依存構造解析に基づく中国語離合詞処理

出羽達也
(株)東芝 研究開発センター
tatsuya.izuha@toshiba.co.jp

１．　はじめに

中国語には二つ以上の語素(形態素)から構成される語が多いが，その中には，語素間に他の成分を挿入することのできる「離合詞」と呼ばれる語がある．例えば，「卒業する」という意味の中国語動詞「毕业」を用いて「卒業できない」ことを表現するときには，不可能を表す「不了」を間に挿入し，「毕不了业」となる．間に他の成分を挿入することができるという点だけに着目すれば句であるという解釈も考えられるが，離合詞を構成する語素の中には他に独立の語としての振る舞いが見られないものがあること，離合詞全体の持つ意味が各語素の意味の単純な合成とは異なることから，一般の句とは異なり，語に準じた取扱いが必要となる．このように離合詞は他の言語にあまり見られない中国語独特の現象であるため，特別な処理が必要となる．離合詞の数は 2,000 種類前後もあると言われており，中国語の高精度な解析を実現するには避けて通ることはできない．

２．　依存構造解析に基づく離合詞処理

挿入成分がある離合詞を認識するには最低限以下の 3 つの処理が必要であると思われる．

(0) 離合詞の候補となる２つの形態素の組のリストを用意しておく．
(1) リストを参照して，離合詞の候補を文中から検出する．
(2) 検出した２つの形態素の組が実際に離合詞であるかどうかを判定する．

(0)は手法の如何にかかわらず必須であると思われるが，(1)(2)にはさまざまなアプローチが考えられる．例えば，自然言語処理の観点から離合詞を取り上げた先駆的な研究である[Fan94]では以下のようなアプローチを採っている．

(1) 前語素を検出したら，文の後方に向かって後語素をサーチする．
(2) 離合詞に挿入可能な形態素列パターンを予め用意しておき，検出した前語素と後語素の間の成分と照合する．

中日翻訳に適用した小規模評価では 93%の文が正しく翻訳されたと報告されている通り，形態素列のパターンマッチに基づく手法で多くのケースをカバーできる．しかし，このような手法には，複雑な挿入成分のパターンを予め網羅するのが難しいという限界がある．例えば，「吃了一个大亏(一つ大損をした)」という例では，離合詞「吃亏(損をする)」に「了一个大」が挿入されているが，このような挿入成分にマッチするパターンは予め用意することができなかったと報告している．

そこで本稿では，形態素列のパターンとして予め網羅するのが難しい複雑な挿入成分に対処するために，以下のような方法を提案する．

(S1) 挿入成分を持つ離合詞を含んだ文を形態素解析する．
(S2) 離合詞の内部構造(語構成)に対応した品詞を前語素と後語素に与え，依存構造解析する．解析には，離合詞のための特別な文法は使用しない．
(S3) 離合詞の内部構造に対応した依存関係(動詞－補語，動詞－目的語)にある形態素の組を取り出し，離合詞候補リストと照合する．
(S4) 離合詞候補リスト中にマッチするものがあれば，取り出した形態素の組を一つのノードに統合する．統合前の形態素の依存先ノードは，依存元を統合ノードに変更する．その際，連体修飾性の依存関係は，依存アークのレベルや依存先ノードの品詞を変更するなどして，動詞を依存先とする適当な依存関係に変更する．

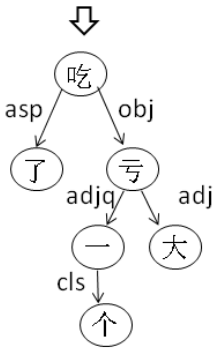
先ほどの「吃了一个大亏」というフレーズを例にとり，図１を参照しながら上述の処理を説明する．まず，形態素解析結果は (1)のようになる．離合詞「吃亏」は次節で述べるように動詞－目的語関係の内部構造であるから，「吃」には動詞”VV”，「亏」には名詞”NN”という品詞が与えられている．続いて依存構造解析を行う．本稿では，統語構造を明確に示すため，いったん句構造解析を行ってから，それを依存構造に変換する．その過程を (2)に示す．なお本稿では，品詞タグおよび句構造タグは Penn Chinese Treebank の定義に準拠している([Xue00])．　(2)で得られた依存構造に対して，　(3)以降で離合詞処理を行う．動詞－補語(アークラベル”vrd”に対応)または動詞－目的語(”obj”)の依存関係にある形態素の組は「吃」と「亏」である．「吃亏」は離合詞であるから，離合詞リストとの照合でマッチすることが期待できる．そこで(4)に示すよう

