2A2-1

文脈情報を用いた機械翻訳サービスの連携

Context-based Coordination of Machine Translation Services

田仲　理恵[*1*2*]　石田　亨[*1*2]　村上　陽平[*2]
Rie Tanaka　Toru Ishida　Yohei Murakami

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

[*2]情報通信研究機構
National Institute of Information and Communication Technology

Machine translation services on the Web are getting popular. For intercultural communication, multiple translation services are often combined. Furthermore, since English is a hub language, it is often necessary to cascade different machine translators to realize translations between non-English languages. As a result, word sense is often changed because of the *inconsistency*, *asymmetry* and *intransitivity* of word selections among machine translation services. Therefore we propose a context-based coordination framework in which context is propagated among cascaded translation services by referring multilingual equivalent terms. We regard machine translation services as black boxes, and achieve substantial quality improvements by Web service coordination. This means that context-based Web service coordination improves the quality of machine translation.

1. はじめに

Web サービスの連携は人工知能にとっても興味深い問題である [Hassine et al. 2006, Singh 2003, Traverso and Pistore 2004, Wu et al. 2003]．例えば，日英翻訳と英独翻訳を合成すると日独翻訳が生まれる．n 言語に対して $n(n-1)$ 個の機械翻訳を開発することは実際的でないため，この連携は現実の問題である．日本や中国などのアジア言語から英語以外の欧州言語への翻訳には，機械翻訳の連携が必須である．例えば，日本の中学校では，ブラジル人の生徒のために，日葡翻訳を必要としている．このような背景から，言語グリッドプロジェクト [Ishida 2006] では，言語・文化の壁を越えたコラボレーションのために，様々な言語資源の活用と連携を目指した活動を行っており，その一環として機械翻訳サービスの連携を目指している．しかし，機械翻訳連携においては，複数のサービスの訳語選択が一貫しないことにより，翻訳途中で語義が変わってしまう問題がある．機械翻訳を用いたコミュニケーションを観察した結果，翻訳の非一貫性，非対称性により訳語選択が一貫せず，参照語の形成が困難となるなど会話の共有基盤が形成されにくいことがわかっている [Yamashita and Ishida 2006]．機械翻訳連携はこの困難に拍車をかけ，日独の折り返し翻訳を行うと “蛸” が “烏賊” に変わるなどの誤訳を生み出してしまう．

自然言語研究では，Linguistic Annotation Language を用いて原文の構文解析結果を訳文に埋め込むことで問題の解決を図っている [Kanayama and Watababe 2003]．しかし，機械翻訳プログラムに手を加えられない Web サービスの利用者にとっては有効な解決法ではない．他方では，Web サービスを有機的に連携させることによって，この問題を解決するアプローチとして，Web サービス合成時に入出力の型を記述して整合を図る *WS-BPEL* などの開発や，プランニング技術を用いて入出力の整合性を自動的に保証する研究も行われている [Liu et al. 2007]．しかし，機械翻訳連携で問題になるのは入出力の型ではない．*WS-Coordination* では，サービスの ID やポートなどの情報をサービス全体に伝達して連携を図っている．本研究ではこのアイデアを利用し，次の課題に取り組む．

多言語の同義語データを作成する
複数言語で生成される訳文の意味の一貫性を保つため，多言語の同義語データを作成する．多言語の同義語データは一部の言語間で人手により提供されているだけであり，既存の言語資源を元に同義語データを自動生成する．

文脈情報を伝播させて連携を図る
翻訳対象の文章，またはその文章を含むテキスト全体から文脈を検出し，機械翻訳サービス間で伝播していくことによって連携を実現する．本研究では文脈の検出方法はすでに存在するとし，検出された文脈を伝播して利用することを目標とする．

本論文の提案手法は，言語処理技術によって機械翻訳連携の品質の向上を図るのではなく，言語処理技術はブラックボックスとして扱い，文脈を用いた Web サービスの連携を実現することによって品質を向上させるというものである．この方法の利点は，機械翻訳の内部に手を加えることなく，存在する Web サービスをそのまま用いて品質の高い連携を実現できることである．本稿では，複数の機械翻訳サービスを連携させた日独折り返し翻訳を例に，提案手法を用いて訳語選択を一貫させることにより翻訳品質が大幅に向上することを示す．この事実は，Web サービス連携の技術が，自然言語処理の範疇と考えられてきた機械翻訳の品質向上に寄与することを示している．

