2C4-3

機械翻訳サービスのための制約に基づく訳語選択

Constraint-Based Word Selection for Machine Translation Service

松野 淳[*1] 石田 亨[*1] 松原 繁夫[*1]
Jun Matsuno Toru Ishida Shigeo Matsubara

[*1]京都大学情報学研究科社会情報学専攻
Department of Social Informatics, Kyoto University

The inconsistency of word selection is one of the problems occurring in word selections in machine translations and interferes with understanding the expression representing the same thing. We propose the word selection based on the constraint optimization to solve this problem. The consistent word selection considering multiple sentences is performed by solving this constraint optimization problem. The result of evaluation shows that the inconsistency of word selection occurred in 13 percent of noun words contained in multiple sentences and the suitable consistent word selection is performed for 84 percent of them. This means that the word selection considering multiple sentences can improve the quality of machine translation.

1. はじめに

インターネット技術の普及により，世界的にインターネット人口は増加の一途をたどっており，エンドユーザレベルでのグローバル化がますます進むものと思われる．エンドユーザレベルでのグローバル化は，情報のグローバル化をも意味し，インターネット上は多様な言語で表記された膨大な情報で溢れかえっている．そのような状況下で，ある言語を別の言語に翻訳する機械翻訳は有用なサービスである．しかし，n 言語の全ての言語の組み合わせに対して n(n-1)/2 個の直接的な機械翻訳を開発することは現実的ではないため，英語をハブ言語として機械翻訳サービスを連携させることが必要となってくる．

言語グリッドプロジェクト [Ishida 2006] では，言語・文化の壁を乗り越えた異文化コラボレーションのために，言語サービスの利便性と有用性を高めるための活動を行っている．言語グリッドプロジェクトの目的の一つに，標準言語を扱う既存の言語サービスを連結することが挙げられる．この目的を達成することにより，世界中の多様な言語を用いるユーザが，他言語のサービスを母国語で利用できるという恩恵を受けることができる．機械翻訳サービスを連携させている機械翻訳連携もこの目的にかなっており，機械翻訳が存在しない言語間での機械翻訳によるインタラクションを可能とする．

しかしながら，機械翻訳を介したコミュニケーションでは，訳語選択の非一貫性が問題となり，同一物を指す表現を相手が理解することを妨げる [Yamashita and Ishida 2006]．単独の機械翻訳でこのような問題が生じるのであれば，機械翻訳を連携させた場合には，もっと大きな問題となってしまう．この問題を解決するために，原語と機械翻訳により選択された訳語からなる単語の組を文間文脈情報として伝播させる手法が提案されている [Tanaka et al. 2009]．しかし，単語の訳語選択が，その単語が現れている文のうちの 1 文のみに依存しており，文内で得られた原語と訳語の組を文間文脈として伝播してしまっている点が問題として挙げられる．

本研究では，翻訳の対象として文書を扱うとし，一貫しない単語の訳語を，一貫している単語の訳語に基づいて一意に決定するために，訳語選択問題を制約最適化問題として定式化する．本制約最適化問題では，原文書中の各名詞単語に対応して変数があり，各変数の訳文書中の訳語を要素として含む有限集合として各変域がある．複数文で現れる単語に対応する変数については，それぞれの文で共起している単語に対応する変数との間に制約があるため，単語が現れる全ての文を考慮に入れた訳語選択が行われる．また，最適化問題であるため，各単語に対応して単一の訳語選択が行われ，一貫した訳語選択が可能である．つまり，本制約最適化問題を解くことは，複数文を考慮に入れた一貫した訳語選択が行われることを意味する．機械翻訳連携における訳語選択に対しても，各機械翻訳の訳語選択問題を順番に制約最適化問題として定式化し解くことにより適用可能である．

連絡先: 〒 606-8501 京都府京都市左京区吉田本町 京都大学大学院情報学研究科社会情報学専攻 石田・松原研究室, 075-753-5396

2. 背景

2.1 機械翻訳の問題点

文書を対象とした機械翻訳の訳語選択において生じる問題に，訳語選択の非一貫性 [Yamashita and Ishida 2006] が存在する．訳語選択の非一貫性とは，同じ語の訳語が周囲の語によって変化する問題である．図 1 では，単語“ right ”はその出現文によって，“ 右 ”および“ 権利 ”と訳されており，一貫した訳語選択が行われていない．このような問題が生じることにより，同一物を指す同じ単語が異なる意味として理解される恐れがあり，文書の理解を妨げる要因となる．複数の機械翻訳を連携した場合には，単一の機械翻訳の場合と比べて，原語の訳語が一貫して選択されることはさらに困難となる．

