統計翻訳手法を用いた類義語の自動抽出

揚石　亮平†　　三浦　孝夫†

† 法政大学大学院 工学研究科 電気工学専攻　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †ryohei.ageishi.08@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

あらまし　本研究では，類義語を自動的に抽出する方法について論じる．一般的に，類義語あるいは関連語を抽出することは容易なことではなく，人手によるコストを避けることは難しい．本研究では複数のコーパスから文対応を生成し，統計翻訳の手法により自動的に類義語を抽出する方法を提案する．実験により本手法の有用性を確認する．
キーワード　統計翻訳，類義語, Wordnet

Automatic Extraction of Synonyms Based On Statistical Machine Translation

Ryohei AGEISHI† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †ryohei.ageishi.08@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

Abstract　In this investigation we discuss how to extract synonyms automatically. Generically, it is not easy to extract synonyms or related words without hand-coding. However, we generate corresponded sentence pair from some newspaper corpuses, and propose how to extract synonyms automatically using by statistical machine translation. We show some experimental results to see the effectiveness.
Key words　Statistical Machine Translation，Synonym, Wordnet

1. ま え が き

近年，インターネットの普及により，誰もが容易に複数の言語による情報を得られるようになった．これらを容易に理解するために機械翻訳に注目が集まり，様々な研究が行われている．現在では大規模対訳コーパスが利用可能なことから統計機械翻訳 [1](Statistical Machine Translation，SMT) が著しい発展を遂げている．

SMT 手法のうち，Brown ら [2] による IBM Model では，英仏対訳コーパスから確率的な仏英単語辞書を語彙確率として学習する．仏文を英文に翻訳する際には，IBM Model は入力仏文に対して単語ごとにすべての語彙確率を計算し，最も確率が高い英文を出力する．

一方で類義語抽出に関する統計的研究は多く行われていない．類義語が単語の意味や概念を扱うものであり，人手によるコストを避けることが難しいからである．しかし背景知識を必要とする類義語は存在している．例えば以下のようなニュース記事の例の場合，

- **麻生太郎，総裁選に立つと表明**
- **麻生太郎氏が総裁選に出馬すると表明した**

この 2 文では**立つ**と**出馬する**の意味が類似する．特定の背景知識を仮定しないとき，例えば大辞林によれば，**立つ**と**出馬する**はそれぞれ**座ったり横になったりしていた人が足を伸ばして自分の体を垂直の姿勢にする**，**馬に乗って出かける**という意味を有し，同義語・類義語ではない．一方，これら 2 語はニュースという領域において類義語である．本研究では類義語を**一般的類義語** (背景知識がなく類義であるもの) と**領域依存型類義語** (背景知識があって類義であるもの) の 2 つに分類し，領域依存型類義語に着目して考える．領域依存型類義語は文書分類や情報検索において重要である．例えば，領域依存型類義語辞書を情報検索に用いると，与えられた検索語の類義語から領域に依存して情報を拡大することが可能である．

本研究の目的は，領域依存型類義語の自動抽出を行うことである．統計的手法はコーパスに依存した学習で有用であり，我々は，統計翻訳手法による領域依存型類義語の自動抽出法を提案する．統計翻訳手法を用いて特定領域のコーパスから語彙確率の学習を行い，ある単語に対して語彙確率の高い語をその語の類義語と見なす．本研究では固有名詞の重要性を考慮した複数コーパスからの文対応生成方法についても提案する．揚石ら [7] は確率過程モデルによる形態素解析結果を固有名詞に関して再推定を行うことで，形態素解析精度が改善されることを示した．固有名詞が重要な役割を果たしている例である．本稿

では提案手法が効果的であることを実験により検証する．

第 2 章で，関連研究について述べ，第 3 章では，ニュースコーパスの文対応生成方法を述べる．第 4 章では，統計翻訳手法を説明し，類義語の抽出方法を述べる．第 5 章では，実験結果を述べ，本手法の有効性を示す．

2. 関 連 研 究

類義語に関する研究としては，George ら [3] による Wordnet がある．Wordnet は英語の概念辞書であり，Synset 概念を導入し類義関係を語の組で表現する．各 Synset が 1 つの概念に対応する．Wordnet では各語が複数の Setset に属してよいため，語の上下関係を表現できる．辞書の構築は手作業で行われているため，大きな変更が起こってしまうと莫大な人手コストがかかるという問題がある．

Wordnet を日本語化したものとして，Kanzaki ら [4] による日本語 Wordnet がある．日本語 Wordnet は Wordnet と同様の構造を持つ．日本語 Wordnet の例として，検索語**立つ**を入力した結果を表 1 に示す．

