻訳ソフトの辞書機能は驚異的。
５０語以上でも全て瞬時に正訳するだろう

2-8　事例７

前述のように、特許は新技術に関係するので、新語が多い。新語には、医化学系のカタカナ表記の名詞句が圧倒的に多く翻訳者泣かせの単語（用語）である。

例えば、下記は、私が最近の翻訳作業で出会った単語であるが、全て正訳だった。

アラミド繊維強化プラスチック、ガラス繊維強化プラスチック、直鎖状低密度ポリエチレン、エチレン酢酸ビニール共重合樹脂、エチレンプロピレンゴム、アルカリ賦活方法、

2-9　事例８

他に、自励発振半導体レーザ、油含有水、遷移金属系材料、血液凝固系解析装置、非水系電解液、分離閉じ込めヘテロ構造、音導管、多孔相互連結構造、ディミング処理、側方透過像、キャニーフィルタ、波長透過特性、シリコン貫通ビア、疎水膜、

いま話題の新語、STAB 細胞やノバルティス社の高血圧治療薬ディオバンも正訳する。

2-10　**辞書として活用した所感**

書籍の辞書や CD-ROM の辞書、Weblio（JSM 科学技術用語日英対訳辞書など）が対応できる単語や用語は、インターネットから無料で使用できる翻訳ソフトは対応（正訳）している。

ところが、上記の｛新語｝や｛複合語｝や｛カタカナ用語｝に対して、書籍の辞書や CD-ROM の辞書、Weblio（JSM 科学技術用語日英対訳辞書など）は全滅しているが、インターネットから無料で使用できる一部の翻訳ソフトは何とか対応している。

但し、辞書の訳語を鵜呑みにせず、英英辞典や用語辞典を用いて、単語の本来の意味を調べて訳すことは翻訳の鉄則。

前述のように、訳語候補を、インターネットの検索機能から、実際の使用例や出現数を調べて比較することも重要。

それらをふまえても、この辞書機能は、翻訳者として待望していたレベルにある。
－ 翻訳ソフトの輝かしい成果 － である
－ 翻訳者として感謝したい －

私は、この優れた辞書機能を日本翻訳連盟（ＪＴＦ）などをつうじて、多くの翻訳者に伝えたい。

③　**名詞句から形容詞・副詞句に拡充**

インターネットから無料利用できる
翻訳ソフトに対して、翻訳者として感謝したい辞書機能は、名詞句の場合だけである。

形容詞句や副詞句や動詞句は、試訳では今一歩である。例えば、

3-1　**数量表現**が苦手

An aperture is reduced to several one-tenths.

「口径を数十分の一に減らす」

It is a pack one pack earlier.

「それは、一つ前のパックである」

3-2　**体調表現**が苦手

例えば、

（今朝から、目がゴロゴロする）

というような体調の翻訳も苦手である。

3-3　**（名詞句＋形容詞・副詞句）の辞書**

但し、数量や体調の表現に新語は少ない。上記のような紛らわしい表現や語句でも、既成の辞書を総動員すれば、なんとか対応できると、翻訳者としての体験から言える。

数量や体調に関連した用語や単語を、翻訳ソフトに、用例翻訳として組み込めば｛名詞句＋形容詞・副詞句｝に限定した優れた辞書になると期待できる。

④　**動詞句の語尾変化に追従**

例えば、（バッグはいくらですか）は正訳する。ところが、バッグは、いくらしましたか、いくらでしょうか、いくらだ、いくらだった、いくらだったか

上記のような動詞句の語尾変化に正確に追従して訳すことが苦手のようである。

この課題について、翻訳者の立場で深く追求していない。しかし、過度に広くないと感じている。この課題は、国語の文法の領域であり、翻訳者だけの立場で問題点を指摘できない。国文法の先生の協力が必須。

⑤　**今後の課題**

前述のように、名詞句の場合、新語であっても、正訳あるいは適訳してくれる。

形容詞句・副詞句、そして動詞句も、同じ発想で進めば実用可能と考える。

一つの文章は、主語（名詞句）と動詞（動詞句）があり、形容詞句と副詞句が修飾し、前後を接続詞で連結すると、単純にいうとこういう構造である。翻訳という作業は、これを無意識に行っているにすぎない。

前述のように、翻訳原稿は長文になると、「てにをは」を含めた文法ミスが生じる。

翻訳ソフトは、原文が文法に準じていることを想定しているので、限界がある。

そこで、下記の開発を要請する。

短文（単語数２０程度）で、

体調と数量の表現が正確な翻訳ソフト

原稿に文法ミスや不明朗な文体が存在しても、短文ならば、ユーザが入力時に、ミスに気付いて再入力すると期待できる。1 文は無理でも、部分訳ならば正訳を目指す

日本は人口が減少し、労働力不足が懸念され、海外労働者の受け入れが必要。協働作業に言語が障壁になるので、翻訳ソフトの実用化が望まれる。

日本語が流暢な中国人でも、日中翻訳時に、中間に英訳文があると助かると言う。

日本語が流暢なドイツ人も同様に言う。

従って、短文でいいから、

日本語　←　→　英訳 の正訳が望まれる。

それをベースにして、各国語への翻訳が期待できる。その機能を、パソコン、電子辞書、スマートフォンなどに搭載することにより、翻訳ソフト（機械翻訳）が実生活で役立つことを願っている。

一般発表

「F タームに基づいたオントロジーの構築」

F タームに基づいたオントロジーの構築

福田悟志[†1]　難波英嗣[†1]　竹澤寿幸[†1]　乾孝司[†2]
岩山真[†3]　橋田浩一[†4]　藤井敦[†5]

†1 広島市立大学大学院 情報科学研究科
†2 筑波大学大学院 システム情報工学研究科　†3 日立製作所 中央研究所
†4 東京大学大学院 情報理工学系研究科　†5 東京工業大学 情報理工学研究科

1. はじめに

本稿では，特許データベースから様々な文献に利用できるようなオントロジーを構築する手法について述べる．オントロジーとは，文献の検索や高度な言語処理に重要な情報源である．しかし，オントロジーを人手で構築し，更新することは非常にコストがかかる．一方で，テキストデータベースからシソーラスやオントロジーを自動構築する様々な手法が提案されているものの，人手による構築作業に取って代わるレベルまでには至っていない．そこで本稿では，最小減の労力で効率的にオントロジーを構築する枠組みについて述べる．

オントロジーを効率的に構築するため，我々は特許分類コード体系のひとつである F タームに着目する．F タームとは，特許を目的・利用分野・材料といった様々な観点から分類することを目的として日本国特許庁が構築した特許の分類体系のひとつである．F タームの詳細については 3 節で述べるが，実は，F タームの構造そのものがオントロジーに近い体系になっている．そこで本研究では，F タームの体系をオントロジーの構築に流用する．これをベースに，ブートストラッピング法と機械学習を組み合わせた手法を用いて，特許との親和性を保持しながら，学術論文など他のジャンルの文献にも利用可能なオントロジーの構築を目指す．

本論文の構成は以下のとおりである．次節では，関連研究について述べる．3 節では，F タームに基づくオントロジー構築手法を提案する．提案手法の有効性を確認するために行った実験について 4 節で報告し，5 節で本論文をまとめる．

2. 関連研究

2.1. 用語間関係の判別

大量のテキストデータから用語間の関係を判別する手法はこれまでに数多く提案されている．一般的に，論文や特許などのテキストベースを対象とする場合，X を上位語，Y を下位語とした時，「Y などの(等の)X」といった定型表現を用いる手法が一般的である[1, 2, 3, 4]．上記の定型表現を用いることで，X と Y は上位下位関係であることを判別することができる．この他にも，安藤ら[5]は，「Y という X」「Y のような X」「Y といった X」などのパターンも上位下位関係を判定するために有用であることを分析している．Kozareva ら[6]は，「X such as Y」「X are Y that」「X including Y」「X like Y」「such X as Y」という 5 種類の上位下位関係を表す表層パターンを用いることで，X と Y の位置関係を判別している．しかし，大量のテキストデータを対象に様々な用語対を判別する場合，上記で述べたパターン以外にも有用なものは数多く存在すると考えられる．また上位下位関係以外の様々な関係(例 : 部分全体関係)においても有用な判別パターンが存在すると考えられるが，これらを網羅的に人手で収集することは困難である．そのため本研究では，ブートストラッピング法[7, 8]により，複数の用語間関係を判別するために有用なパターンを網羅的に収集する．

2.2. ブートストラッピング法

本研究では，シードインスタンス集合を入力とした，パターン抽出とインスタンス抽出の 2 つのフェーズを繰り返すブートストラッピング法を考える．シードとして(X, Y)を与えた場合，パターン抽出フェーズにおいて，X と Y に挟まれているパターンを抽出する．そして，抽出したパターン集合から，「Y などの X」のようにインスタンスと共起しやすいパターンをいくつか選択する．インスタンス抽出フェーズでは，選択したパターンと共起するインスタンスを獲得する．そして，獲得したインスタンス集合からいくつか選択し，再びパターンを抽出する．このような処理を，停止条件が満たされるまで繰り返す．

本研究では，上位下位，部分全体，定義域属性関係という 3 種類の関係を対象に，そのカテゴリに属する新たな用語対を抽出することを目的としている(3.1 節で詳しく述べる)．このように複数のカテゴリを対象とした場合，特定のカテゴリに属するいくつかの用語対をシードとして用いる事が一般的である．

複数のカテゴリを対象にしたブートストラッピング法によるパターン・インスタンスの獲得を行う研究は数多く存在する．Krishnan ら[9]は，医学分野の特許文書集合を対象に，ブートストラッピング法を用いることで用語間の Treatment 関係(例:「A cure B」)と Causal 関係(例 :「A impact B」)を表す動詞(句)を抽出している．小町ら[10]および伊藤ら[11]は，Tchai アルゴリズムと呼ばれるブートストラッピング法を用いて，旅行，金融，番組名，芸能人などのカテゴリを対象に，Web 検索履歴から関連性の高いキーワード群を抽出している．Abe ら[12]は，動名詞を伴う複数の事態間関係(行為効果関係，部分全体関係)を収集するために，ブートストラッピング法を用いている．また，Kiso ら[13]は，ブートストラッピング法における重要な問題の一つである「各カテゴリにおける良質な(シード)インスタンスをどのように発見するのか[a]」を，HIS アルゴリズム[14]と組み合わせることで解決している．

