平成 21 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
及び
海外調査

平成 22 年 3 月

一般財団法人 日本特許情報機構

(正誤表 2010/5/19 反映版)

目　次

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	山梨英和大学教授
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
〃	越前谷　博	北海学園大学助教
〃	綱川　隆司	静岡大学学術研究員
〃	安田　圭志	(独)情報通信研究機構
〃	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	元多摩美術大学講師
〃	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	〃
〃	大塩　只明	〃
〃	塙　　金治	〃
〃	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

AAMT（アジア太平洋機械翻訳協会）の特許翻訳研究会は、Japio（日本特許情報機構）からの委託をうけ、本年度も 8 回にわたる委員会を開催し、活発な研究調査活動を行ってきた。本報告書は、この委員会での活動報告である。

本委員会が始まった平成 15 年以降、年ごとに特許情報の有効活用、その機械翻訳に対する言語処理技術への関心が急速に高まってきている。本年度は、本委員会と直接の関係はないが、Japio 提供の英語・日本語特許文書を使った機械翻訳システムの Shared Task が、国立情報学研究所（NII）が主催する NTCIR において組織され、多くの研究チームが特許の機械翻訳システムを実際に構築、その性能を比較することになった。このことが、本年度、より一層特許翻訳への関心を高めることとなった。本委員会でも、特許の機械翻訳に関する技術の中で、システムの性能をどのように評価するかは大きな関心となっていることから、NTCIR の特許翻訳タスクを組織したグループと合同の会合をもち（7 月 30 日）、意見の交換を行った。NTCIR 側は、今後も機械翻訳、とくに特許の機械翻訳を取り上げてタスク設定していく意向であり、その性能評価に大きな課題が残されていることは認識しており、本委員会との交流を継続していくこととなった。

本年度は、本委員会にとって大きな催しがいくつもあった年であった。まず、カナダ・オタワで開催された MT Summit XII に合わせて、特許の機械翻訳に関する国際ワークショップを開催した。このワークショップは今回が 3 回目であり、過去 2 回の開催と同様、その組織は本委員会が主導権をとることとなった。参加者も 40 名を超え、過去のワークショップを上回るものとなった。さらに、本年度最初のこころみとして、本委員会活動を外部に向けて発信するためのシンポジウムを開催し（11 月 27 日、於総評会館）、63 名 の参加を得た。このシンポジウムでは、特許における中国語の問題を取り上げたが、世界の情報流通が多言語化する中、特許の中国語への翻訳、あるいは、中国での特許を取り巻く状況への関心が高く、これがシンポジウムに予想以上の参加者を得た理由の一つであろう。同じく本年度最初の試みとして、Japio 主催の 2009 特許・情報フェア＆コンファレンスでの本研究会の活動に関するプレゼンテーションも行った（11 月 6 日、於科学技術館）。

以上のように、本年度はこれまでにも増して特許情報への関心、また、機械翻訳への関心が高まった年であった。この関心の高まりを反映して、報告書も非常に充実したものとなった。この報告書が、さらなる関心を引き起こすことを願っている。

2．機械翻訳自動評価における名詞句チャンキングの利用

北海学園大学　越前谷　博
沖電気工業株式会社　下畑　さより

2.1　はじめに

近年、GIZA++[1]、SRILM[2]、そして、moses[3]といったツールキットの利用により、統計機械翻訳の研究が大きく進展している。そのような状況において、機械翻訳システムに対する評価作業は不可欠である。しかし、評価作業を人手で行う場合、コストや一貫性を保つことの困難さが問題となる。そこで、翻訳作業を効率よく行うために、評価作業を自動化する機械翻訳自動評価指標の研究が盛んに行われるようになった。機械翻訳自動評価は、機械翻訳システムの開発と評価を短いサイクルで繰り返すことを可能とし、機械翻訳の進歩に大きく寄与するものと考えられる。

現在、最も広く利用されている自動評価指標として BLEU[4]が挙げられる。BLEU は単語 n-gram に基づく手法であり、容易に利用可能である。また、複数の訳文を一括して評価する場合には、人間の評価との間で高い相関を示すことが知られている。しかし、人間の評価作業は基本的には訳文 1 文ごとを評価しており、複数の訳文をまとめて評価しているわけではない。したがって、文単位で人間と同様な評価能力を持った自動評価指標の実現が期待されているが、BLEU やその他の自動評価指標の評価精度は文単位においては十分とはいえず、人間の評価作業とは大きな隔たりがある。

そこで、本報告では、文単位でより高い精度を有する翻訳自動評価指標として、名詞句チャンキングを用いた自動評価指標を提案する。本手法では、チャンキングを用いて MT 訳と参照訳中の名詞句を抽出し、MT 訳と参照訳間において対応する名詞句を決定する。その結果、ある単語が他方の文に複数出現していても、対応名詞句の情報に基づき対応関係の正しい単語を決定することが可能となる。その結果に基づき、単語レベルのスコアを決定する。更に、出現する名詞句を一般化し、名詞句の並びに着目することで句レベルのスコアを得る。単語レベルのスコアに大局的な情報である句レベルのスコアを取り入れることは、より良い自動評価指標の実現に向け、有効と考えられる。また、チャンキングを用いることで、構文解析ツールのような深い解析を要求することなく、構文レベルの情報を利用することが可能となる。その結果、文として不完全な MT 訳が対象となった場合の悪影響を構文解析ツールに比べ回避することができると考えられる。そして、多言語を対象とした場合においても、チャンキングは特定の言語への依存性が構文解析ツールに比べ低く、適用容易性は高くなると考えられる。NTCIR-7 データ[5]を用いた性能評価実験の結果、提案手法は先行研究である自動評価手法 IMPACT[6][7][8]に比べ、より高い相関係数を出力した。したがって、チャンキングを利用した本手法の有効性が確認された。

2.2　チャンキングを用いた機械翻訳自動評価手法

2.2.1　対応名詞句の決定

本手法では、MT 訳と参照訳に対して、チャンキングにより名詞句を決定する。そして、対応する名詞句を PER(Position independent word Error Rate)[9]スコアに基づき決定する。PER は bag-of-words として WER を計算する評価基準である。そのため、語順の制約は考慮されない。PER スコアは小さいほど評価が高いが、ここでは式(1)、(2)に示すようにスコアが大きいほど評価が高くなるよう補正している。式(1)は参照訳中の名詞句を基準とした再現率を示している。式(2)は MT 訳中の名詞句を基準とした適合率を示している。そして、式(1)と式(2)の F 値である式(3)の値を名詞句間の類似度とする。

$$PER\_R = \frac{\text{MT訳中の名詞句と参照訳中の名詞句の一致語数}}{\text{参照訳中の名詞句の語数}} \qquad (1)$$

$$PER\_P = \frac{\text{MT訳中の名詞句と参照訳中の名詞句の一致語数}}{\text{MT訳中の名詞句の語数}} \qquad (2)$$

$$sim(NP_X, NP_Y) = \frac{(1+\gamma^2)PER\_R \times PER\_P}{PER\_R + \gamma^2 PER\_P} \qquad (3)$$

$$\gamma = \frac{PER\_P}{PER\_R} \qquad (4)$$

式(1)から式(4)に基づき対応名詞句を決定する。その処理過程を以下に示す。

(1) MT 訳中の名詞句において、参照訳中の全名詞句との類似度を求める。
(2) 参照訳中の全名詞句の中から類似度が最も高い名詞句を選択する。
(3) (2)より抽出された参照訳中の名詞句において、MT 訳中の全名詞句との類似度を求める。
(4) MT 訳中の全名詞句の中から類似度が最も高い名詞句を選択する。
(5) MT 訳中の選択された名詞句と参照訳中の選択された名詞句が一致する場合、その名詞句の組み合わせを対応名詞句とする。
(6) (1)から(5)の処理を MT 訳中の全名詞句について行う。

全名詞句間で類似度が 0.0 の場合は対応名詞句が存在しないものとする。また、最大となる類似度が複数存在する場合には、一致語数の多い名詞句の組み合わせ、一致語の文字数の多い名詞句の組み合わせの順で一意に対応名詞句を決定する。いずれの方法においても、対応する名詞句が決定できない場合には、対応名詞句は存在しないものとする。上述した対応名詞句の決定の具体例を図 1 に示す。

図 1 では、チャンキングにより MT 訳から名詞句として“ the amount ”、“ the crowning fall ”、そして、“ the end ”が抽出される。また、参照訳からは名詞句として“ it ”、“ the end part ”、“ the amount ”、そして、“ crowning drop ”が抽出される。次いで、MT 訳中の名詞句と参照訳中の全名詞句間において、式(1)から式(4)に基づき類似度を求める。例えば、MT 訳中の名詞句“ the end ”と参照訳中の全名詞句間で類似度を求めると、参照訳中の名詞句“ the end part ”との類似度が

$PER\_R = \frac{2}{3} = 0.6667$、$PER\_P = \frac{2}{2} = 1.0$ より、$sim(NP_X, NP_Y) = \frac{(1+1.4999^2)\times 0.6667\times 1.0}{0.6667 + 1.4999^2 \times 1.0} = 0.7429$

となり、最大となる。また、参照訳中の名詞句“the end part”とMT訳中の全名詞句間で類似度を求めると、MT訳中の名詞句“the end”との類似度が最大となる。したがって、この2つの名詞句間の類似度は互いに最大となるため、対応名詞句となる。この処理をMT訳中の他の名詞句について行うことで、MT訳中の名詞句“the amount”、“the crowning fall”、“the end”は参照訳中の名詞句“the amount”、“crowning drop”、“the end part”にそれぞれ対応付けられる。

(1) **名詞句の抽出**

MT**訳**：
in general , [NP the amount] of [NP the crowning fall] is large like [NP the end] .

参照訳：
generally , the closer [NP it] is to [NP the end part] , the larger [NP the amount] of [NP crowning drop] is .

(2) PER**スコアに基づく対応名詞句の決定**

MT**訳**:
in general , [NP the amount] of [NP the crowning fall] is large like [NP the end] .

0.7429　1.0000　0.3714

参照訳：
generally , the closer [NP it] is to [NP the end part] , the larger [NP the amount] of [NP crowning drop] is .

図1　対応名詞句の決定の具体例

2.2.2　単語レベルのスコア

本手法では、2.2.1より決定された対応名詞句に基づき、対応関係にある単語をより正確に決定したうえで、単語レベルのスコアを計算する。具体的には、まず、LCS（Longest Common Subsequence：最長共通部分列）の値に相当する共通単語、即ち、LCS経路を得る。LCS経路が複数存在する場合、決定された対応名詞句を用い、以下の式(5)より一意にLCS経路を決定する。

$$RS = \sum_{c \in LCS} (\sum_{w \in c} weight(w))^{\beta} \quad (5)$$

$$weight(w) = \begin{cases} 2 & \text{対応名詞句内の単語} \\ 1 & \text{未対応名詞句内の単語} \end{cases}$$

式(5)の w は共通単語であり、c は共通単語が連続して出現する共通部分である。βは共通部分を構成する共通単語の数に対する重みづけのためのパラメータである。式(5)の LCS は任意のLCS経路を意味する。したがって、対応名詞句中の単語が数多く含まれているほど式(5)の RS（Route Score）の値は大きくなる。図2にLCS経路決定の具体例を示す。

図2の(1)において、MT訳と参照訳の間のLCSは7となる。しかし、LCSが7であるLCS経路は複数存在するため、共通単語を一意に決定することができない。そこで、式(5)に基づき、より正確に共通単語を決定する。図2の(1)の提案手法においては、βの値が2.0の場合、RS は32

$(=1^{2.0}+(2+2+1)^{2.0}+2^{2.0}+1^{2.0}+1^{2.0})$となる。ここで、対応名詞句中の単語はNP1の“the”、“amount”、そして、NP2の“crowning”であり、この3単語の重みは2となる。一方、IMPACTが選択したLCS経路についても式(5)の*RS*を求めると19（$=(1+1)^{2.0}+(2+1)^{2.0}+2^{2.0}+1^{2.0}+1^{2.0}$）となる。提案手法とIMPACTの共通単語を比べると、IMPACTではMT訳のNP1中の“the”は参照訳の最初の“the”に対応すると位置づけられている。それに対して、提案手法では対応名詞句の単語の重みを大きくすることで参照訳のNP1中の“the”が選択されている。このように対応名詞句の情報を利用することで、より正確に対応関係にある単語の決定が可能となる。次いで、提案手法では、IMPACTと同様に、一度決定された共通単語を除き、更なるLCSに基づく共通単語を決定する。図2の(2)においては、(1)より決定された共通部分“, ”、“the amount of”、“crowning”、“is”、そして、“.”を除き、残された単語に対して更にLCSと式(5)に基づき共通部分を決定する。その結果、“the”とNP3中の“the end”が共通部分となる。

(1) LCS**経路の決定処理1**：
LCS=7

提案手法

MT**訳**:

in general **,** [$_{NP1}$ **the amount] of** [$_{NP2}$ the **crowning** fall] **is** large like [$_{NP3}$ the end] **.**

参照訳：

1^2 $(2+2+1)^2$ 1^2 1^2 2^2

generally **,** the closer [$_{NP}$ it] is to [$_{NP3}$ the end part] , the larger [$_{NP1}$ **the amount] of** [$_{NP2}$ **crowning** drop] **is .**

.....

IMPACT

MT**訳**:

in general **,** [$_{NP1}$ **the amount] of** [$_{NP2}$ the **crowning** fall] **is** large like [$_{NP3}$ the end] **.**

参照訳:

$(1+1)^2$ $(2+1)^2$ 2^2 1^2 1^2

generally **, the** closer [$_{NP}$ it] is to [$_{NP3}$ the end part] , the larger [$_{NP1}$ the **amount] of** [$_{NP2}$ **crowning** drop] **is .**

(2) LCS**経路の決定処理2**：
LCS=3

提案手法

MT**訳**:

in general , [$_{NP1}$ *the amount*] *of* [$_{NP2}$ **the** *crowning* fall] *is* large like [$_{NP3}$ **the end**] .

参照訳：

generally , **the** closer [$_{NP}$ it] is to [$_{NP3}$ **the end** part] , the larger [$_{NP1}$ *the amount*] *of* [$_{NP2}$ *crowning* drop] *is* .

図2　LCS経路の決定の具体例

提案手法では、このようにして得られる全共通部分を用いて単語レベルのスコアを算出する。

その際の計算式は以下の式(6)から式(9)となる。

$$R_{wd} = \left(\frac{\sum_{i=0}^{RN} (\alpha^i \sum_{c \in CC} length(c)^\beta)}{m^\beta} \right)^{\frac{1}{\beta}} \quad (6)$$

$$P_{wd} = \left(\frac{\sum_{i=0}^{RN} (\alpha^i \sum_{c \in CC} length(c)^\beta)}{n^\beta} \right)^{\frac{1}{\beta}} \quad (7)$$

$$score_{wd} = \frac{(1+\gamma^2) R_{wd} P_{wd}}{R_{wd} + \gamma^2 P_{wd}} \quad (8)$$

$$\gamma = \frac{P_{wd}}{R_{wd}} \quad (9)$$

式(6)と式(7)はそれぞれ再現率と適合率を示し、その F 値を式(8)と式(9)より求める。式(6)、(7)の $\sum_{c \in cc} length(c)^\beta$ は個々の共通部分ごとに求めた値の総和である。したがって、β の値が 2.0 の場合、図 2 の(1)においては、13（$=1^{2.0}+3^{2.0}+1^{2.0}+1^{2.0}+1^{2.0}$）となる。ここでは単語の重みは全て 1 である。式(6)、(7)の α^i は LCS の決定処理の回数に伴うパラメータであり、カウンタ i の初期値は 0 である。パラメータ α の値が 0.5 の場合には、図 2 の(1)においては、i の値は 0 であるため、$\alpha^i \sum_{c \in cc} length(c)^\beta$ は 13（$=0.5^0 \times 13$）となる。同様に図 2 の(2)においては 2.5（$=0.5^1 \times (1^{2.0}+2^{2.0})$）となる。共通単語はこれ以上存在しないため LCS の決定処理はこれで終了となる。それに伴い、式(6)、(7)の RN は 1 となる。その結果、図 2 においては式(6)、(7)の分子は 15.5（＝13＋2.5）となる。更に、式（6）の R_{wd} は 0.1969（$=\sqrt{15.5/20^{2.0}}$）、式（7）の P_{wd} は 0.2625（$=\sqrt{15.5/15^{2.0}}$）となる。したがって、式（8）の $score_{wd}$ は 0.2164（$=\frac{(1+1.3332)^2 \times 0.1969 \times 0.2625}{0.1969+1.3332^2 \times 0.2625}$）となる。その際の γ の値は 1.3332（=0.2625/0.1969）である。

このように、本手法では、チャンキングを用いることにより、より正確に共通単語を決定したうえで単語レベルのスコアを得ることが可能となる。また、複数参照訳を用いた場合には、最大の R_{wd}、P_{wd} を用いて式（8）の $score_{wd}$ を求める。

2.2.3 句レベルのスコア

本手法では、更に、句レベルのスコアを単語レベルのスコアに対して取り入れる。句レベルのスコアでは、チャンキングにより得られた名詞句のみを抽出し、それらを一つの単語として一般化したうえでスコアを計算する。名詞句は他の句に比べ出現頻度が高く、文全体の大局的な情報をスコアに反映させるために有効と考えられる。

図 3 に名詞句に基づく一般化の具体例を示す。図 3 の(2)では、番号が付与された名詞句が対応名詞句を表している。番号が付与されていない名詞句“NP”は未対応名詞句を示している。図 3

においては、MT 訳と参照訳間の全ての名詞句を抽出し、それを一般化した場合、MT 訳においては“NP1 NP2 NP3”、参照訳においては“NP NP3 NP1 NP2”が得られる。これらの一般化された MT 訳と参照訳間において、名詞句の重みを全て 1 とし、式（10）から式（13）を適用することで句レベルのスコアを求める。

(1) **対応名詞句の決定**

MT**訳**:

in general , [$_{NP1}$ the amount] of [$_{NP2}$ the crowning fall] is large like [$_{NP3}$ the end] .

参照訳：

generally , the closer [$_{NP}$ it] is to [$_{NP3}$ the end part] , the larger [$_{NP1}$ the amount] of [$_{NP2}$ crowning drop] is .

(2) **名詞句の抽出**

MT**訳**：

NP1　NP2　NP3

参照訳：

NP　NP3　NP1　NP2

図 3　名詞句に基づく一般化の具体例

$$R_{np} = \left(\frac{\sum_{i=0}^{RN} (\alpha^{i} \sum_{cnp \in CCNP} length(cnp)^{\beta})}{(m_{cnp} \times \sqrt{m_{no-cnp}})^{\beta}} \right)^{\frac{1}{\beta}} \quad (10)$$

$$P_{np} = \left(\frac{\sum_{i=0}^{RN} (\alpha^{i} \sum_{cnp \in CCNP} length(cnp)^{\beta})}{(n_{cnp} \times \sqrt{n_{no-cnp}})^{\beta}} \right)^{\frac{1}{\beta}} \quad (11)$$

$$score_{np} = \frac{(1+\gamma^{2}) R_{np} P_{np}}{R_{np} + \gamma^{2} P_{np}} \quad (12)$$

$$\gamma = \frac{P_{np}}{R_{np}} \quad (13)$$

式（10）は参照訳を基準とした再現率、式（11）は MT 訳を基準とした適合率を示している。また、式（10）と式（11）の分母は、対応名詞句の数に対し、未対応名詞句の数の平方根を積の重みとして用いている。これにより、未対応名詞句が対応名詞句に対して数多く存在する場合、未対応名詞句の重みが軽減される。文が凝縮された名詞句の並びにおいては、一つの未対応名詞句の存在がスコアの低下に大きな影響を与える。そのため本処理では、未対応名詞句の重みを軽減することで、未対応名詞句の数が多い場合でもスコアが過度に低くならないよう補正している。なお、未対応名詞句が一つも存在しない場合には、分母が 0 となるため、その場合には、未対応

名詞句の重みを 1 として計算する。

図 3 においては、α が 0.5、β が 2.0 の場合、R_{np}、P_{np} は共に 0.7071（$=\sqrt{(1\times 2^{2.0}+0.5\times 1^{2.0})/(3\times 1)^{2.0}}$）となる。その結果、$score_{np}$ は 0.7071 となる。なお、複数参照訳が用いられた場合には、個々の参照訳との $score_{np}$ の平均値を用いる。

2.2.4 最終スコア

2.2.2 の単語レベルのスコア $score_{wd}$ と 2.2.3 の句レベルのスコア $score_{np}$ を以下の式(14)に用いることにより、最終的なスコアを計算する。

$$score = \frac{score_{wd} + \delta \times score_{np}}{1+\delta} \qquad (14)$$

式(14)におけるパラメータ δ は単語レベルのスコア $score_{wd}$ と句レベルのスコア $score_{np}$ の重みを制御するためのパラメータである。δ が 1 の場合には、$score_{wd}$ と $score_{np}$ の重みの関係は 1 対 1 となる。δ として 0.7 を用いた場合、図 1 から図 3 で用いた MT 訳と参照訳の間の最終的なスコアは 0.4185（$=\sqrt{(0.2164+0.7\times 0.7071)/(1+0.7)}$）となる。

このように、本手法ではチャンキングを用いることにより、MT 訳と参照訳間における単語レベルで正しい対応単語を決定し、更に、句レベルの大局的な情報に基づくスコアを取り入れることで、より良い機械翻訳自動評価基準を実現する。

2.3 性能評価実験

2.3.1 実験データ

MT 訳には、特許文 100 文の日本文を NTCIR-7 に参加した 12 グループによる 12 の機械翻訳システムが翻訳した英文 100 文を用いた。表 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 は統計機械翻訳、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、参照訳には、入力文 100 文に対して 4 名のバイリンガルが作成した 4 つの参照訳を用いた。

表 1　グループ名とその手法

Group	tori	HIT2	JAPIO	KLE	MIT	NAIST-NTT
手法	SMT	SMT	RBMT	SMT	SMT	SMT
Group	NiCT-ATR	NTT	Kyoto-U	MIBEL	Moses	tsbmt
手法	SMT	SMT	EBMT	SMT	SMT	RBMT

2.3.2 実験方法

本実験では、2.3.1 で述べた MT 訳と参照訳を用いて提案手法によるスコア計算を行う。また、そのスコアがどの程度、人間による評価と一致するかを求めるために、ピアソンの相関係数とスピアマンの順位相関係数を用いる。その際、人手評価については、3 名のバイリンガルが 12 の機

械翻訳システムにより得たMT訳に対して、Adequacyと Fluencyの観点より5段階評価を行い、3 名の評価値のメジアン値を最終的な評価値として用いる。また、提案手法の有効性を検証するために、翻訳自動評価手法IMPACTを用いて同様の実験を行う。

提案手法においては、式(6)、(7)、(10)、(11)のパラメータαとβには 0.1 と 1.1 をそれぞれ用いる。また、式(14)のパラメータδには0.3を用いる。これらの値は他のデータによる性能表実験[10]に基づき決定した。そして、今回、チャンキングには Shallow Parser[11]を用い、名詞句の決定を行う。

2.3.3 実験結果

表2にAdequacyにおけるピアソンの相関係数、表3にFluencyにおけるピアソンの相関係数、表4にAdequacyにおけるスピアマンの順位相関係数、表5にFluencyにおけるスピアマンの順位相関係数をそれぞれ示す。表中の“Avg.”は、機械翻訳システムごとの相関係数の平均である。“All”は、全12の機械翻訳システムによる1,200文のMT訳に対する自動評価のスコアと人手評価をデータとした場合の相関係数である。

表2と表4のAdequacyにおいて提案手法は、“JAPIO”を除く全ての相関係数において、IMPACTを上回った。表3と表5のFluencyにおいては、いくつかの相関係数においてIMPACTを下回ったが、“Avg.”及び“All”では全てIMPACTを上回る相関係数を示した。これらの結果は提案手法の有効性を示すものである。

表2　Adequacyにおけるピアソンの相関係数

	tori	HIT2	JAPIO	KLE	MIT	NAIST-NTT	NiCT-ATR
提案手法	0.7868	0.4976	0.5971	0.5702	0.6575	0.6740	0.7674
IMPACT	0.7639	0.4487	0.5980	0.5371	0.6371	0.6255	0.7249
	NTT	Kyoto-U	MIBEL	Moses	tsbmt	Avg.	All
提案手法	0.7654	0.7210	0.6365	0.7778	0.5703	0.6684	0.6842
IMPACT	0.7007	0.7125	0.5981	0.7621	0.5345	0.6369	0.6574

表3　Fluencyにおけるピアソンの相関係数

	tori	HIT2	JAPIO	KLE	MIT	NAIST-NTT	NiCT-ATR
提案手法	0.5855	0.3774	0.5689	0.4662	0.5735	0.5338	0.7191
IMPACT	0.5581	0.3407	0.5821	0.4586	0.5768	0.4852	0.6896
	NTT	Kyoto-U	MIBEL	Moses	tsbmt	Avg.	All
提案手法	0.5797	0.6423	0.3254	0.5908	0.4319	0.5329	0.5572
IMPACT	0.5612	0.6320	0.3492	0.6034	0.4166	0.5211	0.5469

表 4　Adequacy におけるスピアマンの順位相関係数

	tori	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NiCT -ATR
提案手法	0.7454	0.5018	0.5837	0.5144	0.6503	0.6536	0.6756
IMPACT	0.7336	0.4881	0.5992	0.4741	0.6382	0.5841	0.6409
	NTT	Kyoto-U	MIBEL	Moses	tsbmt	Avg.	All
提案手法	0.7280	0.7261	0.5963	0.7637	0.6076	0.6455	0.6758
IMPACT	0.6703	0.7067	0.5617	0.7411	0.5583	0.6164	0.6515

表 5　Fluency におけるスピアマンの順位相関係数

	tori	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NiCT -ATR
提案手法	0.5703	0.3243	0.5446	0.4193	0.5742	0.5055	0.6453
IMPACT	0.5481	0.3285	0.5572	0.3976	0.5960	0.4317	0.6334
	NTT	Kyoto-U	MIBEL	Moses	tsbmt	Avg.	All
提案手法	0.5648	0.6620	0.3336	0.6234	0.4485	0.5180	0.5553
IMPACT	0.5471	0.6454	0.3222	0.6319	0.4358	0.5062	0.5489

2.4　まとめと今後の予定

性能評価実験より、チャンキングを用いた提案手法が従来手法の IMPACT に対して、文単位の自動評価においてより高い相関係数を示すことを確認した。しかし、Fluency においては、いくつかの機械翻訳システムで IMAPCT を下回った。提案手法では、MT 訳と参照訳間の一致単語をより正確に決定することが可能となり、それが Adequacy の相関係数のより大きな向上をもたらしたと考えられる。以上の結果より、今後は Fluency の相関係数の向上のための改良が必要である。

また、本報告では MT 訳として英文を用いた実験を行ったが、今後は更に他の言語文を用いて提案手法の検証を行う予定である。

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] Och, Franz Josef. and Ney, Hermann. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, Vol.29, No.1, pp.19-51, 2003.

