文脈辞書を用いた頑健な多義性解消

那須川 哲哉
日本アイ・ビー・エム株式会社 東京基礎研究所
nasukawa@trl.ibm.co.jp

1 はじめに

ネットワーク技術と計算機の能力の向上により, 世界中の電子化された文書に容易にアクセスできるようになった現在, 膨大な文書から必要な情報を効率良く得るための基盤技術として, 自然言語処理の必要性はますます高まっている. 例えば, World Wide Web (WWW) 上の外国語の文書を母国語で読みたいという欲求を満たすため, パーソナルコンピュータ上で動く安価で高速な機械翻訳ソフトが大きな市場を形成している. しかし, その翻訳品質は充分なレベルに達しておらず, 品質向上のための高度な自然言語処理技術が求められている. また, 大量の文書を分類したり, そこから必要な情報を検索したり, 文章内容を要約したりするアプリケーションなど, 自然言語処理の適用範囲が広がっている.

大量の文書から効率的に情報を得ることを目的としたアプリケーションの場合, 多義語の語義を正しく決定する処理能力の重要度が比較的高くなる. 例えば機械翻訳であれば, 翻訳結果を斜め読みして概要を把握できることが重要であり, 完全な翻訳文を出力できない限りは, 表現に凝った文章を生成するよりも, 意味的に正しい訳語が表示され, 文中の言葉を拾い読みするだけで, 何に関して記述された文章かを把握できるような翻訳結果を出力することが望ましい. また, 大量の文書を分類したり, そこから必要な情報を検索したりするためのインデックスを作成する場合にも, 例えば人名にも地名にもなり得る「山口」のような多義語が, 処理対象の文書中でどの意味で用いられているかを正しく認識する能力が重要である.

このような観点から, 本稿では, 多義語の語義を決定する処理の精度向上に取り組んだ結果について報告する.

2 文脈辞書を用いた多義性解消

2.1 文脈辞書

多義語の語義決定は, 自然言語処理における主要テーマの一つであり, 様々な研究が存在するが, いずれの手法も, 文章が記述する対象世界に関する何らかの知識を必要とする点では共通している. 多義語の語義決定の知識は, 1980年代までは専ら人手で構築されていたが, 多人数で作業する場合の品質の一貫性の問題やコストの問題などから, 1990年代には, コーパスから自動的に抽出した知識を利用する手法が盛んに研究開発されるようになった. ところが自動的に抽出した知識は, 抽出源のコーパスの分野に偏る傾向があり, 分野の異なる文書においては適用率が低かったり, 誤った語義を選択する原因となる場合も多い. そこで, 知識の適用率を高め, 語義決定精度を向上させるためには, 処理対象の文脈を考慮して文脈に即した知識を適用し, 曖昧性解消結果を文脈内で共有する枠組を取ることが望ましい.

多義語は同じ文脈内では同じ語義を取る傾向が強いため, 同じ語が同じ文脈内で同じ語義を取るような制約を適用して語義決定を行なうことで, 語義決定の精度が向上するという結果が得られている [1, 2]. その際, 文脈の範囲をどう設定するかが, 処理精度を向上させる重要なポイントである. 文脈の範囲を狭めるほど, 文脈内で同じ語義を取る傾向は強くなる. しかし, その反面, 文脈内の語数が減ると, 同じ語が繰り返し出現する割合が低くなり, 語義決定結果を共有できる割合が低下する. 多義語の振る舞いを実際の文書で調査してみると, 英文計算機マニュアルの分析結果 [2] では, 100文を越える程度の大きさの章であれば, その中に出現する多義語の約80%は同じ章の中で複数回出現するという結果が得られている. その中で同じ多義語が同じ語義を取る割合は, 名詞の場合で97%程度と非常に高いが, 異なる複数の語義を取る場合でも, 局所的な文脈では同じ語義を取ることが多い. 例えば, ある英文計算機マニュアルの277文からなる章においては, "number" という語が5回出現し, 「番号」「数」「数字」という3種類の意味で使われている. この文章中で "number" が各々の意味を取る文の位置を見てみると,

番号 : 第88文
数 : 第143文, 第144文, 第147文
数字 : 第209文

と, 近い位置では同じ意味で用いられる傾向がある. 従って, 多義語が同じ文脈内で同じ語義を取るという文脈制

約を適用する際に, 固定された範囲の文章内で常に同じ語義を取る機構よりは, 位置的により近い語ほど同じ語義を共有し易くなる機構の方が処理精度が向上すると考えられる.