上記で述べた研究におけるアプローチは全て Espresso アルゴリズム[15]に基づいている．これは，近年注目されている非常に精巧なブートストラッピング法のひとつである．次節では Espresso アルゴリズムの詳細を述べる．

2.3. Espresso アルゴリズム

公開公報から新たな用語間関係を収集するためのブートストラッピング法として, 本研究では Espresso アルゴリズムを適用する．Espresso アルゴリズムは，少量のシードインスタンスを用いて反復的に表層パターンの抽出を行い，多くの新たなインスタンスを収集する手法である．このアルゴリズムでは，従来のブートストラッピング法で問題となっていた意味ドリフトを考慮している．意味ドリフトとは，反復過程において複数のカテゴリで出現するようなパターン(ジェネリックパターン)やインスタンスを獲得してしまい，徐々にシードと関連性の低いものに移り変わっていく現象である．従来では，ジェネリックパターンを排除することで意味ドリフトを抑えることを行っていたが，獲得できるインスタンスの数が減少するという問題が新たに発生し，その結果，精度は高いが再現率が十分でないという欠点があった．このような意味ドリフトによる問題を軽減するために，Espresso アルゴリズムでは，スコアリング関数を用いて相互再帰的にインスタンスとパターンのスコアを定義している．これは，信頼度の高いパターンと頻繁に共起するインスタンスは信頼度が高く，信頼度の高いインスタンスと共起するパターンは非常に信頼性があるという考えに基づいている．パターン p およびインスタンス i ($i=\{x,y\}$ (*x,y*: インスタンスにおける用語))のスコアをそれぞれ$r_\pi(p)$，$r_\iota(i)$とした時，以下の式を用いて信頼度を計算する．

$$r_\pi(p)=\frac{1}{|I|}\sum_{i\in I}\frac{pmi(i,p)}{\max pmi}r_\iota(i) \quad (1)$$

$$r_\iota(i)=\frac{1}{|P|}\sum_{p\in P}\frac{pmi(i,p)}{\max pmi}r_\pi(p) \quad (2)$$

$$pmi(i,p)=\log\frac{|x,p,y|}{|x,*,y||*.p,*|} \quad (3)$$

ここで，P はパターン集合，I はインスタンス集合であり，$|P|$と$|I|$はパターンとインスタンスの数を表す．$|x,p,y|$はインスタンスを伴うパターン p の頻度を表している．また，$|x,*,y|$はインスタンスの頻度，$|*.p,*|$はパターンの頻度を表す．*pmi(i,p)*はインスタンスとパターン間の自己相互情報量(PMI: Pointwise Mutual Information)を表しており，max *pmi* は全てのインスタンスとパターンの組み合わせの間における *pmi* の最大値である．なお，$r_\pi(p)$，$r_\iota(i)$の初期値はそれぞれ 1 である．

Espresso アルゴリズムでは，反復過程において(1)式と(2)式を適用することで，精度を高く保ちながら再現率を大幅に向上させている．本研究では，Espresso アルゴリズムによるブートストラッピング法を用いることで，特許との親和性を保持した新たな用語間関係を獲得することを目指す．

2.4. 機械学習による関係判別

Espresso アルゴリズムを用いることで, シードインスタンスとの関連性の高いパターンを獲得することができるが，獲得したパターンを用いてインスタンスを抽出する場合，特定のカテゴリに特化したようなインスタンスが必ず獲得されるとは限らない．例えば，上位下位関係のカテゴリに属するシードインスタンスを用いて「Y といった X」というパターンが獲得されたとする．このパターンを用いて新たなインスタンスを獲得する場合,「キーボードといった入力装置」という文から(入力装置，キーボード)というインスタンスが獲得されるが,「キーボードといった入力部」という文が存在する場合，(入力部，キー

[a] 一般的には，特定のカテゴリ内で頻出するものを選択する方法，人手により選別する方法，ランダムに選択する方法が挙げられる．

ボード)というインスタンスが抽出される．このインスタンスは，部分全体関係にあるものと考えられるため，上位下位関係のカテゴリから除去する必要があるが，収集した全てのインスタンスを人手で判定することは困難である．

Girju ら[16]は，C4.5 と呼ばれる分類器を用いてインスタンスを分類する手法を提案している．この手法では，Iterative Semantic Specialization (ISS)手法により，パターンによる分類ルールを学習しており，高い精度と再現率を示している．しかし，訓練用データの作成に多大なコストを要しており，人手によるタグ付けを行う必要がある．また，対象としている用語間関係が部分全体関係のみである．本研究では，F タームにおける複数の用語間関係を対象としており，ブートストラッピング法により獲得した複数のカテゴリにおけるパターン集合を組み合わせて機械学習に適用することでインスタンスの判別を行う点で異なる．また，本研究で用いる訓練用データは，F タームに基づいたオントロジーを用いるため，非常に信頼性が高いという点が挙げられる．

3. F タームに基づくオントロジーの構築

本節では, F タームに基づくオントロジーの構築手法について述べる．本研究では，以下のステップによりオントロジーの構築を行う．

- Step 1: F タームからの知識抽出
- Step 2: Step 1 で得られた知識(用語間関係)をシードとして新たな用語間関係を獲得

各ステップにおける詳細を次節で述べる．

3.1. F タームからの知識抽出

3.1.1. F タームとは

F タームは，特許を目的・効果・構成などの様々な観点から分類することを目的とした分類体系であり，技術分野を示すテーマコードと観点の集合から構成される．ここでは，機械翻訳分野の F タームを例に説明する．機械翻訳には 5B091 という 1 つのテーマコードが, また「言語」(AA00),「処理対象要素」(AB00),「翻訳方式」(BA00)などの 9 個の観点が設けられている．ある機械翻訳システムについて考えた場合，そのシステムの対象言語は何か，どんな仕組みで翻訳するのか，などの属性が存在するが，これがそれぞれ「言語」(AA00)や「翻訳方式」(BA00)などの観点にあたると考えて良い．

F タームでは，観点が階層化されており，例えば，「言語」(AA00)という観点には，この観点を具体的に示す「・多言語間」(AA01)や「・2 言語間」(AA03)といった F タームコードが存在する．F タームコード間で一般/具体関係がある時には，ドットレベル記法で表すことになっている．図 1 の例では,「翻訳方式」(BA11)の下位分類として「直接翻訳」(BA12)と「間接翻訳」(BA13)があり，さらに「間接翻訳」の下位分類には「トランスファー方式」(BA14)と「ピボット方式」(BA17)がある．

BA11	・翻訳方式
BA12	・・直接翻訳
BA13	・・間接翻訳
BA14	・・・トランスファー方式
BA15	・・・・意味解析
BA16	・・・・・文脈解析
BA17	・・・ピボット方式

図 1 テーマ”5B091(機械翻訳)”の F タームコードの例

3.1.2. F タームからの知識抽出

本研究で構築するオントロジーでは, 3 種類の用語間の関係「上位・下位」「属性・定義域・値域」「全体・部分」を扱う．図 1 のドットレベル記法では明示されていないこれらの関係を人手で判断し，図 2 のような知識を獲得する．

関係 1	属性: 方式 定義域: 機械翻訳 値域: 直接翻訳，間接翻訳
関係 2	上位: 間接翻訳 下位: トランスファー方式
関係 3	上位: 間接翻訳 下位: ピボット方式
関係 4	属性: 利用技術 定義域: トランスファー方式 値域: 意味解析
関係 5	属性: 利用技術 定義域: 意味解析 値域: 文脈解析

図 2 図 1 から得られる知識

3.2. F タームからの知識をシードとして利用した用語間関係の獲得

Step 1 で得られた知識(用語間関係)をシードとし，F タームオントロジーには存在しない新たな知識を公開公報データベースから自動的に収集する. F タームからの用語間関係をシードとして利用した用語間関係の獲得は，以下のステップから構成される．

- Step 2-1: Espresso アルゴリズムによるパターン・インスタンスの獲得
- Step 2-2: 機械学習による Step 2-1 で得られたインスタンスのクリーニング
- Step 2-3: 新たなインスタンスが獲得されなくなるまで Step 2-1，Step 2-2 を繰り返す

図 3 に上記のステップの流れを概略図として示す．本手法におけるパターン，インスタンス獲得はパターンマッチを用いて行う．また，インスタンス抽出フェーズでは，パターンの前後に存在する名詞(句)を抽出する．Step 2-1 および Step 2-2 における詳細を次節で述べる．

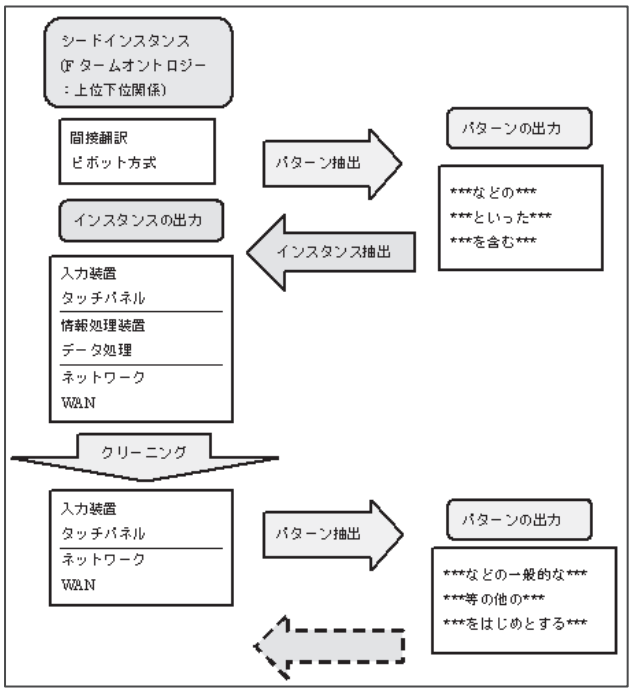


図3　F タームからの知識をシードとして利用した用語間関係の獲得の概要

3.2.1. Espresso アルゴリズムによるパターン・インスタンスの獲得

Espresso アルゴリズムにおけるシードインスタンスとして，本研究では，3.1 節で構築した F タームオントロジーから特定の関係(上位下位，部分全体，定義域属性)に属する少数の用語対を用いる．そして，そのインスタンス間に出現しているパターンを獲得し，(1)式を用いてパターンの信頼度を計算する．次に，パターンの前後に出現する用語対をインスタンスとして獲得し，パターンの信頼度と(2)式を用いてインスタンスの信頼度を計算する．その後，獲得されたインスタンスを用いてパターンを収集する．このように, F タームの体系を流用したインスタンスおよびパターンの抽出と信頼度の計算を繰り返すことで，特許との親和性を保持しながら，特定の関係に属するインスタンスを獲得することができる．