2. 文脈を用いた連携

2.1 機械翻訳連携で生じる問題例

機械翻訳サービスの連携において生じる問題点は，非一貫性，非対称性，非遷移性の 3 つに分類できる．図 1 にそれぞれの例を示す．(a) は，同じ語の訳語が周囲の語によって変化する非一貫性の問題例である．単語 “paper” は文章によって “論文” と訳される場合と “紙” と訳される場合があり，一連の対話でこの現象が起こると問題となる．(b) は，訳語を元の言語に訳し戻した場合に原語に戻らない非対称性の問題例であり，入力文の単語 “パーティー” が “党” に変わっている．これは，

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
(*現所属 日本電気株式会社 C&C イノベーション研究所)

〈 Case 1 〉
アメリカ人 (英): Please add that picture in this paper.
⇒ 訳文 (日): どうぞ，その写真をこの論文の中に追加しなさい.
〈 Case 2 〉
アメリカ人 (英): Please send me this paper.
⇒ 訳文 (日): どうぞ，この紙を私に送りなさい.

(a) 訳語選択の非一貫性

日本人 (日): 私たちは昨日パーティーをしました.
⇒ 訳文 (英): There was a party yesterday.
アメリカ人 (英): How was the party?
⇒ 訳文 (日): 党はどうだったか？

(b) 訳語選択の非対称性

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

(c) 訳語選択の非遷移性

図 1: 機械翻訳サービスの連携において生じる問題

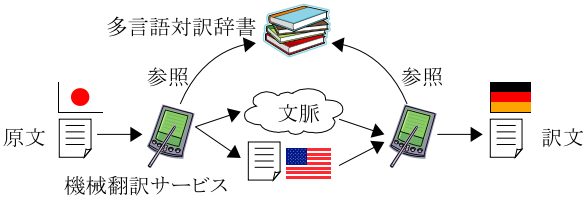

図 2: 文脈情報を用いた連携のフレームワーク

英単語 “party” が両方の意味を含むためである．(c) の非遷移性の例文では，入力文の単語 “欠点” が英単語 “fault” を経て，責任という意味のドイツ語 “Schuld” に変わっている．この単語には欠点の意味はない．これは，英単語 “fault” に欠点，責任などの意味があり，後者に対する訳語として “Schuld” が選択されたためである．(a) と (b) は対話での例だが，機械翻訳を直列に繋げて使用していると考えると，問題は全て，多義語に対する訳語選択に一貫性がないために生じているといえる．

2.2 サービス連携のフレームワーク

文脈を考慮した連携を行うため，各機械翻訳サービスが多言語対訳辞書を参照しながら文脈を伝播する図 2 のモデルを提案する．各サービスは通常は，受け取った入力文のみを見て訳文を生成する．このようなサービスをラッピングし，訳文生成時に考慮した文脈を次のサービスに伝播する機能と，受け取った文脈に従って訳語選択を行う機能を追加する．異なるサービスは異なる言語を扱うため，文脈の伝播時に多言語の同義語データを格納した多言語対訳辞書を参照する．多言語対訳辞書は通常の対訳辞書を多言語に拡張したもので，見出し語に対して複数の言語の訳語が得られる辞書である．

3. 多言語対訳辞書

多言語の同義語の集合である多言語対訳辞書は，一般の対訳辞書を複数組み合わせて作成する*1．先行研究において，二言語間の対訳辞書を組み合わせて二言語の概念の対応を取る

