単言語パラレルテキストからの同義語獲得

下畑 光夫, 隅田 英一郎
ATR 音声言語コミュニケーション研究所
{mitsuo.shimohata,eiichiro.sumita}@atr.jp

1 はじめに

同じ意味を共有する異なる表記の語は同義語と呼ばれ, 数多くの同義語が存在することが知られている. 辞書, シソーラスなど人手により同義語情報が作成されているが, 作成に多大な手間を要するという問題があり, 同義語情報の自動獲得が大いに期待されている.

本論文では, 単言語パラレルテキストから同義語を獲得する方法について述べる. 単言語パラレルテキストとは, 同等の内容を持つ同一言語で記述されたテキスト[1]と定義する. 同等の内容を共有するために, 単言語パラレルテキスト間において包含する語ならびにそれら語間の関係も類似性が高い. この性質により同義語獲得に適したデータであるといえる.

単言語パラレルテキストは一般的にはほとんど公開されていないが,Barzilay ら [1],Shinyama ら [2] が, ニュース記事を利用して自動作成している. ニュース記事の日付や含まれている単語の重なりなどを利用して, 同一もしくは類似の事件を記述した記事のクラスタリングを行っている.

また, 単言語パラレルテキストは,“同一の原文を複数人が翻訳して得られる異なる訳文” という方法でも得ることができる. このタイプのデータは, 機械翻訳 (MT) の自動評価に利用するという目的でいくつかが人手で作成されている.MT の評価型ワークショップとして,NIST MT evaluation[2]や IWSLT[3]が過去に開催されており, それらのワークショップでは自動評価用の単言語パラレルテキストが作成されている.

本論文の提案手法の基本アイデアは, 「単言語パラレルテキスト間で対応する同義語は, 同じ局所的文脈を持つ」というものである. そして, 局所的文脈として獲得対象語の前後 1 単語という緩い定義を採用している. この定義により, 全体的構造が大きく異なるテキスト間からも獲得が可能という特長を実現している. しかし, この緩い制約だけでは誤った獲得が多く発生するため, さらに「品詞の一致」と「外部出現の禁止」という 2 つの制約を付与することで獲得精度を高めている.

本論文では, 一般に利用可能な 2 種類の単言語パラレルテキストを用いた評価実験についても報告する.

[1]ここでの「テキスト」は, 文書, 記事, 文, 句など様々な単位を想定している.
[2]http://www.nist.gov/speech/tests/mt/index.htm
[3]http://www.slt.atr.co.jp/IWSLT2004/

2 獲得手法

同義語獲得は, 単言語パラレルテキストの対から (1) 局所的文脈が一致する,(2) 品詞が一致する,(3) 外部出現がない, の 3 つの制約を満たす語対を獲得することで行う. 以下の節でこれら 3 つの制約について述べる.

2.1 局所的文脈の一致

単言語パラレルテキスト対の間に存在する同義語の対は, 同じ文脈で用いられていると考えられる. 逆に言えば, 同じ文脈で用いられている語の対が同義語対を成す可能性が高い. そこで, 同義語対の認識に局所的文脈情報を用い, その定義を「獲得対象語の前後 1 語ずつ」とした. 図 1に前後 1 語の文脈情報を用いた語対抽出を示す. 図中,2 組の語対はゴシックで書かれた単語 (同義語部) が前後を同じ単語 (文脈語) で挟まれている.

本制約では, 前後 1 単語というごく近接の文脈しか考慮しないため, 図 1のように文構造が大きく異なるテキスト対からも獲得できるという利点を有している. しかし, この制約は偶然に文脈語が一致したに過ぎない誤った語対を多く抽出してしまう. そこで, 以下の節で述べる 2 つの制約により, 誤った語対を除去して精度を高める.