ここで, F タームからのシードインスタンスを用いて獲得するパターンについて，「X<パターン>Y」と「Y<パターン>X」という 2 種類の組み合わせから，それぞれのパターンを抽出する．例えば，F タームにおいて上位下位関係を持つ(間接翻訳，ピボット方式)というシードインスタンスから，「ピボット方式<パターン>間接翻訳」および「間接翻訳<パターン>ピボット方式」という組み合わせにより，パターンをそれぞれ抽出する．この処理を他の 2 種類の関係においても適用するため，合計 6 種類のパターン集合が獲得される．その後，そのパターンを抽出した組み合わせからインスタンスを新たに抽出する．そのため，6 種類のインスタンスが獲得される．ここで，より精巧なパターンおよびインスタンスを獲得するために，獲得されたパターン・インスタンスの選定を行う．具体的には，獲得したパターンまたはインスタンス集合において, 2 種類以上の関係に属するパターン(インスタンス)を除去する．これは，複数の関係に出現するパターンは，特定の関係のみに出現するインスタンスを発見する場合において有益でないと考えられるためである(インスタンスの選定においても同様)[17].

3.2.2. 機械学習による獲得したインスタンスのクリーニング

機械学習によるインスタンスの判別に対する概要を図 4 に示す．本研究では，各カテゴリにおけるシードインスタンスを用いて獲得したそれぞれのパターン集合を組み合わせて用語間の関係を判別する．各セル内の値は，「Y<パターン>X」という表現が公開公報データベース中で何回出現しているかを示している．これらの値の組み合わせにより，各カテゴリにおいて収集したインスタンスの用語間が統計的に成立しているかどうかを機械学習により判断する．

機械学習による用語間関係の判別を行うときに重要となることは，カテゴリ間で収集したパターンをどのように組み合わせるかということである．本研究では，各カテゴリにおいて獲得したパターンが他の同様の組み合わせによるカテゴリ内にどのくらい存在しているかを統計的に判定することに焦点を当てる．そのため，「下位－上位」「部分－全体」「属性－定義域」のカテゴリで収集したインスタンスの判定を行う場合には，これらのカテゴリ内で獲得したパターンを組み合わせる．同様に，「上位－下位」「全体－部分」「定義域－属性」のカテゴリにおけるインスタンスの判別では，これら 3 種類のカテゴリにおいて獲得されたパターンを組み合わせる．

カテゴリ	パターン	X=入力装置 Y=キーボード	X=入力部 Y=キーボード	X=入力装置 Y=発明
下位<パターン>上位	YといったX	134	6	0
	Yなどの他のX	24	1	0
	YのようにX	1	0	3
部分<パターン>全体	Y等からなるX	30	34	0
	Yで構成されたX	5	1	0
	Yを含めたX	0	1	0
属性<パターン>定義域	YにおけるX	1	0	156
	Yにおいて、X	0	0	46
	Yにより、X	0	0	2

図 4 機械学習によるインスタンス内の用語間の関係判別

3.3. 複数言語による用語間関係の判別

本アプローチの特徴は，言語に依存していないという点である．ブートストラッピングアプローチにおけるパターン抽出では，パターンマッチにより 2 つの用語間に存在するものを抽出しており，インスタンス抽出では，パターンの前後にある名詞(句)を抽出している．また，機械学習アプローチでは，パターンとインスタンスの組み合わせの出現回数から統計的に用語間の関係を判別している．そのため，シードインスタンスとして用いる F タームを対象の言語に翻訳することで，様々な言語で記述された特許文書に対しても同様のアプローチを適用できると考えられる．

さらに，複数の言語を対象に，それぞれの抽出したパターンやインスタンスを比較することで，より正確な用語間の関係を判別することができると考えられる．例えば，「スマートフォン等の携帯端末」と「スマートフォン等のバッテリー」という文について考える．上記の 2 文は日本語公開公報に存在するため，(携帯端末，スマートフォン)だけでなく，(バッテリー，スマートフォン)に対しても上位下位関係が成立してしまう．しかし英語特許中に，「mobile computers, such as smartphones」という文が存在し，「batteries, such as smartphones」が存在しない場合，(携帯端末，スマートフォン)は上位下位関係として成立するが，(バッテリー，スマートフォン)は上位下位関係でない確率が高くなると考えられる．

ここで，どのように複数の言語間を対応付けるのかについて考える．専門用語の訳語推定法については，統計的機械翻訳モデルを用いて専門用語の訳語を推定する手法および既存の対訳辞書を利用した要素合成法を併用して専門用語の訳語を推定する手法が提案されている[18]．本研究では，抽出したパターンやインスタンスを専門用語とみなしたとき，統計的機械翻訳モデルを用いる手法について着目する．これは，統計的機械翻訳では，対象とする言語に関する文法的知識を必要としないため，容易に翻訳システムを構築することができるためである．また，統計的機械翻訳ツールには GIZA++[b]を使用し，入力言語を日本語，出力言語を英語としたとき，約 90%の精度で翻訳することができると報告されている．

本研究でシードインスタンスとして用いた F タームには日本語版を翻訳した英語版がある．これをブートストラッピングアプローチのシードインスタンスとして用いる．そして 2 種類の言語(日本語，英語)によるパターンとインスタンスをそれぞれ抽出する．そして，統計的機械翻訳モデルにより，言語間のパターンとインスタンスに対して対応付けを行う．その後，機械翻訳により，抽出したインスタンスの関係を判別する．このように，複数の言語による，より多くの根拠を利用することで，正確な用語間の自動判別を実現できると考えられる．

4. 実験

3 節で提案した手法のうち，各カテゴリにおけるシードインスタンスを用いて獲得したパターンと，機械学習によるインスタンスの判別の有効性を調べるために実験を行った．なお，本実験では，3.1 節と 3.2 節で述べた手法に対する実験のみを行った．

4.1. 実験方法

4.1.1. 実験条件

本実験のブートストラッピングアプローチにおける実験設定は以下のとおりである．

- シードインスタンスとして，F タームコードリストから，各カテゴリに属する 20 個の(名詞(句)で構成されている)用語対をそれぞれ人手で選択し，パターンを抽出する．
- インスタンス抽出フェーズにおいて，ランク付けされた上位 50 パターンを用いる．

本実験では，1 回目の反復により獲得したパターンとインスタンスを用いて提案手法の有効性を確かめる．また，機械学習に用いるパターンとして，Espresso アルゴリズムによる信頼度によってランク付けされた各パターン集合から，それぞれ上位 100 パターンを使用した(表 1)．

4.1.2. 実験データ

本実験では，以下の 2 種類のデータを用いた．

- 日本国特許全文データ：公開公報 1993-2012 年 (396,532 文書，約 14GB)

[b] http://www.fjoch.com/GIZA++.html

表 1 機械学習に用いるパターンの例

下位<P>上位	部分<P>全体	属性<P>定義域
を行う	を格納する	を備える
等の他の	で構成された	を提供する
といった	等からなる	を有することを特徴とする
を保持する	として多用されている	を備えたことを特徴とする
上位<P>下位	全体<P>部分	定義域<P>属性
、すなわち	に格納されている	を実行する
、例えば、	が有する	を実現する
を用いているが、	には、複数の	の基本的な
システムにおける	が記憶する	を行うように制御する

- F タームから獲得した用語間関係リスト : 11,842 個 (102 テーマ)

特許全文データに関して，本研究では，情報分野に関連する IPC コード G06F，G06K，G06T，G11C が付与されているデータを対象に実験を行った．これに関連し，上記の IPC コードを F タームのテーマの範囲に含むものを F タームリストから抽出を行った．その結果，F ターム全 2,790 テーマ中，102 テーマが選出され，これらのテーマに関連する F タームコード 11,842 個が抽出された．

機械学習に用いる訓練用データとして，シードインスタンスを除いた F タームコードを使用した．機械学習に用いるインスタンスにおいて，それぞれが名詞(句)のみで構成されているもののみを対象とした．その結果，上位下位，部分全体，定義域属性関係において，2,719 個，664 個，415 個の F タームコードを機械学習の素性に用いた．

4.1.3. 評価方法

評価用データは，以下の手順により作成した．

1. 各カテゴリに対する反復過程においてランク付けされた 6 種類のインスタンス集合からそれぞれ上位 100 個および上位 500 件から 600 件までのインスタンスを選択する[c]．
2. 各インスタンスに対して，そのカテゴリが表す関係として本当に正しいかどうか人手で判定する．

[c] 上位に出現したインスタンスは一般的な表現のものが多く，下位にランク付けされた結果ほど特徴的なインスタンスが出現している傾向があった．そのため，各ランクの位置におけるインスタンスに対して，本手法がどのくらい性能を示すのかについて調査した．

評価尺度として，精度と再現率を用いた．また，機械学習を用いない場合をベースラインとする．

4.2. 実験結果

実験結果を表 2 に示す．ベースラインと比較すると，「下位－上位」「属性－定義域」「定義域－属性」に関して，再現率を 70-80%程度に保ちながら精度を向上させていることがわかる．この結果から，機械学習による用語間関係の判別は有効であることがわかる．しかし，部分全体関係に関して，精度が向上しないまま再現率が大幅に低下している．これは，部分全体関係を判別するような特徴的なパターンが上位に出現していないからだと考えられる．また，本実験では，機械学習に用いるパターンの数や組み合わせ方法，ブートストラッピングアプローチにおけるパラメータ(シードインスタンス数，インスタンス抽出に用いるパターンの数)を固定していたため，これらの最適な値の設定を調査する必要がある．最後に，本手法により獲得したパターンおよびインスタンスを表 3 に示す．

