

平成 20 年度 AAMT/Japio 特許翻訳研究会

報 告 書

機械翻訳及び辞書構築に関する研究

平成 21 年 3 月

財団法人 日本特許情報機構

# 目 次

1 .	はじめに .....	1
	辻井 潤一 東京大学・AAMT / Japio 特許翻訳研究会委員長	
2 .	機械翻訳の評価手法	
2 . 1	NTCIR-7 データを用いた機械翻訳自動評価規準のメタ評価 .....	2
	越前谷 博 北海学園大学 江原 暉将 諏訪東京理科大学	
	下畑 さより 沖電気工業(株) 藤井 敦 筑波大学	
	内山 将夫 (独)情報通信研究機構 山本 幹夫 筑波大学	
	宇津呂 武仁 筑波大学 神門 典子 国立情報学研究所	
2 . 2	文分解法によるコーパス規模と Bleu 値との関係 .....	14
	宮澤 信一郎 秀明大学 岡田 勇 創価大学	
3 .	翻訳辞書の自動構築	
3 . 1	専門用語抽出アルゴリズムを用いた対訳コーパスからの対訳辞書の自動生成 .....	23
	範 曉蓉 東京大学 二宮 崇 東京大学	
	中川 裕志 東京大学	
3 . 2	第 3 言語を介した対訳辞書の構築 .....	30
	梶 博行 静岡大学	
3 . 3	対訳特許文書からの専門用語対訳辞書生成：機械学習による異種情報源の統合 .....	41
	森下 洋平 筑波大学 宇津呂 武仁 筑波大学	
	山本 幹雄 筑波大学	
3 . 4	対訳辞書および日中漢字の対応関係を用いた特許文書における専門用語の中日翻訳 .....	52
	王 向莉 東京大学 綱川 隆司 東京大学	
	辻井 潤一 東京大学	
4 .	翻訳テキストのアラインメント	
	句の依存関係確率モデルを用いた統計的対訳文アラインメント .....	58
	中澤 敏明 京都大学 黒橋 禎夫 京都大学	
5 .	特許文の構造的な特徴と構文解析	
	特許文への構文解析適応 .....	65
	横山 晶一 山形大学	

## 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	東京大学大学院教授・AAMT 前会長
副委員長	横山 晶一	山形大学大学院教授
〃	江原 暉将	諏訪東京理科大学教授
委員	宮澤 信一郎	秀明大学教授
〃	梶 博行	静岡大学教授
〃	黒橋 禎夫	京都大学大学院教授
〃	宇津呂 武仁	筑波大学大学院准教授
〃	二宮 崇	東京大学講師
〃	橋本 力	山形大学大学院助教・(独)情報通信研究機構
〃	越前谷 博	北海学園大学助教
〃	隅田 英一郎	(独)情報通信研究機構・(株)国際電気通信基礎技術研究所
〃	熊野 明	(株)東芝
〃	下畑 さより	沖電気工業(株)
〃	潮田 明	(株)富士通研究所
〃	三浦 貢	日本電気(株)
事務局	村上 嘉陽	AAMT/Japio 特許翻訳研究会東京事務局・(株)ナビックス
〃	河田 容英	〃 〃 〃
〃	高田 佳代子	〃 〃
オブザーバー	中川 裕志	東京大学教授
〃	安藤 進	多摩美術大学講師
〃	綱川 隆司	東京大学大学院
〃	星野 綾子	〃
〃	範 暁蓉	〃
〃	王 向莉	〃
〃	守屋 敏道	(財)日本特許情報機構
〃	渡邊 豊英	〃
〃	藤城 享	〃
〃	埴 金治	〃
〃	大塩 只明	〃
〃	三橋 朋晴	〃
〃	柿田 剛史	〃
〃	星山 直人	〃

# 1. はじめに

東京大学大学院情報理工学系研究科 教授  
マンチェスター大学コンピュータ科学科 教授(兼任)  
AAMT/Japio 特許翻訳研究会委員長  
辻井 潤一

機械翻訳の研究・開発には、50年にわたる長い歴史がある。非常に楽観的にその可能性が議論された時期、逆に悲観的な見解の時期、ほとんどの研究・開発が中断された時期という、両極を行き来しながら、やっと、技術の客観的な評価ができる時期に来たように思える。現在は、機械翻訳の可能性といったことが議論されることはなく、技術をどのように使って全体の翻訳コストを低減するかに議論が集中するようになった。この中で、特許翻訳というのは使用目的が明確で、かつ、市場価値も明らかであるという点で、技術の格好の適用分野となっている。

実際、多言語社会のヨーロッパでは、翻訳は特許業務の主要な部分を占めており、デンマークでは、特許のための機械翻訳を運用することでコストを軽減し利益を上げている会社もある。

多言語社会とは無縁であったアジアでも、中国や韓国の台頭により、特許における翻訳の問題が注目を浴びている。この地域では、英語を中間においた多言語状況が続くと思われるが、経済規模の拡大とそれに伴う技術開発の活発化とともに、英語を中核にしながらも、地域内での技術情報の交流が進んでいる。

このような状況のもとで、AAMT (アジア太平洋機械翻訳協会)の特許翻訳研究は、Japio(日本特許情報機構)からの委託を受け、本年度も8回にわたって活発な研究調査活動を行った。本年度は、単なる調査の段階から、より技術的な焦点を明確にした研究活動を展開し、Japioからの提供された辞書、テキストコーパスを大学、研究機関、民間企業の研究グループがそれぞれのシステムで処理することで、独創的な研究活動が展開できたことであろう。また、本研究会と並行して行われた国立情報学研究所主催のNTCIRの機械翻訳システムのコンペティションと緊密な連携をとることで、研究会メンバーの範囲を超えて広い範囲の調査活動を行うことが可能となった。

具体的には、NTCIRでのシステム評価の結果を参考にして、翻訳の質を定量的に評価するための手法に関して、調査と具体的な提言をまとめたこと、特許翻訳への統計的翻訳手法の適用、専門用語辞書の半自動的な構築手法、より柔軟な訳語の選択機構に向けての予備実験を行った。これらは、現在の統計的機械翻訳の主流となっている米国の研究開発と比べても、独創的な観点を持つものであり、今後の発展が期待できるものである。

本研究会は、平均月1回のペースで委員会に属する委員が各グループでの研究成果や知見をもちよる形で活動している。したがって、委員会独自に大規模な研究開発を実施するものではないが、日本における主要な機械翻訳グループからの委員が参集する本研究会は、日本における機械翻訳研究、特に特許翻訳に独自の貢献をするものと期待している。本報告書は、平成20年度の調査研究の成果をまとめたものである。諸賢のご意見、ご批判をいただければ幸いである。

## 2.1 NTCIR-7 データを用いた機械翻訳自動評価基準のメタ評価

北海学園大学 越前谷 博  
諏訪東京理科大学 江原 暉将  
沖電気工業株式会社 下畑 さより  
筑波大学大学院 藤井 敦  
(独) 情報通信研究機構 内山 将夫  
筑波大学大学院 山本 幹雄  
筑波大学大学院 宇津呂 武仁  
国立情報学研究所 神門 典子

### 2.1.1 はじめに

現在の情報化社会において不可欠な機械翻訳の普及を背景に、機械翻訳システムに対する自動評価基準へのニーズが高まっている。今日、業界標準となっている自動評価基準として BLEU[1] が挙げられる。BLEU は人手による主観評価との間で高い相関が得られることが知られている。しかし、BLEU は複数の文からなるドキュメント単位での使用を前提としており、文単位での自動評価においては信頼性が不十分であることが問題点として指摘されている。ドキュメント単位において高い相関が得られたとしても、その中の個々の文において十分な相関が得られないということは、人間が行う言語処理とは大きな隔たりが存在していることを意味する。そこで、我々は、今後、より信頼性の高い自動評価基準を実現する上で、既存の自動評価基準がどれほどの性能を有するのかを把握することは非常に重要であるとの考えから、文単位での自動評価の観点より、NTCIR-7 データを用い、これまでに提案されている代表的な機械翻訳自動評価基準のメタ評価を行った。本報告ではその結果について詳細を報告する。NTCIR-7 データに関しては[2]を参照されたい。

### 2.1.2 翻訳評価基準

今回使用した機械翻訳自動評価基準を以下に示す。

- IMPACT(recursive acquisition of Intuitive comMon PArts ConTinum) [3]
- ROUGE-L(Recall-Oriented Understudy for Gisting Evaluation)[4]
- BLEU(Bilingual Evaluation Understudy)
- NIST(National Institute of Standards and Technology)[5]
- NMG\_REF(Normalized Mean Grams)[6]
- NMG\_COR[7]
- Charniak score[7]
- WN
- NMG\_WN
- METEOR(Metric for Evaluation of Translation with Explicit ORdering)[8]

- GTM(General Text Matcher)[9]
- WER(Word Error Rate)[10]
- PER(Position Independent Word Error Rate)[11]

これらの 13 の自動評価基準について、それぞれその概略を以下に述べる。

### 2.1.2.1 IMPACT

IMPACT は大きく 2 つの処理からなる。始めに、2 つの文  $X$ 、 $Y$  の間の LCS (最長共通部分列) を一意に決定する。次いで、一意に決定された LCS を除き、2 つの文  $X$ 、 $Y$  の間の LCS を再帰的に決定することでスコアを求める。LCS の一意の決定処理は、複数の LCS 経路が存在する場合、以下の式(2.1.1)、(2.1.2)より得られるスコアが最も高い LCS 経路を選択することで行う。

$$pos = (1.0 - | \frac{posX}{m} - \frac{posY}{n} |) \leq 1.0 \quad (2.1.1)$$

$$Score = ( \sum_{c \in LCS} (length(c)^\beta \times pos) )^{\frac{1}{\beta}} \quad (2.1.2)$$

式(2.1.1)の  $pos$  は LCS 経路を構成する共通部分の 2 つの文  $X$  と  $Y$  中における相対位置のずれを示し、値が大きいほどずれが小さいことを意味する。式(2.1.2)の  $Score$  は個々の共通部分の構成単語数の大きさと式(2.1.1)の  $pos$  に基づき決定される。したがって、式(2.1.1)と式(2.1.2)より共通部分の相対位置のずれが小さく、かつ、共通部分の構成単語数が大きい LCS 経路が選択されることになる。式(2.1.1)の  $m$ 、 $n$  はそれぞれ、参照訳の単語数と MT 訳の単語数である。

上述した処理より決定された LCS を除き、2 つの文  $X$ 、 $Y$  の間に LCS が存在しなくなるまで再帰的に LCS を決定する。そして、以下の式よりスコアを計算する。

$$R_{IP} = \left( \frac{\sum_{i=0}^{RN} (\alpha^i \sum_{c \in CC} length(c)^\beta)}{m^\beta} \right)^{\frac{1}{\beta}} \quad (2.1.3)$$

$$P_{IP} = \left( \frac{\sum_{i=0}^{RN} (\alpha^i \sum_{c \in CC} length(c)^\beta)}{n^\beta} \right)^{\frac{1}{\beta}} \quad (2.1.4)$$

$$IP(X, Y) = \frac{(1 + \gamma^2) R_{IP} P_{IP}}{R_{IP} + \gamma^2 P_{IP}} \quad (2.1.5)$$

$$\gamma = \frac{P_{IP}}{R_{IP}} \quad (2.1.6)$$

式(2.1.3)、(2.1.4)の  $i$  は再帰処理の回数をカウントする際に用いられる。また、式(2.1.3)、(2.1.4)では、一意に決定された LCS を  $CC$  で表す。LCS が存在しなくなった時点でカウントは終了する。 $c$  は LCS を構成する個々の共通部分であり、 $length(c)$  は共通部分の構成単語数を示している。パラメータ  $(0.0 < \alpha < 1.0)$  は個々の  $CC$  に対する重み付けであり、値が小さいほど語順の違いがスコアに反映れることになる。パラメータ  $(0.0 < \beta < 1.0)$  は個々の共通部分の構成単語数に対する重

み付けであり、値が大きいほど共通部分が個々に独立して存在する場合よりも連続して存在する場合の方がスコアは高くなる。式(2.1.5)、(2.1.6)より参照訳と MT 訳においてそれぞれ  $R_{IP}$  (再現率)、 $P_{IP}$  (適合率) を求め、その F 値を自動評価のスコアとしている。また、IMPACT が出力するスコアは大きいほど評価が良い。

#### 2.1.2.2 ROUGE-L

ROUGE-L は、2 つの文  $X$  と  $Y$  の間の LCS (最長共通部分列) に基づく評価基準である。LCS の値に対して参照訳の単語数と MT 訳の単語数を用いて正規化することにより再現率と適合率を求める。そして、その再現率と適合率の F 値を求め、その値をスコアとする。ROUGE-L が出力するスコアは大きいほど評価が良い。

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad (2.1.7)$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \quad (2.1.8)$$

$$F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta_2 P_{lcs}} \quad (2.1.9)$$

$$\beta = \frac{P_{lcs}}{R_{lcs}} \quad (2.1.10)$$

#### 2.1.2.3 BLEU

BLEU は最も多く利用されている自動評価基準である。BLEU は以下の式の通り、n-gram マッチ率に基づく手法である。

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (2.1.11)$$

$$w_n = 1/N \quad (2.1.12)$$

$$p_n = \frac{\sum_i \text{翻訳文中}i\text{と参照訳}i\text{で一致した}n\text{-gram数}}{\sum_i \text{翻訳文}i\text{の中の全}n\text{-gram数}} \quad (2.1.13)$$

式(2.1.11)の  $BP$  は MT 訳が参照訳よりも短い場合のペナルティとなっている。また、BLEU は文単位ではなくドキュメント単位での使用を前提としているため、本実験におけるメタ評価で使用する BLEU では、文単位で使用するための変更を行っている。すなわち、 $p_n=0$  の場合には  $BLEU=0$  となってしまうため、 $p_n = 0$  であるような最大の  $n$  を  $N$  として使用している[12]。BLEU が出力するスコアは大きいほど評価が良い。

#### 2.1.2.4 NIST

NIST は BLEU と同様に n-gram マッチ率を用いた手法であるが、以下の式で計算される情報量で重み付けしている点が異なる。また、ペナルティ関数も BLEU のペナルティ関数と異なっている。NIST が出力するスコアは大きいほど評価が良い。

$$NIST = BP \cdot \sum_{n=1}^N \frac{\sum_i \left( \sum_{\substack{\text{翻訳文}i\text{と参照訳に} \\ \text{共通する}w_1 \dots w_n}} \text{Info}_i(w_1 \dots w_n) \right)}{\sum_i \text{翻訳文}i\text{中の全}n\text{-gram数}} \quad (2.1.14)$$

$$\text{Info}(w_1 \dots w_n) = \log_2 \frac{\text{評価コーパス中の}w_1 \dots w_{n-1}\text{数}}{\text{評価コーパス中の}w_1 \dots w_n\text{数}} \quad (2.1.15)$$

#### 2.1.2.5 NMG\_REF、NMG\_COR

NMG は以下の式で求められ、構文的自然性に着目している点で他の n-gram マッチ率を用いる評価基準とは異なっている。NMG\_REF は参照訳を用いた場合を示し、NMG\_COR は大規模な目的言語コーパスを用いた場合の NMG である。また、以下の式(2.1.16)にある  $grams(w_i)$  は使用コーパス中に存在する最大の gram 数である。NMG が出力するスコアは大きいほど評価が良い。

$$NMG(C) = \log_e \left( \sum_{i=1}^n grams(w_i) / n \right) \quad (2.1.16)$$

#### 2.1.2.6 Charniak score

Charniak score は構文解析ツール Charniak parser が出力する第 1 候補の評価値をその文の構成単語数で割った値をスコアに用いた評価基準である。Charniak score が出力するスコアは小さいほど評価が良い。

#### 2.1.2.7 WN

WN は文の構成単語数そのものをスコアに用いた評価基準である。長文は翻訳精度が悪くなる傾向があり、WN が出力するスコアは小さいほど評価が良い。

#### 2.1.2.8 NMG\_WN

NMG\_WN は上述した NMG\_REF、NMG\_COR、そして、WN を以下の式に示すように組み合わせた評価基準である。NMG\_WN が出力するスコアは大きいほど評価が良い。

$$NMG\_WN = NMG\_REF + NMG\_COR - 0.05 * WN \quad (2.1.17)$$

#### 2.1.2.9 METEOR

METEOR は、再現率  $R$  と適合率  $P$  に基づく F 値に対して単語の非連続性に対するペナルティ関数  $Pen$  を利用した評価基準である。式(2.1.19)のペナルティ関数  $Pen$  にある  $m$  は MT 訳と参照訳との間で一致した単語数であり、 $c$  は一致した各単語を対象として語順が同じものを一つのチャンクとして統合した場合のチャンク数である。したがって、MT 訳と参照訳が完全一致の場合には  $c=1$  となり、語順が全て逆の場合には  $c=m$  となる。 、 、 の値はパラメータである。

$$F = \frac{P \times R}{\alpha \times P + (1 - \alpha) \times R} \quad (2.1.18)$$

$$Pen = \gamma \times (c / m)^\beta \quad (2.1.19)$$

$$METEOR = (1 - Pen) \times F \quad (2.1.20)$$

また、本実験でのメタ評価で使用する METEOR では、マッチングモジュールとして、完全一致、porter stemmer による stemming、WordNet による類義語に基づくマッチングを使用している。METEOR が出力するスコアは大きいほど評価が良い。

#### 2.1.2.10 GTM

GTM は以下の式(2.1.21)の  $size(M)$ に基づく再現率と適合率の F 値を用いてスコアリングを行う評価基準である。GTM ではチャンクのサイズがスコアに反映されることになる。GTM が出力するスコアは大きいほど評価が良い。

$$size(M) = \sqrt[e]{\sum_{r \in M} length(r)^e} \quad (2.1.21)$$

$$precision = \frac{size(M)}{|C|} \quad (2.1.22)$$

$$recall = \frac{size(M)}{|R|} \quad (2.1.23)$$

#### 2.1.2.11 WER

WER は音声認識の分野で利用されている評価指標であり、以下の式の通り、編集距離に基づきスコアリングを行う。WER が出力するスコアは小さいほど評価が良い。

$$WER = \frac{\sum_i (\text{挿入語数}_i + \text{削除語数}_i + \text{置換語数}_i)}{\sum_i \text{参照訳}_i \text{の語数}} \quad (2.1.24)$$

#### 2.1.2.12 PER

PER は bag-of-words として WER を計算する評価基準である。したがって、語順の制約は考慮されない。PER が出力するスコアは小さいほど評価が良い。

$$PER = 1 - \frac{\sum_i \text{翻訳文}_i \text{と参照訳}_i \text{間の一致語数}}{\sum_i \text{参照訳}_i \text{の語数}} \quad (2.1.25)$$

### 2.1.3 性能評価実験

#### 2.1.3.1 実験データ

MT 訳には、NTCIR-7 の 15 グループによる機械翻訳システムがそれぞれ特許翻訳文を日英翻訳した訳文 100 文を用いた。表 2.1.1 にそれぞれのグループ名と機械翻訳システムの手法の一覧を示す。表中の SMT は統計ベース翻訳、RBMT はルールベース翻訳、EBMT は用例ベース翻訳をそれぞれ意味する。また、本実験での全訳文に対する参照訳は 1 つである。

表 2.1.1 グループ名とその手法

Group	tori	FDU-MCandWI	HIT2	JAPIO	KLE	MIT	NAIST-NTT	NICT-ATR
手法	SMT	SMT	SMT	RBMT	SMT	SMT	SMT	SMT

Group	NTNU	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt
手法	SMT	SMT	SMT	EBMT	SMT	SMT	RBMT

人手による主観評価は全訳文（英文）を Adequacy と Fluency の観点より 3 名が 5 段階評価で行った。そして、それぞれの MT 訳において、3 名の評価のメジアンを最終的な人手による主観評価の値として用いた。

#### 2.1.3.2 評価方法

13 の機械翻訳評価基準に対する性能評価には、ピアソンの相関とスピアマンの順位相関を用いた。自動評価基準が出力した文単位のスコアと人手による Adequacy の主観評価の間の相関、自動評価基準が出力した文単位のスコアと人手による Fluency の主観評価の間の相関をピアソンの相関とスピアマンの順位相関を用いてそれぞれ求めた。

#### 2.1.3.3 実験結果

今回行った 13 の自動評価基準における実験結果を表 2.1.2 から表 2.1.5 に示す。表 2.1.2 は自動評価基準によるスコアと人手の主観評価との Adequacy におけるピアソンの相関係数を示している。表 2.1.3 は自動評価基準によるスコアと人手の主観評価との Fluency におけるピアソンの相関係数を示している。また、表 2.1.4 は自動評価基準によるスコアと人手の主観評価との Adequacy におけるスピアマンの順位相関係数を示している。表 2.1.5 は自動評価基準によるスコアと人手の主観評価との Fluency におけるスピアマンの順位相関係数を示している。

表中の相関係数は全て絶対値を用いている。相関係数は値が高いほど、自動評価基準のスコアと人手評価の間の相関が高いことを示している。表の 1 列目はメタ評価に用いた機械翻訳自動評価基準の名称を記している。表中の“Charniak”は“Charniak score”を指している。表の 1 行目には 15 のグループ名を記している。その際、各グループにおける相関係数は、MT 訳 100 文に対する自動評価のスコアと人手評価との相関係数である。表 1 行目の“Avg.”は 15 グループの相関係数の平均である。また、“All”は 15 グループの全 MT 訳 1,500 文に対する自動評価のスコアと人手評価との相関係数である。自動評価基準の比較については、列単位で相関係数を比べる。列ごとに、最も高い相関係数には下線を引き、2 番目に高い相関係数は斜体で表記している。

表 2.1.2 と表 2.1.4 の“NTNU”の相関係数が空欄となっている。これは、“NTNU”における人手による 5 段階評価が全て同じ値であったため、表 2.1.2 のピアソンの相関係数、表 2.1.4 のスピアマンの順位相関係数共に相関係数を得られなかったことによる。したがって、表 2.1.2 と表 2.1.4 については、“Avg.”は“NTNU”を除いた 14 グループの平均となっている。“All”については、“NTNU”も含んだ相関係数となっている。

