様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月　５日現在

機関番号：13901
研究種目：若手研究(B)
研究期間：2010～2011
課題番号：22700143
研究課題名（和文）
膠着語の統計的語形変化処理
研究課題名（英文）
Statistical inflectional processing of agglutinative languages
研究代表者
小川　泰弘（OGAWA YASUHIRO）
名古屋大学・情報科学研究科・助教
研究者番号：70332707

研究成果の概要（和文）：日本語やウイグル語，ウズベク語などの膠着語においては，名詞や動詞に接尾辞が接続する際に様々な語形変化を起こす．従来においては，そうした語形変化を人手によるルールを用いて処理していた．それに対して本研究では，語形変化を一種の翻字として捉えることにより，統計的なアプローチを導入した．具体的には統計的機械翻訳システムを応用し，語形変化ルールの自動生成を試みた．これにより，従来の統計的機械翻訳とは異なる知見を得るとともに，統計的語形変化処理の有効性を示した．

研究成果の概要（英文）：Japanese, Uighur and Uzbek languages are agglutinative and they have various types of inflection. Traditional ways to deal with the inflection are rule-based methods. In contrast to this, we considered inflection is a kind of transliteration and introduced a statistical approach. Thus we tried automatic generation of inflection rules based on statistical machine translation techniques. From this, we obtained the results showing the effectiveness of statistical inflectional processing.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,400,000	420,000	1,820,000
2011 年度	1,300,000	390,000	1,690,000
総　計	2,700,000	810,000	3,510,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：膠着語，ウイグル語，ウズベク語，形態素解析，音韻変化

１．研究開始当初の背景

これまでに研究代表者は，日本語形態素解析システムの構築，日本語−ウイグル語機械翻訳システム，日本語−ウズベク語機械翻訳システムの開発に取り組んできた．

(1) 膠着語の特徴

日本語・ウイグル語・ウズベク語は，言語学的には膠着語に分類される．膠着語においては，助詞や助動詞を含む接辞が重要な役割を果たす．例えば，語幹に接続して新たな語を派生したり，語の文中における文法的機能を示したりする．そのため，膠着語に対する自然言語処理においては，接辞の取り扱いが重要なテーマとなる．特に，接辞が語幹に接続する際に語形変化が起きる場合は，その処理には工夫が必要である．例えば，日本語においては「書く」に過去を示す接尾辞「た」が接続する場合は「書いた」という語形変化が起きる．

(2) 従来の研究

こうした分野は形態論もしくは形態構文論と呼ばれ，日本語における動詞の活用処理もその一種と言える．日本語形態素解析の初期においては，活用処理が大きな問題であり，久光らの「日本語形態素解析における効率的な動詞活用処理」(1994)などに上げられる各種の手法が提案されてきた．日本語以外の言語も扱える汎用的な手法としては 2-レベル規則が提案されている．これは，実際に出現する語の形である表層形と，それに対応する形態素の列である基底形との間での相互の変換ルールを記述できる仕組みであり，あらゆる言語の語形変化処理が可能であると主張されている．

しかし，そうした従来の語形変化処理は，いずれもルールに基づく手法であり，各音素がどのような条件でどのように変化するかのルールを，その言語のエキスパートが記述する必要があった．実際に，研究代表者は，日本語−ウイグル語機械翻訳や日本語−ウズベク語機械翻訳の研究において日本語・ウイグル語・ウズベク語各言語における語形変化について調査し，そのルールを記述してきたが，言語の特徴を理解するために，多大な労力が必要であった．しかも，こうした語形変化は言語ごとに異なるため，他の膠着語を扱うためには，また新たにルールを作成し直す必要がある．

(3) 統計的手法

一方，現在の自然言語処理においては，統計的手法が各分野で取り入れられている．そこで本研究においても，統計的手法が適用可能ではないかと考えた．以上のような背景から，ルールを言語ごとに作成するのではなく，統計的翻字を用いた語形変化処理を実現する本研究の着想に至った．

２．研究の目的

本研究の最終的な目的は，ある膠着語に対して，適切な事例データを入力することにより，その言語用の語形変化処理モジュールを作成するメタシステムの開発である．語形変化処理には，入力された語を語幹と接辞に分割する解析処理と，語幹と接辞の列から出力文を作成する生成処理の二つがある．従来は，いずれも人手で記述したルールを基に処理されてきたが，本研究では，統計的翻字手法を応用し，自動的に語形変化処理モジュールを作成することを目指した．