01983264-v: 起きる, 立っち, 起立, 立上がる, 立つ, 起き上がる, 立上る
01848718-v: 去る, 退場, 離れさる, 出立, 発す, 走去る, 出発, 走りさる, 離れ去る, 発つ, 立ちさる, 立去る, 離去る, 立ち去る, 出立つ, 立つ, 消え失せる
02618688-v: 成り立つ, 発つ, 成立つ, 立つ
01546111-v: 立ち上がる, 起きあがる, 起上る, 起きる, 立っち, 起上がる, おっ立つ, 押っ立つ, 起立, 立上がる, 立つ, 起き上がる, 起き上る, 立上る, 起つ
02734488-v: 御座る, 御座ある, ござ在る, 位置, 御座在る, いらっしゃる, 居る, ござ有る, 御座有る, 在る, 立つ, 御座居る, ある, 有る

表 1　日本語 Wordnet の例

表 1 中の各数字は Synset の ID を表す．表 1 から，立つには 5 つの Synset ID がある: ”01983264-v”(立つ) , ”01848718-v”(出発する), ”02618688-v”(成る), ”01546111-v”(立ち上がる) , ”02734488-v”(ある)．Synset ID は一般的な類義語を反映しており，領域に依存した類義語は含んでいない．

類義語の抽出を領域依存で自動的に行った研究としては，中渡瀬 [9] の手法がある．中渡瀬の手法はコーパスから複合語を取り出し複合語の修飾関係などを考慮する．複合語を分割し，2 部グラフで表現することにより類義語を抽出する．例えば，**国際戦略**，**国際不況**，**世界不況**，**世界戦略**という 4 語がある場合，(**国際**，**世界**) と (**戦略**，**不況**) という 2 部グラフを構成する．このとき各グループ (**国際**，**世界**)，(**戦略**，**不況**) が類義語であると考える．中渡瀬の手法には複合語が不可欠であるため，本研究が考える単語ごとの類義語の抽出には適さない．

3. ニュース記事の文章対応

本章では 2 つのニュースコーパスから文の対応を得る方法について論じる．次章で述べる統計翻訳手法を用いるためには，対訳コーパス (Parallel Corpus) が必要である．対訳コーパスとは，同等の意味を持つ複数の言語で書かれた文対集合のことをいう．本研究では日本語から日本語への翻訳モデルを学習し，得られた語彙確率を類似度として用いるため，日本語と日本語の対訳コーパスが必要である．しかし大規模な日本語間対訳コーパスは存在しないため，本研究では自動的な文対応生成方法を提案する．

3.1 記事の対応生成

本節では記事の対応生成方法を述べる．本研究において記事とは，見出し，本文，日付からなるものとする．本研究で用いた朝日新聞と読売新聞の例を表 2，表 3 に示す．

＼ＡＦ＼	２００７０９０１	２００７０９０１
＼Ｔ１＼	麻生太郎氏が総裁選に 出馬を表明	「最強」法大か，粘りのＯＳか アメフット・ライスボウル
＼Ｔ２＼	麻生太郎氏が総裁選に 出馬すると発表した． その足で千葉駅に向かい， 麻生氏は演説を行った．…	アメリカンフットボール日本一を 決める第２４回日本選手権は …

表 2　朝 日 新 聞

＼ＹＦ＼	２００７０９０１	２００７０９０１
＼Ｔ１＼	千葉でひき逃げ事件の 容疑者を逮捕	麻生氏総裁選に立つ
＼Ｔ２＼	３１日未明に発生した ひき逃げ事件の容疑者， として千葉県警は３８歳 男性を逮捕した． …	麻生太郎氏は総裁選に 立候補することを発表した． それに伴い，麻生氏は 千葉駅前で熱弁を振った． 総裁候補が千葉で演説を 行うのは初めてのこと．…

表 3　読 売 新 聞

表中の＼ＡＦ＼，＼ＹＦ＼が日付を，＼Ｔ１＼が見出しを，＼Ｔ２＼が本文を表す．本研究では，対応する記事は同一の日付であり，見出しの内容が類似すると考える．本研究では同一の日付の見出しを取り出し，見出しの比較を行うことで記事の対応を得る方法を用いる．

見出しの比較を行う方法について述べる．表 2，表 3 から見出しを抽出しそれぞれ形態素解析を行うと，

(A)　麻生 (固有名詞) 太郎 (固有名詞) 氏 (接尾) が (助詞) 総裁選 (名詞) に (助詞) 出馬 (名詞) を (助詞) 表明 (名詞)

(B) 「(記号) 最強 (名詞) 」(記号) 法大 (固有名詞) か (助詞) , (記号) 粘り (名詞) の (助詞) ＯＳ (固有名詞) か (助詞) アメフット・ライスボウル (固有名詞)

(C)　千葉 (固有名詞) で (助詞) ひき逃げ (名詞) 事件 (名詞) の (助詞) 容疑者 (名詞) を (助詞) 逮捕 (名詞)

