文中の接続助詞「が」に着目した翻訳単位間の意味関係の推定

酒井 浩之
豊橋技術科学大学 知識情報工学系
sakai@smlab.tutkie.tut.ac.jp

松原 茂樹
名古屋大学 情報基盤センター
matubara@nagoya-u.jp

増山 繁
豊橋技術科学大学 知識情報工学系
masuyama@tutkie.tut.ac.jp

稲垣 康善
豊橋技術科学大学
inagaki@tut.ac.jp

1　はじめに

話者が話し始めると，それに追従しながら翻訳を行う同時翻訳において，入力文を翻訳単位（同時翻訳の処理単位.「単語」や「節」など.）に分割し，それぞれを翻訳したのち連結することで，同時翻訳性能の飛躍的な向上が期待できる [5][1]．例えば,「搭乗開始時刻の方ですが一時間遅れとなりますので十二時二十分を予定しております。」という文を「搭乗開始時刻の方です」「一時間遅れとなります」「十二時二十分を予定しております。」という 3 つの翻訳単位に分割し，それぞれの翻訳文「The boarding time」「also one hour behind schedule」「it is going to be twelve twenty.」を連結して訳文を生成する．（図 1 の例文 1 を参照.）

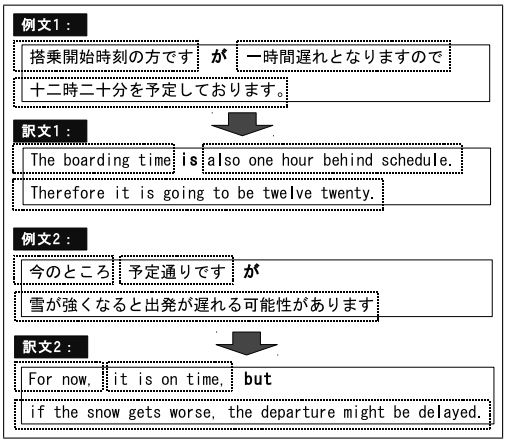

図 1: 入力文の分割・翻訳・連結による訳文生成

同時翻訳の翻訳単位としては，文に比べて十分に小さく，単語よりも大きな言語単位を採用することが考えられる．Kashioka らは日英翻訳の処理単位として「節」を用いることを提案している [1]，また，入力文を翻訳単位に分割する手法として，笠らは，長い文を機械学習手法によって翻訳単位に分割する手法を提案している [8]．しかし，それぞれの翻訳単位の訳文を連結する際に，翻訳単位間の意味関係の推定を行う必要がある．例えば，図 1 の例文 2 では,「今のところ予定通りですが雪が強くなると出発が遅れる可能性があります。」の翻訳単位の翻訳文,「it is on time」と「if the snow gets worse, the departure might be delayed.」の間に「but」を挿入したほうが自然な訳文を生成できる．これは,「今のところ予定通りです」と「雪が強くなると出発が遅れる可能性があります。」の間に出現している接続助詞「が」が「逆接」の意味を表しているからである．しかし，例文 1 の「搭乗開始時刻の方です」と「一時間遅れとなります」の間に出現している接続助詞「が」は「提示」の意味を表しているため，翻訳文を連結する際に「but」を挿入することはできない．

[1]http://www.el.itc.nagoya-u.ac.jp/~matubara/kaken/si/

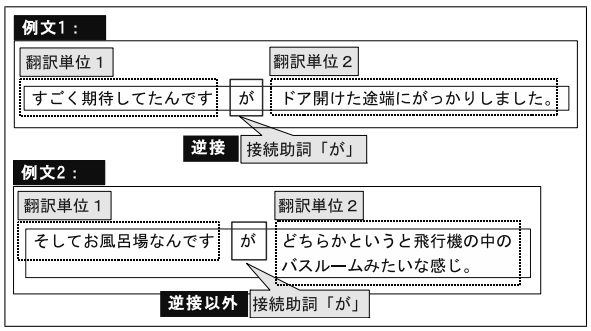

図 2: 接続助詞「が」の意味関係の推定

そこで，本研究では接続助詞「が」に着目してそのあいまい性を解消し，翻訳単位間の意味関係の推定を行う．具体的には，図 2 のように，入力された文の接続助詞「が」の前後の翻訳単位が「逆接」であるか「逆接以外」の意味関係であるかを推定する．推定された結果を使用することで，翻訳単位の翻訳文を連結して全体の訳文を生成する際に，適切な接続詞を挿入することができるようになり，自然な訳文を生成できることが期待できる．

