

平成 25 年度 AAMT/Japio 特許翻訳研究会

報 告 書

機械翻訳及び機械翻訳評価に関する研究

及び

ワークショップ報告

平成 26 年 3 月

一般財団法人 日本特許情報機構

目 次

1. はじめに	1
辻井 潤一 マイクロソフト研究所アジア 首席研究員/ 東京大学大学院情報理工学系研究科 名誉教授	
2. 特許文の言語的特徴	
2.1 特許文における入れ子構造の調査	2
高橋 尚矢 山形大学 横山 晶一 山形大学	
2.2 中日平行コーパスおよび Moses を用いた中国語文パターンの半自動抽出	7
江原 暉将 山梨英和大学	
3. コーパスからの言語知識獲得	
3.1 パテントファミリーにおける対訳文対非抽出部分を利用した専門用語訳語推定方式の評価と分析	15
豊田 樹生 筑波大学 龍 梓 筑波大学	
董 麗娟 筑波大学 宇津呂武仁 筑波大学	
山本 幹雄 筑波大学	
3.2 単語間結合度に基づく複単語表現のアライメントの改善	32
塩田 嶺明 京都大学 中澤 敏明 京都大学	
黒橋 禎夫 京都大学	
3.3 テキスト中の専門用語から Wikipedia への対応付けにむけて	38
綱川 隆司 静岡大学 梶 博行 静岡大学	
4. コンパラブルコーパスによるドメイン適応	43
範 暁蓉 東京大学 二宮 崇 愛媛大学	
5. 英文作成システムのレビューとシステム構想の提案-----MT への応用を目指して	51
宮澤信一郎 秀明大学 江原 暉将 山梨英和大学	
岡田 勇 創価大学 宮崎 瑞之 秀明大学	
Dawn L. Miyazaki 早稲田大学	
6. 機械翻訳評価	
6.1 拡大評価部会の活動概要	61
江原 暉将 山梨英和大学	
6.2 NTCIR PATMT データと WMT METRICS TASK データにおける英文データを対象とした自動評価法のメタ評価	62
越前谷 博 北海学園大学 須藤 克仁 NTT コミュニケーション科学基礎研究所	
磯崎 秀樹 岡山県立大学 江原 暉将 山梨英和大学	
6.3 NTCIR-10 特許審査評価の結果	71
後藤 功雄 NHK 放送技術研究所	
6.4 中日特許文評価用テストセットの作成	78
長瀬 友樹 (株)富士通研究所 王 向莉 (財)日本特許情報機構	
江原 暉将 山梨英和大学	
7. 第 14 回機械翻訳サミット(Machine Translation Summit XIV)および第 5 回特許翻訳ワークショップ	83
(The 5th Workshop on Patent Translation) 参加報告	
横山 晶一 山形大学 梶 博行 静岡大学	
二宮 崇 愛媛大学	

AAMT/Japio 特許翻訳研究会委員名簿

(敬称略・順不同)

委員長	辻井 潤一(※2)	マイクロソフト研究所アジア 首席研究員/ 東京大学大学院情報理工学系研究科 名誉教授
副委員長	横山 晶一(※2)	山形大学大学院教授
	梶 博行	静岡大学大学院教授
委員	江原 暉将(※1)	山梨英和大学教授
	宮澤 信一郎	秀明大学教授
	黒橋 禎夫	京都大学大学院教授
	宇津呂 武仁(※2)	筑波大学教授
	二宮 崇	愛媛大学大学院准教授
	越前谷 博(※2)	北海学園大学大学院教授
	網川 隆司	静岡大学大学院助教
	潮田 明(※2)	元奈良先端科学技術大学院大学客員准教授
	後藤 功雄(※2)	NHK 放送技術研究所
	熊野 明	東芝ソリューション(株)
	下畑 さより	沖電気工業(株)
	三浦 貢	日本電気(株)
	須藤 克仁(※2)	NTT コミュニケーション科学基礎研究所
	隅田 英一郎	(独)情報通信研究機構
オブザーバー	中川 裕志	東京大学教授
	範 暁蓉	東京大学大学院 中川研究室
	安藤 進	元多摩美術大学講師
	呉 先超	バイドゥ株式会社
	磯崎 秀樹(※2)	岡山県立大学
	鈴木 博和(※2)	(株)東芝
	長瀬 友樹(※2)	(株)富士通研究所
	守屋 敏道	(財)日本特許情報機構
	松田 成正	(財)日本特許情報機構
	大塩 只明	(財)日本特許情報機構
	埴 金治	(財)日本特許情報機構
	三橋 朋晴	(財)日本特許情報機構
	柿田 剛史	(財)日本特許情報機構
	土屋 雅史	(財)日本特許情報機構
	星山 直人	(財)日本特許情報機構
	王 向莉	(財)日本特許情報機構
		(※1：拡大評価部会部会長、※2：拡大評価部会メンバー)
事務局	浅利 茂	(株)インターグループ
	平林 剛	(株)インターグループ
	川井 佐知子	(株)インターグループ

平成 25 年度 AAMT/Japio 特許翻訳研究会・活動履歴

平成 25(2013)年 5 月 10 日

第 1 回 AAMT/Japio 特許翻訳研究会・拡大評価部会
(於キャンパス・イノベーションセンター東京)

平成 25(2013)年 6 月 7 日

第 2 回 AAMT/Japio 特許翻訳研究会 (於キャンパス・イノベーションセンター東京)

平成 25(2013)年 7 月 26 日

第 3 回 AAMT/Japio 特許翻訳研究会 (於キャンパス・イノベーションセンター東京)

平成 25(2013)年 9 月 2 日~9 月 6 日

第 14 回機械翻訳サミット (於フランス (ニース) Acropolis Conference Centre)

平成 25(2013)年 10 月 11 日

第 4 回 AAMT/Japio 特許翻訳研究会・拡大評価部会
(於キャンパス・イノベーションセンター東京)

平成 25(2013)年 11 月 29 日

第 5 回 AAMT/Japio 特許翻訳研究会 (於キャンパス・イノベーションセンター東京)

平成 26(2014)年 1 月 24 日

第 6 回 AAMT/Japio 特許翻訳研究会・拡大評価部会
(於キャンパス・イノベーションセンター東京)

平成 26(2014)年 3 月 7 日

第 7 回 AAMT/Japio 特許翻訳研究会 (於キャンパス・イノベーションセンター東京)

平成 26(2014)年 3 月 31 日

『平成 25 年度 AAMT/Japio 特許翻訳研究会報告書 機械翻訳及び機械翻訳評価に関する研究及び
ワークショップ報告』完成

1. はじめに

マイクロソフト研究所アジア 首席研究員
東京大学大学院情報理工学系研究科 名誉教授
AAMT/Japio 特許翻訳研究会委員長

辻井 潤一

機械翻訳は、Web 翻訳の普及に伴い、特に特別な技術ではなくなった。これは、機械翻訳が社会に受け入れられ、一般に使われる技術になったということで、機械翻訳にとっては望ましいことである。ただ、機械翻訳がすでに完成した技術のように捉えられ、それに対する研究開発の努力が低減するとしたら、この発展途上にある技術にとって望ましいことではない。

実際、Web 翻訳を使ったことがある人なら、現在の機械翻訳技術は完成とは程遠いものであることを認識しておられることと思う。英語とヨーロッパ語族の間の翻訳のように、語族が近い言語間の翻訳の精度は、実用に耐えるレベルとなっているが、日本語と中国語、日本語と英語のように語族の異なる翻訳となると、理解できない翻訳が出力されることのほうが多い。比較的短い文ではうまく翻訳できていても、一般のテキストに現れる長い文になると、質の低下は顕著になる。

特許のような科学技術に関わる文書の翻訳は、当該の科学技術分野の知識を持たない場合には、人間の翻訳家にとっても難しいものとなる。機械にとって、その翻訳が難しいのは当然のことであろう。

本研究会は、この特許の翻訳に機械翻訳の技術を適用する際の問題点を把握し、その解決のための技術の方向を見極めるための活動を行っている。学会などの研究会とは異なり、特許翻訳の実務に携わる Japio（日本特許情報機構）の支援と参加を得て、実務に従事する立場からの経験を積極的に議論の対象とすることで、企業の機械翻訳開発や大学の研究をより現実的なものとしていく役割を担ってきた。

このような研究会の意義は年を追うごとに広く認識されるようになり、我々の活動がもととなって、中国や韓国、ヨーロッパ、北米でも同様な活動が活発化してきている。特に、本年度は、本研究会が中心となって、機械翻訳に関する最大の国際会議であるMTサミット（フランス、ニースで開催）にて、特許の機械翻訳に関するワークショップを開催するなど、国際的なリーダーシップをとることができた。

本報告書は、2013 年度に行った我々の活動を総括し、次年度以降の活動のための礎となることを願って、研究会参加メンバーがまとめたものである。本報告書が、知財の国際化に伴い、ますますその重要性を増している特許の多言語翻訳システムの開発、運用、利用に興味を持つ人たちの交流をさらに強めることに貢献できることを願っている。

2. 1 特許文における入れ子構造の調査

山形大学 高橋 尚矢
横山 晶一

1. はじめに

近年、国際的な特許の共有化に伴い国際特許の申請数も増加し続けている。特許文の検索や翻訳などの作業には多くの人手が必要であり、そのため作業を自動化または半自動化することが求められている。これらを解決するためには特許文中に含まれる情報を的確に抽出することが要求される。したがって、特許文に対する正確な係り受け解析が不可欠である。

特許文の課題や解決手段の部分は、200文字を超える長大な一文になることが多い。しかも単語同士の係り受けが複雑であり、意味が明確でないことがある。これは特許文独特の記述や専門用語の多さなどが原因である。そのため通常文に比べると係り受けが曖昧になりやすく、解析の誤りが発生しやすいという特徴がある。

本稿では、係り受けの中でも入れ子構造(nested structure)を取るものに着目した。係り受けの複雑化の要因の一つである入れ子構造を正しく解析できれば、機械翻訳する際に長文を一気に翻訳するのではなく小さなまとまりごとに翻訳でき、精度が高くなると考えた。

今回は“主辞(head word)+英数字”という構成の語に着目し入れ子構造の検出を目指す。理由として並列構造をつくる要素には“装置1”のような構成を持つ語が頻出するからである。またこのような語は係り受け先の特が難しいため長文では機械翻訳しづらい。よって小構造ごとに機械翻訳するという前提のもとで入れ子構造の検出をおこなう方針である。

本稿は、主として[1]に基づき、その内容に加筆したものである。

2. 関連研究

松山ら[2]は特許文ではなく法令文書を対象にした並列構造解析をおこなっている。あらかじめ法令文書の重要語を決め、それらをもとに並列構造を検出した。具体的には並列構造を pf1(前方句),key(並列となるキーワード),pb(後方句)と定義し、この型に当てはまる部分を検出するというものである。しかし階層的並列構造は検出が難しい、前方句と後方句の長さが違いすぎると並列構造の検出に失敗しがちである、など改良の余地がある。また法令文書では並列を意味する語が細かい規則で設定されている [3]のに対し、特許文ではそのような規則が無いいため並列階層の特定が困難である。

我々は、昨年は特許文の並列構造解析について主辞(昨年度は“接尾辞”と表記したが主辞が正しい)に着目した手法[4]で解析を試みた。特許特有の主辞を抜き出しそれをもとに係り受け修正システムを作成し一定の結果が得られた。ただし主辞を限定したことによ

り全体の誤り修正の精度はさほど上がらなかった。新たに発見した問題点としては入れ子構造が正しく解析されなかった点があげられる。今回はその入れ子構造問題に取り組む。

3. 入れ子構造

図1に示したのは公開番号特開 2003-176610 の文章から【解決手段】部分を一部抜粋したものである。今回の実験では下線部に注目した。このような入れ子構造を持つ特許文を KNP[5]で解析したものを図2に示す。この解析結果では独自性の強い複合語をうまく扱えていないほか、中央の四角で囲った箇所（「支持脚4と、支持脚4の」の部分）が並列構造を間違えて捉えているという問題がある。本来であれば「拘束部5 “とを” 有して」とあるように2つ以上の事柄を並列させていなければならないが、この解析結果では拘束部5のみをピックアップする形である。支持脚4と拘束部5の並列関係が成り立たない出力となっているため不適切な出力となる。

もしこれが簡潔な文ならば図3のように正しく解析できることは確認済みである。図3の実験では「一端非拘束状態で手摺体2に折り畳み可能に連結され、床面3上での起立姿勢において手摺体2をほぼ水平な使用姿勢に保持可能な支持脚4」を『要素A』に、「支持脚4の折り畳み動作を規制し、該支持脚4の起立姿勢を維持する拘束部5」を『要素B』に置き換えて構文解析した結果である。

上記の結果を踏まえ、長文内の入れ子構造を“主辞+英数字”をキーワードとして検出を試みた。対象は32特許文の【課題】【解決手段】【請求項】のすべての項目とした。しかし“主辞+英数字”の構造を持つ語同士の並列関係はあまり得られず、且つ KNP で解析が失敗した例は図2の例を含め3件しか得られなかった。

考察として、“主辞+英数字”の構造を持つ語同士の並列関係は短い文の場合は正しく解析できた。今回失敗した3件の例はいずれも長文中の並列構造だったため、解析の正誤に影響を与えるのは“主辞+英数字”の構造を持つ語群よりも文章の長さによるものが大きいと判断した。

【解決手段】縦壁からなる取付基部1に垂直回転自在に連結され、前傾姿勢で床面3上に収容可能な手摺体2と、一端非拘束状態で手摺体2に折り畳み可能に連結され、床面3上での起立姿勢において手摺体2をほぼ水平な使用姿勢に保持可能な支持脚4と、支持脚4の折り畳み動作を規制し、該支持脚4の起立姿勢を維持する拘束部5とを有して構成する。また、収納姿勢において、支持脚4の非拘束端4aが手摺体2先端より後方に位置するように構成する。

図1 実験で使用した特許文

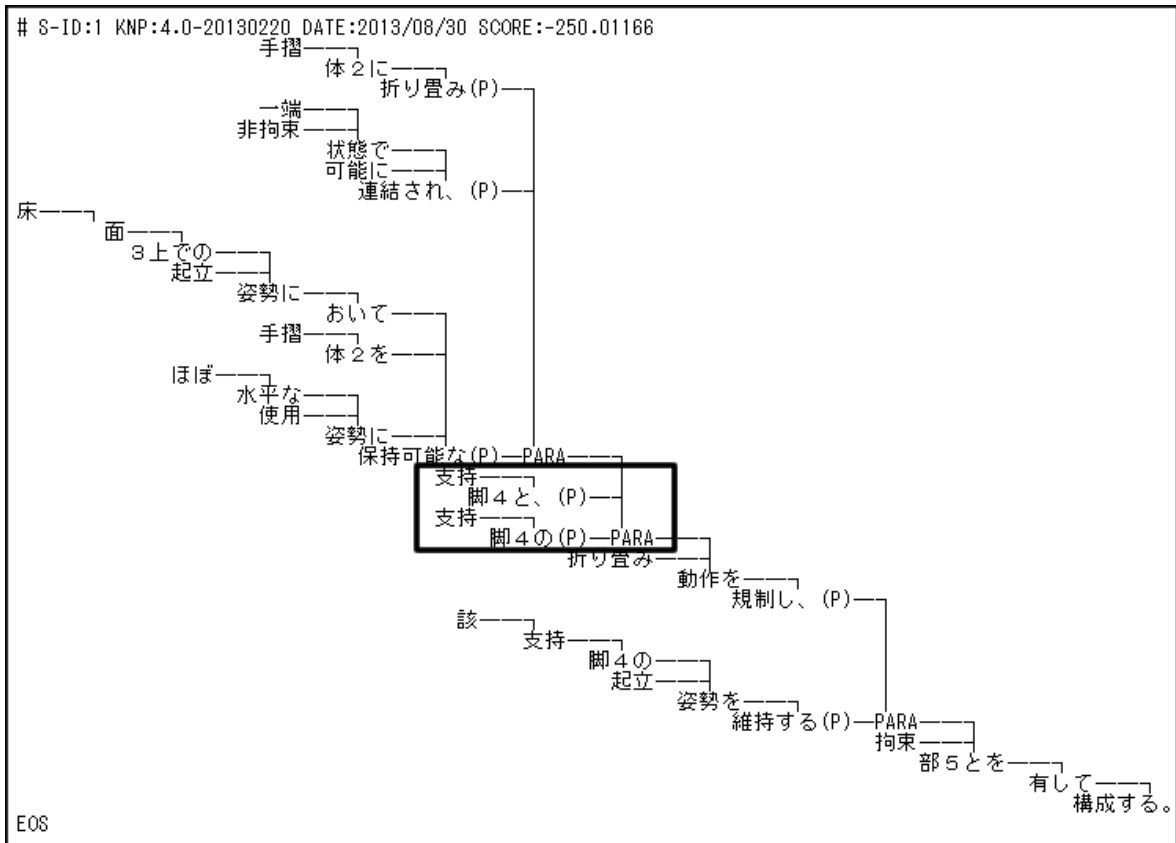


図2 KNPによる特許文の構文解析結果

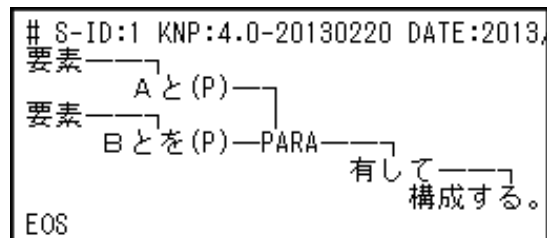


図3 単純化した文の構文解析結果

4. 今後の方針

今回の調査で、構造解析システムは特許文のような長文の解析が不完全であることが改めてわかった。特に並列構造を正しく捉えることがシステムでは難しい。複合語など未定義語の問題は辞書の拡充などで対応できるかもしれないが、並列構造の場合においては長文だからこそ起こりうる失敗例だといえる。ただし最小の並列構造ごとに分解すれば構造解析システムでも正しく解析をおこなえる可能性を示せた。

また“主辞+英数字”の構造を持つキーワードによる並列構造を KNP で解析した際に、失敗例があまり得られないことがわかった。並列構造の解析ミスが少ないというわけではないが、解析が上手くいかない原因が“主辞+英数字”の語群による影響が大きいと言にくい。

今後の方針として二格に着目した並列構造の解析をおこなう予定である。図4、図5は四角で囲んだ部分が並列構造と判定されているが、いずれも並列構造はとらない。図4は、『培土板の上部に、培土板の前方に向けてほぼ水平に延び、下方からの迫り上げ土圧により後部の取付部を支点として弾性をもって変位する可動邪魔板を設ける。』

のような『AにBを設ける』という文章で、図5は、

『種子落下筒21の後方側に、機体の進行方向に沿い所定間隔を開けて対向する左右のガイド板30、31を有し、その間に機体の進行方向に板面を向けたゴム板製の遮蔽板32が配置された覆土装置28を設置する。』

のような『AにBを設置する』という文章が正しい解析である。

参考文献

- [1]高橋尚矢、横山晶一：特許文における入れ子構造の調査、Japio YEAR BOOK (2013) pp.266-269
- [2]松山宏樹、白井清昭、島津明：法令文書を対象にした並列構造解析、言語処理学会 第18回年次大会(2012)
- [3]岩本秀明、長野馨、永井秀利、中村貞吾、野村浩郷：法律文における並列構造の特徴とそれに基づく制限言語モデルについて、情報処理学会自然言語処理研究会(1993)
- [4]横山晶一：接尾辞に着目した特許文の並列構造解析、Japio Yearbook(2012) pp.250-253
- [5]日本語構文・格解析システム KNP：<http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP>

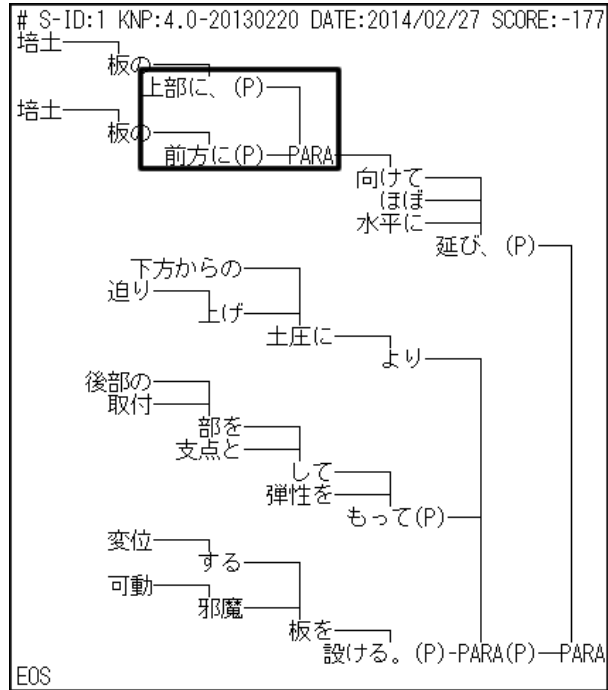


図4 ニ格を並列構造と誤解析した例1

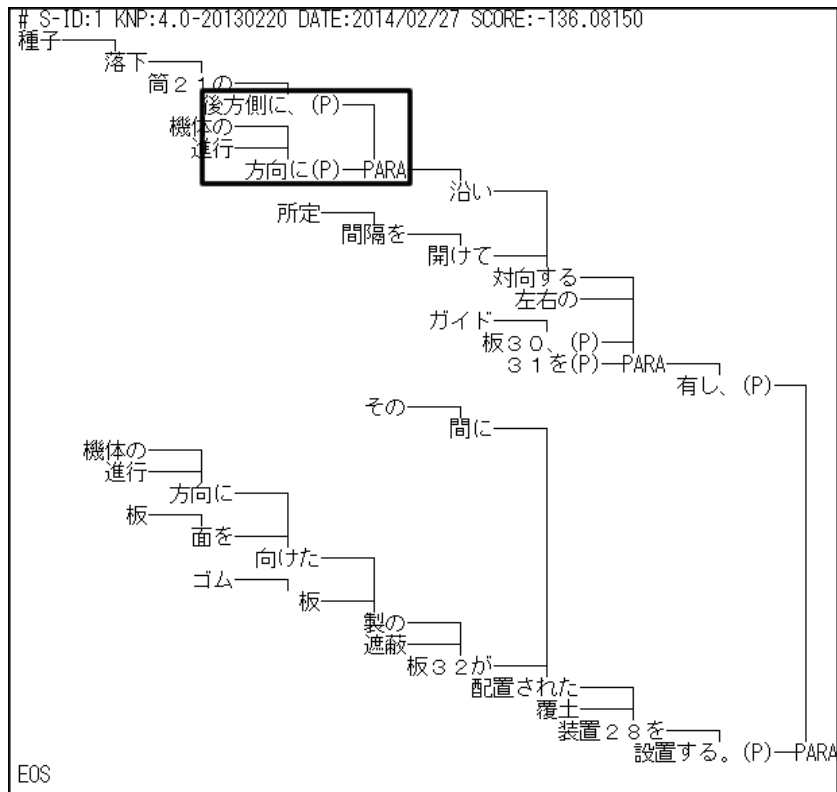


図5 ニ格を並列構造と誤解析した例2

2. 2 中日平行コーパスおよび Moses を用いた

中国語文パターンの半自動抽出

山梨英和大学 江原暉将

2.2.1 はじめに

機械翻訳において正確な翻訳を行うためには、適切な訳語を出力することと入力言語の構文構造を出力言語の構文構造に適切に変換することが必要である。前者は語彙的曖昧性解消、後者は構造的曖昧性解消と呼ばれ、機械翻訳の 2 大課題とされている[1]。特許文書では、専門用語が多く使われ、かつ文長が長く複雑な構文構造を持つという特徴がある。これらの特徴は、語彙的曖昧性解消と構造的曖昧性解消を共に困難にさせ、機械翻訳の精度を低くする要因となっている。

複雑な構文構造を正確に捉え、出力言語の構文構造に適切に変換する方法の一つとしての文パターンの利用がある。特許文は構文が複雑であるが、類似した構文が繰り返して出現するという特徴がある。このような場合、頻出する文パターンを用意しておき、翻訳に利用することで構造的曖昧性解消を適切に行うことができる。文パターンを利用するというアイデアは古くからあり[2, 3]、近年、特許文書に適用したものとして[4, 5, 6, 7, 8]などがある。

文パターンの利用でまず問題になるのが、頻出する文パターンをどのようにして収集するのかということである。人手で収集する方法はコストがかかるという難点がある。一方、構文解析器を利用して自動または半自動で収集する方法が考えられるが、そもそも構文解析の精度が低いから文パターンを利用するのであり、この方法では本末転倒になってしまう。

日本語の文パターンを抽出する一方法として、文末表現に着目する方法がある[9]。日本語は典型的な主要素後置型言語であるため、文末に着目することで文パターンが抽出できる。例えば「本発明は魚釣用スピニングリールに関し、釣糸の巻取り操作時に於ける糸燃れの蓄積を防止し、併せて釣糸のダメージの軽減を図った魚釣用スピニングリールを提供することを目的とする。」という文の文末に着目することにより「・・・を提供することを目的とする。」という文パターンが得られる。この方法は日本語では有効であるが、主要素後置型言語ではない英語や中国語に適用することはできない。

[10] では、中国語の頻出する文パターンを抽出する一方法として、2 段階の方法を提案している。まず、日本語と中国語の文対応コーパスを用意し、その日本語部分の文パターンを文末表現を手掛かりにして求める。次に、同一の文末表現を持つ日本語文に対応する中国語文に共通する表現に着目して中国語の文パターンを求める。しかし[10]の方法では、日本語文パターンは自動的に求まるものの、対応する中国語文パターンは手動で求めている。本報告では、統計的機械翻訳の中で用いられている単語対応機能を用いて、対応する中国語の文パターンを半自動的に求める手法を提案する。

2.2.2 中国語文パターンの抽出

以下のステップで中国語頻出文パターンを抽出する。

- a) 日中文対応コーパスを作成する。
 - b) 上記コーパスの日本語部分から文末表現に着目して日本語文パターンを抽出する。
 - c) 統計的機械翻訳ツールの Moses[11]を用いて、日本語の文末表現に対応する中国語表現を求める。
 - d) 得られた中国語表現をキーとして中国語文パターンを作成する。
- これらを順に説明する。

2.2.2.1 日中文対応コーパスの作成

日中のパテントファミリーを用いて、日中文対応コーパスを作成する。本文で用いる元データはパテントファミリーの要約部分であり、138,065 件の日中文書対応からなる¹。まず、この文書対応コーパスから文対応コーパスを作成しなければならない。その方法は、各文書の中国語部分を市販の機械翻訳システムを用いて、日本語に翻訳し、この機械翻訳結果と対応する日本語部分とを比較して対応スコアの高い文対を求めるものである。ここで、日本語文 A と日本語文 B の対応スコア $s(A, B)$ を以下のようにして計算する。文 A に含まれるキーワードの集合を K_A とし、文 B に含まれるキーワードの集合を K_B とする。また、一般的に集合 S の要素数を $\#(S)$ とすると、

$$s(A, B) = \frac{2 \times \#(K_A \cap K_B)}{\#(K_A) + \#(K_B)}$$

で対応スコアが計算される。 $0 \leq s(A, B) \leq 1$ が成立する。 K_A と K_B が完全に一致すれば $s(A, B)$ は 1 となり、まったく異なれば 0 となる。ここで、キーワードとしては、漢字またはカタカナを含む単語を用いた。また、単語分割には ChaSen を用いた。対応スコアを求める機械翻訳文の元となった中国語文と対応する日本語文を対にすることで文対応コーパスが得られる。この文対応付けの具体例を付録 1 に示す。

文対応コーパスにおいて対応スコアが 0.5 以上の対応のみを求めて最終的な文対応コーパスとした。得られた文対は 43,404 であった。

2.2.2.2 日本語文パターンの抽出

得られた文対応コーパスから日本語部分を抽出し、文末表現を用いて日本語文パターンを求めた。その結果、付録 2 に示すようなデータが得られた。元データが要約部分であるため「提供する。」や「特徴とする。」およびそれらの類似パターンが上位にきている。付録 2 に示す 56 パターンだけで 43,404 件のうち 5,773 件(約 13%)をカバーしている。