(D)　麻生 (固有名詞) 氏 (接尾) 総裁選 (名詞) に (助詞) 立つ (動詞)

となる．形態素解析結果から見出しには助詞を除くと名詞，固有名詞が多く使われることが分かる．したがって本研究では名詞，固有名詞の一致度によって見出しの比較を行う．一致度 M を以下の式で定義する．

$$M_1 = \frac{\sum_i \sum_j Match(w_i^F, w_j^E)}{max(\sum_i C(w_i^F), \sum_j C(w_j^E))} \quad (1)$$

(1) 式において，w^F は朝日新聞中の名詞または固有名詞，w^E は読売新聞中の名詞または固有名詞を表す．また，Match(w_i^F, w_j^E) は一致した名詞または固有名詞の頻度を表す．$C(w_i)$ は生起頻度であり，語 (w_i) が該当文書で出現する頻度を表す．したがって (1) 式は，名詞または固有名詞の相対一致度を表す．一致度 M を用いて (A) と (C)，(A) と (D)，(B) と (C)，(B) と (D) の比較を行うと，(A) と (C) で M=0/6=0.0，(A) と (D) で M=2/6=0.33，(B) と (C) で M=0/5=0.0，(B) と (D) で M=0/5=0.0 となる．よって，見出し (A) を含む記事と見出し (D) を含む記事が対応すると考える．本研究では (1) 式を用いて，すべての同一の日付の見出しの比較を行い，閾値を用いることで記事の対応を得る．

3.2 対応記事中の本文対応生成

本節では，得られた記事の対応からの本文対応生成方法について述べる．前節に示した見出し対応のうち，(A) と (D) の記事対応が存在すると仮定する．

表 2 の見出し (A) の記事本文
(a1)　麻生太郎氏が総裁選に出馬すると発表した
(a2)　その足で千葉駅に向かい，麻生氏は演説を行った
と表 3 の見出し (D) の記事本文
(d1)　麻生太郎氏は総裁選に立候補することを発表した
(d2)　それに伴い，麻生氏は千葉駅前で熱弁を振った
(d3)　総裁候補が千葉で演説を行うのは初めてのこと
の対応を考える．(1) 式を用いて対応する可能性のある文は，(a1) と (d1)(M=0.75)，(a2) と (d2)(M=0.4)，(a2) と (d3)(M=0.4)) である．(1) 式を用いると (a2) に対して (d2) と (d3) の 2 つの候補が存在するが，(a2) が実際に対応する文は (d2) である．(a2) の対応文が (d2) であることは，**演説**という名詞よりも**千葉**や**麻生**という固有名詞が一致することから判断可能である．さらに**麻生**というより重要な語が一致する点からも判断可能である．**麻生**という語が**千葉**という語よりも重要であるということは，表 3 から**麻生**は 2 記事中 1 回，**千葉**は 2 記事中 2 回出現することから分かる．本研究では，固有名詞が一致することに重みをつけ，固有名詞ごとに重みを変化させるために，TF*IDF 法(注1) [8] の IDF 値による重みを用いる．本研究では，固有名詞 w_i の IDF_{w_i} を以下の式のように定義する．

$$IDF_{w_i} = log\frac{D}{\{d : d \ni w_i\}} \quad (2)$$

(2) 式において，D は記事数であり，$\{d : d \ni w_i\}$ は固有名詞 w_i を含む記事数である．本研究では，w_i の朝日新聞，読売新聞での IDF 値をそれぞれ $IDF_{w_i}^F$，$IDF_{w_i}^E$ とする．固有名詞 w_i の IDF 値を以下の式を用いて計算する．

$$IDF_{w_i} = \alpha IDF_{w_i}^F + (1-\alpha) IDF_{w_i}^E \quad (3)$$

α は比例定数である．例えば，(3) 式において α=0.5 とし，表 2，表 3 中の固有名詞，**麻生**と**千葉**の IDF 値を計算すると，IDF(麻生)=0.5log2/1+0.5log2/1=log2，IDF(千葉)=0.5log2/1+0.5log2/2=0.5log2 となり，**麻生**という固有名詞の重みが**千葉**という固有名詞の重みより重くなる．(3) 式を用いて一致度 M を以下のように再定義する．ただし固有名詞の IDF 値は (3) 式を用い，名詞の IDF 値は 1 (注2) として計算する．

$$M_2 = \frac{\sum_i \sum_j IDF_{w_i^F = w_j^E} \times Match(w_i^F, w_j^E)}{max(\sum_i IDF_i^F \times C(w_i^F), \sum_j IDF_j^E \times C(w_j^E))} \quad (4)$$