表 2.1.2 自動評価基準によるスコアと人手評価との Adequacy におけるピアソンの相関係数

Ade	tori	FDU MCandWI	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NICT -ATR	NTNU
IMPACT	<u>0.6765</u>	<i>0.3461</i>	<u>0.3924</u>	<i>0.4319</i>	<u>0.4617</u>	<u>0.4796</u>	<i>0.5336</i>	<u>0.6861</u>	
ROUGE-L	0.6602	0.3075	<i>0.3741</i>	0.4305	0.4254	0.4390	0.4877	<i>0.6679</i>	
BLEU	0.5665	0.2747	0.1918	0.2319	0.3852	0.2892	0.4729	0.5214	
NIST	0.4313	0.1534	0.1884	0.2124	0.2247	0.1303	0.2425	0.3635	
NMG-REF	0.4362	0.1417	0.0038	0.2024	0.2086	0.1847	0.2209	0.4469	
NMG-COR	0.0291	0.1316	0.0899	0.2742	0.1247	0.0049	0.0897	0.0229	
Charniak	0.1358	0.1397	0.0625	0.0027	0.1288	0.1118	0.1463	0.0467	
WN	0.5003	0.2995	0.3122	0.4317	0.3684	<i>0.4586</i>	0.4886	0.4227	
NMG-WN	<i>0.6652</i>	<u>0.3632</u>	0.2364	<u>0.5746</u>	<i>0.4400</i>	0.4395	<u>0.5354</u>	0.5396	
METEOR	0.3898	0.1966	0.1412	0.2630	0.2275	0.2060	0.2617	0.4393	
GTM	0.4094	0.1894	0.1723	0.3088	0.1678	0.1887	0.2367	0.3692	
WER	0.6630	0.3215	0.3715	0.3918	0.4166	0.3537	0.4422	0.6518	
PER	0.3351	0.0983	0.1002	0.1688	0.1370	0.1168	0.1843	0.3617	

Ade	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt	Avg.	All
IMPACT	<u>0.6350</u>	<u>0.4939</u>	<u>0.6218</u>	<i>0.5007</i>	<u>0.7149</u>	<i>0.5262</i>	<u>0.5357</u>	<u>0.6270</u>
ROUGE-L	<i>0.6106</i>	0.4603	<i>0.6097</i>	0.4700	<i>0.7043</i>	0.5154	<i>0.5116</i>	<i>0.6171</i>
BLEU	0.4988	0.2276	0.4595	0.4264	0.5304	0.3593	0.3883	0.4522
NIST	0.4517	0.3005	0.3755	0.2128	0.2974	0.3877	0.2837	0.4183
NMG-REF	0.3607	0.2287	0.3325	0.1798	0.3585	0.2724	0.2556	0.4110
NMG-COR	0.0403	0.2221	0.1994	0.0960	0.1023	0.2908	0.1227	0.1696
Charniak	0.0586	0.2402	0.1741	0.2423	0.0847	0.0577	0.1166	0.2726
WN	0.5211	0.3560	0.4268	<u>0.5558</u>	0.5154	0.4565	0.4367	0.3622
NMG-WN	0.5715	<i>0.4868</i>	0.5191	0.2376	0.5919	<u>0.5603</u>	0.4829	0.4942
METEOR	0.4142	0.0271	0.4453	0.3087	0.3411	0.4475	0.2935	0.4489
GTM	0.4174	0.3049	0.4381	0.2302	0.2788	0.3675	0.2914	0.4253
WER	0.5763	0.4238	0.5684	0.4500	0.6975	0.4495	0.4841	0.5660
PER	0.3028	0.1922	0.3924	0.1174	0.2307	0.3018	0.2171	0.3738

表 2.1.3 自動評価基準によるスコアと人手評価との Fluency におけるピアソンの相関係数

Flu	tori	FDU- MCandWI	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NICT -ATR	NTNU
IMPACT	<b><i>0.5455</i></b>	0.2380	<b><u>0.3087</u></b>	<b><i>0.4478</i></b>	0.3804	<b><u>0.5113</u></b>	<b><u>0.4876</u></b>	<b><u>0.7094</u></b>	0.0457
ROUGE-L	0.5372	0.2027	0.2784	0.4395	0.3506	0.4833	<b><i>0.4481</i></b>	<b><i>0.7039</i></b>	0.0361
BLEU	0.4153	<b><i>0.2709</i></b>	0.1975	0.2831	0.3135	0.3694	0.3746	0.5353	0.1669
NIST	0.4038	0.0539	0.0403	0.2895	0.2441	0.3064	0.2790	0.4513	0.0052
NMG-REF	0.3333	0.0454	0.0507	0.2835	0.1969	0.2929	0.2234	0.5625	0.0396
NMG-COR	0.0124	0.0206	0.0778	0.2539	0.1471	0.1290	0.0232	0.0376	<b><i>0.2062</i></b>
Charniak	0.0814	0.0203	0.0603	0.0284	0.1387	0.0754	0.0409	0.0217	<b><u>0.2538</u></b>
WN	0.4730	<b><u>0.3305</u></b>	<b><i>0.2974</i></b>	0.3371	<b><i>0.3923</i></b>	0.3571	0.4450	0.3313	0.1461
NMG-WN	<b><u>0.5741</u></b>	0.2321	0.2530	<b><u>0.5326</u></b>	<b><u>0.4638</u></b>	<b><i>0.4848</i></b>	0.4398	0.5565	0.1904
METEOR	0.3750	0.1040	0.0986	0.3391	0.2131	0.3626	0.2481	0.5407	0.0116
GTM	0.4003	0.0567	0.0264	0.3337	0.2433	0.3209	0.2320	0.4493	0.0634
WER	0.4862	0.2019	0.2717	0.3584	0.3023	0.3449	0.4239	0.6687	0.0226
PER	0.3281	0.0155	0.0200	0.2098	0.1702	0.2155	0.1493	0.4240	0.0170

Flu	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt	Avg.	All
IMPACT	<b><u>0.5048</u></b>	<b><i>0.4706</i></b>	<b><i>0.5691</i></b>	<b><i>0.2759</i></b>	<b><u>0.5751</u></b>	<b><i>0.4232</i></b>	<b><u>0.4329</u></b>	<b><u>0.6035</u></b>
ROUGE-L	0.4909	0.4275	<b><u>0.5730</u></b>	0.2575	<b><i>0.5682</i></b>	0.4134	<b><i>0.4140</i></b>	<b><i>0.5986</i></b>
BLEU	0.4634	0.3471	0.3845	0.1990	0.4159	0.3597	0.3397	0.4412
NIST	0.4005	0.2921	0.3112	0.1601	0.3128	0.3996	0.2633	0.4815
NMG-REF	0.3408	0.2758	0.2728	0.1314	0.3221	0.3094	0.2454	0.4819
NMG-COR	0.0211	0.2776	0.1950	0.2536	0.0989	0.3045	0.1372	0.2362
Charniak	0.0873	0.2346	0.0880	0.1178	0.0134	0.0225	0.0856	0.3119
WN	0.3839	0.3514	0.4248	<b><u>0.4413</u></b>	0.4772	0.2770	0.3644	0.3076
NMG-WN	0.4345	<b><u>0.5342</u></b>	0.4876	0.0049	0.5453	<b><u>0.4618</u></b>	0.4130	0.5372
METEOR	0.2912	0.1753	0.3621	0.1461	0.3421	0.3491	0.2639	0.5091
GTM	0.3745	0.3240	0.3594	0.1992	0.3075	0.3756	0.2711	0.5018
WER	<b><i>0.4934</i></b>	0.4122	0.5326	0.2673	0.5567	0.3665	0.3806	0.5151
PER	0.2874	0.2325	0.3220	0.0274	0.2566	0.3284	0.2002	0.4488

表 2.1.4 自動評価基準によるスコアと人手評価との Adequacy におけるスパイマンの順位相関係数

Ade	tori	FDU- MCandWI	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NICT -ATR	NTNU
IMPACT	<b><i>0.6073</i></b>	0.3115	<b><u>0.4138</u></b>	<b><i>0.4616</i></b>	<b><i>0.3911</i></b>	<b><i>0.4349</i></b>	0.4722	<b><u>0.6011</u></b>	
ROUGE-L	0.5809	0.2786	<b><i>0.3975</i></b>	0.4371	0.3472	0.3853	0.4080	<b><i>0.5908</i></b>	
BLEU	0.4868	0.2941	0.2016	0.3056	0.2727	0.2397	0.3867	0.4567	
NIST	0.3591	0.1785	0.1952	0.2349	0.1436	0.1138	0.1843	0.3631	
NMG-REF	0.3307	0.1882	0.0026	0.2540	0.1172	0.1669	0.1764	0.3845	
NMG-COR	0.1153	0.1037	0.0642	0.2355	0.1331	0.0012	0.0978	0.0061	
Charniak	0.1757	0.1772	0.0899	0.0169	0.1892	0.1079	0.1875	0.1083	
WN	0.5586	<b><i>0.3224</i></b>	0.3821	0.3178	0.3843	<b><u>0.4847</u></b>	<b><i>0.5106</i></b>	0.5009	
NMG-WN	<b><u>0.6721</u></b>	<b><u>0.3563</u></b>	0.3169	<b><u>0.5488</u></b>	<b><u>0.4030</u></b>	0.4241	<b><u>0.5288</u></b>	0.5390	
METEOR	0.3670	0.1737	0.1813	0.2810	0.1681	0.1979	0.1916	0.4326	
GTM	0.3571	0.2071	0.2077	0.3273	0.1195	0.1888	0.1974	0.3636	
WER	0.5530	0.2680	0.3790	0.3879	0.3265	0.3052	0.3450	0.5234	
PER	0.2726	0.0966	0.0844	0.1715	0.1146	0.1437	0.1365	0.3512	

Ade	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt	Avg.	All
IMPACT	<b><u>0.6284</u></b>	<b><i>0.4360</i></b>	<b><u>0.6093</u></b>	0.4710	<b><u>0.6926</u></b>	<b><i>0.5708</i></b>	<b><u>0.5073</u></b>	<b><u>0.6167</u></b>
ROUGE-L	<b><i>0.5933</i></b>	0.4032	<b><i>0.5982</i></b>	0.4410	<b><i>0.6838</i></b>	0.5336	0.4770	<b><i>0.6003</i></b>
BLEU	0.4662	0.2004	0.4439	0.3452	0.4702	0.3661	0.3526	0.4247
NIST	0.4354	0.3229	0.3381	0.1654	0.2721	0.3489	0.2611	0.3874
NMG-REF	0.3032	0.1731	0.3118	0.1476	0.2961	0.2704	0.2231	0.3644
NMG-COR	0.0057	0.2346	0.2129	0.1201	0.0209	0.2346	0.1133	0.1791
Charniak	0.0338	0.2594	0.1653	0.2629	0.0761	0.0159	0.1333	0.2502
WN	0.5584	0.3698	0.4162	<b><u>0.6290</u></b>	0.5504	0.3966	0.4558	0.3725
NMG-WN	0.5889	<b><u>0.5204</u></b>	0.5031	<b><i>0.4884</i></b>	0.5823	<b><u>0.5716</u></b>	<b><i>0.5031</i></b>	0.5114
METEOR	0.4302	0.1776	0.4535	0.3002	0.3304	0.4686	0.2967	0.4378
GTM	0.3980	0.2969	0.3952	0.1908	0.2511	0.3481	0.2749	0.3861
WER	0.5285	0.3137	0.5200	0.4403	0.6666	0.4562	0.4295	0.5240
PER	0.3064	0.1928	0.3311	0.0846	0.2286	0.2869	0.2001	0.3335

表 2.1.5 自動評価基準によるスコアと人手評価との Fluency におけるスピーアマンの順位相関係数

Flu	tori	FDU- MCandWI	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NICT -ATR	NTNU
IMPACT	<b><u>0.5513</u></b>	<b><u>0.2450</u></b>	<b><u>0.2904</u></b>	<b><u>0.4605</u></b>	0.3353	<b><u>0.4800</u></b>	0.4226	<b><u>0.6517</u></b>	0.0802
ROUGE-L	0.5275	0.2105	0.2712	0.4319	0.2954	0.4344	0.3812	<b><u>0.6455</u></b>	0.0701
BLEU	0.3984	0.2640	0.1902	0.3130	0.2636	0.3592	0.2807	0.4935	0.1543
NIST	0.3772	0.0839	0.0356	0.2430	0.1893	0.2924	0.2476	0.4485	0.0132
NMG-REF	0.2745	0.0458	0.0333	0.2718	0.1114	0.2859	0.2020	0.5470	0.0305
NMG-COR	0.0960	0.0455	0.1115	0.2641	0.1373	0.1197	0.0356	0.0085	0.1768
Charniak	0.0572	0.0725	0.0456	0.0324	0.2045	0.0736	0.0688	0.0021	<b><u>0.1980</u></b>
WN	0.4645	<b><u>0.3051</u></b>	<b><u>0.3224</u></b>	0.3058	<b><u>0.3993</u></b>	0.3421	<b><u>0.4653</u></b>	0.3503	0.1737
NMG-WN	<b><u>0.5533</u></b>	0.2407	0.2858	<b><u>0.5536</u></b>	<b><u>0.4503</u></b>	<b><u>0.4617</u></b>	<b><u>0.4496</u></b>	0.5215	<b><u>0.2163</u></b>
METEOR	0.3691	0.1025	0.1155	0.3443	0.1863	0.3605	0.2020	0.5343	0.0132
GTM	0.3883	0.1048	0.0294	0.2840	0.1951	0.3077	0.2088	0.4402	0.0701
WER	0.4344	0.1429	0.2289	0.3567	0.2403	0.2932	0.3432	0.6049	0.0731
PER	0.3009	0.0008	0.0373	0.1401	0.1527	0.2122	0.1348	0.4118	0.0305

Flu	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt	Avg.	All
IMPACT	<b><u>0.4707</u></b>	<b><u>0.3747</u></b>	<b><u>0.6133</u></b>	0.2758	<b><u>0.6212</u></b>	<b><u>0.4554</u></b>	<b><u>0.4219</u></b>	<b><u>0.5953</u></b>
ROUGE-L	0.4565	0.3288	<b><u>0.6050</u></b>	0.2622	<b><u>0.6158</u></b>	0.4135	0.3966	<b><u>0.5793</u></b>
BLEU	<b><u>0.4592</u></b>	0.2831	0.4291	0.1309	0.4202	0.4029	0.3228	0.4447
NIST	0.4284	0.2766	0.2915	0.0916	0.3110	0.4079	0.2492	0.4339
NMG-REF	0.3154	0.2007	0.2741	0.0747	0.3066	0.3327	0.2204	0.4167
NMG-COR	0.0674	0.2292	0.1765	<b><u>0.3396</u></b>	0.0578	0.2737	0.1426	0.2199
Charniak	0.1083	0.1897	0.1180	0.0490	0.0507	0.0534	0.0883	0.2266
WN	0.3519	0.3284	0.3819	<b><u>0.4783</u></b>	0.4610	0.2902	0.3613	0.3098
NMG-WN	0.4457	<b><u>0.5149</u></b>	0.4838	0.1848	0.5670	<b><u>0.4838</u></b>	<b><u>0.4275</u></b>	0.5159
METEOR	0.3133	0.2262	0.3539	0.1565	0.3417	0.3851	0.2670	0.4616
GTM	0.3835	0.2985	0.3233	0.1364	0.3064	0.3906	0.2578	0.4307
WER	0.4557	0.2767	0.5235	0.2786	0.6035	0.3282	0.3456	0.4934
PER	0.3238	0.2097	0.2711	0.0076	0.2675	0.3304	0.1887	0.3748

#### 2.1.4 まとめと今後の予定

実験結果より、得られた相関係数には大きなばらつきが見られる。例えば、表 2.1.2 において、最も高い相関係数は、“ Moses ” における “ IMPACT ” の 0.7149 である。それに対して、“ FDU-MCandWI ” では、最も高い相関係数を示した “ NMG-WN ” でも 0.3632 であった。同様な傾向は表 2.1.3 から表 2.1.5 にも見られた。したがって、今後は安定した、かつ、より高い相関係数を得るための自動評価基準の実現に向けた研究を行う予定である。

また、今回のメタ評価では、使用した参照訳の数が全 MT 訳において 1 つのみであったため、複数の参照訳を用いた際のメタ評価を行う予定である。

#### 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] Papineni, Kishore., Roukos, Salim., Ward, Todd. and Zhu, Wei-Jing. BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pp.311-318, 2002.
- [2] Fujii, Atsushi., Utiyama, Masao., Yamamoto, Mikio. and Utsuro, Takehito. Overview of the Patent Translation Task at the NTCIR-7 Workshop. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp.389-400, 2008.
- [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] Lin, Chin-Yew. and Och, Franz Josef. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics, pp.606-613, 2004.
- [5] 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>
- [6] Ehara, Terumasa. Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation, Proceedings of the Eleventh Machine Translation Summit Workshop on Patent Translation, pp.13-18, 2007.
- [7] 江原暉将. 句レベルの統計的後編集と翻訳精度の評価, 平成 19 年度 AAMT/Japio 特許翻訳研究会報告書, pp.2-14, 2008.
- [8] Banerjee, Satanjeev., and Lavie, Alon. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or

Summarization, pp.65-72, 2005.

- [9] Turian, Joseph P., Shen, Like. and Melamed, I. Dan. Evaluation of Machine Translation and its Evaluation, Proceedings of the 11th Machine Translation Summit, pp.386-393, 2003.
- [10] Leusch, Gregor., Ueffing, Nicola. and Ney, Hermann. A Novel String-to-String Distance Measure With Applications to Machine Translation Evaluation, Proceedings of the 9th Machine Translation Summit, pp.240-247, 2003.
- [11] Su, Keh-Yih., Wu, Ming-Wen. and Chang, Jing-Shin. A New Quantitative Quality Measure for Machine Translation Systems, Proceedings of the 14th International Conference on Computational Linguistics, pp.433-439, 1992.
- [12] 江原暉将. 新しい機械翻訳自動評価基準を目指して, 平成 18 年度 AAMT/Japio 特許翻訳研究会報告書, pp.2-11, 2007.

## 2.2 文分解法によるコーパス規模と Bleu 値との関係<sup>1</sup>

秀明大学 宮澤 信一郎

創価大学 岡田 勇

### 2.2.1 はじめに

機械翻訳システムに大規模コーパスを用いる手法が提案されているが、その翻訳品質に対しては、多くの課題が残っている。しかし、現実にはコーパスを用いた機械翻訳の重要性が高まるにつれてある程度の品質をどのように保証するかは喫緊の課題となっている。特に、コーパスの規模と翻訳性能との関係について理解することは集めるべきコーパスの量に関する見通しを与えてくれるなど、重要な知見をもたらすだろう。

そのため本研究では、コーパスの規模とそれに基づく機械翻訳システムの品質評価について議論することにしたい。機械翻訳の品質には大きく人手評価と自動評価に分かれるが、人手評価には莫大なコストと時間がかかるため、自動評価でいかに正確な品質評価が可能であるかについて研究が進められてきている。さらに、隅田(2005) は Bleu の提案によって機械翻訳研究には自動評価結果を示すことが当然視されている現状を指摘している。つまり、規模に対応した品質の大きな見通しを与えるためには、自動評価でどのくらいの性能を測定できるかが重要となる。そこで、品質を自動評価する方法について議論する。

自動評価とは、あらかじめ正解文が設定されている場合では、正解文と翻訳文との類似性に基づいて品質を評価するのが一般的である。本研究では n-gram を用いた類似性指標を用いた文分解法と名づけた自動評価手法について提案する。

ところで、類似性指標は文単位の評価を計測するものであって、それらを集約して翻訳システムの性能をどう評価するのであろうか。内元(2003)や吉岡(2005) では、交差検証法(n-fold cross validation) が性能評価の方法として使用されている。これは、対象となるデータセットを n 分割し、そのうちの 1 分割を残りのデータセットから推定した時の誤差を全ての分割に対して集め、その自乗和平均を取ることでデータセットの評価を行う手法である。本研究ではこれを参考に、修正交差検証法と名づけた新たな方法を提案する。

我々は、日英対訳特許文書から生成されたコーパスを用いる。内山(2007) は、特許の機械翻訳が日英・英日の機械翻訳の対象として適切であると主張している。また、森下(2008) は対訳特許文から専門用語辞書を半自動獲得する手法について新たな手法を提案している。このように日英対訳特許文は機械翻訳研究に良く用いられている。さらに、特許という対象の実用性からも、機械翻訳の性能とコーパスの規模との関係を議論することは有益である。

我々は、機械翻訳の性能とコーパスの規模との関係に関する見通しを持つために、全てのコーパス文書対の類似度を効率的に計算することにした。本研究では、大規模なコーパス文書対の全

---

<sup>1</sup> この報告は、岡田勇、宮澤信一郎、大規模コーパスの効率的な Bleu 値計算方法の提案，情報処理学会研究報告，2009-NL-189-(1), 1-6、がオリジナルである。

での組み合わせに対する Bleu 値を計算するための効率的な手法を提案し、その実行結果について報告する。

2 節では、本研究の翻訳品質の評価指標に関する議論を整理し、Bleu ならびにそれを構成する n-gram を基本としたコーパス性能の評価法について検討する。3 節では、提案手法の詳細を説明する。4 節では、研究対象と提案手法を実装した結果について報告する。5 節では、結果について議論し、本研究をまとめる。

## 2.2.2 翻訳品質の評価法

翻訳品質をどう評価するのかについては多くの研究がある。このうち、自動評価手法としては、あらかじめ正解文が設定されている場合では、正解文と翻訳文との類似性に基づいて品質を評価するのが一般的である。ここでは翻訳品質の評価法として個々の文についての類似性評価の方法と、それを全体として集約する方法について議論する。

### 2.2.2.1 類似性の評価

類似性を表す指標としては、Papineki(2002) の Bleu 値をはじめ多くが提案されており、それらの性能についても多くの研究がある。今村(2004) は、機械翻訳の品質評価について複数の評価指標を比較しており、「完全訳を重視した場合、BLEU, WER の効果が高く、理解可能訳を重視した場合、WER, NIST の効果が高」との知見を得ている。岸田(2006) は、Bleu 以外に、WAMU という指標を定義し、それらを用いて言語横断検索の性能を回帰モデルで予測する方法を提案している。江原(2007)では、Bleu の欠点として、単語の文字面の一致に基づく点と構文的な自然性が認識できない点を上げ、新たな評価指標の開発を試みているが、基本的には n-gram の概念を用いている。このように Bleu には多くの議論がなされているが、それに代わる有効な指標が、現状では提案されていない。また本研究目的に照らし、厳密な正確性ではなく、限られた時間と資源制約で概算値を求めるほうが望ましいことから n-gram 一致率を基にする単純な指標である Bleu 値を、翻訳品質の評価指標として取り上げる。

Papineki(2002) による Bleu 値の定義は次の通りである。

長さ  $c$  の候補文  $C$  に対する長さ  $r$  の正解文  $R$  の n-gram Bleu 値は

$$BP = \exp(1 - \max\{1, r/c\})$$

なる BP を用いて、

$$Blue(R, C) = BP \exp\left(\sum_{i=1}^n \log(p_i) / n\right)$$

で与えられる。ここで  $p_n$  とは n-gram 正解率 (一致率) のことである。

英文の品質評価に際して Bleu 値を用いる場合、金山(2003)によると、人手評価との比較において、英単語を単位とする 4-gram で計算することが望ましいことが示されている。日英対訳特許文において、4-gram による Bleu を用いるとすると、定義から、 $Bleu > 0$  であるための必要十分条件が、4-gram で一致する単語列を有することであるから、文の 4-gram の出現パターンを検討することが本質的に重要である。

次に、Bleu の計算を効率的に行うための方法について述べる。Bleu の定義から、4-gram までの句の一致率を数え上げることが重要である。そのため、文を事前に 4-gram に分割しソーティングすることで効率的に Bleu を計算できる文分解法と名づけた方法を提案する。