¹ 本コーパスは、日本語部分が 453,616 文で構成され、中国語部分が 394,744 文で構成されている。

2.2.2.3 Moses の実行および日本語の文末表現に対応する中国語表現の抽出

2.2.2.1 で得られた 43,404 文対を用いてフレーズベース Moses を用いて学習処理を行った。その結果、Giza++での単語対応データが得られ、さらにフレーズテーブルが作成された。この単語対応データとフレーズテーブルから日本語の文末表現に対応する中国語表現を抽出する。ここでは、例として頻度の高い「提供することを目的とする。」という日本語表現を用いて手法を説明する。

フレーズテーブル(phrase-table.0-0)に含まれる日本語表現が「提供することを目的とする」に一致するデータは 123 データであった。Moses が出力する inverse phrase translation probability $\phi(f|e)$ と direct phrase translation probability $\phi(e|f)$ の積が大きい順に 20 データを表 2.2.1 に示す。適切な中国語表現が求まっているが、フレーズテーブルからの抽出であるので連続な表現に限られる。[10]では、「本发明以提供・・・为目的。」のように分離した表現があることが指摘されており、フレーズテーブルを用いた方法では、そのような不連続表現は抽出できない。

表 2.2.1 フレーズテーブルから抽出した「提供することを目的とする」に対応する中国語表現

中国語表現	$\phi(f e)*\phi(e f)$
的 目的 在于 提供 一种	0.01950640
的 目的 是 提供 一种	0.01484489
的 目的 在于 提供	0.01419362
的 目的 在于 提供 一种 即使	0.01019112
发明 的 目的 在于 提供 一种	0.00770700
的 目的 在于 提供 在	0.00636943
发明 的 目的 在于 提供	0.00530786
本 发明 的 目的 在于 提供	0.00489955
的 目的 旨在 提供	0.00424629
目的 在于 提供 一种 可以 抑制	0.00318471
本 发明 的 目的 在于 提供 显示	0.00318471
的 目的 是 要 提供 一种	0.00318471
的 目的 是 提供 无	0.00318471
的 目的 是 提供 无 需	0.00318471
的 目的 是 提供 无 需 通过	0.00318471
的 目的 是 提供 一种 用于	0.00318471
的 目的 在于 提供 在 具有	0.00318471
的 目的 在于 提供 显示	0.00318471
的 目的 在于 提供 显示 出	0.00318471
发明 的 目的 在于 提供 显示	0.00318471

次に単語対応データを用いて中国語表現を求める手法について述べる。Moses の学習フェーズでは Giza++による単語対応データが得られる。単語対応データには原言語(中国語)から目標言語(日本語)に向けての対応(src-ref.A3.final)と目標言語から原言語に向けての対応(ref-src.A3.final)がある。この両者のデータから「提供することを目的とする」に対応する中国語表現を抽出した²。対応には分離した対応も存在する。分離した部分には(+)を挿入して抽出した。たとえ

² 日本語表現中の句点(.)の対応は抽出から除外した。

ば、図 2.2.1 では中国語側で「提供」と「目的」が分離して対応しており³、「提供することを目的とする」に対応する中国語表現は「提供 (+) 目的。」となる。

	ロック	状態	に	なる	まで	の	時間	が	短い	PLL	回路	等	を	提供	する	こと	を	目的	と	する	。	
以																						
提供																						
达到																						
锁定																						
状态																						
的																						
时间																						
短																						
的																						
PLL																						
电路																						
为																						
目的																						
。																						

図 2.2.1 Giza++による分離した単語対応例

こうして抽出された中国語表現は異なり数で 398 件であった。頻度の大きい中国語表現を表 2.2.2 に示す。単語対応データから抽出した中国語表現はフレーズテーブルから抽出したそれに加えて分離した表現が抽出できる。ただ、分離した表現でもたとえば「的 目的 (+) 提供 (+) 种」は実際の中国語原文を調べてみると「的 目的 {是|在于|为|是 再次} 提供 一 种」と表現することができ、連続した表現の範囲で記述可能である。一方、前述した「提供 (+) 目的。」は図 2.2.1 に示すように連続した表現の範囲では記述できない。

³ 「する」は NULL に対応している。

表 2.2.2 単語対応データから抽出した「提供することを目的とする」に対応する中国語表現
(中国語表現が空欄のデータは単語対応がすべて NULL のデータである。)

頻度	中国語表現
30	的 目的 在于 提供
26	
24	的 目的 (+) 提供 (+) 种
20	目的 在于 提供
20	提供 一种
17	本 发明 的 目的 是 提供 一种
15	的 目的 (+) 提供
14	本 发明 的 目的 在于 提供 一种
14	本 发明 提供 一种
14	一种
13	目的 (+) 提供
10	本 发明 的 目的 是 提供 一种 (+) 的
9	本 发明 的 目的 在于 提供 一种 (+) 的
8	本 发明 提供
7	目的 在于 提供 一种
7	本 发明 的 目的 在于 提供 (+) 的
7	的 目的 是 提供 一种
7	的 目的 在于 提供 (+) ,
7	提供 (+) 目的 。
6	目的 在于 提供 (+) 。
6	的 目的 在于 提供 (+) 。
6	的 目的 (+) 提供 (+) 。
6	提供 (+) 的
5	目的 是 提供 一种
5	目的 是 提供
5	的 目的 (+) 提供 一种
5	提供 一种 (+) 的
5	提供 (+) 种 (+) ,
5	提供 (+) 种

2.2.2.3 中国語文パターンの抽出

表 2.2.1 や表 2.2.2 に例示する中国語表現を手掛かりに中国語文パターンを求めることができる。たとえば、「提供 (+) 目的 。」をキーとして原文を検索することで、中国語文パターン「以 提供 NP 为 目的 。」が求められる。ここで NP は名詞句を示す変数(非終端記号)である。さらに対応する日本語と合わせて

「以 提供 NP 为 目的 。」→「NP を 提供 する こと を 目的 と する 。」

という翻訳規則を作成することができる。

2.2.3 まとめと今後の課題

機械翻訳の 2 大課題である語彙的曖昧性解消と構造的曖昧性解消のうち、後者に関して、その精度を向上できる可能性がある文パターンの自動抽出法について述べた。主要素後置型言語であ

る日本語では、文末表現に着目することで、文パターンが自動的に得られる。このことと統計的機械翻訳の手法を利用して、日中文対応コーパスから中国語の文パターンを求める方法を提案した。統計的機械翻訳の学習過程で得られる単語対応データとフレーズテーブルデータから中国語のキー表現が得られ、それを利用して中国語文パターンを求める。現在、最後の部分は手動で行っており、その意味で、本手法は半自動手法である。

今後の課題として以下のことが考えられる。

- ・本手法は、最終部分の中国語文パターンの作成を手動で行っている。この部分を自動化する。
- ・特許文献の各段落や分野によって頻出する文パターンが異なる。これらを網羅的に調査して文パターン辞書を構築する。
- ・文パターン辞書を実際の機械翻訳システムに組み込んで利用する。
- ・日本語でも文末表現以外の文パターンが考えられる。例えば「基体粒子上に、金属硫化物、金属フッ化物、金属炭酸塩又は金属リン酸塩からなる薄膜の少なくとも1層を含む多層膜を成膜してなることを特徴とする多層膜被覆粉体。」では、「ことを特徴とする」というパターンが文末ではなく、文の中間に存在する。このような場合でも文パターンを効率的に抽出する方法を考案する。

参考文献

- [1] 江原暉将、田中穂積：機械翻訳における自然言語処理、情報処理、自然言語処理技術の応用 特集号、Vol.34, No.10, pp.1266-1273, Oct., 1993.
- [2] Hiroyuki Kaji, Yuuko Kida, Yasutsugu Morimoto : Learning Translation Templates from Bilingual Text, Proceedings of COLING-92, pp.672-678, Aug, 1992.
- [3] 加藤直人:定型パターンを含む文の機械翻訳手法、情報処理学会論文誌、Vol.36, No.9, pp. 2081 - 2090, Sept, 1995.
- [4] 船守茉美：中国公開特許公報の日本語への機械翻訳、特技懇 262 号、pp.3-10, Aug, 2011.
- [5] Minah Kim : Current Status of Korea's Machine Translation for Patent Domain Users, 第 1 回特許情報シンポジウム資料, Dec, 2010.
- [6] Wang Dan : Making Effective Use of Machine Translation for Patent Documents: Practice of CPIC, 第 2 回特許情報シンポジウム資料, Nov, 2012.
- [7] 张冬梅, 刘小蝶, 晋耀红：基于模板的汉英专利机器翻译研究, 计算机应用研究, Vol.30, No.7, pp.2044-2047, July, 2013.
- [8] Jin'ichi Murakami et. al : Pattern-Based Statistical Machine Translation for NTCIR-10 PatentMT, Proceedings of the 10th NTCIR Conference, pp.350-355, June, 2013.
- [9] 特許庁：日本語特許出願書類の中国語への機械翻訳に関する調査報告書、Feb, 2011.
- [10] 江原暉将：中国語特許文書から文パターンを抽出する一方法、Japio YEAR BOOK 2013, pp.270-275, Nov, 2013.
- [11] Philipp Koehn, Franz J. Och, Daniel Marcu : Statistical Phrase-Based Translation, Proceedings of HLT-NAACL 2003, pp.48-54, May-June, 2003.

付録 1 文対応例

下記の文書は日本語部分が 3 文から成り、中国語部分も 3 文から成る。本文記載の手法によって文対応を取ると日本語部分の 1 文目は中国語部分の 1 文目に、日本語部分の 2 文目は中国語部分の 2 文目に、日本語部分の 3 文目は中国語部分の 3 文目に対応する。

日本語部分：

ハイブリッド・カーにおける多相交流回転機による駆動装置（内燃機関）へのアシストおよび電池への充電を電池の状態に応じて行えるようにする。

車両に搭載された電池の放電時および充電時の電流、電圧を測定し、プログラム制御回路がこの測定された電流、電圧の情報に基づき、電池と車両の駆動装置に連結された多相交流回転機との間の交流直流もしくは直流交流変換を行うインバータを介して駆動装置に補助動力を与えるときの放電電流を制御する。

電池の充電状態に応じて充放電電流が制限されるので、電池の充電量の多少にかかわらず一様に充放電が行われることが避けられ、充電効率を高めるとともに、電池の使用寿命を増大することができる。

中国語部分：

本发明能够根据电池的状态来进行由混合式汽车多相交流电机产生的、给驱动装置（内燃机）的加速以及给电池的充电。

对车辆上装载的电池其放电时及充电时的电流、电压进行测定，程序控制电路根据此测得的电流、电压信息，通过用于实行电池与连接于车辆驱动装置的多相交流电机之间的交流变直流或直流变交流变换的变换器，对在向驱动装置提供辅助动力时的放电电流进行控制。

由于是根据电池的充电状态来限制充放电电流，避免了不论电池充电量的多少而同样实行的充放电，因此在提高充电效率的同时，还能够使电池的寿命延长。

中国語部分の機械翻訳結果：

当發明は電池の状態によって混じる適当な自動車の多相の交流電気機械の發生の、から駆動装置（ディーゼル・エンジン）の加速およびに電池の充電に行うことができる。

車の上で積載する電池のその放電する時と充電する時の電流、電圧に対して測定を行って、プログラムの制御回路はこのによって電流、電圧の情報を測って、電池を実行して車の駆動装置の多相の交流電気機械間のにつながって次第に変化する直流に交際するあるいはとずっと次第に変化して変換の変換機を交流することに用いるを通じて(通って)、駆動装置に動力に協力することに提供する時の放電する電流に対してコントロールを行う。

なことが電池の充電する状態によって放電する電流を担当することを制限しにくるため、電池が電氣量のいくらを担当して同様に実行したのが担当して放電するのであろうとを免れて、そのため充電する効率を高める同時に、また電池の寿命に延長することができる。

付録2 頻出する日本語文パターン

文節数	頻度	文パターン	文節数	頻度	文パターン
5	27	製造方法を提供することを目的とする。	3	41	ことを目的としている。
5	21	その製造方法を提供することにある。	3	39	低減することができる。
4	406	提供することを目的とする。	3	39	向上させることができる。
4	77	提供することを課題とする。	3	38	提供することができる。
4	41	有することを特徴とする。	3	38	高めることができる。
4	38	含むことを特徴とする。	3	35	提供を目的とする。
4	34	提供することを目的としている。	3	31	目的とするものである。
4	31	製造方法を提供することにある。	3	30	半導体装置の製造方法を提供する。
4	29	ことを目的とするものである。	3	29	実現することができる。
4	24	なることを特徴とする。	3	29	抑えることができる。
4	22	含有することを特徴とする。	3	28	抑制することができる。
4	22	備えることを特徴とする。	3	27	形成することができる。
4	20	形成されていることを特徴とする。	3	25	確保することができる。
3	628	ことを特徴とする。	3	24	ことを可能とする。
3	522	ことを目的とする。	3	22	製造することができる。
3	210	提供することにある。	3	21	方法を提供すること。
3	162	その製造方法を提供する。	2	1667	ことができる。
3	109	行うことができる。	2	970	提供すること。
3	104	ことを課題とする。	2	639	特徴とする。
3	101	得ることができる。	2	577	目的とする。
3	80	ことが可能となる。	2	414	製造方法を提供する。
3	75	製造方法を提供すること。	2	271	ことにある。
3	72	防止することができる。	2	124	ことができるようにする。
3	68	することができる。	2	123	提供することである。
3	64	図ることができる。	2	116	課題とする。
3	62	ことを特徴としている。	2	106	可能となる。
3	48	適用することができる。	2	105	方法を提供する。
3	42	電子機器を提供すること。	2	104	ことが好ましい。

3. 1 パテントファミリーにおける対訳文対非抽出部分を利用した 専門用語訳語推定方式の評価と分析

筑波大学大学院システム情報工学研究科

豊田 樹生, 龍 梓, 董 麗娟

宇津呂 武仁, 山本 幹雄

3.1.1 はじめに

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必要不可欠なものである。この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。本論文では、日米パテントファミリーを情報源として、専門用語対訳辞書を生成する手法を提案する。従来より、日米パテントファミリーの対応特許文書中において、「背景」および「実施例」の部分の日英対訳文対の対応付けを行い、これを情報源として専門用語の対訳辞書を生成する手法が提案されている。例えば、NTCIR-7 特許翻訳タスク[1]において配布された日英 180 万件の対訳特許文を用いて、対訳特許文からの専門用語対訳対獲得を行った研究がある。この研究では、句に基づく統計的機械翻訳モデル[2]を用いることにより、対訳特許文から学習されたフレーズテーブル、要素合成法、Support Vector Machines(SVMs) [8]を用いることによって、専門用語対訳対獲得を行った。しかし、上述の方式では、対訳文対が抽出される部分は、「背景」及び「実施例」全体の約 30%であり、約 70%は利用されていなかった。そこで、本論文では、「背景」および「実施例」のうちの残りの 70%の部分を言語資源として、既存の対訳辞書を用いて専門用語の訳語推定を行う方式の有効性を実証する。特に、人手で作成された辞書である英辞郎及びその部分対訳辞書に加えて、全体の約 30%を訓練例として学習したフレーズテーブルを併用して要素合成法を適用し、専門用語の訳語推定を行う方式、および、フレーズテーブルのみを用いて要素合成法を適用し、専門用語の訳語推定を行う方式を提案する。また、要素合成法を適用する際には、訳語候補が相手言語側特許文書中に存在するか否かの検証を行うことにより、高精度な訳語推定を実現する。その結果、提案手法の要素合成法により、高い適合率・再現率を達成することができた。

3.1.2 日英対訳特許文

本論文では、フレーズテーブルの訓練用データとして、NTCIR-7 の特許翻訳タスク[1]で配布された約 180 万対の日英文対応データを使用した。なお、この文対応データは以下に示す手順で作成されたものである。

1. 1993-2000 年発行の日本公開特許広報全文と米国特許全文を得る。
2. 米国特許の中から日本に出願済みのものを優先権番号より得て、日英対訳特許文書を取得する。
3. 日英対訳特許において日英間で比較的直訳されている関係となっている度合いが大きい「背景」及び「実施例」の部分抽出する。

4. 抽出した部分に対して、文献[7]の手法によって日英間で文対応をつける。

3.1.3 句に基づく統計的機械翻訳モデルのフレーズテーブル

本論文で用いるフレーズテーブルでは、日英の句の組、及び、日英の句が対応する確率を推定し記述する。このとき、前節で述べた文対応データに対して、句に基づく統計的機械翻訳モデルのツールキットである Moses[2]を適用する。Moses によってフレーズテーブルを作成する過程を以下に示す。

1. 単語の数値化、単語のクラスタリング、共起単語表の作成などの処理を文対応データに対する前処理として行う。
2. 文対応データを利用し、最尤な単語対応を英日・日英の両方向において得る。
3. 英日・日英両方向における単語対応を利用し、ヒューリスティックスを用いることにより、対称な単語対応を得る。
4. 対称な単語対応を用いて、可能な全ての日英の句の組を作成する。そして、各組に対して、「文単位の句対応制約」の条件に対する違反の有無をチェックする(違反しない句の組を有効な対応とみなす)。
5. 文対応データにおける日英の句の対応の数を集計する。このとき、各句の対応に翻訳確率を付与する。

手順(4)について、以下に「文単位の句対応制約」の条件を示す。

日本語文の形態素列中の形態素を文頭から順に V_1, V_2, \dots, V_n 英文の単語列中の単語を文頭から順に W_1, W_2, \dots, W_n とし、日本語句を $P_J (= V_p \cdots V_{p'})$ とし、英語句を $P_E (= W_q \cdots W_{q'})$ とする。ここ

で、日英句の組 $\langle P_J, P_E \rangle$ が含まれるある一つの対訳文対 $\langle T_J, T_E \rangle$ 中において得られているあらゆる

単語対応 $\langle V_i, W_j \rangle$ について、「 $p \leq i \leq p' \Leftrightarrow q \leq j \leq q'$ 」が成り立つ場合に、 P_J と P_E は対訳文対

$\langle T_J, T_E \rangle$ において「文単位の句対応制約」に違反しない、と定義する。

3.1.4 要素合成法による訳語推定

3.1.4.1 既存の対訳辞書およびフレーズテーブル

本論文においては、既存の対訳辞書として、「英辞郎」¹²に加えて、英辞郎の訳語対から作成した部分対応対訳辞書[5]、及び、前節で述べたフレーズテーブルを用いる。両者における見出し語数および訳語対数を【表 1】に示す。部分対応対訳辞書生成の手順は以下のとおりである。まず、既存の対訳辞書から、日本語及び英語の用語がそれぞれ2つの構成要素(具体的には、日本語の場

¹ <http://www.eijiro.jp/>

² 本論文では、英辞郎 Ver. 79 及び Ver. 131 を用いる。

合はJUMAN³による形態素解析によって得られる形態素列、英語の場合は単語列) からなる訳語対を抽出し、これを別の対訳辞書 P_2 とする。次に、 P_2 中の訳語対の第一構成要素から前方一致部分対応対訳辞書 B_P を作成し、第二構成要素から後方一致部分対応対訳辞書 B_S を作成する。本論文においては、英辞郎についてはVer.131を使用し、前方一致部分対応対訳辞書及び後方一致部分対応対訳辞書については、Ver.79及びVer.131を統合したものをを用いた。

3.1.4.2 訳語候補のスコア

訳語候補のスコアを $Q(y_S, y_T)$ とする。このとき、 y_S は日本語専門用語を、 y_T は生成された訳語候補を表し、 y_S は構成要素 s_1, s_2, \dots, s_n に、 y_T は構成要素 t_1, t_2, \dots, t_n に分解できると仮定する。すると、

$Q(y_S, y_T)$ は、対訳辞書スコア $\prod_{i=1}^n q(\langle s_i, t_i \rangle)$ とコーパススコア $Q_{\text{corpus}}(y_T)$ の積で定義される。実際には、

ある訳語候補が2つ以上の系列の訳語対から生成される場合があるので、本論文では、以下に示すように、それぞれの系列のスコアの和によって $Q(y_S, y_T)$ を定義する。

$$Q(y_S, y_T) = \sum_{y_S = s_1, s_2, \dots, s_n} \prod_{i=1}^n q(\langle s_i, t_i \rangle) \cdot Q_{\text{corpus}}(y_T)$$

このとき、対訳辞書スコアはこの構成要素同士のスコアの積によって求まり、コーパススコアは訳語候補が目的言語側のコーパスに出現するか否かによって求まる。

3.1.4.2.1 対訳辞書スコア

構成要素の訳語対 $\langle s, t \rangle$ の対訳辞書スコア $q(\langle s, t \rangle)$ は、訳語対が英辞郎、前方一致部分対応対訳辞書 B_P 、または、後方一致部分対応対訳辞書 B_S に含まれる場合のスコア q_{man} 及び訳語対がフレーズテーブルに含まれる場合のスコア q_{smt} の和によって求まる。

$$q(\langle s, t \rangle) = q_{\text{man}}(\langle s, t \rangle) + q_{\text{smt}}(\langle s, t \rangle)$$

$$q_{\text{man}}(\langle s, t \rangle) = \begin{cases} 1 & \langle s, t \rangle \text{が英辞郎, } B_P, \\ & \text{または, } B_S \text{に含まれる場合)} \\ 0 & \text{(それ以外の場合)} \end{cases}$$

$$q_{\text{smt}}(\langle s, t \rangle) = \begin{cases} P(t|s) & \langle s, t \rangle \text{がフレーズ} \\ & \text{テーブルに含まれ、} \\ & \text{かつ、} P(t|s) \geq p_0 \\ & \text{である場合)} \\ 0 & \text{(それ以外の場合)} \end{cases}$$

上記の定義においては、訳語対 $\langle s, t \rangle$ が英辞郎、 B_P 、または、 B_S に含まれる場合、 $q_{\text{man}}(\langle s, t \rangle)$ は1

³ <http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN>

となり、それ以外の場合は 0 となる⁴。一方、 $\langle s, t \rangle$ がフレーズテーブルに含まれる場合は、翻訳確率の下限 p_0 のパラメータに従い、スコアを決定する。このパラメータ p_0 は、6 節において、評価用データ以外の調整用データを用いて最適化される。

3.1.4.2.2 コーパススコア

コーパススコアは、訳語候補 y_T が目的言語側のコーパスに出現する場合にのみ 1 となり、出現しない場合には 0 となる。

$$Q_{\text{corpus}}(y_T) = \begin{cases} 1 & (y_T \text{ が目的言語側コーパス} \\ & \text{に出現する場合}) \\ 0 & (y_T \text{ が目的言語側コーパス} \\ & \text{に出現しない場合}) \end{cases}$$

3.1.4.2.3 例

例として、専門用語“並列態様”の対訳“parallel mode”を獲得する様子を【図 1】に示す。本論文では、まず、この日本語専門用語“並列態様”を構成要素 s_1 の“並列”と s_2 の“態様”に分解し、これらを既存の対訳辞書及びフレーズテーブルを利用して目的言語に翻訳する。そうすると、 s_1 からは t_1 として“parallel”, “concurrent”, “multiple” が、 s_2 からは t_2 として“aspect”, “mode”, “form” が生成され、さらに各々に訳語の参照元に応じたスコアが付与される。次に、前置詞句の構成を考慮した語順の規則にしたがって、それらの構成要素の訳語を結合し、訳語候補を生成する。このとき、各訳語候補の対訳辞書スコアは t_1 と t_2 の積となる。例えば、“parallel aspect”の対訳辞書スコアは $(1.0+0.8) \times 1.0 = 1.8$ となる。

最後に、これら訳語候補を対訳辞書スコア順に、目的言語側のコーパスに対して照合を行い、もし照合すればそのコーパススコアは 1、照合しなければ 0 となる。この場合、結果的に、訳語候補のスコアが最も高い“parallel mode”が獲得される。

⁴ 文献[6]においては、英辞郎中の訳語対のスコアは構成要素の数が多いほど大きくなり、一方、部分対応対訳辞書中の訳語対のスコアは、英辞郎における部分対訳対の頻度が大きいほど大きくなるというスコア体系が採用されている。本論文において、このスコア体系の評価を行なった結果では、本論文で用いているスコア体系との間で大きな性能差は観測されなかったため、本論文では、より簡易なスコア体系を採用した。

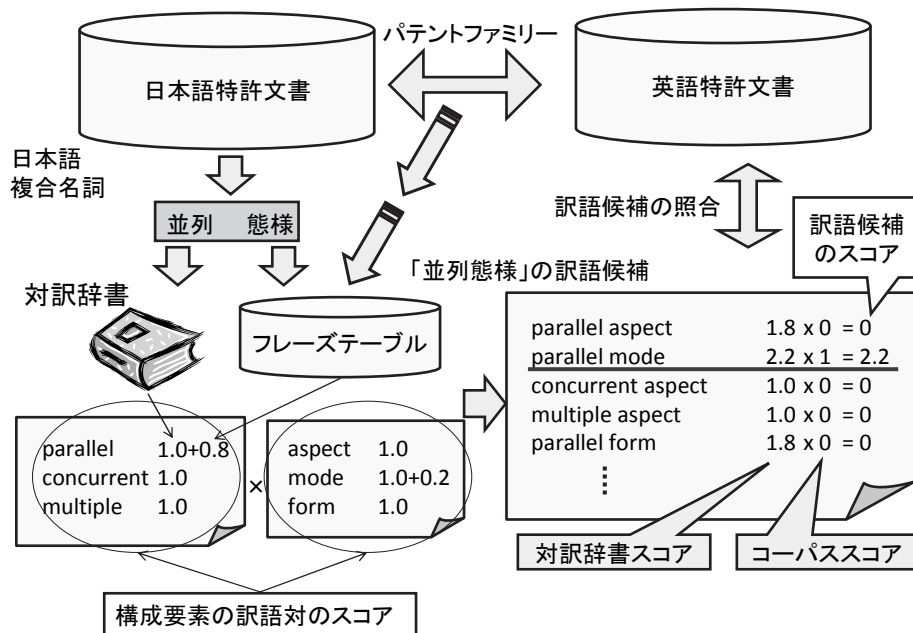


図 1: 要素合成法を用いた日本語専門用語「並列態様」の訳語推定

3.1.5 対訳文対非抽出部分における訳語推定

本論文で用いる日米パテントファミリーの日本語側 D_J は、「背景」 B_J , 「実施例」 M_J , および、「背景・実施例以外の部分」 N_J から構成されている. そして, これらの部分のうち, 「背景」 B_J および「実施例」 M_J は, 対訳文対抽出部分 PSD_J , 及び, 対訳文対非抽出部分 $NPSD_J$ に分割される. また, 英語側の特許文書の全体 D_E に対しても, 同様に, 「背景」 B_E , 「実施例」 M_E , および, 「背景・実施例以外の部分」 N_E から構成され, 「背景」 B_E および「実施例」 M_E は, 対訳文対抽出部分 PSD_E , 及び, 対訳文対非抽出部分 $NPSD_E$ に分割される. この特許文書の構成の例を【

図 2】に示す.

$$\begin{aligned}
 D_J &= \langle B_J, M_J, N_J \rangle \\
 B_J \cup M_J &= \langle PSD_J, NPSD_J \rangle \\
 D_E &= \langle B_E, M_E, N_E \rangle \\
 B_E \cup M_E &= \langle PSD_E, NPSD_E \rangle
 \end{aligned}$$

「数値演算処理装置」に関する日英対訳特許文書

	日本語側	英語側	
実施例	PSD 0001 ⋮	【実施例】 まず…ニューラルネットワークを 1つの適用例として説明する。 ⋮	EMBODIMENTS Description is now made …with reference to an exemplary neural network. ⋮
	NPSD	しかしながら、図45に示す構成に おいては、フラグSTOPおよびEN Dの少なくとも一方が“1”の場合に は、NOR回路300からレジスタ ファイル(図33に示すレジスタフ ァイルは220)およびローカルメモ リ11への数値のデータの書込みが 禁止されるため、…処理対象アド レスの演算ユニット間の不一致の 発生を防止することができ、全ての 演算ユニットを並列態様で動作さ せることができる。	In the structure shown in FIG. 45, however, writing of numeric data from the NOR circuit 300 to the register file (220 shown in FIG. 33) and to the local memory 11 is inhibited when at least one of the flags STOP and END is “1”. …Thus, it is possible to avoid mismatching between the addresses to be processed in the arithmetic units, thereby driving all arithmetic units in a parallel mode.
	⋮	⋮	

要素合成法適用
→parallel mode

照合
して発見

図2：日米パテントファミリー中の「実施例」に
おける対訳文対非抽出部分の例

ここで、本論文では、英訳語推定対象となる日本語専門用語 t_J を抽出するにあたっては、対訳文抽出部分 PSD_J 中の日本語専門用語の英訳語の多くは対訳文対から学習したフレーズテーブル中に含まれると予測し、「背景」 B_J 及び「実施例」 M_J における対訳文対非抽出部分 $NPSD_J$ を抽出元とした。

次に、その日本語専門用語 t_J に対して、英語側の「背景」 B_E 及び「実施例」 M_E を英語側コーパスとみなして要素合成法を適用し、英語訳語候補の集合 $TranCand(t_J, B_E \cup M_E)$ を作成した⁵。

$$\begin{aligned}
 & TranCand(t_J, B_E \cup M_E) \\
 & = \{t_E \in B_E \cup M_E \mid t_J \text{ に対して要素合成法} \\
 & \quad \text{により } t_E \text{ を生成し } Q(t_J, t_E) > 0\}
 \end{aligned}$$

そして、この $TranCand(t_J, B_E \cup M_E)$ を用いて、以下の関数 $CompoTrans_{\max}$ によりスコア最

⁵ ここで、比較評価として、英語側の「背景」 B_E 及び「実施例」 M_E における対訳文対非抽出部分 $NPSD_E$ のみを英語側コーパスとみなして要素合成法を適用する評価実験も行ったが、英語側コーパス中において適切な訳語候補を照合できる割合が下がったため、本論文においては、英語側の「背景」 B_E 及び「実施例」 M_E を英語側コーパスとみなして要素合成法を適用する方式を採用した。

大となる訳語候補を得る.

$$\begin{aligned} & \mathit{CompoTrans}_{\max}(t_J, B_E \cup M_E) \\ &= \mathop{\text{argmax}}_{t_E \in \text{TransCand}(t_J, B_E \cup M_E)} Q(t_J, t_E) \end{aligned}$$

以上の手順により, 日英対訳特許文書の英語側の「背景」 B_E 及び「実施例」 M_E から英語専門用語 t_E を獲得する.