2　関連研究

2 つの文の意味関係の推定を行う研究として，山本らは談話処理のうち文間の接続関係を同定するタスクを設定し，入力として接続詞を持つ文とその前文の連続 2 文を与え，この接続詞を与えない場合に接続関係を同定する手法を提案している [11]．山本らは，大量の Web 文書を用いて与えられた 2 文に最も近い用例を探すことで 2 文間の接続関係を推定している．そして，大量の Web テキストを用例として利用することで，接続関係を推定するための規則を作ることなく接続関係を同定している．し

かしながら，本研究による手法の適用先として，話者が話し始めるとそれに追従しながら翻訳を行う同時翻訳を想定しており，手法の即時性が要求される．そのため，山本らのように，大量の Web 文書等の大規模コーパスを用いて，与えられた 2 文に最も近い用例を検索することで意味関係を推定するアプローチは，本研究には適さないと考える．

意味関係の推定を単語単位で行う研究として，田中らは，名詞句「NP_1 の NP_2」の意味関係を動詞 v と NP_1,NP_2 間の格 c_1, c_2 の 3 字組 (v, c_1, c_2) で記述し，その (v, c_1) または (v, c_2) を単文中の係り受け情報を用いて抽出する手法を提案している [9]．例えば「琵琶湖の写真」の意味関係は，(撮る，を$_{\text{目的格}}$，に$_{\text{目標格}}$) となる．また，文構造のあいまい性を解消する研究として，Nakov らは，前置詞句の修飾先が名詞であるか動詞であるかを判別するタスクにおいて，大量の Web コーパスを訓練データとして使用する手法を提案している [6]．それらに対して，本研究では接続助詞「が」に着目して 1 文における翻訳単位間（複数の文節で構成される）の意味関係の推定を行っているが，接続助詞「が」を含む翻訳単位間の意味関係を推定するには，名詞句や前置詞句よりも大きな言語単位を考慮する必要がある．

3 接続助詞「が」の意味関係の推定

3.1 学習データの自動生成

本研究では, 入力として接続助詞「が」を含む文を入力し, その意味関係が「逆接」であるか「逆接以外」であるかを推定する．図 2 の例では，例文 1 の「すごく期待してたんですが ドア開けた途端にがっかりしました。」の接続助詞「が」は「逆接」を表しているため,「逆接」と判定する．例文 2 の「そしてお風呂場なんですが どちらかというと飛行機の中のバスルームみたいな感じ。」の接続助詞「が」は,「提示」を表しているため,「逆接以外」と判定する．

判定には機械学習手法を用いるが，機械学習のための十分な量の学習データを人手で作成するには膨大なコストが必要となる．そこで，本研究では，学習データを自動的に作成し, それを使用して機械学習を行う．具体的には, 図 3 のように,「しかし」のような「逆接」の意味をもつ接続詞を文頭にもつ文とその前の文に出現する単語列を正例,「そして」や「それでは」のような接続詞を文頭にもつ文とその前の文に出現する単語列を負例として, 学習データを自動生成する[2]．ここで, 本手法の正例で使用した「逆接」の意味をもつ接続詞の一覧を表 2 に, 負例で使用した接続詞の一覧を表 1 に示す．

[2]使用した接続詞は，山本らの研究 [11] において「累加」「逆接」「転換」に分類された接続詞から，学習データとして使用した読売新聞 308, 388 記事に 20 回以上出現したものの中から選択した．