*1 EuroWordNet[Vossen 1998] のように，人手で同義語データが作成されている言語もある．

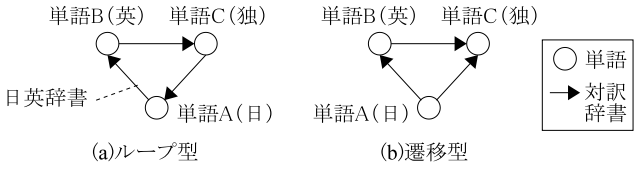

図 3: 三角形の形状

手法が提案されている [Tokunaga and Tanaka 1990]．本研究ではこの手法を三言語での同義語（以降では三つ組と呼ぶ）を取得する手法に拡張する．まず，単語の関係を，単語を頂点，単語間の対訳関係を有向枝で表したグラフで表現する．対訳辞書には辞書引きの方向があるため，枝は有向となる．このようなグラフにおいて，三角形を形成する 3 つの単語を三つ組とみなす．三角形には，図 3 に示すようにループ型と遷移型の 2 種類がある．ループ型は，ある言語を起点に辞書引きを 3 回行い，元の言語に戻ってくる経路である．遷移型は，ある言語から別言語への経路が 2 通りある形状である．このような三角形の経路に従って三つ組が取得できる．

ただし，共通の意味を持たない三つ組も存在する．例えば，単語 A が語義 C_1 と C_2，単語 B が C_2 と C_3，単語 C が C_3 と C_1 を持つ場合には，三角形は形成されるが，3 つの単語に共通の語義は存在しない．各単語が n 個ずつの語義を等確率で持つとして計算すると，3 つの単語に共通の意味が存在する可能性は $n = 2$ に対して 0.83，$n = 3$ に対して 0.91 となり，n が大きくなるにつれて 1 に近づく．実際には語義の使用頻度に偏りがあるため，共通の意味がない確率は，理論上は存在するが無視できるほど小さいといえる．よって，この手法を 4 言語以上に拡張するには，三言語で獲得した三つ組を組み合せればよい．例えば，日・英・独・仏であれば，日英独の三つ組と英独仏の三つ組を取得し，それらを組み合わせて四つ組とする．

本研究と類似の研究に，第三言語を介して対訳辞書を作成する研究がある [田中他 1998]．この研究では，日英辞書と英仏辞書を接続して日仏辞書を作成する際に，仲介した英語の多義性により日本語と異なる意味の仏語が得られる問題を扱っており，仏英辞書および英日辞書を用いた逆引きにより解決を図っている．我々の手法は，各言語間に対訳辞書が存在する場合に，より信頼性の高い同義語を獲得する手法であり，前提と目的が異なっている．いずれかの言語間の辞書が不足している場合にも翻訳連携を行うのであれば，[田中他 1998] で提案されている逆引き法のような同義語獲得手法が必要となる．

4. 文脈情報を用いた連携

図 4 に連携の詳細な処理過程を，図 5 に処理を実現するアルゴリズムを示す．機械翻訳サービスは入力文を受け取って訳文を返すだけのブラックボックスとみなし，文脈を参照して問題のある訳語を検出し，別の訳語で置き換えることで連携を実現する．元の機械翻訳サービスと訳語の検出および補正機能を含めてラッピングした全体を新たな機械翻訳サービスとし，実行時にはラッピングされたサービス同士を接続する．伝播する文脈は，図 4 に示すように翻訳対象の文の文脈（文内の文脈）と翻訳対象の文を含む文書全体の文脈（文間の文脈）の二種類に分類できる．文間の文脈が利用できる場合には，複数の文章にわたって同一の文脈に従った翻訳を行うことができる．