3.1.6 評価

3.1.6.1 評価対象日本語専門用語の選定

提案手法を評価するため, 以下の3通りの比較を行った.

- (i) 「英辞郎のみ」・・・要素合成法における対訳辞書として, 英辞郎及びその構成要素から生成される辞書を用いる.
- (ii) 「フレーズテーブルのみ」・・・要素合成法における対訳辞書として, フレーズテーブルを用いる.
- (iii) 「英辞郎及びフレーズテーブル」・・・要素合成法における対訳辞書として, 英辞郎及びフレーズテーブルを用いる.

はじめに, パテントファミリー1,000組を取り出し, そこから 61,133 例の日本語複合名詞を抽出した.⁶次に, これら 61,133 例の日本語複合名詞に対して4節で述べた要素合成法を訳語推定手法として適用し, 表2-(1)に示すように, 以下の5つのカテゴリに分類した.

また, このとき, (c) に関しては集合の大きさが最大になるように, 要素合成法において用いるフレーズテーブルの翻訳確率の下限值 p_0 (式(1))を0とした.

- (a) 日本語複合名詞が英辞郎に含まれ, その訳語が英語側特許文書に含まれる.
- (b) (a) 以外の場合で, 日本語複合名詞がフレーズテーブルの日本語側と完全一致する.⁷
- (c) (a), (b) 以外の場合で, 日本語複合名詞に対して訳語推定手法を適用した結果, その訳語が英語側特許文書に含まれる.
- (d) (a), (b), (c) 以外の場合で, 日本語複合名詞に対する訳語が英辞郎及び, 訳語推定手法によって生成されるが, 英語側特許文書に含まれない.
- (e) (a), (b), (c), (d) 以外の場合で, 日本語複合名詞に対する訳語が英辞郎及び訳語推定手法により生成されない.

以下では, まず, 表2-(1)の(c)欄に示すように, 上記(i)-(iii)の3通りの手法を適用した結果, 訳語候補が英語側特許文書に含まれる日本語複合名詞の集合を求める.

⁶ MeCab(IPAdic)を用いて形態素解析を行った後, 名詞・接頭辞・接尾辞・未知語のいずれかの品詞の最長の形態素列に加え, 数字・アルファベットの列が接続することを許容したものを名詞句として抽出する.

⁷ フレーズテーブルの生成に用いるコーパスに対し, 複数の組み合わせで, 形態素解析器(MeCabもしくはChasen)の適用・一定の形態素数以上の文の除去などの処理を行い, 何通りかのコーパスを作成した. これらのコーパスから生成されたフレーズテーブルのうち最も規模の大きいフレーズテーブルに対して日本語側と完全一致するか否かを調べた.

- ・対訳辞書として上記(i)「英辞郎のみ」を用いた場合について、求められた日本語複合名詞の集合を集合 E (4,003 例) とする.
- ・上記(ii)「フレーズテーブルのみ」を用いた場合について、求められた日本語複合名詞の集合を集合 P とする.
- ・上記(iii)「英辞郎及びフレーズテーブル」を用いた場合について、求められた日本語複合名詞の集合を集合 EP とする.

次に、集合 E , $(P \cup EP) - (E \cap (P \cup EP)) = (P \cup EP) - E = EV$ より,
 E' (202 例), EV' (373 例) をそれぞれ作成する. ここで、集合 E 中の日本語複合名詞については、対訳辞書として上記(i)「英辞郎のみ」を用いることにより、大半の日本語複合名詞の英訳語を正しく推定できることが分かっているため、集合 M に対しては、集合 E に含まれる日本語複合名詞を除外し、集合 $(P \cup EP) - (E \cap (P \cup EP)) = (P \cup EP) - E = EV$ を作成した後、これらを母集団として評価用の日本語複合名詞の集合 EV' を作成した.

加えて、上記の集合に含まれる日本語複合名詞を以下のように分類した.

- (c1) 専門用語であり、かつその適切な訳語が目的言語側に存在する.
- (c2) 専門用語であるが、その適切な訳語が目的言語側に存在しない.
- (c3) 専門用語以外の、形態素解析における区切り位置が誤っているもの、複合名詞抽出規則の誤り、などに該当する日本語複合名詞.
- (c4) 専門用語以外の一般複合名詞.

加えて、【表 3】、【表 4】、【表 5】、【表 6】 にこれらの例を示す.

そして、評価用セット EV' , E' を上述の分類に基づき、以下のように分割した.

集合 EV' のうち(c1) に属する日本語複合名詞を抽出した集合を EV'_1 ($|EV'_1| = 179$) とする. また、集合 EV' のうち(c2) に属する日本語複合名詞を抽出した集合を EV'_2 ($|EV'_2| = 29$) とする.

集合 E' のうち(c1) に属する日本語複合名詞を抽出した集合を E'_1 ($|E'_1| = 181$) とする. また、集合 E' のうち(c2) に属する日本語複合名詞を抽出した集合を E'_2 ($|E'_2| = 4$) とする.

また、【図 15】に EV' における c1~c4 のフレーズテーブルの翻訳確率に対して下限値を設け、この下限値を変化させていったときの分布を示す.

3.1.6.2 評価結果

要素合成法においてフレーズテーブルの翻訳確率に対して下限値を設けた. この節では、これらの下限値を変化させたときの再現率・適合率の推移および訳語対正解数・棄却数の推移をグラフにして報告する.

まず、集合 E'_1 , $E'_1 \cup E'_2$, E' を用いて、上記(i)「英辞郎のみ」の場合の再現率・適合率を求めた. その評価結果を順に、【図 3】、【図 4】、【図 5】に示す.

集合 E'_1 を用いた場合、適合率・再現率ともに90.1%となった.

集合 $E'_1 \cup E'_2$ を用いた場合、適合率88.1%・再現率90.1%となった.

集合 E' を用いた場合、適合率80.7%・再現率90.1%となった。

また、集合 E_1' 、 $E_1' \cup E_2'$ 、 E' を用いて、上記(i)「英辞郎のみ」の場合の訳語対の正解数・棄却数の推定値を求めた。

その結果を順に、【図 9】、【図 10】、【図 11】に示す。

集合 E_1' を用いた場合、正解数は3230、棄却数は357となった。

集合 $E_1' \cup E_2'$ を用いた場合、正解数は3230、棄却数は436となった。

集合 E' を用いた場合、正解数は3230、棄却数は773となった。

つぎに、集合 EV_1' 、 $EV_1' \cup EV_2'$ 、 EV' を用いて、上記(ii)「フレーズテーブルのみ」、および、上記(iii)「英辞郎及びフレーズテーブル」の場合の再現率・適合率を求めた。

その評価結果を順に、【図 6】、【図 7】、【図 8】に示す。

集合 EV_1' を用いた場合、要素合成法(dic=P)は再現率13%~40%の範囲において、適合率90%以上を達成した。要素合成法(dic=EP)は再現率27%~43%の範囲において、適合率90%以上を達成した。

集合 $EV_1' \cup EV_2'$ を用いた場合、要素合成法(dic=P)は再現率10%~50%の範囲において、適合率80%以上を達成した。要素合成法(dic=EP)は再現率25%~51%の範囲において、適合率80%以上を達成した。再現率39.0%~43.0%の範囲において、要素合成法(dic=P)は適合率87.7%、要素合成法(dic=EP)は適合率86.5%となり、要素合成法(dic=P)のほうがわずかに高い適合率を示した。

集合 EV' を用いた場合、要素合成法(dic=P)は再現率29%~40%の範囲において、適合率50%以上を達成した。要素合成法(dic=EP)は再現率25%~51%の範囲において、適合率50%以上を達成した。再現率39.0%~43.0%の範囲において、要素合成法(dic=P)は適合率52.2%、要素合成法(dic=EP)は適合率55.8%となり、要素合成法(dic=EP)のほうがわずかに高い適合率を示した。

また、集合 EV_1' 、 $EV_1' \cup EV_2'$ 、 EV' を用いて、上記(ii)「フレーズテーブルのみ」、および、上記(iii)「英辞郎及びフレーズテーブル」の場合の訳語対の正解数・棄却数の推定値を求めた。

その結果を順に、【図 12】、【図 13】、【図 14】に示す。

集合 EV_1' を用いた場合、要素合成法(dic=P)は、正解数1000前後の時に棄却数50以下を、正解数2000前後の時に棄却数540以下を達成した。また、評価結果において適合率95.2%を示した点(下限値0.2)において正解数988・棄却数50となった。

要素合成法(dic=EP)は正解数1000前後の時に棄却数85以下を、正解数2000前後の時に棄却数420以下を達成した。また、評価結果において適合率91.7%を示した点(下限値0.2)において正解数1345・棄却数122となった。

集合 $EV_1' \cup EV_2'$ を用いた場合、要素合成法(dic=P)は、正解数1000前後の時に棄却数120以下を、正解数2000前後の時に棄却数825以下を達成した。また、評価結果において適合率87.7%を示した点(下限値0.15)において正解数1142・棄却数160となった。

要素合成法(dic=EP)は、正解数1000前後の時に棄却数170以下を、正解数2000前後の時に棄却数

635以下を達成した。また、評価結果において適合率86.5%を示した点(下限値0.2)において正解数1345・棄却数210となった。

集合 EV' を用いた場合、要素合成法(dic=P)は正解数1000前後の時に棄却数890以下を、正解数2000前後の時に棄却数3155以下を達成した。また、評価結果において適合率52.2%を示した点(下限値0.15)において正解数1142・棄却数1045となった。要素合成法(dic=EP)は正解数1000前後の時に棄却数875以下を、正解数2000前後の時に棄却数2705以下を達成した。また、評価結果において適合率55.8%を示した点(下限値0.2)において正解数1345・棄却数1066となった。

3.1.7 おわりに

本論文では、パテントファミリーから専門用語の対訳辞書を生成する方法について述べた。

NTCIR-7 特許翻訳タスクにおいて配布された対訳特許文対を訓練例として学習したフレーズテーブル、および、既存の対訳辞書に訳語対が登録されていない日英専門用語を対象として、人手で作成された辞書である英辞郎及びその部分対応対訳辞書とフレーズテーブルを併用した要素合成法およびフレーズテーブルのみを用いた要素合成法を適用した。提案手法の要素合成法により、高い適合率・再現率を達成することができた。今後の課題として、日本語専門用語獲得の際に、機械学習などを用いて、専門用語以外の語を効率よく除外していくことがあげられる。

表 1: 英辞郎における見出し語数及び訳語対数

辞書	見出し語数		訳語対数
	英語	日本語	
英辞郎	1,631,099	1,847,945	2,244,117
前方一致部分対応対訳辞書	47,554	41,810	129,420
後方一致部分対応対訳辞書	24,696	23,025	82,087
フレーズテーブル (日本語側エントリと完全一致する語の除外に使用)	33,845,218	33,130,728	76,118,632
フレーズテーブル(推定精度の評価に使用)	17,466,471	17,121,115	36,058,684

表 2: 対象日本語複合名詞の集合=全日本語複合名詞 61,133 個の場合

訳語推定手法		要素合成法		
		英辞郎のみ	フレーズテーブルのみ	英辞郎及びフレーズテーブル
(a)	英辞郎の英訳が英語側特許文書に含まれる	5,450(8.9%)		
(b)	フレーズテーブルの日本語側と完全一致	32,516(53.2%)		
(c)	推定された訳語が英語側特許文書に含まれる	4,003(6.5%) (集合 E)	14,500(23.7%) (集合 P)	14,846(24.3%) (集合 EP)
(d)	英辞郎または訳語推定手法により、英訳語候補生成可能であるが英語側特許文書中には含まれない	397(0.7%)	948(1.6%)	944(1.6%)
(e)	英辞郎または訳語推定手法により生成不能	18,767(30.7%)	7719(12.6%)	7,327(12.0%)
合計		61,133(100%)		

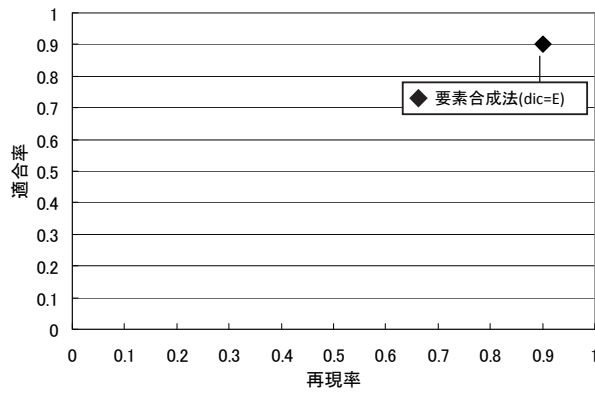


図 3: 集合 E_1' における評価結果

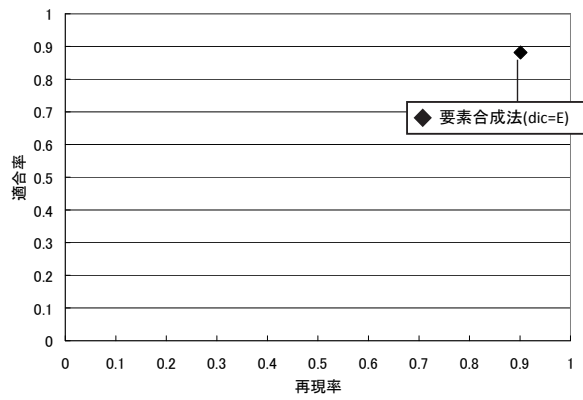


図 4: 集合 $E_1' \cup E_2'$ における評価結果

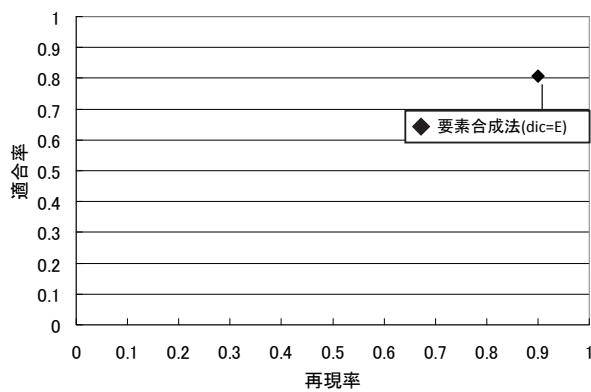


図 5: 集合 E' における評価結果

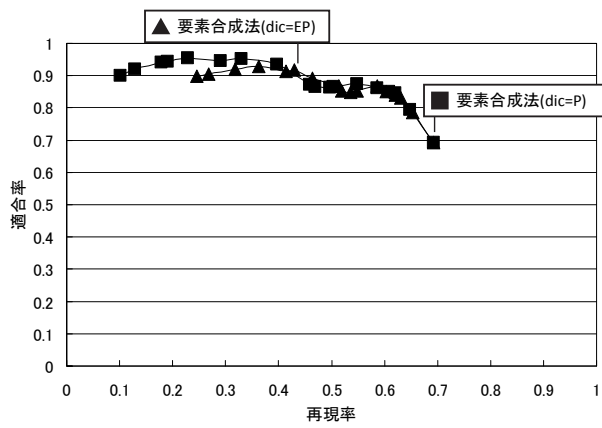


図 6: 集合 EV_1 における評価結果

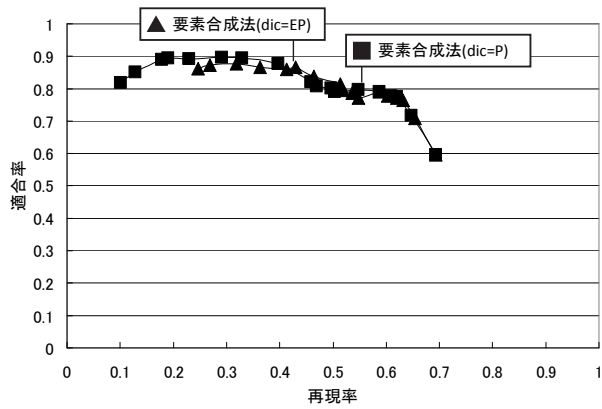


図 7: 集合 $EV_1' UEV_2'$ における評価結果

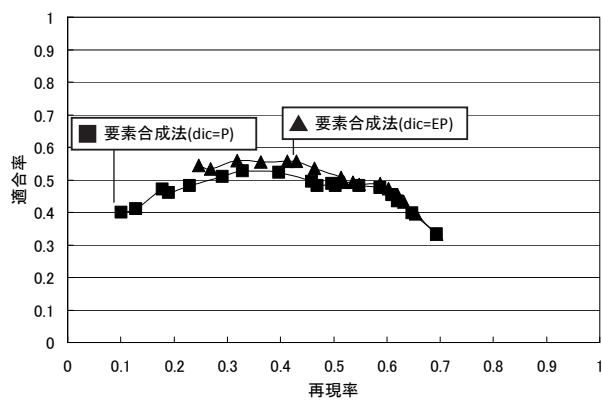


図 8: 集合 EV' における評価結果

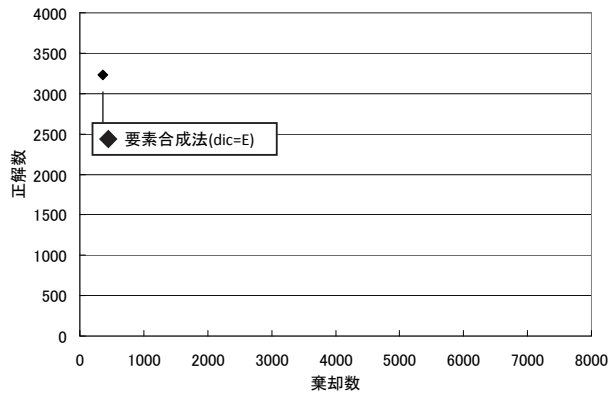


図 9: 集合 E_1' における棄却数-正解数の推移

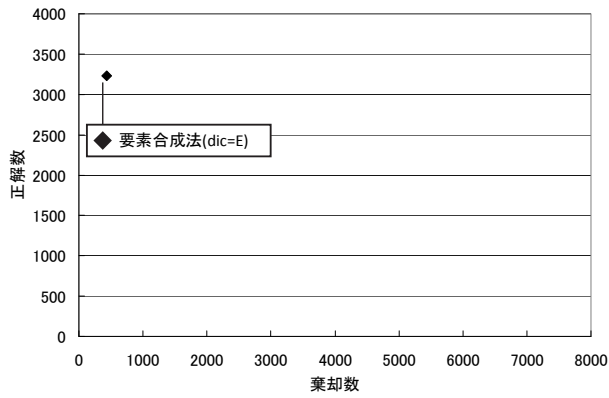


図 10: 集合 $E_1' \cup E_2'$ における棄却数-正解数の推移

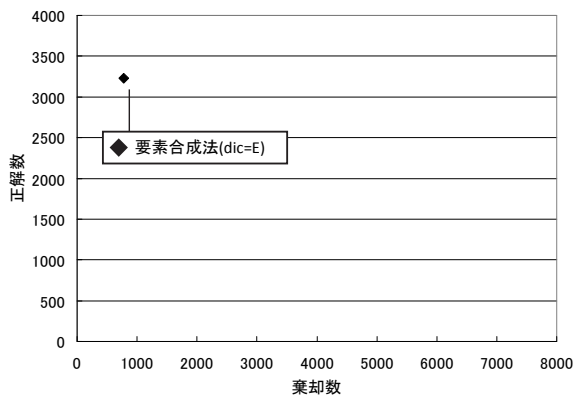


図 11: 集合 E' における棄却数-正解数の推移

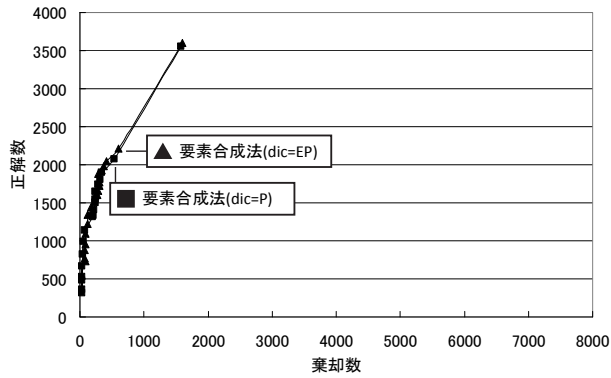


図 12: 集合 EV_1 における棄却数-正解数の推移

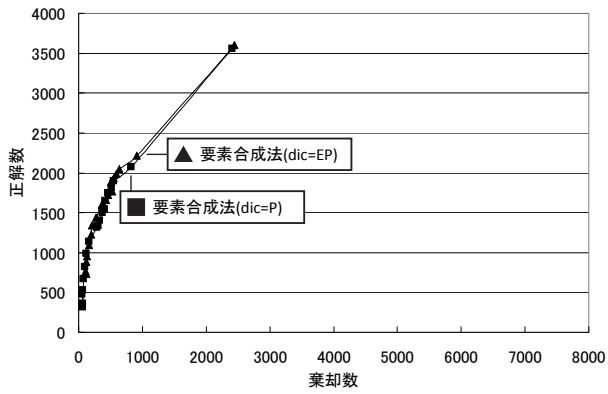


図 13: 集合 $EV_1 \cup EV_2$ における棄却数-正解数の推移

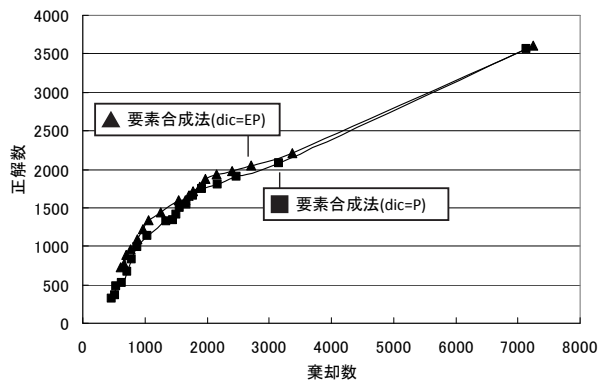


図 14: 集合 EV' における棄却数-正解数の推移

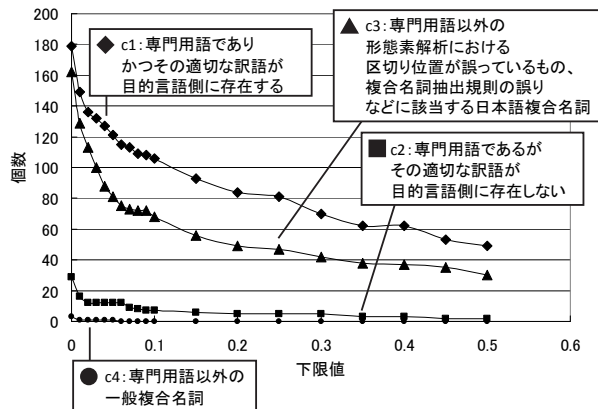


図 15: 集合 EV' における c1~c4 の分布

表 3: c1 において生成された訳語が参照用訳語と一致しなかった場合の例

日本語複合名詞	生成された訳語	参照用訳語
折流通路	flow passage	serpentine flow passages
操作環境構築手段	The	operational environment constructing means

表 4: c1 において生成された訳語が参照用訳語と一致した場合の例

日本語複合名詞	生成された訳語
非反転トライステート回路	non-inverting tri-state circuits
3次元連続断層像	three-dimensional sequential tomographic images

表 5: c3 に該当する日本語複合名詞の例

日本語複合名詞	抽出元の文
よう低消費電力	バッテリー駆動ができるよう低消費電力の...
誘電性液晶装置	本実験例では強誘電性液晶装置の...

表 6: c4 に該当する日本語複合名詞の例

日本語複合名詞	抽出元の文
具体的仕様	図2~図4の溝の具体的仕様を示した...
全列車	ステップ408で同一給電区間内に在線する全列車の情報...

参考文献

- [1] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the Patent translation task at the NTCIR-7 Workshop. In *Proc. 7th NTCIR Workshop Meeting*, pp. 389-400, 2008.
- [2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177-180, 2007.
- [3] 梁冰, 宇津呂武仁, 山本幹雄. 対訳特許文を用いた同義対訳専門用語の同定と収集. 言語処理学会第17回年次大会論文集, pp. 963-966, 2011.
- [4] 森下洋平, 梁冰, 宇津呂武仁, 山本幹雄. フレーズテーブルおよ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, Vol. J93-D, No. 11, pp. 2525-2537, 2010.
- [5] 外池昌嗣, 木田充洋, 高木俊宏, 宇津呂武仁, 佐藤理史. 要素合成法を用いた専門用語の訳語候補生成・検証. 言語処理学会第11回年次大会論文集, pp. 17-20, 2005.
- [6] 豊田樹生, 高橋佑介, 牧田健作, 宇津呂武仁, 山本幹雄. 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の利用. 情報処理学会研究報告, Vol. 2012-NL-208, 2012.
- [7] M. Utiyama and H. Isahara. A Japanese-English patent parallel corpus. In *Proc. MT Summit XI*, pp. 475-482, 2007.
- [8] V. N. Vapnik. *Statistical Learning theory*. Wiley-Interscience, 1998.