表 2 実験結果

上位 1-100 件	提案手法		ベースライン	
	精度	再現率	精度	再現率
下位<P>上位	0.576	0.792	0.480	1.000
上位<P>下位	0.526	0.788	0.520	1.000
部分<P>全体	0.500	0.314	0.510	1.000
全体<P>部分	0.698	0.448	0.670	1.000
属性<P>定義域	0.973	0.706	0.510	1.000
定義域<P>属性	0.893	0.848	0.790	1.000
上位 500-600 件	提案手法		ベースライン	
	精度	再現率	精度	再現率
下位<P>上位	0.513	0.848	0.460	1.000
上位<P>下位	0.400	0.829	0.410	1.000
部分<P>全体	0.417	0.227	0.440	1.000
全体<P>部分	0.529	0.383	0.470	1.000
属性<P>定義域	0.857	0.732	0.410	1.000
定義域<P>属性	0.700	0.672	0.470	1.000

4.3. 考察

前節でも述べたように，機械学習によるインスタンスの判別性能をさらに向上させるためには，そのカテゴリに対するより特徴的なパターンを用いる必要があると考えられる．例えば，「を添付した」というパターンは「部分<パターン>全体」の組み合わせからのみ獲得できる特徴的なパターンであるが，Espresso アルゴリズムによる信頼度の値は低かった．このような特徴的なパターンの信頼度を向上させるための方法として，情報利得を用いたリランキング方法が考えられる．

表 3 本手法により獲得されたインスタンス

	下位<P>上位		部分<P>全体		属性<P>定義域	
上位 1-100 件	上位：処理 下位：送信	上位：情報 下位：アドレス	全体：データ 部分：タグ	全体：テーブル 部分：データ	定義域：プログラム 属性：機能	定義域：処理 属性：手段
上位 500-600 件	上位：処理 下位：書き込み	上位：操作 下位：削除	全体：レジスタ 部分：アドレス	全体：テーブル 部分：キー	定義域：ＩＣカード 属性：手段	定義域：システム 属性：構成
上位 1-100 件	上位：入力装置 下位：キーボード	上位：処理 下位：表示処理	全体：テーブル 部分：ブロック	全体：キャッシュ 部分：アドレス	定義域：受信 属性：受信手段	定義域：アクセス 属性：手段
上位 500-600 件	上位：アドレス 下位：物理アドレス	上位：構成 下位：組合せ	全体：画面 部分：検索結果	全体：記憶部 部分：テーブル	定義域：情報処理装置 属性：形態	定義域：記憶 属性：記憶手段
	上位<P>下位		全体<P>部分		定義域<P>属性	
上位 1-100 件	上位：情報 下位：電子メール	上位：処理 下位：取得	全体：キーボード 部分：キー	全体：レジスタ 部分：値	定義域：アクセス 属性：手段	定義域：アクセス 属性：方法
上位 500-600 件	上位：情報 下位：識別情報	上位：情報 下位：識別情報	全体：メモリセルアレイ 部分：メモリセル	全体：ＲＯＭ 部分：制御プログラム	定義域：アプリケーション 属性：機能	定義域：認証 属性：手段
上位 1-100 件	上位：不揮発性メモリ 下位：フラッシュメモリ	上位：コンピュータ 下位：サーバ装置	全体：チャネルＭＯＳトランジスタ 部分：ゲート	全体：プリンタ 部分：印刷データ	定義域：移行 属性：手段	定義域：バックアップ 属性：機能
上位 500-600 件	上位：情報 下位：個人情報	上位：情報 下位：購入者	全体：フラッシュメモリ 部分：ブロック	全体：制御部 部分：制御プログラム	定義域：リフレッシュ 属性：手段	定義域：ジョブ 属性：要求

情報利得を用いることで，特定のパターンに対するカテゴリへの曖昧さ(エントロピー)を把握し，曖昧さが少ないパターンを選定することができると考えられる．表 4 に，3 種類のカテゴリ(「下位－上位，部分－全体，属性－定義域」または「上位－下位，全体－部分，定義域－属性」)を対象とした各カテゴリにおけるパターンのリランキング結果の例を示す．なお，リランキング方法として，Espresso アルゴリズムによる信頼度と情報利得値を掛け合わせ，値の高い順に並べ替えている．また，これらのパターンをインスタンス抽出フェーズに用いることで，Espresso アルゴリズムによるパターン集合だけでは獲得できない特徴的なインスタンスを抽出することができると考えられる．

表 4 情報利得による各カテゴリ内のパターン集合のリランキング結果

下位<P>上位	部分<P>全体	属性<P>定義域
システムの	として添付された	により生成された
手段の	を添付した	を通じて
に代わる	を除いた当該	にて
装置及びその	を除いた前記	により、前記
上位<P>下位	全体<P>部分	定義域<P>属性
を行い、	に添付されている	を受け付ける
部における	に添付された	の構成を示す
は従来の	から分離された	の一実施例を示す
手段による	に記憶された	の一実施例の

5. おわりに

本研究では，特許データベースを対象に，ブートストラッピング法と機械学習を組み合わせた手法を提案した．本手法では，F タームに基づいたオントロジーをシードインスタンスとして使用しており，特許との親和性を保持した新たな用語間関係を獲得できることを示した．また，本手法は言語に非依存である．今後は米国特許を対象とした実験を行い，統計的機械翻訳を用いて 2 種類の言語間を対応付けることで正確な用語間の自動判別を行うことを目指す．

参考文献

1. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora, *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 539-545, 1992.
2. 相澤彰子: 類語関係抽出タスクにおけるコーパス規模拡大の影響, 情報処理学会研究報告 自然言語処理, NL-175, pp. 91-98, 2006.
3. Nanba, H.: Query Expansion using an Automatically Constructed Thesaurus, *Proceedings of the 6th NTCIR Workshop Meeting*, pp. 414-419, 2007.
4. Kozareva, Z. and Hovy, E.: Learning Arguments Supertypes of Semantic Relations using Recursive Patterns, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1482-1491, 2010.
5. 安藤まや, 関根聡: 上位語・下位語を含む連体修飾表現の言語的分析, 言語処理学会第 10 回年次大会, 2004.
6. Kozareva, Z. and Hovy, E.: A Semi-Supervised Method to Learn and Construct Taxonomy using the Web, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1110–1118, 2010.
7. Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (SCL'95)*, pp. 189-196, 1995.
8. Abney, S.: Understanding the Yarowsky algorithm, *Journal of Computer Linguistics*, Vol. 30, No. 3, pp. 365-395, 2004.
9. Krishnan, A., Cardenas, A.F. and Springer, D.: Search for Patents using Treatment and Causal Relationships, *Proceedings of the 3rd International Workshop on Patent Information Retrieval*, 2010.
10. 小町守, 鈴木久美: 検索ログからの半教師あり意味知識獲得の改善, 人工知能学会論文誌, Vol. 23, No. 3, 2008.
11. 伊藤淳, 戸田浩之, 廣嶋伸章, 望月崇由, 鈴木智也, 筧捷彦: クエリログをコーパスとした意味知識獲得法の改善, 第2回データ工学と情報マネジメントに関するフォーラム(DEIM2010), 2010.
12. Abe, S., Inui, K. and Matsumoto, Y.: Acquiring Event Relation Knowledge by Learning Co-occurrence Patterns and Fertilizing Co-occurrence Samples with Verbal Nouns, *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pp. 497-504, 2008.
13. Kiso, T., Shimbo, M., Komachi, M. and Matsumoto, Y.: HITS-based Seed Selection and Stop List Construction for Bootstrapping, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2, pp. 30-36, 2011.
14. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Enviroment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604-632, 1999.
15. Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pp. 113-120, 2006.
16. Girju, R., Badulescu, T. and Moldovan, D.: Automatic Discovery of Part–Whole Relations, *Journal of the Computational Linguistics*, Vol. 32, No. 1, pp. 83-135, 2006.
17. Curran, J.R., Murphy, T. and Scholz, B. Minimising.: Semantic Drift with Mutual Exclusion Bootstrapping, *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, pp. 172-180, 2007.
18. 森下洋平, 梁冰, 宇津呂武仁, 山本幹雄: フレーズテーブル及び既存対訳辞書を用いた 専門用語の訳語推定, 電子情報通信学会論文誌 D, Vol. J93-D, No. 11, pp. 2525-2537, 2010.

一般発表

「特許文書からの化学物質情報の抽出」

特許文書からの化学物質情報の抽出

Recognizing Chemical Information from Patent Document

池田 紀子† 田中 一成†

Noriko IKEDA† Kazunari TANAKA†

†株式会社富士通研究所 〒211-8588 川崎市中原区上小田中 4-1-1

†FUJITSU LABORATORIES LTD. 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa 211-8588, Japan

E-mail: †{nona, tanaka.kazunari}@jp.fujitsu.com

化学物質は、様々な規則が用いられて、構造の表記や名称の命名が行われている。よって、一つの化学物質について、構造や名称の記載が多様である。そのため、化学分野の専門家であっても、技術文書の化学物質名から、異表記を同定し、その構造を示す化学式を認識することが難しい場合がある。技術文書の調査や分析を行う場合、化学物質名に対応する化学式を得ることができれば、たいへん有用な情報となる。そこで、特許文書をコーパスとして用いて、有機化合物の化学物質名に対応する化学式の抽出を試みた。さらに、命名規則を用いて、得られた化学物質名と化学式の対応関係を部品化し、データベースを作成した。それらの部品を組み合わせて利用することで、特許文書から化学式を抽出できなかった化学物質名について、新たに化学式を生成する手法を考案した。この手法について実証実験を行ったので報告する。

キーワード 特許情報、化学物質、構造、名称、化学式、命名法、抽出

Abstract There are various kinds of methods for the chemical formula and the nomenclature of the chemical substance. So it is difficult to remember all the structures and the names even if it is a specialist of the chemistry. We tried to recognize chemical information in a corpus of patent document. And then we stored them in the database. Moreover, the relation between the chemical name and chemical formulae of organic compound was able to be increased by the chemical nomenclature rules and stored in the database. The new chemical formula was able to be designed by combining the data from the database.