本研究では (4) 式を用いて，すべての対応生成した記事の比較を行い，閾値を用いて対応記事中の本文の対応を得る．

4. 類義語の抽出

本章では，前章で得られた文対応から統計翻訳手法を用いて類義語を抽出する方法を述べる．

4.1 統計翻訳手法による類義語抽出

本節では，本研究で用いた統計翻訳モデル [1] について述べる．統計翻訳では，ある起点言語 *F(Source Language)* からある目標言語 *E(Target Language)* への翻訳を，以下の式のように最尤翻訳文 $\hat{e}$ を求める問題に置き換える．

$$\hat{e} = \arg\max_e P(e)P(f|e) \quad (5)$$

実際に翻訳を行う際には，(5) 式において，$P(e)$ と $P(f|e)$ を求めることにより，翻訳文を生成する．ここで，$P(e)$ を言語モデル (*Language Model*)，$P(f|e)$ を翻訳モデル (*Translation Model*) とよぶ．本稿では，起点言語と目標言語に同一言語を用いるが，翻訳モデルによって領域依存型類義語抽出のための仕組みを構築する．

翻訳モデルの代表的なものとして，Brown ら [2] による IBM Model がある．IBM Model には 5 つの定義 (モデル) が存在するが，本研究では最も構造が単純な IBM Model1 を用いる．

IBM Model1 では，翻訳モデル $P(f|e)$ を計算するため，起点言語と目標言語の単語の対応位置関係をアラインメントとして定義する．起点言語を日本語，目標言語を英語とした図 1 において各単語を繋いでいる線がアラインメントを表現している．

アラインメント a を用いて $P(f|e)$ を以下のように定義する．

$$P(f|e) = \sum_a P(f, a|e) \quad (6)$$

(注1)：TF*IDF 法とは，情報検索などに用いられる文章中の重要語を抽出するための重み付け手法である．TF は文章中の単語の出現頻度を表し，IDF は一般語 (多くのドキュメントに出現する語) のフィルタとしての役割を果たす．

(注2)：実験から最も IDF 値の低い語が”東京”の 1.05 であったため，名詞の重みを 1 とした

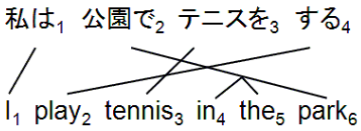

図 1　日本語-英語のアラインメント例

$$P(f,a|e) = \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m} P(f_j|e_{a_j}) \quad (7)$$

ϵ は比例定数であり，m は起点言語文の長さ，l は目標言語の長さ，f_j は起点言語文で j 番目に位置する単語，e_i は目標言語文で i 番目に位置する単語を表す．アラインメント a_j は起点言語文で j 番目に位置する単語が目標言語文で対応する単語の位置を表す．例えば，図 1 において，**する**と *play* が対応するため，$a_4 = 2$ となる．

(7) 式において，求めるべきパラメタは $P(f_j|e_{a_j})$ であり，$P(f_j|e_{a_j})$ が単語 e_{a_j} が単語 f_j に翻訳される語彙確率を表す．パラメタ $P(f_j|e_{a_j})$ の推定は，対訳コーパスを用いて，ラグランジュの未定乗数法により求める．詳細は [2] を参照されたい．

4.2　類似度に対する重み付け

本研究では，朝日新聞と読売新聞を用いた IBM model1 をベースラインとし，領域依存型類語語の抽出を行う．対応生成した 2 つのニュース記事文の片方を起点言語，もう一方を目標言語として語彙確率を得る．ここで，語彙確率 $P(f_j|e_{a_j})$ を単語 e_i に対する単語 f_j の類似度と考える．例えば，(a1) と (d1) を用いて動詞の類似度を計算すると以下のようになる．

P(発表する | 立候補する) = 0.5
P(出馬する | 立候補する) = 0.5

IBM Model1 では語彙確率を対訳コーパスの起点言語・目標言語のアラインメント頻度から求めるため，出現しやすい単語の語彙確率が高くなりがちである．この問題を回避するため出現しやすい単語の重みとして IDF 値を用いる．前章で，単語 w_i の IDF 値を (3) 式のように定めたが，本章では，語彙確率 $P(f_j|e_{a_j})$ の重みとして用いるため，起点言語 f_j の IDF 値が重要となる．単語 f_j の IDF 値である $IDF^F_{f_j}$ を重みとして用いた類似度 $Sim_{e_i}(f_j)$ を以下のように定義する．