3. 2 単語間結合度に基づく複単語表現のアライメントの改善

京都大学 塩田 嶺明

科学技術振興機構 中澤 敏明

京都大学 黒橋 禎夫

3.2.1 はじめに

コーパスベースの機械翻訳では単語アライメントの結果から翻訳知識を獲得するため、高い翻訳精度を実現するには単語アライメントの品質を確保する必要がある。英仏などの似通った言語間では既に高いアライメント精度が実現されているが、日英など言語構造の大きく異なる言語間では、これらと比較すると十分な精度を達成できていないと言えない。

アライメントが困難となる原因の一つに、複数の単語で一つの意味を表す複単語表現がある。複単語表現は言語ごとに様々な種類があり、機能表現や慣用句、固有名詞などがその例である。複単語表現の一部には、構成語単体では全体の意味と結びつかない性質をもつものがある。この性質は、特に“かもしれない”や“in order to”などの機能表現に顕著にみられる。また、このような性質を持つ表現は、単語単体では相手言語側に明確に対応する語をもたないことが多い。特に膠着語である日本語では、長い機能表現が動詞に後続することが多いため、アライメントの難しさが顕著になっている。

近年主流となっているフレーズベースの統計的機械翻訳システム[1]では、単語アライメントの結果に対してヒューリスティクスを適用することで複数の単語からなるフレーズを獲得している。しかし、ここでは機械的に隣接する単語を連結してフレーズを作っているだけである。つまり、意味的なまとまりのある単語の集合と単なる単語 **n-gram** は明示的に区別されていないため、複単語表現の問題に十分に対処できていないと言いがたい。

本研究では、既存の依存構造解析を利用したアライメントモデルに単語列上の情報を加え、複単語表現のアライメントの改善を目指す。具体的には、隣接する二単語間の結合度を定義し、結合の強い単語同士は近接するフレーズに対応されるようなアライメントモデルを構築する。実験の結果、ベースラインに比べアライメント精度が向上することが確認できた。

3.2.2 関連研究

複単語表現の機械翻訳での扱いに関する先行研究では、複単語表現を事前に獲得して利用する方法が多く見られる。Lambert ら[2]は、双方向のアライメントの非対称性に着目して複単語表現の候補とそのスコアを獲得している。スコアが閾値以上のフレーズをアライメントや翻訳において一単語として扱っている。Ren ら[3]は、対訳コーパスの片方の言語側から対数尤度比を用いて複単語表現を獲得し、単語アライメントの結果からその対訳を獲得する。得られた複単語表現とその対訳を利用して SMT システムを補強している。Liu ら[4]は、単言語コーパスの各文を複製して対訳コーパスのように扱い、単語アライメントを行う。そこから同じ単語同士の対応を取り除くと共起しやすい単語同士の対応が得られる。これを用いて任意の二単語の共起しやすいの

コアを獲得し、アライメントや翻訳において素性として利用している。

3.2.3 アライメントモデル

本研究では中澤ら[5]のアライメントモデルをベースラインとし、隣接する二単語間の結合度に基づき、つながりの強い単語同士が近接するフレーズに対応されるようなアライメントモデルを構築した。

3.2.3.1 ベースラインシステム

中澤らのアライメントモデルでは、言語構造が異なる言語間でのアライメントを改善するため、依存構造木上でフレーズの依存関係を考慮することで語順の違いを吸収し、単言語での単語の派生を導入することで相手言語側に明確に対応する語を持たない孤立機能語に対応している。アライメントは以下のステップで行われる。まずは既存の単語アライメントツールによって単語レベルの対応を推定し、ヒューリスティックなルールを用いて依存構造木上で連続なフレーズの対応を獲得する。これを出発点とし、サンプリングを複数回繰り返しながらアライメントを修正していく。

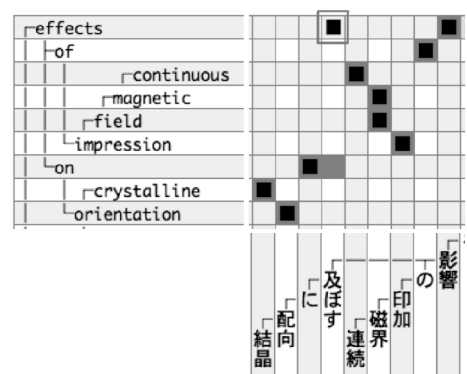


図 1:アライメント誤り例

このモデルでも、フレーズの言語的なまとまりとしての適切さは考慮されていない。図 1 は、複単語表現に関するアライメント誤りの例である。ここでは、“に 及ぼす”がかたまりとして扱われておらず、“及ぼす”が誤って“effects”と対応してしまっている。

また、中澤らは依存構造解析において意味主辞に基づく依存構造木を利用している。意味主辞は統語主辞に比べ内容語同士の依存関係が言語間で保持されやすいという特長がある。一方、例えば図 4 の“なければ なら ない”のように、動詞の子になる単語が増えるため、これらの単語を適切に扱う必要がある。

3.2.3.2 アライメント距離確率モデル

本研究では、先行研究のように複単語表現を事前に獲得して利用するのではなく、隣接する単語間の結合度を用いて複単語表現に対応する。こうすることで、扱う表現を限定することなく様々な表現に柔軟に対応できる上に、他の言語対にも容易に拡張することができる。

まず、複単語表現を構成する単語は結合度が強く、そのような二単語間では相手言語側でも同じもしくは近接するフ

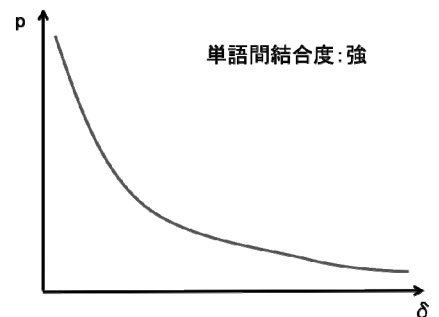


図 2:単語間結合度が強い場合の確率分布

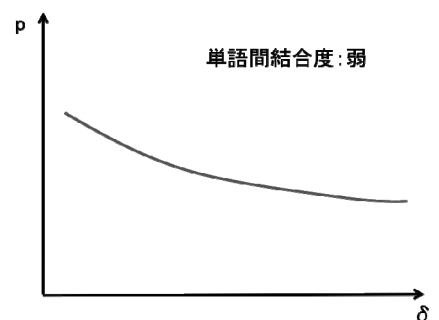


図 3:単語間結合度が弱い場合の確率分布

レーズと対応しやすいと仮定する。この仮定に基づき、隣接する二単語間の結合度を定義し、結合度に対して相手言語側での二単語の対応先の距離を次のように確率化する。

$$P(\delta | r(\text{bigram}))$$

なお、 $r(\text{bigram})$ はある二単語の単語間結合度であり、 δ は対応先の距離である。以下、対応先の距離をアライメント距離、上記確率をアライメント距離確率と呼ぶことにする。アライメント距離は、相手言語側で対応するフレーズの単語列上の最短距離とする。ただし、二単語が同じ対応に含まれる場合はアライメント距離を0とする。また、

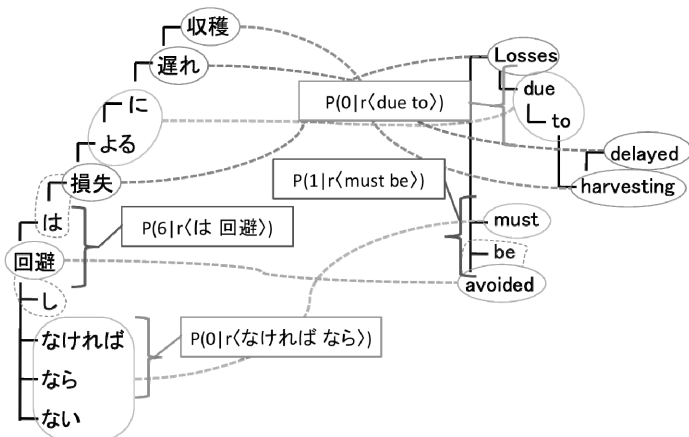


図 4:提案モデル概要

二単語に NULL 対応が含まれる場合は、1 単語目が NULL 対応、2 単語目が NULL 対応、両方 NULL 対応の 3 パターンに分けて扱う。なお、ここで隣接する二単語とは、単語列で連続かつ依存構造木上で連続もしくは兄弟の関係になっている二単語を指すこととする。

アライメント距離確率は、前述の仮定から、単語間結合度が強い場合はアライメント距離が小さいほど確率が高く、アライメント距離が大きいと確率は低いという図 2 のような分布になると考えられる。一方、単語間結合度が弱い場合はこの傾向は弱まり、図 3 のような分布になると考えられる。

図 4 に提案モデルの概要を示す。この図では日英の対訳文が依存構造木で表されており、破線でつながれたフレーズがアライメントを表す。また、アライメント距離確率の例をいくつか示している。例えば、“なければ なら” の二単語に着目する。この二単語はいずれも“must”と対応している。この二単語は“なければ なら ない”という複単語表現の一部であり、結合は強いいため、同じフレーズと対応する確率が高いと考えられる。一方、“は 回避” の二単語のアライメント距離は 6 であるが、この二単語は結合が弱いため、確率はそれほど低くならないと考えられる。

3.2.4 単語間結合度

隣接する単語間の結合度のスコアとしては、以下に説明する 3 種類のものを用いた。一つ目は、**Bigram 確率**である。

$$P(w_n | w_{n-1})$$

2つ目は順方向 **Bigram 確率**と逆方向 **Bigram 確率**の幾何平均をとったもの (**Bidirect**) である。こうすることで、より共起しやすい二単語のスコアを高くすることができる。

$$\left(P(w_n | w_{n-1}) P(w_{n-1} | w_n) \right)^{\frac{1}{2}}$$

3つ目は、両方向の **Bigram 確率**に加え、両単語の出現確率も組み合わせたもの (**Uni&Bi**) であ

る。これにより、コーパス中に高頻度で出現する機能表現に対応する。なお、実験では **Bigram** 確率に重みをおくため $\alpha=1/10$ とした。

$$\left(P(w_{n-1}) P(w_n)\right)^\alpha \cdot \left(P(w_n|w_{n-1}) P(w_{n-1}|w_n)\right)^{\frac{1}{2}}$$

上で述べたスコアは、アライメント実験でも使用する **JST** 日英論文抄録コーパス[6] (約 100 万文) から計算した。表 1 に、単語間結合度のスコアの例を示す。このうち、複単語表現を構成する“なければ なら”や“due to”などはいずれの方法でも高いスコアとなっている。一方、“による”は **Bigram** 確率のスコアは低い、逆方向 **Bigram** 確率や単語の出現確率も考慮すると高いスコアとなる。また、つながりの弱い“は 回避”や“be avoided”はいずれの方法でも低いスコアとなっている。

アライメント距離確率を計算する際、単語間結合度の値そのものに対して行うとスパースになってしまう。これに対処するため、単語間結合度をいくつかの範囲に分割して、その範囲内でアライメント距離確率の計算を行っている。この際、範囲に含まれる **Bigram** の頻度がアライメントを行うコーパス内で均等になるように分割を行っている。また、実際の確率分布に近づけるために、隣接する範囲の確率値を用いてスムージングを行っている。本研究では実際のデータに基づいてアライメント距離確率の計算を行ったが、確率分布を仮定してパラメータ推定を行う方法も考えられ、これについては今後の検討課題とする。

表 1: 単語間結合度のスコア例

	Bigram 確率	Bidirect	Uni&Bi
なければ なら	0.845376	0.602368	0.097682
ならない	0.591857	0.197442	0.042676
による	0.068922	0.261949	0.103429
は 回避	0.000044	0.000936	0.000240
due to	0.989265	0.191227	0.059808
be avoided	0.003486	0.045504	0.008718

3.2.5 アライメント実験

3.2.5.1 実験設定

実験に使用したコーパスは、前節で述べた **JST** 日英論文抄録コーパスである。このうち、文 ID 順に最初の 30 万文を用いて実験を行った。まず、日英それぞれの文に対して構文解析を行う。日本語に関しては、形態素解析器 **JUMAN** と構文解析器 **KNP** を利用した。英語に関しては、**nlparser** を用いて句構造解析を行った結果に対し、フレーズの主辞を定義するルールを適用することで単語依存構造に変換した。アライメントは中澤らのモデルをベースラインとし、これに提案モデルを組み込んだものと比較実験を行った。提案モデルとしては、単語間結合度を利用する方法と、単語間結合度は利用せずアライメント距離確率を $P(\delta)$ とする方法を用いた。単語間結合度を利用する場合は、範囲の分割数を 10 または 20 とした。

アライメントの評価には人手で正解を付与した 500 文に対して、次の式で計算される **Precision**、**Recall**、**Alignment Error Rate (AER)**を用いた。

$$\text{Precision} = \frac{|A \cap P|}{|A|} \quad \text{Recall} = \frac{|A \cap S|}{|S|} \quad \text{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

ただし、A はシステムの実出力 (図 5 の ■ または □)、S は必要な正解 (図 5 の濃い灰色のマス)、P は日本語の接尾辞や英語の冠詞などのように、あっても誤りではない正解 (図 5 の薄い灰色のマス) である。AER はアライメントの総合的な精度の良さを表す指標であり、数値が小さいほど精度が良いことを意味する。

3.2.5.2 結果と考察

実験結果を表に示す。まず、単語間結合度を利用せずにアライメント距離確率のみを計算する方法でも、ベースラインに比べ AER が 0.6 ほど改善している。これは、ベースラインシステムにおいて単語列上の情報がほとんど利用されていなかったためである。単語間結合度を利用した場合はさらに改善がみられ、単語間結合度を利用しない場合に比べ最大で AER が 0.3 ほど改善している。全体的に Precision が上昇する傾向がみられ、結合の強い単語同士がまともに対応づけられたことによりアライメント誤りが抑制できていると考えられる。

図 5 に、提案手法によってアライメントが改善された例を示す。ベースラインでは、“large amount of ⇔ 多量の” の対応のうち、“of” のみ対応先が誤っている。提案手法ではこれらの対応先の距離が遠いため “of ⇔ を” の対応が外れ、“of” が正しく “多量の” に対応付けられている。

逆に提案手法でもアライメントを改善できなかった例としては、構文解析の誤りにより複単語表現を構成する単語が依存構造木上で不連続になっているケースがあった。この問題に関しては、構文解析時にも単語間の結合を考慮し、結合の強い単語同士が依存構造木上で連続になるように

表 2: アライメント実験の結果

	Prec.	Rec.	AER
ベースライン	91.40	82.67	12.74
単語間結合度なし	91.47	84.07	12.10
Bigram 確率 (10 分割)	91.84	83.80	12.06
Bigram 確率 (20 分割)	91.85	83.89	12.00
Bidirect (10 分割)	91.41	84.03	12.15
Bidirect (20 分割)	91.91	84.11	11.86
Uni&Bi (10 分割)	91.82	84.18	11.87
Uni&Bi (20 分割)	91.99	84.13	11.81

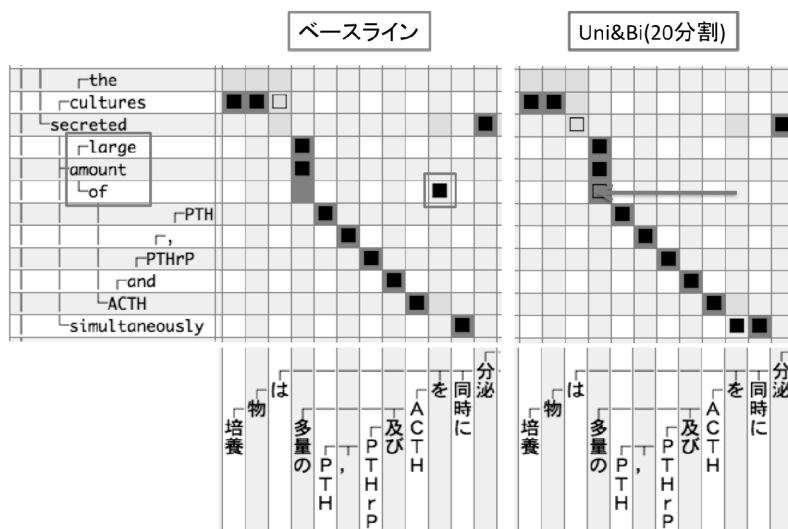


図 5: アライメントが改善された例

すれば改善が期待できる。また、アライメント結果を利用して構文解析を改善する方法も考えられる。他のアライメント誤りの例としては、孤立機能語が他の大きな対応に吸収されてしまうような副作用もみられた。この問題については、単語の品詞など文法情報を利用することで改善できると考えられる。

3.2.6 おわりに

本研究では、複単語表現のアライメントを改善するため、単語間結合度が強い単語対ほど近接するフレーズに対応されやすくなるようなモデルを提案した。実験の結果、ベースラインに比べアライメント精度は改善し、特に複単語表現を構成する単語が誤って遠くの単語と対応するアライメント誤りを抑制する効果を確認した。

今後は、提案モデルを日中など他の言語対に適用してアライメントを行い、効果を調べる予定である。また、翻訳精度を改善する方法の検討も今後の課題である。具体的には、翻訳において複単語表現を過不足なくカバーする適切なフレーズが選択されるような枠組みを検討する必要がある。

参考文献

- [1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In HLT-NAACL2003: Main Proceedings, pp. 127-133, 2003.
- [2] Patrik Lambert and Rafael Banchs. Grouping Multi-word Expressions According to Part-Of-Speech in Statistical Machine Translation. In Proceedings of the EACL Workshop on Multi-Word-Expressions in a Multilingual Context. Trento, Italy. pp. 9-16, 2006.
- [3] Zhixiang Ren, Yajuan Liu, Jie Cao, Qun Liu, and Yun Huang. Improving statistical machine translation using domain bilingual multiword expressions. In Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, pp. 47-54, 2009.
- [4] Zhanyi Liu, Haifeng Wang, Hua Wa, and Sheng Li. Improving statistical machine translation with monolingual collocation. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 825-833, 2010.
- [5] Toshiaki Nakazawa and Sadao Kurohashi. Alignment by bilingual generation and monolingual derivation. In Proceedings of COLING 2012, pp. 1963-1978, 2012.
- [6] Masao Uchiyama and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 72-79, 2003.

3. 3 テキスト中の専門用語から Wikipedia への対応付けにむけて

静岡大学 綱川 隆司

梶 博行

3.3.1 はじめに

Wikipedia は Web 上の百科事典で、多くの人々の共同作業によって編集されている。規模的には冊子体の百科事典を大きく凌駕し、記述内容の質や信頼度も向上していることから、毎日、世界中で多くの人々がアクセスしている。全体が一つのハイパーテキストとなっており、関連する項目を次々とたどれることが大きな特長である。

Wikipedia 記事以外の一般文書から Wikipedia 記事にリンクをはることを“wikification”という。Wikification が実現されると、任意の文書からワンクリックで Wikipedia 記事にとぶことが可能になり、Wikipedia の利便性が更に向上する。Wikipedia は科学技術分野の記事も充実しつつあるので、特許文書の wikification により審査官や発明者、技術者が特許の内容を理解する助けになると期待される。本稿では、特許文書の wikification の実現に向け、特許文書固有の課題を明らかにする。

3.3.2 Wikification とその技術課題

Wikipedia 記事には、記事中の重要語句をアンカーとし、それについて説明した記事に内部リンクと呼ばれるハイパーリンクがはられている。Wikification は、一般のテキストに対して Wikipedia 記事のように Wikipedia 記事へのハイパーリンクを付与するタスクである[1]。

Wikification は(1)アンカーの決定、(2)リンク先記事の決定の2つのステップから構成される。予め Wikipedia 記事全体から抽出した内部リンクデータ（アンカー文字列とリンク先記事の組）を利用するが、以下のような課題がある。

(1) アンカーの決定

Wikipedia 記事中でアンカーとして採用されたことのある文字列がアンカーの候補となる。しかし、それらの中には年号としての数字、略称等、あるいは文脈によっては重要語とならない語が含まれるため、入力テキストの文脈において重要な語であるかどうか判断することが必要である。また、Wikipedia 全体でアンカーとして採用されたことはないが入力テキスト中では重要な語をどう扱うかという問題もある。

Wikipedia の内部リンクデータによらない方法として、tf-idf や C-value [2] 等の指標を用いた重要語抽出方法が挙げられる。

(2) リンク先記事の決定

Wikipedia の内部リンクデータによると、一つのアンカー文字列に対しリンク先記事の候補が複数存在することも多い。候補の中から入力テキストの文脈における意味に対応する適切な記事を選択することが必要である。そのために語義曖昧性解消 (Word Sense Disambiguation) の手法が利用される。例えば、アンカーの周辺文脈と、その語句の各意味に対応する情報（語句の説明文等）

の類似性を測り、最も近い意味を選択する。曖昧性解消に用いる手掛かりによって、リンク先記事の決定方法は以下の二つに大別される。

- ローカルな方法：周辺文脈などを用いて、アンカーごとに独立に曖昧性解消を行う。テキスト全体として一貫性のない解が得られる可能性がある。
- グローバルな方法：アンカーごとの手がかりに加え、テキスト内の各アンカーのリンク先相互の関係なども考慮して同時に曖昧性を解消する。テキスト全体として一貫性のある解が得られることが期待されるが、一つのアンカーの意味を誤って判定した場合、他のアンカーにも誤りが波及する可能性がある。

3.3.3 Wikificationに関する主な研究

Mihalcea and Csomai [1] は、アンカーを決定するため、重要度のランク付けを行った。ランク付けの方法として、従来から用いられる tf-idf、 χ^2 独立検定のほか、wikification に特化した方法として keyphraseness を提案し、従来方法より有効であることを示した。ある語 w の keyphraseness は、次式で定義される。

$$\text{Keyphraseness}(w) = (\text{w がアンカーとして出現した文書数}) / (\text{w が出現した文書数})$$

また、リンク先記事の決定に関しては、知識に基づく方法とデータに基づく方法の二つを比較した。前者は Lesk アルゴリズム [3] の応用で対象文書とリンク先記事の類似性を用い、後者は Wikipedia から収集したアンカーの周辺文脈とリンク先記事をトレーニングデータとしてリンク先記事の分類器を構築した。

Milne and Witten [4] は、アンカーがある記事にリンクされる確率(commonness)を Wikipedia の内部リンクデータから求めるとともに、アンカーを含む記事とリンク先記事の記事に共通してリンクする記事の数から求めた関連性指標を提案し、それらを組み合わせてリンク先記事の曖昧性解消を行った。アンカー w に対するリンク先記事 t の commonness $P(t|w)$ は次式で定義される。

$$P(t|w) = (t \text{ にリンクするアンカー } w \text{ の出現回数}) / (\text{アンカー } w \text{ の出現回数})$$

アンカー w に対して $P(t|w)$ が最も大きい t にリンクすること、すなわち最尤推定による性能は曖昧性解消評価のベースラインとして用いられる。

Cucerzan [5] は、アンカーを含む記事のカテゴリと内部リンクを素性とし、アンカーとリンク先記事候補の素性ベクトルを比較することで、リンク先記事の曖昧性解消を行った。Ratinov [6] はさらにこれらを組み合わせ、アンカー周辺の文脈や記事の素性ベクトルをローカルな素性として、またアンカーとリンク先記事の共通リンク先文書数を用いた二つの関連性指標をグローバルな素性として機械学習を行い、曖昧性解消の精度を改善した。

3.3.4 特許文書と wikification

特許文書の wikification においてアンカーとすべき語の多くは専門用語と考えられる。wikification を行うにあたっては専門用語として重要な語を抽出することが必要になる。例えば、

(1) つづいて、CPU 2はキャッシュ・システム 3に対してリード要求をおこない、タグRAM 3 1にリード・アドレスを供給する。

という文には、“CPU”、“キャッシュ”、“リード”、“RAM”、“アドレス”という専門用語が出現するが、どの語をアンカーとするかが課題となる。

また、特許文書に出現する専門用語はその専門分野においては曖昧性をもたないことが多いが、Wikipedia 全体としてみると必ずしもそうとは限らない。例文(1)中の“キャッシュ”や“リード”に対して、リンク先記事候補として以下のようなものが存在する。

キャッシュ：『キャッシュ (コンピュータシステム)』『キャッシュメモリ』『現金』
『ウェブアーカイブ』等、12 記事

リード：『リード (楽器)』『リードギター』『リード (カメラ)』『リード』等、36
記事

これらの候補のうちどれが適切か判断するためには、3.3.2 の(2)で述べた方法のほか、この文を含む特許と各候補記事の専門分野と比較する方法が考えられる。

3.3.5 特許文書への wikification の適用にむけた事前調査

特許文書に対して wikification を適用するために、特許文書中で wikification が必要な専門用語がどの程度の割合で存在するか、またそれらの曖昧性がどの程度あるかについて調査した。

3.3.5.1 調査方法

NTCIR8 PATMT 訓練データの 2003-2005 年度のテキストからランダムに選択した 10 万文の日本語文を調査対象とした。各文に含まれる文字列で、既知のアンカーと一致するものを抽出し、それらに対して

- Keyphraseness の値
- リンク先記事候補と Wikipedia 中でリンクされた頻度

を集計した。Keyphraseness 閾値 θ として 0 から 0.1 までのいくつかの値を設定し、keyphraseness が θ 以上のアンカー数と各アンカーのリンク先記事候補数を求めた。

3.3.5.2 調査結果

(1) アンカー候補

表 1 に、keyphraseness 閾値 θ を 0 から 0.1 まで変化させたときの 1 文あたりのアンカー数と 1 アンカーあたりのリンク先記事候補数の平均を示す。

$\theta=0$ のときは、日本語 Wikipedia に出現したことがあるすべてのアンカーに対して Wikipedia 記事を対応付けた場合に相当する。Keyphraseness の小さい語は重要でない語であることが多く、アンカーとして適さないケースが見ら

表 1 1 文あたりのアンカー数と
1 アンカーあたりのリンク先記事候補数

Keyphraseness 閾値 θ	1 文あたりのアンカー数	1 アンカーあたりのリンク先記事候補数
0	8.590	5.045
0.01	3.640	5.757
0.03	2.719	5.027
0.05	2.130	4.976
0.07	1.777	4.215
0.1	1.413	4.03

れた。次の例文において、各下線部がアンカー、その右の括弧内の数字が keyphraseness の値を示す。

(2) ダイオード(0.514)の6 4のアノード(0.512)は、トランジスタ(0.496)6 3のエミ(0.0013)ッタと接続(0.0006)され、その(0.00002) カソード(0.463)は、抵抗(0.014)2 3 および6 1の接続点に接続される。

“ダイオード”や“アノード”といった専門用語の keyphraseness の値が比較的高い
のに対し、“接続”や“その”といった、まれにしかアンカーとして扱われない語の keyphraseness の値は極めて低い。したがって、 θ に0より大きい値を設定することでアンカーとして不適切な語をある程度除外することができる。 $\theta=0.01$ とした場合に1文あたりのアンカー数は半数以下の3,640個となり、以下 θ の値を大きくするごとにアンカー数は減少する。また、keyphraseness の大きい語の方がリンク先記事候補数が少ないという傾向も明らかになった。

(2) リンク先記事候補

例文(2)中のアンカー候補に対するリンク先記事候補と、各記事候補に対する commonness $P(t|w)$ の値を表2に示す。Commonness は文の内容によらない値であるため、アンカー周辺の文脈によって適切なリンク先記事が変わる場合、他の語義曖昧性解消手法を適用する必要がある。例えば、文(2)中の“抵抗”に対しては、リンク先記事として『抵抗器』が最も適切と考えられるが、commonness の値は全候補中4位である。

表2 文(2)中のアンカーとリンク先記事候補

アンカー w	リンク先記事候補 t	$P(t w)$
ダイオード	ダイオード	0.995
	発光ダイオード	0.005
アノード	アノード	1
トランジスタ	トランジスタ	0.998
	バイポーラトランジスタ	0.002
エミ	中原絵美	1
接続	接続	0.846
	乗換駅	0.154
その	マールス	1
カソード	カソード	1
抵抗	抵抗	0.488
	電気抵抗	0.192
	抗力	0.182
	抵抗器	0.103
	抵抗制御	0.010
	造波抵抗	0.005
	レジスタンス運動	0.005
	転がり抵抗	0.005
抵抗値	0.005	

図1に、 $\theta=0.01$ および $\theta=0.1$ のときのアンカー候補を対象としたリンク先記事候補数(1, 2, ..., 9, 10)のヒストグラムを示す。 $\theta=0.01$ の場合でリンク先記事候補が1つのみのアンカーが70,468個あるのに対し、2つのものが54,052個、3つのものが42,616個ある。 $\theta=0.1$ とするとアンカー候補は減少し、リンク先記事候補数が多いアンカーほど減少率は大きい。しかし、相当数のアンカー候補が複数のリンク先記事候補をもち、特許文書の wikification においてもリンク先記事の決定が大きな課題となることがわかる。

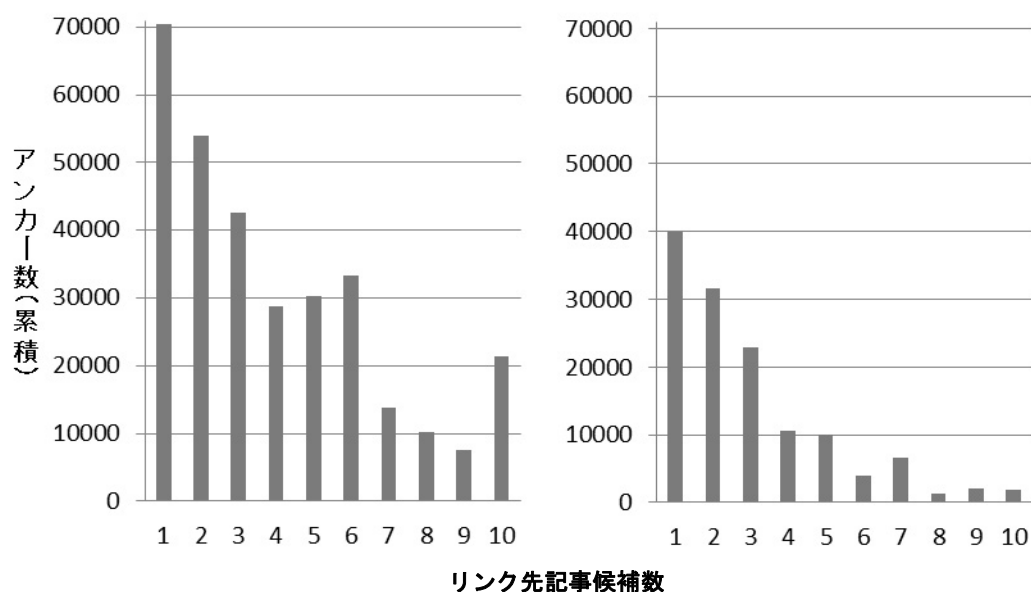


図 1 リンク先記事候補数ごとの累積アンカー数 (左: $\theta=0.01$ 、右: $\theta=0.1$)

3.3.6 おわりに

特許文書に対する wikification の実現に向けて、特許文書中からアンカー候補を抽出し、それらに対応するリンク先記事候補数を調査した。その結果、特許文書においてもアンカーのリンク先記事の曖昧性解消が大きな課題であることが明らかになった。また、アンカーの keyphraseness 指標を用いることにより不適切なアンカー候補をある程度除外することができ、その結果リンク先記事の曖昧性解消の課題が軽減されることを確認した。今後の課題は、特許文書に適したアンカーの決定方法およびリンク先記事の決定方法について、F タームや国際特許分類などの利用を含めて検討することである。

参考文献

- [1] Mihalcea, R. and A. Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 233–242.
- [2] Frantzi, T.K. and S. Ananiadou. 1996. Extracting nested collocations. In *Proceedings of 16th International Conference on Computational Linguistics*, pages 41-46.
- [3] Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24-26.
- [4] Milne, D. and I. H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the Wikipedia and AI Workshop of AAAI*, pages 25-30.
- [5] Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of ACL 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708-716.
- [6] Ratnov, L., D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375-1384.