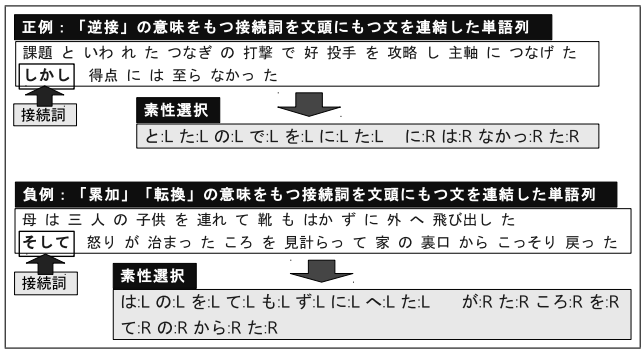

図 3: 学習データの自動生成

表 2: 正例で使用した接続詞

しかし，でも，ところが，だが，しかしながら けれども，けれど，ですが，だけど，けど

3.2 意味関係の推定

自動的に作成された学習データを使用して，接続助詞「が」の意味関係の推定を行う分類器を作成する．ここで, 素性として機能語（助詞，助動詞，形式名詞）を選択する．そして，学習データにおける接続詞を含む文に出現する機能語に「R」のフラグを付与し，また，接続詞を含まない文（接続詞を含む文の前に連結した文）に出現する機能語に「L」のフラグを付与する（図 3 を参照）．本論文では，読売新聞 308, 388 記事を使用し，学習データとして 118, 724 個を自動的に生成した．そして，SVM[10] で分類器を生成した．なお，カーネルとして線形カーネルを使用した．また，実装にあたり，SVM^{light} [3]を使用した．

次に，分類器を使用して接続助詞「が」の意味関係を判定する．ここで，テストデータとして名古屋大学同時通訳データベース [7][4][4]の独話データベースを用いた．テストデータとして独話データベースを用いた理由は，独話データには接続助詞「が」を含む 416 事例において,「逆接」の意味関係である場合が 157 事例,「逆接以外」の意味関係（「提示」等）である場合が 259 事例であるため, 手法適用の必要性が高いからである．それに対して，対話データでは，接続助詞「が」を含む 206 事例において,「逆接」の意味関係である場合が 16 事例,「逆接以外」の意味関係である場合が 190 事例であり，手法適用の必要性が著しく低い．

まず，テストデータに素性を適用する範囲を決定する．これは，独話の中には長い文が含まれることがあり，そのような文には意味関係の推定には無関係な単語列も含まれていることが多いため，意味関係の推定に有効な単語列に限定することで，有効ではない素性が使用されることを防ぐ．具体的には，以下に示すように係り受け解析の結果を用いることで，接続助詞「が」を含む文節より

[3]http://svmlight.joachims.org
[4]http://sidb.el.itc.nagoya-u.ac.jp/

表 1: 負例で使用した接続詞

また，そして，それに，しかも，それから，そのうえ，それと，おまけに，ちなみに
なお，なぜなら，ただし，ただ，と同時に，あわせて，併せて，ましてや，それどころか
次いで，まずは，次に，そのうえで，だとすれば，とすれば，そういえば，それだけに
だって，というのも，実は，本当は，もっとも，ともすれば，そもそも
では，それでは，ところで，さて，それにしても，ともあれ，そしたら，ならば，それなら

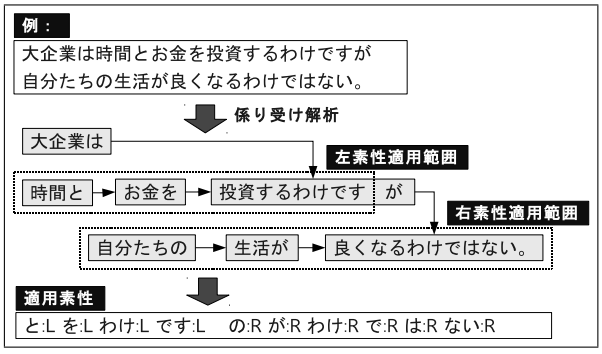

図 4: 素性適用範囲の決定

左側の素性適用範囲（以降，左素性適用範囲と定義）と，右側の素性適用範囲（以降，右素性適用範囲と定義）を決定する（図 4 を参照）．

Step 1: 接続助詞「が」を含む文節を検索する．（以降，接続助詞「が」を含む文節を P_L とする．）

Step 2: P_L の直前に出現し P_L に係っている文節を P_L の直前に追加し，それを P_L と再定義する．その後，係り元がなくなるまで，P_L に係っている文節を P_L の直前に連接していく．