このように, 多義語が同じ文脈内で同じ語義を取るという文脈制約を適用するためには, 処理対象文章内での多義語の語義決定に関する情報を蓄積しておき, 蓄積された情報から最も妥当と判断される語義を同じ文脈内の同じ多義語に対し選択できるように記録する機構があれば良い. 本稿ではこのような記録機構を文脈辞書と呼ぶ. 基本的には, 一つの文脈を構成する全ての文を処理する過程で文脈辞書を構築し, 各多義語の語義決定において, 選択すべき語義が文脈辞書中に記述されていれば, その語義を優先的に選択する. この処理によって, 同じ文脈における同じ多義語の語義の一貫性を保つことができ, 結果的に多義語の語義決定の精度を向上させることができる.

2.2 複合語辞書を用いた語義決定

文脈辞書を構築するためには, 何らかの語義決定の仕組みが必要であり, そのためには何らかの知識が必要となる. しかしながら複雑な知識の構築は, 作業コストの高さや複数の作業者による整合性の維持といった問題から困難である. 従って, 広範囲の分野の文書を対象とした自然言語処理システムを構築する上では, 知識の表現形式は可能な限り単純であることが望ましい. その観点から, ここでは, 特別な知識を用いずに通常の辞書中の情報を用いた語義決定手法を提案する.

例えば, 機械翻訳システムにおける最も基本的な辞書としては,

apple： りんご
trial： 試験
trial： 裁判

のように, 原言語の語と対象言語の訳語を対にした訳語辞書があげられる. また, 語のカテゴリ分類を行なうシステムであれば,

新潟： 地名
山口： 地名
山口： 人名

のように, 語と対応するカテゴリ名からなる辞書が最も単純な辞書と考えられる.

このように単純な構造の辞書を用いて処理する場合, 対応する訳語やカテゴリ名が複数存在する語 (上の例では “trial” 及び「山口」) が多義語となる.

このような単純な辞書を用いても, 辞書中に含まれている複合語の情報を活用することで, 多義語の語義を決定することが可能になる.

一般的に, 単純な構造の辞書を用いた自然言語処理システムの精度を上げようとする場合には, 辞書中に複合語の情報を登録する傾向が見られる. 例えば, 前出の単純な辞書データでは, 対応する訳語が「試験」「裁判」と複数存在する多義語 “trial” を訳し分ける方法が無く, 例えば辞書に記述された順番に依存して, 常に同じ訳語が選択されることになる. この問題を解消するためには,

civil trial： 民事裁判

のように “trial” を含む複合語を登録し, 複合語として文中に出現した場合には, 正しい訳語が選択されるようにする. また, 語のカテゴリ分類に関しても,

山口県： 地名
山口洋子： 人名

のように, 複合語の場合には曖昧性が無い場合が多い.

このように, 多義語であっても複合語の中では語義が確定している傾向が強いという性質が見られる [4] ため, この性質を利用することで, 辞書中の複合語の情報が適用された際に, 同じ文脈における多義語の語義を推定することが可能になる. 例えば, “civil trial” という語の訳語が「民事裁判」であるという辞書情報が適用された文脈の中では, “trial” は「裁判」の意味で用いられてると推定することができ, 地名である「山口県」が存在する文脈中では「山口」が地名を示していると推定することができる.

3 アルゴリズム

本節では, 前節で示した文脈辞書を用いた多義性解消の具体的な実現方法を示す.