4. コンパラブルコーパスによるドメイン適応

東京大学 範 暁蓉

愛媛大学 二宮 崇

4.1 はじめに

近年、機械翻訳の分野では統計的機械翻訳 (SMT) の研究が盛んに行われている。SMT は、ある分野に存する大量の対訳データから翻訳モデルを学習する手法である。しかし、ある分野 (例えば新聞記事分野) で学習された SMT モデルは異なる分野 (例えば化学、電子などの分野) で利用する時に、性能が大きく低下するという問題がある。元々翻訳の対象としていた分野 (新聞記事分野) のデータ分布と新しい翻訳対象の分野 (化学分野や電子分野) のデータ分布が違うことが原因である。分野によって異なるデータ分布の問題を解決する技術は一般にドメイン適応 (domain adaptation) と呼ばれ、機械学習から自然言語処理の諸タスクにおいて広く盛んに研究されている。本研究では機械翻訳のためのドメイン適応について研究を行う。ドメイン適応において、元々解析の対象としていた分野は「ソースドメイン (source domain)」と呼ばれ、新しい解析対象の分野は「ターゲットドメイン (target domain)」と呼ばれる。

本稿では、まず、ソースドメインとターゲットドメインが異なる時、SMT 性能に影響する三つの原因について述べる。次に従来のパラレルコーパスとコンパラブルコーパスによるドメイン適応の研究について述べる。最後に提案手法について述べ、その実験結果を報告する。機械翻訳においては、翻訳元言語は「原言語 (source language)」と呼ばれ、翻訳先言語は「目的言語 (target language)」と呼ばれる。本稿でも原言語と目的言語という用語を用いる。

本稿の構成は以下のようになっている。4.2 節では、ドメインが異なる時、SMT 性能に影響する三つの原因とパラレルコーパスによる分野適応手法について説明する。4.3 節では、コンパラブルコーパスによるドメイン適応を提案し、実験の結果について述べる。4.4 節で本稿の主旨をまとめ、今後の課題について述べる。

4.2 背景

4.2.1 異分野において SMT の性能が低くなる原因

ソースドメインから学習された SMT モデルをターゲットドメインにおいて利用するとき、性能が低くなる原因は主に三つある。

(1) 未知語 (OOV) 問題

各分野に対して、分野特有の語彙や表現が存在する。この特有の語彙や表現はソースドメインから学習された SMT に対して、未知語 (OOV: Out-of-Vocabulary) となっている。

(2) 多義語問題

複数の訳語を持つ多義語が多く存在している。翻訳モデルにおいてもっとも適切な訳語を選択す

表 1 Philipp Koehn の分野適応の実験の結果

方法	BLEU(%)
Large out-of-domain training data	25.11
Small in-domain training data	25.88
Combined training data	26.69
Two translation models	27.64

ることになるが、ソースドメインから学習された翻訳モデルはターゲットドメインにおける分布と異なるため、正しい訳語が選択されないという問題が生じる。例えば、“本”という日本語には一般的に「book, piece, this」などの英語の訳語がある。ソースドメインから学習された翻訳モデルの中に、訳語候補「book, piece, this」に対して、対訳確率は $p(\text{book}) > p(\text{piece}) > p(\text{this})$ の順になる。しかし、ターゲットドメインが特許分野となる場合、対訳確率は $p(\text{this}) > p(\text{piece}) > p(\text{book})$ の順になる可能性が高い。これは特許においては「本手法」が「this method」などに翻訳されることが多いためである。

(3) 新訳語問題

最後に、ある単語に対して、ターゲットドメインに対する訳語がソースドメインに存在しないことがあり、これが SMT の性能低下の原因となる。

Irvine ら (2013) の調査研究によると、OOV 問題と新訳語問題が、ドメイン適応時に SMT の性能が低下する主な原因となる。

4.2.2 パラレルコーパスによるドメイン適応

ターゲットドメインに対する SMT の性能を改善する最も有効な方法はターゲットドメインの対訳コーパスを使用することである。ターゲットドメインのデータが大量に存在する場合には、ターゲットドメインのデータのみを用いて学習すればよいことになるが、ドメイン適応における多くの問題では、ソースドメインに比べてターゲットドメインのデータが圧倒的に少ないことが多い。そのため、ターゲットドメインのデータとソースドメインのデータを併用する手法がよく研究されている。

Philipp Koehn (2007) らはソースドメインのデータとターゲットドメインデータを併用するドメイン適応の実験を行った。表 1 はその結果を示している。表によると、データを簡単併用 (Combined training data) すると、BLEU 値 (%) は 1.58 上がっている。モデル (Two translation models) を併用すると、BLEU 値 (%) は 2.53 上がっている。

しかし、学習用の対訳データは一般的に、数 10 万文から 100 万文程度のペアが必要である。このような量の対訳データ、特に特殊ドメイン (例えば、特許分野) の対訳データを収集することは非常に困難である。Smith (2010) らは Wikipedia から対訳データを収集する手法の研究

を行った。彼らの実験によると、21%の原言語にたいして、英文の訳語が存在している。一方、コンパラブルコーパスはパラレルコーパスよりも大量に存在している。そして、コンパラブルコーパスによるドメイン適応の研究も盛んに行われている。

4.3 コンパラブルコーパスコーパスによるドメイン適応

4.3.1 関連研究

これまでのコンパラブルコーパスによるドメイン適応では、主に対訳辞書構築のための資源としてコンパラブルコーパスがよく用いられている。Hal ら (2011) はコンパラブルコーパスからターゲットドメインの OOV の訳語を推定し、獲得した対訳辞書をソースドメインから学習された SMT の翻訳モデルに加えると、BLEU 値 (%) が 0.5 から 1.5 ほど上がっている。しかし、この手法では既存の対訳辞書を必要としており、また、獲得した対訳辞書を SMT システムの翻訳モデルに加える際の対訳確率の設定も複雑である。

Ravi ら (2011) は原言語の単言語コーパスと目的言語の単言語コーパス(それぞれ対訳の関係にない単言語コーパス)を利用して学習する統計的機械翻訳の手法を提案した。彼らの手法では、目的言語を生成し、生成された目的言語を暗号化することにより原言語が生成される確率的生成モデルを提案しており、暗号解読 (deciphering) により、原言語文 f から目的言語文 e への翻訳を実現する。次式は暗号解読モデルの式を表している。

$$P_{\theta}(f) = \sum_e P(e)P_{\theta}(f|e) \quad (1)$$

目的言語文 e が生成され、 e が暗号化された結果、原言語文 f が生成される確率的生成モデルとなっている。 $P(e)$ は目的言語の単言語コーパスから言語モデルとして学習され、暗号化文の確率 $P_{\theta}(f)$ を最大化することによりパラメータ θ が学習される。

$$\hat{\theta} = \operatorname{argmax}_{\theta} \prod_f P_{\theta}(f) = \operatorname{argmax}_{\theta} \prod_f \sum_e P(e)P_{\theta}(f|e) \quad (2)$$

目的言語の言語モデル $P(e)$ と原言語の単言語コーパスが与えられたとき、翻訳確率 $P_{\theta}(f|e)$ を調整し、EM アルゴリズムで $P_{\theta}(f)$ が最大となる翻訳モデルを探索することによりパラメータ θ が得られる。翻訳は、 $P(e)P_{\theta}(f|e)$ を最大化する e を求めることにより行われる。

$$\hat{e} = \operatorname{argmax}_e P(e)P_{\theta}(f|e) \quad (3)$$

Ravi らの手法は、英文を暗号化した簡単な暗号文に対して 82.5%の精度で元の英文に解読できている。しかし、この手法は、Google Translate で英語をスペイン語に翻訳して得られたパラレルコーパスから学習された SMT の翻訳精度と比較すると、BLEU 値 (%) が 40 下がっ

ており、翻訳手法として、まだ実用に供せられる精度には至っていない。Ravi らの手法は (1) 処理できる語彙数が少ない (500 ぐらい)、(2) 短い文 (4 トークンの文) しか処理していない、(3) 計算コストが高いという問題点がある。

4.3.2 ターゲットドメインにおける暗号解読モデルの学習

暗号解読による機械翻訳モデルにおいて $P_\theta(f|e)$ を推定するとき、まず、原言語 f の訳語候補を決定する必要がある。すべての単語を訳語候補とすると非常に大きな計算コストを要するため、まず訳語候補の数を減らすことが重要である。そのために既存の辞書、対訳資源を利用することが考えられるが、Ravi らは、EM での学習中に、単語の頻度と学習中の翻訳確率に基づいて、訳語候補を徐々に増やす手法を提案している。

ソースドメインのデータ分布とターゲットドメインのデータ分布が異なるが、この二つのデータ分布はある共通分布を持っている (Hal ら 2006)。本研究は、ソースドメインの平行コーパスから学習された翻訳モデルに基づいて、ターゲットドメインのコンパラブルコーパスから暗号解読モデルを学習する手法を提案する。

本研究では、Ravi ら (2011) の研究と同じ確率的生成モデルを採用する。Ravi ら (2011) の確率的生成モデルは以下になる。

- (1) ターゲット言語の文 e を確率 $P(e)$ で生成する。
- (2) 隣接する二つの単語間に NULL 文字を挿入する。
- (3) 文列 e 中の各単語 e_i (NULL を含む) からソース文 f の単語 f_i が翻訳確率 $P_\theta(f_i|e_i)$ で生成される。
- (4) ソース文 f において隣接する二つの単語 f_{i-1} と f_i の順を変更する。
- (5) NULL を削除して、ソース文 f が出力される。

本研究では Ravi ら (2011) の研究と三つの異なるところがある。

(1) 研究の目標

Ravi ら (2011) の研究では、単言語コーパスを用いた SMT モデル (翻訳モデルと部分並び替えモデル local reordering model) の学習を目的としておりドメイン適応のための手法ではない。本研究は、ソースドメインにおいて平行コーパスを用い、ターゲットドメインにおいてコンパラブルコーパスを用いて暗号解読モデルを学習し、ターゲットドメインのための翻訳モデルを得ることを目的としている。

(2) 訳語候補の選択基準

ある単語 w の訳語候補が二つの部分で構成されている。 w がソースドメインに含まれていると、まず、ソースドメインから学習された翻訳モデルの中の訳語が訳語候補となる。このとき、 w に対する訳語候補の数を一定数まで (今回は 50 とした) とする。 w が未知語のときはすべての未知語が訳語候補になる。未知語の場合は大きくても 150 個の訳語候補までとする。

(3) コンパラブルコーパス

Ravi ら (2011) の手法は対訳対応のない単言語コーパスを仮定しているが、実際の実験に使用された単言語コーパスはパラレルコーパスから生成されているため完全な単言語コーパスとはなっておらずコンパラブルコーパスに近い性質を持っていると考えられる。本研究では、完全なコンパラブルコーパスを使用した。

4.3.3 実験

本節では、提案手法の有効性を確認するためターゲットドメインの翻訳モデルの抽出実験の手順および結果について述べる。

実験の手順を以下に示す。

- (1) パラレルコーパスから SMT 翻訳モデルを学習する。
- (2) ターゲットドメインの単言語コーパスから目的言語の言語モデルを学習する。
- (3) 学習された SMT 翻訳モデルから単語の対訳データを抽出する。
- (4) 対訳データの閾値を設定し、原言語の訳語候補を決定する。
- (5) コンパラブルコーパスから両言語の未知語 (OOV) を抽出し、頻度による訳語候補を決定する。
- (6) EM アルゴリズムでターゲットドメインの翻訳モデル $P_{\theta}(f|e)$ を学習する。
- (7) ターゲットドメインの翻訳モデル $P_{\theta}(f|e)$ を SMT システムに入れ、評価実験を行う。

実験に使用したパラレルコーパスとコンパラブルコーパスは次のとおりである。

(1) パラレルコーパス

日英新聞記事対応付けデータ(JENAAD)¹は元ドメインのパラレルコーパスとして使用した。前処理された JENAAD の 149,165 文対のうち、1000 文をチューニングデータとして使用した。残りの文を訓練データとした。

(2) コンパラブルコーパス

二つのコンパラブルコーパスを使用した。一つは Wikipedia から、情報処理に関する日本語と英語の文から生成されたコーパスである。日本語が 9,732 文、129,539 語、4,483 トークン、英語が 8,914 文、95,738 語、6,235 トークンである。もう一つは、グーグル・パテント・サーチで生成された機械翻訳に関する日英特許コンパラブルコーパスである。日本語が 8,658 文、147,137 語、2,763 トークン、英語が 9,585 文、170,343 語、3,916 トークンである。

日本語文に対して、全角符号は半角に変更する前処理と単語分割を行った。英単語に対して、トークナイザ (tokenizer) と小文字化 (lowercase) は行わなかった。機械翻訳システムには、通常のフレーズベース機械翻訳を使用した。以下に使用したツールの一覧を示す。

- 統計的機械翻訳エンジン : Moses (Koehn ら, 2007)

¹ http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/jea/index-ja.html

独立行政法人情報通信研究機構作成

表 2 提案手法の実験の結果

手法	BLEU(%)	
	コンパラブルコーパス(Wiki)	コンパラブルコーパス(特許)
パラレルコーパスから学習した翻訳モデルによる翻訳	12.13	9.75
コンパラブルコーパスから学習した翻訳モデルによる翻訳	10.47	9.32
翻訳モデルの線形組み合わせによる翻訳	12.40	10.08

- 単語アライメント：GIZA++²
- 言語モデル：KenLM (Kenneth Heafield, 2011)
- 日本語の単語分割：Mecab³
- 英語の前処理：Moses の tokenizer.perl、lowercase.perl と escape-special-chars.perl⁴

4.3.4 実験結果

次の比較実験を行った。

- (1) ソースドメインから学習した翻訳モデルを用い、テストデータを翻訳する。
- (2) ターゲットドメインの翻訳モデル $P_{\theta}(f_i|e_i)$ を用い、テストデータを翻訳する。
- (3) ソースドメインから学習した翻訳モデルとターゲットドメインの翻訳モデルを線形補間し、テストデータを翻訳する。

テストデータについて、コンパラブルコーパス(Wiki)に対して、今回手で作った 100 文対のテストデータを使用し、コンパラブルコーパス(特許)に対して、原文とグーグル・パテント・サーチでの翻訳文から 100 文対を選択し、テストデータを使用した。翻訳方法は日本語から英語とし、テストデータ中の日本語文を翻訳した。

表 2 は今回の比較実験の結果である。翻訳結果は機械翻訳の自動評価尺度 BLEU 値を用いた。実験の結果から、ターゲットドメインの翻訳モデルだけでは性能が向上しないことがわかったが、ソースドメインの翻訳モデルと線形補間をとることにより性能が向上されることが確認できた。

4.4 まとめ

本稿では、暗号解読モデルによる機械翻訳手法の調査を行い、この手法を用いてコンパラブルコーパスからターゲットドメインの翻訳モデルを学習する手法を提案した。実験により、この手法はドメイン適応に有効であることがわかった。しかし、コンパラブルコーパスの精度と EM アル

² <http://www.statmt.org/moses/giza/GIZA++.html>

³ <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

⁴ <http://www.statmt.org/moses/?n=moses.baseline>

ゴリズムの計算量のため、今回の実験では、単語対応の翻訳モデルだけを構築し、その結果、フレーズベース機械翻訳システムに対して、精度を大きく改善することは出来なかった。今後は、精度向上のため、フレーズ対応の翻訳モデルの学習に向けて研究を行いたい。

参考文献

Ann Irvine, John Morgan, Marine Carpuat, Hal Daum' e III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics (TACL)*.

Hal Daum' e III and Daniel Marcu. 2006. Domain adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

Hal Daum' e III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Heafield, Kenneth. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, StatMT'07*, pages 224–227, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 12–21, Portland, Oregon, USA, June. Association for Computational Linguistics.

Zellig Harris. Distributional Structure. Katz, J.(ed.) The Philosophy of Linguistics. Oxford University Press, pages 26-47, 1985.

5. 英文作成システムのレビューとシステム構想の提案 -----MT への応用を目指して-----

秀明大学 宮澤信一郎 山梨英和大学 江原 暉将
創価大学 岡田 勇 秀明大学 宮崎 瑞之
早稲田大学 Dawn L. Miyazaki

5.1. はじめに

英文作成支援システムとは、コンピュータによって人間の英文作成を支援するシステムである。この点、自動的に翻訳を行う機械翻訳とは異なる。英文作成支援システムは、機械翻訳がまだ完全でない現況下において、その重要性は変わらない。特に大量の対訳コーパスを利用できる現代においては文作成支援システムは新しい段階に入っているといえる。我々は、将来的に英文作成支援システムと機械翻訳システムを相互に補完するものと考えている。

本論では初めに既存のシステムのレビューを行い、それらを踏まえて、次に我々が考える新方式の英文作成支援システムを提案する。

5.2. 英文作成システムのレビュー

5.2.1 英文作成支援システムのタイプと具体例

ここでは既存のシステムを6つのタイプに分類する、6つのタイプとはテンプレートタイプ、キーワード・構文タイプ、検索エンジン活用タイプ、FEPタイプ、特定用途タイプ、教育支援タイプである、

5.2.1.1 テンプレートタイプ

- (1) 東芝ソリューション「The 翻訳 2009 ビジネス」の「英文作成テンプレート」や、富士通「ATLAS 翻訳パーソナル 2007」の「おまかせ記文」、アルクの「カクスケ」等である、
- (2) 「英文作成テンプレート」では、はじめに対象となる分野を選択すると用意されたテンプレートが表示される。いくつかの箇所が選択可能になっており、ダブルクリックすると候補が表示されるので、選択するとテンプレートの固有の部分の単語が埋められて英文が完成する。

5.2.1.2 キーワード・構文タイプ

キーワード・構文タイプとしては ESCORT や SCOPE などがある,

- (1) ESCORT (2) (3) は名古屋大学のグループが行っている英文作成支援ツールのプロジェクトである。ESCORT は英語キーワードを入力するタイプである。複数のキーワードを入力することができる。その際、単にキーワードを含む文を出力するのではなくキーワード間に構文的な関係が存在する英文を出力する点が特徴である。また、提示される英文は、キーワード間の関係の種類に応じてグループ化される。コーパスは英文論文から取り出された文を格納しており、用途として研究者が英語で論文を作成する場面を想定している。2008 年の時点で、約 71 万文の英文コーパスを有している。ESCORT は下記 URL で実際に運用されている。

<http://escort.itc.nagoya-u.ac.jp/>

使いかたは、想定する語順の通りに英語キーワードを入力すると、その順番にキーワードを持った英文が表示される。また単語の代わりに -n (名詞), -v (動詞), -i (前置詞) のようにして品詞を使用することもできる。単語を自動的に補って検索するには、コストと称するワイルドカードを設定することも出来る。

- (2) 同グループは、フレーズ検索システム SCOPE も公開している。

<http://scope.itc.nagoya-u.ac.jp/>

SCOPE は、英語論文において特定の章構成に頻出する入力英単語を含むフレーズを検索できる。検索には、検索したい英単語を一つ以上入力する。次に論文の標準的な構成である「序論」、「関連研究」、「提案手法」、「評価実験」、「結論」の 5 つの章構成で自分が英作したい部分の章構成を選択する。そして検索ボタンをクリックすることによりフレーズを検索できる。

5.2.1.3 検索エンジン活用タイプ

検索エンジン活用タイプとしては「早稲田大学グループのシステム」や TransAid や「安藤らのシステム」などがある,

- (1) 早稲田大学グループが行っているシステムがある (4)。検索エンジン特有の機能を利用するシステムである。検索エンジンは GoogleAPI を使用している。作成した英文の汎用性を調べるための様々な検索テクニックをシステム側で自動的に構築することによって、語彙の組み合わせの相性を検討する機能や、対訳コーパスから効率的に訳語の用法を発見できる機能を実現するシステムを提案している。以前は Web 上にシステムが公開されていたが、現在は見当たらない。
- (2) TransAid(5)(6)は、日本語の文章と、それを市販の機械翻訳システムで翻訳した結果を入力と

する。その翻訳結果を用いて、インターネット・コーパスから例文を抽出し、先の機械翻訳システムの出力を訂正したり洗練したりして、特定の目的に合った英語にするシステムである。

- (3) 安藤らのシステム(7)(8)は、検索エンジンを用いたシステムである。Google Web APIs を使って検索式を自動生成する。さらに検索された結果からヒット数や用例文などを抽出し、ヒット数に基づくテーブルやグラフ表示、用例の KWIC 表示など英作文に利用しやすい形で提示する方式である。

5.2.1.4 FEP タイプ

- (1) 「英作文メイキング」(9)(10)(11)(12)(13)は NEC の研究グループのプロジェクトで製品化された。入力された日本語を、かな漢字変換と同様のインタフェースでインタラクティブに英語に変換していくシステムである。単語および熟語の日英辞書引き機能があり、対話的な日英構文変換機能によりユーザの英文作成を支援する。ユーザは訳語選択と単文範囲の変換をインタラクティブに繰り返して英文を作成する。任意のアプリケーションプログラムと組み合わせて使用できるフロントエンドプロセッサ型ツールでもある。

5.2.1.5 特定用途タイプ

- (1) 英文アブストラクト作成支援システム(15)は、英文アブストラクト作成支援に特化したシステムである。具体的には次のようなアブストラクト作成に特化した機能を持っている。

- ・ 作成目的および論理展開タイプに応じたアブストラクト用例の検索機能
- ・ 日本人がおかしやすい文法エラーと正解例の参照機能
- ・ 論文アブストラクトのテンプレート参照機能

支援ツールを実現するために必要な言語資源も開発された。

5.2.1.6 教育支援タイプ

- (1) 高松らのシステム(16)は、大学院生程度の学生を支援対象として想定した論文執筆などのテクニカルライティングを支援するためのシステムである。用例検索に焦点をあてる教育を主眼としたシステムであり、英作文支援に必要な機能を検討し、それらの機能を備えている。
- (2) 三好らのシステム(17)は、学習者の要求に応じた内容の英作文学習を支援することを目的とする。文として不完全である学習者からの入力に応じた用例検索と、検索結果の解析、および問い合わせ対話を通じた学習支援方法を有する。

5.2.2 既存システムのまとめ

実用的な英文作成支援システムは大きく二つのタイプに分類できる。一つはテンプレート型であり、英文メール作成支援システムに多い方式である。このようなタイプは、ビジネスレターといった定型化されている分野に適しているが、そのような分野は限られている。

もう一つはコーパス利用型であり、その方式は、日本語のキーワードを入力して、日英対訳コーパスから引いた英語の例文を参考にして、人力で英文を作成するものである。このタイプではコーパスの品質がシステムの性能を左右する。実用的なコーパスサイズは数十万文でも不十分といわれている。このタイプではコーパスのサイズを拡大すると、作成コストよりも品質の向上が見込めなくなるので、おのずとサイズは限界を迎えることになる。コーパスを利用する第二の方式は、キーセンテンスを入力してコーパス中に存在する類似文を検索するものであり、翻訳メモリーとして知られている(19)。コーパスには単言語コーパスと対訳コーパスの2種類がある。単言語コーパスを用いる方法は、大量のコーパスデータを入手できるため、表現の一般性を調べるのに適している。対訳コーパスを用いる方法は、文の対応付けがされた日本文と英文を比較しながら用法を参照できるので、ユーザが意味を把握しやすいという利点を持つ。

5.3. 新たなシステム構想の提案

コーパスを利用するタイプで生じる最大の問題は、結局は人手で翻訳するために翻訳者の能力にかなり依存するということにある。機械翻訳がまだ実用段階にない現在においては、より優れた英文を作成するためには、どのような例文を参照すればよいのか。このような観点は言語の特性を考慮した専門的な議論が求められる。しかし、一般にコーパス利用型はキーワード一致による参照文がほとんどである。しかし、単語そのものは辞書を用いて適切な単語に翻訳できる場合が多い。我々の関心である日英翻訳に限ると、文型や語順に対する適切な提案がコーパスなどのシステムからなされることが重要なのではないだろうか。

例をあげて説明する。「私はあなたがどこから来たのか知らなかった」という日本文を、単語の辞書のみを用いて語順をそのままに「翻訳」すると "I you where from came know didn't." となってしまう。この場合は、SV(=not+verb.)O(=wh-phrase) という文型の例示が重要で、これを利用することで正しい英文に近づき、構造が正確になり、ネイティブにとっても理解しやすくなる。

我々は本稿において、コーパス利用型のうち、現在主流である「キーワード抽出型」の日英翻訳支援システムに代わり、「文型提示型」の日英翻訳支援システムを提案する。文型提示型のシステムには、文型類似文辞書が必要となる。本論文で提案する翻訳支援システムも一種の翻訳メモリーであるが、キーセンテンスとコーパス中の文との間の類似性の計算方法に独自性がある。そのために、従来の翻訳メモリーでは利用していない係り受け解析を用いている。

5.3.1 方法の概要

本システムは以下の3段階で構成される。

- (1)日英対訳コーパスを用意しておく。
- (2)日本文を入力させる。
- (3)日英対訳コーパスを日本文の類似度順に表示する。

この方式であれば、テンプレート型の様に場面を考えたり、キーワード検索型コーパス利用方式のようにキーワードを考えたりする必要がない。また、コーパスを類似度順に得られるので、ユーザの負担が少ない。

5.3.1.1 日英対訳コーパス

日英対訳コーパスには特許コーパスを使う。特許コーパスには二つの特徴があるからである。第一の特徴は日本文と英文の作成者がそれぞれ科学技術の専門家と翻訳の専門家ということである。これによって良質な文を得られる。第二の特徴はデータが多いということである。これによって多様な文を得られる。

5.3.1.2 日本文の類似度

日本文の類似度は文末の類似度によって評価する。日本語の言い回しが文末で決まるからである。文末の類似度は形態素解析と経験式によって評価する。

具体的には以下の通りである。翻訳対象の和文を **a**，データベースから取り出した文を **b** と定義する。まず、**a** と **b** を形態素解析，文節解析，ならびに係り受け解析する。これによって，**a** と **b** が文節係り受け関係に分解される。次に，文節に対して，受け属性と係り属性（表 1）を自動付与する。属性は，日本語の係り受けの性質を考慮して筆者らが独自に設定したものである。最後に，語形と属性を使いながら文間の距離（相違度）を経験式によって計算する。経験式では，文間の距離を「文節間の非類似性に基づく距離」と「文節が深くなるほど小さくなる値」の積の総和の最小値で表す。最小値は動的計画法で求める。なお，文節の深さとは，根文節（日本語では文末の文節）からの係り受けの深さである。これによって，深い文節の距離を軽視するような評価が可能となる。すなわち，浅い文節（文末に近い文節）の距離を重視するような評価が可能となる。以上の方法によって，文型の類似性を大枠で捕らえることができる。