2.2 品詞の一致

品詞を利用した制約として, 得られた語対の同義語部分の品詞が一致しなければならないこととした. これは, 品詞により同義語部の型の一致を図ったものである. 本論文では, 同義語の獲得対象は名詞, 動詞, 形容詞, 副詞としている. 動詞, 形容詞, 副詞の場合は, 各々の品詞からなる 1 単語を同義語の構成要件とし, 名詞の場合は

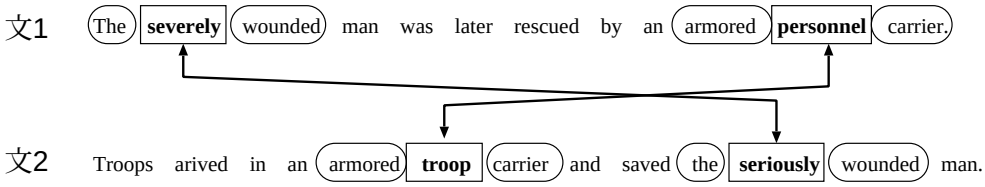

図 1: 局所的文脈を利用した同義語対獲得

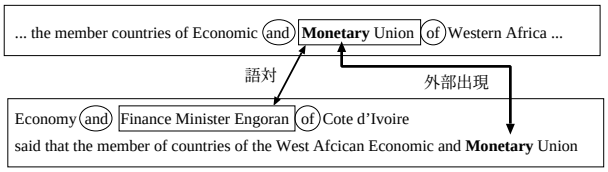

図 2: 外部出現の検出

複合語の存在も考慮して任意長の名詞列を構成要件とした.

なお, 同義語部が名詞の場合, 前接の文脈語が名詞または形容詞である時はその文脈語は複合語の先頭部と見なして同義語部に編入する. また, 後接の文脈語が名詞である場合も複合語の末尾部と見なして同義語部に編入する.

2.3 外部出現の除外

獲得された語対について, 一方のテキストから取られた同義語部の単語が他方のテキストの関係ない部分 (同義語部および文脈語以外) に現れるような現象を外部出現と呼ぼう. 外部出現が見られる場合, その語対はテキスト対から誤った部分を捕えており, 正しい部分はその外部に出現した語である可能性が高い. したがって, 外部出現が見られる語対は獲得対象から除外する. これは, 多くの内容語は 1 テキストにおいてたかだか 1 回しか出現しないということを前提としている. 事実,MTC データ (3.1節) においては, 名詞, 動詞, 形容詞, 副詞の 95.2% が 1 文中に 1 度しか出現しないことを確認している.

図 2に外部出現が見られるテキスト対を示す. テキスト対からは,“Monetary Union” と “Finance Minister Engoran” という名詞対が抽出される. しかし, 前者の名詞に含まれる “Monetary” という語は図中下部のテキストにおいて同義語部および文脈語に含まれずかつ外部に出現している. したがって, この名詞対は除外される.

3 実験

実験では, 機械翻訳の自動評価用に作成された Multiple-Translation Chinese (MTC) コーパスと Web 上のニュース記事を自動クラスタリングしてできたデータ (Google News) の 2 種類を用いた. 両データはテキストデータであるため, これらに Charniak パーザ [3] を適用し, 各単語に対する品詞情報を取得する. なお,Charniak パーザから得られる構文的情報については, 一切利用していない.

	MTC	Google News
テキストクラスタ数	993	61
テキスト数	10,655	394
語数	302,474	176,482
テキスト／クラスタ	10.7	6.46
語／テキスト	28.4	447.9

表 1: 実験データの統計量

表 1に, 両実験データの統計量を記す. 表中, テキストとは MTC データでは文,Google News データでは記事を指す.

3.1 MTC データ

Linguistic Data Consortium (LDC) から, 機械翻訳の自動評価用データ[4]がリリースされている. 中国語のニュース記事を基とした 993 文に対し,11 種類の人手による英語翻訳を付与したものである.

この MTC データに対して同義語獲得を適用したところ, 名詞が 2,942 対, 動詞が 887 対, 形容詞が 311 対, 副詞が 92 対獲得できた. その一部を下に示す.

名詞	press conference foreign funds	news conference foreign capital
動詞	complete	finish
形容詞	notable	significant

この獲得比率は, データ中の各品詞に属する単語の出現率と概ね一致する. 得られた語対について次の 2 種類の評価を行った.

- **元テキスト[5]なし評価**
  得られた語対のみを評価者に示し, 同義語と考えて

[4]LDC Catalog Number LDC2002T01
[5]同義語獲得の元となった単言語パラレルテキストを指す.