3.1 文脈辞書の構築

文脈中に存在する多義語に関して文脈辞書に記述する語義情報は, ある多義語を含む複合語の辞書情報を適用した際に, その多義語に対する単語の辞書情報と複合語の辞書情報との単純な文字列比較処理により得ることができる.

例えば, 機械翻訳システムにおいて “civil trial” という複合語が辞書に登録されている場合, この複合語が文脈中に存在すれば, その中に含まれる “trial” という多

義語に関する語義情報を文脈辞書に記述することができる．文脈辞書に記述する内容は，“civil trial” の訳語（仮に「民事裁判」とする）と “trial” の訳語の各候補（仮に「試験」及び「裁判」とする）と文字列比較することにより，一致文字数のより多い訳語（この場合は「裁判」となる）が，同文脈中の訳語として，より確からしいという情報である．ここではこの確からしさを数値化し，選好度として扱う．その結果

```
trial : 裁判 (saiban)  2.0
        試験 (shiken)  1.0
```

のように，各訳語候補に対する文脈中での選好度を一元化して加算し，比較することが可能になる．従って，文脈全体を処理した上で，各語義候補に対する選好度を比較し，選好度最大の語義を選択することにより，文脈全体において最も確からしい語義を決定することができる．選好度は，各語義に対する確信度を示すものであるため，一致する文字列が長いほど，また，複合語中の単語数が多いほど，高い値を付けるような評価式により計算する．

3.2 文脈辞書の適用

文脈辞書が構築されれば，語義決定の情報が文脈辞書中に記述されるため，文脈辞書中で選好度の最も高い語義を選択しつつ処理を行なうことで，語義決定の精度を向上させることができる．また，文脈辞書中に選好度を記録する際に，各語義候補を支持する複合語の文中出現位置も記録しておけば，文脈辞書参照時に，処理対象の多義語の位置を参照し，位置的により近い複合語の支持する選好度の重みを高くした上で，各語義候補の選好度を比較することにより，その多義語の存在する局所的な文脈において最も適切な語義を決定することが可能になる．

3.3 処理の流れ

文脈辞書を参照して語義を決定するためには，語義決定処理の前に文脈辞書を構築しておく必要がある．従って，文章全体を一旦処理して予め文脈辞書を構築した上で，再び文章の先頭から各文を処理することにより，文章中の全ての多義語の処理において，文脈全体の情報を反映した文脈辞書を参照することが可能になる．

しかし，オンラインで WWW のホームページを翻訳する場合など，即時性が要求される際には，前もって文脈辞書を構築することなく，文章の先頭から一文ずつ処理することになる．その場合には，処理対象文以前の文脈の情報のみを反映した文脈辞書を参照して多義語の語義を決定する．その際，各多義語の語義選択結果を記録しておくことで，最終的に（もしくは処理途中で），文脈辞書の支持する語義と各多義語の語義選択結果との差分を計算した内容をユーザに提示することができる．その結果ユーザが再処理すべきと判断すれば，ユーザの希望に応じて必要な文を再処理することで，文脈全体の情報を反映した処理結果をユーザに提供することが可能になる．

4 分析結果

前節で示した処理の有効性を評価するため，実際の自然言語処理システムの辞書内容と，処理対象となる実際の文章を調査した．

市販されている機械翻訳システム[1] の一時点での辞書内容を調査したところ，登録されている 25 万件以上の翻訳対（原言語と訳語のペア）のうち，35%以上が複合語の翻訳対であった．また，単語の約 50%が多義性を含んでいたのに対し，多義性を含む複合語は 15%未満であった．さらに，多義性を含む単語全体の約 35%が，複合語の辞書見出しに含まれていた．複合語の辞書見出しに含まれる多義語の割合は，訳語候補数が多いほど高いという傾向が見られ，特に，訳語候補数が 5 語以上の多義語のうち，72%以上の語が複合語の辞書見出しに含まれているという結果が得られた．