ラッピングされたサービスでは，まず始めに既存の機械翻訳サービスの機能により入力文を翻訳する．アルゴリズムにおい

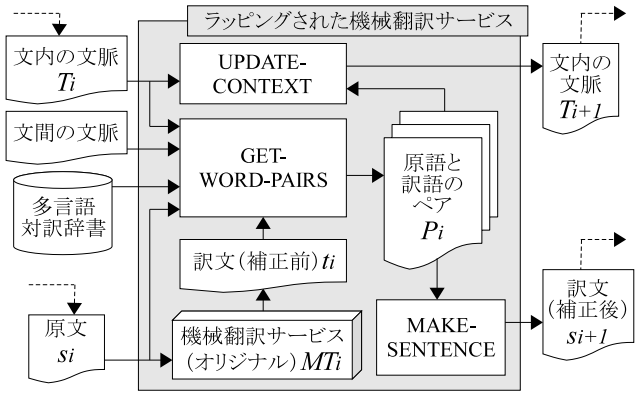

図 4: 連携の詳細な処理過程

ては，i 番目の機械翻訳 MT_i を入力文 s_i と訳文 t_i のペアの集合として定義しており，補正前の訳文 t_i を得ている（10 行目）．その後，GET-WORD-PAIRS を用いて得られた訳語が伝播された文脈に合っているかどうかを調べ，異なる場合には多言語対訳辞書の語を用いて入れ替えを行う（17〜32 行目）．GET-WORD-PAIRS では，原文 s_i の単語 o_i1 つずつについて，GET-TRANSLATED-WORD により訳文 t_i 中のどの語に翻訳されたかを調べる．この部分は，形態素解析と対訳辞書などにより実装できるため詳細な処理は省略する．訳語が c_{i+1} であると分かると，次に文脈と合っているかどうかを調べる．アルゴリズムにおいては，文内の文脈のみを利用した最もシンプルな実装を記述している．文脈は，n 言語の多言語の組である n-tuple の集合 T_i として表す．ある n-tuple $(w_1, ..., w_n)$ において，$w_j (1 \leq j \leq i)$ は各サービスの入力 s_j に含まれており，すでに確定した訳語である．i+1 番目以降の語はその同義語，つまり i 回目以降の翻訳で訳語として使用すべき語を示す．原語 o_i と訳語 c_{i+1} の両方を含む n-tuple が文脈 T_i の中に含まれていれば，訳語はそのまま用いる．そうでなければ，c_{i+1} は過去の翻訳で用いられた訳語と同義ではないと判断し，o_i を含む n-tuple のうちいずれかを選択して置き換えを行う．選択方法は，単語の出現頻度や翻訳のログなど様々である．

原語と補正した訳語のペア P_i が作成されると，MAKE-SENTENCE により補正後の訳文 s_{i+1} を作成する（処理の詳細は省略する）．また，P_i の訳語を用い，関数 UPDATE-CONTEXT（38〜43 行目）により文脈を更新する．文脈の更新は，T_i のうち，P_i に含まれる訳語を含む n-tuple のみを残すように絞り込むことで行い，作成された新しい集合 T_{i+1} を次のサービスに伝播する新しい文脈とする．なお，ここでは機械翻訳サービスを完全なブラックボックスとみなす場合の実現方法を示したが，内部の処理過程を修正できる場合にも，同様の手順により連携を行うことが可能である．

5. 実装と評価

日英独の三つ組を作成し，翻訳サービスの連携を行った．簡単のため，品詞は名詞に限定した．使用した辞書および取得した三つ組を表 1 に示す．遷移型は日本語を起点としている．ループ型で 15,627 組，遷移型で 13,757 組の三つ組が取得でき，重複を除いて統合すると合計で 21,914 組となった．作成した三つ組がどの程度の文書の翻訳に用いることができるかを，Web5 億文コーパスから取得した日本語の単語の出現頻度データ [河原, 黒橋 2006] を用いて調べたところ，三つ組はコーパスに登場する名詞の 58 %，全品詞の 40 %をカバーできるという結果になった．また，出現頻度の高い単語を含む順に三つ組を使用した場合には，約 6000 組で名詞の 50 %，全品詞の 38 %をカバーできるという結果になった．比較的少ない三つ組で文書を効率よくカバーできていることがわかる．

