既存の構文意味辞書を利用した他の目的言語用辞書の構築支援

片岡 明　　山田 節夫　　林 実　　横尾 昭男
日本電信電話株式会社 NTT サイバースペース研究所
{kataoka, syamada, mhayashi, ayokoo}@light.hil.ntt.co.jp

1 はじめに

近年, WWW の発達などにより, 英語圏のみならず, 全世界を相手とした情報交換が可能となっており, 様々な言語対に対する機械翻訳システムへの期待が高まっている.

機械翻訳システムでは, 原言語の構文構造を目的言語の構文構造に変換する際に, 構文意味辞書として結合価パタン対辞書が有効である [4]. 結合価パタン対 (以下, パタン対と呼ぶ) は, 原言語の用言と, 用言が取る格要素とを原言語パタンとして記述し, それに対応する目的言語パタンを対にして持たせることにより, 原言語用言の語義毎に目的言語表現を規定するものである.

日英パタン対辞書 [4] として, 大規模な辞書が人手で既に構築されているが, 実用的な規模のパタン対辞書の構築には膨大な費用と時間が必要である.

これまで, パタン対を対訳コーパスから自動的に学習する研究がなされている [1, 5, 6]. また, 白井ら [3] は, 人手で日英パタン対辞書を構築する際に, 対訳コーパスから自動作成したパタン対を提示する支援を行なっている. ところが, 対訳コーパスは一般に入手が困難であり, その作成にも膨大なコストがかかる.

そこで本研究では, 対訳コーパスと比較して容易に入手可能な目的言語の単言語コーパスと, 既に構築されている日英パタン対辞書とを利用してパタン対を自動作成し, そのパタン対を人に提示することで, 辞書構築を支援することを目的とする.

本稿では, パタン対を自動作成するために, 日英パタン対辞書の日本語パタンの格要素と目的言語文の単語とを, 意味カテゴリの比較によって対応付ける手法について述べる. 目的言語を中国語として評価実験を行ない, 手法の有効性を確認した.

2 パタン対辞書の構築支援

ある日本語用言について, 人が目的言語の単言語コーパスを参照してパタン対を作成する状況を想定する.

対象とするパタン対は, 日英パタン対を例に取ると, 以下のような構成となる.

例)
N1:<主体>が N2:<宿, 席>を 取る
⇒ N1 reserve N2

ここで, N1, N2 などは格ラベルで, それぞれの格要素の役割 (N1: 動作主, N2: 対象, など) を表す. また, <主体>, <宿>などは, 格要素に取り得る単語の制約条件を表す. これは, 意味体系 [2] 中の意味カテゴリで指定したものであり, 以下では条件カテゴリと呼ぶ.

本稿では, 人が行なうパタン対辞書の構築作業として, 以下の手順を想定する.

1. 日本語用言の目的言語訳語を対訳辞書より得て, その訳語を含む文を目的言語コーパスから得る
2. 得られた目的言語文からパタン対を作成すべきか判定する
3. 目的言語文の構造から, 目的言語パタンを作成する
4. 目的言語パタンに対応する日本語パタンを作成する
5. 日本語パタンの条件カテゴリを決定する
6. 各文から作成されたパタン対を比較し, 同一のパタン対とした方が良いものをまとめる

この手順のうち, 3.-5. について, 自動作成したパタン対を提示することで作業者の負担を減らし, 辞書構築の効率を上げることを目指す.

パタン対を自動作成するために, まず, 意味カテゴリを比較することで, 目的言語文の単語を日英パタン対辞書の日本語パタンの格要素に対応付ける. 次に, 格要素に対応付けられた単語の意味カテゴリを条件カテゴリとした日本語パタンを作成し, 各単語の文中での位置や機能語を元にした目的言語パタンを作成する.

本稿では, 目的言語文の単語と日本語パタンの格要素との対応付けについて述べる.