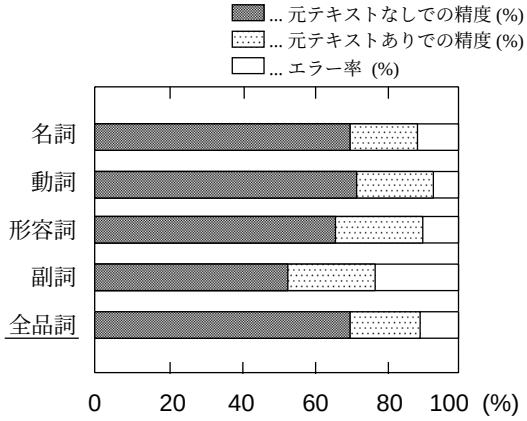

図 3: LDC データに対する結果

よいか否かを判定する.“同義語” の範囲は, 一般の辞書に記載されている程度を想定する.

- **元テキストあり評価**
  上記の元テキストなし評価において “誤り” とされた語対について, 元テキスト対も合わせて評価者に示して再度評価を行う.“誤り” と評価された語対が, 示されたテキスト対の中で正しい位置から獲得されているかどうかを評価する. したがって, この評価で “正解” とされた語対は, テキスト対から文脈的に正しい位置から語対を抽出することに成功したが, その語対の異なり方が同義語の範囲を超えたことを表している.

図 3に評価結果を示す. 品詞全体を通じての同義語獲得精度は 70.0% となった. また, 元テキストありで評価した場合の正解も含めた精度は 89.5% であった. これは本手法がテキスト対の中から 9 割近い精度で適切な部分を捉えていることを示す. しかし, テキスト対から文脈上適切な語対を取り出したとしても, 意訳であるとか文脈に深く依存するなどの理由で同義語とはいえない場合が少なからず生じている.

また, 得られた同義語対を WordNet[6] と照合し, 一般的な同義語知識がどの程度含まれるかを検証した. なお, 名詞については獲得知識の中に多くの固有名を含んでいるため, この検証からは除外した. 獲得した動詞, 形容詞, 副詞の対の語形変化を基本形に変換すると, 合わせて 1,001 対が得られる. この 1,001 対の内,951 対 (95.0%) は WordNet に語対を成す 2 語が登録されていた. さらに, この 951 対の内,205 対 (21.6%) については WordNet で同義語として登録されていた. つまり, 一般的な同義語知識は 2 割程度含まれていることが分かる.

[6]http://www.cogsci.princeton.edu/~wn/

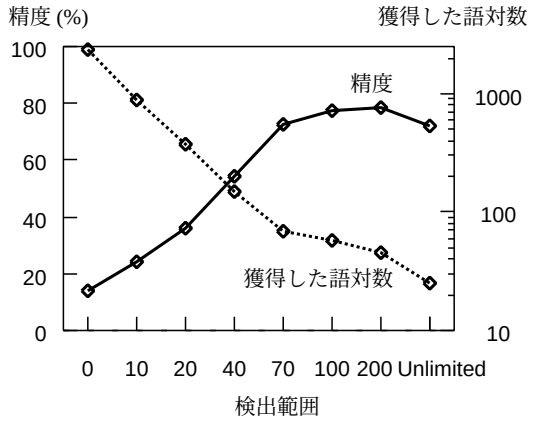

図 4: Google News データにおける外部出現の検出範囲と精度

3.2 Google News データ

Google, Inc. より,Web 上で公開されているニュース記事を自動クラスタリングする Google News [7]が提供されている. Google News では, 約 4,500 のニュースサイトから記事を収集し, 同じ事件を扱っている記事をクラスタリングする. この Google News から 2004 年 12 月 16 日から 2005 年 1 月 11 日に渡り,61 の事件について 394 記事を収集した.

Google News データでの実験では, 記事全体を一つのテキストと見なして同義語獲得を適用した. 本手法は局所的文脈情報を基としておりテキスト全体の構造には依存しないという点から, このような設定での適用が可能である. また, この設定により, テキスト対中の任意の場所の同義語対を獲得できるという利点も生じる.

しかし, この設定を採用した場合に, 外部出現の検出範囲をテキスト, すなわち記事全体とすると検出範囲としては広すぎると思われる. そこで, 検出範囲を文脈語の外側± n 単語とし,n を 0(外部出現検出なし) から unlimited(テキスト全体) までの範囲で 8 箇所設定した. なお, 他の品詞は獲得数が十分でないという点から, 獲得および評価対象は名詞のみとした.