Algorithm COORDINATE-TRANSLATION
-SERVICES(MT, s_1) **return** s_{n+1}

1: MT /* An ordered list of cascaded machine translation services combined ($MT = \{MT_1, MT_2, ..., MT_n\}$) */
2: s_i /* Original input sentence of MT_i */
3: t_i /* Output sentence of $MT_i (t_i \neq s_{i+1})$ */
4: $MT_i = (s_i, t_i)$ /* An original machine translation service; a set of pairs of s_i and t_i */
5: o_i /* A word in sentence s_i */
6: T_i /* A set of n-tuples $(w_1, ..., w_n)$, where w_k is included in $s_k (k \leq i)$; All n-tuples are registered in n-Lingual Dictionary */
7: P_i /* A set of pairs (o_i, o_{i+1}), where $o_i \in s_i$ and $o_{i+1} \in s_{i+1}$, modified translated word of o_i */
8: $T_1 \leftarrow \{(w_1, ..., w_n) | w_1 \in s_1\}$
9: **for each** MT_i **in** MT **do**
10: $t_i \leftarrow MT_i(s_i)$
11: $P_i \leftarrow$ GET-WORD-PAIRS(s_i, t_i)
12: $s_{i+1} \leftarrow$ MAKE-SENTENCE(P_i)
13: $T_{i+1} \leftarrow$ UPDATE-CONTEXT(T_i, P_i)
14: **end loop**
15: **return** s_{n+1}
16:
17: **function** GET-WORD-PAIRS(s_i, t_i) **return** P_i
18: c_{i+1} /* A translated word of o_i, $c_{i+1} \in t_i$ */
19: $P_i \leftarrow \phi$
20: **for each** word o_i **in** s_i **do**
21: $c_{i+1} \leftarrow$ GET-TRANSLATED-WORD(o_i, t_i)
22: **for each** n-tuple $(w_1, ..., w_n)$ **in** T_i **do**
23: **if** $(o_i, c_{i+1}) \subset (w_1, ..., w_n)$ **then**
24: $P_i \leftarrow P_i \cup \{(o_i, c_{i+1})\}$
25: **end if**
26: **end loop**
27: **if** $(o_i, c_{i+1}) \notin P_i$ **then**
28: $o_{i+1} \leftarrow i$+1th word in n-tuple selected from $\{(w_1, ..., w_n) | o_i \in (w_1, ..., w_n)\}$
29: $P_i \leftarrow P_i \cup \{(o_i, o_{i+1})\}$
30: **end if**
31: **end loop**
32: **return** P_i
33:
34: **function** UPDATE-CONTEXT(T_i, P_i) **return** T_{i+1}
35: $T_{i+1} \leftarrow \phi$
36: **for each** pair (o_i, o_{i+1}) **in** P_i **do**
37: $T_{i+1} \leftarrow T_{i+1} \cup \{(w_1, ..., w_n) | (w_1, ..., w_n) \in T_i$ and $(o_i, o_{i+1}) \subset (w_1, ..., w_n)$
38: **end loop**
39: **return** T_{i+1}

図 5: 機械翻訳連携アルゴリズム

表 1: 使用した辞書および作成した三つ組

(a) 使用した対訳辞書

対訳辞書名	見出し語数
ジーニアス和英辞書	31,926（名詞）
コンサイス和独辞書	38,487（全品詞）
オックスフォード英独辞書	31,180（名詞）
クラウン独和辞書	34,255（名詞）

(b) 作成した三つ組

型	三つ組の数
ループ型	15,627 組
遷移型	13,757 組
合計（重複を除く）	21,914 組

〈 評価値が 4（Most）から 5（All）に改善された例 〉
原文 A: トラックが道を塞いでいた.
連携技術なしの訳文 B: トラックは方法を妨げた.
連携技術ありの訳文 C: トラックは道を妨げた.
〈 評価値が 3（Much）から 5（All）に改善された例 〉
原文 A: 社長は労働者を使う.
連携技術なしの訳文 B: 大統領は労働者を使う.
連携技術ありの訳文 C: 社長は労働者を使う.

図 6: 連携技術の適用による効果

次に，日英翻訳・英独翻訳・独英翻訳・英日翻訳を連結させた日独折り返し翻訳に対し，作成した三つ組を用いて連携を行い，予備評価を行った．品質の評価は，日本語の原文 A と，連携技術を用いずに翻訳した折り返し日本語文 B，連携技術を用いた折り返し日本語文 C をそれぞれ比較し，原文 A の意味を B および C がどの程度表現できているかを 5 段階（5: All, 4: Most, 3: Much, 2: Little, 1: None）で評価するという方法で行った．日本語の原文は NTT の自然言語グループが提供する機械翻訳の性能評価用例文[*2]を用い，訳文 B と C が異なる結果となった文 100 文をランダムに選択した．評価者は日本人 3 人とした．訳文 B の評価結果と訳文 C の評価結果には，Welch の検定により 98 %の信頼性で違いが現れた．