3 手法

3.1 格要素の対応

目的言語文の単語を格要素に対応付けたときに与える得点について説明する．日本語パタンの格要素の条件カテゴリと，目的言語文の単語の意味カテゴリとの比較を行なうことで得点を決定する．ここで，目的言語の単語の意味カテゴリは，人手で目的言語の単語に意味カテゴリを付与した単語意味辞書から得る．また，目的言語文は，形態素解析がされているとする．

目的言語文の単語を格要素に対応付けたときの得点を決定するために，まず，意味カテゴリ cat が，条件カテゴリ $cond$ を満たすか否かの判定関数 $F(cond, cat)$ を定義する．

$$F(cond, cat) = \begin{cases} 1 & \text{if } cond = cat, \text{または}, \\ & cond \text{ が } cat \text{ の上位カテゴリ} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

以下の評価関数 $P(case, w)$ により，単語 w を格要素 $case$ に対応付けたときの得点を決める．

$$P(case, w) = \frac{1}{|cat|} \sum_j \sum_i p(cond_i) \cdot F(cond_i, cat_j) \quad (2)$$

ここで，$cond_i$ は $case$ の i 番目の条件カテゴリを，cat_j は w の j 番目の意味カテゴリを，$|cat|$ は w の意味カテゴリ数を示す．$p(cond_i)$ は，意味カテゴリと条件カテゴリの組に与える得点であり，条件カテゴリ $cond_i$ が下位の意味カテゴリで記述されているほど高い得点を与える．意味体系の木構造を深さにより 4 段階 (根〜深さ 1，深さ 2〜3，中間ノード，葉ノード) に分け，それぞれの段階に 1, 2, 3, 4 点の得点を与える．

また，多義を持っている単語が有利にならないよう，単語の意味カテゴリ数で除することで平均を取る．

3.2 日本語パタンと目的言語文の対応

前節で定義した評価関数より日本語パタンと目的言語文との間で類似度を計算する．目的言語文の単語と日本語パタンの格要素とのすべての組合せの中で，$P(case, w)$ の和が最も高いものを類似度とする．日本語パタン pat と，目的言語文 s の類似度 $Sim(pat, s)$ は次式より求める．

$$\because \quad Sim(pat, s) = \max \sum_i P(case_i, v_{ij}), \quad v_{ij} = w_j, v_{kj} \neq v_{lj}, k \neq l \quad (3)$$

ここで，$case_i$ は pat の i 番目の格要素を示し，w_j は s に含まれる j 番目の単語を示す．また，v_{ij} は，$case_i$ に対応付けたときの単語 w_j を示す．ある単語を，異なる格要素に同時に対応付けることはできないため，$v_{kj} \neq v_{lj}, k \neq l$ の条件をつけている．

各日本語パタンごとに目的言語文との類似度を求め，類似度が最大となる日本語パタンに目的言語文を対応付ける．

例えば，目的言語が中国語である場合，図 1 の例では，類似度の大きい pat_1 に中国語文 s が対応付けられる．

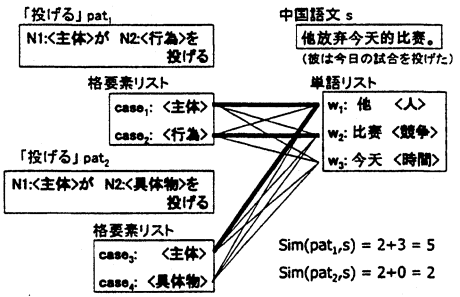

図 1: 日本語パタンと中国語文の対応付けの例

4 評価実験

本手法によって，コーパスから得られた目的言語文の格要素をどの程度正しく対応付けられるかを評価する．実験では，目的言語を中国語として，以下の各辞書，コーパスを使用した．

- 日英パタン対辞書: 日英機械翻訳システム ALT-J/E の一般パタン対辞書 [4] (約 13,000 パタン対)
- 日中対訳辞書: 小学館日中辞典 (日本語見出し約 83,000 語)
- 意味体系: ALT-J/E の一般名詞意味体系 [2] (約 2,700 カテゴリ)
- 中国語単語意味辞書: 中国語名詞約 55,000 語に人手で意味カテゴリを付与したもの
- 中国語コーパス: 小学館日中辞典の用例文 (約 12 万文，形態素解析済み)

評価対象

評価対象として，日英パタン対辞書と日中対訳辞書に共に見出し語がある日本語用言からランダムに

20 用言を選択した. 手法の評価のために, コーパスから得られる中国語文が 100 文以上のものから 20 用言を選択した.