[2] Stolcke, Andreas. SRILM – An Extensible Language Modeling Toolkit, 7th International Conference on Spoken Language Processing, pp.901-904, 2002.

[3] Koehn, Pilipp., Hoang, Hieu., Birch, Alexandra., Callison-Burch, Chris., Federico,

Marcello., Bertoldi, Nicola., Cowan, Brooke., Shen, Wade., Moran, Christine., Zens, Richard., Dyer, Chris., Bojar, Ondrej., Constantin Alexander. and Herbst, Evan. Moses: Open Source Toolkit for Statistical Machine Translation, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp.177-180, 2007.

[4] 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.

[5] 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.

[6] 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.

[7] 越前谷博、江原暉将、下畑さより、藤井敦、内山将夫、山本幹雄、宇津呂武仁、神門典子. NTCIR-7 データを用いた機械翻訳自動評価基準のメタ評価, 平成 20 年度 AAMT/Japio 特許翻訳研究会報告書, pp.2-13, 2009.

[8] Echizen-ya, Hiroshi., Ehara, Terumasa., Shimohata, Sayori., Fujii, Atsushi., Utiyama, Masao., Yamamoto, Mikio., Utsuro, Takehito. and Kando, Noriko. Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR-7, Proceedings of the 3rd Workshop on Patent Translation, pp.9-16, 2009.

[9] Su, Keh-Yih., Wu, Ming-Wen. and Chang, Jing-Shin. A New Quantitative Quality Measure for Machine Translation Systems, Proceedings of the fifteenth International Conference on Computational Linguistics, pp.433-439, 1992.

[10] 小山田崇、越前谷博、荒木健治. 単語情報及びフレーズによる大局的情報を用いた機械翻訳自動評価手法, 情報処理学会研究報告(Vol.2010-NL-195(No.3)), pp.1-7, 2010.

[11]Sha, Fei. and Pereira, Fernando. Shallow Parsing with Conditional Random Fields. Proceeding of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics 2003, pp.134-141, 2003.

３．１　日英対訳特許からの専門用語対訳辞書生成における同義語集合作成に関する調査

筑波大学大学院システム情報工学研究科
森下　洋平、宇津呂　武仁、山本　幹雄

3.1.1　はじめに

[森下08] [Morishita08]では、対訳特許文からの専門用語対訳対獲得を目的として、NTCIR-7 の特許翻訳タスク[Fujii08]で配布された日英180 万件の対訳特許文を用いて評価を行った。これらの手法では、フレーズベース統計的機械翻訳モデル[Koehn07]、要素合成法[外池07]、Support Vector Machines [Vapnik98](SVM) による機械学習を用いることによって、専門用語対訳対獲得の適合率を改善させている(図1)。また、これらの手法は対訳特許文から抽出した日本語専門用語に対し訳語候補を作成し、トークン単位の評価を行う。そのため、ある専門用語対訳対を獲得する際に1 対訳文しか考慮していない。しかし、実際は多くの専門用語は対訳特許文中に複数出現しているため、異なる訳が存在する。そのため、複数の文で獲得された専門用語対訳対に対し、専門用語間の同義、異義関係を見極めることが辞書作成に必要不可欠である。

そこで、本論文では複数の文で獲得された専門用語対に対し、同義専門用語集合を作成する手法を提案する[森下10]。提案手法においては、文字列の類似度や共起語の類似度を用いて、適用可能な範囲は小さいが100%の精度で同義専門用語を同定可能な決定的規則を作成した。この決定的規則により同義専門用語の5.8%を同定することができた。さらに、決定的規則によって同定できない同義専門用語に対し、SVM を用いた同義・異義判定を行った。その結果、決定的規則により同定できない同義専門用語に対して、適合率93.2%、再現率23.1%で同定を行うことができた。さらに、決定的規則およびSVM の判定結果を用いて、同義専門用語集合の同定を行った。その結果、同義集合に対して適合率94.8%、再現率36.6%で同定を行うことができた。

3.1.2　同義専門用語集合同定の流れ

図 1 に、同義専門用語集合同定の流れを示す。

(1) 180 万件から得られた対訳対の中から、同義専門用語候補集合を作成する。
(2) 同義専門用語候補集合に含まれる日英対訳対の全組み合わせを作成する。全組み合わせ中、15.6%が同義専門用語組、84.4%が異義専門用語組となった。それらに対し、同義集合を確実に選定するための決定的規則により同義専門用語および異義専門用語を同定する。決定的規則により、全組み合わせ中 0.9%の同義専門用語、2.2%の異義専門用語に対し、同定を行った。
(3) (2) で同定できない 96.9%の組み合わせに対して、機械学習による同義、異義の同定を行う。同義専門用語集合に対し、SVM による同義判定の結果、適合率 93.2%、 再現率 23.1%で同定を行った。また、異義判定の結果、適合率 93.7%、 再現率 74.9%で同定を行った。

(4) (2) および(3) の同定結果をもとに、同義専門用語集合を同定する。

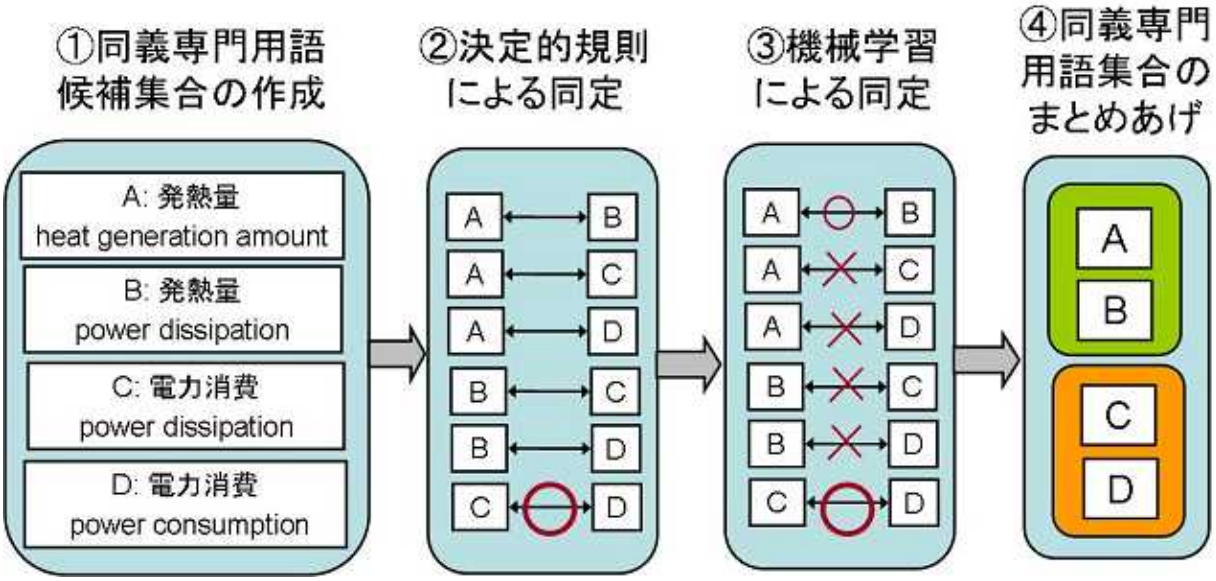


図 1　同義専門用語集合同定の流れ

3.1.3　同義専門用語候補集合の作成

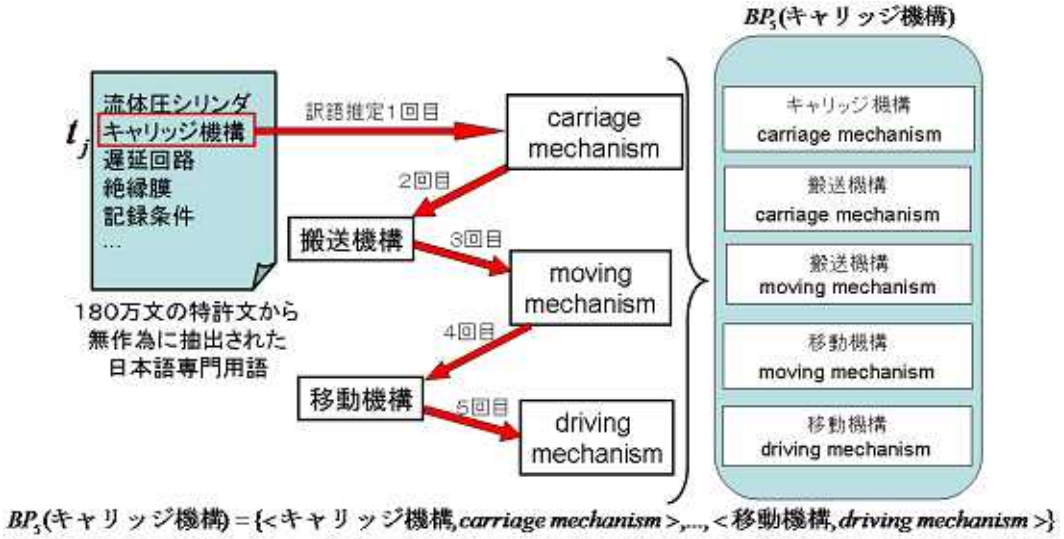


図 2　同義専門用語候補集合の作成

図 2 に、同義専門用語候補集合作成の流れを示す。

1. 180 万文の特許文から日本語専門用語t_Jを無作為に抽出する。抽出した日本語専門用語t_Jを、要素一個の初期集合$T_J^1 = \{t_J\}$とする。
2. 日本語用語集合T_J^1の訳語推定を行い、英語用語集合T_E^2を作成する。以下の式を用いて訳語推定を行う。式中の$PS(t_J)$は、t_Jが出現する対訳文を示す。また、$r_1(t_J, <s_J, s_E>, PT_{JE}) = t_E$は、$t_J$が出現する対訳文$<s_J, s_E>$の中で、$s_E$側に出現する$t_J$の訳語候補のうち、日英方向のフレーズテーブル$PT_{JE}$の順位が 1 位となる英語訳語$t_E$を示す[1]。

[1]得られた対訳対t_J - t_Eに対し、対訳特許文180 万件中の頻度が6 以上1000 以下でない場合、それらを除外した。

$$T_E^{i+1} = \bigcup_{t_J \in T_J^i} \bigcup_{<S_J, S_E> \in PS(t_J)} r_1(t_J, <s_J, s_E>, PT_{JE})$$

3. 手順 2 と同様に、英語用語集合T_E^2を訳語推定して日本語用語集合T_J^3を作成する。以下の式を用いて訳語推定を行う。式中の$PS(t_E)$は、t_Eが出現する対訳文を示す。また、$r_1(t_E, <s_J, s_E>, PT_{EJ}) = t_J$は、$t_E$が出現する対訳文$<s_J, s_E>$の中で、$s_J$側に出現する$t_E$の訳語候補のうち、英日方向のフレーズテーブル$PT_{EJ}$の順位が１位となる日本語訳語$t_J$を示す[2]。

$$T_J^{i+1} = \bigcup_{t_E \in T_E^i} \bigcup_{<S_J, S_E> \in PS(t_E)} r_1(t_E, <s_J, s_E>, PT_{EJ})$$

4. 2、3 の処理を繰り返し、k 回訳語推定を行うことにより得られた対訳専門用語を集めた集合をBP_k とする（本研究では、$k = 5$ とした）。
5. 日英対訳対が正しい対応でない場合、手動で除外する。

本論文では、50 個のt_J を用いて、50 個のBP_5 を作成した。単一の日英対訳対が2つ以上の異なるBP_5 に出現する場合、片方のBP_5 を廃棄し、再度作成した。また、BP_5 中に含まれる日英対訳対に対し、語義ID を人手により付与し、日英対訳対が同義ならば同一の語義ID を付与した。表 1に、50 個のBP_5 中に含まれる日英対訳対および語義ID の種類数を示す。手順4 までで、合計2118 対の日英対訳対が得られたが、手順5 で正しい対応でないものを除外したところ、合計1921 対の日英対訳対が得られた。

また、各BP_5 に含まれる日英対訳対$\langle t_J, t_E \rangle$間の同義、異義関係を判定するために、各BP_5 の中に含まれる全組み合わせ$\langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$組を作成した。表 2に、$BP_5$ 中に含まれる$\langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$組数を示す。 表 2に示す全組を評価対象とし、決定的規則による同義、異義組の同定を3.1.5節で行い、機械学習による同義、異義組の同定を3.1.6節で行う。

表 1　全同義専門用語候補集合に含まれる日英対訳対数

	合計	$BP_5$1 個あたりの平均
日英対訳対	1,921	38.4
語義 ID の種類数	257	5.1

表 2　$\langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$組数

	各BP_5から作成した全組数(割合)	$BP_5$1 個あたりの平均
合計	106,461(100%)	2129.2
同義組合計	16,615(15.6%)	332.3
異議組合計	89,846(84.4%)	1796.9

3.1.4　素性

決定的規則や機械学習に用いる素性として、表 3 に示す素性を用いた。素性は大きくわけて、文字列素性と共起語素性に分類される。

---

[2]得られた対訳対t_J - t_E に対し、対訳特許文180 万件中の頻度が6 以上1000 以下でない場合、それらを除外した。

3.1.4.1 文字列素性

以下では、特に表 3に示す $f5, f6, f7$ 素性について説明する。 $f5$ 素性は、文字列の非共有箇所に対し要素合成法の同一訳が存在するか否かを求める。例えば、「アース電位」「接地電位」という対応の場合、非共有箇所である「アース」「接地」に対し、要素合成法で同一訳が得られれば素性の値は真となる。 $f6$ および $f7$ の素性は、非包含箇所の文字列から、同義異義関係を求める。例えば、 $f6$ で「プリンタ」「プリンター」という対応の場合、非包含箇所である「ー」から人手で作成した規則により同義と判定されれば、素性の値は真となる。また、人手による規則は評価事例を用いて作成した。

3.1.4.2 共起語素性

表 3に示す $f8$ の素性は、共起語一致数を示す。 $f8(t_J^i, t_J^i)$ の場合は、対訳特許文180 万件内で、t_J^i, t_J^j と共起する日本語フレーズの中で、φ^2 尺度の値が0.00001 以上かつ上位1000 位以内のものをそれぞれ求め、一致数を求めた。 $f8(t_E^i, t_E^j)$ の場合も、同様に行った。

表 3 決定的規則および SVM 学習に用いた素性

素性タイプ	素性名	定義
文字列素性	$f1$: 文字列が同一	$f1(t_X^i, t_X^j)$: t_X^i, t_X^j が同一ならば真となる
	$f2$: 編集距離類似度	$f2(t_X^i, t_X^j) = 1 - \frac{ED(t_X^i, t_X^j)}{max(\lvert t_X^i \rvert, \lvert t_X^j \rvert)}$: ED は t_X^i と t_X^j の間の編集距離，$\lvert X \rvert$ は X に含まれる文字数を示す
	$f3$: バイグラム類似度	$f3(t_X^i, t_X^j) = \frac{\lvert bigram(t_X^i) \cap bigram(t_X^j) \rvert}{max(\lvert t_X^i \rvert, \lvert t_X^j \rvert)+1}$: $bigram(X)$ は，X に含まれる文字単位のバイグラム
	$f4$: 同一の形態素 (単語) 数の割合	$f4(t_X^i, t_X^j) = \frac{\lvert morph(t_X^i) \cap morph(t_X^j) \rvert}{max(\lvert t_X^i \rvert, \lvert t_X^j \rvert)}$: $morph(X)$ は，X に含まれる形態素
	$f5$: 非共有箇所に対し要素合成法の同一訳が存在	$f5(t_X^i, t_X^j)$: t_X^i, t_X^j で文字列が一致しない箇所 x_i,x_j に対して，要素合成法による訳語推定を行い，同一訳が存在する場合，素性の値は真となる．
	$f6$: 同義包含関係あり	$f6(t_X^i, t_X^j)$: t_X^i, t_X^j の一方がもう一方に包含されており，かつ非包含箇所が「ー」「s」「es」など，t_X^i, t_X^j が同義と判定できる文字列
	$f7$: 異義包含関係あり	$f7(t_X^i, t_X^j)$: t_X^i, t_X^j の一方がもう一方に包含されており，かつ $f6(t_X^i, t_X^j)$ の値が真でない．
共起語素性	$f8$: 共起語一致数	$f8(t_X^i, t_X^j) = \lvert cooccur(t_X^i) \cap cooccur(t_X^j) \rvert$: $cooccur(X)$ は，対訳特許文 180 万件内で X と共起し，かつ ϕ^2 尺度の値が上位 1000 位以内の単語
翻訳素性	$f9$: 要素合成法の共通訳が存在	$f9(t_Z^i, t_Y^j)$:要素合成法により，t_Z^i を訳語推定し t_Y^j が得られる．または t_Y^j を訳語推定し t_Z^i が得られる
	$f10$: フレーズ翻訳テーブルの共通訳が存在	$f10(t_Z^i, t_Y^j)$:フレーズ翻訳テーブルにより，t_Z^i を訳語推定し t_Y^j が得られる．または t_Y^j を訳語推定し t_Z^i が得られる
決定的規則で用いる素性	$f11$: 同義文字列素性	$f11(t_J^i, t_J^j, t_E^i, t_E^j)$: $\{f1(t_J^i, t_J^j)$ が真 または $f5(t_J^i, t_J^j)$ が真 または $f6(t_J^i, t_J^j)$ が真$\}$ かつ $\{f1(t_E^i, t_E^j)$ が真 または $f5(t_E^i, t_E^j)$ が真 または $f6(t_E^i, t_E^j)$ が真$\}$
	$f12$: 同義共起語素性	$f12(t_J^i, t_J^j, t_E^i, t_E^j)$: $f8(t_J^i, t_J^j)$ の値が 800 以上かつ $f8(t_E^i, t_E^j)$ の値が 100 以上
	$f13$: 異議文字列素性	$f13(t_J^i, t_J^j, t_E^i, t_E^j)$: $f7(t_J^i, t_J^j)$ が真 または $f7(t_E^i, t_E^j)$ が真

$X \in \{J, E\}$, $(Z, Y) \in \{(J, E), (E, J)\}$

3.1.5 決定的規則による同定

表 3 に、決定的規則に用いる素性を示す。決定的規則として、同義の決定的規則と異義の決定的規則を定義した[3]。決定的規則が成り立った場合に、同義または異義を同定する。決定的規則により、全$\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle$組中、3.1%の同義組および異義組に対し同定を行い、残りの96.9%に対し3.1.6節に示す機械学習による同定を行う。

3.1.5.1 同義の決定的規則

同義の決定的規則に用いる素性として、同義文字列素性と同義共起語素性の2 つを定義した。2 つのうちいずれかの素性の値が真である場合、$\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle$を同義と判定する。図 3 に、決定的規則によって同定できた同義専門用語集合の割合を示す。同義決定的規則は、全組の0.9%に適用された。また、同義組に対する同義同定の適合率は100%(966/966)、 再現率は5.8%(966/16,615) となった。

以下では、同義決定的規則の各素性の性能を調べる。同義文字列素性の、同義組に対する同義同定の適合率は100%(810/810)、 再現率は4.9%(810/16,615) となった。また、同義共起語素性の、同義組に対する同義同定の適合率は 100%(290/290)、 再現率は 1.7%(290/16,615) となった。

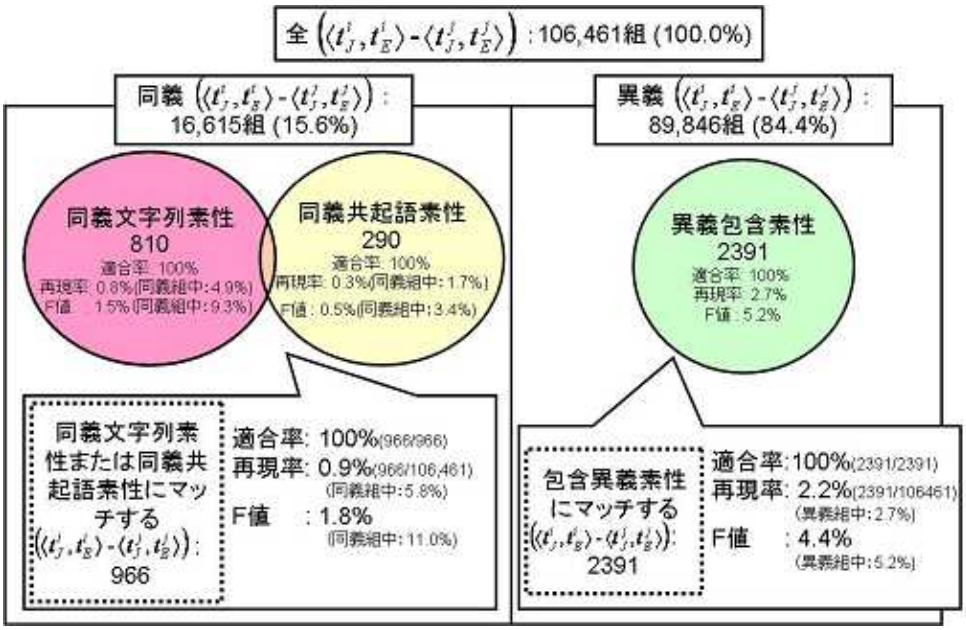


図 3 決定的規則による同定結果

3.1.5.2 異義の決定的規則

異義の決定的規則に用いる素性として、異義包含素性を定義した。異義包含素性の値が真である場合に、$\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle$を異義と判定する。異義の決定的規則は、全組の2.2%に適用された。また、異義組に対する異義同定の適合率は100%(2391/2391)、 再現率は2.7%(2391/89,846) となった。

表 4 機械学習による同義・異義判定の性能評価(%)

[3] 本論文では，評価事例を用いて規則の作成を行った．今後は，評価事例以外の事例を用いて，規則を作成する．

(1) 同義の判定

	適合率	再現率	F 値
ベースライン	89.2 (4742/5376)	33.6 (4742/15648)	48.8
SVM - 適合率が最大となる下限	**93.2** (3617/3879)	23.1 (3617/15649)	37.0
SVM - F 値が最大となる下限	71.0 (8922/12561)	57.0 (8922/15649)	**63.3**

(2) 異義の判定

	適合率	再現率	F 値
ベースライン	88.8 (86821/97728)	99.3 (86821/87455)	93.8
SVM - 適合率が最大となる下限	**93.7** (65488/69891)	74.9 (65488/87455)	83.2
SVM - F 値が最大となる下限	91.5 (85647/93625)	97.9 (85647/87455)	**94.6**

3.1.6 機械学習による判定

3.1.5節で示した決定的規則による同定ができない103,104 組の$\langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$組に対し、10 分割交差検定を用いて機械学習による判定を行う。SVM で評価に用いる素性は、訓練・評価事例以外を用いて表 3 の中から最適な組み合わせを決定した。また、信頼度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、再現率、F 値がそれぞれ最大となる結果を

表 4 に示す。

表 4 に、SVM による同義判定結果を示す。ベースラインは、t_J^i - t_J^iが同一またはt_E^i - t_E^iが同一のものとした。ベースラインに対し、SVM で適合率が最大となるモデルとの比較を行ったところ、再現率が下がったものの、適合率が有意水準1%で改善された(89.2%から93.2%に改善)。表 5に、SVM による改善例を示す。

表 5 SVM による同義、異義判定の改善例

(1) 同義判定

ベースライン:t_J^i-t_J^j が同一または t_E^i-t_E^j が同一

SVM:適合率が最大となる下限を用いたモデル

(a)SVM のみが同義と判定し正解

$\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle$	人手による同義・異義判定	ベースラインによる判定	SVMによる判定	f8: 共起語一致数(log)
〈アース電位‖ground potential〉-〈接地電位‖grounded potential〉	同義	異義	同義	(t_J^i, t_J^j): 0.69 (t_E^i, t_E^j): 3.29

(b) SVM のみが異義と判定し正解

$\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle$	人手による同義・異義判定	ベースラインによる判定	SVMによる判定	f8: 共起語一致数(log)
〈薬液‖liquid〉-〈冷却媒体‖liquid〉	異義	同義	異義	(t_J^i, t_J^j) : 0 (t_E^i, t_E^j) : 0.61