機械翻訳連携においては，訳語選択の非遷移性という問題が存在する [Tanaka et al. 2009]．訳語選択の非遷移性とは，翻訳の途中で訳語の意味が変化してしまう問題である．図 2 では，原語“ 欠点 ”の訳語として，英単語“ fault ”を経て，“ 責任 ”という意味のドイツ語“ Schuld ”が選択されている．これは，英単語“ fault ”に“ 欠点 ”，“ 責任 ”などの意味があるために起こる問題である．

2.2 文脈を用いた機械翻訳

機械翻訳における訳語選択の非一貫性を解決するために，原語と訳語の組を文間文脈として機械翻訳サービス間で伝播する手法が提案されている [Tanaka et al. 2009]．各機械翻訳サー

原文 (英):Russia has reserved its right to calim ”teritories discovered by Russians”. Peru has formally reserved its right to make a claim.
→訳文 (日):ロシアは”その右の”領土はロシア人によって発見されたと主張するために予約してします。ペルー共和国を正式にその権利を主張するために予約しています。

図 1: 英日機械翻訳における“ right ”の訳語選択の非一貫性

原文 (日):彼女の欠点は大きな問題だ。
→英文 (英):Her fault is a big problem.
→訳文 (独):Ihre Schuld ist ein grosses Problem.
(日本語訳:彼女の責任は大きな問題だ。)

図 2: 日独機械翻訳連携における訳語選択の非遷移性

ビスは，通常受け取った入力文のみを基に訳文を生成するが，訳文生成時の文脈を次のサービスに伝播し，次のサービスは受け取った文脈に従って訳語選択を行う．このようにすることで，文間文脈を考慮に入れた一貫した訳語選択が可能となる．図 1 の例を用いて説明すれば，1 文目で“ right ”の訳語として“ 右 ”が選択されたので，(right, 右) からなる単語の組を文間文脈とし次の翻訳サービスに伝播し，2 文目の翻訳の際には文間文脈に従った訳語選択，つまり“ right ”の訳語として“ 右 ”を選択する．

しかしながら，この手法では，訳語が 1 文のみ即ち文内文脈を基に決定されているにも関わらず，その文脈を文間文脈として伝播している点が問題として挙げられる．提案手法が，機械翻訳を用いたコミュニケーション支援を目的としており，即時性が求められているということもあるが，初めから全ての文が把握できる文書翻訳に対しては，適当な手法だとは考えられない．文書翻訳においては，文書翻訳文全体を考慮に入れた上での一貫した訳語選択が望まれる．もし，各機械翻訳において，訳語が一貫していない単語に対して一貫した適切な訳語選択を行うことができれば，機械翻訳を連携させることにより訳語の意味が変化してしまうことを防ぐことができるため，訳語選択の非一貫性を解決するだけではなく，非遷移性をも解決することにつながる．

3. 機械翻訳における制約最適化に基づく訳語選択問題の定式化

ここでは，言語 L_1 から言語 L_2 への機械翻訳を考えるものとして，機械翻訳における訳語選択問題を制約最適化問題 [Larrosa and Dechter 2003] として定式化する．まず，n 個の変数 $x_1, \ldots, x_n$ があり，それぞれに対応する変域を $D_1, \ldots, D_n$ とする．各変数は L_1 文書中に現れる名詞単語のうち，L_2 文書中で訳語が名詞単語として得られる各名詞単語に対応して存在しており，n はそのような名詞単語の総数である．D_i は，L_1 単語 x_i の L_2 文書中での訳語を要素として含む有限集合である．各変数の値は，対応する変域集合に含まれる要素である訳語から選択される．L_1 文書中で x_i と x_j ($1 \leq i < j \leq n$) が，同文共起しているならば x_i, x_j 間に制約 f_{ij} が存在し，その制約を“ x_i の訳語と x_j の訳語は意味的関連性がある ”として表し，$var(f_{ij}) = \{x_i , x_j\}$ とする．ここで，目的関数を表すために Wikipedia を用いた semantic relatedness の計算手法を用いる [Gabrilovich and Markovitch 2007]．具体的には，x_i の訳語が言語 L_2 の Wikipedia の各記事に出現した回数から，tf/idf score を用いて x_i の訳語と各記事に対する関連の強さ (重み) を決定し，各記事に対して重み付けされた x_i の訳語ベクトル v_i を得る．そして，x_i の訳語と x_j の訳語ベクトル v_i と v_j をコサイン相関値により比較することで，x_i の訳語と x_j の訳語の意味的関連性を定量的に表すことができる．f_{ij} を x_i の訳語と x_j の訳語の意味的関連性を表わす定量値が a であった場合に，$1 - a$ を出力する関数とすると，目的関数は以下のように表わされる．