$$Sim_{e_i}(f_j) = P(f_j|e_{a_j}) \times IDF^F_{f_j} \quad (8)$$

また，本研究では共起性に関する重みも用いる．先ほどと同様に，IBM Model1 は語彙確率をアラインメント頻度から求めるため，共起しやすい単語の語彙確率が高くなりがちである．例えば，4.1 節で用いた t(発表する | 立候補する) = 0.5, t(出馬する | 立候補する) = 0.5 の場合を考える．(d1) 中の語の共起頻度を求めると，**立候補する**と**発表する**は共起している (共起頻度=1) が，(d1) 中に**出馬する**は存在しないため**立候補する**と**出馬する**は共起していない (共起頻度=0)．語彙確率の重みとして共起逆頻度を用いると，t(発表する | 立候補する) = $0.5 \times 0 = 0$，t(出馬する | 立候補する) = $0.5 \times 1 = 0.5$ となり (d1) 中の語**立候補する**の類義語を**出馬する**にできる．目標言語中の単語 e_i と単語 $e_{i'}$ の共起頻度 $k(e_{i'}|e_i)$ を条件付き確率によって以下のように求める．

$$k(e_{i'}|e_i) = \frac{C(e_{i'}, e_i)}{C(e_i)} \quad (9)$$

共起頻度を重みとして使用するため，対数逆頻度を用いて，類似度 $Sim_{e_i}(f_j)$ を以下のように定義する．

$$Sim'_{e_i}(f_j) = P(f_j|e_{a_j}) \times IDF^F_{f_j} \times (-logk(f_j|e_i)) \quad (10)$$

実験では，IBM Model1 をベースラインとし，(8) 式，(10) 式と用いる重みを変更して類似度を決定し，類義語の抽出を行う．

5.　実　　験

本研究では，2 つの実験を行う．実験 1 では，ニュース記事から本文対応生成を行い，実験 2 では，得られた文対応を用いて類義語の抽出を行う．本研究では類義語の抽出を行う単語の品詞を動詞に限定する．本研究の手法では記事の対応生成を行う際に，名詞または固有名詞の一致度を用いて対応生成を行うため，名詞と固有名詞に関しては類義語を得ることが困難であることに注目したい．

5.1　準　　備

本研究では，新聞コーパスとして朝日新聞 2007 年版と読売新聞 2007 年版を用いる．記事数として，朝日新聞には 153246 件，読売新聞には 343142 件の記事を用いる．月毎の内訳を表 4 に示す．また，抽出した類義語の評価のため，手作業で対応付けした 350 行をテストデータとして用いる．

月	朝日新聞記事数 (件)	読売新聞記事数 (件)
1 月	11989	26795
2 月	12038	26567
3 月	13483	29647
4 月	12823	28437
5 月	13471	29331
6 月	13563	30092
7 月	12009	27912
8 月	11973	29294
9 月	12809	28814
10 月	13446	30200
11 月	13169	28569
12 月	12473	27484
合計	153246	343142

表 4　朝日新聞，読売新聞の月別記事数

日本語形態素解析ツールとして MeCab [5] を用いる．IBM Model1 を学習する際には，本研究では起点言語を朝日新聞，目標言語を読売新聞とし，繰り返し回数を 5 回とする．類似度の重みに用いる IDF 値と共起頻度は，得られた文対応から動詞のみを取り出し計算する．

5.2 評価方法

本研究では，統計翻訳の評価関数である Word Error Rate (WER) [6] と，Position independent word Error Rate(PER) を用いる．

WER は語順を考慮した不一致率であり，参照訳との編集距離によって計算する．

$$WER = \frac{\sum_i (\text{挿入語数 } i + \text{削除語数 } i + \text{置換語数 } i)}{\sum_i \text{参照訳 } i \text{ の語数}} \quad (11)$$

PER は語順を無視し，文を単語集合とした不一致率である．PER を用いることで，WER では翻訳が正しくても正解と語順が著しく異なる場合に結果が悪くなってしまうため，語順を考慮せず，語のレベルでその翻訳が正しいかを判断することができる．

$$PER = 1 - \frac{\sum_i (\text{翻訳文 } i \text{・参照訳 } i \text{ 間の一致語数})}{\sum_i \text{参照訳 } i \text{ の語数}} \quad (12)$$

ともに誤り率であるので，数値が低いほど良い結果となる．

また，見出しの対応が記事の対応であるかを評価するため，M_1 を用いて対応付けを行った見出しの対応からランダムで 100 件を抽出し，記事対応数により評価を行う．記事が対応しているかどうかの判断は手作業により行う．記事対応数が多ければ，同一日時の見出しの対応は記事の対応といえる．

最後に，抽出した類義語の評価法を定義する．本研究では，手作業で対応付けしたテストデータに対して抽出した類義語を用いて拡張を行い，翻訳精度である PER, WER の改善率により評価を行う．拡張は単語 A に対し { 単語 A, 類義語 B} の形で行う．類義語は類似度最上位のものを使用する．類義語を用いることにより翻訳精度を改善できれば，得られた類義語は正しいといえる．