#### 2.2.2.2 コーパスの評価

コーパスの規模を変えたときの翻訳性能の変化を調べるために、コーパスの評価をいかに行うか議論したい。交差検証法では、コーパス全体をテスト文と参照文に分割し、参照文からテスト文がどのくらい再現できるかを評価するものである。一方、コーパスによる機械翻訳とは、コーパスの中から最も翻訳文に近い候補文をいかに抽出できるかにかかっている。このような優れた翻訳システムが構築できたとすると、抽出された候補文は、目的となるテスト文との類似性が最も高いものであるとみなして差し支えない。つまり、全てのコーパスを構成する文の中で最も高い類似性を持つ文の類似度がコーパスの性能を表すものとなる。そのため、通常はコーパスの規模が大きくなれるほど、そのコーパスの性能は向上する。一方で、コーパスを大きくするにはコストがかかるので、コストと性能のトレードオフが存在し、そのバランスについて議論できることが重要となる。

このような考え方に基づき本論文では、規模  $s$  のコーパス  $C$  の翻訳性能  $R(s)$  を次のように定義することとする。

まず、コーパスの対象とすることができる文のサイズに比べて  $s$  が十分小さい場合は、どの文を対象とするかでいくつもの  $R(C)$  を計算できるため、規模  $s$  のコーパスを複数作成して計算した  $R(C)$  の平均を性能とする。そこで、以下ではある規模  $s$  のコーパスを構成したものとして、その性能を定義する。

交差検証法では、コーパスの一部をテスト文としていたが、本研究では、コーパスを構成する文ではない文集合からテスト文を構成するものとし、その集合を  $T$  とする。なぜなら一般的に  $Bleu(x, y) \cong Bleu(y, x)$  であるので、交差検証法では対称性が生じやすく、テストデータをコーパスから独立したほうが望ましいと考えられるからである。 $T$  の全ての要素  $t$  に対して、 $C$  の全ての要素との Bleu 値を計算し、そのうち最大となるものを  $c_t$  とする。

すなわち、

$$c_i = \arg \max_{c \in C} Bleu(t, c)$$

である。ある優れた翻訳システムを使用すれば、 $t$  に対するコーパス候補文として  $c_t$  を選択す

ることが出来るはずであるから、これをコーパスの性能の基礎とする。そこで、 $Bleu(t, c_t)$  を全てのテスト文に対して行った平均をコーパス  $C$  のテスト文集合  $T$  に対する性能  $R(C|T)$  と定義する。すなわち、

$$R(C|T) = ave(Bleu(t, c_t)) \quad \forall t \in T$$

である。これを用いて規模  $s$  のコーパス性能  $R(s)$  を計算する。

この計算は、規模が異なるごとに行う必要があり、計算量爆発の原因となりうるため、それを回避する手段を提案する。その本質は、テスト文と最大規模のコーパス文とのすべての組み合わせに対する Bleu を最初に計算しておいてから、任意の規模でのコーパスの性能を計算するというものである。次節において詳細に説明する。

### 2.2.3 提案手法とアルゴリズム

ここでは、Bleu 値を効率的に計算する文分解法と、コーパスの性能を計算する方法を提案し、そのアルゴリズムを説明する。

#### 2.2.3.1 文分解法による Bleu の効率的な計算方法

あるテスト文  $x$  に対する候補文  $y$  における Bleu 値を  $Bleu(x, y)$  と定義し、この効率的な計算方法を説明する。そのため前処理として、対象文ごとに、全ての 4-gram の句と最後の 3,2,1 単語からなる句を、ソートしてファイルに保存する。例えば、文  $x$  として "The system works in the front of the neck freely." 文  $y$  として "The system runs in the front of the head." とする場合を考えよう。  $x$  に対して表 1 のようなファイル  $F_x$  を、  $y$  に対して表 2 のようなファイル  $F_y$  を生成する。

表 1 : 文  $x$  から生成されるファイル  $F_x$

freely  
front of the neck  
in the front of  
neck freely  
of the neck freely  
system works in the  
the front of the  
the neck freely  
the system works in  
works in the front

表2：文  $y$  から生成されるファイル  $F_y$

front of the head  
 head  
 in the front of  
 of the head  
 runs in the front  
 system runs in the  
 the front of the  
 the head  
 the system runs in

この前処理により、Bleu 値を計算するための作業領域として、対象文の量に関して約 4 倍のメモリ領域を用いることになる。

次に、 $F_x$  のすべての行  $l_x$  に対して以下の動作を行う。 $l_x$  の第 1 単語と一致する行が  $F_y$  にある場合、その行  $l_y$  が  $l_x$  と第何単語まで一致するか調べ、その値  $k$  に対し、 $G_{F_x}(k)$  に 1 を加える。ただし  $l_y$  が複数あっても追加する値は最大のもの一つのみとする。こうして出来た  $G_{F_x}(k)$  から n-gram 一致率を算出することができ、

$$p_k = \frac{\sum_{k < i} G_{F_x}(i)}{\#l_y - k + 1}$$

となる。ただし、 $l_y$  の個数を  $\#l_y$  とする。以上から Bleu を計算することが出来る。上記の例では  $Bleu(x, y) = 0.459$  となる。この方式で計算すると計算量は、単語数のオーダーとなり効率的である。

### 2.2.3.2 修正交差検証法によるコーパス性能の効率的な計算方法

任意の規模のコーパスについて性能評価を効率的に行うため、テスト文  $T$  と最大コーパスを構成する場合のコーパス文  $C$  に関する全ての  $Bleu(T, C)$  を計算し、Bleu 値表を作成する。これを用いて任意の規模  $s$  のコーパスの性能  $R(s)$  を以下の方法で計算する。

最大コーパスを分割して、規模  $s$  のコーパス  $C_i$  を作成する。Bleu 値表のうち  $C_i$  を構成する文の行を取り出し、テスト文ごとに最大 Bleu 値を算出し、その平均を  $R(C_i | T)$  とする。これを全てのコーパス  $C_i$  に対して行った平均を  $R(s)$  とする。この方法によると計算量は(テスト文の規模 × 最大コーパスの規模) のオーダーとなるので効率的といえる。

### 2.2.4 使用データと計算結果

ここでは使用データについて説明し、提案手法を実装した結果について紹介する。

#### 2.2.4.1 使用データ

日英対訳特許文として、(財)日本特許情報機構から提供された公開特許公報要約と PAJ 対訳データの 2003 年公開全件 348,061 件の中から、英文対訳で特許概要を説明している部分 (タグ <SDOABLA="E">) を対象とする。この部分は実質的に特許の概要を説明している部分である。ここは「解決すべき問題(PROBLEM TO BE SOLVED:)」という部分と「解決(SOLUTION:)」という部分に分かれる。前者はほぼ To 不定詞で始まり問題を名詞句の形で表現している。後者で具体的な概要が文として記述されている。そこで全データのうち、後者の部分が 1 文以上あるもの 348,058 件のうち、327,680 件を対象にした。この部分の 2 文目以降は、1 文目と同じ専門用語や名詞句の使用が散見されることから、1 文目に対する Bleu 値が高くなりやすくコーパスの性能に影響を与えやすい。しかし、これらの文は同時に翻訳されるのが普通であるから、全ての文書対に対する Bleu 値の計算を想定する場合は、コーパスに含めるべきではなく、今回の対象には適さないため、英文説明の 1 文目のみを抽出し対象データとする。

このデータに対し 32768 = 215 文ごとにグループ 0 から 9 を作り、各グループごとに 0 から順に、文に番号を付与する。各グループに対して、文番号 0 - 99 の文をテスト文とし、文番号 10000 - 19999 の文をコーパスを構成する文の候補とする。表 3 にデータの基本統計量を示す。

表 3 : テスト文とコーパス文の基本統計量

	テスト文	コーパス文
規模	1000	100000
1 文あたりの 平均単語数	54.10	52.78
総単語種数	4744	62948

#### 2.2.4.2 計算結果

前節で提案した方法を Linux 上にて perl で実装した結果、最も処理時間のかかる、全組み合わせに対する Bleu の事前計算であっても、1 日程度のオーダーで計算可能であることが分かった。また、コーパスの規模別の性能は表 4 と図 1 にまとめる。また、規模別の n-gram 一致率の平均の推移を表 4 と図 2 に、性能と n-gram 平均一致率との相関を表 5 にまとめた。ここで相関とは、n-gram の一致数と性能という二つの変量に関する全部で 10 種類ある規模ごとに計算した相関係数のことである。

#### 2.2.5 議論とまとめ

本研究では、大規模コーパスの翻訳性能を少ない記憶領域と計算時間で測定する方法を提案装した。その結果、計算すべき 1000 × 10 万件のうち、Bleu が正となるデータも 32; 628 件と

少なく、重複計算を防ぐことで1日程度の計算時間で測定できた。

表4ならびに図1から、本研究で定義したコーパスの性能の推移は規模が5000以上であると、コーパス規模の対数に関する線形回帰で表すことが出来ると予測できる。回帰直線と決定係数を求めると

$$R = -0.0486 + 0.0157 \log s$$

となり、決定係数は0.997の高い水準となっている。全ての規模に対する回帰直線の決定係数も0.92と高い。この回帰直線がどのくらいの規模に関して有効なのかは、今後の議論であるが、日英対訳特許における近似的なコーパスの性能予測としては有効であろう。さらに、性能予測曲線は規模の増加に伴いやがて収束することが推測できるので、全体としてS字曲線を描くと思われる。この点も今後検討したい。

また、表5からコーパスの性能とn-gramとの相関は概ね高く、3-gramでは99.9%となっている。つまり、Bleuで計測した大まかな性能は3-gram一致率で計測する場合と近似することを意味する。

本研究は日英対訳特許文書という限定されており、ただちに一般的な性能評価に結びつかない。表3から明らかなように、テスト文の総単語種類数は5000弱であるが、その100倍の規模であるコーパス文のそれは6万を超えていることから、対象が専門用語の多い特殊性を有することが示唆される。これをどのように拡張して、一般性を持たせるかは今後の課題となろう。

さらに、コーパスの性能として4-gramまでの一致率を使ったBleuを用いたが、他の指標についても検討する必要があるかもしれない。例えば、特許文は特定の専門用語が散見され、その存在によって性能を大幅に低下することがありうる。例えば"I watch an orange under the table."という文と"I watch an apple under the table."という文は4-gramで一致する部分がないのでBleu = 0であるが、文の構造は完全に再現され、たった一つの名詞の置き換えで正解文と一致できる良質な翻訳であると判定しても良い。日英対訳特許でもこれと同種の問題が起きる可能性は高く、これに対処できるような指標は検討すべきであろう。

## 参考文献

[今村 2004] 今村賢治, 隅田英一郎, 松本裕治, 機械翻訳自動評価指標の比較, 第10回言語処理学会年次大会, 452-455.

[江原 2007] 江原暉将, 新しい機械翻訳自動評価基準 NMG の提案, Japio 2007 YEARBOOK, (財)日本特許情報機構, 238-265.

[金山, 2003] 金山博, 荻野紫穂, 翻訳精度評価手法 BLEU の日英翻訳への適用, IPSJ SIG Notes, 2003(23), (2003-NL-154-19), 131-136.

[岸田 2006] 岸田和明, 三田図書館・情報学会, 2006年度研究大会.

- [森下 2008] 森下洋平, 宇津呂武仁, 山本幹雄, 対訳特許文書からの専門用語対訳辞書半自動獲得におけるフレーズテーブルと既存対訳辞書の併用, 情報処理学会研究報告, 2008-NL-187,91-98.
- [Papineki 2002] Papineki, K., et.al., Bleu : a Method for Automatic Evaluation of Machine Translation, Proc. ACL 2002, 311-318.
- [隅田 2005] 隅田英一郎, 佐々木裕, 山本誠一, 機械翻訳システム評価法の最前線, 情報処理, 46(5),552-557.
- [内元 2003] 内元清貴, 関根聡, 村田真樹, 井佐原均, 用例に基づく手法と機械学習モデルの組み合わせによる訳語選択, 自然言語処理, 10(3),87-114.
- [内山 2007] 内山将夫, 山本幹雄, 藤井敦, 宇津呂武仁, 特許情報を対象とした機械翻訳:共通基盤による評価タスクを目指して, IPSJ SIG Notes,2007(76), (2007-NI-180-23), 133-138.
- [吉岡 2005] 吉岡篤志, 徳久雅人, 村上仁一, 池原悟, 名詞句パターン辞書を用いた日英機械翻訳の試作- bi-gram による訳出選択の場合, 平成 17 年度電気・情報関連学会中国支部連合大会論文集, 305-306.

表 4 : 規模別のコーパス性能と n-gram の平均一致率

規模	性能	1-gr.	2-gr.	3-gr.	4-gr.
100	0.00066	0.0185	0.0018	0.0004	0.0000
200	0.00117	0.0207	0.0024	0.0007	0.0001
500	0.00233	0.0238	0.0036	0.0013	0.0002
1000	0.00374	0.0263	0.0047	0.0020	0.0004
2000	0.00579	0.0287	0.0060	0.0029	0.0007
5000	0.00979	0.0313	0.0079	0.0043	0.0013
10000	0.01370	0.0325	0.0095	0.0057	0.0020
20000	0.01823	0.0340	0.0114	0.0073	0.0028
50000	0.02488	0.0356	0.0143	0.0098	0.0042
100000	0.02999	0.0369	0.0164	0.0115	0.0055

表 5 : 性能と n-gram との相関係数

n-gram	相関係数	F 値
1-gram	0.912	0.0002
2-gram	0.993	0.0000
3-gram	0.999	0.0000
4-gram	0.994	0.0000

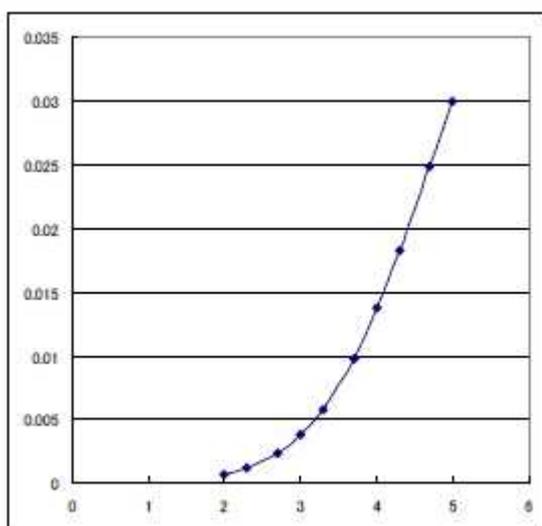


図 1 : コーパスの規模(横軸)と性能(縦軸)の関係

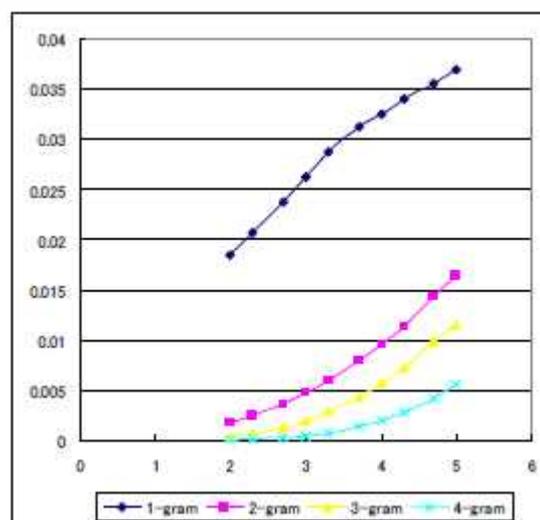


図 2 : コーパスの規模(横軸)と平均 n-gram 一致率(縦軸)の関係

## 3.1 専門用語抽出アルゴリズムを用いた対訳コーパス

### からの対訳辞書の自動生成

東京大学 範 暁蓉

二宮 崇

中川 裕志

#### 3.1.1 はじめに

対訳辞書は機械翻訳において非常に重要なリソースであり、伝統的なルールベース翻訳システムだけではなく、近年盛んに研究されている統計的機械翻訳においても構成的ではない複合語や専門用語の対応を可能とする翻訳を実現するために、非常に質の高い対訳辞書が必要とされている。特許文の翻訳は実社会から大きく期待されているが、特許文には特許文特有の非常に多くの専門用語や複雑な複合語が出現し、既存の対訳辞書で新しく出現する専門用語や複合語に対応することは難しい。

本稿では、専門用語抽出システムと対訳コーパスを用いて品質の高い対訳辞書を自動的に生成する手法を提案し、日英の特許コーパスに対し適用した結果について報告する。一般に利用可能な単語対応付けシステムを用いれば、対訳コーパスから自動的に単語対は得られるが、そのままでは対訳辞書に供せられる質を得ることは難しい。本研究では、専門用語抽出システムを用いて専門用語レベルの対応付けを与え、専門用語対の対応スコアを再ランク付けすることにより非常に質の高い対訳関係の語句を抽出する手法を提案する。本稿では、まず、従来の対訳語の抽出技術を調査し、次に専門用語抽出アルゴリズムを用いた対訳コーパスからの対訳抽出技術を説明し、それを特許文に適用した結果を報告する。

本稿の構成は以下のようになっている。3.1.2 節では、従来の対訳抽出技術に関する調査を説明する。3.1.3 節は、専門用語抽出アルゴリズムを用いた対訳コーパスからの対訳辞書の自動生成手法を説明し、3.1.4 節では、特許文からの用語抽出の実験・結果について報告する。3.1.5 節で本稿の主旨をまとめ、今後の課題について述べる。

#### 3.1.2 従来の対訳語抽出技術に関する調査

対訳語抽出は今までに様々な手法が提案されているが、前提とするリソースの違いによりそれらの手法を大きく分類することができる。例えば、既存の対訳辞書を用いる/用いない手法、中間言語を介する/介さない手法などのように前提とするリソースに大きな違いがある。ここでは抽出元となるリソースの違いにより既存の手法の分類および調査を行う。具体的には、抽出元が、対訳辞書、パラレルコーパス、コンパラブルコーパス、ウェブである場合に分類する。

対訳辞書から対訳語の抽出 (張玉潔ら, 2005)

三つの言語 A と B と C に対し、A と B の対訳辞書がなく、A と C、B と C 間の対訳辞書が存在するとき、言語 C を介して、A と B の対訳辞書を構築することができる。この方法で

は、AC 辞書を用いて A 言語から C 言語への訳語候補を生成し、次に BC 辞書を用いて C 言語から B 言語への訳語候補を生成することにより A 言語から B 言語への訳語候補を生成する。この方法で抽出された対訳辞書は精度が高いが、非常に多くの訳語候補が生成される。また、辞書中に存在しない対訳語を抽出することができず、複合語の対訳語抽出が難しい。

#### パラレルコーパスから対訳語の抽出

近年多くのパラレルコーパス（文単位の対応付けがなされている対訳コーパス）が利用可能になり、それらを用いた対訳語抽出の研究が行われている。パラレルコーパスは大きくわけて二種類存在し、一つは翻訳された文学作品や多言語による商品の説明書であり、もう一つは翻訳に関する研究のために構築された対訳コーパスである。パラレルコーパスに対し単語アライメント（対訳文の単語対応付け）を行えば単語単語の対訳対が自動的に得られるため、パラレルコーパスを前提とした単語アライメントの性能向上に関する研究が多い。また、複合語対訳対の抽出にも単語アライメントの結果に基づいて抽出することが可能である。

単語単語対訳対の抽出は、頻度、文脈情報、統計手法を組み合わせた単語アライメントにより行われ、以下の方法がある。

- (1) 辞書や単語、意味クラスおよび品詞の語彙知識を利用して抽出する方法 (Ker&Chang, 1997)。
- (2) 統計的な手法で抽出する方法。
  - 統計翻訳モデル (IBM 1-5 モデル) を基づく抽出方法。この手法は EM アルゴリズムを利用して計算する (Brown et al., 1993)
  - HMM モデルを基づく抽出手法 (Toutanova et al., 2002)
  - Och のテンプレートを基づく手法。この手法は一对多の単語対応を解決した。(Och&Ney, 2002; Och&Ney, 2003)
  - 相互情報量を利用する方法 (Gale&Church, 1991)

Och の手法は多くの研究者により用いられており、Giza++はこの手法に基づいて実装された単語アラインメントツールである。複合語対訳対の抽出は単語アラインメントに基づいて行われるが良い精度を達成することが難しい。

パラレルコーパスからの対訳語抽出は非常に高い精度で抽出することが可能であり、例えば、単語対の抽出精度は 99%に達する。そのためパラレルコーパスを利用する研究が盛んに行われている。しかし、パラレルコーパスの構築は難しく、利用可能なパラレルコーパスの量は十分とはいえない。

#### コンパラブルコーパスからの対訳語抽出

二つの言語の文章データが内容的に非常に似通っている場合、この 2 言語の文書を 1 組にしてコンパラブルコーパスという。コンパラブルコーパスからの対訳語の抽出方法には以下の手法がある (Chiao&Zweigenbaum, 2002; Gamallo, 2007)。

- (1) 二つ言語の単語  $w_1$  と  $w_2$  に対し、もし  $w_1$  の周りの単語が  $w_2$  の周りの単語と対訳対の

関係にあれば、w1 と w2 も対訳対する。この方法は対訳辞書に大きく依存し、対訳辞書の多義性の問題がある。

- (2) パラレルコーパスから単語対応を求め、それを利用して手法(1)と同じ方法でコンパラブルコーパスから対訳辞書を抽出する。
- (3) 少量のパラレルコーパスからテンプレートを抽出し、そのテンプレートを用いてコンパラブルコーパスから対訳辞書を抽出する。

コンパラブルコーパスから抽出された対訳辞書の精度はパラレルコーパスより低いが、コンパラブルコーパスはパラレルコーパスに比較して大量に存在する。コンパラブルコーパスからの抽出方法とパラレルコーパスからの抽出手法をうまく組み合わせれば大規模な高精度の対訳辞書が自動的に抽出される。

#### ウェブから対訳語の抽出

ウェブ上に様々な言語で記述されたウェブページが大量に存在する。ニュースや新聞も毎日更新されており、パラレルコーパスやコンパラブルコーパスより、非常に巨大なリソースである。ウェブページからの対訳語抽出に関しては次の研究がある。(Brill et al., 2001; Nazar et al, 2008)

- (1) ある一つの対訳対をシードとしてサーチエンジンのキーワードとして入力して、ウェブ上にこの対訳語に関する両言語の文書を抽出する。
  - (2) コンパラブルコーパスからの抽出手法を用いて、対訳対応を抽出する。
- しかし、得られるテキストはコンパラブルコーパスよりも対訳関係が弱いため、これらの手法により得られる対訳語の精度は低い

#### その他の研究

その他に次のような対訳語抽出の研究がある。

- (1) Wikipeda から対訳辞書の構築
- (2) 複数の単一言語コーパスから対訳語を抽出する手法

### 3.1.3 専門用語抽出アルゴリズムを用いた対訳コーパスからの対訳辞書の自動生成手法

本研究は、専門用語抽出システムと単語アライメントを組み合わせ、質の高い専門用語辞書の生成手法を提案する。図 1 は提案手法の全体像を表しており、次のステップで対訳抽出される。

