

4P-9 コーパスからの接続尤度情報抽出の一手法

野村直之 奥井伸司

(株)日本電子化辞書研究所

1. はじめに

形態素解析には、辞書引きの直後に接続検定を行って形態素列を絞りこむという手法が広く用いられている。この際に、曖昧性解消の効果を高める目的で、接続可/不可の2値から多値に拡張した接続マトリクスが提案されている。しかし、この多値の接続マトリクスを、実在の入力文における多種類の曖昧性に対して妥当な曖昧性解消の効果を発揮させるように作成し保守・改良するのは、非常に困難である[坂井88]。これに対し本稿では、1つの接続マトリクスにヒューリスティクスを直接埋め込まずに、2値のマトリクスが残した接続品詞の可能性の中から、相対尤度情報によって絞り込みを行なうという2段階に分けた選択方法を前提とし、その後段で利用する情報の抽出方法とその知識表現形式を提案する。

2. 日本語文における接続品詞の選好尤度

接続品詞ペア間の相対尤度情報の有用性を定性的に判断する目的で、具体例をとりあげて考察する。

例1 (設計者が予期しなかった機械翻訳の誤訳の例)

「それは十分だ。」→ "The winding is enough."

例1の入力文には、動詞「それる」の連用形の名詞用法が存在し得る(そちらが正解となるようなコンテキストを作れる)。この用法が偶々代名詞に優先して選択された結果上記の訳文となったものであるが、例1の解析を必須的に不適格とする強い文法的/意味的制約は存在しない。

この問題に対処するため、「平仮名の代名詞と助詞が隣り合った時には他の可能性に優先してそれらを選択する。」というルールを、解析エラーが発生した時だけ場あたりに記述しているのが現状の商用自然言語解析システム開発の一面面であると考えている。このようなエラー主導の対処では、同様の選好尤度が別の多くのコンテキストでは逆転している可能性、即ち解析の平均正解率が低下する危険性を見積ることが出来ない[Su91]。また、同様のルールがどの程度存在するのかの見積りを行なうにも至らず、システム改良の工学的的方法論が確立しているとはいえない。

例1における選好尤度のもう1つの特徴は、「設計者が予期しなかった」こと、すなわち母国語話者が正解以外の曖昧性に気づかない類の知識であった点である。そこで、コーパスという実データからこの種の選好尤度の知識を収集することが有効だと自然に判断される。コーパスからの知識抽出によれば、コーパスサイズに比例して知識量/精度を増せる可能性や、漸進的なシステムの改良[Muraki91]、対象分野における同種知識の網羅性など、上述の問題点を解決できるという期待がある。

個別語彙の隣接データそのものを事例として蓄積することは、現在利用可能な10数万文規模の解析結果付きコー

パスを資源として想定すると非現実的である。このため、互いに隣接する形態素をなんらかの範疇にまとめる必要がある。そこでその最初の試みとして、従来から分かち書きの有効な手掛かりと考えられている文字種と接続品詞を範疇として、曖昧性の中で選択された正解と不正解とのヒストグラムデータから接続品詞ペア間の相対尤度情報を抽出する枠組を以下のように設計した。

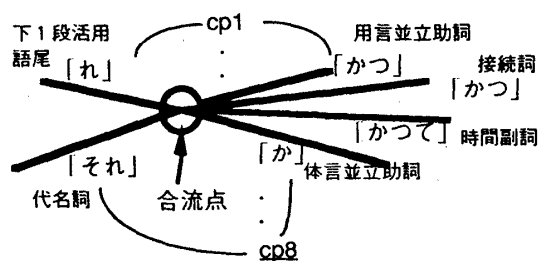
3. コーパスからの接続品詞ペア間相対尤度情報の抽出方式

日本語の形態素解析における2値接続マトリクス上の選好尤度情報を蓄積する目的で、適切なデータ量を見積った。人手で解析結果の正解を選択した10数万文規模の付きコーパスを資源として利用するという前提から、以下の要件を必要と考えた。

[必要条件] 1) 既存のコーパスから全自動で抽出する。

2) コーパスのサイズに比して有意な尤度差が得られるように母数を設定する(母数が大きすぎてヒストグラムにならなくなってしまっては困る)。

従来から日本語文分かち書きのための有効な手掛かりと考えられてきている文字種と接続品詞の数は高々数千程度とみられる。ここで、2値接続マトリクスの適用後にどの程度の接続品詞ペアの曖昧性が存在し得るかを図1をもちいて観察してみる。



CP = Connecting number Pair i (i=1,2,...,1214)

図1 「それかつてを頼って」の形態素全探索結果

形態素境界「...れ/か...」において接続可能な組み合わせは、図1に示すように、 $2 \times 4 = 8$ 個である。接続品詞は数百個程度設定するのが通例であるが、2値接続マトリクスが許容する接続品詞ペアの組み合わせはその2乗の数値にはならず、高々数千程度と見積られる。一例として[EDR88]における2値接続マトリクスによればその接続品詞ペア数は1214個である。

ここで、仮に、形態素境界に存在する接続品詞ペアの曖昧性の組み合わせのすべてを母数とすることを想定してみる。すなわち、図1におけるcp1,...,cp8という8個の接続品詞ペアのセットを1つの固定的な統計データ蓄積の枠とし、それと一部でも異なるものがあれば、互いに全く無関係に