Keyword patent document, chemical substance, structure, name, chemical formula, nomenclature, recognition

1. はじめに

化学物質には様々な表記法や命名法があるため、化学物質について、特許や論文などの化学技術情報を検索・抽出・分析する場合には、化学物質名の異表記の問題を避けて通ることができない。例えば、有機化合物の「1-メトキシ-2-プロパノール」という化学物質名については、以下のような命名法による名称[1,2]や表記法による化学式や登録番号がある。

①命名法による名称

・置換命名法（構造を表現する体系名）：1-メトキシ-2-プロパノール

・付加命名法（体系名）：1-メトキシ-2-ヒドロキシプロパン

・基官能命名法（構造を表現しない慣用名）：プロピレングリコールメチルエーテル

②表記法による化学式(元素構成を表現)

・組成式（元素組成を表現）：C4H10O2

・示性式（基を連結して分子構造を表現）：CH3OCH2CH(OH)CH3

③表記法による登録番号

・CAS 登録番号：107-98-2

化学の学問分野は、有機化学、無機化学、高分子化学などを始めとし、多くの専門分野が存在する。同じ化学物質であっても、専門分野や着目する性質などによって、名称が異なったり、表記が異なったりする。このため、様々な命名法や表記法が使われているので、専門家にとっても全ての名称や表記を把握することは難しい。

化学物質名の異表記問題を辞書で対処すると、次の限界が指摘されている[3]。

・日々、新しい物質が誕生
・表記に関する基準や方針が時代とともに変化
・後発医薬品により商品名が増加
・書き手が勝手に作成

さらに、同一の化学物質以外にも、以下のような問題も存在する。

・同じ元素組成でも、構造が異なる異性体が、複数存在

・人手で作られた辞書は高額な使用料がかかる場合もあり、調査や分析の支援として商用利用するには、知的財産権の問題が大きな障壁

上記のような背景は、特許や論文の調査を行う際、さらに、大きな障害となる。特に、研究領域の拡大から、他分野の技術を調査する研究者や、専門知識の十分でない知財部門の担当者にとっては、コストや品質に影響する大問題である。

本報告では、化学物質情報の理解を支援する目的で、化学物質名と化学式の対応関係を抽出する手法について考案し、実証実験を行ったので報告する。

2. 化学物質名と化学式の対応関係を抽出

次の手法を考案した。

①特許文書をコーパスとして用いて、化学物質名に対応する化学式を抽出し、データベースを作成

②有機化合物の命名規則を用いて、①で得られた化学物質名と化学式の対応関係を部品化し、データベースに蓄積

③部品を組み合わせて利用することで、直接、化学式を抽出できなかった化学物質名について、新たに化学式を生成

2．1　特許文書から直接抽出

特許庁は 1993 年以降の特許文書を電子化して発行している。この特許文書をコーパスとして用いて、化学物質名と化学式の対応関係の抽出を試みた。化学物質の中から、炭素(C)骨格に水素(H)が結合した構造を基本構成とする有機化合物を選択した。有機化合物について、化学物質名と化学式の対応関係を次のルールを用いて抽出した。

①片仮名、英数、「酸」などの一部の漢字、括弧が連続して並ぶ文字列を抽出

②括弧書きを利用して、化学物質名と化学式の対応関係を抽出

例えば、「プロパン(CH3CH2CH3)」のように書かれている場合に、化学物質名の「プロパン」と化学式の「CH3CH2CH3」が対応しているとして抽出

③化学式を英数文字と括弧のみと限定して、炭素(C)と水素(H)を含む場合のみを抽出

しかし、上記ルールでは、多数の誤った対応関係が出現した。それらは、想定外の漢字や平仮名が出現したために、文字列の途中や、間違った対応関係が抽出されたからである。そこで、これらの誤りを減らすために、抽出した対応関係から、以下を削除することにした。

・開き括弧と閉じ括弧の数が合わない
・数字＋単位のみ
・化学物質名と化学式の両方が英数字のみ

上記を抽出ルールに追加したところ、9630 の化学物質名と化学式の対応関係が抽出できたので、化学式の部品データベースに蓄積した。なお、1.で例とした、「1-メトキシ-2-プロパノール」に対応する化学式も抽出できた。

2．2　化学式の部品

特許文書から抽出できた化学物質名と化学式の対応関係のみから、化学式を特定できる化学物質名の数は少ない。例えば、「1-メトキシ-2-プロパノール」CH3OCH2CH(OH)CH3 に骨格の炭素(C)が 1 つ増加した「1-メトキシ-2-ブタノール」CH3OCH2CH(OH) CH2CH3 については、特許文書から対応する化学式が抽出できなかった。

命名規則を用いて、化学的に意味を持つように、「1-メトキシ-2-ブタノール」を分割すると、「1-メトキシ」と「2-ブタノール」に部品化できる。部品化した化学物質名に対応する部品の化学式は各々抽出できた。「1-メトキシ」CH3O と「2-ブタノール」CH3CH(OH) CH2CH3 の化学式の部品を組み合わせることによって、未抽出の「1-メトキシ-2-ブタノール」と化学式の対応関係を生成できると考えた。

2．3　原子価

元素は、原子価(ある原子が何個の他の原子と結合するかを表す数)を持っており、炭素(C)は 4、水素(H)は 1、酸素(O)は 2 である。化学物質では、分子が結合する場合、原子価に応じて、外れる原子があるので、単純に部分構造の化学式を組み合わせても正しい化学式が生成できるとは限らない。

2．4　部品から生成する化学式

例えば、「1-メトキシ」CH3O と「2-プロパノール」CH3CH(OH)CH3 から、1-メトキシ 2-プロパノール CH3OCH2CH(OH)CH3 の化学式を生成する場合

を考える。単純に足し合わせた場合、CH3OCH3CH(OH)CH3となり、誤りである。実際に結合する場合、原子価による制約で、プロパノールの水素が 1 つ引き抜かれて結合するので、CH3OCH2CH(OH)CH3となる。

化学物質の構造を理解しやすいように、1-メトキシ-2-プロパノール(1)と 1-メトキシ-2-ブタノール(2)について、部品から生成する構造式(化学式)を図 1 に示す。

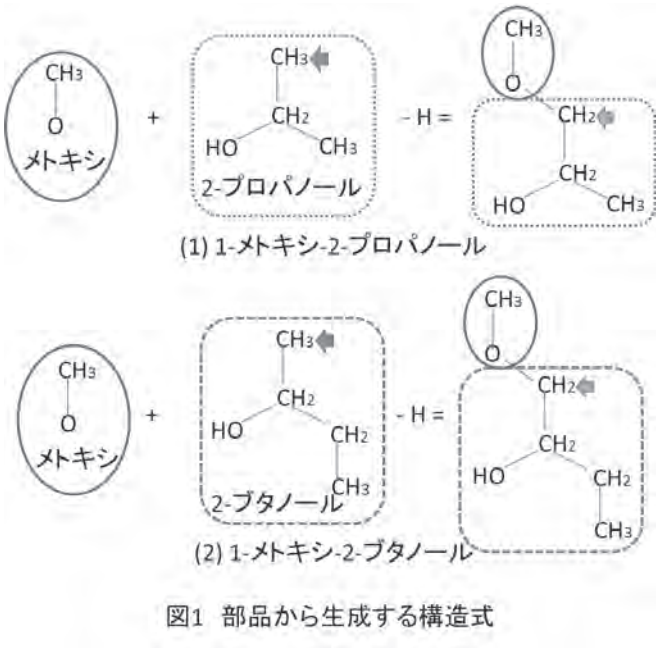


図1 部品から生成する構造式

3. 化学式の部品データベースの作成

3．1 化学式の部品化

化学物質名と化学式の対応関係を増やすために、有機化合物の命名規則などを化学式の部品生成ルールとして利用することにした。部品化ルールとしては以下のような 10 個の基本ルールを使用した。

①原子価を考慮
②水素 1 個を削除し、部分名の語尾を「タン(アン)」から「チル(イル)」に変換
③水素を「水酸基、ヒドロキシ基」(OH)に置き換える場合、部分名の語尾を「ン」から「ノール(オール)」に変換
④水素 1 個をハロゲン(塩素、フッ素…)1 個に置き換える場合、「クロロ、塩化」、「フルオロ、フッ化」…に置き換え
⑤同じ基が 2 個では「ジ」、3 個では「トリ」に置き換え

これらのルールを再帰的に適用することで、大量の部品を生成することができる。ルールを適用する再帰回数の限度を 3 回と設定して部品を作成すると、特許文書から抽出できた 9630 の対応関係から、約 20 倍近くの 178838 の化学物質名と化学式の対応関係が生成できた。これらを化学式の部品データベースに蓄積した。

3．2 追加した部品化ルール

前述の基本ルールでは、抽出できない複雑な構造に対応できるように、以下のオプションの部品化ルールを追加した。

・エポキシ：隣り合う炭素から水素 1 個ずつを削除して、酸素 1 個を追加
・ジオン:炭素 2 個から各水素 2 個を削除して、酸素 2 個を追加
・「アミン」から「アミノ」に変換：窒素から水素 1 個を削除

3．3 化学式の差分から部品生成

上記の部品化ルールでは、未知部品の抽出を想定していない。そこで、差分から部品のバリエーションを増やす方法を考案した。差分は、化学物質名と化学式の対応関係の既知部分を再帰的に削除した残りとした。この差分の化学物質名と化学式の対応関係も、化学式の部品データベースに蓄積した。

図 2 に示すように、抽出された化学物質名と化学式の対応関係から、既知部分の化学式と部分名を引いていき、最後に残った部分の化学式と部分名を対応付けて、化学式の部品データベースに登録し、さらに、部品のバリエーションを増やした。

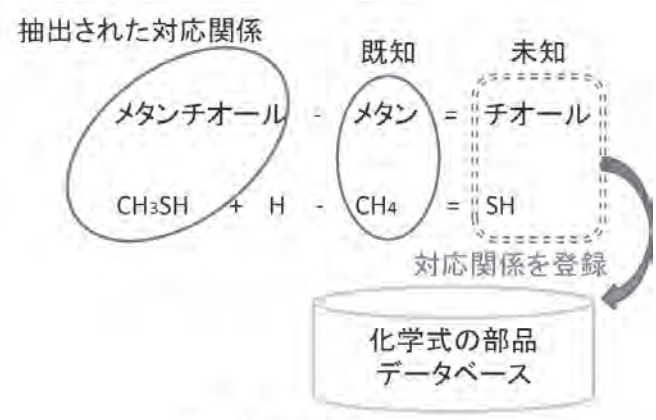


図2 差分による部品化例

4. 化学式の生成

4．1　化学式生成ルール

次のルールで、化学式の生成を試みた。

①目的とする化学物質名の文字列と前方一致で一致する部分名を探索、なお、文字列では、化学式特有の括弧やカンマなどの記載を解析
②一致する部分名が見つかった場合、一致する文字列を目的とする化学物質名から削除
③残りの文字列と前方一致する部分名を繰り返し探索
④体系的名称を用いて、見つかった部分名に対応する部分化学式をつなげることで、化学式を生成