5.3 実験結果

5.3.1 実　験　1

ニュース記事の文対応生成に関する結果を示す．始めに，記事の見出し対応を一致度 M_1 を用いて閾値を変化させた場合の記事数を表 5 に示す．また対応記事中の朝日新聞，読売新聞の本文数を表 6 に示す．

どの閾値においても 4 月の対応件数が最も多くなっている．また，一記事あたりの平均本文数は，朝日新聞で 49.1 文，読売新聞で 39.7 文であった．

次に，PER, WER を表 7 に示す．本来，記事同士が対応するかどうかはその見出しの要旨が同じどうかで判断すべきであるが，定量的な実験が困難なため，本研究では朝日新聞を翻訳文，読売新聞を参照訳として WER と PER を用いて評価を行う．

最も PER の値が良いのは $M_1 = 0.8$ としたときであり，PER の値は 0.40 である．最も WER の値が良いのは $M_1 = 1.0$ としたときであり，WER の値は 0.74 である．

また，$M_1 = 0.8$ で対応付けを行った見出しから 100 件をランダムに抽出し，手作業により見出しの対応，記事の対応を評価した結果を表 8 に示す．

M_1	0.5	0.6	0.7	0.8	0.9	1.0
1 月	5538	2021	791	446	158	137
2 月	7097	3291	1029	519	185	162
3 月	8656	3444	1289	733	283	245
4 月	34344	16170	5859	3002	1209	922
5 月	7299	3028	1194	768	327	295
6 月	5878	2534	1027	603	256	221
7 月	17041	5143	1667	843	316	251
8 月	5977	2345	852	475	189	166
9 月	8957	3806	1367	723	287	240
10 月	9029	4835	1731	948	373	332
11 月	8537	63594	1099	631	205	185
12 月	4895	2196	899	560	223	207
合計	123248	52407	18804	10251	4011	3363

表 5　実験結果:ニュース記事見出し月別対応記事数 (件)

閾値	朝日新聞本文数 (文)	読売新聞本文数 (文)
0.5	7160307	5963342
0.6	3145875	2507245
0.7	1037026	833717
0.8	460683	375253
0.9	161648	132916
1.0	120703	92837

表 6　実験結果:対応記事中の本文数

閾値	PER	WER
0.5	0.57	0.94
0.6	0.50	0.87
0.7	0.43	0.80
0.8	0.40	0.74
0.9	0.41	0.70
1.0	0.42	0.67

表 7　実験結果:ニュース記事見出しの PER，WER

	記事対応	記事非対応
$M_1 = 0.8$	83	17

表 8　実験結果:$M_1 = 0.8$ の見出し 100 件中における記事の対応

よって記事対応数=83 である．

次に，$M_1 = 0.8$ を用いて見出しの対応生成を行った記事から，本文対応生成を一致度 M_2 を用いて閾値を変化させた場合の本文数，単語数，動詞数を表 9 に，PER，WER を表 10 に示す．

最も PER の値が良いのは $M_2 = 0.5$ としたときであり，PER の値は 0.34 である．最も WER の値が良いのは $M_2 = 0.9$ としたときであり，WER の値は 0.82 である．

5.3.2 実　験　2

類似度 $P(e_i|f_j)$, $Sim_{e_i}(f_j)$, $Sim'_{e_i}(f_j)$ を用いて類義語の抽出を行った結果を示す．記事見出しの一致度を $M_1 = 0.8$，記事本文の一致度を $M_2 = 0.9$ として得られた記事本文対応を用いる．この場合，類似関係を抽出できた動詞は 2045 種類あった．このうち頻度順に上位 30 件を表 11 に示す．

M_2	0.5	0.7	0.9
対応本文数 (文)	25485	13539	7629
対応本文中の朝日新聞単語数 (延べ語)	2307423	1279467	743125
対応本文中の朝日新聞単語数 (語)	23245	18744	13776
対応本文中の朝日新聞動詞数 (延べ語)	137216	71706	39621
対応本文中の朝日新聞動詞数 (語)	3387	2686	2005
対応本文中の読売新聞単語数 (延べ語)	2835007	1623426	987459
対応本文中の読売新聞単語数 (語)	24584	19900	15363
対応本文中の読売新聞動詞数 (延べ語)	151376	80459	45739
対応本文中の読売新聞動詞数 (語)	3534	2799	2045

表 9　実験結果:ニュース記事本文対応数，単語数

閾値	PER	WER
0.5	0.34	0.86
0.7	0.35	0.84
0.9	0.36	0.82

表 10　実験結果:ニュース記事本文の PER，WER

順位	動詞	頻度	順位	動詞	頻度
1	なる	4034	16	決める	383
2	告示する	1600	17	比べる	380
3	立候補する	1230	18	発表する	365
4	行う	1229	19	みる	329
5	する	1048	20	伴う	327
6	ある	760	21	決まる	317
7	減る	637	22	受ける	304
8	目指す	473	23	見る	281
9	投開票する	447	24	続く	269
10	よる	434	25	公開する	269
11	予定する	424	26	巡る	264
12	増える	412	27	当選する	257
13	上回る	408	28	表明する	255
14	届け出る	399	29	できる	250
15	除く	385	30	争う	248