に，2つのノードを統合した新ノード「吃亏」を作る．続いて，「吃」と「亏」を依存元としていたノード「了」「一」「大」の依存元を統合ノードに変更する．このとき，数量句による連体修飾を表すアークラベル”adjq”と，形容詞句による連体修飾を表すアークラベル”adj”はそれぞれ”advq”, ”adv”といった連用修飾性のものに変更する．

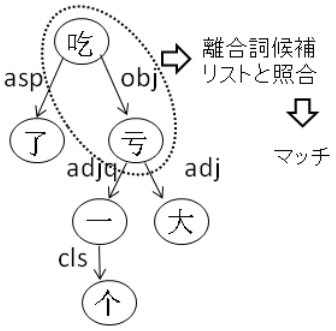
(1) 吃/VV 了/AS 一/CD 个/M 大/JJ 亏/NN

(2)

```
(VP (VV 吃)
    (AS 了)
    (NP (QP (CD 一)
            (CLP (M 个))
        (ADJP (JJ 大))
        (NP (NN 亏))))
```



(3)



(4)

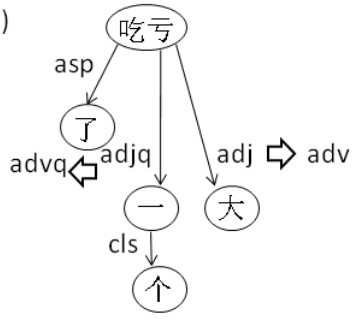


図１　離合詞処理の流れ

このような処理のメリットとして，以下の2点が考えられる．

- 特定の依存関係にある形態素の組だけ調べればよいので，離合詞候補の探索コストが小さい．
- 数多くのパターンを用意しなくても，様々な挿入成分を網羅的にカバーできる・

一方，デメリットとしては以下の2つが考えられる．

- 依存構造解析を行う計算コストが大きい．
- 依存構造解析の精度は必ずしも高くなく，解析を誤ると離合詞を検出できなくなる．

実際には，挿入成分が比較的単純な離合詞を，従来のパターンベースの手法でロバストに認識しておき，そこで漏れた離合詞を依存構造解析ベースの手法で救うというような利用方法が現実的であろう．

依存構造解析ベースの離合詞認識手法(S1)～(S4)が有効であるためには，以下の前提が成り立っている必要がある．

(a) 離合詞が挿入成分を持つとき，前語素と後語素は，離合詞の内部構造に対応した品詞を持つ語と同じ統語的振る舞いをする．

(b) 離合詞の内部構造に対応した品詞を前語素と後語素に与えて依存構造解析を行ったとき，前語素と後語素は直接依存関係にあり，かつその関係は内部構造と同じである．

そこで次節以降では，上記のことが成り立っているかどうかを検証していく．

３．　離合詞の内部構造

[Lu90]によると，離合詞はそれを構成する語素間の関係に基づいて4種類に分類することができる．

(1) 動詞－目的語関係（[Lu90]では「動賓関係」）
(2) 動詞－補語関係（[Lu90]では「動補関係」）
(3) 並列関係
(4) 主述関係

このうち本稿では，語彙が豊富で使用頻度も大きい(1)(2)のみを取り扱うことにする．

４．　離合詞の前語素･後語素と挿入成分の統語的関係

[Lu90]では，離合詞の前方成分と後方成分の間に挿入可能な成分として 11 種類の語句を挙げている．本節では，それらについて第２節で挙げた２つの前提(a)(b)が成り立つことを確認していく．なお以下の議論において，離合詞の後方成分として，あるいは挿入成分としてしばしば補語が登場する．中国語の動詞と補語の統語的関係を論じる際には，どちらを主辞と捉えるか議論の分かれるところであるが，本稿においては，動詞が主辞であるという前提で議論を進める．

3-1. 動詞－補語型離合詞

動詞－補語型の離合詞については，挿入可能な成分は可能不可能を表す ”得 de” ”不 bu” に限られる．そしてこれは一般の動詞と補語の間に ”得 de” ”不 bu” が挿入され可能不可能を表す現象(結果補語・方向補語の可能形)と一致していることが[Lu90]でも指摘されている．Penn Chinese Treebank では，結果補語・方向補語の可能形を”VPT” (potential form V-de-R or V-bu-R) というタグで表現しているが，図２に示すように，一般の動詞－補語も離合詞も同じ扱いをしている．図２の句構造を依存構造に変換すると図３のようになる.なお,句構造から依存構造に変換する際には，CFG 規則 VPT→VV+AD+VA(または VV) の右辺第１項を主辞とし，主辞と右辺第２項，第３項との間の関係ラベルをそれぞれ”pt”，”vrd”としている．

以上のことから,動詞－補語型の離合詞については，(a)(b)が確認できた.