日本語 20 用言に対して,

- 日英パタン対辞書から, 日本語パタン: 59 パタン
- 日中対訳辞書から, 日中の用言訳語ペア:144 ペア

が得られた. また, コーパスから, 日本語用言の訳語を含む中国語文 3,793 文が得られた.

正解データの作成

正解データとして, 以下の手順により, 人手で日本語パタンと中国語文の対応付けを行なう.

1. 日本語用言ごとに, 日英パタン対辞書より日本語パタンを得る. また, コーパスより中国語文を得る.
2. 得られた各文に対して, 文中の中国語用言と同じ語義を持つ日本語パタンを選択する. 同じ語義の日本語パタンが無い文に対しては対応付けを行なわない.
3. 文中の単語を, その文中での格要素としての役割に応じて日本語パタンの格要素に対応付ける.

その際, 対応付けの手法の評価のために, 形態素解析誤り, 中国語単語意味辞書の不備, 日英パタン対辞書の不備により対応付けが行なえない文は, 正解データに含めない.

人手で対応付けを行なった結果, 中国語 478 単語 (426 文) が日本語パタンの格要素に対応付けられた. 対応付けが行なわれなかった文は, 格要素となる単語を持たない文や, 中国語訳語が, 他の日本語用言に対応する語義で使用されていた文である.

実際の辞書構築の際には, 得られた文からパタン対が作成可能か人が判定を行なった後, 2 節の作業 3.-5. について支援を行なうことを想定しているので, 対応付けが行なわれた 426 文を評価対象とする.

評価

正解データにおいて単語が格要素に対応付けられた 426 文に対して, 本手法による対応付けを行なった.

辞書の構築支援では, 正しいパタン対の提示が効率に寄与する. また, 誤ったパタン対を提示されても, 誤りの判定に時間がかからなければ, 提示しないときと効率は同じであると考えられる. そこで, 単語と格要素の対応をどれだけ取れたのかを見るために, 中国語単語が対応付けられた格要素の格ラベルと正解データとを比較して, 次式の再現率で評価する.

$$\text{再現率} = \frac{\text{正解と本手法とで格ラベルが一致する単語数}}{\text{正解で格要素に対応付けられた単語数}}$$

正解データでは, 格要素 N1 に対応付けられた単語が 46.9%と最も多かった. したがって, すべての単語を格要素 N1 に対応付けたときの再現率 46.9%がベースラインとなる.

評価結果を表 1 に示す. 表中の「単語数」は格要素に対応付けられた単語数, 「一致数」は対応付けられた格要素の格ラベルが正解データと一致する単語数を表す.

表 1: 評価結果

	単語数	一致数	再現率
正解	478	-	-
本手法	592	324	0.680
ベースライン	732	224	0.469

5 考察

本手法の再現率は, ベースラインより上回っているため, 日本語パタンの格要素の制約条件を利用することが, 対応付けに有効であることが分かった.

誤りの原因を分類した結果を表 2 に示し, 以下それぞれの原因について考察する.

表 2: 誤りの分類

原因	単語数
a. 同一条件の格要素	74
b. 他の単語が対応付けられる	37
c. 単語の多義	11
d. 格要素の省略	14
その他	18
計	154

a. 同一条件の格要素

日本語パタンにおいて, 同一の条件カテゴリが複数の格要素に指定されていた場合, ある単語をどちらの格要素に対応付けるべきか決定することができない. 以下の例では, 日本語パタンの格要素 N1 と N2 とが同一条件となっているため, 中国語文の単語「敌军」をどちらに対応付けるべきか意味カテゴリからは決定できない.