4．2　化学式生成と問題点

例として、「メトキシプロパノール」の化学式の生成を示す。

①まず、前方一致で探索すると、「メトキシ」を得る。
②次に、「メトキシプロパノール」から「メトキシ」を引いた「プロパノール」を前方一致探索すると、「プロパノール」の化学式を得る。
③再帰的に全ての組み合わせを探索すると、複数の化学式を得る。「プロパノール」は、「ヒドロキシ基」(OH)が結合する場所に、3 つのバリエーションがある。よって、「メトキシプロパノール」の化学式では、以下を得る。

- 1-メトキシ-1-プロパノール
  [CH3OCH(OH)CH2CH3]
- 1-メトキシ-2-プロパノール
  [CH3OCH2CH(OH)CH3]
- 1-メトキシ-3-プロパノール
  [CH3OCH2CH2CH2(OH)]

また、抽出誤りに起因するゴミも混入し、さらに、多くの化学式を得るので、一意に集約できない。これは、化学構造上の元素の位置関係が不明確であるために起こる問題である。

5. 実証実験

5．1　実験データ

1)　入力データ

「特定化学物質の環境への排出量の把握等及び管理の改善の促進に関する法律の別表第一」[4]から、有機化合物以外の化学物質名や慣用名を除いた、化学物質名 71 個を用いた。

2)　出力データと比較する正解データ

日本化学物質辞書 Web[5]や Wikipedia から引用した化学式を用いた。

5．2　化学式の生成

以下の 4 つの手法を順に用いて、化学式を生成した。

①特許文書から直接抽出した化学物質名と化学式の対応関係
②基本の部品化ルールを適用して得られた部品
③オプションの部品化ルールを追加して得られた部品
④化学式の差分から得られた部品

5．3　実証実験の結果

5．2 の実証実験の結果を、正解の化学式が得られた化学物質名の数とその割合で、表 1 に示す。

表1　実証実験の結果

対応関係作成手法	手法ごと		手法の累計	
	正解数[個]	正解率[%]	正解数[個]	正解率[%]
①特許文書から直接抽出	14	20	14	20
②基本の部品化ルール	12	17	26	37
③オプションの部品化ルール	3	4	29	41
④差分	8	11	37	52

①特許文書から直接抽出できた情報では、約 2 割が正解であった。
②部品化ルールを追加して用いると、約 4 割近くが、正解であった。
③オプションの部品化ルールを追加し、④差分による部品を使うと、約 5 割が正解であった。

5．4　化学物質の大きさの比較

化学物質の大きさの違いによる正解率を比較した。5．2 の結果から、炭素数が 10 個未満の化学物質 41 個を対象とした場合と、炭素数が 10～14 個の化学物質 20 個を対象とした場合を抽出して、正解率の比較を行った結果を表 2 に示す。

表2　化学物質の大きさによる実証実験の結果比較

対応関係作成手法	炭素10個未満の41個			炭素10～14個の20個		
	正解数[個]	手法ごと正解率[%]	手法の累計正解率[%]	正解数[個]	手法ごと正解率[%]	手法の累計正解率[%]
①特許文書から直接抽出	13	32	32	1	5	5
②基本の部品化ルール	10	24	56	2	10	15
③オプションの部品化ルール	3	7	63	0	0	15
④差分	2	5	68	6	30	45

1) 炭素数が 10 個未満の小さい方の化学物質を対象とした場合

①特許データから直接抽出の手法では、約 3 割が正解であった。①～④までの手法を全て用いると、約 7 割が正解であった。

2) 炭素数が 10～14 個の大きい方の化学物質を対象とした場合

①～④までの手法を全て用いると、約 5 割が正解であった。

大きい方の化学物質では、④の差分による部品化による正解率が高かったものの、全体として、小さい方の化学物質の方が、正解率が高かった。

6. 考察

今回の実証実験から、以下の知見が得られた。

1) 特許文書から抽出できる化学物質名と化学式の対応関係をそのまま使うだけではなく、化学物質の命名規則などを部品化ルールとして利用することで、より多くの化学物質名と化学式の対応関係を得られることがわかった。

2) 特許文書から抽出できた情報に部品化ルールを追加することによって、炭素数が少なく小さい構造では正解率があがった。炭素数が多く大きい構造では、正解率の上昇は小さかった。これは、炭素数が多い程、より複雑な構造となるので、化学式の生成が困難になるからである。加えて、複雑な部分構造に対して、慣用名が使われる場合が増えるからである。例えば、「アニソール」は、「メトキシベンゼン」の慣用名である。もし、「アニソール」が、「メトキシベンゼン」と同義であると判明できれば、「メトキシ」と「ベンゼン」に分割できるので、正解の化学式を得ることが可能である。慣用名辞書を整備すれば、さらなる、正解率向上が見込まれる。

3) 酸素含有の環構造の「エポキシ」や二個の二重結合の「ジオン」などは、単純に水素を 1 つ引き抜いて結合する基とは異なるので、単独の部品としての扱いが難しい。オプションの部品としてルール化し、別に部品化することで、機械的に扱うことが可能になった。

4) 差分による部品化に、2)や 3)を再帰的に適用することで、正解率向上が見込まれる。今回、不正解だった「キノキサリニル」を例としてあげる。

①「2-メトキシ-3-クロロ-6-ニトロキノキサリン」C9H6ClN3O3 の対応関係が抽出できている。

②①から、「メトキシ」(CH3O)、「クロロ」(Cl)、「ニトロ」(NO2)を削除すると、「キノキサリン」となる。C9H6ClN3O3 から、前述の部品に相当する元素を削除し、水素 3 個を加えると「C8H6N2」となる。これは、「キノキサリン」の正解データと一致する。

③「キノキサリン」から水素 1 個を削除し、「ン」を「ン」+「イル」=「ニル」に置き換えると、「キノキサリニル」C8H5N2 となり、正解データの対応関係と一致する。

5) 「キノキサリン」は、窒素含有縮合環の慣用名の一つである。さらなる、正解率向上には、オプションの部品化ルールに、窒素、芳香環、縮合環などの追加記述が必要である。これらのオプションの部品化ルールから、少ない対応関係を元に部品を増やして、より多くの対応関係が得られるようになると考えている。

6) 今回の実験では、自動抽出したままのデータを利用し、不正解データへの対策を行っていない。よって、精度評価には至らなかった。

7. おわりに

特許文書をコーパスとして用いることで、有機化合物の化学物質名と化学式の対応関係を抽出できた。また、化学物質の命名規則を元に作成した部品化ルールによって、化学物質名と化学式の対応関係を増やすことができた。実証実験の結果、これらの手法を組み合わせることで、実験データの半分は正解できた。よって、化学物質名と化学式の対応関係の抽出には、本手法が有効だと考える。本手法は発展途上であり、正解率向上には、専門分野ごとのカスタマイズが必須と考える。また、Linked Open Data(LOD)の発展などが、化学情報(製造方法、化学反応、パラメータ、用途など)抽出のブレイクスルーを生み出す可能性になりそうだと考える。本稿では触れなかったが、高分子学会等で報告したように、この対応関係を可視化し、化学物質情報の理解を支援できるようにツール化した[6]。

文　献

[1] 国際純正および応用化学連合 International Union of Pure and Applied Chemistry (IUPAC) で制定した化合物の命名法規則
http://www.iupac.org/nc/home/publications/technical-reports/guidelines-for-drafting-reports/4-nomencl.html

[2] 日本化学会命名法専門委員会編,化合物命名法—IUPAC 勧告に準拠, 東京化学同人(2011)
[3] 藤井敦, 田中るみ子:特許検索における化学物質名の異表記同定に向けた考察, Japio YEAR BOOK 2010, pp.182-187(2010)
[4] 特定化学物質の環境への排出量の把握等及び管理の改善の促進に関する法律施行令, http://law.e-gov.go.jp/htmldata/H12/H12SE138.html
[5] 日化辞 Web, http://nikkajiweb.jst.go.jp/nikkaji_web
[6] 池田紀子, 田中一成:化学系特許の読解支援, 第 61 回高分子討論会, 3Pd042(2012)

一般発表

「特許情報分析のためのマイニング手法と分析ツール Patent Mining eXpress」

特許情報分析のためのマイニング手法と分析ツール Patent Mining eXpress

岩本 圭介
株式会社 NTT データ数理システム
Keisuke Iwamoto
NTT DATA Mathematical Systems, Inc.

概要

特許情報の分析においては、収集／整理した情報に対して集計を行い、集計結果をマッピングする手法が広く用いられている。特許情報をビジュアルな形にまとめることは直感的な理解を促すうえで非常に重要であるが、集計ベースのまとめ上げのみを用いている限り、人間が事前に決めた軸でしか物事を把握することはできない。対象分野のプロフェッショナルであればそこからでも充分な読み解きを行うことが可能であるであろうが、データマイニング・テキストマイニングの技術を適用することで、より広い層の利用者が未知な分野に対して、データに潜むまだ見ぬ法則性や気付きを得ること、またそれを人間の感覚に合致した形で可視化できるようになると考える。
実際には、種々のマイニング手法を目的に応じて使い分けていく。実用的には、年度や出願人等の属性情報を軸とした特徴の抽出と、人間の感覚にマッチした情報の可視化を実現することは非常に有用である。さらに、それらの目的は特許情報特有の事情である、適切な技術用語の抽出が実現できてはじめて成り立つものである。本稿では、それらの目的のために適用可能なマイニング手法を提案する。また、当社 NTT データ数理システムでは、このような一連の特許情報マイニングの手順をカバーするマイニングツール **Patent Mining eXpress** を開発し、製品化した。このツールの概要についても紹介する。