すなわち，複合語の場合には意味の曖昧性が少なく，その中の語義が確定している可能性が高いと共に，曖昧性の高い多義語の多くが複合語に含まれるという結果が得られたことにより，前節で示した手法で有効な語義決定を行なえる見通しが得られた．

次に，本手法を機械翻訳システムに適用した場合に，実際の文章中の語義決定（この場合は訳語選択）においてどの程度の効果があるかを，CNN が WWW で提供[2]している英文ニュース記事を対象にして調査した結果を表 1に示す．1997 年 8 月 25 日から 9 月 3 日までの 7 日分のヘッドニュースを調査したところ，一記事の平均 64.9 文（14.2 単語/文）の中に多義語を含む複合語が 57.3 語存在した．これらの複合語中の多義語に対する語義選択情報が文脈辞書中に記述されることになり，その文脈辞書を参照する結果，参照しない場合と比較して，文中の訳語が向上する文が記事全体の 14.5%を占めるという結果が得られた．また，記事の表示されているホームページには，他のページへのリンク情報など，記事本文とは直接関係の無い文も含まれているため，そのような文を

[1] 日本ＩＢＭ(株)『インターネット翻訳の王様』
[2] http://www.cnn.com/

表 1: 英文ニュース記事の翻訳における複合語の適用状況と訳語選択精度の向上

記事の日付	記事中の文数	記事中単語数	適用した複合語数	適用した複合語に含まれる多義語数	翻訳向上文数	記事ページ全体における翻訳向上文の割合	記事本文における翻訳向上文の割合
08/25	49	736	123	38	10	20.41% (10/49)	28.57% (10/35)
08/26	82	1093	177	71	10	12.20% (9/82)	16.98% (9/53)
08/27	43	537	108	39	8	18.60% (8/43)	33.33% (8/24)
08/28	62	943	160	70	18	29.03% (17/62)	40.48% (17/42)
09/01	78	1129	161	73	3	3.85% (3/78)	4.17% (3/72)
09/02	57	849	122	43	5	8.77% (5/57)	9.80% (5/51)
09/03	83	1168	169	67	12	14.46% (12/83)	15.58% (12/77)
全体	454	6455	1020	401	66	14.54% (64/354)	18.08% (64/354)

除いて, 本文のみで評価すると, 訳語が向上する文の割合は, 本文の 18.1%となった.

5 まとめ

本稿では, 大量の文を高速に処理する自然言語処理応用システムの基盤技術として, 単純な処理の枠組で多義語の語義決定精度を向上させる処理手法を示した.

本手法の評価は機械翻訳システムを対象として行なったが, 同様の処理をキーワードのカテゴリ決定に応用しても効果があるという結果が得られている [3].

また, 文脈辞書に蓄積される情報は, ユーザーが選択した一つの文書における多義語の語義の傾向を示すものであるが, 同じユーザーが選択する複数の文書に何らかの一貫性があるとすれば, 文脈辞書に記述された情報は, 文脈辞書を構築した対象文書のみでなく, 同じユーザーが選択する他の文書を処理する上でも有効であると考えられる. その場合には, 文脈辞書の情報を, 新しい情報ほど大きな重みをつける形で蓄積していき, 個人辞書として利用することで, ユーザーが何も意識せずにシステムを使い込むだけで, 処理精度が向上していく仕組みを実現することが可能になると考えられる.

参考文献

[1] William A. Gale, Kenneth W. Church, and David Yarowsky. One Sense per Discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*. (1992).

[2] Tetsuya Nasukawa. Discourse Constraint in Computer Manuals. In *Proceedings of TMI-93*, pages 183–194. (1993).

[3] 那須川哲哉. 文脈情報を利用したキーワード語義決定. 第 11 回人工知能学会全国大会, pages 348–349. (1997).

[4] David Yarowsky. One Sense per Collocation. In *Proceedings of ARPA Human Language Technology Workshop*. (1993).