連携技術を用いることで，平均で 100 文中 41 文で評価値の向上が見られ，評価値は 100 文全体で 0.47 ポイント上昇した．図 6 では，日本語の単語 “道” が “方法” に誤訳されていた部分，および “社長” が “大統領” に誤訳されていた部分が改善されている．これらの誤訳は仲介した英単語 “way” と “president” が多義であったたために生じていた．また，連携技術を適用した訳文 C では，訳文 B において評価値が 4 であった文の 34 %，3 であった文の 32 %，2 であった文の 49 %，1 であった文の 60 %で評価値が改善されるという結果になった．

6. おわりに

本研究では，機械翻訳サービスの連携における訳語選択の非一貫性，非対称性，非遷移性の解決のため，各々の機械翻訳サービスはブラックボックスとみなした上で連携を図る手法を提案した．また，それにより翻訳品質の大幅な改善が見られることを示した．本研究の貢献は下記の通りである．

対訳辞書を組み合わせて多言語に拡張する
複数の対訳辞書を元に，登場する単語と単語間の対訳関係をグラフで表し，グラフの構造を利用することで多言語の同義語の獲得を行う手法を提案した．

多言語の同義語情報により訳語選択を制御する
作成した多言語の同義語データを用い，伝播された文脈に基づいて訳語を選択するように機械翻訳を制御する手法を提案した．このアイデアの実現として，翻訳対象文の文脈のみを用いたシンプルなアルゴリズムを示した．

提案手法による翻訳品質の改善は，自然言語分野や機械翻訳を用いた多文化でのコラボレーションの分野にとどまらず，様々なドメインにおいて生成される Web サービスの連携において重要な役割を果たすものであるといえる．

*2 http://www.kecl.ntt.co.jp/mtg/resources/index.php

参考文献

[Hassine et al. 2006] Hassine, A. B., Matsubara, S. and Ishida, T.: A Constraint-Based Approach to Horizontal Web Service Composition. *ISWC-06*, pp.130–143 (2006).

[Singh 2003] Singh, M. P.: Distributed enactment of multiagent workflows: temporal logic for web service composition. *AAMAS-03*, pp.907–914 (2003).

[Traverso and Pistore 2004] Traverso, P. and Pistore, M.: Automated Composition of Semantic Web Services into Executable Processes. *ISWC-04*, pp.380–394 (2004).

[Wu et al. 2003] Wu, D., Parsia, B., Sirin, E., Hendler, J. A. and Nau, D. S.: Automating DAML-S Web Services Composition Using SHOP2. *ISWC-03*, pp.195–210 (2003).

[Ishida 2006] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration. *SAINT-06*, pp.96–100, keynote address (2006).

[Yamashita and Ishida 2006] Yamashita, N. and Ishida, T.: Effects of Machine Translation on Collaborative Work. *CSCW-06*, pp.515–523 (2006).

[Kanayama and Watababe 2003] Kanayama, H. and Watanabe, H.: Multilingual Translation via Annotated Hub Language. *MT-Summit IX*, pp.202–207 (2003).

[Liu et al. 2007] Liu, Z., Ranganathan, A. and Riabov, A. V.: A Planning Approach for Message-Oriented Semantic Web Service Composition. *AAAI-07*, pp.1389–1394 (2007).

[Vossen 1998] Vossen, P. (ed.): EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht, Netherlands: Kluwer (1998). See: http://www.hum.uva.nl/ ewn/.

[Tokunaga and Tanaka 1990] Tokunaga, T. and Tanaka, H.: The Automatic Extraction of Conceptual Items from Bilingual Dictionaries. *PRICAI-90*, pp.304–309 (1990).

[田中他 1998] 田中久美子, 梅村恭司, 岩崎英哉.: 第三言語を介した対訳辞書の作成. 情報処理学会論文誌, Vol. 39, No. 6, pp.1915–1924 (1998).

[河原, 黒橋 2006] 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 171-12, pp.67–73 (2006).