日本語パタン:

N1:<主体>が N2:<主体>を 攻める

中国語文: 攻(攻める) 敵軍(敵軍)。

b. 他の単語が対応付けられる

ある条件カテゴリにマッチする意味カテゴリを持つ単語が文に複数あり, 対応付けるべきではない単語が格要素に対応付けられる. 特に, 格要素の条件カテゴリが, 任意の意味カテゴリとマッチすることを意味する<*>であるとき (19/37) は, すべての単語を対応付けることができるため, 対応付けるべき単語を意味カテゴリによって絞り込むことができない.

a. の同一条件の格要素を持つパタンと, b. の条件カテゴリが任意のカテゴリであるパタンは, 日英パタン対辞書の日本語パタン 13,282 のうち,

a. 同一条件を持つパタン: 3,448
b. 条件に<*>を持つパタン: 2,877
(a. と b. との重複は 1,079)

と, 合わせて 5,246 パタン (39%) であった. これらのパタンに対応付けられる文については, 意味カテゴリの比較では, 正しい対応付けが期待できないため, 構文情報を利用する方法などを検討する必要がある.

c. 単語の多義

複数の意味カテゴリを持つ単語には各意味カテゴリの得点の平均値を与えている. 格要素の条件カテゴリが下位の意味カテゴリであれば, マッチしたときの得点は高いが, マッチする意味カテゴリ数が少なくなり得点の平均値が下がる. 逆に, 上位の意味カテゴリであれば, 得点は低いが, マッチする意味カテゴリ数が多くなり得点の平均値は下がらない.

その結果, 意味カテゴリを多く持つ単語が, 上位の条件カテゴリを持つ格要素に誤って対応付けられることがある.

複数の意味カテゴリを持つ単語への得点の与え方をさらに検討する必要がある.

上記の a.-c. は, 目的言語に依存しない原因であるが, 言語に依存する原因として以下が挙げられる.

d. 格要素の省略

中国語では, 日本語と同様に文の主語や目的語などが省略されることがある. そのような文で, 省略された単語が対応付けられるべき格要素に, 他の単語が誤って対応付けられることがある.

中国語における格要素の省略の傾向を調査し, 省略されやすい格要素の得点を低くするなどの方法が必要である.

6 おわりに

既存の利用可能な資源である単言語コーパス, 日英パタン対辞書を利用したパタン対辞書の構築支援を行なうことを目的として, 目的言語文の単語を日本語パタンの格要素に対応付ける方法について述べた. また, 人手による対応付けを正解とした評価実験を行なった結果, 意味カテゴリの比較が対応付けに有効であることが確認できた. 本手法による対応付けからパタン対を作成し, 人に提示すれば, 辞書構築支援に有用であると期待できる. 今回は中国語を対象として実験を行なったが, 本手法は, 目的言語に依存した情報を利用していないため, 他の目的言語のパタン対辞書を構築する際にも適用可能である.

実際の辞書構築作業における効果を評価することが今後の課題である.

参考文献

[1] Hussein Almuallim, Yasuhiro Akiba, Takefumi Yamazaki, and Shigeo Kaneda. Induction of Japanese-English translation rules from ambiguous examples and a large semantic hierarchy. 人工知能学会誌, Vol. 9, No. 5, pp. 730–740, 1994.

[2] 宮崎正弘, 池原悟, 横尾昭男, 白井諭. 日英機械翻訳のための意味属性体系. 電子情報通信学会技術研究報告 NLC-97-12, pp. 29–36, 1997.

[3] 白井諭, 兵藤富子, 上田洋美, 横尾昭男, 池原悟. 日英機械翻訳用構文意味辞書の作成支援. 平成 7 年電気関係学会関西支部連合大会 G14-3, 1995.

[4] 白井諭, 横尾昭男, 中岩浩巳, 池原悟, 宮崎正弘. 日英機械翻訳のための構文辞書. 電子情報通信学会技術研究報告 NLC-97-14, pp. 45–52, 1997.

[5] 田中英輝. 動詞訳語選択のための「格フレーム木」の統計的な学習. 自然言語処理, Vol. 2, No. 3, pp. 49–72, 1995.

[6] 宇津呂武仁, 松本裕治, 長尾眞. 二言語対訳コーパスからの動詞の格フレーム獲得. 情報処理学会論文誌, Vol. 34, No. 5, pp. 913–924, 1993.