(2) 異義判定

ベースライン:t_J^i-t_J^j が同一でない かつ t_E^i-t_E^j が同一でない

SVM:適合率が最大となる下限を用いたモデル

(a)SVM のみが異義と判定し正解

$\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle$	人手による同義・異義判定	ベースラインによる判定	SVMによる判定	f2: 編集距離類似度	f3: バイグラム類似度	f4: 同一の形態素(単語)数の割合
〈管状型部材‖tubular member〉-〈中空筒部材‖tubular member〉	異義	同義	異義	(t_J^i, t_J^j): 0.5 (t_E^i, t_E^j): 1	(t_J^i, t_J^j) : 0.2 (t_E^i, t_E^j) : 1	(t_J^i, t_J^j): 0.33 (t_E^i, t_E^j): 1
〈ファクシミリ送信‖data transmission〉-〈送信データ‖data transmission〉	異義	同義	異義	(t_J^i, t_J^j): 0.1 (t_E^i, t_E^j): 1	(t_J^i, t_J^j) : 0.13 (t_E^i, t_E^j) : 1	(t_J^i, t_J^j): 0.5 (t_E^i, t_E^j): 1

(a) SVM のみが同義と判定し正解

$\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle$	人手による同義・異義判定	ベースラインによる判定	SVMによる判定	f2: 編集距離類似度	f3: バイグラム類似度	f4: 同一の形態素(単語)数の割合
〈Pトランジスタ‖p transistor〉-〈P型トランジスタ‖p-type transistor〉	同義	異義	同義	(t_J^i, t_J^j):0.89 (t_E^i, t_E^j):0.71	(t_J^i, t_J^j): 0.75 (t_E^i, t_E^j): 0.65	(t_J^i, t_J^j) : 0.67 (t_E^i, t_E^j) : 0.50

表 4 に、SVM による異義判定結果を示す。ベースラインは、t_J^i - t_J^i が同一でなく、かつt_E^i - t_E^i が同一でないものとした。ベースラインに対し、SVM で適合率が最大となるモデルとの比較を行ったところ、再現率が下がったものの、適合率が有意水準1%で改善された(88.8%から93.7%に改善)。

同義判定および異義判定にて、SVM で適合率が最大となるモデルとベースラインを比較した。例(2) の場合は、t_J^i - t_J^i が同一なため、ベースラインで異義組である(<薬液,liquid>-<冷却媒体,liquid>)を同義と判定してしまった。一方、SVM では異義と判定することができた。SVM の素性である共起語一致数($f8$) は、t_E^i - t_E^i が一致していることから $f8(t_E^i - t_E^i)$ の値が高いものの、t_J^i - t_J^i の意味が異なるため値が低くなった。そのため、これらの素性が有効に働き、SVM は異義と判定できたと考えられる。例(4) の場合は、t_J^i - t_J^i が異なり、かつt_E^i - t_E^i が異なるためベースラインで同義組である(<Pトランジスタ, p transistor>-<P型トランジスタ,p-type transistor>)を異義と判定してしまった。一方、SVM では同義と判定することができた。SVMの素性である編集距離類似度、バイグラム類似度、同一形態素(単語) 数の割合が有効に働いたためと考えられる。

3.1.7 同義専門用語集合の同定

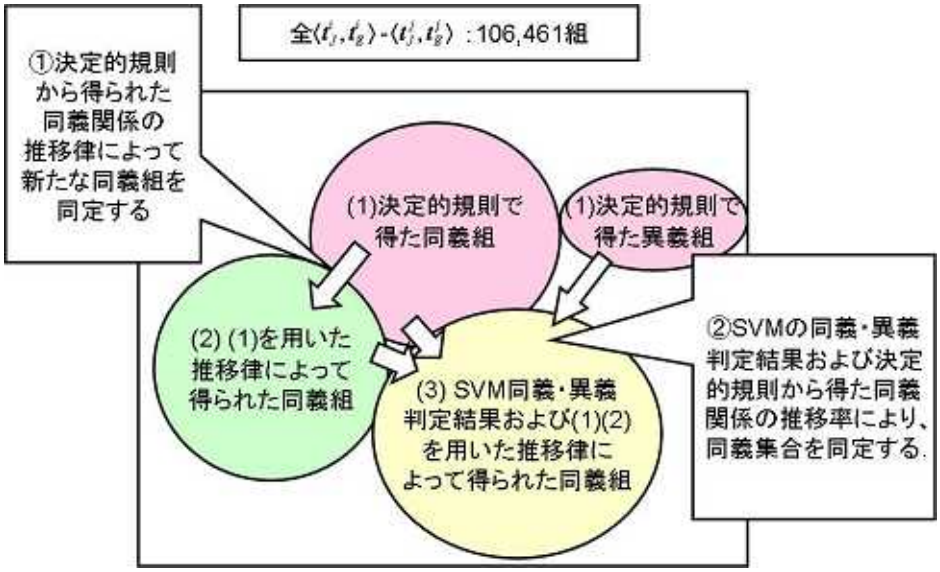


図 4　決定的規則および SVM 判定結果を用いた同義専門用語集合同定の流れ

図 4に、同義専門用語集合同定の流れを示す。3.1.5節に述べた決定的規則による同定結果および3.1.6節に述べたSVM による判定結果から得られた同義関係の推移律を用いて同義集合を同定した。まず、決定的規則から得られた同義関係の推移律によって新たな同義集合を同定する。さらに、SVM の同義、異義判定結果および決定的規則から得た同義関係の推移律により、同義集合を同定した。

3.1.7.1　同義組のための推移律

同義組の集合を *Syn*、異義組の集合を *NonSyn* とする。同義組集合 *Syn* から得られた同義関係の推移律を利用して、新たな同義組を追加する。以下の更新規則が新たに適用できなくなるまで同義組集合 *Syn* の更新を行う。また、更新規則の模式図を図 5の上図に示す。

$$\text{条件式}\quad \langle t_J^k, t_E^k\rangle\text{-}\langle t_J^i, t_E^i\rangle \notin Syn, \langle t_J^k, t_E^k\rangle\text{-}\langle t_J^i, t_E^i\rangle \notin NonSyn,\ \langle t_J^i, t_E^i\rangle\text{-}\langle t_J^j, t_E^j\rangle, \langle t_J^j, t_E^j\rangle\text{-}\langle t_J^k, t_E^k\rangle \in Syn$$

$$\text{更新式}\quad Syn \leftarrow \{\langle t_J^k, t_E^k\rangle\text{-}\langle t_J^i, t_E^i\rangle\} \cup Syn$$

$\langle t_J^k, t_E^k\rangle$ - $\langle t_J^i, t_E^i\rangle$が同義組集合および異義組集合に入っておらず、$\langle t_J^i, t_E^i\rangle$ - $\langle t_J^j, t_E^j\rangle$ と $\langle t_J^j, t_E^j\rangle$ - $\langle t_J^k, t_E^k\rangle$ の両方が同義の場合、$\langle t_J^k, t_E^k\rangle$ - $\langle t_J^i, t_E^i\rangle$を新たな同義組とし *Syn*に加える。これらの更新式は、決定的規則から得られた同義組に対して行う場合と、3.1.6 節で示すSVM から得られた同義組に対して行う場合がある。更新規則により *Syn*に追加される同義$\langle t_J^k, t_E^k\rangle$ - $\langle t_J^i, t_E^i\rangle$組の具体例を、図 5の下図に示す。

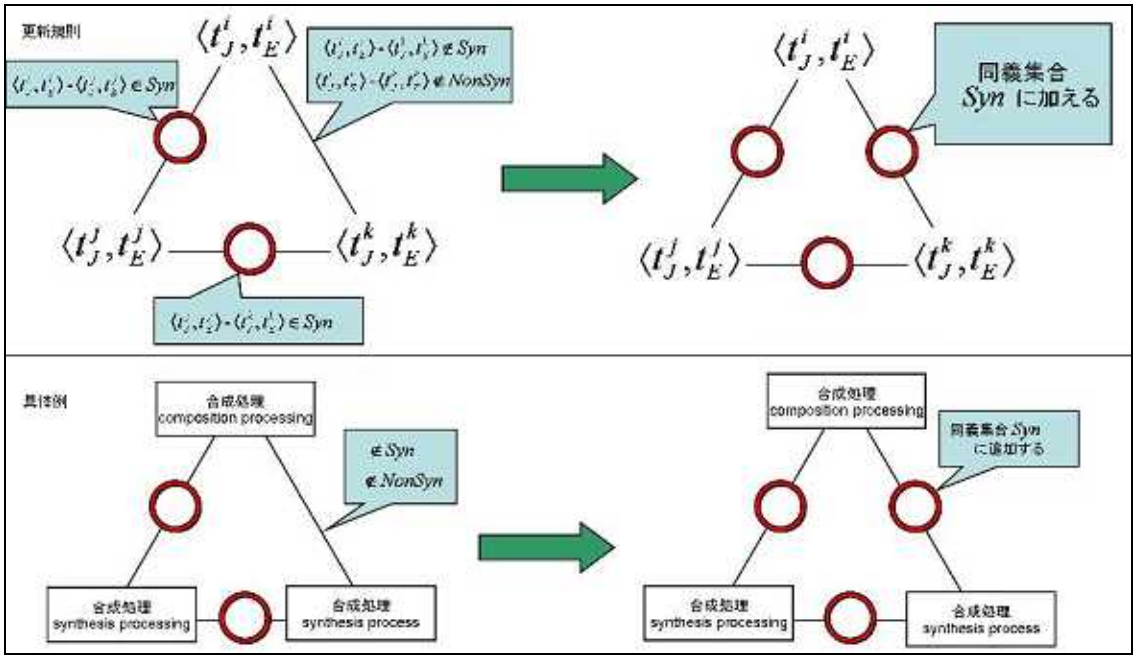


図 5　3.1.7.1 節の更新規則により *Syn* に追加される場合

3.1.7.2　SVM 同義判定で得られた同義組を用いた同義集合の更新規則

SVM 同義判定で得られた同義組$\langle t_J^k, t_E^k\rangle$-$\langle t_J^i, t_E^i\rangle$を用いた同義集合の更新規則を以下に示す。また、更新規則の模式図を図 6 の上図に示す。SVM の信頼度順に、以下の更新規則を適用した。

$\langle t_J^k, t_E^k\rangle$-$\langle t_J^i, t_E^i\rangle \notin Syn$, $\langle t_J^k, t_E^k\rangle$-$\langle t_J^i, t_E^i\rangle \notin NonSyn$

条件式　であり，式 (1) および (2) のいずれも満たさない．

(1)　$\exists j; \langle t_J^i, t_E^i\rangle$-$\langle t_J^j, t_E^j\rangle \in Syn$, $\langle t_J^j, t_E^j\rangle$-$\langle t_J^k, t_E^k\rangle \in NonSyn$

(2)　$\exists j; \langle t_J^i, t_E^i\rangle$-$\langle t_J^j, t_E^j\rangle \in NonSyn$, $\langle t_J^j, t_E^j\rangle$-$\langle t_J^k, t_E^k\rangle \in Syn$

(1)(2) を順に行う

更新式　(1) $Syn \leftarrow \{\langle t_J^k, t_E^k\rangle$-$\langle t_J^i, t_E^i\rangle\} \cup Syn$

(2) 6.6.1 節の推移律が適用できなくなるまで，Syn の更新を行う．

図 6 の(1) では、$\langle t_J^i, t_E^i\rangle$-$\langle t_J^j, t_E^j\rangle$が同義で、$\langle t_J^j, t_E^j\rangle$-$\langle t_J^k, t_E^k\rangle$が異義である。そのため、$\langle t_J^k, t_E^k\rangle$-$\langle t_J^i, t_E^i\rangle$は同義組として矛盾するので、更新式は適用せず *Syn* に加えない。同様に(2)では、$\langle t_J^i, t_E^i\rangle$-$\langle t_J^j, t_E^j\rangle$が異義で、$\langle t_J^j, t_E^j\rangle$-$\langle t_J^k, t_E^k\rangle$が同義であるため、$\langle t_J^k, t_E^k\rangle$-$\langle t_J^i, t_E^i\rangle$を *Syn* に加えない。具体例を、図 6 の下図に示す。一方、更新式の条件を満たす場合、$\langle t_J^k, t_E^k\rangle$-$\langle t_J^i, t_E^i\rangle$を *Syn* に加える。図 7 に *Syn* に追加される$\langle t_J^k, t_E^k\rangle$-$\langle t_J^i, t_E^i\rangle$組の具体例を示す。

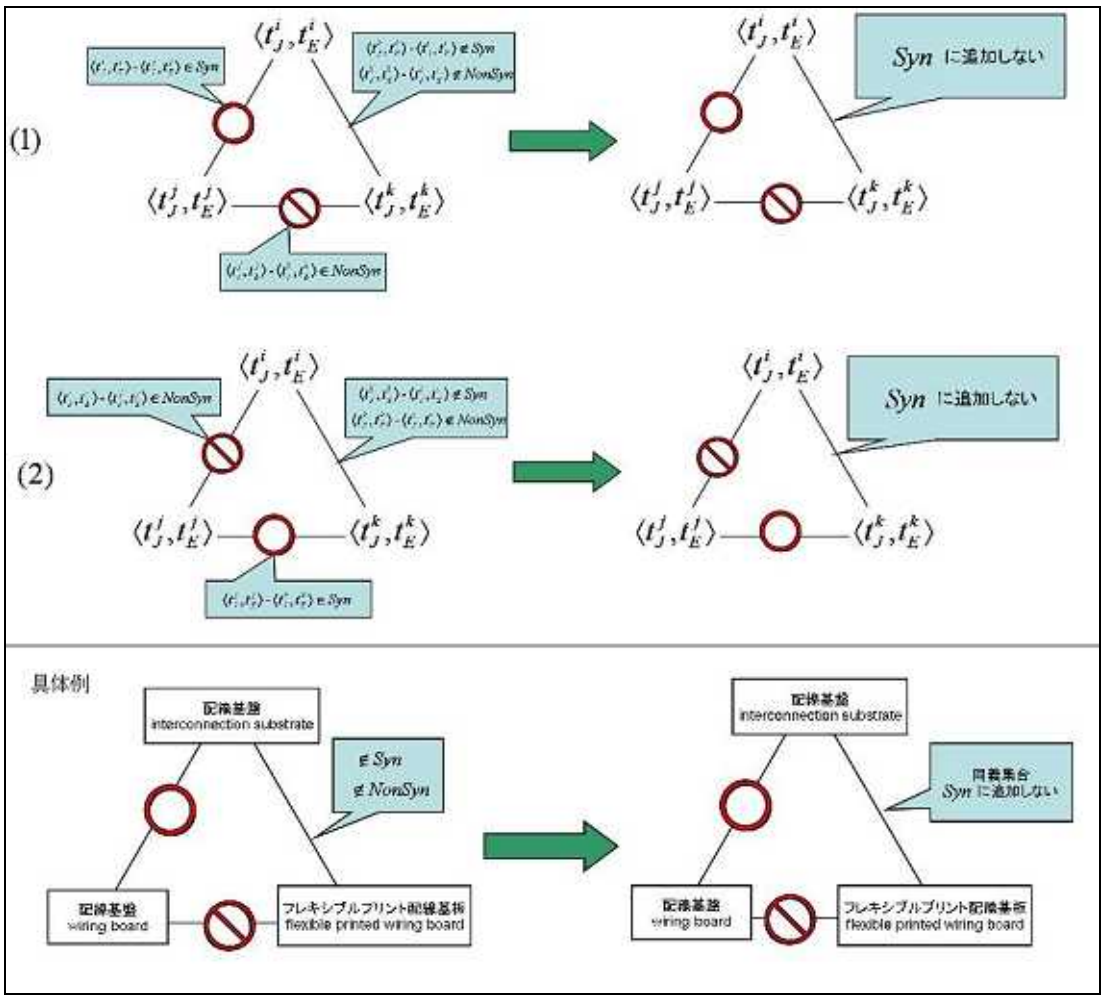


図 6　3.1.7.2 節の更新規則により *Syn* に追加されない場合

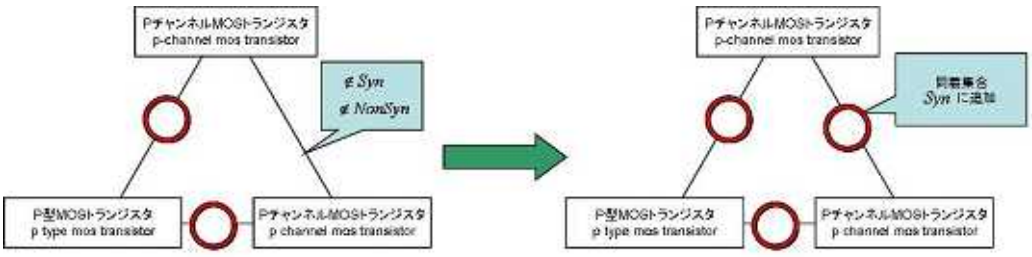


図 7　3.1.7.2 節の更新規則で *Syn* に追加される $\langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$ 組の例

3.1.7.3　SVM 異義判定で得られた異義組を用いた異義集合の更新規則

SVM 異義判定で得られた異義組 $\langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$ を用いた異義集合の更新規則を以下に示す。また、更新規則の模式図を図 8 に示す。

条件式　$\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle \notin Syn,\ \langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle \notin NonSyn$

更新式　$NonSyn \leftarrow \{\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle\} \cup NonSyn$

$\langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$ が同義組集合および異義組集合に入っていない場合、$\langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$ を新たな異義組とし *NonSyn* に加える。

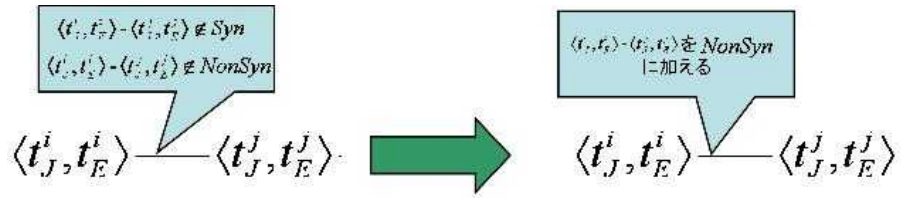


図 8　3.1.7.3 節の更新規則により *NonSyn* に追加できる場合

3.1.7.4　同定手順全体の流れ

1. 決定的規則によって得られた同義組を *Syn*、異義組を *NonSyn* とする。*Syn* に対し 3.1.7.1 節の推移律を用いることにより、*Syn* を更新する。
2. SVM 同義判定で得られた同義組候補集合の要素を $\langle \langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$ - *cf* - “＋”$\rangle$、SVM 異義判定で得られた異義組候補集合の要素を $\langle \langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$ - *cf* - “－”$\rangle$ と記述する。ここで、*cf* は SVM の信頼度を示す。それらの和集合を *Can* とし、要素を $\langle \langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$ - *cf* - *s* $\rangle$ と記述する。決定的規則および推移律によって得られた同義組を *Syn*、決定的規則から得られた異義組を *NonSyn* とし、以下の手順により、*Can* から得られた同義関係の推移律によって *Syn* および *NonSyn* を更新する。また、*Can* が空で要素を選別できない場合、更新を終了する。

   Can の中で、*cf* の値が最大の要素 $\langle \langle t_J^i, t_E^i \rangle$ - $\langle t_J^j, t_E^j \rangle$ - *cf* - *s* $\rangle$ を選別し、さらに *Can* から要素を除外する。*s* = ”＋” の場合、要素に対し 3.1.7.2 節に示す更新規則を用いて *Syn* を更新する。*s* = ”－” の場合、3.1.7.3 節に示す更新規則を用いて *NonSyn* を更新する。

3.1.7.5　評価結果

決定的規則および SVM による判定で得られた同義組および異義組、3.1.7.1 節の推移律、3.1.7.4 節の手順を用いて求めた同義集合、異義集合に対して行った評価結果を以下に示す。

1. 決定的規則から得た同義組および 3.1.7.1 節の推移律を用いた同義集合の同定

   決定的規則で得られた同義組の評価結果を、表 6 の「(a) 決定的規則によって同定した集合」の欄に示す。決定的規則で得られた同義組に対し、3.1.7.1 節の推移律を用いることにより同義組集合 *Syn* を同定する。表 6 の「(c) 決定的規則によって得られた同義組を用いた推移律により新たに得られた同義組と (a) の和集合」の欄に、得られた同義集合 *Syn* に含まれる同義組の評価結果を示す。(1) (a) 決定的規則によって同定した集合の結果と比較して、適合率を 100% に維持したまま、再現率を改善することができた (5.8% から 12.7% に改善)。

2. 決定的規則およびSVM による判定で得られた同義組および異義組、3.1.7.1節の推移律、3.1.7.4節の手順を用いた同義集合、異義集合の同定

決定的規則および推移律により得られた同義集合を *Syn* とし、決定的規則によって得られた異義組を *NonSyn* とする。SVM による判定で得られた同義組および異義組の評価結果を表 6 の「(b)SVM の判定結果を用いて生成した集合」に示す。(b-1) の結果は、同義判定の適合率が最大となる信頼度を用いた結果であり、(b-2) の結果は、同義判定のF 値が最大となる信頼度を用いた結果である。

表 6 の「(d) 6.6.4 節の手順により同定した集合」の欄に、同定した同義集合、異義集合の評価結果を示す。「(d-1) 入力:(c) および(b-1) の同義組、(a) および(b-1) の異義組」では、適合率が最も高くなる入力の組み合わせとして、表 6 に示す(c) および(b-1) の同義組、(a) および(b-1) の異義組を用いた。その結果、適合率94.8%、再現率36.6%で同義組の同定を行うことができた。また、「(d-2) 入力:(c) および(b-2) の同義組、(a) および(b-1) の異義組」では、F 値が最も高くなる入力の組み合わせとして、表 6 に示す(c) および(b-2) の同義組、(a) および(b-1) の異義組を用いた。その結果、適合率84.2%、再現率53.9%で同義組の同定を行うことができた。

表 6　同定した同義集合内に含まれる $\langle t_J^i, t_E^i \rangle$-$\langle t_J^j, t_E^j \rangle$ 組の評価結果(%)

(1). 決定的規則および SVM によって同定した同義組，異義組

	同義組			異義組		
	適合率	再現率	F 値	適合率	再現率	F 値
(a) 決定的規則によって同定した集合	100.0 (966/966)	5.8 (966/16615)	11.0	100.0 (2391/2391)	2.7 (2391/89846)	5.2
(b)SVM の判定結果を用いて生成した集合						
(b-1). SVM の信頼度 : 適合率が最大となる値を使用	93.2 (3617/3879)	23.1 (3617/15649)	37.0	93.7 (65488/69891)	74.9 (65488/87455)	83.2
(b-2). SVM の信頼度 : F 値が最大となる値を使用	71.0 (8922/12561)	57.0 (8922/15649)	63.3	((d-1) および (d-2) で使用せず)		

(2). 3.1.7.1 節の推移律および 3.1.7.4 節の手順を用いて更新した同義集合および異義集合

	同義組			異義組		
	適合率	再現率	F 値	適合率	再現率	F 値
(c) 決定的規則によって得られた同義組および推移律により新たに得られた同義組と (a) の和集合	100.0 (2118/2118)	12.7 (2118/16615)	22.6	(1)(a) 異義組から不変		
(d) 3.1.7.4 節の手順により同定した集合						
(d-1) 入力: (c) および (b-1) の同義組, (a) および (b-1) の異議組 (適合率が最大となる組み合わせ)	**94.8** (6088/6525)	36.6 (6088/16615)	52.8	94.2 (61892/65737)	68.9 (61892/89846)	79.6
(d-2) 入力: (c) および (b-2) の同義組, (a) および (b-1) の異議組 (F 値が最大となる組み合わせ)	84.2 (8954/10642)	53.9 (8954/16615)	**65.7**	94.1 (56373/59930)	62.7 (56373/89846)	75.3

3.1.8 おわりに

本論文では、対訳特許文から獲得された専門用語対訳対を用いて、同義専門用語集合の分析と同定を行った。評価実験においては、まず文字列の類似度や共起語の類似度を用いて、適用可能な範囲は小さいが100%の精度で同義専門用語を同定可能な決定的規則を作成した。この決定的規則により同義専門用語の5.8%を同定することができた。さらに、決定的規則によって同定できない同義専門用語に対し、SVM を用いた同義・異義判定を行った。その結果、決定的規則により同定できない同義専門用語に対して、適合率93.2%、再現率23.1%で同定を行うことができた。さらに、決定的規則およびSVM の判定結果を用いて、同義専門用語集合の同定を行った。その結果、同義集合に対して適合率94.8%、再現率36.6%で同定を行うことができた。

同義専門用語集合同定の研究で、機械学習を用いたものに[Tsunakawa08]らの手法がある。本研究と、Tsunakawaらの研究で大きく異なる点として、Tsunakawaらの研究は既存の辞書であるJST 辞書に含まれる対訳対に対して同義集合の同定を行うのに対し、本研究では対訳特許文から抽出した対訳対に対して同義集合の同定を行う点があげられる。本論文では、対訳特許文に含まれる共起語を用いることにより、同義判定の適合率を向上させた。

3.1.9 参考文献

[森下08] 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，Vol. 2008, No. (2008-NL-187), pp. 91-98 (2008).

[Morishita08] Y. Morishita, T. Utsuro, and M. Yamamoto. Integrating a phrase-based SMT model and a bilingual lexicon for human in semi-automatic acquisition of technical term translation lexicon. In *Proc. 8th AMTA*, pp. 153–162, 2008.

[Fujii08] Fujii, A., Utiyama, M., Yamamoto, M. and Utsuro, T.: Overview of the Patent Translation Task at the NTCIR-7 Workshop, *Proc. 7th NTCIR Workshop Meeting*, pp. 389-400(2008).