表 1 文節属性

受け種別		係り種別	
意味	記号	意味	記号
名詞	N	連用修飾	y
述語	V	連体修飾	t
サ変動詞	NV	連用または連体修飾	ty
述語名詞	NV	係助詞「は」	h
副詞・連体詞	E	終止	s
接続詞	NV		

5.3.2 高速化

入力文を全てのコーパスと比較しては処理が多すぎる。そこで、次の様にする。先ず、コーパス同士を文末の類似度によってクラスタリングしておく。次に、入力文を各クラスタでの代表的な文と比較する。これによって、入力文とクラスタの類似度が分かる。次に、類似度の高いクラスタを選ぶ。次に、入力文を該当クラスタでの全文と比較する。これによって、入力文と全文の類似度が分かる。最後に、該当クラスタの全文を類似度順に並べ替えて表示する。

5.3.3 クラスタリング

文間の距離を用いて文型類似文辞書を作成する。作成手順の概要を図 1 に示す。図 1 でクラスタとは、クラスタリングの手法(最遠隣法)で距離の近い文を集めた集合で、これらのクラスタを構築することで文型ごとの代表的な文を抽出し辞書を構築することができる。クラスタはすべての文と文との間の距離から作られる距離行列に基づいて行われるべきであるが、計算コストが大きいと考えられるので、我々は文末表現の一致する文を集めて仮の文集合を作り、この仮の文集合の中で多くの文を含むものだけに対してクラスタリングを行うことで計算時間を現実的なものに抑える方針である。すなわち、ここで作成されたクラスタが文型となる。クラスタごとに距離行列を用いて最も中心となる文を選び、これを当該クラスタの代表的な文として選択する。これを集めることで文型類似文辞書の索引を作ることができる。それぞれの索引にはクラスタ内の文が表示される。

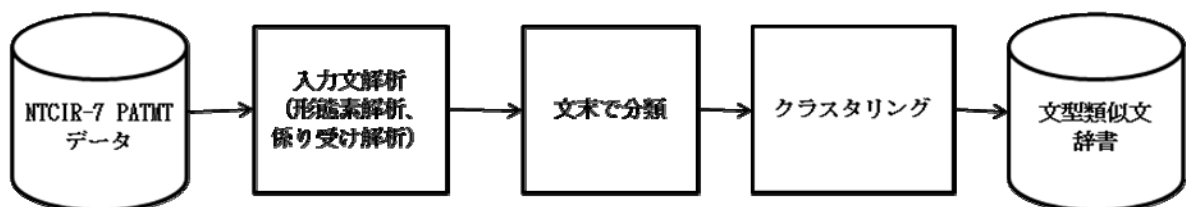


図 1 文型類似文辞書作成過程

5.3.4 クラスタを用いたシステム

前節で構成した文型類似文辞書を用いて日英翻訳支援システムを構築する構想を述べる。概念図を図 2 に示す。まず、翻訳したい日本語文を入力すると形態素解析、係り受け解析が行われる。次に、文型類似文辞書の文型とパターンマッチングが行われ、同一文型の中の日本語文と上記方法で距離計算がなされる。ここで全ての索引のうち、距離が近い順番に日英文のペアが表示される。それぞれのペアから当該クラスタ内の全ての文を閲覧することもでき、それを用いて翻訳者

は英文作成を行う。

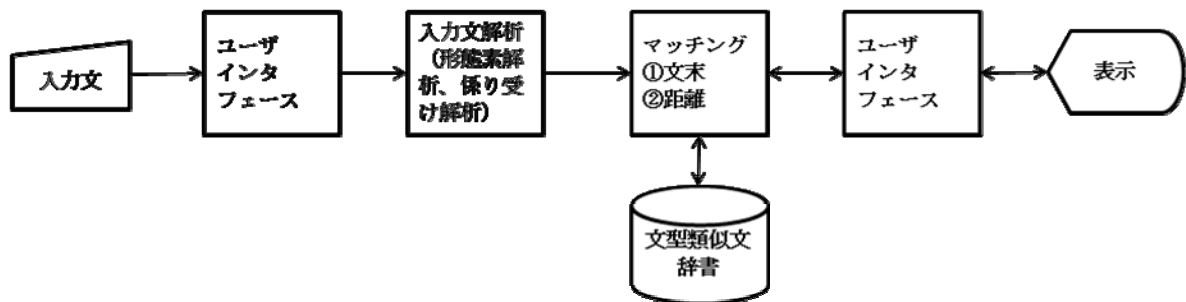


図 2 日英翻訳支援システムの構成

5.3.5 システムの特徴

本節で、本システムの特徴について考察する。はじめに文型類似文辞書を用いる利点について述べる。

(1) 文型の網羅性

本提案の文型類似文辞書は上述の膨大な特許コーパス（180 万文例）から自動抽出されるので、文型の網羅性、完全性が高い。この網羅性はもちろんコーパスの質に負うところが大きいですが、翻訳における有益性を示している。

(2) 文型の自動分析

文型によって英文作成支援をする本があるが(20) (21)，この場合は、翻訳者自身で該当文型を探さなければならず、時間のロスなどコストが大きい。これに対し、提案したシステムでは入力文に対応する文型を自動的に計算するため、複数の候補が表示されユーザの負担が少ない。

(3) 文型類似文の多さ

クラスタによって類似文を集めているため、当該文に対する類似文を複数表示することができる。

(4) 文型類似文の近さによる並び替え

文型類似文が入力文に近い順に自動的に距離計算されて表示されるので、ユーザは参照する類似文を見つけやすい。

(5) 文型とクラスタとの関係

クラスタの生成は文間距離を用いて自動的になされるため、これがいわゆる言語学的な文型と一致するかどうかは別問題となる。しかし、我々は、コーパスが網羅的であるなど適切であれば、我々の定義によって形成したクラスタは対象となる言語分野において使用される文型（特に文末表現）を網羅的に表現する手段として適切であると考えている。

(6) 特許文との関係

本提案システムが対象としている科学技術文（特に特許文）の特性との関連について議論する。科学技術文は複文や長文が多く、現段階では機械翻訳で高品質の英文を作成するのが困難であるにも関わらず、年々増加する特許の国際的な戦略に対応するため、その翻訳コストは重要な課題となっている。人間翻訳の場合はその科学技術分野以外の専門外の人が英訳する場合も多いが、科学技術特有の表現があり難しい。こういった点は、既存のテンプレートによる英文作成支援システムや、キーワードを入力してヒットする文を表示する英文作成支援システムでは困難であることを示している。それに比べ、提案した日英翻訳支援システムは、プロの人間翻訳の実際の類似文を参考に出来るため、既存の専門用語辞書も援用しながら品質の良い英文を作成しやすい。

5.4. まとめと今後の展望

我々は、日英翻訳を行う際に重要となる文型の抽出を容易に行えるようにするため、コーパスを用いる方法のうち、キーワードの一致のみで候補文を抽出する従来型に代わって、文末に注目する方法を提案した。この方法では、文末構造が似ている文をクラスタリング手法でグルーピングし、代表的な文を索引とする文型類似文辞書を構築する。この辞書を用いると、翻訳したい文と同じ文末構造を持つ文の集合を優先的に表示するために、同じコーパスであっても異なる文を候補文とすることができる。このような方法では、いわゆる機械翻訳のように直訳的に翻訳するのではつukれないような、こなれた英文を作成できると予想される。それを確かめるために、文型類似文辞書を用いた日英翻訳システムのプロトタイプを構築し、被験者を用いた評価実験を行う必要がある。

この新方式は機械翻訳にとっても有用と考えている。例えば機械翻訳の前処理に利用できる可能性をもっている。将来的には機械翻訳や特許文の翻訳への応用を図る予定である。

謝辞

支援をいただいた(財)日本特許情報機構に感謝します。

参考文献

- (1) 堺 武志 : “日本語 - 外国語 文例 翻訳 ソフト : ことばさがし” ,
<http://rose.ruru.ne.jp/multiplication/kotobaWin/jp.html> (2010)
- (2) 松原茂樹, 加藤芳秀, 江川誠二 : ” 英文作成支援ツールとしての用例文検索システム E S C O R T ”, 情報管理 51(4), pp. 251-259 (2008)
- (3) 松原茂樹, 江川誠二, 加藤芳秀 : ” 英文用例検索システム E S C O R T : 論文データベースを用いた図書館サービス ”, 情報プロフェッショナルシンポジウム予稿集, pp. 125-129 (2007)
- (4) 大鹿広憲, 佐藤学, 安藤進, 山名早人 : ” Google. を活用した英作文支援システムの構築 ” 『DEWS2005』4B-i8 (2005)
- (5) 赤峯 享, 奥村 明俊, 村木 一至, 亀井 真一郎 : ” 日本語入力による英文作成支援システム : 仮名漢字変換から仮名英語変換へ ”, 情報処理学会全国大会講演論文集 第 46 回平成 5 年前期(3), pp. 119-120 (1993)
- (6) 赤峯 享, 奥村 明俊, 村木 一至 : ” TransAid : 英文書作成支援システム ”, 情報処理学会全国大会講演論文集 第 44 回平成 4 年前期(3), pp. 261-262 (1992)
- (7) 安藤 一秋, 綱嶋 祐一, 溝渕 昭二 : ” WWW 言語資源を利用した英文作成支援ツール : ユーザの語彙レベルに基づく適応型用例提示 ”, 電子情報通信学会技術研究報告. ET, 教育工学 105(632), pp. 139-144 (2006)
- (8) 安藤 一秋, 綱嶋 祐一, 溝渕 昭二 : ” Google Web APIs を利用した英文作成支援ツール ” , 情報科学技術フォーラム一般講演論文集 5(3), pp. 485-486 (2006)
- (9) 土井 伸一, 亀井 真一郎, 田村 真子, 浜田 和彦, 村木 一至 : ” FEP 型英文作成支援ツールにおける対訳例文利用機構 ” , 情報処理学会全国大会講演論文集 第 56 回平成 10 年前期(2), pp. 226-227 (1998)
- (10) 山端 潔, 土井 伸一, 亀井 真一郎, 佐藤 研治 : ” FEP 型英文作成支援ツールにおける対訳的曖昧性解消方式 ”, 情報処理学会全国大会講演論文集 第 52 回平成 8 年前期(3), pp. 29-30 (1996)
- (11) 亀井 真一郎, 佐藤 研治, 浜田 和彦, 村木 一至 : ” FEP 型英文作成支援ツール : 外国語情報発信の効果的インタフェース ” , 情報処理学会全国大会講演論文集 第 51 回平成 7 年後期(3), pp. 77-78 (1995)
- (12) 山端 潔, 石原 佳典, 土井 伸一, 亀井 真一郎, 佐藤 研治, 浜田 和彦 : ” FEP 型英文作成支援ツール : 日英構文変換 ” , 情報処理学会全国大会講演論文集 第 51 回平成 7 年後期(3), pp. 75-76 (1995)
- (13) 土井 伸一, 亀井 真一郎, 山端 潔, 田村 真子, 浜田 和彦 : ” FEP 型英文作成支援ツール : 訳語選択のユーザーインタフェースと辞書記述 ” , 情報処理学会全国大会講演論文集 第 51 回平成 7 年後期(3), pp. 73-74 (1995)
- (14) 高倉 佐和, 古郡 延治 : ” 日本語入力による英文作成支援インタフェース ”, 情報処理学会研究報告. 自然言語処理研究会報告 2002(66), pp. 7-14 (2002)
- (15) 成田真澄 : ” 英文アブストラクト作成支援ツールのユーザ評価 ”, 平成 12 年度 COE 形成基礎研究費研究成果報告(先端的言語理論の構築とその多角的な実証 - ヒトの言語を組み立て演算する能力を語彙の意味概念から探る -), 神田外語大学, pp. 309-318 (2001)
- (16) 高松優, 水野淳太, 岡崎直観, 乾健太郎 : ” 英作文支援のための用例検索システムの構築 ”, 言語処理学会 第 18 回年次大会発表論文集 pp. 361-318 (2012)
- (17) 三好康夫, 岡本竜 : ” 用例に基づく対訳的英作文支援システム ”, JSiSE(教育システム情報学会)-W 第 3 回若

手研究者フォーラム (2000)

- (18) 小倉書店編：“CD-ROM 科学技術論文，報告書その他の文書に必要な英語文型・文例辞典”，小倉書店 (1997)
- (19) Somers, H.：“Translation memory systems,” in *Computers and Translation: a Translator's Guide*, John Benjamins Publishing Co.：pp.31-48 (2003)
- (20) 佐藤元志著，田中宏明監修，古米弘明監修，鈴木穰監修：“英語論文表現例集 with CD-ROM すぐに使える 5,800 の例文”，技報堂出版 (2009)
- (21) 佐藤洋一編著：“科学技術英語論文英借文用例辞典 英作文から英借文へ簡単！英語論文作成法”，オーム社 (2010)

6. 1 拡大評価部会の活動概要

山梨英和大学 江原暉将

昨年度から本研究会の下部組織として「拡大評価部会」を設置し、機械翻訳の評価に関する議論を進めてきた[1]。本部会での議論の焦点は以下の通りである。

1. 「技術調査目的」のために特許文書を機械翻訳する場合の評価
2. 人手評価、自動評価、半自動評価
3. 評価用テストセット
4. 対象とする言語の範囲：日本語、英語、中国語
5. 評価手法の理想形、理想を実現するための課題、課題克服への道程

今年度は3回の部会を開催した。

2013年5月10日 今年度の活動計画

2013年10月11日 中間報告と議論

2014年1月24日 最終報告および報告書の執筆について

活動は、人手評価、自動評価、テストセットの3つのテーマに関し分担して行った。概要を以下に示すが、詳細については本章の各記事をご覧ください。

人手評価に関しては、NTCIR-10における審査官経験者による人手評価について分析した。特許審査においては先行文献調査が必要であり、「技術調査目的」の中でも審査業務を直接反映した場合の人手評価を取り上げた。

自動評価に関しては、WMTおよびNTCIRにおける機械翻訳結果に対して既存の自動評価手法を適用し人手評価との相関を分析した。各種自動評価手法の得失を評価するとともに、アジア言語と欧米の言語とで評価に違いが出るかを分析した。

テストセットに関しては、中国語から日本語への機械翻訳評価を念頭に、パテントファミリーから文対応を取得した。さらに各文対応に評価の観点を付与してテストセットとして整備した。

機械翻訳結果の評価に対しては、まだまだ課題が多く、2014年度も引き続き部会の活動を継続する予定である。

参考文献

[1] 江原暉将：機械翻訳の評価について、平成24年度AAMT/Japio特許翻訳研究会報告書、pp.37-39、2013年3月。

6. 2 NTCIR PATMT データと WMT METRICS TASK データにおける

英文データを対象とした自動評価法のメタ評価

北海学園大学 越前谷 博

NTT コミュニケーション科学基礎研究所 須藤 克仁

岡山県立大学 磯崎 秀樹

山梨英和大学 江原 暉将

6.2.1 はじめに

近年、機械翻訳の進展に伴い、機械翻訳のための様々な自動評価法が提案されている。その中で最も多く使用されている手法は BLEU^[1]であることは多くの研究者が認めるところである。しかし、BLEU については文単位や異なるアーキテクチャーを持つ機械翻訳システムが混在している場合、その評価精度が十分ではないという問題点も同時に指摘されている。そこで、本報告では、現在提案されている自動評価法がどの程度の評価精度を保証するものなのか、また、それぞれの自動評価法を有効的に利用するために、個々の自動評価法がどのような利点や欠点を持っているのかを明らかにすることを目的に行った、自動評価法のメタ評価について報告する。今回のメタ評価では、最も広く利用されている言語である英文に翻訳された文、即ち、英文データを対象として分析を行った。メタ評価の結果、個々の自動評価法の利点や欠点を含めた知見を得ることができた。本報告で得られた知見は、使用目的に応じてどの自動評価法を使用すべきかを決定する際の有益な情報になると考えられる。

6.2.2 使用した自動評価法

今回のメタ評価で使用した自動評価法は 4 つである。まず、自動評価法におけるデファクトスタンダードである BLEU と NIST^[2]を使用した。更に、日本で開発された自動評価法である IMPACT^[3]と RIBES^[4]を使用した。いずれの手法も機械翻訳が出力する翻訳文と人手で作成される正解訳である参照訳間の比較により評価結果をスコアとして算出するものである。これら 4 つの自動評価法がどの程度の評価精度を示し、どのような特徴を有するのかを検証した。

6.2.2.1 BLEU (Bilingual Evaluation Understudy)

BLEU は現在最も広く利用されている自動評価法である。その特徴は n グラム適合率をベースにしている点である。以下に BLEU の計算式を示す。

$$p_n = \frac{\sum_{c \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{c' \in \{\text{Candidates}\}} \sum_{n\text{-gram}' \in C'} \text{Count}(n\text{-gram}')}$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right)$$

p_n は翻訳文における n グラム適合率を示している。 N の値としては4が最も良いとされている。 BP は文の長さに対するペナルティを示している。 BP を用いることで、翻訳文の構成単語数が非常に小さい場合、参照訳と翻訳文が大きく異なっても n グラム適合率が高くなり、その結果、スコアが高くなるという問題を回避することができる。具体的には翻訳文の構成単語数 c が参照訳の構成単語数 r より小さい場合、 $e^{(1-r/c)}$ が計算される。そして、 N を1から4まで変化させた際の n グラム適合率の幾何平均を求め、 BP を付与して得られた値を評価スコアとして出力する。このように BLEU の計算法は非常にシンプルであることから高速処理が可能である。しかし、 n グラム適合率の幾何平均を用いているため、値が0となる n グラム適合率が存在すると全体のスコアが0になる。このような傾向は構成単語数の小さくなる文に適用した場合に高くなるため、文単位の評価には適さないとされている。また、 w_n は $1/N$ である。BLEU のスコアは0.0 から1.0の値を示し、高い値ほど評価が高いことを意味する。

6.2.2.2 NIST

NIST も BLEU と同様、広く利用されている自動評価法の一つである。NIST の特徴は、BLEU と同じく n グラム適合率に基づいているが、情報量を利用している点にある。以下に NIST の計算式を示す。

$$Info(w_1 \dots w_n) = \log_2 \left(\frac{\text{the \# of occurrences of } w_1 \dots w_{n-1}}{\text{the \# of occurrences of } w_1 \dots w_n} \right)$$

$$Score = \sum_{n=1}^N \frac{\sum_i \left(\frac{\sum_{\text{翻訳文と参照訳に共通する } w_1 \dots w_n} Info_i(w_1 \dots w_n)}{\sum_i \text{翻訳文中の全 } n\text{-gram 数}} \right)}{\cdot \exp \left\{ \beta \log^2 \left[\min \left(\frac{L_{sys}}{L_{ref}}, 1 \right) \right] \right\}}$$

式 $Info$ が情報量を示している。例えば、2つの単語列”the dog”と”the cat”においてすべての参照訳における出現回数がそれぞれ1と2であり、単語”the”の出現回数が10であった場合、相互情報量 $Info$ は”the cat”よりも”the dog”の方が高くなる。したがって、Score では”the cat”よりも”the dog”の方が重みづけされることになる。また、NIST では n グラム適合率を統合する際には単純和を用いている。NIST におけるペナルティは BLEU と同様に、翻訳文の構成単語数が参

照訳の構成単語数よりも小さい場合に課せられる。NIST のスコアは 0.0 以上の実数値で示され、値が高いほど、評価が高い。n グラム適合率における N の値としては BLEU の 4 に対して、5 が使用される。更に、NIST も BLEU と同様に、文単位の評価には適さないとされている。

6.2.2.3 IMPACT (Intuitive common Parts ConTinum)

近年では日本でも自動評価法がいくつか提案されている。その中の一つである IMPACT は文単位の自動評価に着目している点が特徴である。IMPACT の計算式を以下に示す。

$$pos_w = \left(1.0 - \left| \frac{posX(c)}{m} - \frac{posY(c)}{n} \right| \right)$$

$$RS = \left(\sum_{c \in LCS} (length(c)^\beta \times pos_w) \right)^{\frac{1}{\beta}}$$

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

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

$$IMPACT = \frac{(1 + \gamma^2)PR}{\gamma^2 P + R} \quad \gamma = \frac{P}{R}$$

IMPACT は式 RS と式 pos_w に基づき共通部分列を一意に決定する。その際、最長共通部分列 (LCS) を用いる。この処理を再帰的に繰り返すことで全ての共通部分をスコアに反映させている。しかし、語順の異なる共通部分については式 P と式 R においてパラメータ α によりスコアにどの程度反映させるかの制御を行っている。 P は適合率、 R は再現率を意味する。そして、 P と R の F 値をスコアとして使用する。式 pos_w 中の m と n はそれぞれ翻訳文と参照訳の構成単語数である。また、 c は任意の共通部分を示しており、 $posX(c)$ と $posY(c)$ は共通部分 c の翻訳文と参照訳における出現位置を示す。したがって、式 pos_w は任意の共通部分の相対的な位置のズレを表している。式 RS 中の $length(c)$ は任意の共通部分 c の構成単語数を示している。 β は共通部分の構成単語数に基づく重みづけパラメータである。また、式 P と R における RN は共通部分列の決定処理の回数である。例えば、共通部分列が 2 つ存在し、それらの語順が異なる場合、 RN は 2 となる。IMPACT のスコアは 0.0 から 1.0 の値で示され、値が大きいほど評価が高い。

6.2.2.3 RIBES (Rank-based Intuitive Bilingual Evaluation Score)

RIBES も日本で提案された自動評価法であり、その特徴は日英など構造が大きく異なる言語間

の自動評価を高い精度で行うことを目的としている点にある。以下に RIBES の計算式を示す。

$$\tau = \frac{2 \times Conc}{nC_2} - 1$$

$$P = n/h$$

$$RIBES = \frac{(\tau + 1)}{2} \cdot P^\alpha \cdot BP^\beta$$

RIBES は語順に着目した評価法であり、Kendall の順位相関係数に基づき語順の一致度を得る。式 τ は次のように決定される。参照訳と翻訳文で共通する単語が n 個あるとする。共通する単語が参照訳において $w_1, w_2 \dots w_n$ の順に並んでいるとして、これを $[1, 2, \dots, n]$ という整数のリストで表す。共通する単語の翻訳文での語順は、この $1, \dots, n$ の整数の順列で表せる。この順列から整数 2 つのペアを取りだすと、 nC_2 個のペアができる。このうち、昇順のペアの数を $Conc$ とすると、これらの値より順位相関係数 Kendall の τ が求まる。そして、この τ を用いて式 RIBES よりスコアを計算する。 τ の値は -1 から +1 の値を取るため、1 を加えたものを 2 で割ることでスコアが 0 から 1 の値となるように正規化を行っている。また、RIBES では翻訳文と参照訳間の共通単語のみに着目するため、共通単語が非常に少ない場合には、過剰に高いスコアを与えることがある。そこで、ペナルティとして P^α をスコア計算に導入している。 P は翻訳文の構成単語数における共通単語数の割合である。また、 BP は BLEU と同様のペナルティで文の短い翻訳文の適合率が過度に高くなることを防いでいる。 α と β はパラメータであり、それぞれ 0.25 と 0.1 が使用されている。

6.2.3 メタ評価

メタ評価は前節で述べた 4 つの自動評価法を用いて行った。その際に用いたデータはコンテスト型ワークショップである NTCIR の特許翻訳テストコレクション（以下、NTCIR）と Workshop on Statistical Machine Translation の Metrics Tasks（以下、WMT）が提供しているデータである。これらのデータから英文に翻訳されたものを対象としてメタ評価を行った。

6.2.3.1 実験データ

使用した全ての英文（翻訳文、参照訳）における tokenization は NTCIR の “Automatic Evaluation Procedure at NTCIR-10”¹⁰ に記載されている手順に基づき行った。使用した参照訳の数は全データにおいて 1 である。これらの翻訳文と参照訳を用いて、それぞれの自動評価法はスコアを算出する。そして、自動評価法のスコアと人手評価による評価値を比較することで自動評価法の精度を求める。その際には System-level と Segment-level において自動評価法の精度を求める。人手評価においては文単位の評価値のみが提供されているため、システム単位の評価値は新たに決定する必要がある。そこで、システム単位の評価値は文単位のスコア値の平均値を用

いた。ただし、NTCIR-7 データにおいては、3 名による人手評価の値が存在しているため、始めに 3 名の値の平均値を文単位の評価値とし、更に、それらの平均値をシステム単位の評価値とした。また、NTCIR-9 データにおいては、Acceptability の観点からの人手評価は高い方から“AA”、“A”、“B”、“C”、“D”となっているため、それぞれ 5、4、3、2、1 の 5 つの値に変換した。

6.2.3.2 実験方法

自動評価法の評価精度を求めるために System-level と Segment-level の観点より自動評価法のスコアと人手評価の評価値との間の相関係数を算出した。System-level の相関係数には Spearman の順位相関係数^[6]を、Segment-level の相関係数には Kendall の順位相関係数^[7]を用いた。なお、順位相関係数の計算式はデータに同順位がある場合の計算式を用いた。また、それぞれの自動評価法のシステムはインターネット上で配信されているものをダウンロードして用いた。これは再現性の観点から自作可能な自動評価法であっても、提供されているものを用いるべきとの考えに基づいている。BLEU と NIST については“mteval-13a.pl”（以下、BLEU-v13a、NIST-v13a）、IMPACT については“IMPACT.rb”（以下、IMPACT-v4.0.2）、RIBES については“RIBES.py”（以下、RIBES-v1.02.4）をそれぞれ関連サイトからダウンロードして使用した。また、IMPACT と RIBES のパラメータ値についてはデフォルト値を用いた。

6.2.3.3 実験結果

メタ評価における実験結果を表 1 から表 8 に示す。“ () ”の数値は相関係数の計算に用いたデータ数である。表 1 と表 2 は NTCIR-7 データにおける System-level と Segment-level の相関係数である。

NTCIR-7 においては、Adequacy と Fluency の 2 つの観点からの人手評価が存在しているためそれぞれの人手評価の評価値と自動評価法のスコアとの相関係数を求めた。表 3 と表 4 は NTCIR-9 データにおける System-level と Segment-level の相関係数である。NTCIR-9 については、Adequacy と Acceptability の 2 つの観点からの人手評価が存在しているためそれぞれの相関係数を求めた。表 5 と表 6 は WMT12 データにおける System-level と Segment-level の相関係数である。また、表 7 と表 8 は WMT13 データにおける System-level と Segment-level の相関係数である。WMT12、WMT13 データにおける“en”、“cs”、“de”、“es”、“fr”、そして“ru”はそれぞれ英語、捷語、独語、西語、仏語、そして、露語を意味する。したがって、“de-en”とは原文が独語であり、翻訳文が英語であることを示している。

表 1 NTCIR-7 データにおける System-level の相関係数

	Adequacy(15)	Fluency(15)	Avg.
BLEU-v13a	0.5873	0.5935	0.5904
NIST-v13a	0.5568	0.5396	0.5482
IMPACT-v4.0.2	0.8545	0.8250	0.8398
RIBES-1.02.4	0.9545	0.9018	0.9282

表 2 NTCIR-7 データにおける Segment-level の相関係数

	Adequacy(1,500)	Fluency(1,500)	Avg.
BLEU-v13a	0.3330	0.3491	0.3411
NIST-v13a	0.2911	0.3233	0.3072
IMPACT-v4.0.2	0.4275	0.4316	0.4296
RIBES-1.02.4	0.3679	0.3388	0.3534

表 3 NTCIR-9 データにおける System-level の相関係数

	Adequacy(19)	Acceptability(14)	Avg.
BLEU-v13a	-0.0364	0.1810	0.0723
NIST-v13a	-0.2512	-0.1587	-0.2050
IMPACT-v4.0.2	0.3513	0.2824	0.3169
RIBES-1.02.4	0.6917	0.6714	0.6816