1. はじめに

データマイニング・テキストマイニングは、大量のデータから隠れた傾向や法則性を抽出するための技術であり、意思決定を支援する目的で行われる。
特許情報を分析し、その結果を自社の研究開発の方向性の決定に役立てようという試みは広く認知されているが[2]、必ずしも分析者が対象となる技術分野に精通しているかというとそうではない。そのような場合でも、マイニングによる特徴抽出、及び可視化によって、より人間の感覚に合致したアウトプットの提示が可能となる。
2 章で、そのための特徴抽出と可視化の手法について述べ、3 章でそれらの手法を用いて実現した特許分析ツール **Patent Mining eXpress** を解説する。最後に 4 章で今後の開発の方向性について示す。

*株式会社 NTT データ数理システム
〒160-0016 東京都新宿区信濃町 35 番地信濃町煉瓦館 1 階
E-Mail：iwamoto@msi.co.jp

2. 特許情報分析のためのマイニング手法

2.1. 特徴抽出

特許明細データの集合に対して、もしくは明細データの 1 件 1 件に対して特徴的な単語を抽出することは、理解の上の手掛かりを提供するものとして非常に有用である。

ここで注意すべきことは、ある単語が「特徴的」であるかどうかということは、比較対象があってはじめて論じることができる、という点である。比較対象を何にするか、データの母集団をどのように定義するかで抽出の結果は異なってくるため、どのような由来のデータを扱っているか、ということは分析の際に常に意識する必要がある。本節では「特徴的」であることを評価するための手法について述べる。

2.1.1. 属性を切り口として評価する手法

特許明細に含まれるデータのうち、要約・発明の課題や解決手段・請求項といった部分がテキスト情報として解釈できるものであるが、それ以外の出願日・出願人・発明者といった属性情報を切り口として特徴単語の抽出を行うことで、例えばある出願人 A 社に特有な技術はどのようなものか、もしくは昨年度特に出現しているキーワードは何か、といった観点で抽出を行うことができる。この場合、比較対象は何か、という観点では「A 社と A 社以外」「昨年度とそれ以前」とを比較することになる。

ここで、単語wと属性rを固定して、次の 4 種類の量を考える。

- 属性rのデータで
  単語wが出現した頻度：a
- 属性r以外のデータで
  単語wが出現した頻度：b
- 属性rのデータで
  w以外の単語が出現した頻度：c
- 属性r以外のデータで
  w以外の単語が出現した頻度：d

これは、単語と属性との頻度のクロス表として、表 1 のように表せる。

表 1 単語･属性頻度クロス表

	属性r	属性r以外
単語w	a	b
単語w以外	c	d

例えば、$a/(a+b)$という量を考えると、これは単語w以外の出現状況を全く考慮に入れていない指標ではあるが、単語wの総出現回数における同単語の属性rでの出現回数での割合であるため、その属性で出現しない単語では最大値1.0を示し単純に理解ができるという点で有用なものである。

更に、有意かどうかという明確な評価基準が求められる場合、有意性の検定が可能な次のような量を考えると良い。

- χ 二乗値
- Fisher の直接確率

表 1 のクロス表において、縦・横の合計を記したものが表 2 である。ただし総単語数を$N = a + b + c + d$とした。合計の個数（周辺度数）を固定した場合、もし「単語wを含むか否か」という事象と「属性rであるか否か」という 2 つの事象が独立であり関連性が無いのであれば、表の ”?” の部分には$a + c$個のうち$(a + b)/N$ 程度の割合が該当するはずであり、$\frac{(a+c)(a+b)}{N}$がその期待値であるといえる。しかし、これら 2 つの事象に関連性があれば、実際の値aはこの期待値から外れた値になることが予測できる。したがって、期待値に対するずれの割合$\left\{\frac{(\text{期待値}-\text{実測値})^2}{\text{期待値}}\right\}$の総和を評価することで、単語$w$と属性$r$との関連性を評価することができる。

表 2 合計頻度を固定する

	属性r	属性r以外	合計
単語w	?		$a+b$
単語w以外			$c+d$
合計	$a+c$	$b+d$	N

これがχ二乗値であり、次のように書くことができる[4]。

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

このχ^2 は自由度 1 のχ^2分布にしたがい、この分布に対する有意水準を与えることで「単語wと属性rとが独立である」という仮説を棄却するか否かを決定することができる。有意水準5%の棄却域が$\chi^2 \geq 3.841$、有意水準 1%の棄却域が$\chi^2 \geq 6.635$であるため、これらの場合は仮説を棄却し「独立ではない」、すなわち単語wと属性rの出現に関連があるとみなす。ただし、特徴的に「多く」出現しているような場合にのみ意味があると考えられるので、$ad-bc>0$であるような単語のみを採用する。

また Fisher の直接確率[1]は、表 2 のように合計頻度を固定した上で、その合計頻度と矛盾しないような単語・属性頻度クロス表の数値を決めた場合、観測された状態以上に偏りが発生するケースがどの程度の確率で生じるかという値を評価するものであり、この値が有意水準P以下であればそのP値において「単語wと属性rとが独立である」という仮説を棄却できる。

表 3 合計頻度と矛盾しない頻度

	属性r	属性r以外	合計
単語w	$a+n$	$b-n$	$a+b$
単語w以外	$c-n$	$d+n$	$c+d$
合計	$a+c$	$b+d$	N

実際に表 2 と矛盾しないよう、また表 2 の状態よりも偏るよう（単語w・属性rのセルの頻度がaよりも高くなるよう）頻度の値を決めた場合、表 3 のようになり、これは値nを決めると一つの決め方が定まる。クロス表の各セルの値は0以上でなければならないので、nは$0 \sim \min(b,c)$の値を取りうる。nを固定したとき、その状態が実現する確率は次の$P(n)$で表せる。

$$P(n) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{N!\,(a+n)!\,(b-n)!\,(c-n)!\,(d+n)!}$$

この$P(n)$を$n = 0 \sim \min(b,c)$にわたって和を取り、その値を確率値として有意水準と比較する。値が大きければ「今、このデータに現れている程度の偏りは稀なことではない」と考えられ、逆に小さければ「この程度偏ることは稀なケース」であるといえる。

2.1.2. 分布の偏りを評価する手法

属性値のような切り口が与えられているのではなく、特許明細 1 件 1 件に対して、特許 1 件毎の特徴語を知りたいといった場面もある。そのような場合でも、特許 1 件 1 件の違いを属性とみて前節の手法を適用することは可能である。実用的には出願番号等の 1 件毎にユニークな情報を属性として用いればよい。

また、

- 1 件の特許の中で
  多く語られている単語は重要
- 他の特許に
  あまり出現していない単語は重要

と考え、特許jにおける単語iの出現回数$tf_{i,j}$と単語iが出現するデータ数の逆数（さらに対数を取る場合が多い）であるidf_iの積で評価を行う tf-idf 法も広く用いられている。また、次式で示される指標 BM25 も知られている。

$$BM25_{i,j} = idf_i \frac{tf_{i,j}\cdot(k+1)}{tf_{i,j} + k\left(1-b-b\dfrac{|D|}{avgdl}\right)}$$

ただし、b及びkはパラメタ、$|D|$は特許明細jの単語数、avgdl は明細の平均文書長である。b=0 及び$k=0$とした場合は tf-idf と同一であるが、$|D|$

の項の効果により文書長の影響を考慮した指標になっている。

2.2. 可視化

前節の方法で例えば出願人毎、もしくは特許明細 1 件毎の特徴単語が明らかになったとして、それをどのような形で示して理解に繋げるか、という点が次の課題として存在する。

例えば出願人であるメーカー毎の特徴語が表 4 のように得られたとする。「即席麺」「インスタント麺」という分野における特許明細データから抽出を行ったものであるが、表 4 の内容を実際に確認すると、A 社・B 社は「風味向上方法」など共通で出現している課題もあって麺自体の製法がターゲットになっているのに対して、C 社は容器の製法といった面で即席麺・インスタント麺に関わっているということが分かる。ただし、そのこと自体は表 4 の内容を読んで理解し解釈しなければ見えてこないものであり、表 4 がそのまま理解の上での助けになっているとは言い難い。

表 4 表形式によるリストアップ

A 社	B 社	C 社
減塩	水分含量	紙製シート
横断面	**顆粒状**	低コスト
顆粒状	乾燥処理	どんぶり形状
風味向上方法	湯切り孔	外周面
ギ酸	**風味向上方法**	成形芯

ここで強調したいものは物事同士の「関連」、また近さ・遠さが定義できるものについてはそれらの間の「距離関係」である。表 4 も出現人と抽出された技術用語との間の「関連」を示したものである。このような表形式、もしくは比較的単純な棒グラフ・折れ線グラフ・円グラフといった表現手段に加えて、人間の間隔に合致したアウトプットを得るための分析方法について述べる。

2.2.1. 繋がりを可視化するための手法

物事同士に関連がある、というその繋がりを強調して図示を行うのがネットワークグラフである。ネットワークグラフは「実体」に対応するノードと、その「実体」間に関連があることを表すリンクとで構成される。

ノードが表す「実体」として、単語や出願人、発明者等の要素を考えることができる。単語と出願人を「実体」ととらえ、表 4 の状況をネットワークグラフにしたものが図 1 である。単純なものではあるが、A・B 社と C 社が異なる「固まり」になっていること、また A 社と B 社は「風味向上方法」「顆粒状」といった単語を共有していることが容易に理解できる。

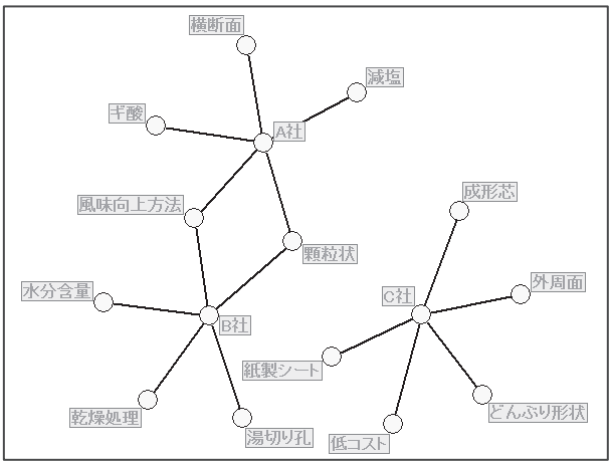


図 1 表 4 のネットワーク図表現

何をもって関連があるとみなすか、その指標としては、属性-単語間の関係では前節で挙げたような関連性の指標を用いることができるが、属性同士・単語同士については 1 件のデータにおけるそれらの同時発生の状況（共起）をもとに図示を行うことがよく行われている。出願人同士の同時発生を確認することは特許の共同出願状況を解明することになる。単語同士の共起状況を図示することは、同時に語られやすい単語のグループを明らかにし、テキスト内にどんな話題が出現しているかという「かたまり」を浮かび上がらせることができる。