本研究においては，その中でも日本語・ウイグル語・ウズベク語の三つの言語に関して，メタシステムを通じて解析・生成モジュールを作成し，性能評価を通じて本研究の有用性を具体的に明らかにした．具体的には，以下の二つの目的を達成した．

(1) 統計的手法適用による精度の検証：

上述のように，これまでの語形変化処理は主にルールに基づく手法が適用され，統計的手法は適用されてこなかった．その原因の一つは，ルールに基づく手法で高精度の結果が得られる一方で，統計的手法では，それに匹敵する精度が得られないと予想されていたからである．

本研究では，実際に統計的手法を適用することにより，どの程度の精度が得られるかを確認するとともに，統計的機械翻訳と統計的語形変化処理において，どのような違いがあるかを比較することが目的の一つである．

(2) 言語学的側面からの検証：

上記の(1)の意義は自然言語処理の立場からものであるが，本研究では，基底形が何かという言語学的な問題に答えることも目的の一つである．「研究開始当初の背景」の項において，「書く」に「た」が接続して「書いた」となる例を上げた．この例では，過去を示す接尾辞の基底形が「た」であるとしているが，何が基底形であるかは諸説あり，「た」の他にも「tta」や「ita」が基底形だと主張する説もある．そうした説においては，例えば，基底形「た」が条件により変化すると説明されるが，なぜ「た」が基底形なのかの根拠を示していない．

ルールに基づく手法では，基底形を変えるたびにルールも変更する必要があり，どの基底形が適切であるかの検証は容易ではない．しかし，統計的手法に基づく本研究では，基底形を変えた比較実験により，基底形ごとの精度を比較することが可能となる．それにより，どの基底形が適切であるかについて，自然言語処理の観点からの指標を示す．

３．研究の方法

本研究の基本的なアプローチは，統計的機械翻訳の一種である統計的翻字である．従来の研究では，表層形から基底形，基底形から表層形への変換のルールを人手により記述してきたが，本研究では，こうした変換を統計的翻字として捉え，統計的機械翻訳用のモジュールである GIZA++, SRILM, Moses などの既存のシステムを使用することにより，効率的に統計的翻字を実現した．具体的には以下の手法で研究を実施した．

なお，日本語の語形変化処理に関しては，文字単位で処理するよりも音韻論的アプローチを用いた方が性能が良いため，本研究では音韻論的な観点から日本語の語形変化を記述した派生文法を基本的な文法として採用した．

(1) 学習用データの構築：

従来の統計的機械翻訳や統計的翻字では

対象となる双方の言語データが明確であったが，本研究においては，その一方が明確でない．すなわち，何が基底形かが問題となる．よって，本研究においては，まず基底形を定め，それに基づくデータを作成した．具体的には，各言語のコーパスを形態素解析し，必要に応じて形態素を変形させることにより，基底形の列からなる学習データを作成した．しかし，それだけでは語幹と接辞の組合せにおいて出現しないものがあるため，語幹と接辞の間で可能な組合せを網羅的に作成することも試みた．その際に，ウイグル語とウズベク語に関しては母語話者にデータの作成を依頼した．

(2) 統計的語形変化処理に関する実験：
上記の(1)で構築した学習用データを用いて，日本語・ウイグル語・ウズベク語それぞれにおいて統計的語形変化処理に関する実験を行った．その際には以下の三つの比較実験を行った．

① 翻訳モデルの比較：
翻訳モデルとは，原言語文(翻訳する文)と目的言語文(翻訳された文)のペア，すなわち対訳文から学習されるモデルであり，例えば「猫」が「cat」に翻訳されるなどの情報は，この翻訳モデルに含まれている．今回の実験においては，基底形が原言語文に，表層形が目的言語文に相当する．
統計的機械翻訳においては，コーパスからランダムに抽出した対訳文を訓練データとして用いることが多いが，語形変化処理の場合もそれで良いのか，また学習データの分量はどの程度必要か，という観点から比較実験を行った．

② 言語モデルの比較：
言語モデルとは，目的言語文の流暢さを表すものであり，言語モデルが良いと機械翻訳の出力文が自然な文となる．一般に統計的機械翻訳においては，言語モデルの学習に使用するデータが多ければ多いほど性能が向上するとされているが，統計的語形変化処理の場合にも同じ結果が言えるのか，また，データの作成にはどのような方法が良いのか，という観点から比較実験を行った．