$$f^*(X) = \sum_{\{x_i, x_j\} \in V} f_{ij}(X)$$

(集合 V はその要素に，制約が存在する変数の組を含む．) このとき，$f^*(X)$ を最小にするような全ての変数に対する値の割り当てが，この問題における解であり，求めるべき訳語の組である．図 3 に制約最適化に基づく訳語選択問題の定式を示す．

変数集合 $X = \{x_1, \ldots, x_n\}$(x_i:品詞が名詞である原語)
変域集合 $D = \{D_1, \ldots, D_n\}$(D_i:x_i の訳文書中の訳語を要素とする集合)
制約集合 $C = \{f_{ij} \mid x_i, x_j$ が原文書中で同文共起している $\}$
コスト関数 f_{ij}:訳言語の Wikipedia に基づいて，x_i の訳語と x_j の訳語の意味的関連性を定量値として出力する関数
目的関数:$f^*(X) = \sum_{\{x_i, x_j\} \in V} f_{ij}(X)$
(集合 V は要素として，変数間に制約が存在する変数の組を含む)

最適解:$minf^*(X)$ とする全ての変数に対する値の組み合わせ

図 3: 機械翻訳における制約最適化に基づく訳語選択問題の定式

このように定式化することによって，訳語選択の非一貫性が生じうる単語，つまり複数文で現れる単語の訳語選択は，それぞれの文で共起している全ての単語の訳語選択の影響を受ける．また，制約最適化問題では，各変数の値が一意に決定されるため，制約最適化に基づいて定式化された訳語選択問題を解くことは，複数文を考慮に入れた一貫した訳語選択が行われることを意味する．

例として，図 1 で表わされた英語文から日本語文への機械翻訳における英単語“ right ”に対する訳語選択問題を定式化する．まず，英単語“ Russia ”,“ right ”,“ territory ”,“ Russian ”,“ Peru ”にそれぞれ対応して変数 x_1, x_2, x_3, x_4, x_5 と変域 $D_1 = \{$ ロシア $\}$, $D_2 = \{$ 右，権利 $\}$, $D_3 = \{$ 領土 $\}$, $D_4 = \{$ ロシア人 $\}$, $D_5 = \{$ ペルー共和国 $\}$ がある．また，x_1x_2 間，x_1x_3 間，x_1x_4 間，x_2x_3 間，x_2x_4 間，x_2x_5 間，x_3x_4 間に制約があり，それぞれの制約に対するコスト関数を f_{12} (x_1 , x_2) , f_{13} (x_3 , x_3) , f_{14} (x_1 , x_4) , f_{23} (x_2 , x_3) , f_{24} (x_2 , x_4) , f_{25} (x_2 , x_5) , f_{34} (x_3 , x_4) とする．このとき形成される制約ネットワークは，図 4 のように表すことができる．この問題の最適解は，コスト関数の総和である目的関数 $f^*(X)$ を最小にするような全ての変数に対する値 (日訳語)