図 4に実験結果を示す. 検出範囲を拡大すると, 獲得語対数が減少して精度が向上するが, 検出範囲が 100 を超えるあたりで精度は頭打ちとなる. 検出範囲が 100 を超える領域での精度は平均で 76.0% であった. また, 精度が 13.8% から 76.0% まで改善されたことは, 外部出現の除外の大きな効果を示している.

[7]http://news.google.com/

4 関連研究

本研究と同じく, 単言語パラレルテキストから同義語 (語彙的パラフレーズ) を獲得する研究を 3 つ取り上げる. Barzilay らは, 本手法と同じく同義語を挟む局所的文脈を利用して獲得している [4]. 利用する制約は局所的文脈のみであるが, 参照する範囲を最大 3 語とし, 正例, 負例を用いた学習により検出精度の高い文脈の選別を行っている.

Pang らは, パラレルテキストから構文情報を利用して top-down のラティスを構成することにより, 同義語を獲得している [5]. 構文情報を拠り所として語の対応をとるため, 局所的文脈が一致しなくても同義語獲得が可能という利点がある. その反面, パラレルテキストは構文的に類似している必要があるという制約が加わる.

下畑らも, 局所的文脈情報として前後 1 語を採用している [6]. 2 つのパラレルテキスト対において DP マッチングを適用して置換オペレータを抽出し, それを語彙的パラフレーズとしている. その際に, 編集距離が 2 以下のテキスト対は獲得対象から除外される. これは, テキスト全体にわたる語順を考慮しており, テキスト対が語彙と構文双方で類似している必要がある.

また, 単言語テキスト (非パラレル) を用いた単語クラスタリングについても触れておこう. 単語クラスタリングでは, 何らかの観点で類似の関係をもつ語群を意味的に類似である語であるとしてクラスタリングする. そのような観点としては, 対象語が出現する文書, 共起する単語, 修飾する形容詞と名詞などが一般的である [7]. しかし, 同義語を獲得するという課題に対しては, 非パラレルコーパスはその中に上位語や対義語といった様々な語も含まれることとそれらの関係も多様であるという点が欠点となる. そのため, この方式では同義語だけをその他の上位語や対義語から分離することが困難である [8].

5 まとめ

本論文では, 単言語パラレルテキストから同義語を獲得する方法について述べた. 提案手法では, 局所的文脈の一致を基とし, 品詞の一致と外部出現の除外という制約を加えて同義語を獲得する. 本手法は, 必要となる知識が少ない上, テキストの全体的な構造に依存しないという特長を有している.

MTC データを用いた実験では, 提案手法は 70.0% の精度で同義語を獲得することができた. また, テキスト対から同義語に該当する適切な場所を抽出するという観点では 89.5% の精度を達成でき, 提案手法の基本性能の高さを示すことができた. さらに, 自動クラスタリングされたデータである Google News に対して 76.0% の精度で同義語を獲得することができ, 本手法のロバスト性を示すことができた. 特に, 外部出現の除外による制約は, 精度を 13.8% から 76.0% まで改善させており, 大きな効果があることを示した.

謝辞

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである.

参考文献

[1] R. Barzilay and L. Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proc. of HLT-NAACL 2003*, pp. 16–23, 2003.

[2] Y. Shinyama, S. Sekine, K. Sudo, and R. Grishman. Automatic paraphrase acquisition from news articles. In *Proc. of the 2nd International Conference on Human Language Technology Research*, 2002.

[3] E. Charniak. A maximum-entropy-inspired parser. In *Proc. of the 1st NAACL*, 2000.

[4] R. Barzilay and K. McKeown. Extracting paraphrases from a parallel corpus. In *Proc. of the 39th ACL*, pp. 50–57, 2001.

[5] B. Pang, K. Knight, and D. Marcu. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proc. of HLT-NAACL 2003*, pp. 181–188, 2003.

[6] 下畑光夫, 渡辺太郎, 隅田英一郎, 松本裕治. パラレルコーパスからの機械翻訳向け同義表現抽出. 情報処理学会論文誌, Vol. 44, No. 11, pp. 2854–2863, 2003.

[7] C. D. Manning and H. Schütze, editors. *Foundations of Statistical Natural Language Processing*, pp. 265–314. MIT Press, 1999.

[8] 山本和英. テキストからの語彙的換言知識の獲得. 言語処理学会第 8 回年次大会, pp. 639–642, 2002.