1. 専門用語抽出 (TermExtractor)
2. リセグメント処理
3. 専門用語付き単語アライメント
4. 単語アライメント
5. リランキング

図中では日英のパラレルコーパスを想定しており、対訳文の対である日本語文と英語文が入力される (図左上と図右上)。

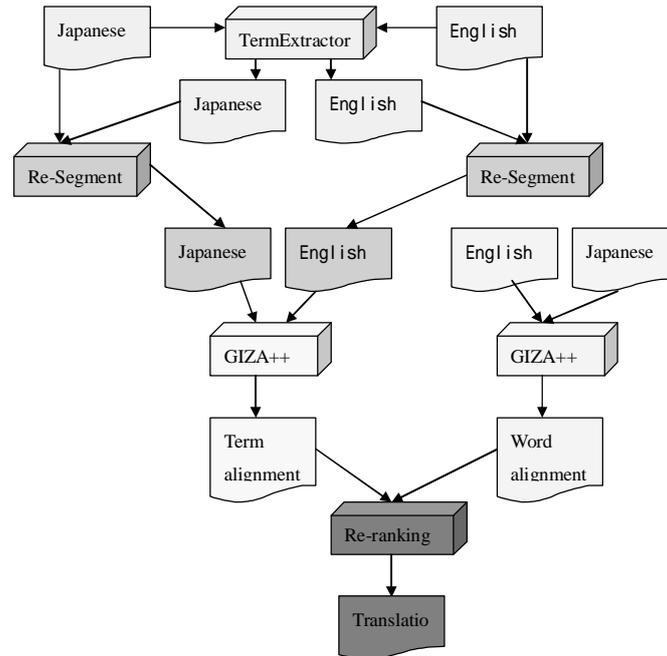


図 1. 提案手法の抽出アルゴリズム

本アルゴリズムでは最初にそれぞれの文を専門用語抽出システムで処理し、日本語および英語の専門用語が抽出される。本研究では言選Web (TermExtractor) を用いて専門用語を抽出する。言選Webは一般に利用可能な専門用語抽出システムであり、日本語、中国語、英語、フランス語、ドイツ語、イタリア語、スペイン語、フィンランド語、スウェーデン語に対応する。本研究では、日本語の言選Webと英語の言選Webを用いた。

次に、専門用語抽出システムにより得られた専門用語を用いて、日本語文と英語文の単語境界を再計算し (リセグメント処理)、単語アライメントシステム (GIZA++) に入力する (専門用語付き単語アライメント)。日本語や英語はトークナイザーにより単語に分割されているが、これらの単語は専門用語よりも短い単位の単語であるため、そのまま単語アライメントを行うと専門用語に対する対訳確率を得ることは難しく、結果として専門用語の抽出が難しくなる。例えば、「組合せ範疇文法を書く」「write a Combinatory Categorical Grammar」という日英の対訳文の対が与えられた時、トークナイザーは「組合せ | 範疇 | 文法 | を | 書く」と分割するため、「組合せ /Combinatory」「範疇/Categorical」「文法/Grammar」に対する対訳確率やそれらの句の対訳確率が計算されるが、学習が不十分であったりノイズが多い場合には「組合せ範疇文法」に対する

「Combinatorial Categorical Grammar」という専門用語の対訳語は得られない。一方、「組合せ範疇文法」や「Combinatory Categorical Grammar」が専門用語であることがわかっていれば、それらを一つの単語として単語アライメントを行うことにより、より精度の高い専門用語の対訳辞書が得られること、および、統計モデルの学習が容易になることが期待される。本研究で提案するリセグメント処理は、専門用語と判定される単語列を一つの単語として区切り直す処理である。例えば、上記の例では、入力として「組合せ | 範疇 | 文法 | を | 書く」を与えたとき、「組合せ範疇文法 | を | 書く」と単語境界を区切り直す。

表 1 言選 Web で得られた専門用語

順位	日本語用語		順位	英語用語	
	用語	重要度		用語	重要度
1	装置	231979076.92	1	method	1924924798.54
2	製造方法	196882071.51	2	device	243593927.71
3	画像形成装置	103913088.99	3	manufacturing method	226783016.55
4	半導体装置	58227570.41	4	apparatus	76048463.50
5	制御装置	36336752.48	5	image forming apparatus	55033626.33
6	画像処理装置	30031743.03	6	system	53561371.66
7	記録媒体	27930678.85	7	semiconductor device	50802151.44
8	表示装置	26675671.25	8	program	37869450.18
9	方法	26202837.07	9	control method	32413791.38
10	プログラム	25879734.87	10	method of manufacturing	26542475.31
11	製造装置	23677249.52	11	display device	22816450.59
12	記録装置	19022693.66	12	manufacturing	26336501.49
13	制御方法	18220829.11	13	recording medium	17173321.02
14	遊技機	17278888.05	14	image forming device	15427921.08
15	画像表示装置	12268117.13	15	game machine	15038502.03
16	液晶表示装置	11755384.71	16	control device	12632549.65
17	システム	11323757.36	17	image forming method	12481169.99
18	電子機器	10261766.30	18	using	8747302.73
19	処理装置	8791257.97	19	image processing method	8442823.37
20	通信装置	8579941.15	20	vehicle	8192541.72
21	処理方法	8372906.90	21	production method	8140130.48
22	画像形成方法	8117459.66	22	liquid crystal display device	7886150.69
23	情報処理装置	8100336.44	23	producing	7863396.49
24	画像処理方法	7274028.95	24	image display device	7673459.91
25	撮像装置	6806003.22	25	program	7431868.57
26	形成方法	6314294.98	26	manufacturing semiconductor device	7399380.43
27	電気光学装置	6278752.47	27	method of manufacturing semiconductor device	6713068.93
28	印刷装置	5920852.82	28	manufacturing method therefor	6615284.23
29	基板処理装置	5390500.11	29	electronic device	6487897.04
30	インクジェット記録装置	5291824.63	30	structure	6476402.13

最後に、本研究で提案するリランキング手法を適用する。上述のリセグメント処理により得られる専門用語の対訳対は専門用語が頻出しない場合にはデータスパースネスのため却って得られる専門用語の質が低下することがある。そこで、リセグメント処理を行っていない文書対に対しても単語アライメントシステム (GIZA++) を適用し、専門用語を構成する各単語の対訳確率を用いて、専門用語の対訳確率のスムージングを行う。

### 3.1.4 特許文からの対訳用語抽出の実験

特許データ日英 364,449 件のタイトルに対し、本研究で提案する手法を用いて、対訳用語の抽出実験を行った。英語の分かち書きは TreeTagger に付属するスクリプト、日本語の分かち書きは茶筌を用いた。専門用語抽出ツールは中川研究室の言選 Web を用いた。

表 1 は特許コーパスの日本語文および英語文に対し言選 Web が出力する上位 30 個の日本語の専門用語と英語の専門用語である。結果を見ると、日本語の専門用語の精度が非常に高い。しかし、英語の専門用語では”-ing”のような三つの動名詞を抽出された。これは stemming 処理をして

表 2 提案手法により得られた対訳語の例

順番	
1	地図データ送信方法 map data transmission method 1
2	研削用水溶性加工油 water-soluble machining oil 1
3	金属屑塊成形装置 forming waste metal block 1
4	写真プリント品質管理ツールセット photographic print quality control tool set 1
5	成形用積層発泡シート laminated foamed sheet 1
6	穴付物干し竿装置 clothespole device 1
7	熱収縮性複合シート heat-shrinkable composite sheet 1
8	白黒画像形成装置 black-and-white image forming device 1
9	立体視用表示方法 stereoscopic display method 1
10	薄板状材料積層体 thin flat laminar body 1
11	軟磁性樹脂成型品 magnetic resin molded article 1
12	レーザービーム径 measuring diameter of laser beam 1
13	ドレッシング工具 dressing tool 1
14	砕氷攪拌ドラム crushed ice stirring drum 1
15	空間像計測方法 measuring spatial image 1
16	プリモールドパッケージ用リードフレーム pre-mold package 1
17	プリモールドパッケージ pre-mold package 1
18	テーマ管理システム theme management system 1
19	ブラシレスDCモータ制御方法 brushless dc motor control method 1
20	通知情報表示方法 notified information displaying method 1

いないことが原因と考えられる。

表 2 は提案する手法によって得られた対訳用語を示している。上位 20 個の結果を見ると、一部対応がとれていない専門用語対もあるが、長い専門用語の精度はとても良い。

### 3.1.5 まとめと今後の課題

本稿では、従来の対訳語抽出技術の調査を行い、専門用語抽出とアライメント技術を組み合わせる方法を提案し、日英の特許コーパスを用いてその性能を評価した。今後は、この手法の精度向上に関する研究を行う。

### 参考文献

張玉潔、馬青、井佐原均：英語を介した日中对訳辞書の自動構築、*自然言語処理*、Vol. 12, No. 2, pp. 63-85, 2005.

Ker, S. J. and J. S. Chang. 1997. A Class-based Approach to Word Alignment. *Computational Linguistics*, Vol. 23, No. 2, pp. 313-343.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263-311.

Toutanova, K., H. T. Ilhan and C. D. Manning. 2002. Extensions to HMM-based Statistical Word Alignment Models. In *Proc. of EMNLP-2002*, pp. 87-94.

Och, F. J and H. Ney. 2002., Improved Statistical Alignment Models, In Proc. of ACL 2002, pp. 440-447.

Och, F. J. and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models . Computational Linguistics, Vol. 29, No. 1, pp. 19-51.

Gale, W. A. and Kenneth W. Church. 1991. Identifying word correspondences in parallel texts. In Proc. of the Fourth DARPA Speech and Natural Language Workshop, pp. 152-157.

Chiao, Y. and P. Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In Proc. of COLING 2002, pp. 1-5.

Gamallo Otero, P. 2007. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In Proc. of Machine Translation Summit XI.

Brill, E., G. Kacmarcik and C. Brockett. 2001. Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs, In Proc. of NLPRS 2001, pp. 393-399.

Nazar, R., L. Wanner and J. Vivaldi. 2008. Two-Step Flow in Bilingual Lexicon Extraction from Unrelated Corpora. In Proc. of EAMT 2008, pp. 140-149

## 3.2 第3言語を介した対訳辞書の構築

静岡大学情報学部情報科学科

梶 博行

【要約】一方の言語が同一である2つの対訳辞書から新しい言語対の対訳辞書を生成する新しい方法を提案する。媒介となる第3言語の語の多義性によって生み出される誤った訳語を除去するため、第1言語と第2言語それぞれの単言語コーパスを用いることが特徴である。すなわち、それぞれの言語のコーパスから関連の強い語のペアを抽出して、第1言語の関連語ペアと第2言語の関連語ペアのアラインメントをとることにより、第1言語の見出し語の各関連語が第2言語のどの訳語を支持するかを決定する。誤った訳語を支持する関連語は非常に少ないので、誤った訳語を除去することができる。日英と中英の対訳辞書から日中对訳辞書を生成する予備実験を行い、提案方法のフィージビリティを確認した。

### 3.2.1 はじめに

機械翻訳やクロス言語情報検索のニーズがさまざまな言語対に広がってきている。しかし、大規模な対訳辞書が利用できる言語対は限られている。このため、新しい言語対の対訳辞書を低コストで作成することが課題となっている。

大規模な対訳辞書が存在しない言語対であっても、それぞれの言語と第3の言語（たいていは英語）の間に大規模な対訳辞書が存在するケースが多い。そこで、第1言語 - 第3言語の対訳辞書と第2言語 - 第3言語の対訳辞書から第1言語 - 第2言語の対訳辞書を合成することが考えられる。第1言語の語と第2言語の語が対訳関係にあるとき、それらに共通の第3言語の訳語が存在するからである。しかし、媒介となる第3言語の訳語が多義語である場合、誤った訳語が混入するという問題がある。

この問題に対し、Tanaka and Umemura (1994) は、媒介となる第3言語の訳語の数が多いほど正しい訳語である可能性が高いという仮説をたて、その有効性を実証した。しかし、この考え方だけでは十分な精度が達成できないことも明らかである。Tanaka and Umemuraの方法を基本としながら、意味カテゴリが一致するペアを優先する方法 (Bond, et al., 2001) あるいは品詞や構成文字の関連性を考慮する方法 (Zhang, et al., 2007) が提案されているが、十分な解決には至っていない。

本稿では、第1言語と第2言語のコーパスを利用して誤った訳語を除去する新しい方法について述べる。ここで、両言語のコーパスは同じ分野のコーパスであればよく、パラレルコーパスやコンパラブルコーパスである必要はない。第3言語を介して対訳辞書を作成しようとする言語対では、通常、大規模なパラレル/コンパラブルコーパスの存在が期待できないことを考慮した結果である。

なお、3.2.2.1 で述べるように、提案方法では、誤った訳語だけでなく、対象分野では使用されない訳語も除去される。したがって、分野を限定しない一般的な対訳辞書を生成することは

きない。これは提案方法の限界であるが、分野に適応した対訳辞書が生成されるということはむしろ優れた特徴であるといえる。

### 3.2.2 提案方法

#### 3.2.2.1 基本アイデア

提案方法は任意の言語対に適用可能であるが、本稿では英語を介して日本語 - 中国語の対訳辞書を生成する場合について述べる。図 3-2-1 に、日英対訳辞書と中英対訳辞書の結合によって得られる日本語と中国語の語の対応関係を例示する。日本語の見出し語「工場」に対して、「工場」と「厂」は正しい訳語であるが、「植株」と「作品」は誤った訳語（以下、ノイズという）である。これらは、媒介となる英語の語 “works” や “plant” の多義性に起因している。このようなノイズを除去することが必要である。

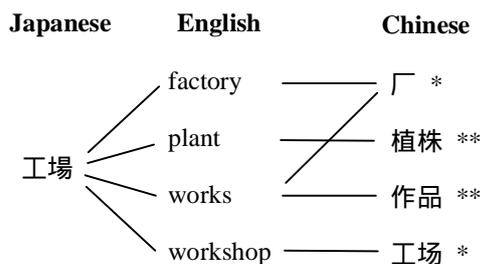
このため、2つの言語の単言語コーパスを用いた教師なしの語義の曖昧性解消方法として考案したアルゴリズム（Kaji and Morimoto, 2005）を用いる方法を提案する。このアルゴリズムは、それぞれの言語のコーパスから相互に関連の強い語のペアを抽出し、対訳辞書を介してアラインメントをとることにより、第1言語の各見出し語に対してその関連語と訳語の関連行列を反復計算するアルゴリズムである。このアルゴリズムによると、例えば、「工場」の関連語である「バルブ」や「製造」は、正しい訳語である「工場」や「厂」との関連度が大きく、ノイズである「植株」や「作品」との関連度が小さくなる。日本語コーパスから抽出される「工場」の関連語と中国語コーパスから抽出される「工場」や「厂」の関連語の間には対訳関係が成立するものが多いのに対し、「工場」の関連語と「栽植」や「作品」の関連語の間には対訳関係が成立するものがほとんどないからである。したがって、一定数以上の関連語と高い関連度をもつ訳語のみ選択することにより、ノイズを除去することができる。なお、この方法によれば、正しい訳語であっても、コーパス中で使用されないものはノイズと同様に除去される。

図 3-2-2 に示すように、提案方法は (1) 日英・中英対訳辞書の結合、(2) 各言語のコーパスからの関連語ペアの抽出、(3) 関連語 - 訳語関連行列の反復計算、(4) 訳語のスクリーニングの各ステップから構成される。以下、日中対訳辞書を生成する場合について記述するが、日本語と中国語を入れ替えると中日対訳辞書を生成する場合になる。

#### 3.2.2.2 日英・中英対訳辞書の結合

日英対訳辞書と中英対訳辞書から、共通の英訳語をもつ日本語の語と中国語の語のペアをすべて求めることにより、日中対訳辞書を生成する。この辞書にはノイズが含まれる。

#### 3.2.2.3 関連語ペアの抽出



\*Correct translation, \*\*Incorrect translation

図 3-2-1 英語を介した日本語と中国語の語の対応

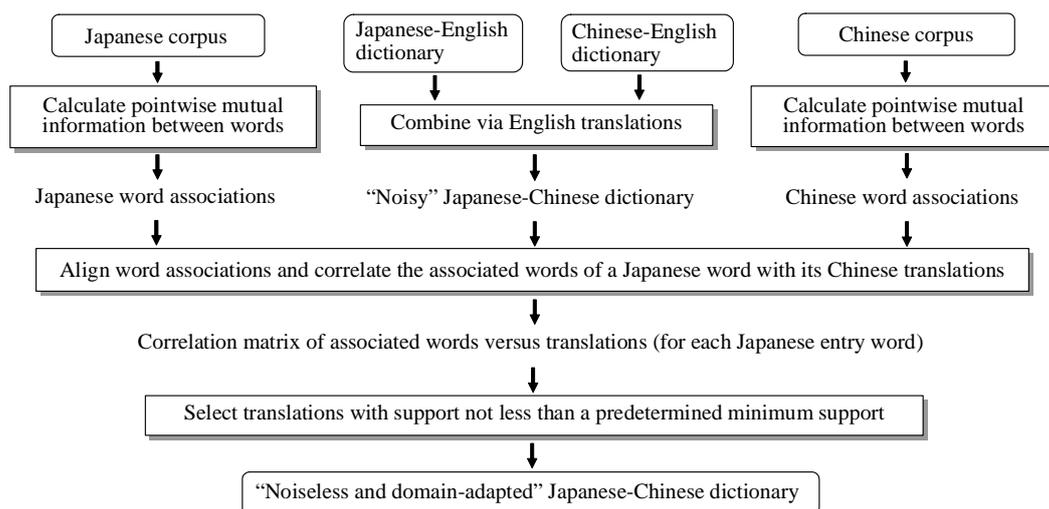


図 3-2-2 提案方法の概要

日本語コーパス、中国語コーパスそれぞれから関連語ペアを抽出する。関連語ペアとは、予め定めた閾値以上の相互情報量をもつ語のペアである。2つの語  $x$  と  $x'$  の相互情報量  $MI(x, x')$  は次式で定義される (Church and Hanks, 1990)。

$$MI(x, x') = \log \frac{\Pr(x, x')}{\Pr(x) \cdot \Pr(x')}, \quad [1]$$

ここに、 $\Pr(x)$  は  $x$  の生起確率、 $\Pr(x, x')$  は  $x$  と  $x'$  の共起確率である。

語と語の共起確率を求めるため、語と語の共起頻度をカウントする。本方法の目的にはトピック的な関連語が有効と思われるので、前後の文にまたがる比較的大きなウィンドウを設定し、ウィンドウ内に共起する頻度をカウントする。対象とする関連語ペアは内容語のペアであるが、両言語のウィンドウの範囲が釣りあうよう、ウィンドウサイズは機能語を除き内容語の数で指定する。3.2.3の実験ではウィンドウサイズを $\pm 25$  内容語とした。

### 3.2.2.4 関連語 - 訳語関連行列の反復計算

関連語と訳語の関連は、基本的には、対訳辞書を介した関連語ペアのアラインメントに基づいて決定することができる。すなわち、見出し語  $x$  とその第  $i$  関連語  $x'(i)$  のペア  $(x, x'(i))$  が第  $j$  訳語  $y(j)$  とその関連語  $y'$  のペア  $(y(j), y')$  と対応づけられるとき、第  $i$  関連語  $x'(i)$  と第  $j$  訳語  $y(j)$  が関連をもつと考えることができる。しかし、これには次のような問題がある。

- (a) アラインメントの失敗： 両言語のコーパスが内容的に対応していない、あるいは対訳辞書のカバレッジが十分でないため、アラインメントがとれる関連語ペアは必ずしも多くない。
- (b) アラインメントの曖昧性： 対訳辞書を介したアラインメントでは、偶然対応がつく関連語ペアもあって、関連語ペアの対応にはしばしば曖昧性が生じる。

これらの問題点を解決するため、関連語と訳語の関連度を[2]~[4]式のように再帰的に定義し、[5]式を初期値として反復計算する。式中の添え字  $n$  は反復計算のサイクルを示す。反復計算の収

束の様子などの詳細は ( Kaji and Morimoto, 2005 ) を参照されたい。

- 見出し語  $x$  の第  $i$  関連語  $x'(i)$  と第  $j$  訳語  $y(j)$  の関連度 :

$$C_n(x'(i), y(j)) = MI(x'(i), x) \cdot \frac{C'_n(x'(i), y(j)) + \alpha \cdot C''_n(x'(i), y(j))}{\max_k [C'_n(x'(i), y(k)) + \alpha \cdot C''_n(x'(i), y(k))]}, \quad [2]$$

ここに、 $MI(x'(i), x)$  は  $x'(i)$  と  $x$  の相互情報量、 $\alpha$  は関連度の第 1 因子に対する関連度の第 2 因子の相対的な重みを表すパラメータである。

[2]式からわかるように、関連度  $C$  は「関連語と見出し語の相互情報量」と「2つの因子  $C'$  と  $C''$  の重みつき和を ( 最大値が 1 になるように ) 正規化した値」の積である。

- 関連度の第 1 因子 :

$$C'_n(x'(i), y(j)) = \sum_{x'' \in A(x, x'(i))} C_{n-1}(x'', y(j)), \quad [3]$$

ここに、 $A(x, x'(i))$  は  $x$  と  $x'(i)$  に共通の関連語の集合である。

[3]式からわかるように、第 1 因子  $C'$  は他の関連語と当該訳語の関連度を用いて定義されている。これは、互いに関連する関連語は同じ訳語と強い関連をもつという仮説に基づいており、上述の問題点(a)を解決する。

- 関連度の第 2 因子 :

$$C''_n(x'(i), y(j)) = \max_{y'} \left( MI(y', y(j)) \cdot \sum_{x'' \in B((x, x'(i)), (y(j), y'))} C_{n-1}(x'', y(j)) \right), \quad [4]$$

ここに、 $B((x, x'(i)), (y(j), y'))$  は、 $x$  と  $x'(i)$  に共通の関連語であって、その中国語訳の少なくとも一つが  $x$  の中国語訳  $y(j)$  と  $x'(i)$  の中国語訳  $y'$  に共通の関連語であるものの集合である。

[4]式からわかるように、第 2 因子  $C''$  は、関連語の 3 つ組のアラインメントを構成する関連語と訳語のペアに対してのみ定義される。これは、関連語ペアのアラインメントと比較して関連語の 3 つ組のアラインメントが信頼できることに着目したもので、上述の問題点(b)を解決する。

- 関連度の初期値 :

$$C_0(x'(i), y(j)) = MI(x'(i), x). \quad [5]$$

### 3.2.2.5 訳語のスクリーニング

関連語 - 訳語関連行列を 2 値化する。すなわち、各行 ( 関連語 ) において最大値をもつ要素を 1、それ以外の要素を 0 にする。

	厂	植株	作品	工場
パルプ	3.98	0.81	0.23	3.75
工程	3.18	0.45	0.58	3.45
飼料	0.54	3.32	0.17	1.18
製造	2.89	0.43	0.61	3.13
:	:	:	:	:

(a) “工場”に対する関連行列

	厂	植株	作品	工場
パルプ	1	0	0	0
工程	0	0	0	1
飼料	0	1	0	0
製造	0	0	0	1
:	:	:	:	:

Support 0.03 0.04 0.00 0.93

(b) 2 値行列と支持度

図 3-2-3 関連語 - 訳語関連行列と訳語の支持度

そして、列（訳語）ごとに、1 の要素の比率を算出する。この比率は、各関連語が1つの訳語を支持すると考えたときに各訳語を支持する関連語の比率を表すので、支持度と呼ぶ。最後に、支持度が予め定めた閾値以上となる訳語を選定する。

関連語 - 訳語関連行列と支持度の例を図 3-2-3 に示す。この例で支持度の閾値を 0.1 とすれば、日本語見出し語「工場」の中国語訳語として「工場」が選定され、ノイズである「栽植」と「作品」は除去される。なお、正しい訳語「厂」も除去されるが、これは使用したコーパスにおいてあまり使用されていなかったためと考えられる。

### 3.2.3 予備実験

#### 3.2.3.1 実験方法

名詞に限定して、EDR 日英対訳辞書と LDC Chinese-English Translation Lexicon から日中対訳辞書と中日対訳辞書を生成する予備実験を行った。使用したコーパスは次のとおりである。

- ・日本語コーパス：毎日新聞記事（2000 年 1 月～2005 年 12 月、632 M バイト）
- ・中国語コーパス：新華社通信記事（1999 年 1 月～2004 年 12 月、473 M バイト）

まず、日本語コーパスと中国語コーパスから名詞の関連語ペアを抽出した。日本語テキストは JUMAN により形態素に分割し、頻度 100 以上の名詞を抽出した。中国語テキストは Nakagawa and Uchimoto (2007) による中国語形態素解析プログラムにより形態素に分割し、頻度 30 以上の名詞を抽出した。両言語とも、前後それぞれ 25 内容語を収容するウィンドウ内に共起する名詞のペアを抽出して頻度をカウントし、名詞のペアそれぞれの相互情報量を計算した。そして、相互情報量が 1.0 以上のペアを関連語ペアとして抽出した。

次に、EDR 日英対訳辞書と LDC Chinese-English Translation Lexicon を結合して作成したノイズを含む日中名詞対訳辞書と日本語、中国語の関連語ペアを用いて、頻度 100 以上の日本語名詞に対し関連語 - 訳語関連行列を計算した。同様に、EDR 日英対訳辞書と LDC Chinese-English Translation Lexicon を結合して作成したノイズを含む中日名詞対訳辞書と中国語、日本語の関連語ペアを用いて、頻度 30 以上の中国語名詞に対し関連語 - 訳語関連行列を計算した。なお、関連行列の反復計算に使用するメモリ量の制約のため、関連語 - 訳語関連行列の関連語数は最大 700 に制限した。

ノイズを含む日中名詞対訳辞書では、頻度 100 以上の日本語名詞 10,003 語のうち 8,284 語が複数の訳語候補をもっており、これらの見出し語の訳語を関連語 - 訳語関連行列を用いてスクリーニングした。同様に、ノイズを含む中日名詞対訳辞書では、頻度 30 以上の中国語名詞 9,288 語のうち 8,426 語が複数の訳語候補をもっており、これらの見出し語の訳語を関連語 - 訳語関連行列を用いてスクリーニングした。

#### 3.2.3.2 実験結果

訳語スクリーニングの結果は支持度に対する閾値によって大きく変わるが、ここでは閾値を 0.1 に設定した。このとき、1 見出し語あたり訳語数は、日本語の 8,284 見出し語の平均で 15.77 から 1.40 に、中国語の 8,426 見出し語の平均で 27.09 から 1.33 に減少した。このように大きく減少した理由の一つとして、ノイズだけでなく対象分野で使われない訳語も除去されるという提案方法

の特質があげられる。

日中対訳辞書のスクリーニング結果の例を表 3-2-1(a)に示す。各見出し語に対し、訳語候補が支持度の降順に示されている。また、閾値(0.1)以上の支持度を得た訳語は太字で示されている。

・見出し語「ホール」

英語の“hall”と“hole”をカタカナ表記した語。“hall”を意味する「**大厅**」は選択されたが、“hole”を意味する「洞」は除去された。「洞」は「オゾン」、「量子力学」などの関連語に支持されたが、支持度は十分でなかったためである。

・見出し語「新聞」

英語の“paper”の多義性に起因するノイズ「**纸张**」(紙)や「**论文**」(論文)がうまく除去された。

・見出し語「電車」

ノイズである「**练**」(訓練)や「**教养**」(教養)とともに正しい訳語「**火车**」が除去された。新華社通信記事コーパス中、「**火车**」が「**列车**」ほど使用されていなかったためと思われる。

日中対訳辞書のスクリーニング結果も同様である。表 3-2-1 (b)に例を示す。

日中対訳辞書のスクリーニングについて再現率と適合率を求めた。ただし、訳語候補が正しい訳語であるかノイズであるかを判定する作業はコストがかかるので、384 見出し語を対象とした。比較のため、従来方法である 1 回逆引き法 (田中ほか, 1998) とその変形版によるスクリーニングについても再現率と適合率を求めた。1 回逆引き法 (IC1 と略記) は、2 語以上の英語の語を介して得られる訳語候補を選択する方法である。IC1 では、どの訳語候補も媒介となる英語の語が 1 語であるような見出し語に対しては訳語が選択されない。そこで、どの訳語候補も媒介となる英語の語が 1 語であるような見出し語に対しては全ての訳語候補を選択する変形版 (IC1' と略記) も用意した。

提案方法と IC1、IC1' の再現率と適合率を表 3-2-2 にまとめた。IC1 では一部の見出し語に対し

表 3-2-1 スクリーニング結果の例

(a) 日中対訳辞書

Entry noun	Translation candidates	Translation candidate	Support
ホール (hall / hole)	大堂, 大厅, 殿, 霍尔, 堂, 厅, 洞, 洞穴, 洞子, 空穴, 孔, 孔洞, 孔隙, 窟, 窟窿, 窿, 漏洞, 穴, 阱, 堀	<b>大厅</b>	<b>0.960</b>
		洞	0.034
		霍尔	0.006
新聞 (news-paper)	报, 报纸, 论文, 论文儿, 纸, 纸头, 纸张, 报界, 报刊, 新闻界	<b>报纸</b>	<b>0.986</b>
		纸张	0.009
		论文	0.001
		纸	0.001
電車 (electric train)	带带, 吊吊, 火车, 教养, 练, 列车, 培训, 培养, 培育, 培植, 训练, 斗车	<b>列车</b>	<b>0.979</b>
		练	0.016
		火车	0.004
		教养	0.001

(b) 中日対訳辞書

Entry noun	Translation candidates	Translation candidate	Support
大厅 (hall)	ホール, 会堂, 会館, 僧堂, 堂, 堂宇, 大広間, 大間, 広座敷, 広敷, 広敷き, and 5 others	<b>廊下</b>	<b>0.791</b>
		<b>ホール</b>	<b>0.203</b>
		会堂	0.003
		大広間	0.001
报纸 (news-paper)	ペーパー, ペーパー, ペーパー, ペーパー, 新紙, 新聞, 新聞紙	<b>新聞</b>	<b>0.993</b>
		ペーパー	0.007
列车 (train)	しつけ, 仕付, 仕付け, 仕立, 仕立て, 列車, 口火, 導火, and 14 others	<b>列車</b>	<b>0.994</b>
		電車	0.006

て結果が得られないので、見出し語ベ-ースの適用率を計算し、表 3-2-2 に含めた。

提案方法は再現率が非常に低いが、これは正しい訳語のうちコーパスで使用されているもののみを選択する方法であるためである。コーパスで使用されている訳語かどうかを手で判定することは極めて困難であるので、コーパスで使用されている/いないに拘わらず、正しい訳語がすべて選択されるべきであるという前提で再現率を計算した。したがって、提案方法の再現率はコーパスで使用されている訳語の比率と解釈したほうがよいかもしれない。適合率について比較すると、提案方法は IC1 や IC1' より優位である。IC1 との差はそれほど大きくないが、IC1 は適用率を犠牲にしていることに注意すべきである。

表 3-2-2 訳語スクリーニングの再現率と適合率

	Proposed method	IC1	IC1'
# entry nouns	384		
# entry nouns for which one or more translations were selected	384	273	384
<b>Applicability ratio</b>	100%	71.1%	100%
# correct translation candidates (S) *	2,270		
# selected translations (T)	553	926	2,485
# selected correct translations (S∩T)	359	565	1,059
<b>Recall</b>	15.8%	24.9%	46.7%
<b>Precision</b>	64.9%	61.0%	42.6%

\* Correctness of 7,410 translation candidates in total for 384 entry nouns was assessed manually. Correct translations missing in the "noisy" Japanese-Chinese dictionary were not included. Therefore, recall is overestimated.

### 3.2.3.3 追加実験1 再スクリーニング

ノイズを含む対訳辞書を用いて計算された関連行列は、ノイズを含まない対訳辞書を用いて計算された関連行列に比べて信頼度が低いかもしれない。なぜなら、ノイズを含む対訳辞書によれば関連語ペアのアラインメントにおける曖昧性が増大するからである。

そこで、再スクリーニングの実験を行った。すなわち、スクリーニングされた対訳辞書を用いて計算しなおした関連語 - 訳語関連行列を用いて、対訳辞書を再度スクリーニングした。なお、第1段階のスクリーニングでは、その目的を考慮し、支持度の閾値を 0.025 という低い値に設定した。いっぽう、第2段階のスクリーニングでは 3.2.3.2 の実験と同じ 0.1 に設定した。

再スクリーニング結果の再現率と適合率を表 3-2-3 に示す。再現率は再スクリーニング前と同程度であるが、適合率が約 5% 向上したことがわかる。

表 3-2-3 追加実験の結果

Method (Minimum support)	Recall	Precision
Basic method (0.1)	17.8%	71.8%
Re-screening (1st stage: 0.025, 2nd stage: 0.1)	16.8%	76.9%
Bidirectional screening - union (0.25)	26.7%	66.8%
Bidirectional screening - intersection (0.025)	18.0%	76.2%

Note: Results for 247 Japanese entry nouns were evaluated.

### 3.2.3.4 追加実験2 双方向スクリーニングの組合せ

対訳辞書は反転することによって逆方向の対訳辞書に変換することができる。そして、

反転した対訳辞書をスクリーニングした結果はもとの対訳辞書をスクリーニングした結果を反転したものと必ずしも一致しない。それゆえ、日中对訳辞書のスクリーニング結果と中日対訳辞書のスクリーニング結果を組み合わせることによってよりよい結果が得られる可能性がある。

#### (1) 双方向スクリーニング結果の和集合

提案方法は、同義の訳語候補があるとき、そのうちの最もよく使用されるもの以外は過小評価されるという欠点をもつ。同義の訳語候補は関連の強い語も共通であるものが多いが、それらの大部分は最もよく使用される訳語との関連度が最大になるからである。例えば、表 3-2-1 において、「火车」の支持度は非常に低かった。その関連語の多くは「列车」との関連度が最大になったためである。

この欠点は反転した辞書のスクリーニングによってある程度解決することができる。比較的使用頻度が小さい語（例えば、上例の「火车」）が見出し語であっても、提案方法は動作するからである。そこで、日中对訳辞書のスクリーニング結果と中日対訳辞書のスクリーニング結果を反転した辞書の和集合をとり、評価した。この実験では、支持度の閾値を比較的大きく 0.25 に設定した。あまり使用されない訳語は、逆方向の辞書のスクリーニング結果を反転させることによって得られると期待してよいからである。

双方向のスクリーニング結果の和集合をとった場合の再現率と適合率を表 3-2-3 に示す。単方向のスクリーニング結果に比べて、適合率が 5%ほど低下したが再現率が大きく向上している。

#### (2) 双方向スクリーニング結果の共通部分

日中对訳辞書のスクリーニング結果と中日対訳辞書のスクリーニング結果の両方に含まれる対訳語ペアは、支持度が高なくても確実性が高いと考えられる。そこで、そこで、支持度の閾値をかなり小さな 0.025 に設定し、日中对訳辞書と中日対訳辞書の両方をスクリーニングし、それらの結果の共通部分をとって、評価した。

双方向のスクリーニング結果の共通部分の再現率と適合率を表 3-2-3 に示す。単方向のスクリーニング結果に比べて、再現率を同程度に保ったまま、適合率が約 4%向上している。

### 3.2.4 改良の方向

3.2.3 に示した実験は提案方法が有望であることを示しているが、改良の余地も大きい。まず、さまざまなパラメータを最適化することが必要である。例えば、関連語ペア抽出におけるウィンドウサイズと語の頻度や相互情報量に対する閾値、関連語 - 訳語関連行列の関連語の数、支持度に対する閾値などのパラメータがある。3.2.3 の実験を通じて、特に関連語ペア抽出におけるパラメータの値がスクリーニング結果に大きく影響することがわかっている。それに比べて、関連語 - 訳語関連行列の計算におけるパラメータ、例えば関連度の第 1 因子、第 2 因子を重みづけるパラメータ  $\alpha$  や反復計算の反復回数などはあまり問題がない。関連語 - 訳語関連行列の反復計算アルゴリズムが比較的広い範囲のパラメータ値に対して安定に動作することが確認されている (Kaji and Morimoto, 2005)。

3.2.3.3 の再スクリーニングと 3.2.3.4 の双方向スクリーニングはともに有効であった。これらを組み合わせることももちろん可能である。なお、双方向スクリーニングにおいて和集合をと

るか共通部分をとるかは、再現率と適合率のどちらを重視するかによる。

提案方法によって生成される対訳辞書が、パラメータの値だけでなく入力として用いる対訳辞書によって大きく変化することは明らかである。提案方法は、日英・中英対訳辞書を結合したノイズを含む日中対訳辞書に含まれる訳語候補の中から選択するだけであり、候補以外の訳語を抽出することはできない。そこで、

3.2.3 の実験で用いたノイズを含む日中対訳辞書に訳語候補を補充した上でスクリーニングする追加実験を行った。表 3-2-4 に、3つの見出し語について、訳語候補補充の前と後のスクリーニング結果を示す。これからわかるように、適切な訳語候補を補充することにより、再現率だけでなく適合率も向上する。実際的な観点からは、EDR 日英対訳辞書と LDC 中英対訳辞書だけでは不十分であり、日英、中英ともよりカバレッジの高い辞書を用いることが必要である。

表 3-2-4 訳語候補を補充した日中対訳辞書のスクリーニング結果の例

Entry noun	Before supplementation	After supplementation
難民 ( <i>refugees</i> )	受難者	难民
先端 ( <i>point, tip</i> )	终端, 小费, 论点	顶端
故郷 ( <i>hometown</i> )	底座, 由来	家乡, 故乡

### 3.2.5 関連研究

第3言語を介して対訳辞書を生成する方法は分野を限定しない一般的な辞書を生成する方法と分野に依存した辞書を生成する方法に分けられる。前者には、Tanaka and Umemura (1994) の1回 / 2回逆引き法とそれに加えて他の手がかりを利用する方法がある (Bond, et al., 2001; Shirai and Yamamoto, 2001; Zhang, et al., 2007)。後者として提案された方法はないが、非パラレルコーパスから出現文脈が類似している語のペアを対訳語ペアとして抽出する方法 (Rapp, 1995; Kaji and Aizono, 1996; Tanaka and Iwasaki, 1996; Fung and Yee, 1998; Rapp, 1999) が適用できることは明らかである。Sammer and Soderland (2007) が、さまざまな言語の単言語コーパスとさまざまな言語対の対訳辞書から PanLexicon と呼ぶ多言語辞書を作成するために、出現文脈の類似度に基づく方法を利用している。

出現文脈の類似度に基づく方法と本稿で提案した方法はともに「ある言語において強い関連をもつ語の訳語はやはりその言語において強い関連をもつ」という仮説に基づいている。2つの方法の違いは、対訳知識の獲得と語義の曖昧性解消というもともとの目的の違いによる。誤った訳語を除去する方法として、出現文脈の類似度に基づく方法には、「文脈類似度の閾値を全ての見出し語に共通に設定することが困難である」という問題がある。これに対し、提案方法は「訳語の支持度に対する閾値を全ての見出し語に共通に設定することができる」という利点をもつ。さらに、提案方法においては支持する関連語とともに訳語が選択されるので、テキスト中の各インスタンスに対し文脈に基づいて訳語を選択することができるという優れた特長をもつ。

### 3.2.6 おわりに

一方の言語が同一である2つの対訳辞書から新しい言語対の“分野適応”対訳辞書を生成する

方法を提案した。この方法はそれぞれの言語の単言語コーパスを必要とするが、単言語コーパスはパラレルコーパスに比べるとはるかに多くの分野で利用可能である。

提案方法の骨子は次のとおりである。2つの対訳辞書を結合して得られる対訳辞書は正しい訳語だけでなく誤った訳語を含んでいる。そこで、見出し語の関連語と訳語の関連語をそれぞれの言語のコーパスから抽出し、対訳辞書を介してアラインメントをとることにより、各関連語が支持する訳語を決定する。ここでは、アラインメントの失敗やアラインメントの曖昧性に対処するため、関連語と訳語の関連度を再帰的に定義し、反復計算する。各訳語の支持度、すなわち各訳語を支持する関連語の比率を求め、予め定めた閾値以上の支持度をもつ訳語を選択することにより、対訳辞書から誤った訳語が除去される。

EDR 日英対訳辞書と LDC 中英対訳辞書、毎日新聞記事コーパスと新華社通信記事コーパスを用いた予備実験を行い、提案方法のフィージビリティを確認した。提案方法に含まれるさまざまなパラメータを最適化することが今後の課題である。また、実用的な観点からは、より高カバレッジの日英・中英対訳辞書を用いることが重要である。

**謝辞：** 本研究の一部は科学技術振興調整費によるプロジェクト「日中・中日言語処理技術の開発研究」および(株)日立製作所との共同研究として実施した。

## 参考文献

- Bond, Francis, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. (2001). Design and construction of a machine-tractable Japanese-Malay dictionary. In *Proceedings of Machine Translation Summit VIII*, pp. 53-58.
- Church, Kenneth and Patrick Hanks. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1): 22-29.
- Fung, Pascale and Lo Yuen Yee. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp. 414-420.
- Kaji, Hiroyuki and Toshiko Aizono. (1996). Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 23-28.
- Kaji, Hiroyuki and Yasutsugu Morimoto. (2002). Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 411-417.
- Kaji, Hiroyuki and Yasutsugu Morimoto. (2005). Unsupervised word-sense disambiguation using bilingual comparable corpora. *IEICE Transactions on Information and Systems*, E88-D(2): 289-301.
- Nakagawa, Tetsuji and Kiyotaka Uchimoto. (2007). A hybrid approach to word segmentation and POS tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 217-220.
- Rapp, Reinhard. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.
- Rapp, Reinhard. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp.

519-526.

- Sammer, Marcus and Stephen Soderland. (2007). Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In *Proceedings of Machine Translation Summit XI*, pp. 399-406.
- Shirai, Satoshi and Kazuhide Yamamoto. (2001). Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *Proceedings of the 19th International Conference on Computer Processing of Oriental Languages*, pp. 174-179.
- Tanaka, Kumiko and Hideya Iwasaki. (1996). Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 580-585.
- Tanaka, Kumiko and Kyoji Umemura. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 297-303.
- 田中久美子, 梅村恭司, 岩崎英哉. 1998. 第三言語を介した対訳辞書の作成. *情報処理学会論文誌*, Vol. 39, No. 6, pp. 1915-1924.
- Zhang, Yujie, Qing Ma, and Hitoshi Isahara. (2007). Building Japanese-Chinese translation dictionary based on EDR Japanese-English bilingual dictionary. In *Proceedings of Machine Translation Summit XI*, pp. 551-557.