表 11　実験結果:類義関係抽出可能動詞 (頻度順)

抽出した類義語の評価のため，テストデータ 350 行中の朝日新聞の動詞 1748 語，586 種類に対し，最も類似度が高い類義語で拡張を行い，読売新聞との比較を行う．類義語を用いる前後での PER，WER の値を表 12 に示す．

	類義語による拡張前	類義語による拡張後
PER	0.69	0.62
WER	0.92	0.90

表 12　実験結果:類義語を用いる前後での PER，WER の比較

PER の値で 7%の改善が見られる．

また抽出した類義語の例として，**立つ** (頻度 50 位) と**打つ** (頻度 92 位) に関する類義語の抽出を行った結果の類似度が高い順に並べた上位 10 件を表 13，表 14 に示す．

表 13，表 14 から提案手法は**立つ**という語に対して，**立候補する**，**出馬する**という類義語を，**打つ**という語に対して，**はねる**，**衝突する**という類義語を抽出可能である．

順位	Model1	Sim	+IDF	Sim	+共起頻度	Sim
1	なる	0.086	立つ	0.086	立候補する	0.088
2	立候補する	0.081	立候補する	0.068	立つ	0.086
3	立つ	0.049	推薦する	0.051	巡る	0.080
4	告示する	0.032	締め切る	0.040	出馬する	0.076
5	推薦する	0.031	巡る	0.039	届け出る	0.069
6	巡る	0.026	なる	0.035	進める	0.068
7	決まる	0.024	出馬する	0.034	出席する	0.056
8	出馬する	0.022	決まる	0.034	スタートする	0.056
9	決める	0.022	進める	0.031	推薦する	0.051
10	届け出る	0.021	スタートする	0.030	向ける	0.050

表 13　実験結果:類義語の抽出 (立つ)

順位	Model1	Sim	+IDF	Sim	+共起頻度	Sim
1	打つ	0.354	打つ	0.721	打つ	0.721
2	死亡する	0.127	来る	0.243	来る	0.288
3	来る	0.108	死亡する	0.234	運ぶ	0.221
4	調べる	0.104	調べる	0.181	調べる	0.210
5	運ぶ	0.063	運ぶ	0.153	死亡する	0.184
6	はねる	0.048	はねる	0.102	はねる	0.097
7	する	0.029	衝突する	0.062	衝突する	0.078
8	衝突する	0.024	乗る	0.038	乗る	0.064
9	よる	0.019	右折する	0.031	右折する	0.063
10	乗る	0.016	渡る	0.030	渡る	0.050

表 14　実験結果:類義語の抽出 (打つ)

5.4 考　　察

実験 1 に関する考察を行う．記事見出しの対応生成では，閾値を $M_1 = 0.8$ としたときに最も PER の値が良く，0.40 である．このとき，WER の値は 0.74 である．見出しは文の形をなしていないものが多いため，PER の値が良いことが望ましい．したがって閾値 0.8 では，6 割の同じ単語を含むが，文としては語順が異なる見出しが対応可能である．なおかつ多くの同じ固有名詞を含むので，同じ趣旨の見出しが対応可能であると考える．例えば，以下のような見出しの対応が得られた．

- 朝日新聞:“「信頼し合って暮らす社会に」天皇陛下が感想 ”
- 読売新聞:“ 天皇陛下が新年の感想「皆が信頼して暮らせる社会を」”

上記の例のように，同じ単語を多く含むが，語順が異なる見出しの対応生成できる．また，対応記事数としては閾値によらず 4 月が最も多かった．これは 4 月に統一地方選挙が行われたため，通常 1 対 1 で記事が対応するはずであるが，複数の地域で同様な見出しが使われ，多対多で対応してしまったためであると考えられる．実際，以下に示すような見出しを対応可能と判定してしまっている．

- 町村議選の候補者　統一地方選　／山梨県
- 町村議選の候補者　統一地方選　／青森県
- 町村議選の候補者　統一地方選　／千葉県
- 町村議選の候補者　無投票当選者　統一地方選　／長野県

これらの見出しは，新聞記事の地域欄によるものであると考え

られる．また，100 件中の記事対応数は 83 件であった．よって同一日時の見出しの対応から記事の対応付けが可能であるといえる．

得られた記事対応からの本文の対応生成では，閾値を $M_2 = 0.9$ としたときに最も WER の値が良く，0.82 である．見出しの対応とは異なり，本文の対応であるため，文として類似することが望ましいので，WER の値が良いことが望ましい．例えば，以下のような本文の対応が得られている．