[Koehn07] Koehn, P., et al.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. 45th ACL, Companion Volume*, pp. 177-180 (2007).

[外池07] 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).

[Vapnik98] Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998).

[森下10] 森下洋平, 宇津呂武仁, 山本幹雄. 対訳特許文からの対訳専門用語獲得における同義専門用語集合の分析と同定. 言語処理学会第 16 回年次大会論文集.

[Tsunakawa08] Tsunakawa, T. and Tsujii, J.: Bilingual Synonym Identification with Spelling Variations, *Proc. 3rd IJCNLP*, pp.457-464 (2008).

３．２　中国語の同義語抽出の性能に関する調査

東京大学　範　暁蓉
二宮　崇

3.2.1　はじめに

人間はある一つの意味を表現するのに様々な表現方法をつかって表現するため、これらの表現を同じ意味として機械に自動的に認識させることは自然言語処理における究極の目標の一つである。そのような目標を実現するための重要な基礎技術の一つとして、同義語のリストを自動的に構築する同義語抽出と呼ばれる基礎技術が研究されている。同義語とは同じ意味を持つと考えられる単語のことであり、同義語抽出は機械翻訳や情報検索などの自然言語処理の様々なタスクにおいて非常に重要なタスクと考えられている。例えば、情報検索のタスクにおいては、ある検索語をキーワードとして検索する時、検索語に関する文章を十分収集するために、その検索語に関する文章だけではなく、検索語の同義語や、類義語に関する文章も検索することが要求される。しかしながら、今までの同義語抽出に関する多くの研究は英語を対象としており、中国語を対象とした研究はほとんど行われていない。

本稿では、中国語における同義語抽出の性能を評価するために、教師なし学習と教師あり学習の二つの手法を適用して、実験を行った結果を報告する。

本稿の構成は以下のようになっている。3.2.2 節では、従来の同義語抽出技術について説明する。3.2.3 節では、まず、中国語の同義語抽出実験に関する手法を説明して、次に、中国語コーパスから抽出を行った実験結果について報告する。3.2.4 では本稿の主旨をまとめ、今後の課題について述べる。

3.2.2　同義語抽出の技術に関する調査

同義語抽出技術では、一般に「Distributional Hypothesis」という仮説に基づき、単語の類似度を計算する。Distributional Hypothesis とは、「同じ文脈を持つ単語は、同じ意味を持つ傾向にある」というものである。単語類似度は次の二つの計算ステップで計算される。まず、各単語に対して、様々な文脈情報 (係り先と係り元と対象語の前後連続単語など) を抽出して、各単語の「素性ベクトル」を生成する。次に、単語の素性ベクトルを利用して、単語間の類似度を計算する。

素性の種類

同義語抽出のために様々な情報から素性を抽出することが考えられる。例えば、辞書やシソーラスの中の単語について、この単語の定義の中に出る単語を素性とする方法が Blodel と Sennelart により提案されている (Blondel & Sennelart, 2002)。WordNet のような意味辞書の中の単語の意味や意味辞書の構造を素性として利用する手法も提案されている (Jarmasz & Szpakowicz, 2003)。同義語抽出において一般によく使われるリソースは、単一言語の大規模コーパスであり、文脈情報を同義語抽出のための素性とする。文脈情報に関して、2 種類の文脈がよ

く用いられる。一つ目は、単語の文脈として、その単語の n 単語前後の単語列 (n-gram) を文脈とするものである。二つ目は、構文情報を用いて取り出した、単語の関係から文脈を取り出すものである。

類似度の計算

類似度計算の手法は、教師なし手法と教師あり手法の二つに分かれ、従来の多くの手法は教師なし手法に分類される。教師なし手法の中では、Cosine 距離、Dice 係数と Jaccard 係数の三つが代表的な手法である。Weeds は博士論文 (Weeds, 2003) の中で、これらの類似度アルゴリズムに関する詳細な説明を与えている。教師あり手法では、同義語の正解データから距離モデルのパラメータを学習する。教師有り手法は正解データを用いるため、一般に教師なし手法より高い精度を実現する。しかし、これらの学習には非常に多くの時間を要し、また、正解データの作業コストが高く、かつ、大量の訓練データが必要となる。そのため、同義語抽出に関する教師あり手法の研究はまだ少ない。清水ら (Shimizu ら, 2008) は新しい同義語抽出の教師あり学習手法を提案した。この手法は従来の教師あり手法よりも必要となる素性が少なくかつ精度が高い。この提案手法を用いると、英語の同義語抽出実験において教師なし手法より精度が約 15%高くなったことが報告されている。

3.2.3 中国語の同義語抽出の調査

中国語は分かち書きされていないという特徴を持つ言語であるため、中国語の同義語抽出は英語よりも難しいと一般に考えられている。しかし、中国語の同義語抽出に関する研究は少なく、中国語における従来の同義語抽出の性能は現在のところあまり知られていない。本稿は中国語の同義語抽出の性能を調査することを目的とする。今回、教師なしと教師あり手法の両方について評価し、特に教師あり手法における抽出性能を中心に評価する。

調査対象の手法

教師なし手法について、類似度の計算には三つの距離尺度、Cosine 距離、Euclidean 距離と Jaccard 係数を用いる。

Cosine 距離は次のように計算する。

$$\frac{w_1 \cdot w_2}{\|w_1\| \cdot \|w_2\|}$$

Euclidean 距離は次のように計算する。

$$\sqrt{\sum_{c \in C(w_1) \cup C(w_2)} \left(wgt(w_1, c) - wgt(w_2, c) \right)^2}$$

ただし、C は素性の集合であり、c は素性の一つである。

Jaccard 係数は次にように計算する。

$$\frac{\sum_{c\in C(w_1)\cup C(w_2)} \min(wgt(w_1,c), wgt(w_2,c))}{\sum_{c\in C(w_1)\cup C(w_2)} \max(wgt(w_1,c), wgt(w_2,c))}$$

教師あり手法には清水らの手法を用いる。

中国語の同義語抽出手法

実験の設定について述べる。4 年分の LDC2007T38 コーパス (2001 年 1 月~2004 年 12 月、単語数は 6 千万以上) から、教師なし手法と清水らの教師あり手法を用いて、中国語の同義語抽出実験を行った。

次に実験方法について述べる。 実験は下記の(1)〜(8)の手順により行った。

(1) コーパスを構文解析する。

文脈情報は N-gram と構文情報の二種類を用いた。予備実験では、構文情報を利用した方が良い結果がでたので、今回の実験には構文情報を素性として用いた。Stanford の中国語パーザーを用いて構文解析を行った。

(2) 構文解析の結果から、候補単語と素性を抽出する。

中国語パーザーの出力は全てをそのまま使えるほど精度が高いわけではないため、高い精度で解析されていると期待される解析結果を部分的に抽出し同義語抽出に用いた。全ての関係の中で、Dobj 関係の精度が一番高いため、候補単語と素性は Dobj 関係から生成した。頻度 30 以上の単語を候補単語として抽出した。

(3) 素性選択基準を利用して、素性の数を削減する。

教師あり手法にとって、素性の数は最も重要なことである。素性の数が多すぎると、計算時間および使用するメモリの点から学習が非常に困難となる。素性の選択は、二つの基準、頻度と文脈の重要度を用いて選択した。まず、頻度 30 以上の素性は有効な素性として抽出した。次に、多くの単語が利用する文脈は最も重要な文脈という文脈の重要度基準で選択した。文脈の重要度は次の式で計算する。

$$df(c) = \left|\{w \mid N(w,c) > 0\}\right|$$

ただし、w は候補単語で、c は素性で、$N(w,c)$ は w と c の共起頻度である。重要度が 30 以上の

素性は有効な素性として抽出した。選択する前の素性の数は 64,045 で、選択後 1,719 になった。

(4) 同義語辞書を利用して、訓練データとテストデータを生成する。

教師あり手法では、訓練データとテストデータが必要となる。中国語の同義語辞書である「同義語詞林」(Mei Gia-Chu et al. 1984)を用いて、訓練データとデストデータを生成した。同義語詞林は 1984 年に発表された中国語の同義語辞書である。この辞書は図 1 のような 5 階層の構造を持つ。レベル 1 からレベル 4 までは単語の分類で、レベル 5 には、12,193 セットの同義語セットと類語セットがある。訓練データとテストデータは同義語セットから生成された。

Level 1

num	cateory
A	人
B	物
C	時間和空間
D	抽象事物
E	特征
F	動作
G	心理活動
H	活動
I	現象與狀態
J	關聯
K	助語
L	敬語

Level 2

num	cateory
Aa	泛稱
Ab	男女老少
Ac	體態
Ad	籍屬
Ae	職業
Af	身份
Ag	狀況
Ah	親人
Ah	眷屬
Ai	輩次
Aj	關系
Ak	品性
Al	才識
Am	信仰
An	丑類
Ao	專名

Level 3

num	cateory
Aa01	人
Aa01	人民
Aa01	眾人
Aa02	我
Aa02	我們
Aa03	你
Aa03	你們
Aa04	他
Aa04	他們
Aa05	自己
Aa05	別人
Aa05	某人
Aa06	誰
Ab01	男人
Ab01	女人
Ab01	男女

Level 4

num	cateory
Aa011	人
Aa012	人民
Aa013	眾人
Aa021	我
Aa022	我們
Aa031	你
Aa032	你們
Aa041	他
Aa042	他們
Aa051	自己
Aa052	別人
Aa053	某人
Aa061	誰
Ab011	男人
Ab012	女人
Ab013	男女

Level 5

num	term
Aa011_01	人
Aa011_01	人士
Aa011_01	人物
Aa011_01	候選人
Aa011_01	人選
Aa011_01	士
Aa011_01	人氏
Aa011_02	人類
Aa011_02	生人
Aa011_03	人員
Aa011_03	口
Aa011_03	人口
Aa011_03	人手
Aa011_03	食指
Aa011_03	人丁
Aa011_03	丁口

図 1 同義語詞林の構造 (ChuRen Huang et al. 2005)

少なくとも 1 つの同義語を持つ単語を抽出し、1812 単語が抽出された。この 1812 単語の中、260 単語は 5 個以上の同義語をもつ。この 260 単語を 5 つのセットに分けて、4 つのセットを訓練データとして用い、1 つのセットはテストデータとして用いた。

清水らの実験では、訓練データのサイズを変えた場合の影響を測定する実験は行われていなかった。今回の実験では、訓練データの量がどの程度学習結果へ影響するかを調査するため、訓練データのサイズを 50, 100,150,200 と変えて、同義語抽出の精度を測定した。

(5) 訓練データから学習する。

(6) 学習された学習器を利用して、テストデータの距離を計算する。

(7) 教師なし手法でテストデータの類似度を計算する。

(8) 最後に、この二つの手法を比較する。

実験の結果

平均精度（Mean Mean Average Precision）、RKL(Average Rank of Last Synonym L)とTop1 の三つの尺度を用いて実験結果を評価した。

表 1 は今回の比較実験の結果である。Jaccard、Cosine と Euclid は教師なし手法で、Mahalanobis は教師あり手法である。結果を見ると、Jaccard 係数が教師なし手法の中で一番良い結果となった。教師あり手法については、訓練データの量が少ないと、教師なし手法の Jaccard 係数よりも精度が低くなることがわかった。実験結果の中では訓練データのサイズが 50 の場合において、Jaccard 計数よりも精度が低くなっている。訓練データの量を増やすと、教師有り学習の方が精度が高くなり、訓練データのサイズが 100 になると全ての教師なし手法よりも精度が良くなっている。Top1 の尺度については訓練データのサイズが 150 である場合が最も精度が高かったが、MAP と RKL においては、訓練データのサイズが最大である 200 の場合において最も精度が高くなった。また、教師あり手法の最高の Top1 の値は 41%であることが分かる。

表 1　実験の結果

教師なし手法	Metric		MAP	RKL(1181)	TOP1
	Jaccard		**0.14455**	**503.60784**	**0.25490**
	Jaccard L2		0.13196	508.07843	0.23529
	Cosine		0.02064	867.94118	0.03922
	Euclid		0.03758	757.54902	0.05882
	Euclid L2		0.05488	782.94118	0.13725
教師あり手法	手法	訓練データのサイズ			
	Maharanobis	50	0.13694	582.15686	0.23529
	L2	100	0.19676	485.74510	0.27451
		150	0.32059	377.56863	**0.41176**
		200	**0.32526**	**344.13725**	0.39216

3.2.4　まとめと今後の課題

本稿では、中国語の同義語抽出の性能調査を行った。教師なし手法と教師あり手法の二つの手法を比較実験によって評価した。教師あり手法は教師なし手法よりも精度が高いことが実験によって示されたが、まだ実用に供する精度には至っていない。今後は、実用に用いられうる中国語の同義語抽出に向けて研究を行いたい。

参考文献

Blondel V. D. and Sennelart P. 2002. Automatic extraction of synonyms in a dictionary. In Proceeding of the SIAM Workshop on Text Mining.

ChuRen Huang, XiangBing Li and JiaFei Hong. 2005. The Robustness of Domain Lexico-Taxonomy: Expanding Domain Lexicon with CiLin.

Julie Elizabeth Weeds, 2003. Measures and Applications of Lexical Distributional Similarity. Ph.D. thesis, University of Sussex, September.

Mario Jarmasz and Stan Szpakowicz, 2003. Roget's Thesaurus and Semantic Similarity. In Proceedings of RANLP 2003. pages 212-219.

Mei, Gia-Chu. 1984. Cilin- Thesaurus of Chinese words. (in Chinese) Hong Kong.

Miller, Uri. 1997. Thesaurus construction: problems and their roots. Information Processing and Management 33(4), pages. 481-493.

Nobuyuki Shimizu, Masato Hagiwara, Yasuhiro Ogawa, Katsuhiko Toyama and Hiroshi Nakagawa. 2008. Metric Learning for Synonym Acquisition. In Proceedings of COLING 2008, pages 793-800.

３．３　コンパラブルコーパスを用いた訳語選択

静岡大学　綱川　隆司
梶　博行

3.3.1　はじめに

本稿では，非パラレルコーパスであるコンパラブルコーパスを訳語選択や統計的機械翻訳に用いる手法を確立するため，コンパラブルコーパスを用いた訳語選択手法において，手掛かりとなる抽出する関連語の範囲の比較検討と，統計的機械翻訳の分野適応への応用について述べる．

翻訳において，複数の意味を持つ語の翻訳を行う場合，それぞれの意味に対応する訳語が異なるため，語義曖昧性の解消を行い，適切な意味を持つ訳語を選んで出力する必要がある．機械翻訳においてこの問題に対処するには，入力文やその分野等の情報を考慮しなければならない．しかし，ルールに基づく機械翻訳システムでは，どの訳語を選択するかのルールを基本的に人手で記述しなければならず，多大な労力を要する上，相互に干渉する複雑なルールを管理するのも困難である．

Kaji and Morimoto (2002) では，翻訳対象語の周辺に現れやすい語と訳語候補との間のスコアの計算を行い，スコアの高い語を関連語として選択し，訳語選択のための「関連語－訳語関連行列」を求める．このスコアを求めるための単語間指標として相互情報量をはじめいくつかの指標を用いることで関連語の範囲を変化させ，実際に訳語選択を行って比較および評価を行う．

一方，統計的機械翻訳 (SMT) などのテキストデータに基づく方法では，データから得られた確率や言語モデル等の情報により訳語の選択を行っている．従って，対象分野のパラレルコーパスを用意するだけで，その分野の語彙や言い回しに適応した翻訳モデルを学習できるという特徴をもつ．しかし，大規模なパラレルコーパスが利用できる分野は限られるという問題がある．

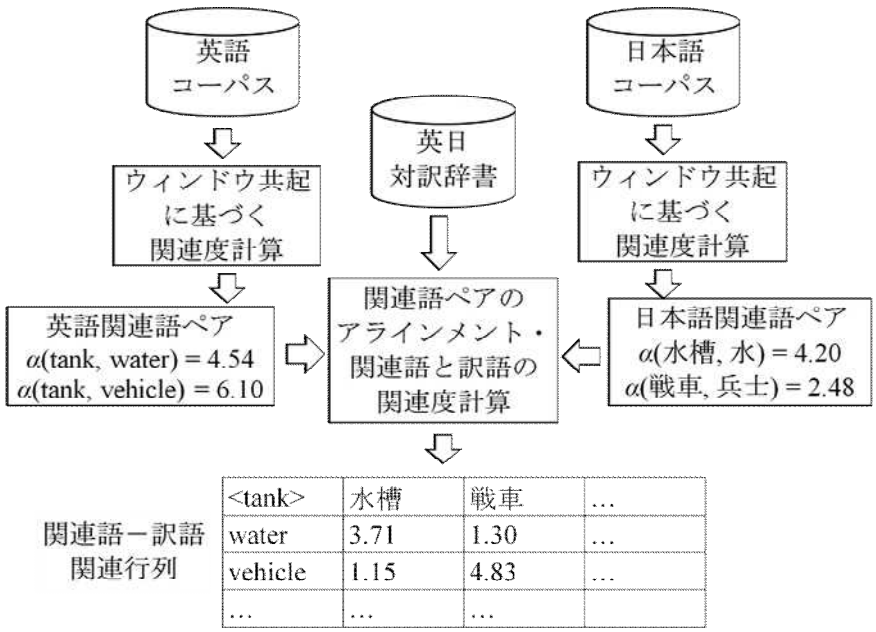
この問題を解決するため，コンパラブルコーパスから翻訳モデルを学習する方法を提案する．さらに，対象分野のコンパラブルコーパスから提案方法で学習した翻訳モデルを他の分野のパラレルコーパスから従来方法で学習した翻訳モデルを組み合わせ，パラレルコーパスが存在しない分野に SMT を適用する方法を示す．

3.3.2　関連語－訳語関連行列

図 1 に関連語－訳語関連行列の計算手法を示す．

まず入力言語および出力言語それぞれの単言語コーパスに含まれる名詞を列挙し，全ての名詞対について共起頻度に基づく指標 α を計算する（3.3.3 節参照）．各名詞について，指標が高い語から順に関連語として抽出する．

次に，関連語－訳語関連行列を以下のように計算する．「相互に関連のある関連語は同じ訳語を支持する」という仮説に基づき，対象語 f の第 i 関連語 $f'(i)$ と，f の j 番目の訳語 $e(j)$ の関連度 $C_f(f'(i), e(j))$ を以下の反復計算で得る．



<tank>	水槽	戦車	…
water	3.71	1.30	…
vehicle	1.15	4.83	…
…	…	…	

図 1: 関連語－訳語関連行列の計算

$$C_f^{(n)}(f'(i), e(j)) = \alpha(f'(i), f) \times \frac{\sum_{f'' \in A(f, f'(i))} \alpha(f'(i), f'') \cdot C_f^{(n-1)}(f'', e(j))}{\max_k \sum_{f'' \in A(f, f'(i))} \alpha(f'(i), f'') \cdot C_f^{(n-1)}(f'', e(k))},$$

ただし，n は反復計算のサイクル，$A(f, f'(i))$は対象語 f と関連語 $f'(i)$に共通の関連語の集合である．すなわち，$A(f, f'(i)) = \{f'' \mid \alpha(f, f'') \geq \theta, \alpha(f'(i), f'') \geq \theta\}$.

反復計算の初期値は以下の式で求める．

$$C_f^{(0)}(f'(i), e(j)) = \begin{cases} \dfrac{\alpha(f'(i), e(j))}{\sum_k \alpha(f'(i), e(k))} & (\Sigma_k \alpha(f'(i), e(k)) \neq 0), \\ 0 & (\text{それ以外}) \end{cases}$$

$$\alpha(f'(i), e(j)) = \begin{cases} 1 & (\exists e', (f, f'(i)) \approx (e(j), e')) \\ 0 & (\text{それ以外}) \end{cases}$$

ここで，$(f, f'(i)) \approx (e(j), e')$は，$f$ と $e(j)$，および $f'(i)$と e'がそれぞれ対訳辞書に訳語対として存在することを示す．

以上から，曖昧性なく対訳関係にある関連語ペアの組が種となって，関連語と訳語の間の関連度が反復計算される．

訳語選択は以下のようにして行う．ある文$\mathbf{f} = f_1 f_2 \cdots f_I$に含まれる単語$f_i$に対して訳語 $e(1), e(2), \ldots, e(J)$があるとき，訳語 $e(j)$に対するスコアを以下で定義する．

$$\mathrm{Score}(f_i, e(j)) = \sum_{\substack{l=i-w_1 \\ (l \neq i)}}^{i+w_1} r(i, l) C_f(f_l, e(j)),$$

$$r(i,l)=\begin{cases}1 & (|i-l|\leq w_0)\\ \dfrac{1}{\sqrt{|i-l|}} & (\text{それ以外})\end{cases}$$

ただし，w_0, w_1はウィンドウサイズとする．また，ある$f_l, e(j)$の組について$C_f(f_l, e(j))$の値が求められていない場合は 0 を代入する．

訳語のうち，最もスコアが高い訳語を選択する．また，全ての訳語についてスコアが 0 の場合は出力なしとする．

3.3.3 抽出する関連語の範囲の最適化

本研究では，訳語選択の手掛かりとなる関連語－訳語関連行列における関連語の範囲を，共起頻度に基づく指標および共起頻度の計算方法の観点から比較検討する．

【共起頻度に基づく指標】

コーパスに含まれる名詞 x, y について，x と y が内容語 $w-1$ 語以下を挟んで出現するとき，x と y が共起すると定義する（wはウィンドウサイズ）．x と y の出現回数をそれぞれn_1, n_2，共起回数を m，コーパスに含まれる単語数を N，2-グラムの総数を M とする．共起頻度に基づく指標αに以下を用いる．

相互情報量 (MI) (Church and Hanks, 1990)

$$\mathrm{MI}(x,y)=\log_2\frac{m/M}{(n_1/N)(n_2/N)}$$

対数尤度比 (LLR) (Dunning, 1993)

$$\mathrm{LLR}(x,y)=-2\big(\mathrm{logL}(m,n_1,r)+\mathrm{logL}(n_2-m,N-n_1,r)-\mathrm{logL}(m,n_1,r_1)-\mathrm{logL}(n_2-m,N-n_1,r_2)\big);$$

$$\mathrm{logL}(k,n,r)=k\log_2 r+(n-k)\log_2(1-r),$$

$$r_1=\frac{m}{n_1}, r_2=\frac{n_2-m}{N-n_1}, r=\frac{n_2}{N}.$$

t-スコア (TScore) (Church et al., 1991)

$$\mathrm{TScore}(x,y)=\frac{m-n_1n_2/N}{\sqrt{m}}$$

Dice 係数 (Smadja, 1993)

$$\mathrm{Dice}(x,y)=\frac{2m}{n_1+n_2}$$

Jaccard 係数 (Smadja et al., 1996)

$$\mathrm{Jaccard}(x,y)=\frac{m}{n_1+n_2-m}$$

Pearson's χ^2指標 (Manning and Schütze, 1999)

$$\chi^2(x,y)=\frac{N(mN-n_1n_2)}{n_1n_2(N-n_1)(N-n_2)}$$

また，複数の指標の特長を組み合わせるため，相互情報量と対数尤度比，および相互情報量とt-スコアを組み合わせた指標を以下のように定義する．

MI&LLR

ある名詞 x について，各名詞 y の相互情報量，対数尤度比での順位をそれぞれ r_1, r_2 とする．各名詞を $\max(r_1, r_2)$ の昇順で並び替え，上位 t 個以外を取り除く．指標の値として相互情報量の値を用いる．

MI&TScore

MI&LLR について，対数尤度比を t-スコアで置き換えたもの．

【共起頻度の計算方法】

出現回数が少ない語は共起する名詞も少なくなり，ウィンドウサイズが一定の条件下ではデータが疎になる．本研究では出現回数に応じてウィンドウサイズを変化させて共起回数を計算する以下の２つの手法について比較を行う．

ウィンドウ拡張

出現頻度 n の名詞について，ウィンドウサイズ w' を以下の式で定義する．

$$w' = \begin{cases} \lfloor w + \log_2(n_t - n) + 1 \rfloor & (n < n_t) \\ w & (n \geq n_t) \end{cases}$$

ただし，n_t は出現頻度の閾値とする．

ウィンドウ拡張（頻度重み付けあり）

ウィンドウ拡張後，２つの名詞が距離 d で共起する際，共起回数を 1 回の代わりに重み付きの回数 $\min(w/d, 1)$ 回をカウントする．

3.3.3.1 実験

英語および日本語のコーパスから関連語－訳語関連行列を各手法により求め，英文に含まれる各名詞の訳語を出力して評価する実験を行った．

関連行列を計算するためのコーパスには，English Gigaword コーパスの New York Times（2004 年, 277MB）および毎日新聞コーパス（2004 年,140MB(UTF-8)）を用いた．関連行列の反復計算に用いる対訳辞書には EDR 電子化辞書の英日・日英対訳辞書，EDICT (Breen, 1995)および英辞郎を組み合わせたものを用いた．評価対象として New York Times（2005 年 1 月）の 157 パラグラフに含まれる延べ 1448 語に対して訳語の出力を行った．1448 語のうち，周辺に手掛かりとなる関連語が一語も出現しない場合，および関連行列が存在しない語（191 語[1]）については出力なしとした．