### 3.3 対訳特許文書からの専門用語対訳辞書生成：

#### 機械学習による異種情報源の統合

筑波大学 大学院 システム情報工学研究科

知能機能システム専攻 森下 洋平

宇津呂 武仁

コンピュータサイエンス専攻 山本 幹雄

##### 3.3.1 はじめに

本研究では、対訳特許文書から専門用語訳語対を獲得する手法として、フレーズベース統計的機械翻訳モデル[1]により学習されるフレーズテーブルと、既存の対訳辞書を用いる要素合成法[3]を併用する。評価実験においては、まずフレーズテーブルにより生成された訳語候補と要素合成法により生成された訳語候補をそれぞれ単独で評価し、また、翻訳資源が異なるこれら二手法によって生成された訳語候補が一致する場合について、その訳語候補を評価する。さらに、Support Vector Machines (SVM)[4]を用いた機械学習により、得られた訳語候補の検証を行う。SVMの素性は、既存の対訳辞書を利用したものや、全対訳文から得られる統計量などを用いた。その結果、SVMを用いることによりフレーズテーブルを用いた訳語候補推定の性能を改善することができた。

##### 3.3.2 日英対訳特許文

本研究では、NTCIR-7の特許翻訳タスクで配布された1,798,571件の文対応データを、フレーズテーブルと統計的共起測定法の学習データとして、またその中の400件の日本語文から抜き出した専門用語を評価対象データとして使用した。特許翻訳タスクとは、ワークグループ参加者に特許文書を対象とした対訳文や対訳文書、評価セットといった機械翻訳における共通基盤となるデータを提供し、さらにそれらを用いて、特定の課題を行うことによって、機械翻訳の研究レベルの底上げを図ったタスクである。なお、配布された文対応データは、以下の手順で得られたものである。

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、日米対訳特許文書を取得する。
3. 日米特許で互いに対応関係にある部分（背景、実施例）を抽出し、文アラインメントをつける。

##### 3.3.3 訳語推定手法

###### 3.3.3.1 英辞郎

専門用語が、既存の辞書に登録されているか否かを調べるために、既存の対訳辞書として、収録

語数約 129 万語である英辞郎<sup>1</sup>Ver.79 を使用した。

### 3.3.3.2 要素合成法

名詞句を構成要素に分解し、既存の対訳辞書(英辞郎)を用いて構成要素ごとに訳語を求め、それらを再構成して全体の訳を得る要素合成法[3]を用いる。要素合成法によって、対象日本語名詞句の訳語候補と、それらが対応するスコアを求める。図 1 に具体的な手順を示す。

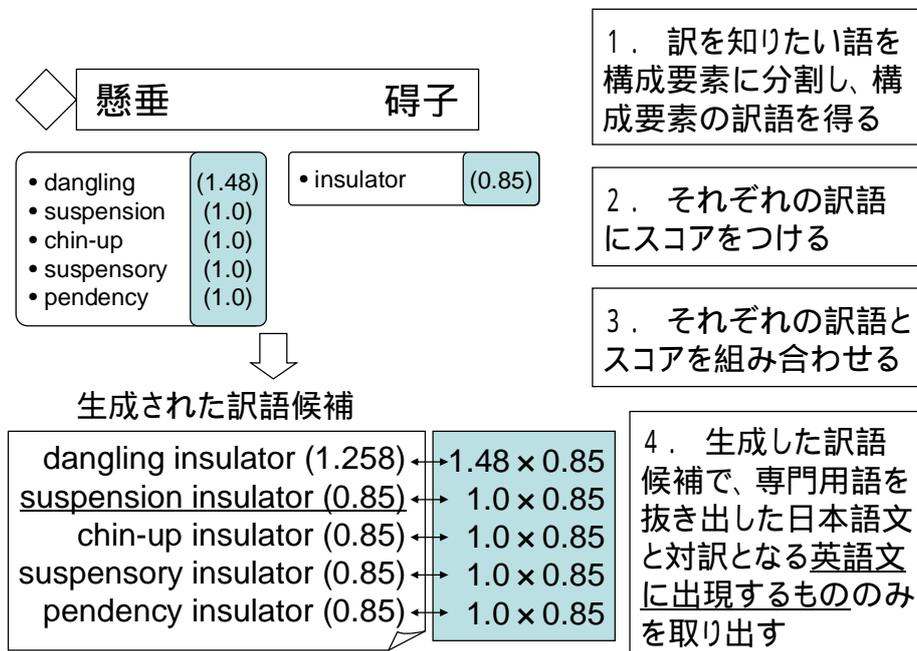


図 1 要素合成法による訳語推定

構成要素の訳語を求め、スコアを付与するのに、要素合成法では 2 種類の辞書を用いる。既存の対訳辞書である英辞郎と、英辞郎から生成した部分対訳辞書である。部分対訳辞書とは、「複合語中の構成要素がどのように訳されるのが自然か」の情報を含む辞書である。例えば、「applied」は、一般的には「応用の」、「応用された」などに訳されるが、複合語としては

applied mathematics 応用 数学

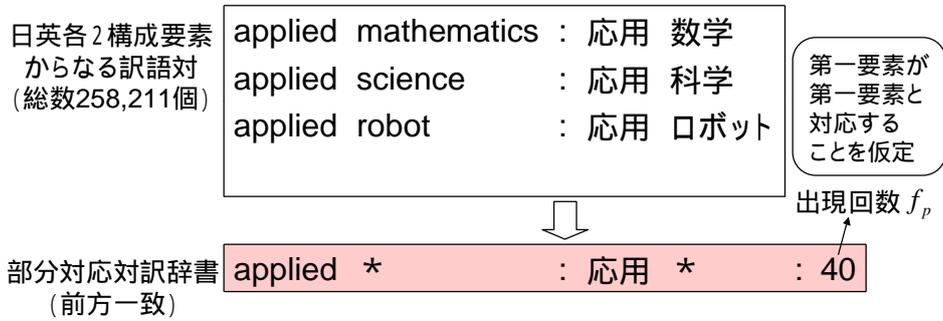
のように、「応用」と訳すことが多い。それにもかかわらず、既存の対訳辞書には applied 応用という対応が登録されておらず、このような組み合わせは数多く存在する。部分対訳辞書は、そのような複合語としての自然な訳を既存の対訳辞書から生成し、登録した辞書である。また、部分対訳辞書には 2 つの種類があり、前節語として自然な訳を登録する前方一致部分対訳辞書、後節語として自然な訳を登録する後方一致部分対訳辞書がある。前方一致部分対訳辞書の具体的な作成方法を図 2 に示す。

<sup>1</sup> <http://www.eijiro.jp/>

日英各2構成要素からなる訳語対を英辞郎から抽出 ( $P_2$ )

日英とも共通の第一要素からなる訳語対を $P_2$  から収集

日英双方の第一要素からなる部分対応訳語対を作成



22

図 2 前方一致部分対応訳語対の作成方法

後方一致部分対応訳語対も同様の手順で作成する。

また、要素合成法のスコアの付け方を図 3 に示す。

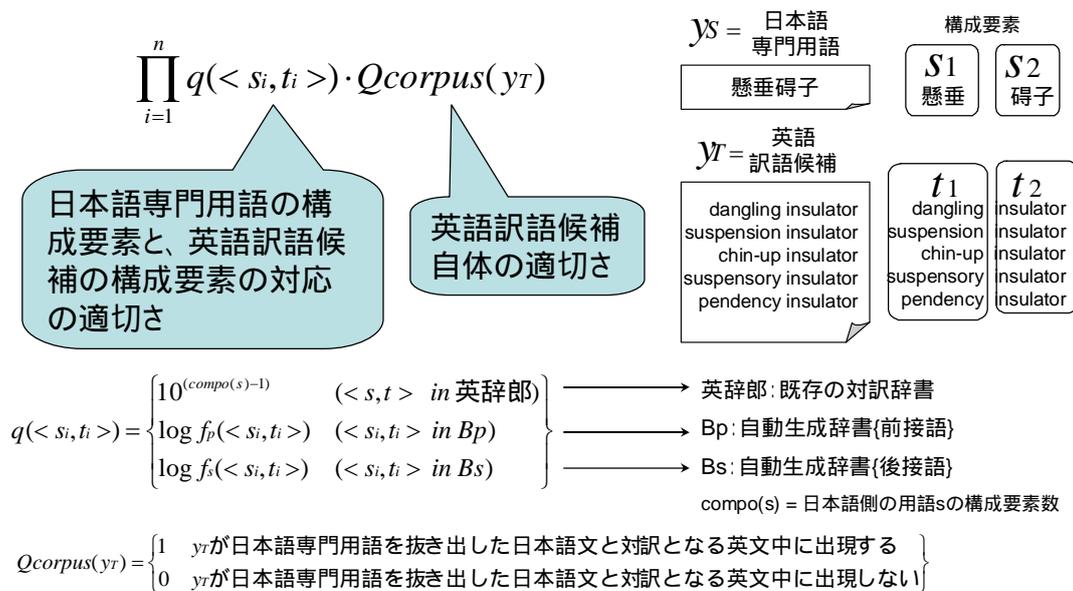


図 3: 要素合成法のスコアの付け方

### 3.3.3.3 フレーズテーブル

フレーズベースの統計的機械翻訳モデルのツールキットである Moses を用いて、3.3.2 節で述べた文対応データから、対応しやすいフレーズペア、およびフレーズペアが対応する確率を示したフレーズテーブルを作成する。以下に Moses がフレーズテーブルを作成する過程を示す。

1. 文対応データの前処理として、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を行う。
2. IBM モデルにより文対応データから単語対応を生成するツールである GIZA++を用いて、最尤な単語対応を得る。英日、日英の両方向で行う。求めた英日最尤単語対応の例を図 4 に示す。

ある	.	.	.	.	.	.	.	.	.	.	.	.
で	.	.	.	.	.	.	.	.	.	.	.	.
同様	.	.	.	.	.	.	.	.	.	.	.	.
と	.	.	.	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.	.	.	.
図	.	.	.	.	.	.	.	.	.	.	.	.
は	.	.	.	.	.	.	.	.	.	.	.	.
以外	.	.	.	.	.	.	.	.	.	.	.	.
それ	.	.	.	.	.	.	.	.	.	.	.	.
	the	remains	are	the	same	as	is	fig	12	$e_0$		

図 4: 英日最尤単語対応

3. 日英両方向の単語対応から，対称な単語対応をヒューリスティックスを用いて得る．対称な単語対応の例を図 5 に示す．

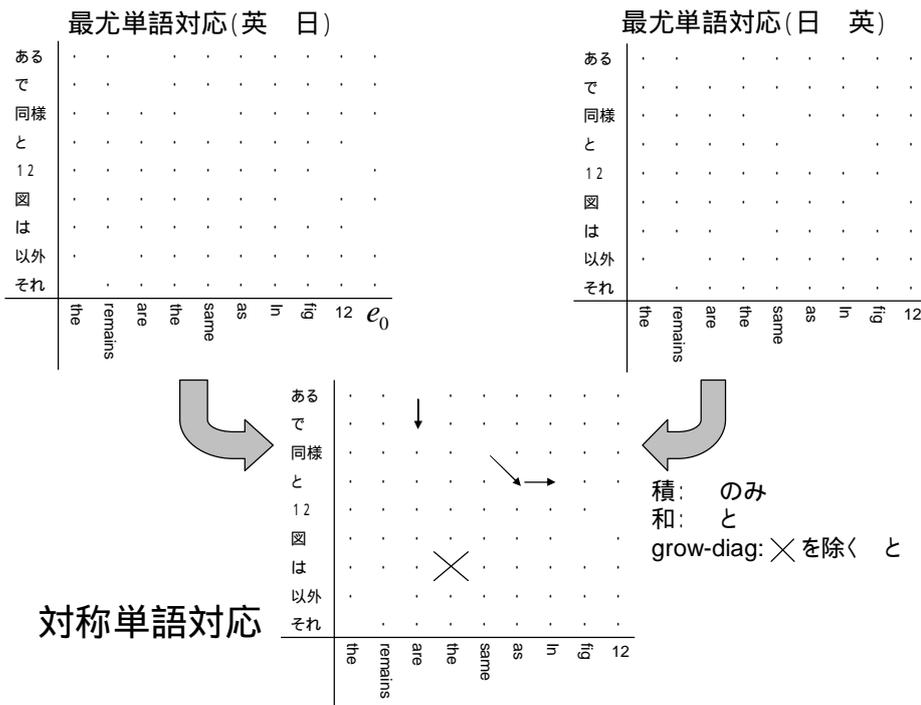


図 5: 対称単語対応

4. 対称な単語対応表に矛盾しない、一貫したフレーズ対応を得る。一貫したフレーズ対応の例を図 6 に示す。対称の単語対応を図のようなテーブルに置き換え、英日で任意の単語対応の固まりを四角で囲んだ時、灰色部分の領域に他の単語対応が存在しないフレーズのペアが一貫したフレーズ対応である。このようなフレーズ対応を全て抽出する。

ある	.	.	.	.	.	.	.	.	.
で	.	.	.	.	.	.	.	.	.
同様	.	.	.	.	.	.	.	.	.
と	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.
図	.	.	.	.	.	.	.	.	.
は	.	.	.	.	.	.	.	.	.
以外	.	.	.	.	.	.	.	.	.
それ	.	.	.	.	.	.	.	.	.
	the	remains	are	the	same	as	in	fig	12

図 6: 一貫したフレーズ対応

5. 学習データからフレーズ対応の数を数えてフレーズ翻訳確率を付与する。  
6. reordering モデル (フレーズの並び替え確率) を計算, 生成する

作成したフレーズテーブルの形式を表 1 に示す。

表 1: フレーズテーブルの形式

日本語フレーズ	英語フレーズ	フレーズ対応のモデルパラメータ					
その結果	as a result	0.242637	0.0629573	0.229377	0.000588454	2.718	
その結果	in consequence	0.132075	0.00664276	0.0014436	4.62987e-05	2.718	
その結果	the cpu 501	0.0227273	4.79776e-06	5.1557e-05	3.97554e-09	2.718	

表 1 は「その結果」という日本語フレーズ周りの対応を示したものである。その結果 as a result という対応が一般的だが、その結果 the cpu 501 というあり得ない対応までフレーズテーブルには登録されている。これは、学習データで「その結果」と「the cpu 501」が同じよ

うな使い方をされていたためである。このような、一般的なフレーズ対応からあり得えにくいフレーズ対応まで網羅したリストがフレーズテーブルである。

フレーズ対応のモデルパラメータは5つとなり、各種確率は、フレーズの英日翻訳確率  $P(ja|en)$ 、英日方向の単語の翻訳確率 (IBM モデル) の積、日英翻訳確率  $P(en|ja)$ 、日英方向の単語の翻訳確率 (IBM モデル) の積、フレーズペナルティ (常に自然対数の底  $e=2.718$ ) となる。今回は、フレーズテーブルのスコアとして、フレーズの日英翻訳確率  $P(en|ja)$  を用いた。

### 3.3.4 各訳語推定手法単独の出力およびその共通部分の性能評価

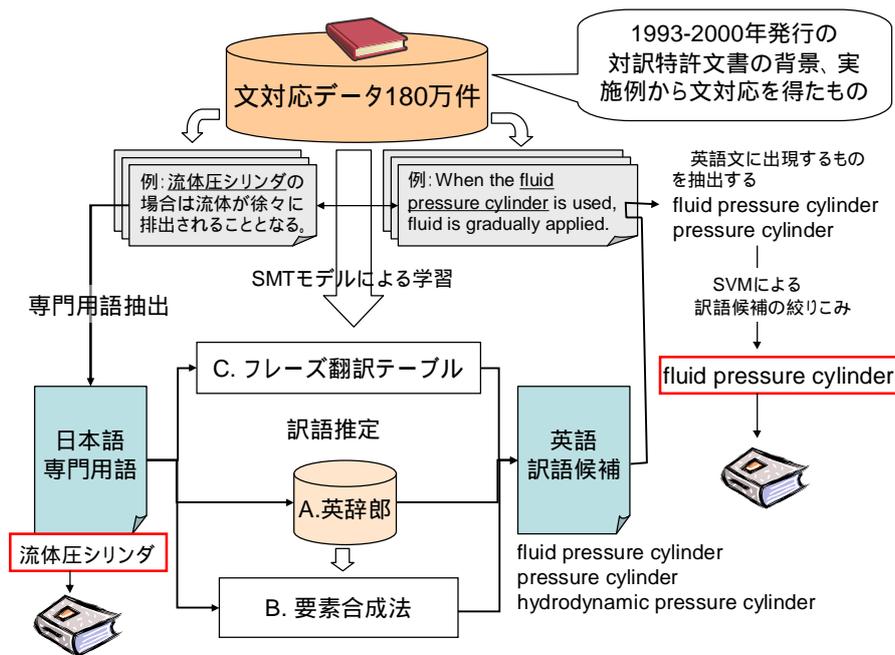


図 7: 評価全体の流れ

全対訳文 180 万件から、IPC 分類が均等になるように、無作為に抽出した対訳文を評価用データとした。そして、以下の日本語専門用語集合 (図 8) に対して、各訳語推定手法単独、もしくは複数の手法を併用した場合の性能を、図 9 に示す。

- (a) 日本語専門用語全体
- (b) 全ての手法により同一の訳語候補を出力できた日本語専門用語
- (c) フレーズテーブルと要素合成法により同一の訳語候補を出力できたが、英辞郎では訳語を生成できなかった日本語専門用語
- (d) フレーズテーブルのみにより訳語候補を生成できた日本語専門用語

集合(d)に対するフレーズテーブルの適合率(88.6%)をベースラインに設定し、これより高い適合率を目指す。三手法全てにより、同一の訳語候補を得られた日本語専門用語集合(b)と、要素合成

法およびフレーズテーブルにより、同一の訳語候補を得られた日本語専門用語集合(c)に対する F 値は 90%を超え、ベースラインより高いものとなった。集合(b)および(c)では、既存の対訳辞書およびフレーズテーブルという、異なる性質を持つ翻訳資源を併用することにより、共通の訳語を得た場合に適合率を改善する結果が得られた。また、集合(b)と(c)を合わせると、全日本語専門用語の 43%に対し、約 95% の適合率を実現している。したがって、以上のような手法により、対訳専門用語の半自動獲得において 高い適合率を実現するという、本論文の目的を達成できることが分かる。

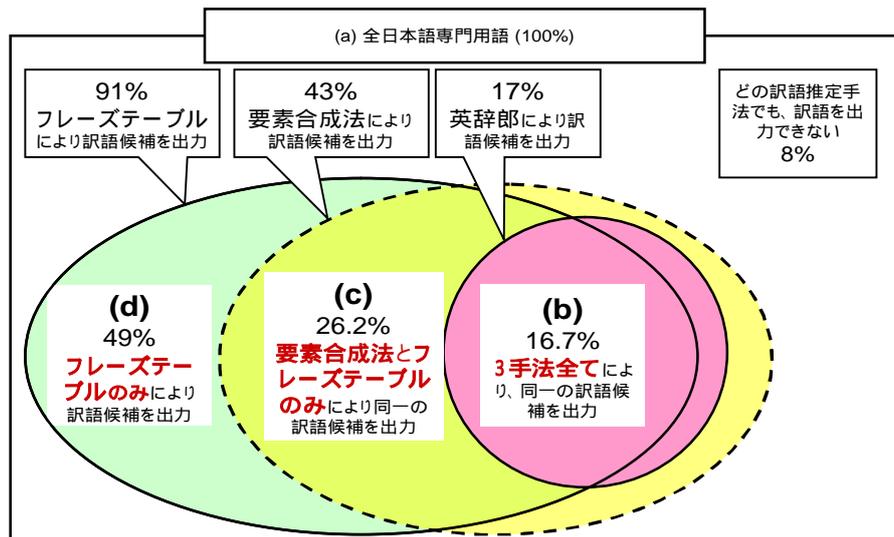


図 8: 英辞郎、要素合成法、フレーズテーブルの訳語候補が英文中に存在した割合

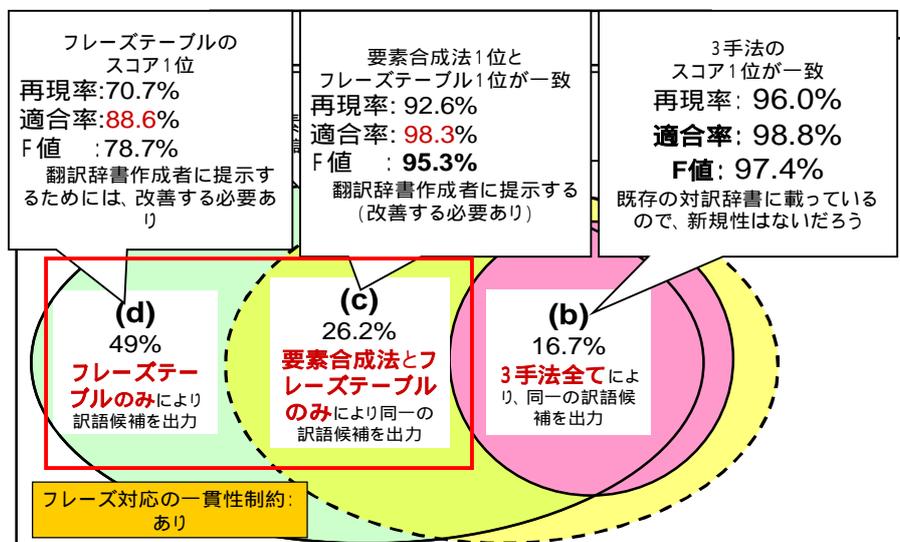


図 9: それぞれの名詞句集合における評価結果

### 3.3.5 SVMによる訳語候補の検証

#### 3.3.5.1 手法

本節では、SVMを用いて、三種類の訳語推定手法によって得られた訳語候補の検証を行う。SVMのツールとして、TinySVM(<http://chasen.org/taku/software/TinySVM/>)を用いた。訓練および評価事例を $\langle t_J, t_E, c \rangle$ と記述する。ここで、 $t_J$ は日本語専門用語、 $t_E$ は少なくとも一つの手法で生成された英語訳語候補、 $c$ は $t_E$ が $t_J$ の正訳か否かを示す。 $t_E$ が正解の場合、 $c=+$ となり、そうでない場合  $c=-$ となる。カーネル関数として、線形カーネルと二次多項式カーネルを比較し、より高い性能が得られた二次多項式カーネルを採用した。評価時においては、事例 $\langle t_J, t_E, c \rangle$ のうち日本語側に日本語専門用語  $x_J$ を持つ事例 $\langle t_J, t_E, c \rangle$ を集めて、クラス  $c$ の判定を行い、十分に信頼できる事例があれば、それを選別する方式で評価を行った。本研究では、SVMの分離平面から、評価事例までの距離を信頼度とし、 $x_J$ を共有する事例の中で分離平面からの距離が最長のものを選択した。

以下の評価では、日本語専門用語集合( $c$ )と、日本語専門用語集合( $d$ )から得られた訓練、評価事例を用意した。これら2種類の集合から得られた訓練、評価事例は別々に扱い、それぞれに対して10分割交差検定を行った。

#### 3.3.5.2 素性

- 単言語素性
  - 日本語専門用語の形態素数
  - 英語訳語候補の単語数
- 二言語素性
  - 英辞郎を利用
    - 要素合成法のスコアと順位
    - 日本語専門用語・英語訳語候補の構成要素の対応が少なくとも一つ英辞郎に存在
  - 対訳文から得られる統計量を利用
    - フレーズテーブルのスコアと順位
    - 相互情報量、 $\phi^2$ 尺度、dice係数、対数尤度比を求めるのに必要な、 $freq(t_J, t_E)$ 、 $freq(t_J, \neg t_E)$ 、 $freq(\neg t_J, t_E)$ の値(直接素性として用いたほうが、よい性能が得られたため)
    - フレーズ対応の一貫性制約

図 10: SVMの素性

SVMにおいて用いた素性を図10に示す。

#### 3.3.5.3 評価結果

フレーズ対応の一貫性制約を用いるにあたって、以下の三つの条件でSVMによる絞りこみを行

った。

- (1) フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外しない。SVMの素性に違反の有無素性を用いない。
- (2) フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外しない。SVMの素性に違反の有無素性を用いる。
- (3) フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外する。SVMの素性に違反の有無素性を用いない。

評価結果を図 11, 12 に示す。訳語候補のうち、信頼性の低いものを識別する手段として、分離平面から評価事例までの距離に下限を設定し、下限に満たない評価事例がある場合はそれらを除いた。下限値の調整の際には、訓練・評価事例以外の事例を用いた。各素性を用いた場合において、適合率が最高となる下限を用いた結果を表枠内の上に、F 値が最高となる下限を用いた結果を下に示す。以上の結果から、SVM を用いた訳語候補の検証により、対訳専門用語の半自動獲得において F 値または適合率を改善することができた。

- 集合c(要素合成法とフレーズテーブルのみにより同一の訳語候補を出力した日本語専門用語)

ベースライン (要素合成法1位フレーズテーブル1位の訳語候補が一致)	SVMによる絞込み (カーネルは2次多項式を使用)	適合率が最高となる下限
再現率: 92.6%(451/487) 適合率: 98.3%(451/459) F値 : 95.3% フレーズ対応の一貫性制約: あり	再現率: 72.9% 適合率: 99.2% F値 : 80.1% フレーズ対応の一貫性制約: SVMの素性として用いる	適合率が0.9%向上! (有意差なし)
	再現率: 98.2% 適合率: 98.2% F値 : 98.2% フレーズ対応の一貫性制約: SVMの素性として用いる	F値が最高となる下限 F値が2.9%向上! (有意差あり)

図 11: SVM の評価結果 : 集合(c)

- 集合d(フレーズテーブルのみにより訳語候補を出力した日本語専門用語)

ベースライン (フレーズテーブルの スコア1位)	SVMによる絞込み (カーネルは2次多項式 を使用)
再現率: 70.7%(619/875) 適合率: 88.6%(619/699) F値 : 78.7%	再現率: 35.1% 適合率: 95.5% F値 : 49.4%
フレーズ対応の一貫性制約: あり	フレーズ対応の一貫性制約: 用いない
	再現率: 90.9% 適合率: 90.9% F値 : 90.9%
	フレーズ対応の一貫性制約: あらかじめ違反するものを除外

適合率が最高となる下限

適合率が5.9%向上! (有意差あり)

F値が最高となる下限

F値が12.2%向上! (有意差あり)

図 12: SVM の評価結果 : 集合(d)

### 3.3.5 おわりに

本研究では、対訳特許文に対して、複数の訳語推定手法を併用し、対訳専門用語獲得の性能を改善する手法を提案した。翻訳知識源が異なる2種類の訳語推定手法を用いることと、SVMを用いて、フレーズテーブルから得た訳語候補を検証し、信頼度の低いものを排除することで、適合率またはF値を改善することができた。

### 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, *Proceedings of the 7th NTCIR Workshop Meeting*, pp. 389--400 (2008).
- [2]. P. Koehn, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177-180, 2007.
- [3]. 外池 昌嗣, 宇津呂 武仁, 佐藤 理史. ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定. *自然言語処理*, Vol. 14, No.2, pp. 33-68, 2007.
- [4]. V. N. Vapnik: *Statistical Learning Theory*, Wiley-Interscience (1998).