の割り当てである．本例では，“ right ”の訳語として“ 右 ”ではなく，“ 領土 ”と意味的関連性が特に強い“ 権利 ”が選択される．

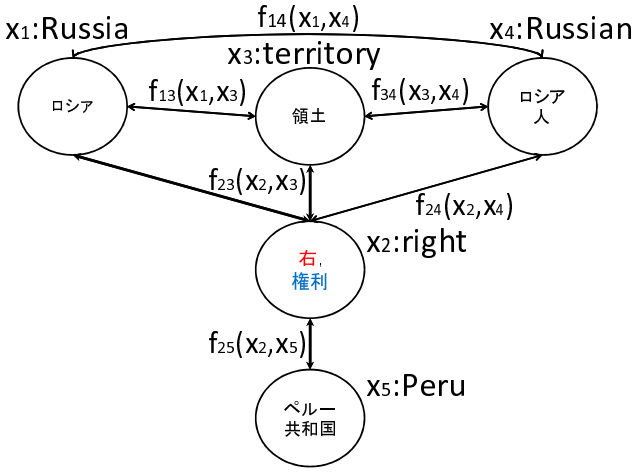

図 4: “ right ”の訳語選択問題を表す制約ネットワーク

4. 実装と評価

4.1 訳語選択システムの実装

機械翻訳における訳語選択問題を制約最適化に基づいて定式化し，分枝限定法により解を求めることにより，一貫した訳語選択を行うシステムを実装した．図 5 は訳語選択システムの処理過程を表したものである．まず，原文書中の 1 文を原文取得機能により得て，その文を機械翻訳サービスを用いて翻訳することで訳文を得る．そして，原文と訳文からなる 2 つの文の組を原文訳文ペア取得機能により得ている．次に，名詞の原語訳語ペア取得機能により多言語対訳辞書を用いて先ほど得た文のペアから，原文に含まれる名詞単語と訳文に含まれるその名詞単語の訳語からなる名詞単語のペア集合を得る．この処理を原文書内の全ての文に対して繰り返し行うことで，最終的に各文に対する名詞の原語と訳語からなるペア集合を得る．ここで得られた各文に対する名詞の原語と訳語からなるペア集合から，制約最適化問題定式化機能により制約最適化に基づいた訳語選択問題を定式化する．最後に，定式化された訳語選択問題の最適解を制約最適化問題求解機能でコスト関数を用いて分枝限定法により得ている．この解は，各名詞の原語に対して唯一つの名詞の訳語からなるペア集合である．

4.2 訳語選択システムの評価

訳語選択システムにより一貫して選択された訳語の評価を行うために，英語 Wikipedia の世界 7 大陸の各記事の各パラグラフの文書に対して，Google Translate を用いた英日機械翻訳を行い，(1)“ 文書内で複数回出現しており，訳語選択の非一貫性が生じなかった名詞単語 ”，(2)“ 文書内で複数回出現しており，訳語選択の非一貫性が生じた名詞単語 ”を各記事ごとにそれぞれカウントした (表 1 参照)．次に，各パラグラフ文書の Google Translate を用いた英日機械翻訳において，訳語選択システムにより各英語名詞単語に対して一貫した訳語選択を行い，(3)“ (2) に分類されていた名詞単語のうち，訳語選択システムにより適切な訳語選択が行われた名詞単語 ”，(4) “ (2) に分類されていた名詞単語のうち，訳語選択システムにより適切でない一貫した訳語選択が行われた名詞単語 ”，(5)

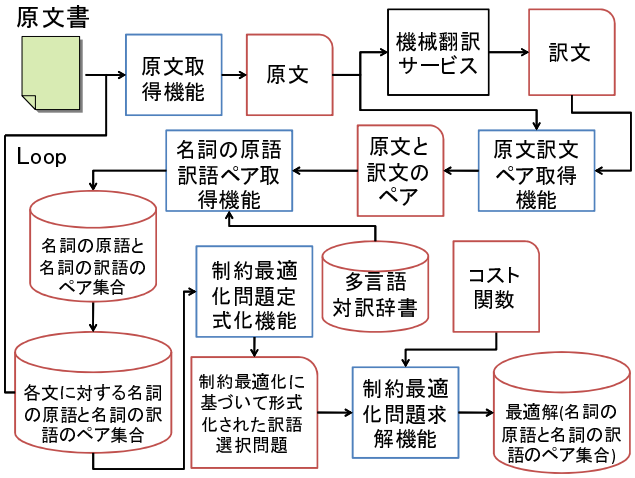

図 5: 訳語選択システムの処理過程

“ (2) に分類されていた名詞単語のうち，訳語を一貫させる必要がないにも関わらず，訳語選択システムにより一貫した訳語選択が行われた名詞単語 ”をそれぞれ記事ごとにカウントした (表 2 参照)．

表 1 は，各記事の各パラグラフの文書内で複数回出現していた名詞単語 635 語のうち 81 語 (13%) において訳語選択の非一貫性が生じており，それらの名詞単語に対しては，訳語選択システムにより複数文を考慮に入れた一貫した訳語選択を可能としたことを示している．また，表 2 は，訳語選択システムにより，訳語が一貫していなかった名詞単語 81 語のうち 68 語 (84%) に対して適切な訳語を与えることが可能であったことを示している．