正解データのヒストグラムを作るという方法である。この場合、そのヴァリエーションは組み合わせ的に増大し、実際に数10の階乗を数回乗じた程度のオーダーになる恐れがある。また、定性的にも、例1における選好尤度と明らかに根を同じくする2つの接続境界の事例「それ／は」、「それ／でも」が別個に扱われてしまう。

このため、曖昧性の存在条件を一部捨象し、2つの接続品詞ペア間の相対尤度のみを以下の手順で相対尤度データを図2の集計テーブルに蓄積することとした。

	cp1	cp2	cp3	cp4	cp5		cp1214
cp1		12	8	6			
cp2							
cp3		3					
cp4							
cp5							
cp1214							

図2 単語境界上に存在する接続ペア間の相対尤度情報を抽出したテーブル

[手順]

- 0 予め、2値の接続マトリクスを用いて辞書引き可能性の全探索を行ない枝分れ図を作成しておく（存在する形態素数に比例する処理時間で実行可能）。
- 1 1つの形態素境界に着目し、すべての接続品詞ペアの曖昧性を検出する。
- 2 正解の接続品詞ペア（予め人手で選択したもの）を図2の集計テーブルの該当行にマッチングさせる。
- 3 不正解の接続品詞ペア全てについて、正解の接続品詞ペアの行の該当列にマッチングさせ、その箱の数値を+1インクリメントする。

なお、上の手順において、左右に未定義語を含む形態素境界はカウントの対象としない。

4. 接続品詞ペア間相対尤度情報の利用可能性

図2の形で出来上がった相対尤度テーブルは、解析の過程の任意のタイミングで、保持している曖昧性に対して直接適用可能である。たとえば例1の入力文「それは十分だ。」では、解析木が完成し、文脈条件から個々の語の曖昧性を解消しようとする試みが失敗に終わった時点で「代名詞+助詞」対「ひらがな表記の動詞連用形+助詞」の相対尤度のデータをテーブルから得てデフォルト選択を行なうこともできる。

一般には、パーザの能力や適用タイミングを制御するのが妥当と予想される。すなわち十分能力の高いパーザ、多数の曖昧性をもった入力（形態素解析結果）を十分高速に

処理できる横型パーザと組み合わせる場合は、パージング後、即ち文法的／意味的適格性によって候補を絞り込んだ後に相対尤度情報を適用するという安全な手順をとるべきであろう。これに対し、能力の低いパーザで速度を稼がねばならないような場合は、多値の接続マトリクスによるアプローチと同様にパージング前に適用することもできる。

極端な場合として、パーザ無しで形態素抽出と形態素選択部しかもたない軽量の言語処理フロントエンドを想定する。本来文法的条件に基づいて絞られるべき曖昧性解消条件を相対尤度テーブルが一部代行し（たとえば「で（格助詞／助動詞）」のヒューリスティクス）、平均的に正解率の高いデフォルトを選択させられるようになる可能性がある。

一方、入力文の構文的複雑度に応じて相対尤度テーブルの適用タイミングをパージングの前後に振り分ける戦略も考えられる。例1のような単文ならばより後方のタイミングとし、長い複雑な構文に対してはパージングの前にある程度相対尤度テーブルによって選択を行なっておくという戦略である。タイミングの決定には、構文的条件以外にも、たとえば相対尤度テーブル自身の情報をも利用することも可能である。すなわち、図2のテーブルに蓄積されたデータについて、絶対頻度／相対頻度のしきい値を設定し、たとえば「出現数の差(cp1-cp2)が10以上で出現数の比(cp1/cp2)が10以上」という条件を満たすものだけをパージング前に適用するという方法である。

上記のような選択情報適用のタイミングの自由度は、単語自身の出現頻度や品詞／意味素性の接続確率等の別の情報についても同様である。相対尤度テーブルも含め、各々別個に設定できるという柔軟性も有している。

5. おわりに

以上、高精度で保守性の高い形態素解析を実現する目的で、予め用意した大規模コーパスから形態素品詞ペアに関する相対尤度情報を抽出し、利用する枠組を示した。手法の1案を示した。今後、本手法の有効性を定量的評価するための実験を行なう予定である。この際にコーパス自身の検証と精度向上への有効なフィードバックの手順を具体的に確立したい。

さらに、評価結果におけるエラーの定性的分析を行ない、接続品詞以外にどのような別種の素性について相対尤度情報が有効となっているかを探索していきたい。

参考文献

- [坂井88] 坂井他「PIVOT J-E:日本語の形態素分解」, 信学全国大会 '88年秋季, D-138
- [Su91] Su, Keh-Yih, "A New Quantitative Quality Measure for Machine Translation Systems", Proceedings of COLING92 (to appear)
- [Muraki91] Muraki, K. "Machine Translation Systems and Large-Scale Electronic Dictionaries", Proc. of the Int'l Workshop on Electronic Dictionaries, EDR TR-031.
- [EDR88] EDR TR-015, 「辞書開発支援システム」 pp.16-「表2-1 日本語連接行列」