- 朝日新聞:“ 第８５回全国高校サッカー選手権は２日，２回戦で県代表・秋田商が神村学園（鹿児島）と対戦し，ＰＫ戦で惜敗した 。 ”
- 読売新聞:“ 全国高校サッカー選手権大会（読売新聞社など後援）は２日，２回戦１６試合が行われ，県代表の秋田商は，横浜市の三ツ沢球技場で神村学園（鹿児島）と対戦し，０－０の同点で迎えたＰＫ戦で惜しくも敗れ，初戦突破はならなかった． ”

上記の例のように，語順が類似するが，動詞が異なる文の対応生成が可能である．また，文対応の PER の値は，見出し対応の PER の値より良いが，これは PER が単語集合としての類似度を示す尺度であり，見出しに比べ本文は長いため，より多くの単語を含んでいるからであると考える．それに伴い，文対応の WER の値は，見出し対応の WER の値より悪化している．

実験 2 の考察を行う．表 13，表 14 から分かるように，**なる**や**する**といった出現しやすい語の類似度を下げることが可能である．これらの単語は日本語に固有の表現で，動詞としての働きではなく動詞を補助的な意味を付与する働きとして用いられることが多い．例えば，**なる**という単語は，$M_1 = 0.8$, $M_2 = 0.9$ として得られた文対応 7629 文中 2965 文に含まれる．したがって，その IDF 値は 0.41 となるので，類似度を下げることが可能である．

また，**打つ**という語に比べ，**立つ**は上位に類義語が抽出可能である．これは 2007 年に参議院議員選挙や統一地方選挙があり，選挙に関連する記事が多く出現していたからであると考える．実際に，**立つ**は得られた文対応 7629 文中の読売新聞に 140 文，**打つ**は 7629 文中に 70 文に出現しており，**立つ**を含む文の方が多く見られた．

最後に，類義語を適用する前後での PER，WER に関する実験についての考察を行う．類義語を用いることで，PER が 7%改善している．例えば，

- 朝日新聞:“ 市は滞納額の上位１００人のうち，支払いの意思表示がない７４人に通知書を送付した。”
- 読売新聞:“ 市保育運営課によると，9月と１０月，滞納額の多い上位１００人のうち，再三の 督促にもかかわらず分割納付に応じず，納付する意思を示さない７４人に財産 差し押さえの事前通知書を郵送した． ”

という文対に対し，朝日新聞の動詞に類義語を用いて拡張すると，

- 朝日新聞:“ 市は滞納額の上位１００人のうち，支払いの意思表示がない７４人に通知書を { 送付した, 郵送した }. ”

となり，単語対応が取れるようになるため翻訳精度が上昇している．このことから類義語を用いることにより朝日新聞の動詞を読売新聞の動詞に書き換えることが可能であるといえる．したがって，領域に依存した類義語の抽出が可能であると考える．

6. 結　　論

本研究では，類義語を自動的に抽出する方法として，統計翻訳モデルを用いた方法を提案した．その際に，複数のニュースコーパスの対応生成を行う方法を述べ，文対応の生成も自動的に行った．得られた文対応から統計翻訳モデルを用いて動詞の類義語自動抽出を行った．実験により得られた類義語を対訳文に用いることで 7 %の精度改善が見られたことから提案手法が領域に依存した類義語を抽出可能であることを確認した．

文　　献

[1] Adam Lopez, Statistical Machine Translation, ACM Computing Surveys, Vol. 40, No. 3, Article 8, Publication date: August 2008.

[2] Brown, P. F., Pierta, S. A. D., Pierta, V. J. D., and Mercer, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. Comput. Linguist. 19, 2, 263.311

[3] George A. Miller. 1995. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[4] Kyoko Kanzaki, Francis Bond, Noriko Tomuro and Hitoshi Isahara. 2008. Extraction of Attribute Concepts from Japanese Adjectives. In LREC-2008, Marrakech.

[5] MeCab: http://Mecab.sourceforge.net/

[6] OCH, F. J., TILLMAN, C., AND NEY, H. 1999. Improved alignment models for statistical machine translation. In Proceedings of EMNLP-VLC. 20.28.

[7] Ryohei Ageishi, Takao Miura, 2008, Named Entity Recognition Based On A Hidden Markov Model in Part-Of-Speech Tagging, ICADIWT 2008.

[8] 北 研二，津田 和彦，獅々堀 正幹，情報検索アルゴリズム，2002，共立出版

[9] 中渡瀬 秀一，複合語からの類義語抽出法，2002，情報処理学会研究報告 pp.39-46