訳語の出力結果に対して，正解（1 点），一部正解（0.5 点）および不正解・出力なし（0 点）の 3 段階で人手による評価を行った．評価は 1 語につき 2 名で行い平均を最終的なスコアとした．実験に用いたパラメータは以下の通りである．

$$w = 10, n_t = 20, t = 400, w_0 = 5, w_1 = 25$$

3.3 節で述べた各手法を適用した結果を表 1 に示す．Dice 係数および Jaccard 係数を用いた場合に最も高いスコアを得た．

表 1: 訳語選択実験結果

手法	平均スコア	出力語	関連語なし
相互情報量	0.31	714	543
対数尤度比	0.43	1190	67
t-スコア	0.41	1203	54
Dice	0.49	1166	91
Jaccard	0.48	1166	91
Pearson's χ^2	0.26	714	543
MI&LLR	0.30	690	567
MI&TScore	0.12	359	408
ウィンドウ拡張	0.31	717	540
頻度重み付け	0.31	716	541

相互情報量や Pearson's χ^2 指標を用いるのに比べ, 他の指標を用いることで関連語が存在しないために出力できなかった語が大幅に減少した．相互情報量では低頻度語に対して高い値が割り当てられる傾向があるため，訳語選択の際に手掛かりとなる周辺の語と関連語が一致しないためと考えられる．

平均スコアでは Dice 係数および Jaccard 係数が最も高く，およそ 39%の語について正解の訳語を選択できた．また，出力語数では t-スコアを用いた場合で最も多かった．

相互情報量を対数尤度比と組み合わせる手法では，スコアの変化はみられなかった．また，t-スコアとでは悪化した．組み合わせに用いた閾値，および Dice 係数など他の手法との組み合わせは今後の課題である．

共起ウィンドウの拡張，および頻度の重み付けについては，結果にほぼ変化が見られなかった．ウィンドウ拡張の対象となった低頻度語が結果にほぼ寄与できなかったと考えられるが，より広いウィンドウや分野情報の導入，また Dice 係数での拡張が改善案として挙げられる．

3.3.3.2 関連研究

単言語における語義曖昧性解消は，辞書やコーパス等のデータを使って教師なし学習を行う手法が提案されてきた (Ide and Veronis, 1998)．品詞等の文法的情報，構文的関係にある語，および周辺に共起する分野に関する語を曖昧性解消の手掛かりとして用いている．本研究ではこれらの一部を利用して，単語の翻訳の曖昧性解消に用いている．

Li and Li (2002) は単語の翻訳の曖昧性解消を対訳辞書の対応付けをもとにブートストラッピングによって分類器を構築し行っている．本研究では辞書の対応関係は反復計算の種として用いコーパスから求めた手掛かり語を考慮に加えている．

Vickrey et al. (2005) は統計的機械翻訳に文脈を考慮した訳語選択を素性として導入し，訳語の選択を試みている．文全体の統計的機械翻訳システムへの導入が大きな課題である．

管理	control	management
システム	**1.451**	1.247
管制	1.199	**1.751**
装置	**1.633**	0.211
端末	0.139	**1.502**
	…	…

⇩

擬似翻訳確率の計算

⇩

P_{pseudo}(control|**管理**)=0.700　P_{pseudo}(management|**管理**)=0.300

図 2: **名詞の擬似翻訳確率の推定**

3.3.4　訳語選択手法の統計的機械翻訳の分野適応への応用

本節では，3.3.2 節で求めた関連語－訳語関連行列から名詞の翻訳確率を擬似的に推定することで，統計的機械翻訳に必要なフレーズテーブルをコンパラブルコーパスから構築する．翻訳対象分野のパラレルコーパスがない場合には，このフレーズテーブルによってその分野に適応した翻訳結果が出力されることが期待される．

3.3.4.1　名詞の擬似翻訳確率の計算

パラレルでないコーパスから名詞間の翻訳確率を求めるため，3.3.2 節で求めた関連語－訳語関連行列から名詞間の“疑似翻訳確率”を推定する手法を提案する．

関連語－訳語関連行列において，対象語の各関連語はそれとの関連度が最大の訳語を支持すると考え，対象語 f の第 j 訳語 $e(j)$への擬似翻訳確率 $P_{pseudo}(e(j)|f)$ を次式で定義する．

$$P_{pseudo}(e(j)\mid f)=\frac{(\mid S(e(j))\mid+\varepsilon)}{\sum_k(\mid S(e(k))\mid+\varepsilon)}.$$

ここで，$S(e(j))$は訳語 $e(j)$ を支持する関連語の集合である．すなわち，

$$S(e(j))=\left\{f'(i)\mid C(f'(i),e(j))>\max_{k\neq j}C(f'(i),e(k))\right\}.$$

なお，εは，$S(e(j))$が空集合であっても擬似翻訳確率が 0 にならないようにするための微小な定数値である．4.の評価実験ではε=0.025 とした．

図 2 には，対象語“管理”に対する関連語－訳語関連行列の一部が例示されている．関連語“システム”,“装置”は訳語“control”を，関連語“管制”,“端末”は訳語“management”をそれぞれ支持し，“control”と“management”への擬似翻訳確率がそれぞれ 0.70，0.30 と推定されている．

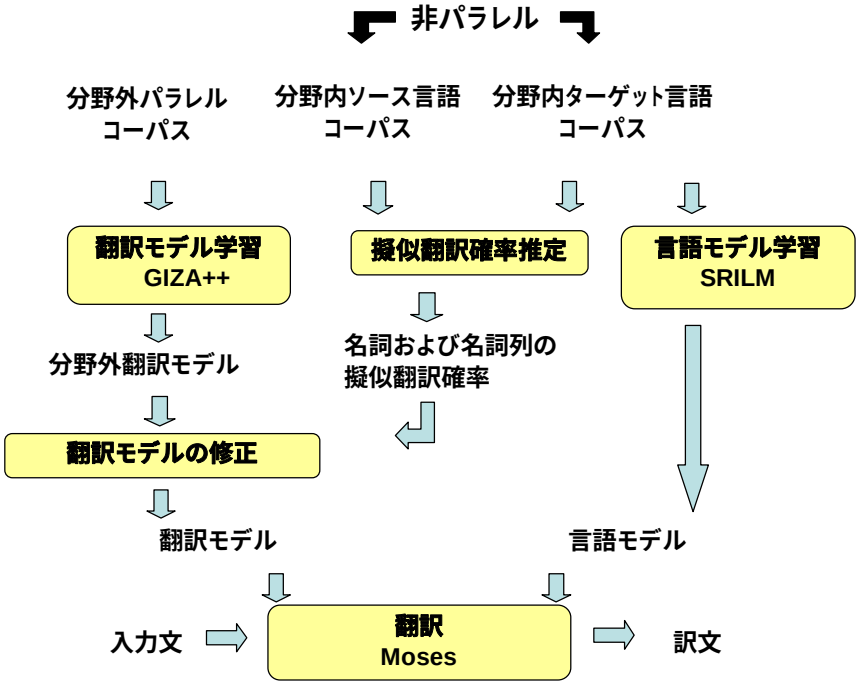


図 3: 擬似翻訳確率を用いた SMT の分野適応

3.3.4.2 名詞列の擬似翻訳確率

関連語－訳語関連行列は対訳辞書に含まれる対象語とその訳語について計算されるので，名詞の疑似翻訳確率だけでは 2 語以上からなるフレーズの擬似翻訳確率を計算することはできない．フレーズを名詞列に限り，2 語以上からなるフレーズの擬似翻訳確率を以下の(1)～(3)のステップにより推定する（図 3 参照）．

(1) 名詞列の抽出とアラインメント

ソース言語とターゲット言語それぞれのコーパスから，２語以上の名詞列を抽出しその頻度をカウントする．なお，より長い名詞列の部分列となっている名詞列も頻度カウントの対象とする．次に，対訳辞書を参照しながら，構成要素の間に対訳関係が成立する名詞列を対応づける．すなわち，f_1 と e_1, f_2 と e_2, .., f_n と e_n それぞれの組の間に対訳関係が成立するとき，２つの名詞列 $F=f_1f_2\ldots f_n$ と $E=e_1e_2\ldots e_n$ を対応づける．

(2) 擬似翻訳確率の計算

(1)の結果，名詞列 $F=f_1f_2\ldots f_n$ と m 個の名詞列 $E(j)=e_1(j)e_2(j)\ldots e_n(j)$（$j$=1, 2, .., m）のアラインメントが得られたとする．このとき，擬似翻訳確率を名詞列 $E(j)$の出現頻度に基づく推定値と構成要素間の擬似翻訳確率に基づく推定値の幾何平均として求める．すなわち，

$$P_{pseudo}(E(j)\mid F)=\sqrt{\frac{\#(E(j))}{\sum_k \#(E(k))}}\sqrt{\frac{\prod_i P_{pseudo}(e_i(j)\mid f_i)}{\sum_k \prod_i P_{pseudo}(e_i(k)\mid f_i)}}$$

ここで，$\#(E(j))$は名詞列 $E(j)$の出現頻度を表す．名詞列のアラインメントがすべて正しければ，$E(j)$の出現頻度の比率を翻訳確率とするのがよい．構成要素間の擬似翻訳確率に基づく推定値

との幾何平均をとる理由は，名詞列のアラインメントの誤りの影響を小さくするためである．対訳辞書はさまざまな文脈で成立する可能性のある対訳関係を含んでいるため，対訳辞書を介した名詞列のアラインメントでは誤ったアラインメントが得られることがある．しかし，誤ったアラインメントの場合，構成要素間の擬似翻訳確率もすべてが大きな値をもつ可能性は小さい．上の式によれば，訳語として正しくない名詞列に大きな翻訳確率を与えることを防ぐことができると考える．

3.3.4.3 擬似翻訳確率を用いた統計的機械翻訳の分野適応

分野外のパラレルコーパスから学習した翻訳モデルを分野内のコンパラブルコーパスを用いて分野に適応させる．図 3 に示すように，GIZA++ (Och and Ney, 2003) 等により分野外のパラレルコーパスから学習した翻訳確率と提案方法により分野内のコンパラブルコーパスから推定した擬似翻訳確率の平均をとる．一方の方法で翻訳確率が推定できないフレーズ対については，他方の方法で推定された値をそのまま採用する．なお，ターゲット言語の言語モデル（N グラム確率）は，SRILM (Stolcke, 2002) を用いて分野内のターゲット言語コーパスから学習する．また，デコーダとして Moses を利用する．

3.3.4.4 評価実験

提案方法により分野適応させた SMT（提案方法 1）と分野外パラレルコーパスから学習した翻訳モデルをそのまま用いた SMT（従来方法）の比較実験を行った．なお，英単語の小文字化は行わなかった．

実験に使用したコーパスと対訳辞書は次のとおりである．

(1) トレーニングコーパス

(a) 分野外パラレルコーパス：Japio の 2003 年の日英特許抄録の物理分野（20,000 抄録(日 5.32MB, 英 4.54MB)）．

(b) 分野内非パラレルコーパス：JST の 1981 年から 2005 年までの日英科学技術文献抄録[1]の基礎化学分野（日 151,958 抄録(90.8MB), 英 102,730 抄録(64.9MB)）．ただし，(2)のテストコーパスとして抽出した部分を除外．

(2) テストコーパス

(1)の(b)のコーパスから抽出した対訳文 1000 文．対訳辞書を参照して抄録対に含まれる日英の文の類似度を計算し(Utiyama and Isahara, 2007)，類似度の高いペアを抽出した．

(3) 対訳辞書

EDR 辞書, 英辞郎, EDICT から名詞のみを抽出してマージした辞書を使用した．日本語が 163,247 語，英語が 93,727 語，日英の対訳関係が 333,656 対含まれる．

翻訳方向は日本語から英語とし，テストコーパス中の日本語文を翻訳した．テストコーパ中の英文はレファレンス訳として利用した．従来方法として次の２つのケースを実行した．

- 従来方法：分野外パラレルコーパスから学習した翻訳モデルをそのまま使用する．
- 従来方法+辞書：対訳辞書の全ての訳語に一様な擬似翻訳確率を与え，分野外パラレルコーパスから学習した翻訳確率との平均をとる．

表 2: 提案手法 1 の結果

	トレーニングコーパス	BLEU
従来方法	分野外パラレルコーパス（物理）	0.1142
従来方法+辞書		0.1294
提案方法(i)	分野内非パラレルコーパス（基礎化学）	0.1330
提案方法(ii)		0.1319
提案方法(iii)		0.1321
提案方法(iv)		0.1327

表 3: 提案手法 2 の結果

	トレーニングコーパス	BLEU
従来方法	分野内パラレルコーパス（基礎化学）	0.1637
従来方法+辞書		0.1632
提案方法(i)	分野内非パラレルコーパス（基礎化学）	0.1682
提案方法(ii)		0.1670
提案方法(iii)		0.1678
提案方法(iv)		0.1671

提案方法については，トレーニングコーパスとして使用する分野内非パラレルコーパスの量が異なる４つのケースを実行した．

(i) (1)の(b)のコーパス全体を使用．

(ii) (1)の(b)のうち日本語抄録は半分を使用．

(iii) (1)の(b)のうち英語抄録は半分を使用．

(iv) (1)の(b)のうち日本語抄録，英語抄録とも半分を使用．

また，分野外パラレルコーパスを分野内パラレルコーパスに置き換えた実験（提案方法 2）を行った．これは，パラレルコーパスは小規模なものしか利用できないがコンパラブルコーパスは大規模なものが利用できる分野においても提案方法が有効であることを示すためである．分野内パラレルコーパスとして，テストコーパスの作成と同様の方法で(1)の(b)のコーパスから抽出した対訳文 20000 文(日 3.61MB,英 3.17MB)を使用した．

各方法による翻訳結果に対して BLEU スコア (Papineni et al., 2002) (4-gram) を算出した．提案方法 1 と提案方法 2 の結果をそれぞれ表 2 と表 3 にまとめた．

実験の結果から，以下の結論を得た．

(1) 提案方法 1 の結果から，提案方法が分野適応に有効であるといえる．なお，単純に従来方法の翻訳モデルに対訳辞書を追加しただけの場合と比較しても BLEU スコアは向上している．

(2) 提案方法 2 の結果から，パラレルコーパスが分野外の場合だけでなく，パラレルコーパスが分野内の場合でも，提案方法により翻訳精度が向上するといえる．分野内パラレルコーパスを利用する場合は，分野外パラレルコーパスを利用する場合と異なり，対訳辞書を組合せただけでは精度は向上しないこともわかった．

(3) コンパラブルコーパス全体を利用した場合と，日英とも半分にした場合を比較すると前者の BLEU スコアのほうが高い．提案方法の効果は一般的にコーパスの量に応じて大きくなると考えられる．しかし，提案方法 1 では，英語抄録のみを半分にした場合の BLEU スコアがさらに高くなっており，より多くのケースについて比較することが必要である．

3.3.4.5 関連研究

Koehn and Knight(2000)は，EM アルゴリズムを用いて非パラレルコーパスから翻訳確率を求め

る方法を提案している．この方法ではターゲット言語コーパス中の訳語の出現頻度に強く影響された結果が得られる．これに対し，提案方法ではソース言語コーパス中の対象語の語義の分布を反映した結果が得られる．

また，Wu et al. (2008) は，対象分野の対訳辞書を用いて，分野外のパラレルコーパスから学習した翻訳モデルを対象分野に適応させる方法を提案している．しかし，対訳辞書は翻訳確率の分野適応に十分な情報を含んでいないと思われる．

3.3.5 おわりに

本稿ではコンパラブルコーパスから構築した関連語－訳語関連行列による訳語選択手法により，以下の２点について実験を行った．

(i) 訳語選択のための関連語の抽出範囲の最適化のための比較検討

(ii) 関連語－訳語関連行列を用いた統計的機械翻訳の分野適応

(i) において，共起頻度に基づく指標では Dice 係数および Jaccard 係数を用いた場合に最もよい結果が得られた．また，指標の組み合わせや共起ウィンドウの拡張手法では性能の向上が見られなかった．

今後の課題としては，Dice 係数を中心とした指標の組み合わせや低頻度語への対処による性能の改善，および文全体の機械翻訳システムへの導入と文全体での翻訳結果の評価が挙げられる．

一方，(ii) については，コンパラブルコーパスから得た関連語－訳語関連行列を用いて，名詞や名詞列の擬似翻訳確率を推定する方法を提案した．提案方法を用いて分野内コンパラブルコーパスから推定した擬似翻訳確率を分野外パラレルコーパス（または分野内小規模パラレルコーパス）から推定した翻訳確率と併用することにより，SMT の BLEU スコアが向上することを実験により確認した．今後の課題として以下の点が挙げられる．

(1) 関連語－訳語関連行列の計算パラメータの最適化

関連語ペアの抽出におけるウィンドウサイズ，共起頻度の閾値，相互情報量の閾値などの値を変更して実験し，最適値を探す．

(2) フレーズテーブルの修正の重み最適化

提案方法で推定された擬似翻訳確率と従来方法で推定された翻訳確率の平均をとっているが，擬似翻訳確率の重みを変える，あるいは異なる素性値として扱う．

(3) 動詞に対する擬似翻訳確率の推定

動詞の訳語を決定の手がかりとしては目的語などが有効であり，構文共起に基づいて関連語を抽出することが望ましい．関連語－訳語関連行列の反復計算のアルゴリズムもそれに合わせて変形することが必要である．

参考文献

Breen, J.W. 1995. Building an Electronic Japanese- English Dictionary. In *Proc. of the Japanese Studies Association of Australia Conference*.

Church, Kenneth W., William Gale, Patrick Hanks and Donald Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, pages 115-164.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22-29.

Ide, Nancy and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):1-40.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Unsupervised Word Sense Disambiguation Using Bilingual Comparable Corpora. In *Proc. of the 19th International Conference on Computational Linguistics*, pages 411-417.

Koehn, Philipp and Kevin Knight. 2000. Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm. *Proc. AAAI 2000*, pp. 711-715.

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, Evan Herbst, 2007. Moses: Open Source Toolkit for Statistical Machine Translation, In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL), demonstration and poster session*, pages 177-180.

Li, Cong and Hang Li. 2002. Word translation dis- ambiguation using bilingual bootstrapping. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 343-351.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

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

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proc. ACL 2002*, pp. 311-318.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143-177.

Smadja, Frank, Kathleen R. McKeown and Vasileios Hatzivassiloglou. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. In *Computational Linguistics*, 22(1):3-38.

Stolcke, Andreas. 2002. SRILM - An Extensible Language Modeling Toolkit. *Proc. Intl. Conf. Spoken Language Processing*, pp. 901-904.

Utiyama, Masao and Hitoshi Isahara. 2007. A Japanese-English Patent Parallel Corpus. *Proc. Machine Translation Summit XI*, pp. 475-482.

Vickrey, David, Luke Biewald, Marc Teyssier and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proc. of the Conference on HLT/EMNLP*, pages 771-778.

Wu, Hua, Haifeng Wang and Chengqing Zong. 2008. Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Corpora. *Proc. COLING 2008*, pp. 993–1000.

３．４「データ駆動型中国語 HPSG パーサのためのツリーバンクの変換」で用いられる中国語ツリーバンク CTB における「把」構造と「被」構造について

東京大学　王　向莉、Kun Yu、辻井　潤一

3.4.1　はじめに

従来、既存の言語理論に基づき、人手で言語処理用の文法を開発するのは一般的な手法であった。しかし、このような手法は十年またはそれ以上の時間がかかり、そして、開発者の巨大な努力が必要である。最近、文法獲得における新しい手法が提案された。それは、ツリーバンクから文法を獲得する手法である(Xia 1999; Chen and Shanker 2004; Chiang 2000; Hockenmaier and Steedman 2002; Miyao 2006; Guo 2009; Cramer and Zhang 2009)。この手法は、コストが低い、網羅性が高い、機械学習のために training data を提供できるなど、いくつかのメリットがある。

われわれは、中国語 HPSG 文法を獲得することを目標としている。中国語の HPSG 文法を既存の中国語ツリーバンクから獲得する手法によって、その目標を実現する。具体的には中国語の既存のツリーバンクを HPSG ツリーバンクに変換し、中国語 HPSG 文法を獲得する。既存の中国語ツリーバンクとしては、Penn Chinese Treebank (CTB), Peking University Treebank(PKU), Tsinghua University Treebank (TSU) がよく知られている。そのなかでも、CTB は解析の一貫性がよく保たれているため、HPSG ツリーバンクへ変換するのに適切な言語資源であると議論されている（Yu et al. 2010)。そこで、われわれは CTB を基礎的な言語資源として選んだ。

ツリーバンクから獲得される文法はツリーバンクの構文解析の精度に依存する。ツリーバンクにおける構文解析が誤っていた場合、誤った文法が獲得されてしまう。この問題の解決策としては、ツリーバンクの誤りを一般化し、ツリーバンクを修正する手法が考えられる。本稿では、1) CTB の「把」構造と「被」構造の一部の解析の誤りを一般化して議論する; 2)望ましい解析を提案する; 3)CTB を修正する手法を提案する；4)修正した結果を評価する。

3.4.2　CTB の「把」構造と「被」構造の解析の問題点

中国語は SVO 型言語で、目的語は一般的に述語の後ろに現れる。しかし、「把」構造では目的語が述語の前に現れ、「被」構造では目的語が主語の位置に、主語が「被」の後ろ、述語の前に位置する。1a は一般的な文、1b は「把」構造文、1c は「被」構造文である。

1a. 我/I 吃/eat 苹果/apple
　　私はりんごを食べる
1b. 我/I 把/Ba 苹果/apple 吃/eat
　　私はりんごを食べる
1c. 苹果/apple 被/Bei 我/I 吃/eat
　　りんごが私に食べられる

CTB では、「把」も「被」も動詞として処理されている。述語の後ろに目的語の trace がつけられ、それに述語前に現れる目的語と同じ数字をつけ、同じものを指すことを表す。図 1 は CTB の「把」構造文の例を示す。目的語の trace は*で表現される。また NP-SBJ-1 は NP-OBJ-1 と

同じ数字１をつけることで、NP-OBJ-1 という trace が NP-SBJ-1 のものであることを表す。このような情報は HPSG 文法を獲得する際に重要な情報である。

CTB では、すべての「把」構造文と「被」構造文に目的語の trace がついている訳ではなく、trace がない「把」構造文と「被」構造文が多く存在する。目的語の trace がないことは「把」の後ろの要素が述語の目的語でなく、述語の後ろに戻せないことを意味する。しかし、図 2 に示す例文(2a)では、「把」の後ろの要素「花子/hanako」は、2b のように述語「当作/treat as」の後ろに戻せないが、「当作/treat as」を「当/treat」と「作/as」の 2 つの単語に分割すれば、「把」の後ろの要素「花子/hanako」は、2c のように述語「当/treat」の後ろに戻せる。したがって、2a の構造は図 3 のように修正する必要がある。「当作/treat as」は１つの単語でなく、２つの単語に分割するべきであり、かつ、述語「当/treat」の後ろに目的語の trace をつけるべきである。2a と同じ構造を持つ文をどのようにしてツリーバンクから探し出し、どのようにその構造を修正するのかが１つの課題となる。

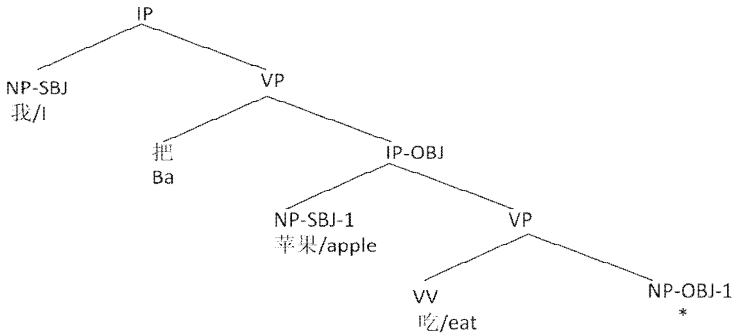


図 1：CTB の「把」構造文の解析 1

2a. 太郎/taro 把/Ba 花子/hanako 当作/treat as 朋友/friend
太郎は花子を友達と思う

2b. *太郎/taro 当作/treat as 花子/hanako 朋友/friend

2c. 太郎/taro 当/treat 花子/hanako 作/as 朋友/friend
太郎は花子を友達と思う

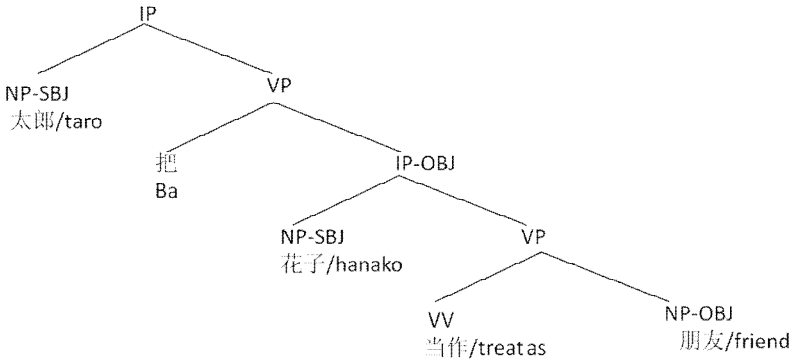


図 2：CTB の「把」構造文の解析 2

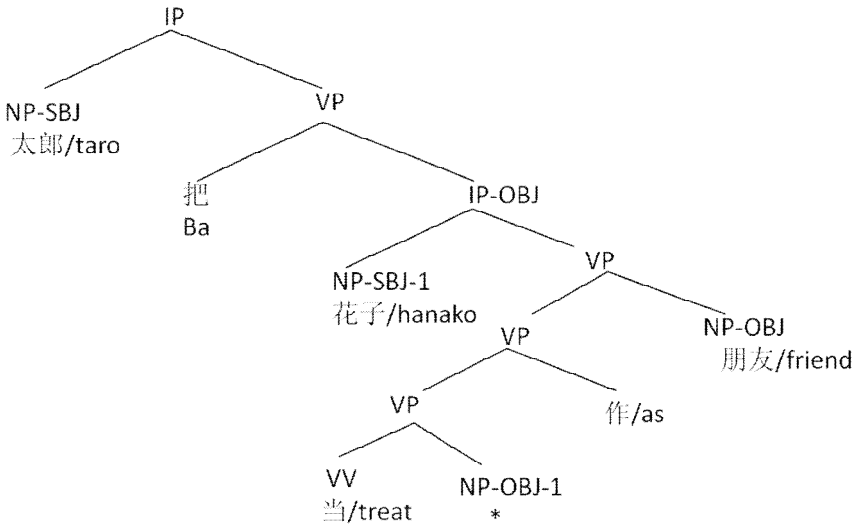


図 3：CTB の「把」構造文の解析 2 の修正

3.4.3 新しい文構造と目的語マーカ

言語学の視点からみると、2a は中国語の１つの構造を代表している。中国語では、「当」のような動詞は、２つの目的語を取ることができない。したがって、3a は正しくない表現である。しかし、「当」のような動詞の後ろに「作」のような単語がついていると、3b のように２つ目の目的語を取ることができる。3c と 3d はこの種の構造の「把」構造と「被」構造である。ここでは、「作」のような役割を果たす単語を目的語マーカ(object marker)と呼ぶ。中国語の目的語マーカは機能語であり、数が少ない。4a の「给」、5a の「入」、6a の「为」などは、「作」と同じような役割を果たし、それぞれの文において、文 4b、5b、6b のように目的語を述語の直後に戻せるため、目的語マーカに属している。