### 3.4 対訳辞書および日中漢字の対応関係を用いた特許文書における

#### 専門用語の中日翻訳

東京大学 王 向莉 網川 隆司 辻井 潤一

##### 3.4.1 はじめに

特許文書には多くの分野の専門用語や独特の語彙・言い回しが含まれている。このことにより特許文書の翻訳においては特定の分野の専門的知識や翻訳先言語の特許文の構成に関する知識が必要となる。特許文書等に含まれる専門用語の翻訳には大規模な対訳辞書が不可欠であり、市販の特許文書翻訳に対応した機械翻訳ソフトウェア等でも大規模な辞書が用意されているが、その作成・メンテナンスには大きなコストがかかる。

本研究では、日本語・中国語間の専門用語の翻訳に着目し、現状では対英語と比較して整備が遅れている日中間の専門用語の翻訳を統計的機械翻訳技術によって実現する。翻訳に用いる知識として、既存の中日対訳辞書に加え、日英および中英の対訳辞書から英語を介して日中間の対訳フレーズ対をより多く獲得し、既存の日中対訳辞書のみではカバーできない語彙に対応する。また、同一のものに対して日中間で（字形の変化を含め）同じ漢字を用いる語彙も多いことから、日中漢字の対応関係も導入する。

##### 3.4.2 日中漢字の対応関係

漢字は古来中国から日本に伝えられたものであり、現在でも日本語と中国語の間では、共通または類似の意味を持っている漢字が多く存在する。また、中国あるいは日本で発明された語彙がもう一方の国へ伝わったために、同じ概念を同一の単語で表すこともある。このため、日本語と中国語は語族が異なるにもかかわらず、お互いに単語の意味をある程度推測することが可能である。

日本語と中国語の漢字を統一的に扱うための文字集合として Unicode が利用可能である。Unicode を用いることで、異なる言語間の漢字間の同一表記による対応関係を得ることができる。しかし、中国語と日本語の間には同じ起源をもつ漢字であっても字形の差や意味の差が存在する場合がある。

表 1 に同じ起源の漢字に対して日本語・中国語間で字形や意味に差異があるかどうかを分類した。字形が異なる場合、これらに関連づけるためには中国語の簡体字と日本語の新字体との間の変換テーブルが必要となる。さらに、同一起源の漢字であっても、意味が全く異なる場合や一部の意味しか共有していない場合がある。このため、単純な漢字の対応付けでは不適切な訳を生じさせる可能性もある（青木ら, 2009）。

本研究では、字形の異なりに加え、漢字の意味的な対応関係を獲得するために、日中対訳辞書から統計的機械翻訳（IBM Model）(Brown et al., 1993; Och and Ney, 2003)を利用して日中漢字間の翻訳確率を獲得し、英語を介した専門用語の中日翻訳に適用した。

表 1： 日中漢字の字形・意味の差異の有無による分類

分類	中国語(括弧内は繁体字)	日本語(新字体)	中国語の意味
同一表記で共通の意味を表す	字, 体, 癌, 椅子, 条件	字, 体, 癌, 椅子, 条件	
字形が異なるが共通の意味を表す	姫, 书(書), 广(廣), 汉字(字), 关节(關節)	姫, 書, 広, 漢字, 関節	
同一表記で異なる意味を表す	走, 娘(嬢), 丈夫	走, 娘, 丈夫,	走: 歩く, 行く 娘: 母, 年上の女性 丈夫: 主人, 夫; 成年男子
字形が異なり, 異なる意味を表す	汤(湯), 手紙(手紙), 告诉(告訴)	湯, 手紙, 告訴	汤: スープ 手紙: トイレtpーパー 告诉: 告げる, 知らせる

表 2： 中国語の漢字に対する日本語の漢字への翻訳確率

翻訳確率	娘	翻訳確率	走	翻訳確率	肌	翻訳確率	站
0.152	女	0.163	足	0.812	筋	0.307	駅
0.109	母	0.100	縁	0.084	肌	0.130	立
0.073	婦	0.071	歩	0.020	腱	0.102	所
0.068	親	0.051	行	0.020	箎	0.086	場
...	...	...	...	...	...	...	...
0.039	娘	0.039	走	0.015	膚	0.014	站

表 2 に、得られた漢字間の翻訳確率を示した。最も上の行の中国語の漢字に対して、それぞれの日本語の漢字に訳される確率を左に記した。中国語の「娘」に対しては日本語の「娘」よりも「女」「母」などの漢字に高い確率が割り当てられている。「走」に対しては、中国語の「走进」が日本語の「足踏みする」に訳されたり、中国語の「走廊」が日本語の「縁側」に訳されたりするために「走」一文字の意味である日本語の「歩」「行」よりも高い確率が「足」「縁」に割り当てられている。中国語の「肌」は日本語の「筋」に対応し、日本語の「肌」「皮膚」を意味するには「肤(膚)」を用いる必要がある。最後の例は中国語の「站」が日本語の鉄道の「駅」に対応する例であり、現在の日本語ではほぼ用いられない「站」には小さい確率しか割り当てられていない。以上のように、同一の起源を持つかどうかにかかわらず、漢字間の意味的な対応関係を得ることができた。

### 3.4.3 英語を介した専門用語の中日翻訳

本研究では、ピボット言語を用いたフレーズベース統計的機械翻訳(内山・井佐原, 2007; 網川ら, 2008)を用いて、特許文書に現れる専門用語の中日翻訳を行う。このための訓練データとして、中日対訳辞書に加え、中英および日英対訳辞書を用いる。各辞書の用語に対してそれ

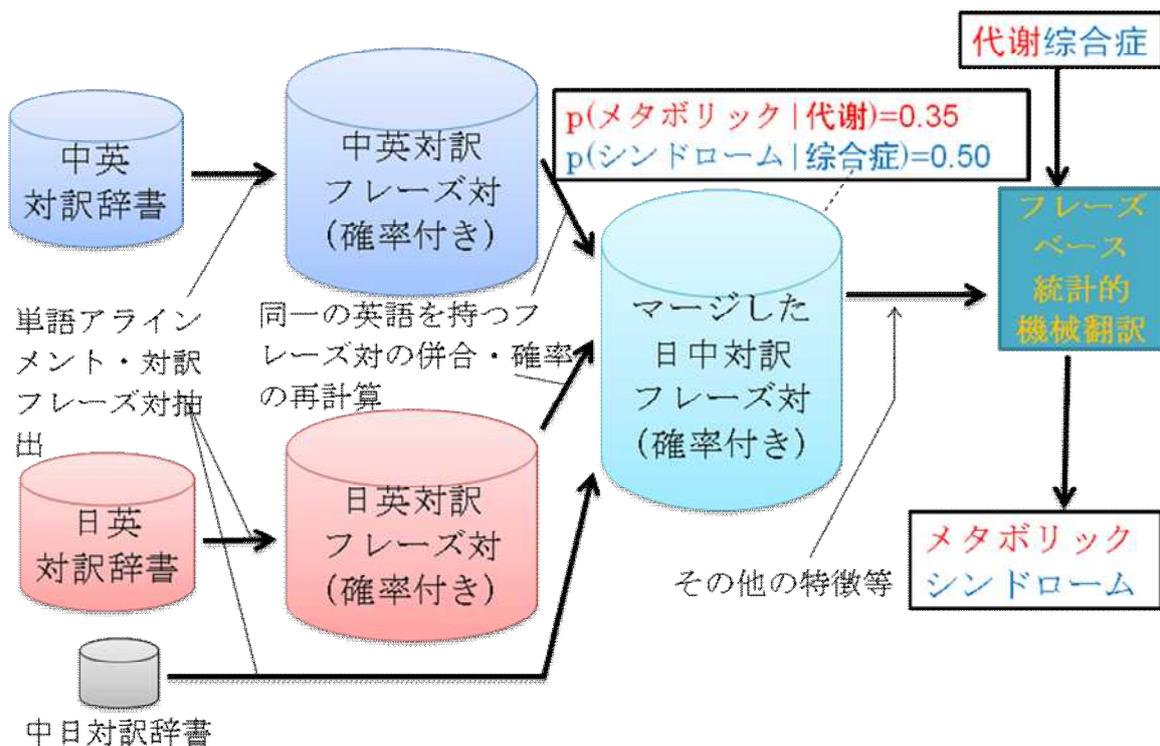


図 1： 英語を介したフレーズベース統計的機械翻訳による専門用語の中日翻訳

それぞれの言語の形態素解析を行って単語分割を行い，各対訳辞書を並列コーパスとみなしてフレーズベース統計的機械翻訳 (Och and Ney, 2003) の枠組みを用いて各言語対のフレーズ対とその翻訳確率を獲得する。中英および日英の対訳フレーズ対からは，以下の式により英語を介して日中対訳フレーズ対を新たに作成し，それを中日の対訳フレーズ対とマージする。

$$p'(j|c) = \sum_e p(j|e)p(e|c), \quad p'(c|j) = \sum_e p(c|e)p(e|j).$$

以上の枠組みを図 1 に示す。

フレーズベース統計的機械翻訳では，対数線形モデルを用いて，以下の式を用いて中国語の専門用語  $c$  を日本語  $j$  に翻訳する。

$$j = \operatorname{argmax}_j \sum_{m=1}^M \lambda_m h_m(j, c).$$

ただし， $h_m(j, c)$  は  $j, c$  間の対訳らしさを表す素性関数であり， $\lambda_m$  は各関数の重みである。

素性関数としては，以下のものを用いた。

- 中日対訳辞書から求めたフレーズ翻訳確率（両方向）
- 中英・日英対訳辞書からマージにより得られたフレーズ翻訳確率（両方向）
- 日本語言語モデル
- フレーズの並び替えスコア (Koehn et al., 2003)
- 日中漢字の対応関係に基づくスコア

日中漢字の対応関係に基づくスコアは，前節で得られた漢字の翻訳確率  $p(j|c)$  を用いてすべてのフレーズ対に対して計算を行った。そのアルゴリズムを図 2 に示す。

入力：中国語の文字列  $c = c_1 c_2 \dots c_{|c|}$ ，日本語の文字列  $j = j_1 j_2 \dots j_{|j|}$ ，ビーム幅  $d$ ，閾値  $\varepsilon$

出力： $c, j$ 間の日中漢字の対応関係に基づくスコア

1.  $c$  または  $j$  が空文字列のときは，1 を返す。
2. 中国語の最初の文字  $c_1$  に対して，漢字の翻訳確率が高い日本語の文字を順に  $d$  個まで求める。ただし，漢字の翻訳確率が閾値  $\varepsilon$  以下のものは除く。
3. 求めた日本語の各文字  $j^{(1)}, j^{(2)}, \dots, j^{(d)}$  について， $j$  からそれを取り除いた文字列をそれぞれ  $j^{(1)}, j^{(2)}, \dots, j^{(d)}$  とする（すべての文字の翻訳確率が閾値  $\varepsilon$  以下のときは， $j^{(i)} = j$  とする）。また， $c' = c_2 \dots c_{|c|}$  とする。 $(c', j^{(1)}), (c', j^{(2)}), \dots, (c', j^{(d)})$  のそれぞれの文字列組について，それぞれ再帰的に本アルゴリズムを適用して得られたスコアを  $s_1, \dots, s_d$  とする。漢字の翻訳確率と再帰的に得られたスコアの積の最大値

$$\max_{i=1, \dots, d} p(j^{(i)} | c_1) \times s_i$$

図 2：日中漢字の対応関係に基づくスコアを求めるアルゴリズム

#### 3.4.4 実験

本手法による専門用語の中日翻訳が特許文書に含まれる専門用語に対してどの程度有効かを調べるために以下の実験を行った。

機械翻訳を行うためのフレーズ対は，以下の対訳辞書から抽出した。

**中英** 万方数据 (Wanfang Data) 英汉-汉英科技大词库 (525,259 訳語対)

(<http://qh.library.hb.cn:85/kjxx/yhcb.htm>)

**日英** JST 機械翻訳辞書 (527,206 訳語対) (<http://pr.jst.go.jp/others/tape.html>)

**日中** EDR 日中英辞書 (596,967 訳語対) (EDR 電子化辞書 ([http://www2.nict.go.jp](http://www2.nict.go.jp/r/r312/EDR/J_index.html)

[http://www2.nict.go.jp/r/r312/EDR/J\\_index.html](http://www2.nict.go.jp/r/r312/EDR/J_index.html)) の日英対訳辞書の一部に中国語訳を付加したもの)

いずれの辞書も専門用語を含み，また日中辞書には一般語も含まれている。形態素解析には JUMAN (<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>) および *cjma* (Nakagawa and Uchimoto, 2007) を用いた。日本語の言語モデルとして，Web 日本語 N グラム (工藤・賀沢, 2007) から構築した 3-グラム言語モデルを用いた。実験にはフレーズベース統計的機械翻訳のツールキットとして Moses (Koehn et al., 2007) を用いた。

翻訳に用いるテストデータとしては，(財)日本特許情報機構 (Japio) から提供された中日特許文献対訳アラインメントファイルを用いた。これは情報科学分野からなる 232 対訳文アラインメントファイルである。

#### 実験結果と考察

表 3 から表 6 に実験の評価を示す。特許文献アラインメントファイルから 210 の中日用語対が得られた。そのうち，完全に一致する用語対は 27 語、ほぼ一致する用語対は 7 語、意味的に一致する用語対は 86 語あり、あわせて全体の 5 割以上を占める。意味的に一致する用語の

うち実際によく用いられる用語は 50 語あり、意味的に一致する用語のうち半分以上を占める。また、その中で一方がカタカナによる表記になっているため完全一致にならない用語は 21 語あり、全体の 42%を占める。一方、実際にはほぼあるいは全く用いられない用語は 36 語あり、そのうち 14 語はカタカナによって完全一致にならない用語であり、全体のほぼ 4 割を占める。

表 3： 実験結果の評価（全体）

専門用語	完全一致する用語	ほぼ一致する用語	意味的に一致する用語	一部が間違いである用語	全部が間違いである用語
210	27	7	86	58	32
100%	12.86%	3.33%	40.95%	27.62%	15.24%

表 4： 実験結果の評価（意味的に一致する用語のうち用いられる用語・用いられない用語）

意味的に一致する用語	よく用いられる用語	用いられない用語
86	50	36
100%	58.14%	41.86%

表 5： 実験結果の評価（表 4 の用いられる用語のうちカタカナによる不一致が原因のもの）

よく用いられる用語	カタカナで不一致である用語	他の用語
50	21	29
100%	42.00%	58.00%

表 6： 実験結果の評価（表 4 の用いられない用語のうちカタカナによる不一致が原因のもの）

用いられない用語	カタカナで不一致である用語	他の用語
36	14	22
100%	38.89%	61.11%

### 3.4.5 おわりに

本研究では、日中漢字の対応関係を中日対訳辞書から獲得し、英語を介した中日翻訳を行うことで特許文書に含まれる専門用語を中国語から日本語へ翻訳することを試みた。テストセットに含まれる 210 語の用語対について、約 13%については抽出した用語対と完全に一致する翻訳結果が得られ、ほぼ一致する用語と意味的に一致するもののうち実際によく用いられる用語を加えて 84 語（40%）について意味的に正しい訳語が得られた。

今後の課題としては、フレーズ対抽出および言語モデル構築に特許文書を用いることで精度の向上を図るほか、特許文書に多く含まれているカタカナ語に対処するために音訳に関する素性を組み込むことが考えられる。

[参考文献]

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263-311, 1993.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48-54, 2003.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of ACL, demo. session*, pages 177-180, 2007.

Tetsuji Nakagawa and Kiyotaka Uchimoto. Hybrid approach to word segmentation and POS tagging. In *Companion Volume to the Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 217-220, 2007.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19-51, 2003.

青木 雅子, 蒋 甥, 井上 聡子, 菊池 俊一. 日中科学技術対訳コーパス作成における訳し分け用語・同形異義語の収集・分析. *言語処理学会第 15 回年次大会発表論文集*, pages 180-183, 2009.

内山 将夫, 井佐原 均. 統計的機械翻訳におけるピボット翻訳の比較. *言語処理学会第 13 回年次大会発表論文集*, pages 187-190, 2007.

工藤 拓, 賀沢 秀人. *Web 日本語 N グラム* 第 1 版, 2007.

綱川 隆司, 岡崎 直観, 辻井 潤一. 日英・中英対訳辞書からの日中対訳辞書の構築. *言語処理学会第 14 回年次大会発表論文集*, pages 464-467, 2008.

## 4 . 句の依存関係確率モデルを用いた統計的対訳文アライメント

京都大学大学院情報学研究科 中澤 敏明  
黒橋 禎夫

### 4.1.1 はじめに

日本語と英語などのように、言語構造の大きく異なる言語の対訳文を精度良くアライメントする際に重要なポイントは二つある。一つは構造の違いを吸収するために構文解析や依存構造解析を利用することであり、もう一つは多対多のアライメントを生成できることである。既存の対訳文アライメント手法の多くは、文を単純な単語列として見た単語アライメントであり、大きな構造の違いは扱えない[1, 4]。またこれらの手法によるアライメント結果にさらに構文情報を追加する手法[3, 6]も提案されているが、やはり不十分である。これに対してアライメントの最初から構文情報を利用する手法もいくつか提案されているが、生成可能なアライメントが1対1のみであるという制約があったり、ヒューリスティックなルールに基づいて多対多対応を生成するなど、ロバスト性に欠ける[2, 8]。本稿では両言語を依存構造解析し、木構造上での語句の依存関係をモデル化した新たな句アライメント手法を提案する。このモデルでは単語列アライメントで扱うのが困難な距離の大きな語順変化にも対応することができるし、モデル内で自然に多対多対応を推定することができる。

### 4.1.2 依存関係確率モデル

本章では、提案するモデルについて説明する。

#### 4.1.2.1 アライメントの単位

以後の説明では原言語に日本語、目的言語に英語を仮定する。ただし、提案手法は言語対によらない汎用的なものであることに注意されたい。まず対訳文を両言語とも依存構造解析する。日本語文に対してはJUMAN およびKNP を用い、英語文に対してはCharniak のパーサを用いて句構造に変換し、これにフレーズのhead を定義するルールを適用することにより依存構造に変換する。これらの処理により、文は1つの内容語と0個以上の機能語とをひとまとまりとした句をノードとする依存構造木に変換される。

次に機能語を、内容語同士をつなぐものとする。つまり機能語は、内容語とその係り先の内容語との間に位置しており、それらをつなぐ枝として働くと考える。なお機能語がない句については、その句の係り受け情報を機能語の代わりとする。係り受け情報としては各パーサによって定義される係り受けタイプと係る方向を利用する。係り受けタイプは日本語では”文節内”や”連用”、”連体”などが、英語では”NP”、”VP”、”PP”などがある。係る方向はその句が係り先の前から係るか、後ろから係るかの情報である。これは英語で同士の前から係る名詞は主語(日本語での”ガ格”)である可能性が高く、後ろから係る名詞は目的語(”ヲ格”)である可能性が高いといった現象を扱うのに有効である。図1に依存構造木の例を示す。提案モデルではこ

母集団に分散があると考える場合、その母集団による効果を変量効果という

When the dispersion in a population is considered to exist, the effect due to the population is called a variable effect

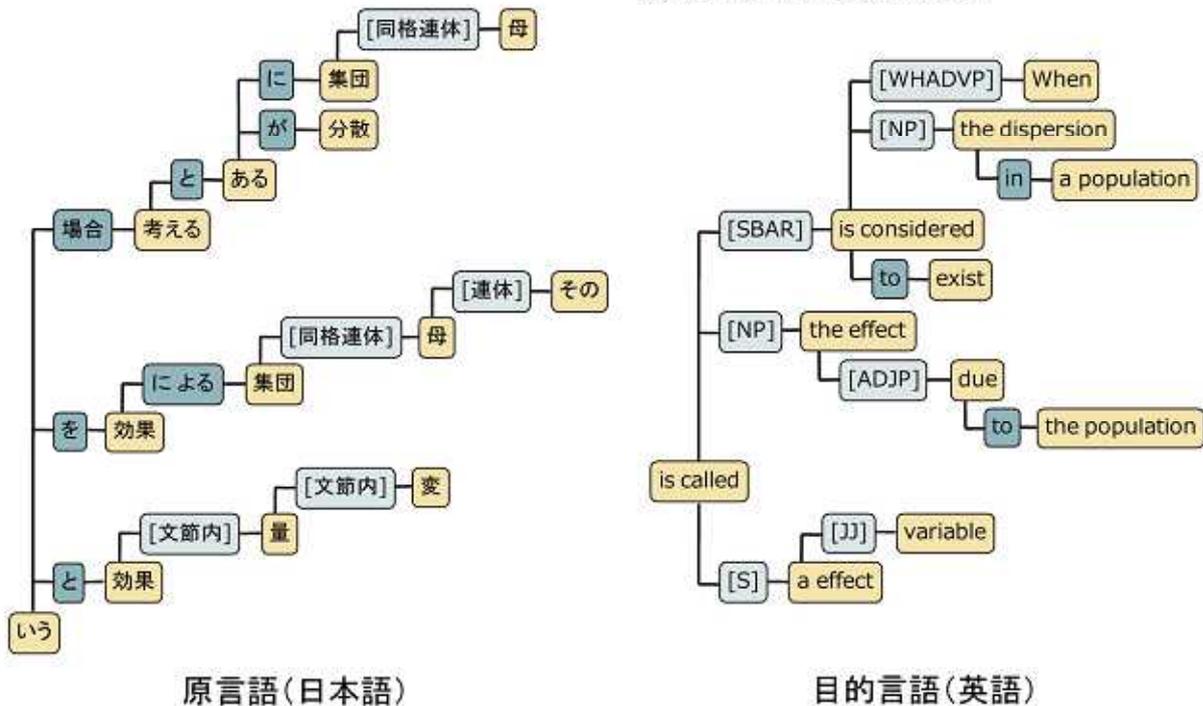


図 1: 依存構造木の例

これら内容語ノード、付属語ノードそれぞれをアライメントの最小単位として扱う。なお、モデル推定時には連続する複数のノードのかたまりもアライメントの単位として考慮することにより、多対多対応を実現する(4.1.3.2.3 章参照)。