(3) に含まれる名詞単語の例としては，“ capital ”が挙げられ，ある文においては，適切でない訳語“ 資本 ”が選択されていたが，訳語選択システムにより，適切な一貫した訳語“ 首都 ”が選択されるようになった．また，一貫していない訳語が同じ意味を表している名詞単語 (例えば，訳語として“ 西側 ”，“ 西 ”が選択されていた“ west ”) についても，訳語選択システムによりいずれの訳語が選択されたとしても (3) に含めることとした．一方，(4) に含まれる名詞単語の例としては，“ party ”が挙げられ，ある文においては適切な訳語“ パーティー ”が選択されていたにも関わらず，訳語選択システムにより，適切でない一貫した訳語“ 党 ”が選択されてしまった．(5) に含まれる名詞単語の例としては，“ area ”が挙げられ，“ area ”を“ 面積 ”または“ 地域 ”と翻訳される文に応じて訳語を選択する必要があるにも関わらず，一貫した訳語として“ 地域 ”を選択してしまった．

5. おわりに

本研究では，文書を対象とした機械翻訳における訳語選択の非一貫性を解決するために，訳語選択問題を制約最適化に基づいて定式化した．同文共起する名詞の原語間に制約を課し，最も制約を満たすような全ての原語に対する訳語の組をその文書中で選択されるべき訳語の組とした．このようにして訳語を求めることにより，訳語選択の非一貫性が生じる単語の訳語選択が，その単語が現れる各文で共起している全ての単語の訳語選択の影響を受けるため，複数文を考慮に入れた一貫した訳語選択が可能となった．

また，訳語選択問題を制約最適化に基づいて定式化および定

記事タイトル	(1)	(2)	総数 (1)+(2)
Asia	61	6	67
Africa	106	13	119
North America	53	9	62
South America	78	9	87
Antarctica	102	18	120
Europe	128	25	153
Australia	26	1	27
合計	554	81	635

(1):各パラグラフの文書内で複数回出現しており，訳語選択の非一貫性が生じなかった名詞単語
(2):各パラグラフの文書内で複数回出現しており，訳語選択の非一貫性が生じた名詞単語

表 1: Goolgle Translate による英日機械翻訳における訳語選択の非一貫性に基づく名詞単語の分類

記事タイトル	(3)	(4)	(5)	総数 (2)
Asia	4	1	1	6
Africa	12	0	1	13
North America	8	0	1	9
South America	8	1	0	9
Antarctica	14	3	1	18
Europe	21	2	2	25
Australia	1	0	0	1
合計	68	7	6	81

“ (2):各パラグラフの文書内で複数回出現しており，訳語選択の非一貫性が生じた名詞単語 ”は，以下の (3)〜(5) に分類される．
(3):訳語選択システムにより適切な一貫した訳語選択が行われた名詞単語
(4):訳語選択システムにより適切でない一貫した訳語選択が行われた名詞単語
(5):訳語を一貫させる必要がないにも関わらず，訳語選択システムにより一貫した訳語選択が行われた名詞単語

表 2: Goolgle Translate による英日機械翻訳において訳語選択システムにより選択された一貫した訳語が適切であったか否かに基づいた名詞単語の分類

式化された問題を解く訳語選択システムを実装した．訳語選択の非一貫性が生じていた名詞単語の割合は 13%であり，それらの名詞単語のうち，訳語選択システムにより正しい訳語が選択された名詞単語の割合は 84%であった．このことから，複数文を考慮に入れた訳語選択は，機械翻訳の品質を向上させる可能性があることが分かった．機械翻訳を連携した場合においても，各機械翻訳での訳語選択問題を順番に制約最適化に基づいて定式化し解くことにより，一貫した訳語選択が可能である．

謝辞 京都大学グローバル COE プログラム「知識循環社会のための情報学教育研究拠点」並びに総務省戦略的情報通信研究開発推進制度から助成を受けた．

参考文献

[Ishida 2006] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration. IEEE/IPSJ Symposium on Applications and the Internet. SAINT-06, pages 96-100(2006).

[Yamashita and Ishida 2006] Yamashita ,N. and Ishida ,T.: Effects of Machine Translation on Collaborative Work. In Proceedings of International Conference on Computer Supported Cooperative Work. CSCW-06, pages 515-523(2006).

[Tanaka et al. 2009] Tanaka ,R., Murakami ,Y. and Ishida ,T.: Context-based approach for pivot translation services. In Proceedings of International Joint Conference on Artificial Intelligence. IJCAI-09, pages 1555-1561(2009).

[Larrosa and Dechter 2003] Larrosa ,J. and Dechter, R.: Boosting search with variable elimination in constraintoptimization and constraint satisfaction problems. Constraints 8, pages 303-326(2003).

[Gabrilovich and Markovitch 2007] Gabrilovich, E. and Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In Proceedings of International Joint Conference on Artificial Intelligence. IJCAI-07,pages 1606-1611(2007).