3a. *太郎/taro 当/treat 花子/hanako 朋友/friend

3b. 太郎/taro 当/treat 花子/hanako 作/as 朋友/friend
太郎は花子を友達と思う

3c. 太郎/taro 把/Ba 花子/hanako 当/treat 作/as 朋友/friend
太郎は花子を友達と思う

3d. 花子/hanako 被/Bei 太郎/taro 当/treat 作/as 朋友/friend
花子は太郎に友達と思われる

4a. 把/Ba 花/flower 献/give 给/to 他/him
花を彼にささげる

4b. 献/give 花/flower 给/to 他/him
花を彼にささげる

5a. 把/Ba 书/book 放/put 入/into 书包/bag
本をかばんの中に入れる

5b. 放/put 书/book 入/into 书包/bag

本をかばんの中に入れる

6a. 把/Ba 他/he 称/call 为/as 老师/teacher

彼を先生と呼ぶ

6b. 称/call 他/him 为/as 老师/teacher

彼を先生と呼ぶ

ここで言及した文構造の「把」構造と「被」構造は、中国語で書かれた文章の中で頻繁に現れる構造で、CTB の中にも数多く存在している。

3.4.4 CTB の修正手法

CTB の中で、3 節で言及した構造を持つ文が多数に存在しているが、正しく解析されていない。この節では、このような文を抽出し、正しい構造に修正する方法を提案する。中国語では、単語とフレーズの境界は非常に曖昧である。ツリーバンクを構築する際に、解析の一貫性を保つために、単語の分割基準を決めなければならない。CTB 単語分割ガイドラインでは、2 つの漢字からなるフレーズを 1 つの単語として処理する場合が多い。「建设 成/build into」と「建 成/build into」はいずれも動詞と目的語マーカからなるフレーズであるが、CTB では、「建设 成/build into」を 7b のように２つの動詞に分けていて、１つの複合動詞 VRD として処理している一方で、「建 成/build into」は 7a のように１つの単語をとしている。われわれは、7a のように処理されている文と 7b のように処理される文をパターン A とパターン B に分けて、処理する。

7a. (建成/VV)

7b. (VRD (VV 建设) (VV 成))

節 3 で提案した構造における目的語マーカは機能語で、数が少ない。われわれは中国語の目的語マーカを列挙したリストを作成した。CTB の修正手法は 3 つにステップからなる。

まず、3 つの抽出パターンを作成して、パターン A に当てはまる文を CTB から抽出する。8a は「把」構造文に、8b と 8c は「被」構造文に対応する。3 つのパターンにある VV の末の漢字は、リストに列挙された目的語マーカのいずれかに一致する。抽出された文の VV は 2 つの単語に分割する。9a は分割する前の文の例で、9b は分割した後の文である。

8a. NP0 + 把 ＋ NP1 ＋ VV

8b. NP0 + 被 ＋ NP1 ＋ VV

8c. NP0 + 被 ＋ VV

9a. 把/Ba 花子/hanako (VV 当作/treat as) 朋友/friend

9b. 把/Ba 花子/hanako (VRD (VV 当/treat) (VV 作/as)) 朋友

次に、10a、10b、10c の 3 つの抽出パターンを作成し、パターン B に当てはまる文を CTB から抽出する。10a は「把」構造文に、10b と 10c は「被」構造文に対応する。3 つのパターンにある VRD の後ろの VV は、リストに列挙される目的語マーカのいずれかに一致する。

10a. NP0 + 把 ＋ NP1 ＋ (VRD VV1 VV2)

10b. NP0 + 被 ＋ NP1 ＋ (VRD VV1 VV2)

10c. NP0 + 被 ＋ (VRD VV1 VV2)

最後に、ステップ1で処理されたパターンAに当てはまる文とパターンBに当てはまる文を合わせて、11a、11b、11cの規則によって、その構造を修正する。たとえば、12aを修正するあと、12bになる。

11a. NP0 + 把 ＋ NP1 ＋ (VP (VP VV1 (NP-OBJ-1 (NONE *))) VV2)

11b. NP0 + 被 ＋ NP1 ＋ (VP (VP VV1 (NP-OBJ-1 (NONE *))) VV2)

11c. NP0 + 被 ＋ (VP (VP VV1 (NP-OBJ-1 (NONE *))) VV2)

12a. 把/Ba 花子/hanako (VRD (VV 当/treat) (VV 作/as)) 朋友/friend

12b. 把/Ba 花子/hanako (VP (VP (VV 当/treat) (NP-OBJ-1 (NONE *))) (VV 作/as))) 朋友/friend

3.4.5 修正精度による評価

修正された文に対して、評価を行った。表1、表2、表3は評価の結果を表している。表1に示すように、提案した手法によって、CTBから968の文を抽出した。抽出されたパターンAに当てはまる文は657文である。表2に示すように、パターンAに当てはまる文のうち、19文は正しく修正できなかった。エラータイプ1は13aのようなエラーを指す。13aでは、「通过/通過」は1つの動詞であるが、その最後の漢字は目的語マーカに含まれているため、誤って分割された。エラータイプ2は13bのようなエラーである。13bでは、動詞「做/する」は1つの漢字からなる動詞で、この漢字は目的語マーカがリストに含まれているため、目的語マーカとされた。表3に修正精度を示す。抽出された968文のうち、正しく修正された文は949文で、修正精度は98%に達している。

13a.把/Ba 草案/draft (VRD (VV 通) (VV 过))

草案が通過する

13b.把/Ba 工作/work (VRD (VV) (VV 做/do)) 在/in 前头/ahead

前もって、仕事をする

CTBから抽出された文の総数	パターンAにあたる文数	パターンBにあたる文数
968	657	311

表1：修正精度における評価1

パターンAにあたる文数	エラーのタイプ1	エラーのタイプ2	正しく修正された文の数
657	11	8	638

表2：修正精度における評価2

CTBから抽出された文の総数	正しく修正されなかった文の数	正しく修正された文の数
968	19	949
100%	2%	98%

表3：修正精度における評価3

3.4.6 おわりに

ツリーバンクから文法を獲得する手法が主流になっている。しかし、ツリーバンクから獲得された文法はツリーバンクの構文解析の精度に依存することが問題点となる。本稿では、既存の中国語ツリーバンク CTB から、獲得される中国語 HPSG 文法の精度をあげる手法を提案した。まず、ツリーバンク中のある構造における解析が誤っていることを指摘し、一般化した構造について正しい解析を提案する。次に、CTB から誤った文を抽出し、修正する手法を提案した。最後に修正結果に対して評価を行った。この評価によると、修正精度は 98%に達しており、提案した手法の有効性が確認された。

[参考文献]

Xia, F. (1999). “Extracting Tree Adjoining Grammars from Bracketed Corpora”. In Proceedings of the 5th NLPRS.

Chen, J. and Shanker, V. (2004). “Automated Extraction of TAGs from the Penn Treebank”. In Proceedings of the 6th IWPT.

Chiang, D. (2000). “Statistical Parsing with an Automatically-extracted Tree Adjoining Grammar”. In Proceedings of the 38th ACL. pp. 456-463.

Hockenmaier, J. and Steedman, M. (2002). “Acquiring Compact Lexicalized Grammars from a Cleaner Treebank”. In Proceedings of the 3rd LREC.

Miyao, Y. (2006). “From Linguistic Theory to Syntactic Analysis: Corpus-oriented Grammar Development and Feature Forest Model”. Ph. D. Thesis. The University of Tokyo.

Guo, Y. (2009). “Treebank-based acquisition of Chinese LFG Resources for Parsing and Generation”. Ph. D. Thesis. School of Computing, Dublin City University.

Cramer, B. and Zhang, Y. (2009). “Construction of a German HPSG Grammar from a Detailed Treebank”. In Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks, ACL-IJCNLP 2009. pp. 37-45.

Yu, K. et al. (2010). “Comparison of Chinese Treebanks for Corpus-oriented HPSG Grammar Development”. Journal of Natural Language Processing (Special Issue on Empirical Methods for Asian Language Processing), April 2010.

4．依存関係確率モデルを用いた統計的句アライメント

京都大学　中澤　敏明
黒橋　禎夫

4.1 はじめに

日本語と英語のように言語構造が著しく異なり、語順変化が大きな言語対において、対訳文をアライメントする際に重要なことは二つある。一つは構文解析や依存構造解析などの言語情報をアライメントに組み込み、語順変化を克服することであり、もう一つはアライメントの手法が 1 対 1 の単語対応だけでなく、1 対多や多対多などの句対応を生成できることである。これは一方の言語では 1 語で表現されているものが、他方では 2 語以上で表現されることが少なくないからである。しかしながら、既存のアライメント手法の多くは文を単純に単語列としてしか扱っておらず[1]、句対応は単語対応を行った後にヒューリスティックなルールにより生成するといった方法を取っている[3]。Quirk ら[6]はアライメントに構造情報を統合しようとしたが、前述の単語列アライメントを行った後に用いるに留まっている。単語列アライメント手法そのものの精度が高くないため、このような方法では十分な精度でアライメントが行えるとは言い難い。

本論文では単語や句の依存関係に注目した句アライメントモデルを提案する。提案手法のポイントは以下の 3 つである。

1．両言語とも依存構造解析し、アライメントの最初から言語の構造情報を利用する
2．アライメントの最小単位は単語だが、モデル学習時に句となるべき部分を自動的に推定し、句アライメントを行う
3．各方向（原言語→目的言語と目的言語→原言語）の生成モデルを二つ同時に利用することにより、より高精度なアライメントを行う

本モデルは二つの依存構造木において、一方の依存構造木で直接の親子関係にある一組の対応について、他方のそれぞれの対応先の依存関係をモデル化しており、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。言い替えれば、本モデルは木構造上での reordering モデルということができる。また本モデルはヒューリスティックなルールを用いずに、句となるべき部分を自動的に推定することができる。ここでいう句とは必ずしも言語的な句である必要はなく、任意の単語のまとまりである。ただし、Phrase-based SMT における句の定義との重要な違いは、我々は木構造を扱っており、単語列としては連続でなくても、木構造上で連続ならば句として扱っているという点である。

また我々のモデルは IBM モデルのような各方向の生成モデルを両方向分同時に用いてアライメントを行う。これはアライメントの良さを両方向から判断する方が自然であり、Liang ら[4]による報告にもあるように、そうした方が精度よいアライメントが行えるからである。

4.2 提案モデル

以降の説明においては言語対として日本語と英語を用いるが、提案モデルはこの言語対に特別

に設計されたものではなく、言語対によらないロバストなものである。

提案モデルは依存構造木上で定義されるものであるので、まず対訳文を両言語とも依存構造解析し、単語の依存構造木に変換する。図１に依存構造木の例を示す。単語は上から下に順に並んでおり、文のヘッドとなる単語は最も左側に位置している。アライメントの最小単位はこれら各単語であるが、モデル推定時に複数単語のかたまりを句として自動的に獲得する。

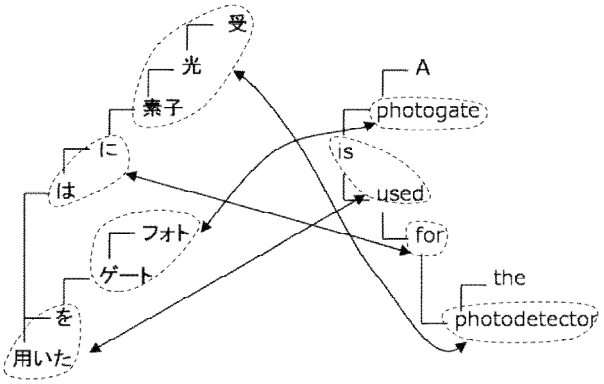


図 1：依存構造木とアライメント例

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 $\hat{\mathbf{a}}$ を

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} p(\mathbf{f} \mid \mathbf{a}, \mathbf{e}) \cdot p(\mathbf{a} \mid \mathbf{e}) \qquad (1)$$

により求める。$p(\mathbf{f} \mid \mathbf{a}, \mathbf{e})$ は lexicon probability と呼ばれ、単語の翻訳確率を表しており、$p(\mathbf{a} \mid \mathbf{e})$ は alignment probability と呼ばれ、語順入れ替えの確率を表している。これに対し、提案モデルは双方向のモデルを同時に利用し、以下の式により最良のアライメントを求める。

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} p(\mathbf{f} \mid \mathbf{a}, \mathbf{e}) \cdot p(\mathbf{a} \mid \mathbf{e}) \cdot p(\mathbf{e} \mid \mathbf{a}, \mathbf{f}) \cdot p(\mathbf{a} \mid \mathbf{f}) \qquad (2)$$

$p(\mathbf{f} \mid \mathbf{a}, \mathbf{e})$ と $p(\mathbf{e} \mid \mathbf{a}, \mathbf{f})$ は句翻訳確率であり、$p(\mathbf{a} \mid \mathbf{e})$ と $p(\mathbf{a} \mid \mathbf{f})$ は依存関係確率である。なお提案モデルは EM アルゴリズムにより学習される。

4.2.1 句翻訳確率

fがN 個の句（$F_1, F_2, \ldots, F_N$）からなり、**e**がM 個の句（$E_1, E_2, \ldots, E_M$）とNULL（E_0）からなるとする。またアライメント $\mathbf{A}^{fe}$ はfの各句からeの各句への対応を表し、$A^{fe}_j = i$は句F_jが句E_iに対応していることを示すとする。提案モデルでは、IBM モデルにおける単語翻訳確率の代わりに、句翻訳確率を考える。ただし、2語以上からなる句はNULL対応にはならないという制限を加える（その句に含まれる各単語がNULL対応になるものとする）。句翻訳確率は以下のようになる。

$$p(\mathbf{f} \mid \mathbf{a}, \mathbf{e}) = \prod_{j=1}^{N} p(F_j \mid E_{A^{fe}_j}) \qquad (3)$$

ここで、句F_jと句E_iが対応付いたと仮定すると、この句の対応に寄与する句翻訳確率は、双方向分の句翻訳確率を掛け合わせるため以下のようになる（この確率の積を句対応確率と呼ぶことにする）。

$$p(F_j \mid E_i) \cdot p(E_i \mid F_j) \qquad (4)$$

4.2.2 依存関係確率

まず **e** のある単語 e_p と、e_p に係る単語 e_c について考え、それらの可能なアライメントのうち e_p が句 E_P に属し、e_c が句 E_C に属しており、E_C が E_P に係っているものを考える。このような状況において、E_P と E_C の **f** での対応句 $F_{A_P^{ef}}$ と $F_{A_C^{ef}}$ の関係をモデル化したものが依存関係確率である。日英などのように語順の大きく異なる言語対であっても、文内の単語や句の依存関係は多くの場合保存されることが多い。提案モデルはこのような傾向を考慮したものである。直接の親子関係にある 2 単語が属する 2 句の対応先の句の関係は $rel(e_p, e_c)$ のように記述することにし、これは e_p が属する句の対応先の句から、e_c が属する句の対応先の句への経路として定義される。この ***rel*** を用いて、依存関係確率は以下のようになる。

$$p(\mathbf{a}\,|\,\mathbf{e}) = \prod_{(e_p, e_c)\in D_{\mathbf{e}-pc}} p_{\mathbf{ef}}(rel(e_p, e_c)) \qquad (5)$$

ここで $D_{\mathbf{e}\text{-}pc}$ は **e** の木構造において直接の親子関係にある全ての単語の組み合わせである。

4.3 トレーニング

提案モデルは 2 つのステップに分けて学習される。これは IBM モデルにおいて、完全に最適解が求まる簡単なモデルからスタートし、徐々により複雑なモデルに移行することに対応する。Step 1 では単語翻訳確率の推定が行われ、Step 2 では句翻訳確率と依存関係確率の推定が行われる。どちらのステップにおいてもモデルは EM アルゴリズムにより学習される。またステップ 1 においては句は扱わず、全て単語単位での学習となる。複数単語の塊＝句は Step 2 において自動的に獲得される。

4.3.1 Step 1

Step 1 では各方向独立に、単語翻訳確率を推定する。これは IBM Model 1 と全く同様の方法により行われる。Step 1 の推定の際には対応の単位は各ノード単体、つまり単語のみであり、句は考慮しない。句は Step 2 の推定から考慮し、句となるべき候補を動的に作り出すことにより実現する。これは Step 1 の段階で可能な句の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。

4.3.2 Step 2

Step 2 では句翻訳確率と依存関係確率の両方を推定する。また二つのモデルを同時に用いて、一つの方向性のないアライメントを得る。Step 1 では計算を効率化することにより、近似を用いずにモデルの推定が完全に行えるが、Step 2 では可能なアライメントを全て考慮することは不可能である。そこで我々は最も良いアライメントを探索するために、まず句翻訳確率のみから初期アライメントを生成し、その後依存関係確率も考慮しつつ、山登り法によってアライメントを徐々に修正するという方法をとる。

さらに Step 2 において新たな句候補の生成を行う。新たな句候補は山登り法によって求められ

た最も良いアライメントの状態から生成され、次のイタレーションから考慮される。つまり、Step 2 のイタレーションが進むに連れ、より大きな句の対応を発見することができる。

全体として、Step 2 の 1 回のイタレーションは、E-step での初期アライメントの生成と山登り法による最適なアライメントの探索、E-step と M-step の間での新たな句候補の生成、M-step でのパラメータの更新の 4 つの要素からなる。

Step 2 での一回目のイタレーションでは、パラメータの初期値を以下のようにする。一回目のイタレーションにおいては全ての句は 1 単語からなるため（2 単語以上からなる句候補が獲得されていないため）句翻訳確率については、Step 1 で求めた単語翻訳確率をそのまま用いる。依存関係確率は、Step 1 の最後のイタレーションで得られた最も良いアライメント結果において依存関係の生起回数を計数し、そこから求めた確率を用いる。

4.3.2.1 初期アライメントの生成（E-step）

依存関係確率は用いず、句翻訳確率のみから初期アライメントを生成する。全ての句候補同士の対応（もしくは NULL 対応）に対して、句対応確率を計算する。これらの中から、句対応確率

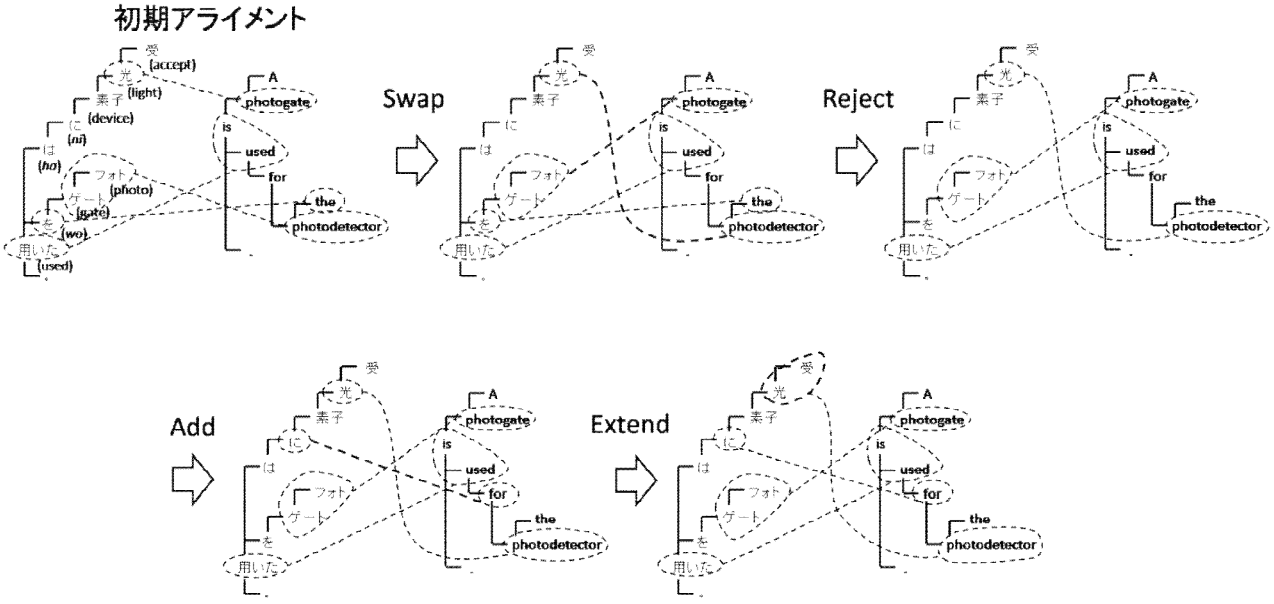


図 2：山登り法の例

の相乗平均が高いものから順に、対応として採用する。この際、各単語は 1 度しか対応付かないようにする。つまりすでに採用されている対応と重なるような対応は採用しない。なお句候補の生成については後で述べる。

4.3.2.2 山登り法（E-step）

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、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。

Swap：任意の 2 つの対応に注目し、それらの対応を入れ替える。例えば図 2 の最初の操作では、

“光 ⇔ photogate”と“フォト ゲート ⇔ photodetector”の対応がそれぞれ“光 ⇔

photodetector”と“フォト ゲート ⇔ photogate”というように対応が入れ替えられている。

Extend：任意の1つの対応に注目し、そのいずれかの言語における句を、親または子方向に1ノード分だけ拡大する。

Add：NULL対応となっている原言語側及び目的言語側のノード間に、新たに対応を追加する。

Reject：すでにある対応を削除し、それぞれNULL対応とする。

図2に山登り法によるアライメント修正過程の例を示す。なお図2は1回以上イタレーションを行ったあとの状態である。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し行い、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。なお修正の途中で得られたアライメントの状態を、確率の高いものから n 保存しておき、仮想的な n-best アライメントとし、パラメータ推定の際に利用する。

4.3.2.3 新たな句候補の生成

山登り法により得られた最も良いアライメント結果のうち、NULL対応となった単語に注目する。NULL対応となった語の親、または子の単語がNULL対応でなければ、その単語とNULL対応の単語とをまとめたものを新たに句として獲得し、Step 2の次のイタレーションから探索範囲に入れる。例えば図2の最終状態においてNULL対応となっている“素子”は、その子の対応である“受 光 ⇔ photodetector”に含まれ、新たに“受 光 素子”という句を作りだし、“受 光 素子 ⇔ photodetector”という対応があるものと考えるさらに親の対応である“に ⇔ for”に含まれ、“素子 に”という句もつくり出し、“素子 に ⇔ for”という対応があるものと考える。これらの新たに考慮される対応には、元の対応の出現期待値（正規化されたアライメントの確率）を分配する。このように、NULL対応に注目することにより動的に句となるべきかたまりを獲得していき、モデルの構築を行う。

4.3.2.4 モデル推定（M-step）

一般的なEMアルゴリズムにおいては、得られた n-best アライメントのそれぞれのアライメント確率を正規化し、各アライメントにおけるパラメータの出現回数をこの正規化された確率値（出現期待値）を用いて計数する。我々もこの方法に従い、全ての対訳文での全てのアライメント結果を集めてパラメータの推定を行う。ただし、正確に全てのアライメントを数え上げることはできないため、山登り法の途中で得られたアライメントのうち、アライメントの確率の高いもの上位 n 個（山登りの回数が n に満たない場合はその全て）を用いる。パラメータ推定は各パラメータの出現期待値の総和を全体の回数で正規化することにより行われる。例えば句翻訳確率は以下のような式により推定する。

$$p(F_j \mid E_i) = \frac{C(F_j, E_i)}{\sum_k C(F_k, E_i)}, \quad p(E_i \mid F_j) = \frac{C(E_i, F_j)}{\sum_k C(E_k, F_j)} \qquad (6)$$

ここで $C(F_j, E_i)$は F_jと E_iがアライメントされた回数である。

ここまでの処理により、EM アルゴリズムの E-step、M-step が終了し、再び E-step に戻る。これを複数回繰り返すことにより、モデルのトレーニングを行う。

4.4 実験と考察

提案手法の有効性を示すためにアライメント実験を行った。トレーニングコーパスとして JST 日英抄録コーパスを用いた。このコーパスは、科学技術振興機構所有の約 200 万件の日英抄録から、内山・井佐原の方法[7]により、情報通信研究機構が作成したものであり、100 万対訳文からなる。このうち 475 文に人手で正解のアライメントを付与し、正解データとした。正解データは Sure(S)と Possible(P)の２段階に分けてアライメントされている[5]。また評価の単位は日本語、英語とも単語とし、適合率・再現率・AER により精度を求めた。