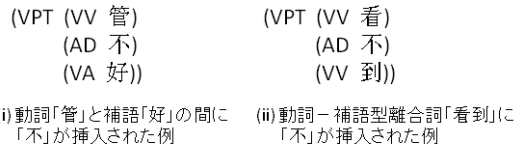


図２ 不可能を表す「不」の句構造表現

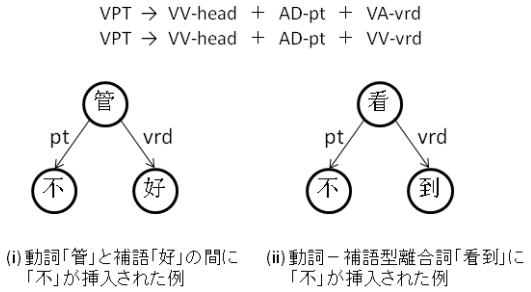


図３ 不可能を表す「不」の依存構造表現

3-2. 動詞－目的語型離合詞

離合詞の前語素を動詞,後語素を名詞とみなしたとき，[Lu90]が挙げた 11 種類の挿入成分すべてに対して，前語素か後語素のいずれかを依存元とする依存構造解析が可能である．このとき，前語素と後語素の間の動詞－目的語関係は依存構造上でも保存されている．

3-2-1. 前語素を依存元とする挿入成分

[Lu90]が挙げた 11 種類の挿入成分のうち，以下のものは前語素(動詞)を依存元として解析できる．

・可能不可能を表わす成分
・結果補語
・方向補語
・動態助詞
・構造助詞「的」

スペースの都合上，上記の５つすべての検証結果を示すことはできないため，ここでは方向補語を取り上げて検証する．動詞－目的語型離合詞「插秧」(田植えをする)の間に方向補語「上」が挿入されたケースを図４に示す．左側の句構造解析結果において，前語素(VV 插)を一般の動詞で，後語素(NN 秧)を一般の名詞で置き換えた構造は数多く見られる一般的な構造である．そしてこれを変換して得られた右側の依存構造では，動詞－目的語の内部構造が保存されている(“obj”アーク)．以上のことから，動詞－目的語型離合詞に方向補語が挿入された場合にも(a)(b)が確認できた．図４右の依存構造において，”obj”アークで結ばれた「插」と「秧」を統合すれば離合詞処理の完了である．

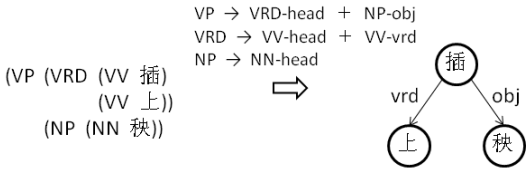


図４ 方向補語(単純型)の挿入

ただし，同じ方向補語でも,「起来」のような複合型の方向補語は特別な取り扱いが必要となる．このような複合型の方向補語が挿入されるときは，方向補語の前語素と後語素の間に離合詞の後語素が入り込む．このような場合のアノテーション仕様については[Xue00]で言及されていないため,仮に図５のような仕様を決めて解析すると，(a)(b)は成立しない.「上|下|进|出|回|过|起|开」＋「去|来」という構成の複合方向補語が挿入されるケース対しては，例外的に図６のような特殊処理が必要となる．なお図５，図６は動詞－目的語型離合詞「打仗」(戦争をする)に複合型方向補語「起来」が挿入されたケースである．

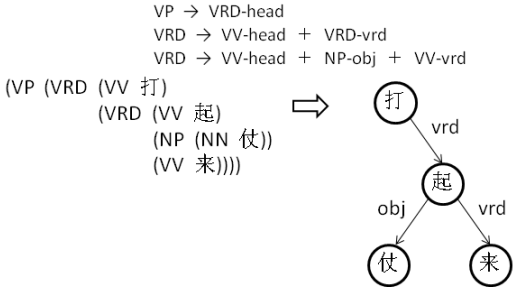


図５ 方向補語(複合型) の挿入

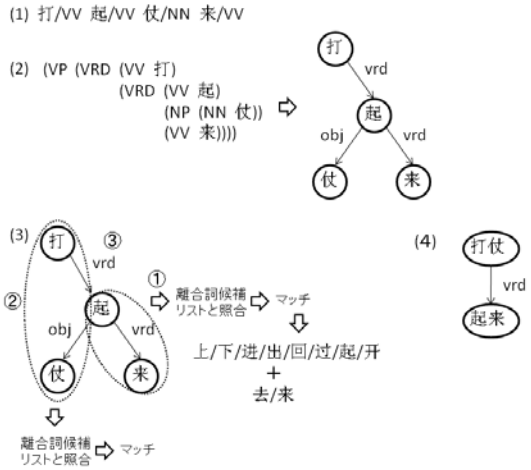


図６　複合型方向補語の挿入に対応した
離合詞処理の流れ

3-2-2. 後語素を依存元とする挿入成分

以下の挿入成分は後方成分を依存元として解析できる.