③ 基底形の比較：
上記の①と②は，統計的手法適用による精度の検証を目的とした実験であるが，③はどの基底形が適切であるかを検証する，言語学的観点からの実験である．各種の基底形を用意することにより，統計的処理の立場からはどの基底形が適切であるか，比較実験を行った．

４．研究成果
(1) 研究全体を通じた成果：
本研究では，まず，日本語の語形変化処理に関して，その基底形から表層形への変換ルールを統計的機械翻訳用の各モジュールを使用して作成した．その際，従来の統計的機械翻訳では得られなかった以下の知見を得た．

① 翻訳モデルの学習データとしては，ランダムに抽出されたデータではなく，語幹と接辞の組合せを網羅したものを用いるべきである．
② 言語モデルの学習データに関して，一般の統計的機械翻訳においてはデータ量を増やせば増やすほど精度が向上するが，統計的語形変化処理においては，比較的少ないデータで精度の上限に達する．
③ 基底形の選択は精度に大きな影響を与える．特に日本語の場合は，文字単位ではなく音素単位で処理した方が精度が高い．また，語形変化のタイプごとに異なる記号を用いるような基底形を選択することにより精度が向上する．

同様の実験を，ウイグル語やウズベク語に対しても行った．ウイグル語やウズベク語も日本語と同様に名詞や動詞に接尾辞が接続するが，語形変化に関しては日本語よりも複雑である．特にウイグル語においては，母音調和という距離が離れた音素の影響を受ける音韻変化があるため，より多くの学習データが必要になることを明らかにした．
ウイグル語やウズベク語に関しては，言語自体があまり知られていないため，以下では日本語の実験に関する研究成果について詳細を示す．

(2) 日本語の統計的語形変化処理に関する実験から得られた成果：
「研究の方法」で示したように，統計的語形変化処理に関する実験では，①翻訳モデルの比較と②言語モデルの比較の二つの実験を行った．
実験において学習されたモジュールの評価には，学習データと同じものをテストデータにしたクローズドテストと，学習データと異なるテストデータを用いたオープンテストを行った．またクローズドテストでは，規則的な変化だけに限定した場合(closed regular)と，すべてのデータを使用した場合(closed all)の２種類のテストを行った．

① 翻訳モデルの比較：
従来の統計的機械翻訳と同様に，ランダムに抽出した学習データ(1. および2.)を用意した．しかし，語形変化処理においては，出

現する音素の分布の偏りが大きいため，ランダム抽出では，学習できない音素の組合せが存在する．そこで，動詞と接尾辞の組合せから網羅的に生成したデータ(3．および 4．)も用意した．よって最終的には，以下の 4 種類の訓練データを用意した．

1．EDR 千文：
EDR コーパスからランダムに 1,000 文を抽出し，そこに出現した動詞句を使用したもの．

2．EDR 全文：
EDR コーパス中の全 208,156 文を抽出し，そこに出現した動詞句を使用したもの．

3．組合
動詞の語幹末尾ごとに EDR コーパス中の出現頻度が高いもの 5 個と，接尾辞 32 個の組合せ．

4．組合(10)
動詞の語幹末尾ごとに EDR コーパス中の出現頻度が高いもの 10 個と，接尾辞 32 個の組合せ．

以上の訓練データを用いて学習したところ，表 1 のような結果が得られた．

表 1：訓練データの比較

	訓練データ	size	closed all	closed regular	open
1	EDR 千文	2,580	89.8%	78.6%	94.3%
2	EDR 全文	531,685	90.6%	98.2%	95.4%
3	組合	1,920	92.6%	99.6%	90.9%
4	組合(10)	3,520	92.5%	99.5%	91.5%

この結果から，ランダムに抽出したデータを使用するという，従来の統計的機械翻訳と同様の方法では，規則的な変化のルールであっても学習できない場合があることが判明した．これにより，統計的語形変化においては，ランダムに抽出したデータではなく，語幹と接辞の組合せデータを網羅したものを使用するのが良いことが判明した．

② 言語モデルの比較
言語モデルの学習には，EDR コーパスから作成したデータ(1．～6．)と，翻訳モデルの実験で使用したデータ 3．(7．)を用意し，以下の 7 個のモデルを比較した．