表 4 NTCIR-9 データにおける Segment-level の相関係数

	Adequacy(5,700)	Acceptability(5,700)	Avg.
BLEU-v13a	0.1222	0.1304	0.1263
NIST-v13a	0.0676	0.0783	0.0730
IMPACT-v4.0.2	0.2258	0.2275	0.2267
RIBES-1.02.4	0.2367	0.2245	0.2306

表 5 WMT12 データにおける System-level の相関係数

	cs-en(6)	de-en(16)	es-en(12)	fr-en(15)	Avg.
BLEU-v13a	0.9429	0.4985	0.8911	0.7707	0.7758
NIST-v13a	0.9429	0.6364	0.8989	0.6909	0.7918
IMPACT-v4.0.2	0.8857	0.7191	0.9213	0.7688	0.8237
RIBES-1.02.4	0.9429	0.6968	0.8339	0.6129	0.7716

表 6 WMT12 データにおける Segment-level の相関係数

	cs-en(6,421)	de-en(7,096)	es-en(5,664)	fr-en(6,933)	Avg.
BLEU-v13a	0.1153	0.1484	0.1243	0.1460	0.1335
NIST-v13a	0.1151	0.1444	0.1312	0.1377	0.1321
IMPACT-v4.0.2	0.1177	0.1404	0.1357	0.1518	0.1364
RIBES-1.02.4	0.0983	0.1299	0.1124	0.1441	0.1212

表 7 WMT13 データ System-level の相関係数

	cs-en(11)	de-en(17)	es-en(12)	fr-en(13)	ru-en(19)	Avg.
BLEU-v13a	0.8159	0.9234	0.8409	0.9107	0.5899	0.8162
NIST-v13a	0.8205	0.8903	0.7465	0.8283	0.7439	0.8059
IMPACT-v4.0.2	0.8659	0.9148	0.8654	0.8393	0.7487	0.8486
RIBES-1.02.4	0.6482	0.8139	0.6868	0.8596	0.6047	0.7226

表 8 WMT13 データにおける Segment-level の相関係数

	cs-en (57,000)	de-en (87,635)	es-en (46,825)	fr-en (54,205)	ru-en (100,120)	Avg.
BLEU-v13a	0.1356	0.1357	0.1472	0.1279	0.1215	0.1336
NIST-v13a	0.1270	0.1277	0.1420	0.1183	0.1148	0.1260
IMPACT-v4.0.2	0.1445	0.1473	0.1573	0.1285	0.1306	0.1416
RIBES-1.02.4	0.1321	0.1436	0.1504	0.1233	0.1227	0.1344

6.2.3.4 考察

表 1 から表 8 より、いくつかの知見が得られた。NTCIR データにおけるメタ評価の結果からは RIBES が System-level においては他の手法よりも非常に高い相関係数を示している。Segment-level においては NTCIR-7 データでは IMPACT、NTCIR-9 データでは RIBES が最も高い相関係数を示している。WMT データにおけるメタ評価の結果からは IMPACT が System-level、Segment-level 共に最も高い相関係数を示している。これらの結果は IMPACT と RIBES のそれぞれの特徴が反映されたものとなっている。IMPACT は文単位での評価精度の向上を目的に提案されており、今回のメタ評価の結果がそのことを裏付けている。また、RIBES は語順の大きく異なる言語間の翻訳文に対する評価精度の向上を目的に提案されており、そのことが NTCIR データの System-level のメタ評価の結果として表れている。

また、メタ評価の結果について機械翻訳システムのアーキテクチャーの観点から検証する。表 9 にそれぞれのデータにおける統計翻訳とルールベースの機械翻訳システムの数を示す。ただし、WMT13 データについてはルールベースの機械翻訳システムの数が不明なため対象から外した。

表 9 機械翻訳システムのアーキテクチャーの内訳

	NTCIR-7	NTCIR-9		WMT12			
		Adequacy	Acceptability	cs-en	de-en	es-en	fr-en
SMT	12	15	12	6	13	9	12
ルールベース	2	4	2	0	3	3	3
SMT の割合	80.0%	78.9%	85.7%	100.0%	81.3%	75.0%	80.0%

一般的に BLEU や NIST などの N グラム適合率に基づく自動評価法は統計翻訳に対しては高いスコアを与え、ルールベースには低いスコアを与えることが指摘されている。表 9 より統計翻訳の割合が比較的低いデータは NTCIR-9 の”Adequacy”と WMT12 の”es-en”である。これらのデータにおけるメタ評価の結果を見ると、表 3 と表 4 より NTCIR-9 の”Adequacy”では RIBES が最も相関係数が高く、表 5 と表 6 より WMT12 の”es-en”では IMPACT が最も相関係数が高くなっており、今回のメタ評価からは、どちらが複数のアーキテクチャーが混在している場合でも高い評価精度を示すことができるかについては一概には言えない。

次いで、メタ評価の結果と人手評価の評価値との関係について述べる。表 10 と表 11 に NTCIR データと WMT データの人手評価の平均値を示す。表 10 と表 11 より WMT データと NTCIR-7 の”Fluency”の評価値は 3.0 以上であり、それら以外 3.0 未満である。そして、これらの違いは今回のメタ評価の結果とある程度の関連性があると考えられる。即ち、人手評価の評価値が比較的高いデータにおいて IMPACT は高い相関係数を得ることができ、人手評価の評価値が中間レベルであれば RIBES が高い相関係数を得ることができると考えられる。したがって、ある程度高い人手評価値が得られる機械翻訳システムには IMPACT が、中間レベルの人手評価値が得られる機械翻訳システムにおいては RIBES が有効であると考えられる。なお、今回のメタ評価において、IMPACT と RIBES との間で 5%未満の有意水準で有意であったのは、この NTCIR-7 の Segment-level での”Fluency”のみであった。

表 10 NTCIR データにおける人手評価の評価値

	NTCIR-7		NTCIR-9	
	Adequacy	Fluency	Adequacy	Acceptability
評価値	2.6	3.2	2.7	2.7

表 11 WMT データにおける人手評価の評価値

	WMT12				WMT13				
	cs-en	de-en	es-en	fr-en	cs-en	de-en	es-en	fr-en	ru-en
評価値	3.2	3.1	3.1	3.2	3.1	3.4	3.3	3.4	3.1

6.2.3 まとめ

本報告では、NTCIR データと WMT データを用いて英文を対象とした自動評価法のメタ評価について述べた。メタ評価の結果、IMPACT と RIBES はそれぞれの特徴に沿った評価が実現できていることが明らかとなった。これらの結果より、使用目的に応じた自動評価法の使い分けのために有効な情報がある程度提供できたと考えられる。しかし、いずれの自動評価法においても文単位の評価精度は得られた相関係数を見る限りにおいては不十分であり、自動評価法の研究における今後の大きな課題である。また、今後は対象を英文だけではなく、他の言語文も用いたメタ評価を行う必要がある。更に、様々な自動評価法が提案されている現状を踏まえ、他の自動評価法も含めたメタ評価を実施する予定である。

謝辞

この研究は国立情報学研究所との共同研究に関連して行われた。

参考文献

- [1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu(2002) “BLEU: a Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318.
- [2] NIST. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, 2002, <http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf>
- [3] Echizen-ya, Hiroshi. and Araki, Kenji. Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, Proceedings of the Eleventh Machine Translation Summit, pp.151-158, 2007.
- [4] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada(2010) “Automatic Evaluation of Translation Quality for Distant Language Pairs,” Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 944–952.
- [5] “Automatic Evaluation Procedure at NTCIR-10”,
<http://ntcir.nii.ac.jp/PatentMT-2/AutomaticEvaluationProcedure/>
- [6] “スピアマンの順位相関係数”, <http://aoki2.si.gunma-u.ac.jp/lecture/Soukan/spearman.html>
- [7] “ケンドールの順位相関係数”, <http://aoki2.si.gunma-u.ac.jp/lecture/Soukan/kendall.html>

6. 3 NTCIR-10 特許審査評価の結果

NHK 放送技術研究所 後藤功雄

1. はじめに

筆者らは2012年1月～2013年6月にかけて評価型ワークショップ NTCIR-10 特許機械翻訳タスク¹[1] を主催した。NTCIR-10 特許機械翻訳タスクでは、特許審査での機械翻訳の有用性に基づく評価（特許審査評価）を実施した。昨年度（2012年度）の報告書²[2]では NTCIR-10 特許機械翻訳タスクでの人手評価方法の概要について説明した。本稿では、特許審査評価の詳細な手順および結果について説明する。その前に特許審査評価の手法概要を説明する。

2. 特許審査評価の手法概要

特許審査官は、審査において既存の特許を調査して同じ技術が存在すれば、その既存の特許を引用して審査対象の特許を拒絶する。そのため、既存の特許に記載された技術的内容である事実を認定する必要がある。既存の特許が外国語で書かれていてその内容が分からない場合には、翻訳する必要がある。この翻訳を機械翻訳で行った場合に、翻訳結果から引用文書の特許を認定するために有用な事実をどれだけ認定できるかに基づいて、機械翻訳の有用性を評価する。特許審査評価では、外国語が日本語の場合と中国語の場合を想定して、日英翻訳および中英翻訳の結果を評価した。本評価は日本知的財産翻訳協会（NIPTA）の協力により実施した。

2. 1 全体の評価の流れ

全体の評価の流れ（図1）について説明する。

（1）準備 評価で利用するデータを構築する。まず審査の結論が拒絶の審決（審査の最終決定を記載した文書）とその審査で引用された特許を取得する。そして、審査において「引用文書から審査官が認定した事実」の根拠となる文を引用文書から抽出する。抽出した文をテストデータとする。なぜなら、この抽出した文は、「審査官が認定した事実」を表しているためである。そして、このテストデータが正しく翻訳されれば、翻訳結果から「審査官が認定した事実」を認定することができるためである。

（2）翻訳 テストデータを機械翻訳する。

（3）評価 翻訳結果から、「引用文書から審査官が認定した事実」をどれだけ認定できて、審査に有用であるかについて評価する。

¹ <http://ntcir.nii.ac.jp/PatentMT-2/>

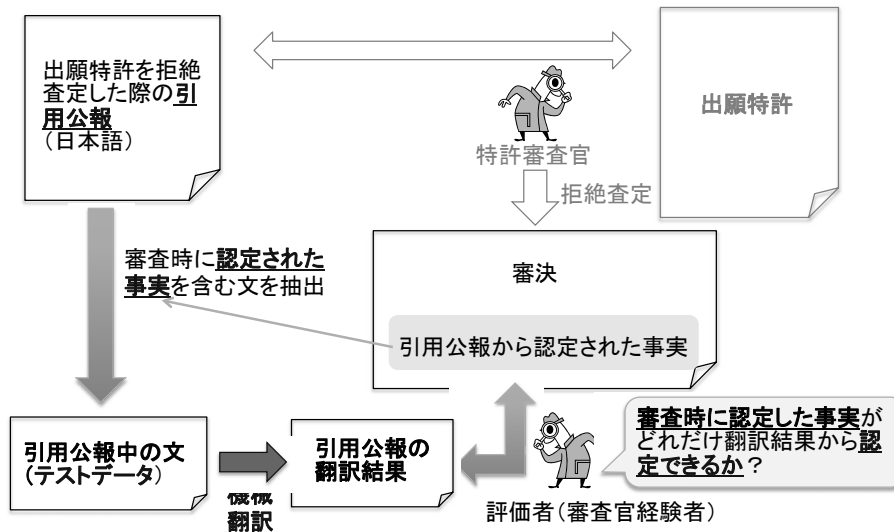


図 1：特許審査評価の概要

2. 2 評価方法

特許審査評価は日英翻訳と中英翻訳に対して行った。特許庁で審査官の経験があり、英語の能力が高い 2 人の評価者が評価した。1 システムあたり各評価者が 20 文書、合計 40 文書（文書の重複を含む）の評価を行なった。評価したシステムは、adequacy の結果が上位のシステムから手法の多様性を考慮して日英、中英それぞれ 3 システムを選択した。

2. 3 評価基準

特許審査評価で用いた評価基準を表 1 に示す。有用性の評価は、過去の審査で審査官が引用文書の特許から認定した事実を、引用文書を機械翻訳した結果からどれだけ認定できるかに基づいて行った。評価は引用文書単位で行なった。

2. 4 データの構築

この評価に必要なデータを、審決を用いて以下のようにして構築した。

- (1) 結論が不成立（拒絶）の審決を取得する。
- (2) 審決中に記載されている「引用文書から審査官が認定した事実の説明」を抽出する。
- (3) 審査官が認定した事実を構成要素単位に分けて、それぞれの内容の根拠となる文を引用文書から抽出する。抽出した文を機械翻訳で翻訳するテストデータとする。

表 2 に、審査官が認定した事実と引用文書から抽出した文の例を示す。表 2 の左端の列は、審決中に記載されている審査官が引用文書から認定した事実である。表 2 の中央の列は、審査官が認定した事実を構成要素単位に分けたものである。表 2 の右端の列は、中央の列の事実を認定する根拠となった引用文書中の文を抽出したものであり、この文が翻訳するテストデータである。29 文書のテストデータを構築した。中英翻訳のテストデータは、日英翻訳のテストデータを日中翻訳して構築した。

表 1：特許審査評価の評価基準

VI	引用発明を認定するために有用な事実が全て認定できて、翻訳結果のみで審査可能
V	引用発明を認定するために有用な事実が半分以上認定できて、審査に有用
IV	引用発明を認定するために有用な事実が1つ以上認定できて、審査に有用
III	IVに至らないが、部分的に事実が認定できて、その文献が審査で無視できないことが分かる
II	一部の事実が認定できたが、審査に有用とはいえない
I	全く事実が認定できず、審査の役に立たない

(引用文書単位の評価)

表 2：審査官が認定した事実と引用文書から抽出した文の例

審査官が認定した事実	構成要素単位に分けた審査官が認定した事実	引用文書から抽出した文(テストデータ)
これらの記載事項によると、引用例には、「内部において、先端側に良熱伝導金属部 43 が入り込んでいる中心電極 4 と、中心電極 4 の先端部に溶接されている貴金属チップ 45 と、中心電極 4 を電極先端部 41 が碍子先端部 31 から突出するように挿嵌保持する絶縁碍子 3 と、絶縁碍子 3 を挿嵌保持する取付金具 2 と、中心電極 4 の電極先端部 41 との間に火花放電ギャップ G を形成する接地電極 11 とを備えたスパークプラグにおいて、中心電極 4 の直径は、1.2～2.2mm としたスパークプラグ。」の発明が記載されていると認められる。	内部において、先端側に良熱伝導金属部 43 が入り込んでいる中心電極 4	また、図 3 に示すごとく、中心電極 4 の内部においては、上記露出開始部 431 よりも先端側にも良熱伝導金属部 43 が入り込んでいる。
	中心電極 4 の先端部に溶接されている貴金属チップ 45	また、中心電極 4 の先端部には、貴金属チップ 45 が溶接されている。
	中心電極 4 を電極先端部 41 が碍子先端部 31 から突出するように挿嵌保持する絶縁碍子 3	上記中心電極 4 は、電極先端部 41 が碍子先端部 31 から突出するように絶縁碍子 3 に挿嵌保持されている。
	絶縁碍子 3 を挿嵌保持する取付金具 2	上記絶縁碍子 3 は、碍子先端部 31 が突出するように取付金具 2 に挿嵌保持される。
	中心電極 4 の電極先端部 41 との間に火花放電ギャップ G を形成する接地電極 11	上記接地電極 11 は、図 2 に示すごとく、電極先端部 41 との間に火花放電ギャップ G を形成する。
中心電極 4 の直径は、1.2～2.2mm	また、上記碍子固定部 22 の軸方向位置における中心電極 4 の直径は、例えば、1.2～2.2mm とすることができる。	

3. 評価の詳細な手順

本節では、特許審査評価の詳細な手順について説明する。評価者が評価を行う際に利用し、評価結果を記入した評価シートの例を表 3 に示す。

表 3：特許審査評価の評価シートと評価結果の例

(1)審査官が引用文書から認定した事実(審決中の記載)							
高誘電体膜をゲート絶縁膜に用いた周辺トランジスタ部のトランジスタと、これと同様の素子構造でゲート絶縁膜に強誘電体膜を用いたメモリセル部のメモリ強誘電体メモリとを同一のチップに形成する半導体装置であって、周辺トランジスタ部の Si 基板 1 表面に、ゲート絶縁膜 9c として高誘電体膜の Ta2O5 膜を、Si 基板 1 界面との間に界面準位等ができていくように、SiO2 膜を介して堆積し、メモリセル部の Si 基板 1 表面に、高誘電体膜である Ta2O5 膜を用いる誘電体膜 9a を介して、BiSr2Ta2O9 膜を用いる強誘電体膜 9b を堆積した半導体装置。							
(2) 審査官が引用文書から認定した事実(審決中の記載)	(3)引用公報の記載(翻訳元のテストデータ)	(4)機械翻訳出力	(5)認定すべき引用発明部分を抽出した結果	(6)英語が不明な表現	(7)技術的内容が正しく理解できない表現	(8)備考	
高誘電体膜をゲート絶縁膜に用いた周辺トランジスタ部のトランジスタと、	図 8(b)に示すように、メモリセル部 21 には図 8(a)の強誘電体メモリを用い、メモリセルを制御するコントロール回路部(周辺トランジスタ部)やロジック回路部には、図 1 に示したような高誘電体膜をゲート絶縁膜に用いたトランジスタを用いている。	As shown in Fig.8 (b), the transistor which used the high dielectric film as shown in Fig.1 for the gate insulating film is used for the control circuit part (peripheral transistor part) which controls a memory cell, or the logic-circuit unit at memory cell portion 21 using the ferroelectric memory of Fig.8 (a).	As shown in Fig.8 (b), the transistor which used the high dielectric film as shown in Fig.1 for the gate insulating film is used for the control circuit part (peripheral transistor part) which controls a memory cell, or the logic-circuit unit at memory cell portion 21 using the ferroelectric memory of Fig.8 (a).	the logic-circuit unit at memory cell portion 21 using the ferroelectric memory of Fig.8 (a). この英訳では、ロジック回路部に強誘電体メモリを用いることになってしまふ。主語と述語が正しくとらえられていない。	この英訳では、ロジック回路に強誘電体メモリを用いることになってしまふので、技術的に意味をなさない(正しくは、高誘電体膜をゲート絶縁膜に用いたトランジスタを用いる)。	・文中の or の前までは正しく訳されており、この部分だけなので、審決と同じ事実認定ができる。	
これと同様の素子構造でゲート絶縁膜に強誘電体膜を用いたメモリセル部のメモリ強誘電体メモリとを同一のチップに形成する半導体装置であって、	本実施形態は、第 1 実施形態等で説明した高誘電体膜をゲート絶縁膜に用いたトランジスタと、これと同様の素子構造でゲート絶縁膜に強誘電体膜を用いた強誘電体メモリとを同一のチップに形成するものである。	A tongue applied configuration forms in the same tip the transistor which used for the gate insulating film the high dielectric film described by a 1st embodiment etc., and the ferroelectric memory which used the ferroelectrics film for the gate insulating film by the same element structure as this.	A tongue applied configuration forms in the same tip the transistor which used for the gate insulating film the high dielectric film described by a 1st embodiment etc., and the ferroelectric memory which used the ferroelectrics film for the gate insulating film by the same element structure as this.	日本語との対応関係が理解できない。	技術的な内容を把握することは困難。	・日本語は明確なのですが、機械翻訳には向かない日本語なのでどうか。	
周辺トランジスタ部の Si 基板 1 表面に、ゲート絶縁膜 9c として高誘電体膜の Ta2O5 膜を、Si 基板 1 界面との間に界面準位等ができていくように、SiO2 膜を介して堆積し、	この時、Si 基板 1 界面との間にいわゆる界面準位等ができていくように、前処理を行なうとか、或いは、界面に薄い(例えば 1nm 程度)SiO2 膜や Si3N4 膜などを介して高誘電体膜(Ta2O5 膜や(Ba, Sr)TiO3 膜などを)を堆積しても良い。	At this time, it may pretreat or may deposit high dielectric films (Ta2O5 film (Ba, Sr) TiO3 film, etc.) on an interface via SiO2 thin (for example, about 1 nm) film, Si3N4 film, etc. so that the so-called interface state density etc. may not be made easily between Si-substrate 1 interfaces.	At this time, it may pretreat or may deposit high dielectric films (Ta2O5 film (Ba, Sr) TiO3 film, etc.) on an interface via SiO2 thin (for example, about 1 nm) film, Si3N4 film, etc. so that the so-called interface state density etc. may not be made easily between Si-substrate 1 interfaces.	・概ね訳されている。	・概ね意味が通じる。		
メモリセル部の Si 基板 1 表面に、高誘電体膜である Ta2O5 膜を用いる誘電体膜 9a を介して、	誘電体膜 9a は、Si 基板 1 表面に強誘電体膜 9b を安定して形成するためのものであり、例えば膜厚 1nm 程度の Si3N4 膜や膜厚 2nm 程度の高誘電体膜である Ta2O5 膜等を用いることができる。	Dielectric film 9a is for stabilizing and forming ferroelectrics film 9b in the Si-substrate 1 surface. For example, Ta2O5 film etc. which are Si3N4 film of about 1 nm of film thicknesses and a high dielectric film of about 2 nm of film thicknesses can be used.	Dielectric film 9a is for stabilizing and forming ferroelectrics film 9b in the Si-substrate 1 surface. For example, Ta2O5 film etc. which are Si3N4 film of about 1 nm of film thicknesses and a high dielectric film of about 2 nm of film thicknesses can be used.	・酸化タンタルと窒化ケイ素が混同されて訳されている。	・技術的には誤り。	・部分的に取り出せば、事実認定は可能。	
BiSr2Ta2O9 膜を用いる強誘電体膜 9b を堆積した	また、強誘電体膜 9b としては、例えば、Bi4Ti3O12、PZT、PLZT、BiSr2Ta2O9、BaMgF4 等や、1nm 程度の膜厚のエピタキシャル成長させた酸化セリウム(CeO2)の上に形成するエピタキシャル成長させた膜厚 20nm 程度の (Ba, Sr)TiO3 膜を用いることができる。	As ferroelectrics film 9b, for example Bi4Ti3O12, PZT, (Ba, Sr) TiO3 film of about 20 nm of film thicknesses which are formed on PLZT, BiSr2Ta2O9, BaMgF4, etc. and the cerium oxide (CeO2) that an about 1-nm film thickness grew epitaxially and which were grown epitaxially can be used.	As ferroelectrics film 9b, for example Bi4Ti3O12, PZT, (Ba, Sr) TiO3 film of about 20 nm of film thicknesses which are formed on PLZT, BiSr2Ta2O9, BaMgF4, etc. and the cerium oxide (CeO2) that an about 1-nm film thickness grew epitaxially and which were grown epitaxially can be used.	・英訳では、強誘電体膜が BiSr2Ta2O9 膜の上に形成されることになってしまふ。 ・もともと、原文もわかりにくい。	・技術的には誤り。	・審決の事実認定を裏付けることは困難。	
BiSr2Ta2O9 膜を用いる強誘電体膜 9b を堆積した	次に、SiO2 膜 3 を除去して Si 基板 1 表面を露出させ、例えば誘電体膜 9a を介して強誘電体膜 9b を堆積する。	Next, SiO2 film 3 is removed, and the Si-substrate 1 surface is exposed, for example, ferroelectrics film 9b is deposited via dielectric film 9a.	Next, SiO2 film 3 is removed, and the Si-substrate 1 surface is exposed, for example, ferroelectrics film 9b is deposited via dielectric film 9a.				
半導体装置。	本発明は、半導体装置及びその製造方法に関する。	The present invention relates to a semiconductor device and a manufacturing method for the same.	The present invention relates to a semiconductor device and a manufacturing method for the same.	・OK	・OK		
(9)総合評価			V				

評価シートは、主に表 3 においてグレーで示されている 9 つの項目からなる。以下、各項目を項目名の最初の丸括弧付き数字で示す。評価者に提示されるデータは項目(1)～(4)であり²、評価者は項目(5)～(9)を記入する。評価者は次の手順で評価を実施する。

1. 項目(4)の機械翻訳出力から、項目(2)で示されている認定すべき引用発明の事実に対応する翻訳結果部分を項目(5)に抽出する。
2. 項目(5)に抽出した部分について、英語が不明な場合に項目(6)にコメントする。
3. 項目(5)に抽出した部分について、技術的内容が正しく理解できない場合に項目(7)にコメントする。
4. 必要に応じて項目(8)に備考を記入する。
5. 上記 1～4 について、1 文書分の文に対して実施し、これらの結果に基づいて項目(9)に表 1 の評価値を記入する。

英語が不明なために引用発明の事実を認定できないのか、あるいは、翻訳結果の内容は理解できても技術的な内容が正しくないために引用発明の事実を認定できないのかという区別ができるように、これらを別項目（言語評価の項目(6)と技術評価の項目(7)）に分けて問題点を指摘するようにした。

評価済の評価シートは項目(5)～(8)の記述から、翻訳結果のどの部分が評価対象として抽出されて、どのような問題があったのかを検証することが可能なデータとなっている。

4. 評価結果

日英翻訳の評価結果を図 2 に、中英翻訳の評価結果を図 3 に示す。日英翻訳で評価した 3 システムの手法は、JAPIO-1：ルールベース機械翻訳 (RBMT)、EIWA-1：RBMT の翻訳結果を統計的機械翻訳 (SMT) で後編集、NTITI-1：SMT である。中英翻訳で評価した 3 システムの手法は、BBN-1：SMT、RWSYS-1：3 つの手法 (SMT, RBMT, および RBMT の翻訳結果を統計的機械翻訳 (SMT) で後編集) のシステムコンビネーション³、SRI-1：SMT である。

図 2 より、JAPIO-1 システムは 6 割以上の文書で評価 VI を獲得し、全文書が評価 V 以上であった。これより、トップの日英翻訳システムの性能は特許審査で有用なレベルであることが分かった。SMT の NTITI-1 システムも評価 V 以上が 6 割以上あり、ある程度有用であることが分かった。また、図 3 より、BBN-1 システムは 2 割の文書で評価 VI を獲得し、9 割弱の文書で評価 V 以上であった。これより、トップの中英翻訳システムの性能も特許審査で有用なレベルであることが分かった。

² 項目(5)にも項目(4)と同じデータをあらかじめ保存しておき、抽出はそこから評価者が行う。

³ 複数の機械翻訳結果の出力を入力として、最終的な翻訳結果を出力する手法。

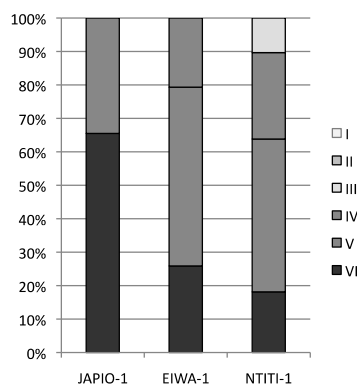


図 2：日英 特許審査評価の結果

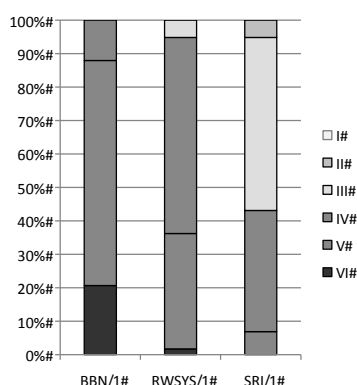


図 3：中英 特許審査評価の結果

中英翻訳のテストデータは特許翻訳の専門家が日英翻訳のテストデータを中国語に翻訳して作成したものであるため、テストデータの技術的な内容は同じである。しかしながら、この評価では、日英翻訳の結果と中英翻訳の結果を公平に比較することはできない。これは次の理由による。実際の日本語特許文書から抽出した文と、それを中国語に翻訳して作成した文では、抽出した日本語文を機械翻訳した場合より、中国語に翻訳して作成した中国語文を機械翻訳した方が低い翻訳品質になる傾向があることが NTCIR-10 の時系列評価において示されている。そのため、本評価の中英翻訳は、実際より不利な条件になっていると考えられるからである。したがって、中英翻訳での本来の有用性は、本評価結果より高いと考えられる。

日英翻訳の *adequacy* および *acceptability* 評価での順位と特許審査評価の順位は一致しており、相関が高かった。

中英翻訳の *adequacy* および *acceptability* での順位と特許審査評価の順位も一致し、相関は高かった。しかし、*adequacy* および *acceptability* では、RWSYS-1 と SRI-1 では統計的に有意な差がない程差が小さく、ほぼ同程度の評価結果であった。それに対して、特許審査評価では、RWSYS-1 と SRI-1 では明らかな差が見られた。このことから、特許審査で重要な要素であるが、*adequacy* や *acceptability* の評価基準では十分に考慮されていない要素があると考えられる。こ

の要素の候補の 1 つとして fluency が挙げられる⁴。自動評価では RWSYS-1 の方が SRI-1 より BLEU スコアが大幅に高かった。文単位の翻訳における RWSYS-1 の BLEU スコアは 0.4006 で SRI-1 の BLEU スコアは 0.3256 であった。BLEU スコアは fluency との関連も強いと考えられるため、有用性には fluency の要素も重要であると予想される。

5. まとめ

NTCIR-10 で実施した特許審査評価の手法概要、詳細な手順、および結果について説明した。特許審査評価で評価した翻訳結果、テストデータ、および評価結果シートは研究用に公開される予定である。このようなデータを一度構築すれば、新しいシステムの評価にも有用である。この評価で評価した翻訳結果を基準として比較することで、他のシステムの特許審査での有用性がある程度調べることができるからである。また、この評価結果のデータは、翻訳結果のうちどの部分が評価対象となり、どのような問題があったのかを検証することが可能である。

謝辞

特許審査評価のテストセット作成および評価に協力していただいた日本知的財産翻訳協会 (NIPTA) 石井理事長他に感謝します。

参考文献

- [1] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. 2013. Overview of the patent machine translation task at the NTCIR-10 workshop. In Proceedings of NTCIR-10, pages 260–286.
- [2] 後藤功雄. NTCIR-9, NTCIR-10 特許機械翻訳タスクでの人手評価. AAMT/Japio 特許翻訳研究会 2012 (平成 24) 年度報告書, pages 79–85.