表 1：アライメント実験結果

	Pre	Rec	AER
Step 1	77.55	33.92	47.20
Step 2-1	84.26	48.38	61.65
Step 2-2	84.85	57.26	68.53
Step 2-3	82.84	61.86	**71.03**
Step 2-4	80.21	63.13	70.88
Step 2-5	78.71	64.10	70.88
Step 2-6	78.02	63.38	70.76
Step 2-7	76.83	64.60	70.39
Step 2-8	71.99	**67.71**	69.85
intersection	**90.51**	45.16	60.31
grow-final-and	79.92	60.06	68.70
grow-diag-final-and	77.80	61.47	68.77

日本語文に対しては形態素解析器 JUMAN および依存構造解析器 KNP を用い、英語文に対しては Charniak の nlparser を用いて構文解析し、ルールにより単語の依存構造木に変換する。また Step2 のパラメータ推定の際に用いるアライメントの数は n=10 とした。

比較実験として、単語列アライメント手法として広く利用されている IBM モデルを実装したアライメントツールである GIZA++[5]を用いてアライメントを行った。各モデルのイタレーション回数などのオプションはデフォルトの設定をそのまま利用した。さらに各方向のアライメント結果を三つの対称化手法により統合した。結果を表１の下部３行に示す。利用した対称化手法は 'intersection'、'grow-final-and'、'grow-diag-final-and'の 3 つである[2]。

一方、提案手法によるアライメント精度を表１の上部に示す。まず'Step 1'に示されているのは、Step 1 のイタレーションを 5 回行った後に学習されたパラメータ(単語翻訳確率) を用いたアライメントの精度である。なおここでのアライメントは、両方向のパラメータを用いて、初期アライメント生成手法と同様にアライメントを生成した結果である。'Step 2-X'は Step 2 の各イタレーション終了時点でのアライメント精度である。'Step 2-1'は句翻訳確率は'Step 1'のものと同じだが、それに加えて'Step 1'のアライメント結果から推定した依存関係確率を用いてアライメントを行っている。つまり、'Step 1' と'Step 2-1'とを比較することにより、依存関係確率を用いることによるアライメント精度の向上が見て取れる。以後 Step 2 のイタレーションを行い、その都度アライメント精度を計測した。結果として、提案手法では単語列アライメント手法よりも AER で 2.3 ポイントのアライメント精度向上を達成した（Step 2-3 と grow-diag-final-and との比較による）。適合率だけを見ると'intersection'が最もよい値を示しているが再現率が極端に低くなっている。

また再現率が最も高いのは'grow-diag-final-and'であるが、同程度の再現率を示している提案手法の結果を見ると、適合率では大きく上回っており、総合的に見て提案手法は単語列アライメント手法よりも優れているということができる。

4.5 まとめと今後の課題

本稿では依存関係確率モデルを用いた統計的句アライメント手法を提案した。提案モデルは木構造上での reordering モデルということができ、シンプルなモデルながらも言語構造の違いを柔軟に吸収し、精度の高いアライメントを実現できた。実験結果から、語順の大きく異なる言語対に対しては既存の単語列アライメント手法では十分な精度を達成することは困難であり、構文解析などの言語情報を利用することが自然であり、高い効果を示すことが証明された。今回は日本語と英語間のアライメント実験のみしか行わなかったが、同様に語順に大きな違いのある日本語と中国語間での実験などを行い、提案手法が言語対によらずロバストな手法であることを示す必要がある。

考察にも述べたとおり、提案手法は依存構造解析に大きく依存しており、依存構造解析誤りが容易にアライメントの誤りにつながってしまう。両言語の解析結果を照らしあわせて、文構造を修正しつつアライメントすることも可能なはずであり、現在検討中である。これが実現できれば、依存構造解析とアライメント双方の精度向上が可能となると考える。

アライメントの精度のみを評価したが、この結果が翻訳の精度にどのように影響するかを調査することは今後の課題である。

参考文献

[1] Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). “The Mathematics of Statistical Machine Translation: Parameter Estimation.” Association for Computational Linguistics, 19 (2), pp. 263–312.

[2] Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., and Talbot, D. (2005). “Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation.” In Proceedings of International Workshop on Spoken Language Translation 2005 (IWSLT'05).

[3] Koehn, P., Och, F. J., and Marcu, D. (2003). “Statistical Phrase-Based Translation.” In HLT-NAACL 2003: Main Proceedings, pp. 127–133.

[4] Liang, P., Taskar, B., and Klein, D. (2006). “Alignment by Agreement.” In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, pp. 104–111 New York City, USA. Association for Computational Linguistics.

[5] Och, F. J. and Ney, H. (2003). “A Systematic Comparison of Various Statistical Alignment Models.” Association for Computational Linguistics, 29 (1), pp. 19–51.

[6] Quirk, C., Menezes, A., and Cherry, C. (2005). “Dependency Treelet Translation: Syntactically Informed Phrasal SMT.” In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 271–279.

[7] Utiyama, M. and Isahara, H. (2007). “A Japanese-English Patent Parallel Corpus.” In MT summit XI, pp. 475–482.

5．規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳(その 2)

山梨英和大学　江原　暉将

5.1　はじめに

これまで、規則方式日英機械翻訳(RBMT)と統計的後編集(SPE)を組み合わせることで翻訳精度の向上を図ってきた[江原、小玉 2005][江原 2006][江原 2008]。これらの比較を表 1 に示す。表 1 には本報告の結果も合わせて載せてある。図 1 に示すシステムアーキテクチャーには変更はないが、各部分を構成する部品や使用しているデータに変化があり、BLEU や NIST で評価した翻訳精度は向上している。

本システムは RBMT 部と SPE 部から構成されている。RBMT 部では入力日本語文を規則方式機械翻訳によって英語文に翻訳する。さらに SPE 部でその英語文を、より精度の高い後編集後英語文に書き換える。SPE 部は訓練データから得られた翻訳モデル[1] と言語モデルを用いて動作する。

5.2　本報告で用いる訓練データと試験データ

本報告で用いるデータは、国立情報学研究所から「NTCIR-8 特許翻訳タスク参加者用テストコレクション」として提供された NTCIR-7 の formal run のためのデータである[Fujii, 2008]。試験データは 1381 文である。訓練データの元データは、日英特許平行コーパスであり、約 180 万文から成る。言語モデル(LM)の訓練データとしては、元データの英語部分を抽出して用いた。よって 180 万文である。翻訳モデル(TM)の訓練データは、次節で述べる方法によって元データから 8 万 2 千文対の日英対応データを選択して用いた。

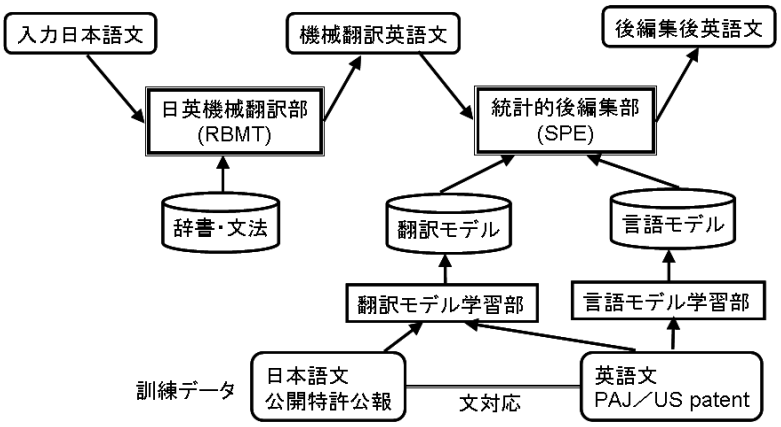


図 1 システムアーキテクチャー

[1] 後編集は英語から英語への書き換えであるから「翻訳モデル」という用語は適切でなく「書き換えモデル」と呼ぶべきであるが、慣例に従って翻訳モデルという用語を用いる。

表 1 規則方式機械翻訳(RBMT)と統計的後編集(SPE)を組み合わせたシステムの推移 [2]

	[江原、小玉2005]	[江原2006]	[江原2008]	本報告
RBMT部分	市販品A	非市販品	非市販品	市販品B
SPE部分	単語レベル(isi)	単語レベル(isi)	句レベル(Moses)	句レベル(Moses)
TM学習器	Giza-pp	Giza-pp	Giza-pp	Giza-pp
TM訓練データ	特開報／PAJ 9万3千文対	特開報／PAJ 9万3千文対	特開報／PAJ 9万3千文対	NII NTCIR-7 8万2千文対
LM学習器	Srilm	Srilm	Srilm	Srilm
LM訓練データ	PAJ 33万文	PAJ 33万文	PAJ 33万文	US patent 180万文
BLEU	0.1607	0.1728	0.2912	0.2998
NIST	4.7184	4.7893	6.3398	7.3058

5.3 翻訳モデル訓練データの抽出方法

基本的な考え方は、元データから試験データにマッチする日本文とそれに対応する英文のみを抽出して用いるという方法である。そのため、本システムは、翻訳モデルの構築と翻訳自体を同時に進める必要があり、現時点での計算機能力では、リアルタイムの翻訳ができないという難点がある。それは置くとして、訓練データ構成法は以下の通りである。

(a) 試験データと元データの日本語部分からのキーワード抽出

試験データと元データの双方の日本語部分を形態素解析し(ChaSen を利用)、カタカナまたは漢字を含む形態素をキーワードとして抽出する。

(b) 訓練文の抽出

試験データに属する各試験文のキーワード集合と訓練データに属する各訓練文のキーワード集合を比較して、類似している訓練文のみを抽出する。この比較においては、試験文に含まれる各キーワードごとにそれを含む訓練文のみを事前に抽出しておいて、その中から類似度の高い文を 10 文ずつ抽出するという方法をとった[3]。ここで、類似度の計算方法は以下のとおりである。試験文のキーワード集合(異なり)を T、訓練文のキーワード集合(異なり)を S とする。また集合 A の要素数を#(A)で表す。このとき、試験文と訓練文の類似度 sim は

---

[2] 使用ツールの詳細は以下のとおり。
言語モデル学習器：http://www.speech.sri.com/projects/srilm/の srilm.tgz ver.1.5.5
翻訳モデル学習器：http://code.google.com/p/giza-pp/の giza-pp-v1[1].0.1.tar.gz
単語レベルデコーダ：
http://www.isi.edu/publications/licensed-sw/rewrite-decoder/index.html の
isi-rewrite-decoder-r1.0.0a/linux/decoder.linux.public (現在ダウンロードできないようである)
句レベルデコーダ：http://sourceforge.net/svn/?group_id=171520 の moses.2007-05-29.gz
BLEU と NIST の計算プログラム：ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl
ただし、BLEU 値を文単位で計算するために計算式を若干変更してある[江原 2007]。

[3] ただし、試験文のキーワードが訓練文に 10 回未満しか出現していない場合は、出現しているすべての文を利用した。0 回の場合は、当該キーワードに適した訓練文がないと判断し抽出はしていない。

$$sim = \frac{2\times\#(T \cap S)}{\#(T)+\#(S)}$$

で計算される。上記の操作をすべての試験文に適用し、得られた訓練データの全体を翻訳モデルの訓練データとする。上記過程で同一の訓練文が複数回抽出される場合は、抽出された回数だけ重複して訓練データに加える。このようにして 1381 文の試験データに対して 82385 文の翻訳モデル訓練データが得られた。なお各試験文に含まれるキーワードの平均個数は 12.34 であった。また、訓練データの日本語部分は、当然、RBMT 部に用いたのと同じシステムを用いて英語文に翻訳し、それを翻訳モデルの原言語部として利用した。翻訳モデルの目標言語部は訓練データの英語部分である。

4 節に示す例文 1 に対して、訓練データの抽出例を説明する。抽出されたキーワードは

{図, 回転, 羽根, 駆動, モータ, 構成, 例, 示す}

の 8 個である。これらのキーワードごとに元データから類似度 sim の大きい順に最大 10 文を選択する。その結果 39 文が訓練文として抽出された。それらを付録 A に示す。訓練文として例文 1 と類似の構文を持ったものが抽出されていることが分かる。そのような意味で、本方式は一種の用例方式による後編集と見ることもできる。

5.4 実験結果

実験結果は既に表 1 に示してある。BLEU=0.2998、NIST=7.3058 であり、これまでより翻訳精度が向上している。ただ、これまでは使用データが PAJ(Patent Abstract of Japan)であったが、今回は NTCIR-7 のデータであり、一概に比較することはできない。

5.5 翻訳例

翻訳例を以下に示す。rbmt に比較して spe の出力英文の品質が向上していることが読み取れる。特に spe では適切な訳語が用いられていることがわかる。しかしながら、例文 2 のように rbmt 部分の出力が英文としてかなり崩れている場合には spe によっても修復されていない。

例文 1

src : 図５は回転羽根２を駆動するモータの構成例を示す図である。

ref : fig . 5 is a diagram showing a structural example of a motor for driving the rotating blade 2 .

rbmt: drawing 5 is a figure showing the example of composition of the motor which drives moving vane 2 .

spe: fig . 5 is a diagram showing an example of a motor for driving the rotator 2 .

例文 2

src: さらに、心線ワイヤ５１の先端部がラミネートフィルム５９により挟まれて保持され、その変形、ピッチの狂いを防止する。

ref: moreover , the front ends of the core wires 51 are sandwiched with laminated films 59 to prevent deformation of the core wires 51 for the purpose of maintaining their relative positions intact .

rbmt: it faces across the tip part of cable core wire 51 with laminate film 59 , it is held , and the deviation of the modification and a pitch is prevented .

spe: the distal end portion of the core wire 51 with the lamination film 59 is maintained , and the offset of the modification and its pitch is prevented .

例文 3

src: この絶縁ハウジング１０の外面に取り付けられるシールドカバー３０を図６に示している。

ref: fig . 6 shows the shield cover 30 , which is to be mounted on the insulative housing 10 .

rbmt: shield cover 30 attached to the external surface of this insulating housing 10 is shown in drawing 6 .

spe: the shield cover 30 attached onto the outer face of the insulating housing 10 is shown in fig . 6 .

5.6 おわりに

規則方式機械翻訳システム(RBMT)と統計的後編集システム(SPE)を組み合わせて、特許文書用機械翻訳システムの精度向上が図れた。得られた BLEU 値は 0.2998 である。今後の課題として、RBMT 部分の構文解析、構文生成精度の改善がある。RBMT 部分で崩れた英文が出力されてしまった場合、SPE 部分で修復するのは困難である。構文解析・生成の精度向上の一手法として文パターンの利用が考えられる。

参考文献

[江原、小玉 2005] 江原暉将、小玉修司：特許文の日英機械翻訳結果と PAJ を比較して翻訳知識を抽出する研究、平成 16 年度 AAMT/Japio 特許翻訳研究会報告書、pp.86-96, March, 2005.

[江原 2006] 江原暉将：規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳、平成 17 年度 AAMT/Japio 特許翻訳研究会報告書、pp.40-44, March, 2006.

[江原 2007] 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, March, 2007.

[江原 2008] 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.2-11, March, 2008.

[Fujii, 2008] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro：Overview of the Patent Translation Task at the NTCIR-7 Workshop, Proceedings of NTCIR-7 Workshop Meeting, pp.389-400, December, 2008.

付録 A　例文１に対して抽出された訓練データ(日本語部分)

図７は、ワード線駆動回路の構成例を示す。
図５は本実施例におけるモータ駆動回路の構成要素を示すブロック図である。
回転羽根６０は、駆動源であるモータ８１に連結されている。
図５は、図４に示すフォーカルプレーンシャッタ羽根をフォーカルプレーンシャッタに組み込んだ例を示す。
図６には、図１で示したシャッタ羽根７の詳細な構成が示され、図７には、駆動レバー８の詳細な構成が示されている。
すなわち、駆動モータ５１の単体の回転数の変化を説明すると、図２８に示すようになる。
（モータ駆動制御回路／モータ駆動制御装置の構成）図３は、モータ駆動制御回路４を示す回路図である。
図５にセンサ駆動パルスとモータ駆動パルスの発生部の構成の一例を示す。
図３にその構成例を示す。
図１４は、パルスモータおよびその制御装置の構成例を示す図である。
図８にディスクを回転駆動する構成を示す。
モータ１１の正逆転により図４に示すようにシャッタ羽根８の開閉が行われる。
このときの構成例は図１１に示すようなものとなる。
また、上記のアジテータローラ４，５，６の回転駆動機構は図３および図４に示すように構成されている。
図８は、他の形状の羽根が設けられた回転子７の構成例を示す側面図である。
これに対して、図６６～図６８に示すものは、いずれも、裏当て部材７２０を回転駆動するモータ７２６がそなえられている。
その構成例を図６に示す。
このときの構成例は図１２に示すようなものとなる。
また、図６（B）に示す構成は、駆動機構２２としてモータ２７を用いたものである。
このときの構成例は図１３に示すようなものとなる。
このときの構成例は図１０に示すようなものとなる。
図９はキャリッジを回転移動させるキャリッジ駆動機構を示す構成図である。
パドルホイール 59 は、図３に示すように、モータ M1 により回転駆動される軸の先端に複数の羽根部材を放射状に取り付けたもので、モータ M1 により矢印ｇ方向へ回転される。
図３はモータ駆動方向制限回路２５の具体的回路構成の一例を示す回路図である。
図４はモータ駆動回路２７の具体的回路構成の一例を示す回路図である。
尚、図１１（ｂ）に示す４０はシャッター羽根である。
シート送りローラ４０は、図２に示すシート送り用のモータ（ＳＦモータ）４１によって回転駆動される。
このタイミング機構Ａはドラム２の回転駆動用モータMに直結されており、図７にその具体例を示す。
これにより、モータ１１３が回転駆動すると、プラテン１００は図１０に示す矢印ａ方向に回転するようになっている。
その構成例を図３に示す。
図４に、ＬＣＤ駆動端子の回路の従来の構成例を示す。
このときの構成例は図１９に示すようなものとなる。
図３に、ＤＣモータのＩ‐Ｔ曲線(駆動電流‐トルク曲線)およびＮ‐Ｔ曲線(回転数‐トルク曲線)を示す。
図１３（A）に示すアジテータ４０Ｂはねじれた羽根１１４を有している。
図１１に、ＤＲＡＭの構成例を示す。
図１７は、対象図形である馬に羽根を付け加える例を示す。
図７はモータ駆動方向制限回路２５Ａの具体的回路構成の一例を示す回路図である。
前記駆動輪１２は図３に示すモータ１７によって一方向に間歇回転される。
図１３（A）に示すアジテータ４０Ｂはねじれた羽根１１４を有している。

6．特許文の訳し分けと動詞の格情報との対応に関する調査

山形大学　横山　晶一

6.1　はじめに

特許文の請求範囲の部分が、日本語の文には余り見られない長大な一文（約 200 字以上）になる場合が多いことは、すでに何度も指摘してきた[1]。また、昨年度の報告書では、構文解析器を特許文に適応させると、ある程度改善が見られることも指摘した[2]。

ここでは、特許文の訳し分けと動詞の格情報との対応に関する調査結果について簡単に述べる。この調査は、奥山[3～5]、鈴木[6,7]ら、研究室のメンバーによるものであるが、ツールの不備等から、まだ基礎的な段階にとどまっている。本稿では、主として鈴木に基づき、現状と問題点を述べる。

6.2　格情報

動詞の格を表す情報には、結合価や述語項構造、格フレームなどがあるが、ここでは格フレームを用いる[8]。ここで用いる格フレームは、汎用の形態素解析システム KNP [9]の格解析結果として出力されるもので、Web 上のコーパスから自動的に構築されたものである。

図１に、動詞「積む」の格フレームの一部を示す。図１上部の「積む（動 1)」では、「選手が経験を積む」といった文に対する格フレームを示し、下部の「積む（動 3)」では、「人が荷物を積む」といった文に対する格フレームを示す。「選手」などの名詞がどのような役割を果たすかが「＜ガ格＞」といった格で示され、語の後の数字は、コーパス中の頻度を示す。

これを英訳すると、「積む（動 1)」では、“ acquire ”（蓄積する）という動詞になり、「積む（動 3)」では“ load ”（荷を載せる）という動詞になる。

積む/つむ:動 1
<ガ格> 選手:17,自分:14,人:10,…
＊<ヲ格> 経験:37342,体験:1363,…
<ニ格> <補文>:70,実際:36,それぞれ:8,…
<デ格> 現場:121,会社:96,分野:86,…
積む/つむ:動 3
<ガ格> 人:12,職人:6,男:6,…
＊<ヲ格> 荷物:3592,石:1074,荷:884,…
<ニ格> 車:739,トラック:100,船:77,…
<デ格> 河原:24,手:6,前:5,…

図１　「積む」の格フレームの一部

このように、異なる格フレームが異なる英訳に対応していれば、格フレームの違いを英訳の違いに反映させることができる。また、日本語の意味の違いも表現できる。

本稿では、この考えに基づき、２つの調査を行った。一つは格フレームと英訳との対応、もう一つは、英訳と日本語文の格構造との対応である。以下にその内容を簡単に述べる。

6.3 調査データ

ここで調査に用いたデータは、Japio 特許情報データベース[10]の C12N 分野（微生物）の 2004 年度に公開された特許の「要約」項から抜き出した 6,251 文である。この部分には、人手による英訳(PAJ)が付されている。データの例を図２に、対応する英訳を図３に示す。

図２の“ ID ”の次の数字は特許の公開番号を表し、“ SOLUTION ”は【解決手段】を表す。右端の「1-2」という数字は、右側が、【解決手段】の文の数、すなわちここには２文あることを示し、左側がその中の第 n 文の当たることを示す。したがって、「本発明は、」で始まる文は、【解決手段】２文の１番目、「また、ここで、」で始まる部分は、２番目の文であることを示している。図３は、それらに対応する英訳である。英訳は一つの要約を一つの ID で表している。本研究では、この英訳を正しいものとして日本語との対応を調査した。

ID:2004000005_SOLUTION_1-2
本発明は、新規なポリペプチド及びそれらのポリペプチドをコードする核酸分子の提供に係る。
ID:2004000005_SOLUTION_2-2
また、ここで、それらの核酸配列を含むベクター及び宿主細胞、異種ポリペプチド配列に融合した本発明のポリペプチドを含むキメラポリペプチド分子、本発明のポリペプチドに結合する抗体、及び本発明のポリペプチドを製造する方法も提供される。
ID:2004000006_PROBLEM_1-1
新規なポリペプチド及びそれらのポリペプチドをコードする核酸分子の提供。

図２　日本語特許データの一部[10]

ID:2004000005 SOLUTION:
The new polypeptides, and the nucleic acid molecules encoding the polypeptides. In addition, vectors and host cells which contain the nucleic acid sequences.
Chimera polypeptide molecules which contain the polypeptides fused with heterologous polypeptide sequences. Antibodies which combine with the polypeptides. And a method for producing the polypeptides.
ID:2004000006 PROBLEM TO BE SOLVED:
To provide new polypeptides and nucleic acid molecules encoding the polypeptides.

図３　図２のデータの英訳(PAJ)（一部改変）

表１　サ変、和語高頻度動詞

サ変動詞	出現頻度	和語動詞	出現頻度
配列/はいれつ	1933	有する/ゆうする	1097
提供/ていきょう	1834	含む/ふくむ	1076
発明/はつめい	823	用いる/もちいる	868
発現/はつげん	376	成る/なる	788
由来/ゆらい	358	得る/える(受け身)	280
利用/りよう	339	行う/おこなう	266
検出/けんしゅつ	331	有る/ある	173
培養/ばいよう	328	有す/ゆうす	169
含有/がんゆう	294	示す/しめす	167
使用/しよう	274	得る/える	166

この 6,251 文の中には、1,500 種類を超える動詞が存在する。サ変動詞と和語動詞の高頻度のものをそれぞれ 10 種、表１に示す。ただし、サ変の場合は、名詞として出現したものも区別せずに示してある。頻度、種類では、サ変の方が（名詞も含めて）多いといえる。

一般にサ変の場合は、比較的意味が限定的で、英訳もそれほどバリエーションがないのに対して、和語動詞は多くが多義で、英訳では困難を伴うことが多い。これらの動詞に対する格フレームの有効性を確かめるために、すでに述べた２種類の調査を行った。

6.4　格フレームと英訳の対応（調査１）

6.4.1　動詞「含む」の調査例

動詞の訳し分けに対して格フレームが有効かどうかを、以下の手順に従って調査した。以下では、動詞「含む」についての解析を例にあげる。

(1) 動詞を含む文の抜き出し

動詞「含む」が含まれる文を 100 文、6,251 文から抜き出す。受動態「含まれる」の形は抜き出さない。たとえば次のような文が抜き出される。

(a) …工程を含むことを特徴とする…
(b) …レセプターを含む溶液を…

(2) 格解析

抜き出した文を KNP に入力する。ここでは格フレームも出力するオプションをつけてある。上記の文に対する解析結果は次のようになる。

(a) 【含む／ふくむ】　動 [34]
　　D 工程を《ヲ》[BGH:×] ことを《外の関係》[[BGH:×]*

★-26.844 点 含む/ふくむ:**動 3** 直受 1 ニ使役 可能 直受 2

(b) 【含む/ふくむ】 動 [34]

D レセプターを《ヲ》[BGH:×] 溶液を《ガ/ガ２/ヲ/外の関係/時間》[BGH:×]*

★-29.504 点 含む/ふくむ:**動 1** 直受 1 ニ使役 可能 直受 2

(3) 対応する英訳の抜き出し

対応する英訳をデータベースから抜き出す。

(4) 解析結果と英訳との比較および格フレームごとの分類

入力文に対応する英訳から、「含む」の英訳を抜き出して分類する。上記(a)の文では、英訳は“comprise”、(b)では“contain”となる。

6.4.2 解析結果と考察

表２ 「含む」の訳し分け対応結果

格番号	英訳	英訳の数	
動 3	contain	44	69
	comprise	14	
	include	10	
	involve	1	
動 1	contain	11	12
	include	1	
動 7	contain	4	7
	comprise	2	
	include	1	
動 5	comprise	3	4
	include	1	
動 16	contain	1	2
	include	1	
動 17	contain	2	2
動 4	contain	1	1
動 15	include	1	1
動 22	contain	1	1
動 34	contain	1	1

表２に「含む」の訳し分け結果を示す。「含む」の格フレーム数は 34 あるが、そのうち 10 種類が 100 文中に出現した。表から分かるように、そのうち半数以上が〈動 3〉に分類された。英語表現は“contain”(65),“comprise”(19),“include”(15),“involve”(1)の４種類であった。

69 文が分類された〈動 3〉の内容は次の通りである。

含む/ふくむ:動 3
＊ <ヲ格> 内容:9872, キーワード:1612, 事項:1034, 項目:616, 氏名:317, ...
<ニ格> コンテンツ:14, 住所:8, タイトル:7, 内容:6, 計画:3, ...
<ノ格> むけ:313, 借り主:203, 下記:53, 関連:51, ...