1．EDR 全文節：
EDR コーパス中の全 208,156 文に出現した文節(動詞句以外も含む)．

2.EDR 全文：
EDR コーパス中の全 208,156 文に出現した動詞句．

3.同重複：
2.と同様であるが，重複する動詞句を省かないもの．

4.同 3-gram：
2.と同様であるが，2-gram ではなく 3-gram モデルを使用したもの．

5.EDR 千文 a：
EDR コーパス中の 1,000 文に出現した動詞句．

6.EDR 千文 b：
5.とは別の 1,000 文に出現した動詞句．

7.組合せ：
訓練データ 3.と同じもの．

以上のデータから学習した言語モデルを用いて学習したところ，表 2 のような結果が得られた．

表 2：言語モデル用データの比較

	言語モデル	size	closed all	closed regular	open
1	EDR 全文節	281,263	91.8%	98.5%	88.9%
2	EDR 全文	22,614	92.6%	99.6%	90.9%
3	同重複	531,685	92.2%	98.9%	93.9%
4	同 3gram	22,614	89.2%	92.8%	92.1%
5	EDR 千文 a	1,022	89.4%	99.2%	92.2%
6	EDR 千文 b	1,063	92.9%	99.3%	94.2%
7	組合せ	1,899	92.7%	98.7%	86.5%

一般の統計的機械翻訳においてはデータ量を増やせば増やすほど精度が向上することが知られている．しかし，表 2 の結果からは，一番精度が高かったのが 6．の EDR1,000 文を用いた場合で，それ以上データを増やしても効果がなかった．これにより，統計的語形変化処理においては，この程度のデータ量で上限に達することが分かった．

これは，データを増やしても性能の向上が期待できないという欠点でもあるが，ウイグル語やウズベク語などの言語資源が手に入りにくい言語であっても，ある程度の量があれば充分な精度を達成できるという利点でもある．

(2) 適切な基底形の検証：

③ 基底形の比較：
上記の①と②は，統計的手法適用による精度の検証を目的とした実験であるが，以下は言語学的観点から，どの基底形が適切であるかを検証する実験である．

まず，日本語を文字単位で処理する平仮名表記と，音素単位で処理するローマ字表記を比較し，ローマ字表記の方が高精度であるという結果が得られた．

また，語幹と接尾辞の接続の際の変化においては，以下の 2 種類の考え方がある．

1．語幹と接尾辞が接続する場合，状況に応

じて，接尾辞の先頭の音素が欠落する．
2．語幹と接尾辞が接続する場合，状況に応じて，接尾辞の先頭に音素が挿入される．

これらについて比較実験を行ったところ，1.の音素が欠落すると考えるモデルの方が高精度であった．
その他，不規則変化する場合に関しては，以下に示すような様々な基底形を比較した．

・「来る」： **ko**，ku，k，ki
・「する」： **se**，su，s，si，sa
・「なさる」などの変則動詞： **nasar**，nasa
・「思う」などの末尾がｗ： **omow**，omo
・完了の「-(i)ta」など： **Ita**，ita，Tta，tta，Ta，ta，Ida，Da，da，Nda，n’da
・命令の「-e/-ro」： e，**ro**，i，E

すべての基底形の組合せの結果を示すことは出来ないため，本研究が参考にした派生文法が提案する基底形(太字)と，一番精度が高かった提案基底形(下線)の比較のみを表３に示す．

表３：異なる基底形の比較

基底形	closed all	closed regular	open
派生文法	84.6%	99.0%	72.7%
提案基底形	92.6%	99.6%	90.9%

この結果，基底形の選択により精度に大きな差が出ることが明らかになった．また基底形の選択基準としては，例外的な変化をする場合には，その変化に固有の文字を使用するべきであるという知見が得られた．
同様な知見を，ウイグル語やウズベク語に関しても獲得した．

3) 今後の展望：
以上の成果は，これまでの研究では得られなかったものである．特に膠着語の研究に関しては，日本語・韓国語・トルコ語を除くと言語資源も少なく，実際に実現されているシステムも少数である．それに対して，ウイグル語やウズベク語などの言語が，比較的少量のデータで語形変化処理できることを示した点に本研究のインパクトがある．
今後の展望としては，作成した統計的語形変化処理システムを統計的機械翻訳へ組み込むことを進めている．統計的機械翻訳には，大量の対訳コーパスが必要であり，その構築も現在は進めており，最終的には膠着語間の汎用的な機械翻訳システムの実現を目指す

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計１件）
(1)小川泰弘，他： 統計的機械翻訳システムを利用した膠着語の音韻変化処理，言語処理学会第 18 回年次大会講演論文集，
広島市立大学（2012.3.16）

６．研究組織
(1)研究代表者
小川 泰弘（OGAWA YASUHIRO）
名古屋大学・大学院情報科学研究科・助教
研究者番号：70332707

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

：