・人称代名詞「我」/「你」/「他」＋「的」
・「谁」/「哪个」＋「的」
・「这」/「那」＋「个」
・「什么」
・数量表現

上記の４種類の成分のうち，人称代名詞＋「的」を取り上げて検証する．図７は，動詞－目的語型離合詞「生气」(腹が立つ)に人称代名詞「我」＋「的」が挿入されたケースである．左側の句構造解析結果において，前語素(VV　生)を一般の動詞で,後語素(NN 气)を一般の名詞で置き換えた構造は数多く見られる一般的な構造である．そしてこれを変換して得られた右側の依存構造では，動詞－目的語の内部構造が保存されている(“obj”アーク)．以上のことから，動詞－目的語型離合詞に人称代名詞＋「的」が挿入された場合にも(a)(b)が確認できた．図５右の依存構造において，”obj”アークで結ばれた「生」と「气」を統合し，連体修飾性のアークラベル”adjde”を”obj”に変更，さらにノード「的」を除去すれば離合詞処理の完了である．

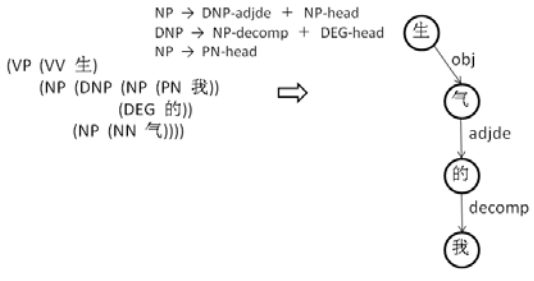


図７　人称代名詞＋「的」による後語素の連体修飾

3-2-3. 前語素と後語素を依存元とする挿入成分

以下の挿入成分は，前半が前方成分，後半が後方成分を依存元として解析できる.

・人称代名詞「我」/「你」/「他」＋「什么」

図８では，動詞－目的語型離合詞「送礼」(プレゼントする)に人称代名詞「他」＋「什么」が挿入され，「他」が前語素(動詞)の間接目的語に，「什么」が後語素(直接目的語)の連体修飾成分になることにより二重目的語構文を構成している．この場合も(a)(b)は成り立っており，”obj”アークで結ばれた「送」と「礼」を統合し，連体修飾性のアークラベル”adjdt”を”obj”に変更すれば離合詞処理の完了である．

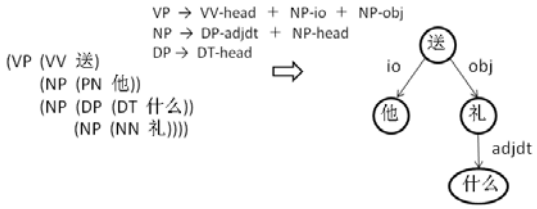


図８　人称代名詞＋「什么」の挿入による
二重目的語構文

[Lu90]の 11 種類の挿入成分はどれか一つだけが現れるというわけではない．図１の例のように，複数の種類の成分が挿入されることもあるが,提案手法では,複雑なパターンを用意することなくこれらを網羅的に処理することができる.

４．おわりに

挿入成分を持つ中国語離合詞の処理のために，依存構造解析に基づく方法を提案した．提案手法は，形態素列パターンに基づく従来の手法と比べて，複雑な挿入成分に対する網羅性が高いという利点をもつ.今後は，認識精度の評価を行い，本方式の有効性を検証する必要がある.

参考文献

[Lu90]　鹿宗世，李清華，大瀧幸子: 中国語離合詞500，東方書店(1990).

[Fan94]　范莉馨，任福継，宮永喜一，栃内香次: 中日機械翻訳における離合詞の処理手法，情報処理学会論文誌，Vol.35，No.9，pp.1702-1713(1994).

[Xue00]　Nianwen Xue，Fei Xia: The Bracketing Guidelines for the Penn Chinese Treebank (3.0)，http://www.cis.upenn.edu/~chinese/parseguide.3rd.ch.pdf (2000)．