Step 3: P_L が係っている文節を P_R と定義し，Step 2 の処理と同様に，係り元がなくなるまで，P_R に係っている文節を P_R の直前に連接する．

Step 4: P_L を左素性適用範囲，P_R を右素性適用範囲とする．

そして，図 4 のように，左素性適用範囲に対しては「L」のフラグが付与された素性を適用し，右素性適用範囲に対しては「R」のフラグが付与された素性を適用する．

4 評価

本手法を実装して評価実験を行った．実装にあたり，形態素解析器として ChaSen[5]，係り受け解析器として CaboCha[3][6]を使用した．評価用の正解データとして，名古屋大学同時通訳データベースの独話データベースから，接続助詞「が」を含む 416 事例を抽出し，人手にて「逆接」と「逆接以外」のラベルを付与した．その結果，416 事例のなかで 157 事例が「逆接」であり 259 事例が「逆接以外」であった．この正解データを使用して正解率 *Accuracy*，精度 $P(x)$，再現率 $R(x)$ を計算した．それぞれの定義を以下に示す．

$$Accuracy = \frac{c_{num}}{a_{num}},\ P(x) = \frac{|C(x)|}{|S(x)|},\ R(x) = \frac{|C(x)|}{|A(x)|}$$

ただし，

c_{num}**:** 本手法によって正しく判定された意味関係の事例数．

a_{num}**:** 正解データの事例数．

$C(x)$**:** $x = but$：本手法によって「逆接」と判定された事例のうち正解であった事例の集合．$x = other$：本手法によって「逆接以外」と判定された事例のうち正解であった事例の集合．

$S(x)$**:** $x = but$：本手法によって「逆接」と判定された事例の集合．$x = other$：本手法によって「逆接以外」と判定された事例の集合．

$A(x)$**:** $x = but$：正解データにおいて「逆接」と判定された事例の集合．$x = other$：正解データにおいて「逆接以外」と判定された事例の集合．

評価結果を表 3 に示し，それぞれの項目について以下に説明する．

位置情報なし: 本手法における素性（機能語）に位置情報 (L,R) を付与しない．

内容語: 素性として内容語（名詞，動詞，形容詞）を使用（位置情報は付与）．

全ての品詞: 素性として全ての品詞を使用（位置情報は付与）．

素性適用範囲なし: 素性の適用範囲の決定の処理を行わず，接続助詞「が」の前後の文字列を全て素性の適用範囲として使用．なお，素性としては機能語を使用．

ベースライン: 全て「逆接以外」と判定

BACT: 機械学習手法として BACT[2] を使用した場合．なお，入力は全ての単語を要素とするリスト構造とした．

5 考察

表 3 より，本手法の正解率は 75.0% であり，内容語を素性とした場合（正解率 54.3%）より良い結果を得た．この結果より，接続助詞「が」を含む翻訳単位間の意味関

[5]http://chasen-legacy.sourceforge.jp/
[6]http://chasen.org/~taku/software/cabocha/

表 3: 評価結果

	Accuracy(%)	P(but)(%)	R(but)(%)	P(other)(%)	R(other)(%)
本手法	75.0	68.5	62.4	78.4	82.6
位置情報なし	66.8	56.9	49.7	71.7	77.2
内容語	54.3	44.0	77.1	74.5	40.5
全ての品詞	69.5	58.3	66.9	78.0	71.0
素性適用範囲なし	72.8	67.2	54.8	75.3	83.8
ベースライン	62.3	0.0	0.0	62.3	100.0
BACT	67.3	57.2	52.9	72.7	76.1