#### 4.1.2.2 モデルの定義

提案するモデルでは、IBM Model [1]のような方向性のある確率モデルを双方向同時に利用する。そこでまず方向性のある確率モデルを定義する。IBM Model では与えられた原言語文  $f$  と目的言語文  $e$  に対して最も良いアラインメント  $\hat{a}$  を

$$\hat{a} = \arg \max p(f | e, a) \cdot p(a | e) \quad (1)$$

により求める。ここで  $p(f | e, a)$  は lexicon probability、 $p(a | e)$  は alignment probability と呼ばれている。提案モデルでは木構造を利用しているため、以下のような確率を考える。与えられた原言語の依存構造木  $T_f$  と目的言語の依存構造木  $T_e$  に対して、IBM Model と同様に以下のように書ける。

$$\hat{a} = \arg \max p(T_f | T_e, a) \cdot p(a | T_e) \quad (2)$$

さらに  $T_f$  が  $J$  個のノード  $f_1, \dots, f_J$  から、 $T_e$  が  $I$  個のノード  $e_1, \dots, e_I$  から構成されているとし、各ノードの係り先(親)のノードを、添字にダッシュを付けることにより表し(例えばノード  $f_j$  の係り先のノードは  $f_j$  )、それぞれの確率を以下のように定義する。

$$p(T_f | T_e, a) = \prod_{j=1}^J p(f_j | e_{a_j}) \quad (3)$$

$$p(a | T_e) = \prod_{j=1}^J p(\text{rel}(e_{a_j}, e_{a_j'}) | \text{rel}(f_j, f_{j'})) \quad (4)$$

$a_j$  は  $f_j$  の対応先のノード番号であり、 $e_{a_j}$  は対応ノードを表す。

$p(f_j | e_{a_j})$  単語翻訳確率であり、IBM Model 1と同様の定義である。 $\text{rel}(f_j, f_{j'})$  は  $f_j$  からみた  $f_{j'}$  の依存関係であり、 $\text{rel}(e_{a_j}, e_{a_j'})$  も同様である。 $p(\text{rel}(e_{a_j}, e_{a_j'}) | \text{rel}(f_j, f_{j'}))$  は依存関係確率であり、原言語側で直接の親子関係になっている二つのノード ( $f_j$  と  $f_{j'}$ ) の対応する目的言語側のノード ( $e_{a_j}$  と  $e_{a_j'}$ ) の依存関係がどうなっているかを表す確率である。提案モデルでは対訳文のアライメント確率  $p(a | f, e)$  を lexicon probability と alignment probability を双方向分すべて掛け合わせたものとして以下のように定義し、この確率が最大となるアライメントを最も良いアライメントとする。

$$p(a | f, e) = p(T_f | T_e, a) \cdot p(a | T_e) \cdot p(T_e | T_f, a) \cdot p(a | T_f) \quad (5)$$

### 4.1.3 トレーニング

モデルのトレーニングではまず単語翻訳確率のみを推定(Step 1) し、これを初期値として依存関係確率と単語翻訳確率を同時に推定(Step 2) する。推定にはEM アルゴリズムを用いる。

#### 4.1.3.1 Step 1

表 1: カテゴリ確率の有無による精度比較

	適合率	再現率	F 値
カテゴリ確率なし	58.25	42.77	49.33
カテゴリ確率あり	<b>76.52</b>	<b>50.23</b>	<b>60.65</b>

基本的にはIBM Model 1と全く同じ方法で、各方向独立に推定を行う。ただし提案モデルでは内容語、機能語が内容語、機能語、NULL のどれに対応するかといった確率(カテゴリ確率) も利用する点が異なる。予備実験ではこの確率の導入により、Step 1でのアライメントの精度が表1 のように向上することがわかった。カテゴリ確率は出現頻度が極端に少ない単語を正しくアライメントするのに有効である。これは出現頻度が少ない語はほぼ確実に内容語であるからである。なおカテゴリ確率はStep 1でのみ利用し、Step 2では利用しない。

Step 1では対応の単位は各ノード単体のみであり、複数ノードのかたまりは考慮しない。つまり、この時点では1対1対応のみしか扱えない。多対多対応はStep 2から考慮し、多ノード候補を動的に作り出すことにより実現する。これはStep 1の段階で多ノード候補全てを考慮すると、アライメント候補数が爆発し、扱えなくなるためである。

#### 4.1.3.2 Step 2

単語翻訳確率と依存関係確率を同時に推定する。Step 2では双方向の確率を両方使い、一つの

アライメント結果を得る。しかしStep 1とは違い近似なしでモデルを完全に推定することはできないため、単語翻訳確率のみを用いて初期アライメントを生成し、山登り法により依存関係確率を考慮した $n$ -best アライメントを探索する。

依存関係確率は、原言語側で直接の親子関係にあるノード ( $f_j$  と  $f_j$ ) の対応する目的言語側のノード ( $e_{aj}$  と  $e_{aj}$ ) の関係 ( $rel(e_{aj}, e_{aj})$ ) に対して与えられる確率である。 $rel(P_2, P_1)$  はあるノード ( $P_1$ ) から別のあるノード ( $P_2$ ) までの経路で表し、離れているノード数分だけ以下の表記を並べることにより表現する。

- $P_2$  が  $P_1$  の子: c
- $P_2$  が  $P_1$  の親: p

例えば図1 において”いう”から見た”場合”の依存関係は”c”、“考える”は”c;c”、“ある”は”c;c;c;c”などとなる。

#### 4.1.3.2.1 初期アライメントの生成

原言語側、目的言語側の全てのノード及び多ノード候補間の単語翻訳確率を計算し、単語翻訳確率の最も高いペアから順に採用する。ただし、すべてのノードは1度しかアライメントされない。つまり対応の採用は排他的に行われるが、そもそも多ノード候補を考慮しているため自然と多対多対応が実現される。多ノード候補の生成については3.2.3章で述べる。なお、単語翻訳確率は両方向の単語翻訳確率を単純に掛け合わせるにより計算される。

#### 4.1.3.2.2 山登り法

初期アライメントの状態から、依存関係確率を考慮しながらアライメントを修正していき、徐々に確率の高いアライメントを探索していく。修正手段としては以下の4種類を考える。

**対応の入れ替え:** 任意の2つの対応に注目し、それらの対応を入れ替える。

**大きな対応への拡大:** 任意の1つの対応に注目し、その原言語側または目的言語側いずれかの対応を、親または子方向に1ノード分だけ拡大する。拡大先のノードにすでに対応が存在する場合、その対応は棄却し、反対言語側のノードはNULL対応となる。

**対応の追加:** 対応がないノード間に新たに対応を追加する。

**対応の棄却:** すでにある対応を棄却する。

修正後のアライメント確率が修正前よりも高くなる場合にのみ修正を実行し、修正された状態から再度修正を行っていく。確率が高くなる修正箇所がなくなるまで修正を繰り返す。最終的に得られたアライメントが、最も確率の高いアライメントとなる。

#### 4.1.3.2.3 多ノード候補の生成

獲得された最も確率の高いアライメントの中に、NULLに対応付けされたノードがあった場合、それらを親、または子のNULL対応でないノードに併合したものを新たな多ノード候補として生成し、次のイタレーションから探索に入れる。この新たな多ノード候補はその対訳文から1回出現したものと数える。





- [2] Colin Cherry and Dekang Lin. A probability mode to improve word alignment. In *Proceedings of the 41<sup>st</sup> Annual Meeting of the Association of Computational Linguistics*, pages 88-95, 2003.
- [3] Brooke Cowan, Ivona Kučerová, and Michael Collins. A discriminative model for tree-to-tree translation. In *Proceedings of the 2006 Conference on EMNLP*, pages 232-241, Sydney, Australia, July 2006. Association for Computational Linguistics.
- [4] Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT-NAACL 2003: Main Proceedings*, pages 127-133, 2003.
- [5] Toshiaki Nakazawa and Sadao Kurohashi. Linguistically-motivated tree-based probabilistic phrase alignment. In *In Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA2008)*, 2008.
- [6] Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271-279, 2005.
- [7] Masao Utiyama and Hitoshi Isahara. A japanese-english patent parallel corpus. In *MT summit XI*, pages 475-482, 2007.
- [8] Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of 39th Annual Meeting of the ACL*, pages 523-530, 2001.

## 5 . 特許文への構文解析適応

山形大学 横山 晶一

### 5.1.1 はじめに

特許文が、詳細の説明や要件において、日本語の文には余り見られない長大な一文（200 字以上）になり、複雑な並列構造や係り受け構造を含んで、正確な構造を解析するのが困難であることは、すでに昨年度やそれ以前の報告書でも繰り返し述べてきた[1]。

昨年度までは、主として並列構造について調査し、出力された係り受け構造を修正するシステムを構築した。今年度も並列構造についてある程度の調査は行ったが、特に特筆すべき成果が得られなかった。

本稿では、上記とは異なる立場からのアプローチについて述べる。これまでは、上記にもあるように、構文解析器（パーザ）の出力を自動修正するシステムを構築してきたが、ここでは、構文解析器自体を、特許に適するように学習させた結果について述べる。

具体的には、Tsujii, Sagae による Ensemble 法[2]を用いて特許文の学習データを作り、構文解析器をそのデータで学習させて、特許文に適応させるという手法をとる。学習データ作成に使用した構文解析器は、KNP[3], 南瓜[4], MST [5], malt [6]の各パーザで、KNP 以外のパーザを学習に用いて適応させた[7-9]。評価は、人手で作成した3分野の正解データと、これらのパーザの結果とを比較する形で行った。その結果、ある程度特許文に特化した改善が見られた。その結果について以下に報告する。

### 5.1.2 Ensemble 法

構文解析器の中には、学習によって、ある分野に特化した出力が得られるように変化できるものがある。その際、問題になるのは学習するための文をどのように作成するかということである。Tsujii, Sagae は、学習するための訓練データを作成する方法として、Ensemble 法を提案した[2]。これは、異なるアルゴリズムに基づく構文解析器を複数用意し、それらに文集合を解析させ、その結果を照らし合わせて、多数決的方法によって、学習のための構文解析情報を出力するという方法である。複数の構文解析器が同一の解析結果を出力するということは、異なる構文規則を用いても、出力が同じ構造になっていることを意味するので、正解データである可能性が高いと考えられる。従って、このデータを、構文解析器を訓練するデータとして使用することが可能である。

### 5.1.3 特許領域への適応

特許領域へ構文解析器を特化するために、上にも述べたように、いくつかの構文解析器を用いた。また、Ensemble 法による訓練データを作成するための対照パーザとして、KNP [3]を用いた。以下に詳細を述べる。

# 適応の流れ

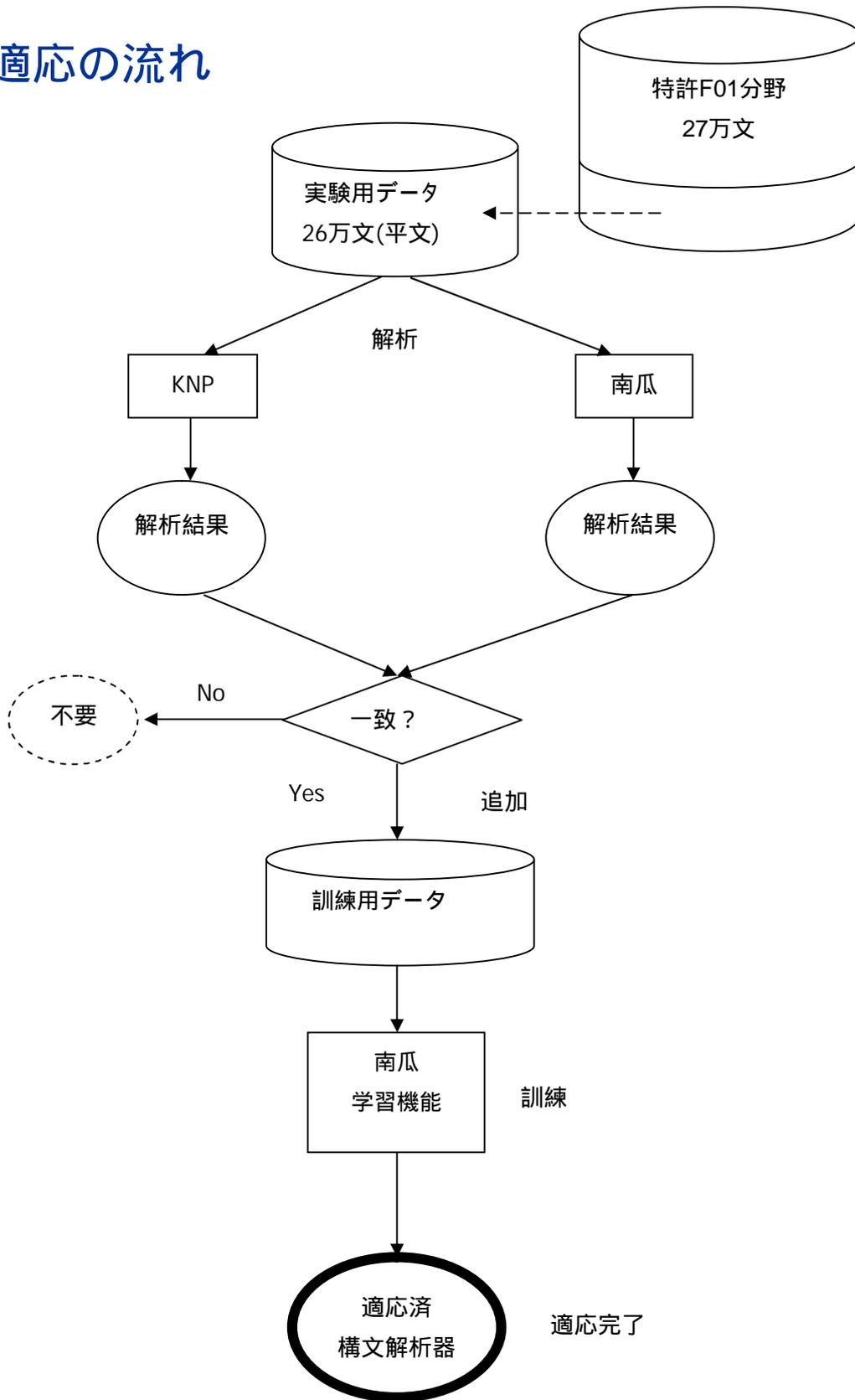


図1 Ensembleによる適応済構文解析器構築の図[7]

### 5.1.3.1 KNP と南瓜による南瓜の適応

図 1 に Ensemble 法による南瓜[4]の適応[7]の概念図を示す。解析に用いる特許文は、AAMT/Japio 研究会データベース[10]の 2004 年度の F01 (機械) 分野 267,405 文のうち、KNP で解析できる、1 文 200 文字以下の 257,818 文である。

訓練用データの作成手順は、図 1 で示すように、次のステップからなる。

- (1) AAMT/Japio 研究会データベース F01 分野の文を用意する
- (2) KNP, 南瓜にそれぞれ解析用データを入力して解析させる
- (3) 出力は、表記のしかたに違いがあるので、同じ表記法になるよう書き換える
- (4) 2 つの解析結果を比較し、完全に一致するものを訓練用データとして保存する

上記の文のうち、7.3%に当たる 18,899 文が一致した。これを訓練データとして用いる。

このデータを用いて学習を行った南瓜の評価を行う。評価には、上記データベースの中から、訓練用データに含まれていない 1,000 文を KNP で解析し、人手で確認、修正を行ったものを評価用データとして用いる。適応前の南瓜と適応後の南瓜の解析結果を評価用データと比較する。比較のしかたは、1 文に含まれる文節すべてに対して、文節の係り先と区切り位置が一致しているかどうかを見るという方法で行った。その結果を表 1 に示す。

表 1 適応済構文解析器と既存「南瓜」の正解データ比較結果の内訳

	既存の南瓜(文)	適応済構文解析器(文)
完全一致	5	6
文節数一致,係り受け不一致	304	303
文節数不一致	691	691
計	1,000	1,000

表から分かるように、この手法で改善された文は、1 文のみであった。その文を以下に示す。左の番号は、係り先の文節で、0 から始まり、これが 1 に係ることなどを示す。

# S-ID:P\_F01\_0207874

ストイキまたはリッチに保持する時間は、数秒ないし数分間で十分である。

\* 1 ストイキまたは

\* 2 リッチに ←

\* 3 保持する ←

\* 6 時間は、

\* 5 数秒ないし

\* 6 数分間で

\* -1 十分である。

適応済構文解析器の係り先

既存の南瓜の係り先

この文の第0文節「ストイキまたは」は、学習前の南瓜では、第2文節「保持する」に係っていたが、適応後には、並列構造である「リッチに」に係っている。たった1文ではあるが、特許文の課題の一つである並列構造[1]を改善したというところには、見るべきものがある。

### 5.1.3.2 KNP, MST Parser, malt Parser による Ensemble 法

同様の実験を、別のパーザを用いて行う[8,9]。ここでは、言語に依存しないパーザである、MST Parser [5]と、malt Parser [6]を用いる。各々について簡単に解説する。

MST Parser は、単語の間にスコアをつけて結んだグラフの中から、スコアが最大となる Spanning Tree (閉路を持たないグラフ) を求めることで構文解析を行う。このパーザでは、大量の正しい構文解析データから単語間のスコア付けにおける重みを学習することで、その言語に適応した構文解析器となる。

malt Parser は、transition-based のパーザで、これまでの構文解析結果と、現在の状態が与えられたとき、次のパーザの動作として最も適切なものを選択するか、これらの選択の系列全体を最適化することによって動作する。アルゴリズムは MST とは全く異なっている。

これら2つのパーザは、大量の構文解析済みのテキストコーパスから学習を行うことで、その言語に適応したパーザとなる。そのために、次のような手順を用いる。

- (1) 京大コーパス[11] (KNP で解析されている) 約4万文を用いて、MST Parser, malt Parser を日本語に適応させる。これを default のパーザとする
- (2) 上記の F01 分野の文 (約25万文) を(1)の default パーザと KNP で解析する
- (3) MST Parser + KNP, MST Parser + malt Parser の2通りの組み合わせで一致する結果を学習する
- (4) (3)の結果を用いて MST Parser を学習させる
- (5) (4)の結果、作成された解析モデルを用いて、G06F (電氣的デジタルデータ処理)、C12N (生物)、F01 (機械) の3分野各1,000文 (計3,000文) について評価する

表2 評価結果[9]

		F01 (機械)	C12N (生物)	G06F (情報)
default	Accuracy	0.9039	0.9176	0.9084
	C.C.	0.1131	0.1781	0.1381
MST + KNP	Accuracy	0.9236	0.9273	0.9219
	C.C.	0.1931	0.2342	0.2162
MST + malt	Accuracy	0.9062	0.9142	0.9088
	C.C.	0.1361	0.1891	0.1431
MST のみ	Accuracy	0.9047	0.9176	0.9081
	C.C.	0.1161	0.1771	0.1371

評価は、次の2つの指標を用いて行う。

$Accuracy = (\text{正解データと一致した形態素数}) / (\text{全形態素数})$

$Complete\ Correctness\ (C.C.) = (\text{正解データと文全体において一致した形態素数}) / (\text{全形態素数})$

表2に評価結果を示す。表で「MSTのみ」と書かれているのは、defaultのパーザによる解析結果からランダムに25,000文を抽出して学習を行ったものである。この中には、MST + KNP, MST + maltで一致したものは含まれていない。この結果は、Ensemble法によって行ったものより低くなっている。

表2では、MST + KNPが最も高い数値を示している。また、いずれの分野においても精度が向上している。これは、特許領域では、別の分野であってもある程度適合できることを示唆している。

この出力結果をさらに向上させるために、格フレーム[12]を用いた改善の可能性を探った。MST Parserでは、辞書の情報は使用していない。従って、格フレームを利用して検索を行えば、結果の改善が期待できる。G06Fの1,000文からランダムに500文を抜き出してエラー分析を行ったところ、名詞 動詞間に誤りがある(すなわち格フレームに誤りがある)文は12文であった。このうち、格フレームを用いて修正可能なのは7文、格フレームにデータが存在しないものが4文、格フレームを用いて修正すると誤りになるものは1文であった。MSTで利用できる文が約25,000文あることから、この数字をそのままあてはめると、600程度の文について修正できる可能性がある。

#### 5.1.4 まとめ

構文解析器を特許文対応にする、パーザ適応という手段を用いて構文解析器を学習、特化させたところ、ある程度の改善が見られた。ただ、この手段を用いて評価用データを作ると、計算時間の割に、少数のデータ(7%程度)しか得られず、やや効率が悪いという欠点があった。

今後は、学習のしかたや評価用データの作り方についてさらに検討していきたい。

また、格フレームを用いて適応パーザをさらに改善する手法も試み、ある程度の成果を得た。今後はこの手法がどの程度有効かをさらに追究する予定である。

#### 謝辞

本研究の一部は、科学技術研究費(基盤研究(C))「動的シソーラスを用いた特許文の解析システム」(課題番号18500102)に基づき行われた。

#### 参考文献

- [1] 横山晶一・橋本力：並列接続詞を含む特許文の係り受け解析修正システムと、特許文の連体修飾節の構造解析、平成19年度AAMT/Japio特許翻訳研究会報告書(2007) pp.68-76
- [2] Jun'ichi Tsujii, Kenji Sagae: Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles, EMNLP-CoNLL (2007) pp.1044-1050

- [3] KNP (<http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/KNP.html>) 京都大学黒橋研究室
- [4] 南瓜 (<http://chasen.org/%7Etaku/software/cabocha/>) 奈良先端科学技術大学院大学松本研究室
- [5] R. McDonald, F. Pereira, K. Ribarov, and J. Hajic: Non-projective Dependency Parsing using Spanning Tree Algorithms, Proceedings of HLT/EMNLP (2005)
- [6] malt Parser (<http://w3.msi.vxu.se/~jha/maltparser/>)
- [7] 後藤広樹: Ensemble 法を用いた特許ドメインへの構文解析適応、山形大学工学部卒業論文 (2009)
- [8] 吉田節行、横山晶一: パーサ適応と格フレームによる特許文構文解析精度の向上、情報処理学会第 71 回全国大会 2S-6 (2009)
- [9] 吉田節行: パーサ適応と格フレームによる特許文構文解析精度の向上、山形大学大学院理工学研究科修士学位論文 (2009)
- [10] AAMT/Japio 特許翻訳研究会データベース (2005)
- [11] 京大コーパス (<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html>)
- [12] 河原大輔、黒橋禎夫: 高性能計算環境を用いた Web からの大規模格フレーム構築、情報処理学会自然言語処理研究会 NL-171 (2006) pp.67-73

禁 無 断 転 載

平成20年度AAMT/Japio特許翻訳研究会報告書  
(機械翻訳及び辞書構築に関する研究)

発 行 日 平成21年3月

発 行 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 ナビックス