ある事象 X が発生したデータの件数が$n(X)$である場合、2 つの事象 A,B の共起の指標としては、次のものがよく用いられる[3]。

- 信頼度＝$\frac{n(A\cap B)}{n(A)}$
  A が発生しているときに、どのくらいの割合（確率）で一緒に B が発生しているか
- サポート＝$\frac{n(A\cap B)}{N}$，N は全データの件数
  全体の中で、A,B が同時に起こるのはどの程度の割合か

2.2.2. 距離関係を可視化するための手法

特許明細データの全体像はどのようになっていて、その中にどのようなまとまりが存在しているか、という点を可視化するために、1 件 1 件の特許明細を 1 データ点と考え、それらの間の位置関係を適切に定義して「地図」のようなアウトプットを作成することも直感的な把握を促すうえで非常に意味がある。

マップ上の位置関係を決めるうえで、特許明細データの間の近さ・遠さを定義する必要がある。そのために、通常はそれぞれの特許明細に対して単語を各成分に取り、単語出現の状況を数値に落とし込んだベクトル表現を作成する（図 2）。

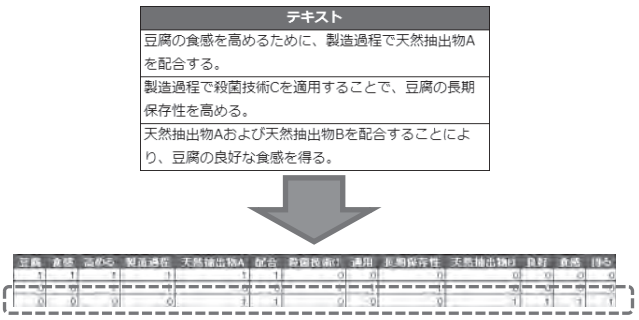


図 2 テキストのベクトル表現

ベクトルの各成分として用いる単語は、特許の特徴をうまく反映するものを使用するのが望ましく、2.1.2 章で解説した指標などを用いて選択を行う。

このベクトル間の距離により、特許明細間の距離を定義することができる。このベクトル表現は極めて多次元の情報であるが、人間が知覚可能な形でこの状況を可視化するために、この多次元空間内での特徴、すなわち距離関係をできるだけ保ったまま 2 次元、もしくは 3 次元の点に落とし込む必要がある。

多次元尺度構成法は、そのための代表的な手法である。多次元空間内での特許明細iと特許明細jとの間の距離の 2 乗d_{ij}を要素とする$t\times t$行列Dを与え、次の手続きで計算を行う。

1. Dに Young-Householder 変換を行った行列Pを求める。
   $$P=-\frac{1}{2}HDH, H=I-\frac{1}{t}ee^T$$
   （ただし、eは全ての要素が 1 の列ベクトル）
2. Pを特異値分解して、直交行列Vと対角行列Λを用いて次のように表す。
   $$P=V\Lambda V^T$$
3. 対角行列Λの固有値の大きい方からd個取出し、それ以外を 0 とした行列をΛ_dとし、取り出した固有値に対応する固有ベクトルを並べた行列をV_dする。
   $Y={\Lambda_d}^{1/2}{V_d}^T$とおくと、$P_d=V_d\Lambda_d{V_d}^T=Y^TY$は、$P$を$\mathrm{tr}((P-Y^TY)^2)$の意味で最小化したものになっている。
   このYをd次元の座標としてプロットする。

3. Patent Mining eXpress

NTT データ数理システム による **Patent Mining eXpress** は、WEB ブラウザ上の簡便な操作で書誌情報・テキスト情報の分析・可視化が可能な特許情報分析ツールである。**Patent Mining eXpress** 全体の画面を図 3 に示す。分析で用いている手法・アルゴリズムは前章にて解説したものを搭載している。

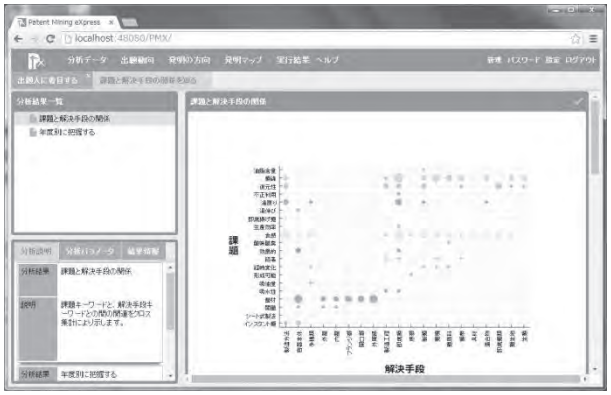


図 3 Patent Mining eXpress 画面

画面上部のメニューにより、図 4 に示す分析メニューを実行することができる。

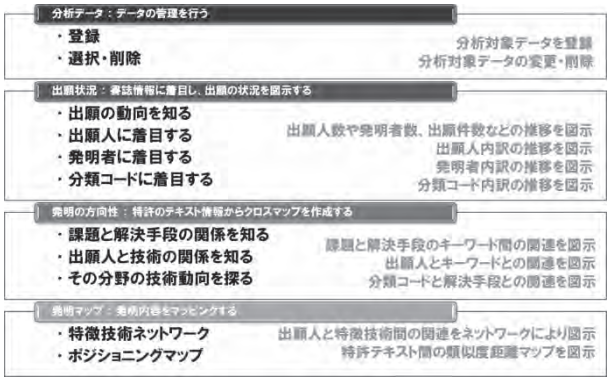


図 4 Patent Mining eXpress メニュー一覧

2.1.1 章で示した属性を切り口とした特徴抽出の結果は、次の図 5 のようなネットワーク図で表現することが可能である。

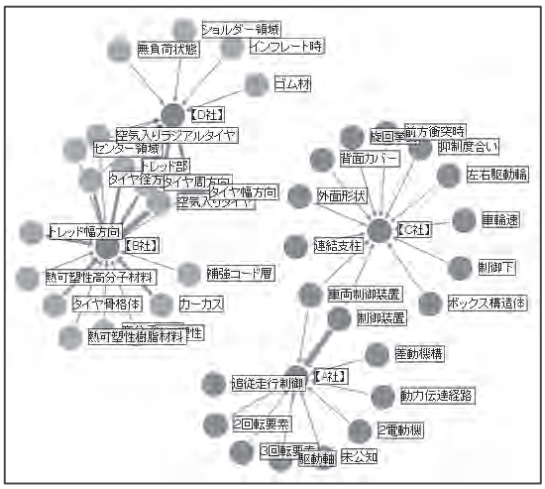


図 5 特徴技術ネットワーク

また、2.1.2 章で示した評価指標によって、単語の重みづけを行い、どのような「課題」に対してどのような「解決手段」が出現しているか、その関係を図 6 のようにクロスマップで示すことができる。

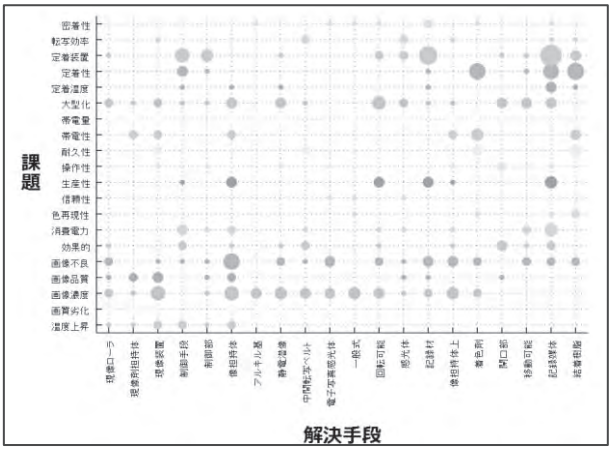


図 6 課題と解決手段の関係を知る

2.2.2 章で示した特許明細間の距離関係は、2 次元平面上の図示として、図 7 のような形で表すことが可能である。
それぞれの点が 1 件の明細に対応しており、マウス操作でそれぞれの特許の特徴単語と原文テキストを参照することが可能である。

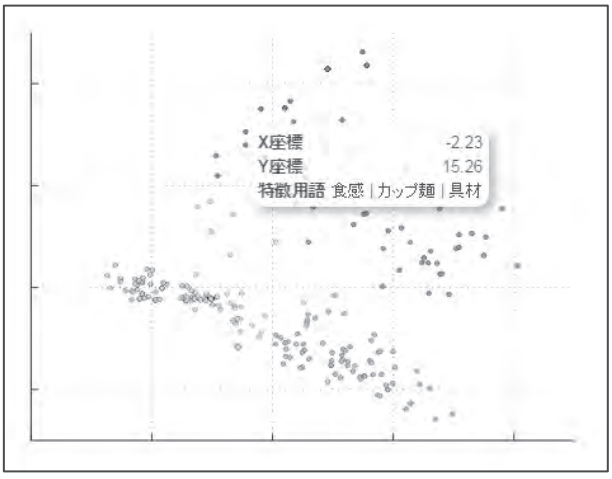


図 7 ポジショニングマップ

4. おわりに

本稿では、特許情報からの特徴抽出と可視化を行う上で有用と考えられるマイニング手法を解説し、それらに基づいて開発を行った特許分析ツール **Patent Mining eXpress** について紹介を行った。

Patent Mining eXpress は、複雑な設定項目を排し操作の簡便性を最重要視して開発を行ったため、今後はその簡便さを確保したままカスタマイズ性を向上させることが課題である。機能と使い勝手の両立を図り、よりよいツールを提供していきたい。

参考文献

[1] 豊田裕貴 菰田文男 編著(2011)『特許情報のテキストマイニング』ミネルヴァ書房

[2] 服部兼敏(2010)『テキストマイニングで広がる看護の世界』ナカニシヤ出版

[3] NTT データ数理システム(2014)『Visual Mining Studio 8.0 技術資料』

[4] NTT データ数理システム(2014)『Text Mining Studio 5.0 技術資料』

平成26年度AAMT/Japio特許翻訳研究会
第3回特許情報シンポジウム 資料集

発 行 日　平成26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