係の推定には，素性として位置情報を付与した機能語を使用することが有効であると考える．SVM^{light} の model ファイルを解析し，テストデータの中でどのような素性が「逆接」の判定に有効であったかを調べた．その結果，「た:L」「ない:R」「まし:L」「まし:R」「ませ:R」といった素性が「逆接」の判定に大きく寄与していた．例えば「それはひとつの政府を作り上げるということはひとりの男の子の夢でござい ました **が**冷戦が始まり まし てそれは達成され ませ んでした。」といった事例では上記の素性がいくつか出現し，正しく「逆接」と判定された．しかし,「私先日い た んです**が**そこにいらしてた方々というのは八つの違う国からいらしてい まし た。」といった事例においても上記の素性がいくつか出現してしまい,「逆接以外」と判定すべきところを「逆接」と判定された．

以上のように，機能語と位置情報 (L,R) を素性とした場合でも比較的良好な結果を得ることはできるが，反例もまた存在する．よって，より高い正解率を達成するためには，本手法とは別のアプローチ（例えば，接続助詞「が」の前後の翻訳単位に極性 (positive，negative) を付与し，極性が反転している場合を「逆接」と判定）も必要であると考える．しかしながら，全て「逆接以外」と判定するベースライン手法に比べて正解率が 12.7% 向上しており，本手法を同時翻訳に適用する意義があると考える．

6 まとめ

本研究では接続助詞「が」に着目してそのあいまい性を解消し，翻訳単位間の意味関係の推定を行った．具体的には，入力として接続助詞「が」を含む文を入力し, その意味関係が「逆接」であるか「逆接以外」であるかを推定した．推定には機械学習手法を用いるが，機械学習のための十分な量の学習データを人手で作成するには膨大なコストが必要となる．そこで，本研究では，新聞記事コーパスから大量の学習データを自動的に作成し, それを使用して機械学習を行った．評価を行った結果，正解率 75.0% を達成し，全て「逆接以外」と判定する場合より良好な結果を得た．

謝辞

本研究は科学研究費補助金 基盤研究 B(20300058) の助成を受けたものである。

参考文献

[1] Kashioka, H. and Maruyama, T.: Segmentation of Semantic Unit in Japanese Monologue, *Proceedings of Oriental COCOSDA 2004*, pp. 87–92 (2004).

[2] Kudo, T. and Matsumoto, Y.: A Boosting Algorithm for Classification of Semi-Structured Text, *EMNLP* (2004).

[3] 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002).

[4] Matsubara, S., Takagi, A., Kawaguchi, N. and Inagaki, Y.: Bilingual Spoken Language Corpus for Simultaneous Machine Interpretation Research, *Proceedings of 3rd International Language Resources and Evaluation Conference(LREC-2002)*, pp. 153–159 (2002).

[5] 松原茂樹: 同時通訳の工学と科学 －次世代自動通訳技術の実現に向けて－, 情報処理, Vol. 49, No. 6, pp. 19–25 (2008).

[6] Nakov, P. and Hearst, M.: Using the Web as an Implicit Training Set: Application to Structural Ambiguity Resolution, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 835–842 (2005).

[7] Ryu, K., Matsubara, S., Kawaguchi, N. and Inagaki, Y.: Bilingual Speech Dialogue Corpus for Simultaneous Machine Interpretation Research, *Proceedings of Oriental COCOSDA-2003*, pp. 217–224 (2003).

[8] 笠浩一朗, 松原茂樹, 稲垣康善: 対訳コーパスを用いた同時翻訳単位の検出, 言語処理学会第 15 回年次大会論文集, pp. 881–884 (2009).

[9] 田中省作, 冨浦洋一, 日高達: 係り受け情報を用いた名詞句「NP_1 の NP_2」の意味関係の候補の抽出, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, pp. 77–84 (2001).

[10] Vapnik, V.: *Statistical Learning Theory*, Wiley (1999).

[11] 山本和英, 齋藤真実: 用例利用型による文間接続関係の同定, 自然言語処理, Vol. 15, No. 2, pp. 21–51 (2008).