⁴ Acceptability は fluency の要素も考慮するが、考慮するのは評価が C 以上の文のみであり、RWSYS-1 と SRI-1 はいずれも半数以上が F (C より低い評価) であったため、訳質の低い半数以上の文で考慮できていない。

6. 4 中日特許文評価用テストセットの作成

(株)富士通研究所 長瀬友樹

山梨英和大学 江原暉将

(財)日本特許情報機構 王向莉

1 はじめに

機械翻訳の評価方法として、評価用例文を用意しておき、その翻訳結果がうまく訳出できたかどうかを文法項目毎にチェックすることで翻訳精度を評価する「テストセット評価」が提案されている¹⁾²⁾³⁾。テストセット評価を用いると、翻訳精度を定量化するだけでなく、翻訳エンジンの具体的課題が明らかになるなど、従来の人間評価や自動評価⁴⁾とは異なる観点での評価が可能になる。

AAMT/Japio 特許翻訳研究会評価拡大部会では、2013年度の活動として、中国語特許文献の和訳評価に特化した「中日機械翻訳評価テストセット」を新規に作成した。本稿では、中日機械翻訳評価テストセット（初版）の作成プロセス、およびその仕様について説明する。

2 特許文テストセット作成手順

テストセットの作成では、対訳例文（原文、訳文）の収集と、収集した各例文への設問の付与をしなければならない。今回は特許文が対象であるので、具体的な作業として以下の2通りの方法を検討した。

- 方法1. ①1 パテントファミリのダウンロード
②文アラインメント&チェック
③頻出表現の抽出&チェック
④テストセット文章の選択（又は作成）
⑤設問の付与

- 方法2. ①特許庁等の報告書から頻出表現を選択
②頻出表現を含む文章（中国語、日本語）の検索、ダウンロード
③テストセット文章の選択（又は作成）
④設問の付与

方法1は、中国語と日本語のパテントファミリを検索し、対応する中国文と日本語文の対応付けを行い、頻出の対訳表現を含む例文を集めてテストセットとする方法である。方法2は、頻出の対訳表現を整理済みの文献から例文と設問のヒントを得る方法である。方法1によれば大量の対訳例文をもとに分野別のテストセットを抽出することも可能だが、文献のダウンロード、文章の対応付け、対応する頻出表現を抽出する過程で多大な工数を必要とする。2013年度はこれらの作業に十分な工数を確保できないことから、やむを得ず方法2を進める方針とした。特許表現からの頻出表現抽出では下記の文献を参考にした。

特許庁：日本語特許出願書類の中国語への機械翻訳に関する調査報告書（平成 23 年）
http://www.jpo.go.jp/shiryoutoushin/chousa/pdf/tokkyo_kikai_honyaku_houkoku/chn_all.pdf

3. テストセット作成状況

1) ステップ① 特許庁等の報告書から頻出表現を選択

参考文献は、日本語キーワードから機械翻訳によって生成した中国語文字列をキーにして中国語の特許文件データベースを検索する場合の日中機械翻訳の課題をまとめた調査報告書である。調査内容には「定型化可能な表現の分析」が含まれており、この中で日本公開特許広報の各記載項目における定型文、定型パターン文、定型フレーズ、主文パターンおよび節パターンが対訳で整理されている。報告書に掲載されている文パターン、節パターンの例を、それぞれ表 1、表 2 に示す。

表 1 文パターンの例

日本語	中国語	頻度
<\$1>ことを特徴とする<\$2>。	<\$2>，其特征在于：<\$1>。	17762
<\$1>ことを特徴とする請求項<\$2>に記載の<\$3>。	如权利要求<\$2>所述的<\$3>，其特征在于：<\$1>。	8111
<\$1>ことを特徴とする請求項<\$2>記載の<\$3>。	如权利要求<\$2>所述的<\$3>，其特征在于：<\$1>。	6872
<\$1>ことを特徴とする請求項<\$2>又は<\$3>に記載の<\$4>。	如权利要求<\$2>或<\$3>所述的<\$4>，其特征在于：<\$1>。	807
<\$1>ことを特徴とする請求項<\$2>のいずれか<\$3>に記載の<\$4>。	如权利要求<\$2>中任<\$3>所述的<\$4>，其特征在于：<\$1>。	764
<\$1>ことを特徴とする<\$2>のいずれか<\$3>に記載の<\$4>。	<\$2>中任<\$3>所述的<\$4>，其特征在于：<\$1>。	764

表 2 節パターンの例

日本語	中国語	頻度
<\$1>を有する<\$2>	具有<\$1>的<\$2>	911
請求項<\$1>に記載の<\$2>であって、	为权利要求<\$1>中所述的<\$2>，	470
請求項<\$1>記載の<\$2>であって、	为权利要求<\$1>中所述的<\$2>，	252
<\$1>を構成する<\$2>	构成<\$1>的<\$2>	178
<\$1>を検出する<\$2>	检测<\$1>的<\$2>	158
<\$1>を有し、	具有<\$1>，	157
<\$1>に設けられた<\$2>	设置于<\$1>上的<\$2>	139
前記<\$1>は<\$2>であり、	所述<\$1>为<\$2>，	136
前記<\$1>を構成する<\$2>	构成所述<\$1>的<\$2>	136

報告書に記載されている定型文、定型パターン文、定型フレーズ、主文パターンおよび節パターンの件数をまとめたものが表3である。これらの文、フレーズ、パターンをテストセットの対訳例文作成の参考とした。ただし、調査報告は日中翻訳を対象としたものであるため、異なる日本語表現に同一の中国語表現が対応する例が多数含まれるという問題がある(表1の2-3行目参照)。今回作成するテストセットは逆方向の中日翻訳を想定したものであるため、重複する中国語表現をひとつにまとめた結果、実際に抽出できた対訳表現数は表3よりも3割程度少ない件数となった。

表3 報告書に掲載されている対訳例の内訳

	要約	請求項	明細書	計
定型文	7	8	12	27
定型パターン	10	10	10	30
定型フレーズ	27	24	33	84
主文パターン文	10	10	10	30
節パターン	16	13	17	46
計	70	65	82	217

2) ステップ② 中日ファミリー特許文献のダウンロード

パテントファミリーの情報をもとに、中日で内容が対応する特許をダウンロードすることができる。この方法で48組の中日ファミリー特許のペアをダウンロードした。

3) ステップ③ 頻出表現を含む文章の検索およびテストセット文章の選択

ステップ②でダウンロードした中日の対訳ファイルから、ステップ①で整理したパターンにマッチする文対を抜き出し、これをテストセットに登録する例文の候補とした。しかし下記の理由により例文候補の数が大きく制限された。

- ① 中日特許対としてダウンロードできたのは請求項と明細書のみで、要約をダウンロードすることができなかった。
 - ⇒ 要約から抽出した文、パターン、フレーズを削除し、請求項と明細書から抽出したもののみをテストセット文候補とした。
- ② 定型文や定型フレーズには「課題」、「本発明」のような語彙レベルの対訳が含まれ、テストセットとしての有効性が低いと思われる登録が少なくなかった。
 - ⇒ 定型パターン文、主文パターン文、節パターンから抽出できた文例のみを、テストセット文候補とした。

上記によりテストセット文の選択を行った結果、テストセット文候補の数は29文にまで減ってしまった。

4) ステップ④ 設問の付与

- ・テストセット文のネイティブ検証

設問付与の前に、テストセット候補文の日本語と中国語が、それぞれ文法的に正しい文章になっているか、また、両言語の文章が内容的に一致しているかどうかについて、日中バイリンガルによってチェックを行った。また、長いテストセット候補文3文については2文に分割し、これによってテストセット候補文は合計32文となった。

- ・構文的妥当性を問う設問の付与

テストセット候補文に対して、1文ずつ構文的妥当性を問う設問を設定した。テストセット候補文を選択するときに用いた対訳パターンの日本語側パターンが、機械翻訳の訳文の中に含まれているかどうかを問う設問を基本として設定した。

- ・語彙的妥当性を問う設問の付与

特許文の品質評価の軸としては、構文的妥当性ととも語彙的妥当性が重要である。テストセットを用いて語彙的妥当性のチェックができるように、テストセット候補文から中国語と日本語の専門用語の対を抽出し、機械翻訳の訳文の中に日本語専門用語が含まれるかどうかを問う設問を設定した。ひとつの例文に2箇所以上の専門用語を含む場合もあるため、32文の候補文から50の語彙を問う設問が追加された。

4. おわりに

これまで説明してきた方法により、規模は小さいながら中日特許文テストセットの初版が完成した。テストセットの仕様は下記のとおりである。

文数	: 32 文	(内訳) 要約	: 7 文
		請求項	: 2 文
		明細書	: 23 文
設問数	: 104	(内訳) 構造についての設問	: 54
		語彙についての設問	: 50

テストセット評価では、テストセットを分野やパート別に作成すれば、分野/パート別の評価が可能である。たとえば、IPCの大分類ごとに作成した複数のテストセットを用いれば、翻訳エンジン毎の分野別の品質特性などを簡単に調べることができる。特許文件が持っている書誌情報を活用して、年代、業種別のテストセットを作成して、詳細な評価をすることも可能である。

今年度の活動から、パテントファミリの情報を用いて内容的に対応している中国語、日本語のファイルをダウンロードするためのツール⁵⁾があれば、容易に大量の特許文テストセットを作成できそうな感触を得た。今後、テストセットの規模を大きくしていく中で様々な可能性を検討していきたい。

References

- 1) Isahara, H. 1995. JEIDA's Test-Sets for Quality Evaluation of MT Systems --Technical Evaluation from the Developer's Point of View--. *Proc. of MT Summit V*.
- 2) Uchimoto, K., K. Kotani, Y. Zhang and H. Isahara. 2007. Automatic Evaluation of Machine Translation Based on Rate of Accomplishment of Sub-goals. *Proc. of NAACL HLT*, 33-40.
- 3) Nagase, T., H. Tsukada, K. Kotani, N. Hatanaka and Y. Sakamoto. 2011. Automatic Error Analysis Based on Grammatical Questions . *Proc. of PACLIC*.
- 4) Papineni K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proc. of ACL*, 311–318.
- 5) Higuchi, S. M. Fukui, A Fujii and T. Ishikawa. 2001. A system for multi-lingual patent retrieval. *Proc. Of MT Summit VIII*.

以上

第14回機械翻訳サミット(Machine Translation Summit XIV) および 第5回特許翻訳ワークショップ (The 5th Workshop on Patent Translation)参加報告

山形大学 横山晶一
静岡大学 梶 博行
愛媛大学 二宮 崇

1. 会議概要

第14回機械翻訳サミット(Machine Translation Summit XIV、以下 MT Summit XIV と略称)は、2013年9月2~6日、フランス、ニース(Nice)の Acropolis Conference Centre で開催された。会場は、ニースの中心街や旧市街に近い静かな場所で、アクセスも便利な場所にある。

本会議は、4日~6日の3日間で、参加者は主催者側発表で229名であった。本会議に先立って2日と3日に6つの Tutorial と4つの Workshop が行われた。Tutorial の参加者は計71名、Workshop の参加者は計135名であった。なお、3日に歓迎レセプション、5日にバンケットが開かれた。

AAMT/Japio 特許翻訳研究会関係では、筆者ら3名の他に、宮澤信一郎(秀明大学)、宇津呂武仁(筑波大学)の各委員、Japio から、松田成正部長、星山直人主任が参加した。

本会議では、2つの招待講演、4つのユーザ講演と1つのパネル討論が行われた。研究論文は73編の応募のうち37編(約50%)が採択され、ユーザ論文は、24編の応募のうち17編(約71%)が採択された。これらの論文は3並列のオーラルセッション(研究2、ユーザ1)とポスターセッション(研究のみ)で発表された。また、FP7プロジェクトのポスターセッションで20あまりのEUプロジェクトの発表があった。

以下に筆者らがオーガナイズした特許翻訳ワークショップと本会議の概要を報告する。

2. 第5回特許翻訳ワークショップ

このワークショップは、AAMT/Japio 特許翻訳研究会(委員長: 辻井潤一)における種々の議論をもとに、MT Summit でワークショップを開いて、この分野の研究者の意見交換ができればよいという趣旨の下で開催された。過去には第1回が Phuket (Chair: 横山晶一)、第2回が Copenhagen (Co-Chair: 辻井潤一、横山晶一)、第3回が Ottawa (Co-Chair: 江原暉将、横山晶一) 第4回が厦門 (Co-Chair: 横山晶一、江原暉将、王丹(CIPO)) で開催されている。

今回は横山晶一、梶博行が Co-Chair として開催した。以下、プログラムの発表順に、概要を簡単に述べる。以下の時間は、そのセッショントータルの時間を示す。間で空いている時間は、休憩や昼食に当てられている。

(1) 招待講演(9:00-10:40) (司会：梶博行)

Chair の横山晶一による開会のあいさつの後、2つの招待講演が行われた。

・Paul Schwander (European Patent Office): Machine Translation at the EPO – Removing Language Barriers from Patent Documentation

EPO の機械翻訳への戦略的な取り組みについて EPO の情報管理国際プログラムの責任者であるポール・シュワンダー氏が講演した。内容は以下のとおり。特許情報は膨大で機械翻訳が必須。審査官は機械翻訳の結果をみて人手翻訳すべき特許を選択している。中国特許庁の出願の伸びが著しく、日本語や韓国語を含むアジア言語の特許情報へのアクセスが EPO でも大きな課題となっている。Google と協力して特許情報の機械翻訳サービス Patent Translate を 2012 年 2 月に開始した。現在、日本語、中国語を含む 21 の言語と英語の間の翻訳がサポートされ、3 億 500 万件の翻訳特許文書にアクセスすることができる。訳文の質を 5 段階で評価するとともに質の向上をはかっている。2014 年末までに 32 言語に拡大する計画で、その中には韓国語やロシア語が含まれる。

・Hitoshi HONDA (Japan Patent Office): Current Status of Machine Translation System in the JPO

日本の特許庁における機械翻訳システムの利用について特許庁・特許情報企画室・特許情報企画調査係長の本田仁氏が講演した。内容は以下のとおり。審査官の業務の 1 ステップである先行技術のサーチにおいて外国語文書へのアクセスを容易にする手段として機械翻訳が果たす役割は大きい。現在、米国や欧州の特許については日本語抄録（人手翻訳）と英語の全文テキスト（原文）を利用することができるが、韓国や中国の特許については英文抄録（人手翻訳）しか利用できない。2014 年度中には韓国、中国特許の機械翻訳による日本語全文テキストを提供する予定である。外国の特許庁との間の審査情報の相互利用にも力を入れており、日本特許の審査情報の日英機械翻訳結果を EPO や USPTO（米国特許庁）に提供している。EPO や USPTO のユーザ（審査官）からはフィードバックがあり、システムの改良に役立っている。また、中国特許庁の出願件数の著しい増加に鑑みて、中日機械翻訳システムの構築に力を入れている。2012 年には、中国と日本の対応する特許のテキストから対訳関係を抽出し、100 万語規模の日中専門用語対訳辞書を構築した。

(2) 一般講演 1(11:10-12:10) (司会：二宮崇)

ここでは 3 件の発表が行われた。

・Svetlana Sheremetyeva: On Integrating Hybrid And Rule-Based Components For Patent MT With Several Levels Of Output

Sheremetyeva による発表。この論文の提案手法では、意味役割付与まで行う構文解析を用い、原言語の意味表現から翻訳先言語の意味表現に変換し、翻訳先言語の文を生成することで機械翻訳を行っている。システムは、NP チャンカーや VP チャンカーを用いる浅いモジュールと意味役割まで扱う深いモジュールから構成され、深いモジュールでは、意味役割のついた述語項構造が用いられている。また、解析結果を修正するインタラクティブモードや、単文の解析を出力す

るモードを備えている。まだ大規模な実験は行われていないが、より多くの知識やルールを記述することで実用的な機械翻訳システムの実現を目指している。

- Itsuki Toyota, Zi Long, Lijuan Dong, Takehito Utsuro, and Mikio Yamamoto: Compositional Translation of Technical Terms by Integrating Patent Families as a Parallel Corpus and a Comparable Corpus

Toyota による発表。パラレルコーパスとコンパラブルコーパスを用いた構成的な対訳辞書生成手法の提案をしている。特許翻訳の学習および評価でよく用いられているパラレルコーパスは、コンパラブルコーパスから対訳文対を抽出することにより生成されており、そのため元のコンパラブルコーパスのうち 30%程度しか用いられていない。従って、パラレルコーパスからの対訳辞書抽出では残りの 70%が有効に利用されていない。この論文の提案手法では、コンパラブルコーパスから対訳専門用語の候補を抽出し、英辞郎およびパラレルコーパスから学習した翻訳テーブルを用い、構成的に対訳候補のスコアを計算し、対訳専門用語を決定する。この手法によりコンパラブルコーパスの残り 70%から対訳専門用語対が得られる。この手法により比較的高い精度でより多くの対訳語対が得られることが実験により示された。

- Shoichi Yokoyama: Analysis of Parallel Structures in Patent Sentences, Focusing on the Head Words

Yokoyama による発表。特許データベースから、並列構造を形成していると思われる主辞(head word)を抽出し、頻度の高い語を用いて、並列構造の誤りを修正した結果、58%程度の誤りを訂正できたという結果を述べた。

(3) 基調講演(14:30-15:30) (司会：横山晶一)

- Philipp Koehn (University of Edinburgh, Great Britain):Advances in Machine Translation for Patent Translation and Beyond

午前中に自らの Tutorial をこなし、わざわざこの基調講演を引き受けてくれたことをまず感謝したい。内容は、本会議の招待講演や、この会議を通じて常に話題になっていたトピックである Domain Adaptation や異なるモデルの combination、最近の Operation Sequence Modelなどを分かりやすく説明し、最後に big data に対する挑戦について述べるといった、近年の MT の進展を解説するものであった。

(4) 一般講演 2(16:00-16:40) (司会：Svetlana Sheremetyeva)

2 件の発表があった。

- Yun Jin and Oh-Woog Kwon, Seung-Hoon Na, and Young-Gil Kim: Patent Translations as Technical Document Translation: Customizing a Chinese-Korean MT System to Patent Domain

韓国のグループによる発表。特許文書と技術文書の類似性に着目して、Domain Adaptation に技術文書をもちいることにより、中韓 MT システムの性能向上を図った結果について述べた。

- Rahma Sellami and Fatiha Sadat, and Lamia Hadrach Belguith: Exploiting Multiple

Resources for Japanese to English Patent Translation

チュニジアとカナダのグループによる発表。NTCIR-10 の日英特許翻訳システムの翻訳性能を改善するために Domain Adaptation や resource の拡張を行うのに、日英で同じと思われる Wikipedia を使用したという内容であった。

(5) まとめ

横山による締めくくりの挨拶。次回も行いたいという趣旨のことを述べた。

3. 本会議

3.1. 招待講演、ユーザプレゼンテーション

招待講演は、4日（今回の主催者である Andy Way の開会挨拶に続いて）、5日にそれぞれ1つずつ行われた。

- Hinrich Schuetze: The operation sequence model: integrating translation and reordering operations in a single left-to-right model (4日)

ミュンヘン大学教授による講演。これまでの PB-SMT など、種々のモデルの利点と欠点を分かりやすく解説するとともに、それらを統合、克服するモデルとしての operation sequence model について解説し、その現状と将来展望について述べた。

- Harold Elsen: Demystifying Machine Translation: Learning from the real world (5日)

ドイツの DELTA International CITS GmbH の managing director による講演。SMT と RMT とのさまざまな比較や、MT と TM の結合、術語、post-edit など、リアルワールドにおけるさまざまな問題点をあげて、MT の将来像はこうあるべきだという持論を述べた。

ユーザプレゼンテーションは、Hans Uszkoreit による “Including the User in Research and Application Improvement” (ただし DCU のメンバーによる代読) のほか、LionBridge, IBM Germany, Microsoft の3社が発表した。

3.2. 一般講演、ポスターセッション

一般講演は、論文採録のところでも述べたように、Research and Development 2つと User Track 1つに分かれて、3つのセッションが並行して進められるという形式で行われた。

話題は、今回は比較的狭い範囲に絞られたような感想を持っている。

まず一つは Domain Adaptation である。これは、たとえば初日の招待講演の直後に行われた Patrick Lehen et al.: Conditional Random Field Using Intermediate Classes for Statistical Machine Translation や、George Foster et al.: Simulating Discriminative Training for Linear Mixture Adaptation in Statistical Machine Translation (ベストペーパー賞を獲得)、Katharina Waeschle et al.: Generative Discriminative Methods for Online Adaptation in SMT などがそれにあたる。

Terminology も主要なテーマである。岡山大の竹内孔一氏らのグループの発表、Koichi Satoh et

al.: Terminology-driven Augmentation of Bilingual Terminologies などである。

また、pre-editing, post-editing も話題の一つで、Guillaume Wisniewski et al.: Analysis of a Large Corpus of Post-Edited Translations: Quality Estimation, Failure Analysis and the Variability of Post-Editation や、Michel Simard et al.: PEPr: Post-Edit Propagation Using Phrase-based Statistical Machine Translation、ポスターで発表された Lucia Morado Vazquez et al.: Comparing Forum Data Post-Editing Performance Using Translation Memory and Machine Translation Output: A Pilot Study などがある。

Estimation もテーマの一つで、Kahif Shah et al.: An Investigation on the Effectiveness of Features for Translation Quality Estimation, Sudip Kumar Naskar et al.: Meta-Evaluation of a Diagnostic Quality Metric for Machine Translation, Ahmed Hamdi et al.: The Effect of Factorizing Root and Pattern Mapping in Bidirectional Tunisian- Standard Arabic Machine Translation, Luis Formigo et al.: Real-Life Translation Quality Estimation for MT System Selection, Pratyush Banerjee et al.: Quality Estimation-guided Data Selection for Domain Adaptation of SMT, Maaxi Li et al.: Listwise Approach to Learning to Rank for Automatic Evaluation of Machine Translation などがある。

3.3. FP7 プロジェクトセッション

EU の The 7th Framework Programme for Research and Technological Development (第7次研究・技術開発フレームワークプログラム, 2007~2013年) の支援を受けたプロジェクトのセッションで、以下のプロジェクトのポスター発表があった。ヨーロッパでは機械翻訳技術の実用化、応用がさまざまな形で進められていることがわかる。

- ACCEPT (ジュネーブ大学ほか) : インターネットコミュニティコンテンツの機械翻訳
- BOLOGNA (CrossLang ほか) : 統計的機械翻訳とポストエディット、翻訳メモリを統合したシラバス翻訳サービス
- CASMACAT (エディンバラ大学ほか) : 翻訳プロセスの認知モデルに基づく次世代翻訳ワークベンチ
- CNGL (ダブリンシティ大学) : 多言語・マルチモーダル・マルチメディアコンテンツの生成、処理、統合
- EUBRIDGE (カールスルーエ工科大学ほか) : テレビのキャプション、大学の講義、EU 議会などの翻訳
- EXCITEMENT (Bar Ilan 大学ほか) : 多言語テキスト推論プラットフォーム
- EXPERT (Pangeanic ほか) : ハイブリッド言語翻訳技術
- FAUST (ケンブリッジ大学ほか) : リアルタイムでユーザーフィードバックに適応する機械翻訳システム
- Let'sMT (コペンハーゲン大学ほか) : 個人に合わせてカスタマイズされた翻訳システムの構築ツール
- MateCat (FBK ほか) : 機械翻訳のコンピュータ支援翻訳への統合

- METANET (DFKI ほか) : 翻訳品質評価 QTLaunchPad、Web コンテンツのためのメタデータ MultilingualWeb-LT など
- MOLTO (イェーテボリ大学ほか) : 制限言語に基づく多言語オンライン翻訳
- MONNET (DELI ほか) : オントロジーの翻訳
- MosesCore (エディンバラ大学ほか) : オープンソース機械翻訳の開発と利用
- PANACEA (Pompeu Fabra 大学ほか) : 言語資源工場 (機械翻訳のための言語資源の獲得、生成、メンテナンス)
- QT (ダブリンシティ大学ほか) : 翻訳品質評価技術
- SIGNSPEAK (アーヘン工科大学ほか) : 手話認識・翻訳
- SUMAT (Deluxe Digital Studios ほか) : AV コンテンツのサブタイトルの機械翻訳
- TaaS (Tilde) : 多言語ターミノロジーの獲得、処理、再利用
- transLectures (バレンシア技術大学ほか) : ビデオ講義の転写・翻訳
- TTC (ナント大学) : コンパラブルコーパスからのターム抽出
- XLIKE (ザグレブ大学ほか) : クロスリンガル知識抽出

3.4. パネルディスカッション

パネルは、The MT Tower of Babel という題で、司会者と 6 名のパネリストが壇上に上がった。MT システムはすでに 50 年以上も前から作られているのに、Google や Bing が出てくるまではほとんど一般の関心をひかなかったのはどういうわけかというやや刺激的な問いかけで、研究者や翻訳会社からのパネラーが討論した。ただ、この種のパネルの常として、時間が足りなかったのが残念であった。

4. まとめと考察

本会議も、ワークショップも、現在の機械翻訳のかかえる問題点が明確になり、大変有意義であった。機械翻訳関係の研究者、ユーザ、メーカー等が世界レベルで一同に会する機会はこの会議以外には非常に少ない。そのためにも今後ともこの会議の継続と発展を祈っている。今回の会議では、free の翻訳者も結構参加していたのは、日本では余り見られない現象であった。

なお、会議の最後に IAMT Award of Honor が、SMT の貢献などに対する業績で、Hermann Ney 教授に授与された。また、この分野で長年活躍してきた John Hutchins 氏に特別表彰が行われ、会場は standing ovation で、氏をたたえた。次回の会議は、EAMT 会長の Andy Way 氏から、AMTA 顧問 (前会長) の Alon Lavie 氏に IAMT 会長が引き継がれ、アメリカ Miami で 2015 年 10 月下旬に ATA National Conference と連動して開催の予定である。なお、AAMT の中岩浩巳会長が IAMT 副会長に就任した。また、EAMT-2014 の会議は、クロアチアの世界遺産都市 Dubrovnik で、2014 年 6 月 16~18 日に行われることも合わせて発表された。

今回のアレンジや論文査読などに委員会の多くのメンバーの助力があったことに感謝する。第 6 回のワークショップも、今後の議論の中で詰めてゆきたい。

————— 禁 無 断 転 載 —————

平成 25 年度 AAMT/Japio 特許翻訳研究会報告書
(機械翻訳及び機械翻訳評価に関する研究及びワークショップ報告)

発行日 平成 26 年 3 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 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