４種類の英語の英和辞典[11]による意味は次の通りである。

contain： (容器・場所の中に)含む、入っている、含有する
comprise： 含む、(部分)からなる、全体を構成する
include：含む、(全体の一部として)もつ、包含する
involve：含む、(必然的に)伴う、必要とする

〈動 3〉には“include”が適切な英語であると考えられるが、表２から分かるように、69 文中 10 文のみしか分類されなかった。いくつかの動詞について同様の試みを行ったが、明確な対応は見出せなかった。

6.5 英訳ごとの格構造の違い（調査２）

前節とは逆に、動詞「含む」を英訳ごとに分類して格解析を行った。その結果を図４に示す。すでに示したように、100 文のうち 65 文が“contain”と訳されている。

contain(65)
<ガ格>培地:6、蛋白質:5、DNA:5、… <ヲ格>配列:8、ポリペプチド:7、DNA:6、… <ニ格>塩基配列:1
comprise(19)
<ガ格>方法:13、試薬:2、液滴:2、… <ヲ格>工程:5、こと:5、緩衝液:2 …
include(15)
<ガ格>方法:3、DNA:2、疾患:1… <ヲ格>工程:3、遺伝子断片:1、癌:1 … <ニ格>下流:1
involve(1)
<ガ格>操作:1 <ヲ格>精製:1

図４　「含む」の英訳ごとの格構造

図４の格関係を見ると、英訳の側から逆に格関係を分類すると、訳し分けの可能性がある程度推察される。たとえば、“contain”は格要素として具体物を取り、“comprise”は抽象物を取るという違いが見られる。この傾向は、調査した他の動詞（結合する、成る、用いる、由来するなど）でも見られた。

6.6 考察と今後の展望

調査に対して人手で関わった部分が多く、調査対象が不十分であったため、明確な結論は出ていない。今後は、alignment の自動化等を通じて調査対象を拡大する予定である。

今回の調査から、格関係を見直す必要性が痛感される。格フレーム自体が Web コーパスから作成されたという経緯を持っており、今回対象にした特許文とはかなりの違いがあると思われる。今後は調査対象を広げるとともに、述語項構造、結合価などの情報と比較していく予定である。

謝辞

データ提供や数々のアドバイスをいただいた AAMT/Japio 特許翻訳研究会のメンバーならびに Japio の方々に感謝致します。

文献

[1] 横山晶一：動的シソーラスを用いた特許文の解析システム、科学技術研究費成果報告書(2007～2009)　（本報告書には、以下の[2～5]を含むほとんどの論文が含まれている）

[2] 横山晶一：特許文への構文解析適応、平成 20 年度 AAMT/Japio 特許翻訳研究会報告書、pp.65-70 (2009)

[3] 奥山真澄：格フレームを用いた特許文の訳し分け、山形大学工学部卒業論文 (2009)

[4] 奥山真澄、横山晶一：格フレームを用いた特許文の訳し分け、情報処理学会東北支部研究会 (2009) 08-6-B1-3

[5] Shoichi Yokoyama, Masumi Okuyama: Translation Disambiguation of Patent Sentences using Case Frames, Machine Translation Summit XII, 3rd Workshop on Patent Translation (2009)

[6] 鈴木勘平：動詞の格情報を用いた特許文の解析、山形大学大学院理工学研究科修士学位論文 (2010)

[7] 鈴木勘平、横山晶一：特許文の訳し分けにおける格フレームの有効性、情報処理学会第 72 回全国大会 (2010) 4W2

[8] 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 (2006) 2006-NL-171

[9] KNP: http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[10] AAMT/Japio 特許翻訳研究会特許情報データベース (2004)

平成21年度 AAMT/Japio特許翻訳研究会

海　外　調　査　報　告

第12回機械翻訳国際会議

（Machine Translation Summit XII）

及び

第3回特許翻訳ワークショップ

（The 3rd Workshop on Patent Translation）

平成22年3月

一般財団法人 日本特許情報機構

2009 年 9 月 7 日

第 12 回機械翻訳国際会議（Machine Translation Summit XII）および 第 3 回特許機械翻訳ワークショップ（The 3rd Workshop on Patent Translation）参加報告

横山　晶一（山形大学）、江原　暉将（山梨英和大学）
宮澤　信一郎（秀明大学）、潮田　明（富士通研究所）

１．出張目的

機械翻訳および特許翻訳に関する調査を目的として、表記国際会議およびワークショップに参加し、発表、討論する。

２．会議概要

第 12 回機械翻訳国際会議（Machine Translation Summit XII、以下 MT Summit XII と略称）は、2009 年 8 月 27～29 日、カナダの首都オタワの、Hotel Chateau Laurier で開催された。会場は、オタワでも最も伝統ある、首都の中心に位置するホテルで、ホテル名は、タイタニックで遭難死した設立者に代わって開設式典を行った、当時のカナダ首相の名にちなむ。

参加者は pre-registration の段階（8 月 17 日現在）で、27 の国から 279 名で、この大会の主催者である Laurie Gerber の話では、最終的には 300 名くらいであったということである。場所的な面から、アメリカ地域からの参加者が最も多く 166 名。うち 118 名がアメリカから、48 名がカナダからであった。アジアからは 31 名で、日本が 15 名、中国が 4 名、香港が 1 名となっている。ヨーロッパからは、アイルランドからの 16 名を筆頭として、74 名が参加した。5 つの招待講演（英、伊、加、米 2）と 3 つのパネル、一般セッションは 89 編の応募論文のうち 48 編が採択され、うち 21 件が通常の発表、27 件がポスターでの発表であった。論文選択の詳細は以下のとおりである。

Submissions/ Acceptance	Accepted Papers by Country
Submitted: 89 Accepted total: 48 = 21 regular papers, 27 posters Rejected: 35 Withdrawn: 6	Canada 8 China 1 France 5 Germany 2 Ireland 6 Japan 6 Korea 1 Mexico 1 Singapore 2 Spain 1 Switzerland 2 UK 3 USA 8 United Nations 1 Vietnam 1

参加者の詳細な内訳は以下の通りである（8 月 17 日現在)。

As of August 17, 2009

participation	
Asia	31
Europe	74
US+CA+BR	166
Africa	2
Middle East	5
sum	278

As of August 17, 2009

Paid	251
waived	28
Sum	279

各トラックの発表や、繰り返しになるがパネルなどの詳細は、IAMT からの資料をそのまま引用すると次のようになる。

Government program track:　16 presentations, one keynote, one panel

Commercial user program track: 15 presentations two keynotes and one panel

5 Invited talks (UK, Italy, Canada, 2 from US)

3 panels (one plenary/commercial – combination of technology for user benefit; one government track (gov't use of MT); one commercial track (introducing technology as part of translator training)

また、これと前後して同じホテルで 6 つのチュートリアル（26 日)、5 つのワークショップ（本会議前、会議中、会議後）が開催された。筆者らが参加したのは、そのうちの第 3 回特許翻訳ワークショップ（The 3rd Workshop on Patent Translation）（共同座長江原暉将山梨英和大教授、横山晶一山形大教授）である。このワークショップは、日本、デンマーク、カナダ、アメリカ、オランダ、ドイツ、中国、韓国などから 37 名の参加者があり、招待講演 2 つと、ユーザからの講演 3 つ、Japio の林昭彦理事長からの挨拶、一般発表 5 つ、パネルディスカッション（Moderator: 潮田明氏（富士通研究所)）が行われた。

本会議では、前日に歓迎晩餐会、2 日目にバンケットも開かれ、各講演では活発な質問が行われた。以下に本会議、ワークショップの概要を報告する。

３．トピックス、主な論文要旨と意見

今回の本会議の特徴は、研究(MT Research)、政府機関による取り組み(Government Users)、商用(Commercial Users)という 3 つの並行セッションが設けられたことである。ごく初期の MT Summit が、やや政策的な会合であった当時に戻ったという印象を受けた。

３．１．基調講演

本会議開催の主催者である Laurie Gerber による Welcome and Conference Overview で、講演の途中で、井佐原 AAMT 会長が、7 月下旬に逝去した田中穂積元会長（IAMT やシンガポールの MT Summit VII の主催者でもあった）の追悼を述べた。

３．２．招待講演

招待講演のうち 3 つの簡単な概要を述べる。

- Johann Roturier: Deploying novel MT technology to raise the bar for quality: Key advantages and challenges

  世界的なウィルス対策ソフトメーカーである Symantec のトップによる講演。MT を採用している理由として、各国のユーザに向けた対応が迅速で低コストで必要であるという理由をあげた。このために、Systran との共同プロジェクトとして、Post Editing を統計的に行う試みや、きちんとした TM を作る試みを紹介した。現在、日、中、仏、独、伊の言語でこれらの試みが進行中であるということも述べた。
- Marco Trombetti: Getting a share of the human translation market with the world's largest Translation Memory

  Translated.net の CEO による講演。この会社は Web を通じて 35,000 人の翻訳者を擁しており、9,000 人のユーザに、80 言語で 10 万語の訳を提供している。これらのサービスを行うためには高速でメンテナンスが容易な TM が必要であり、現在 10 億語の TM を目指している。構築のためには、ドキュメントの固有名詞を隠すなどの処理を自動的に行うことが必要である。最終的には 100 億語規模に拡大して技術的な改善もおこなうとともに、QA Cloud Platform の構築も目指している。
- Pierre Isabelle and Roland Kuhn: MT: The current research landscape

  主催者側、カナダの NRC の研究者による発表である。この講演は、現在の MT の研究と、本会議の発表（ポスターも含む）の概説を結びつけて要領よく解説したもので、この 講演のみを聞くと、全体像が分かるという大変お得なものであった。彼らが特に強調したのは、ユーザの要求と現在の研究とのギャップで、その要素として、1. MT に基づくツール、2. MT System の評価、3. 多言語の問題、4. コーパスの改善とデータマイニング、5. 統計的機械翻訳(SMT)のシステム改善とデコーディング（ここが特に詳しかった）、6. システムの結合、7. 構文と語順変更に分けて述べた。

３．３．パネルディスカッション

パネルディスカッションは、全体のパネルとして、Covering Technologies: What are the benefits for MT users?というものが行われた。Organizer は Mike Dillinger (Translation Optimization Partners)で、パネリストは、Terry Lawler (SDL), Daniel Gervais (MultiCorpora), Alex Yanishevsky (ProMT), Jaap van der Meer (TAUS)である。Organizer に加えて、Discussant として、Paul Bremer (Apptek)が加わっている。内容は、RB + SMT, TM + cloud computing, advanced leveraging + shared TM data, speech recognition + MT など、いろいろな技術の結合が行われているが、それがユーザにとって、質や性能、使い勝手、経済性、安全性などからどのような利益をもたらすかというものであった。

政府機関による取り組み(Government Users)のセッションでは、"Translation in Government" という表題でパネルディスカッションが行われた。Moderator は、Nickolas Bemish（米国防情報局）で、パネリストは Donald Barabe（カナダ公共事業行政サービス省翻訳局次長）、Dan Scott（米国家情報長官室外国語プログラム室ディレクター）、RMK Sinha（インド工科大学教授）、

Chuck Simmons（米航空宇宙諜報センター主任情報アナリスト）、Carl Rubino（米応用機械翻訳センター）である。オーディエンスからも米国防総省や米司法省からの参加者が加わり、政府機関において機械翻訳や各種言語ツールを使うユーザ側の課題について議論が行われた。特に、直接のユーザである翻訳家や分析官に対してどのようにMTなどのツールの使いこなし方を教えたらよいか、言語ツールをユーザ環境へシームレスに統合させることの重要さ、そしてそれらのノウハウを各政府機関間でどう共有すべきかなどについてオーディエンスとパネリストの間で活発な議論が行われた。

もう一つのパネルで、special session として設けられた Preparing Translators for the current technology landscape は、Patricia Phillips Batoma, Roxana Girju, Elizabeth Lowe (University of Illinois), Patricia Minacori (University Paris VII Diderot)という 4 名の女性パネリストによる「翻訳者を志す文系の学生にどうやって技術的なことを教えたらよいか」という面白いパネルであった。参加者は柿田剛史氏(Japio)などを含めてほんの数名であったが、コンピュータに素養のない学生への教育というテーマで議論が行われた。

３．４．一般講演、ポスターセッション

一般講演は、上記のように、主として、研究(MT Research)、政府機関による取り組み(Government Users)、商用(Commercial Users)という 3 つの並行セッションに分かれて進められた。そのうちのいくつかを以下に示す。また、ポスターセッションはオープスペースで 2 回、それぞれ 2 時間にわたって行われた。以下は、だいたい時間順にいくつかのものを述べる。

- Carol van Ess-Dykema, Dennis Persanowsky, Susan Converse, Rachel Richardson, John S. White, Tucker Maney: Metrics to assess translation memory technology
  アメリカの National Virtual Translation Center (NVTC)と Naval Research Laboratory のグループによる発表。国家安全保障のため、英語を中心として、英語へ 50 言語、英語から外国語へ 70 言語ほどの翻訳を、本、放送などについて行っている。TM は、アラビア語、ロシア語、中国語について構築している。
- Hannah Grap: Ranking MT quality: focus on the brand
  商用の MT で、自らが優れていると主張している項目を“ brand ”と名付け、それを Actionable and contribute with brand voice (5)から、Not useful (1)までの 5 段階で評価したもの。
- Larry Rogers: Translation quality: No longer in the eye of the beholder
  翻訳評価において、エラーを記録するための標準を開発しようとする試み。この情報を記録するときに、エラーを、悪い項目、悪い意味、省略、構造的エラー、スペルミス、句読法的エラー、その他の 7 つに分け、統計的モデルを構築しようとした。
- Jordi Carrera, Alex Yanishevsky: Technology for translators: What doesn't kill you, makes you stronger
  ProTM という翻訳会社のグループによる発表（A. Yanishevsky が発表）。TM を使って翻訳者が Post Editing (PE)を行うのに、どう時間短縮を行うかについて論じた。
- Nguyyen Bach, Qin Gao, Stephan Vogel: Source-side dependency tree reordering models with sub-tree movements and constraints
  CMU のグループによる発表。英語からスペイン語、英語からアラビア語（イラク）への翻訳

にあたって、部分木を入れ換えながら翻訳する手法について論じたもの。

- Sirvan Yahyaei, Christof Monz: Decoding by dynamic chunking for SMT
  University of London と、University of Amsterdam のグループによる発表。 SMT を行うにあたり、最も問題視される語順の変更の問題について、目標言語側で変更するのではなく、原言語側で変更しようとする試み。その際に正確な chunking が必要になるが、それを maximum entropy classifier で行っている。
- Midori Tatsumi: Correlation between Automatic Evaluation Metric Scores, Post- Editing Speed, and Some Other Factors
  立見みどり氏(Dublin City University)による発表。Systran で翻訳した文を後編集するときの編集スピードと、BLEU などの評価値との相関を、Pearson correlation を用いて評価したもの。文が単純だと相関が高いという結果が得られているが、相関値は 0.5 程度である。
- Istvan Varga, Shoichi Yokoyama: Transfer rule generation for a Japanese- Hungarian machine translation system
  山形大学のグループによる発表。規則ベースの機械翻訳(RBMT)を構築するときに、日本語とハンガリー語のように、bilingual resource の少ない言語の間でどのように頑健で効率のよい変換規則を作るかについて論じたもの。
- Sherri Condon, Gregory Sanders, Dan Parvaz, Alan Rubenstein, Christy Doran, John Aberdeen, Beatrice Oshika: Normalization for automated metrics: English and Arabic speech translation
  MITRE と NIST のグループによる発表。DARPA の TRANSAC というプロジェクトに基づき、アラビア語と英語の音声翻訳を行ったもの。
- Rod Holland: How to read a machine-translation text
  MITRE の Holland による発表。MT の結果を読み解くのに、知らなくても元の言語にアクセスしたり、図などを参照することによって効果があるとするもの。
- Susumu Bani: The Japan Patent Office's use of MT
  特許庁の番井進氏（特許庁総務部普及支援課特許情報企画室調査第二係長）による発表。日本の特許の Web 化や電子化(IPDL)、MT 化について要領よく述べた後、現状や将来計画について述べた。なお、同様の発表を Workshop でも行っている。
- Satoshi Kamatani, Tetsuro Chino, Kazuo Sumita: Hybrid spoken language translation using sentence splitting based on syntactic structure
  東芝のグループによる発表。釜谷聡史氏が発表を行った。日英の音声翻訳において、構文解析を用いて文の分割を行うことによって翻訳効率を上げるという発表であった。
- Stephen Soderland, Christopher Lim, Mausam Mausam, Bo Qin, Oren Etzioni, Jonathan Pool: Lemmatic Machine Translation
  University of Washington のグループによる発表。文を一種の断片としてコード化（これを Lemmatic encoding と言っている）し、それに基づいて翻訳しようというもの。一種のパターン化、断片化という印象がある。
- Dimitriy Genzel, Klaus Macherey, Jakob Uszkoreit: Creating a high-quality MT system for a low-resource language: Yiddish

Google のグループによる発表。Yiddish は、ドイツを中心とするヨーロッパのユダヤ系の人々の間で話されている言語で、話者が高齢化し、言語資源も少なく、正書法も確立されていない。ドイツ語やヘブライ語からの輸入語も多い。こうした言語を英語に訳す試みについて述べたもの。

- Philipp Koehn, Alexandra Birch, Ralf Steinberger: 462 MT systems for Europe
  University of Edinburgh らのグループによる発表。前回の MT Summit のときの発表をさらに拡張して、EU 内で話されている 23 の言語のほとんどの組み合わせについて MT system を構築し、評価した結果について発表した。
- Philipp Koehn, Barry Haddow: Interactive assistance to human translators using SMT methods
  これも前期と同じグループによる発表。翻訳者の支援のためのインタラクティブなシステムを構築して、その効率について評価した。HP で誰でも試行することができる。
- Michel Simard, Pierre Isabelle: Phrase-based MT in a computer-assisted translation environment
  NRC のグループによる発表。Pierre Isabelle が発表を行った。コンピュータ支援による翻訳環境の中に、句に基づく MT をどのように統合するかについて述べたもの。
- Lucia Specia, Marco Turqui, Zhuoran Wang, John Shawe-Taylor, Craig Saunders: Improving the confidence of MT quality estimates
  Xerox のグループらによる発表。発表は Craig Saunders が行った。参照訳が使えない場合に、文レベルでの MT 出力の質をどのように評価するかという問題について論じた。

３．５．第３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。第１回は Phuket（Chair: 横山晶一）、第２回は Copenhagen（Co-Chair: 辻井潤一、横山晶一）である。

今回は第３回のワークショップで、Chair は江原暉将、横山晶一が共同でつとめた。参加者は約 40 名で、そのうち 2 名が招待講演者、また、user report やパネリストも招待の形をとった。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。

- 招待講演(9:00-10:45)（司会：江原暉将）
  Chair の江原暉将によるこれまでの WS の歴史、今回のプログラムの紹介、Japio 林昭彦理事長の挨拶の後、次の 2 つの招待講演が行われた。

(1) Sophie Mangin: European Machine Translation Programme – Concept, Status and Future Plans
  Munchen などヨーロッパに 5 つの拠点を持つ EPO (The European Patent Office)の現状などを紹介した。EPO においては、特許に用いられる基本言語として、英語、フランス語、ドイツ語の 3 つがある。EPO では、esp@cenet というサービス（日本の IPDL に相当する特許

DB サービス）の一部として、英語とドイツ語、英語とスペイン語の MT が盛んに行われており、その他に、英語とイタリア語、フランス語、ポルトガル語、スウェーデン語との MT が試みられている。ヨーロッパ以外では、EPO, USPTO (The United States Patent and Trademark Office), JPO (The Japan Patent Office), KIPO (The Korean Intellectual Property Office), SIPO (The State Intellectual Property Office of the P.R.C.)による IP5 cooperation という協力の枠組みが進んでおり、この下に MT を含む 10 の foundation が組織されている。この枠組みが将来に向けての情報交換や MT の中心になることが期待されている。

(2) Dan Wang: SIPO's Efforts On Improving Quality of Chinese-English Patent Machine Translation Service

中国 SIPO の王丹氏による講演で、2005 年 4 月に立ち上げ、2007 年 4 月から Single-user version、2008 年 4 月から改良テストを行っている CPMT (China Patent Machine Translation)のプロジェクトについて述べた。興味深いのは、RBMT において、Hierarchical Network of Concepts (HNC)の中で、文を 57 カテゴリに分けて、性能を上げているということであった。なお、RBMT, SMT, EBMT などの統合も将来的には試みる予定である。

・ 一般講演 1 (11:00 -12:00)（司会：Svetlana Sheremetyeva）

(1) Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro: Exploiting Patent Information for the Evaluation of Machine Translation

NTCIR-7 の評価結果に関する発表。発表は内山氏(NICT)が行った。NTCIR-7 における日英言語ペアの test collection と、それを用いた種々のシステムの評価結果について述べた。

(2) Hiroshi Echizen-ya, Terumasa Ehara, Sayori Shimohata, Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Noriko Kando: Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR-7

上記 NTCIR-7 のデータを用いて、AAMT/Japio 特許翻訳研究会のグループが、評価の評価を行った研究に関する発表。発表は江原により行われた。BLEU などの評価値と、人間による評価との相関を求めたものである。

(3) Bin Lu, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang, Oi Yee Kwong: The Construction of a Chinese-English Patent Parallel Corpus

City University of Hong Kong らのグループによる発表。中国語と英語との noisy な parallel corpus を、filter を用いて alignment が取れるようにした研究について発表を行った。

・ ユーザによる講演(13:30-14:30)（司会：横山晶一）

(1) Arti Shah: Translation of Patent Documents at the United States Patent and Trademark Office

USPTO からの発表。翻訳者、文書による翻訳サービス、MT、internal screening system、Systran の利用など、USPTO における現状について述べた。比較的少ない翻訳者で、いろいろな言語に対応しようとしているのが印象的であった。

(2) Young Pyo Kim: KIPO's MT Activities and IP5 Mutual MT Project
韓国の KIPO による発表。2000 年に日韓、2005 年に韓英、2008 年に英韓という MT サービスを立ち上げた韓国の現状と、上記とも関連するが、IP5 への取り組みについて述べた。

(3) Susumu Bani: Current status of MT application in JPO
特許庁の番井進氏による発表。前日の本会議での発表よりは、やや特許翻訳へ絞って、より詳しい計画が説明された。特に、サービス対象を次第に拡大していく方針が示された。

・ **一般講演２**(14:30-15:10)**（司会：梶博行）**

(1) Svetlana Sheremetyeva: An Efficient Patent Keyword Extractor As Translation Resource
デンマークからの発表。翻訳者や MT システムの開発者にとって、キーワードや句といった資源の再利用がいかに有効であるかについてのべたもの。

(2) Shoichi Yokoyama and Masumi Okuyama: Translation Disambiguation of Patent Sentences using Case Frames
横山による発表。文中の語と、動詞との対応を表す格フレームの分類と、翻訳語との違いがどの程度対応しているかを調べたもの。和語動詞の対応には問題があるが、サ変動詞については対応がとれる場合もある。

・ Panel Discussion: Real World Challenges of Patent Translation(15:30 – 17:00) (Moderator:**潮田明**)
各国の特許関連のパネリスト(Georg Artelsmair (EPO), Tao Wang (SIPO), Susumu Bani (JPO), Arti Shah (USPTO), Philipp Koehn (University of Edinburgh))**に対して、現状のサービス、MT の利点と将来の利用計画などについて質問し、それに対する答えに基づいて、討論が行われた。潮田氏の適切なアレンジによって、いろいろな議論が行われた。**

・ **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。

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、政府関係者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。特に今回ワークショップをさらに発展させ、情報を共有しようという気運が盛り上がりつつあるのは心強い。特に今回、政府サイドの発表が多かったことは、安全保障などに MT を利用する機運が高まっていることを示している。日本政府からは特許庁以外に発表がなかったのは残念である。

なお、次回の会議は、AAMT 会長の井佐原均氏が IAMT 会長となって、アモイで 2011 年に開催の予定である。第４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。

平成２１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及び海外調査）

発 行 日　平成22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 ナビックス