# 同義語の類似度に関する考察

寺田昭 † 吉田稔 ‡ 中川裕志 ‡

†(株) 日本航空インターナショナル
‡ 東京大学情報基盤センター

**あらまし** 同義語の同定は，情報検索，テキストマイニングなどの処理を行う上で必要な作業である．本論文では，同義語の類似度をその周辺に出現する語により計算する手法について，詳細な実験を行った結果を報告する．対象としたコーパスは，航空関係の日本語のレポートで，同義語の候補として日本語の漢字・ひらがな，カタカナ，アルファベット及びそれらの略語が含まれている．提案手法は，クエリに対して，コーパスに出現する語の相互の類似度を計算し，類似度の高い同義語候補語をユーザに提示するものである．文脈語，同義語の種類・頻度が類似度の精度に対してどのような影響があるかを調査した．

## 1 はじめに

航空分野のマニュアル，補足情報，業務報告書等に使用される名詞は，漢字/ひらがなだけでなく，カタカナ，アルファベット及びそれらの略語が使用される．例えば，「滑走路」を略語を使用して「RWY」,「R/W」と表現している．このようなテキストを計算機で処理する場合には，同義語辞書が必要であるが，これらの語句は汎用の辞書に載っていない場合が多い．さらに，語句の使用は，統制されているものではなく，また，常に新しい語が使用されるので，一度，分野の辞書を作成しても，それを定期的にメンテナンスする必要がある．これを人手だけで行うのは大変な作業である．

我々は，同義語の類似度をその周辺に出現する語の文脈情報により計算することにより同義語辞書を半自動的に作成するツールを開発している [2]．本論文では，文脈語，同義語の種類・頻度が同義語の類似度の精度に対してどのような影響があるかを調査した．

## 2 定義

本節では，同義語候補語，文脈情報及び類似度の計算アルゴリズムについて説明する．なお，基本的には，単名詞を対象としているが，複合名詞についても，アルゴリズムは同一であり，複合名詞の選択は，専門用語抽出システム [3] が抽出したものを使用した．

### 2.1 同義語候補語

同義語候補語として形態素解析器が出力したアルファベット，名詞，カタカナを対象とした．形態素解析器としては，茶筅[1]を使用し，その中で出現頻度が，100 以上のものを使用した．

[1] http://chasen.naist.jp/hiki/ChaSen/

### 2.2 文脈情報

「同義語は，同じような文脈で使用される」という仮定から，文脈情報により語の類似度を計算する．クエリを q とし，$x_\alpha ... x_2 x_1 q y_1 y_2 ... y_\beta$ をその前後の語の並びとする．前後の語は，形態素解析器が出力した語とする．対象とするクエリ $q$ の文脈語をクエリの前で $x_\alpha ... x_1$，クエリの後ろで $y_1 ... y_\beta$ とすると，window 幅は α，β で，window[α, β] と表現することとする．同義語候補語の window 幅についても，同様の表現とする．window 幅は，同義語，同義語候補語を含む 1 文の範囲内だけを考慮した．文脈語としては，名詞，アルファベット，カタカナ，動詞，形容詞を使用した．文脈情報は，それぞれの同義語，同義語候補語について，window 幅内に出現するコーパス内全ての文脈語の頻度ベクトルを log で補正したものとした．すなわち，文脈語の頻度ベクトルの各要素を $w_i$ とすると，文脈情報は，$\log(w_i + 1)$ になる．

### 2.3 類似度

クエリ (query) の文脈情報を $\boldsymbol{c_q}$，同義語候補語 (synonym) の文脈情報を $\boldsymbol{c_s}$ とする．$\boldsymbol{c_q}$ と $\boldsymbol{c_s}$ をベクトル空間モデルで表し，その類似度をベクトルの余弦で計算した．すなわち，クエリと同義語候補語の類似度 ($sim$) は，次式で計算される．

$$sim(query, synonym) = \frac{\mathbf{c_q} \cdot \mathbf{c_s}}{|\mathbf{c_q}| \cdot |\mathbf{c_s}|} \qquad (1)$$

### 2.4 平均精度

情報検索の性能評価として精度と再現率がよく用いられるが，これらは，与えられたクエリに対する検索結果全体に対する性能を表すものである．同義語の辞書作成には，検索結果の順位が重要である．つまり，検索結果

表 1: window 幅による平均精度 (%) の比較

| FWD \AFT | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | - | 21.4 | 37.8 | 37.1 | 33.0 |
| 1 | 25.6 | 31.5 | 40.7 | 38.7 | 35.3 |
| 2 | 35.5 | 36.9 | 43.1 | 40.1 | 36.8 |
| 3 | 34.8 | 37.8 | 40.9 | 39.2 | 36.1 |
| 4 | 32.2 | 34.6 | 37.4 | 36.3 | 33.3 |

の中で正解のものが上位にあるほど評価値は，高い必要がある．したがって，このような評価尺度を表すものとして平均精度 (average precision) を用いた [1].

### 2.5 コーパス

コーパスとして，航空分野でのパイロットレポートを使用した．このレポートからは，事前に名前等の個人情報は削除し，個人を特定できないようにしてある．レポートの内容には，出発地・到着地などの定型情報とテキストで自由に記述された表題，本文が含まれている．本論文では，1,992 年から 2,003 年までの本文を対象とした結果，6,427 件のレポートが含まれ，そのサイズは，約 6.9M バイトであり，同義語候補語の数は，1,343 であった．同義語抽出のタスクは，クエリの同義語をこれらの同義語候補語の中から選択するものである．

### 2.6 評価用辞書

今回の実験評価のために，同義語辞書を人手で作成した．その結果，辞書の登録数は 406，同義語数は 777 で，平均同義語数は 1.91 であった．

## 3 文脈語の違いによる精度比較

これまでの研究で文脈語は，頻度の大きなものと小さなものを除くと平均精度が良くなることが分かっている [2]．平均精度が高かった頻度として，最小頻度を 50，最大頻度を 600 を用いる．

### 3.1 window 幅による比較

window 幅について，window 幅を同義語候補語の前 (FWD) に 0〜4 語，後 (AFT) に 0〜4 語変化させて実験したところ，表 1 の結果が得られている．平均精度は，window[2,2] が 43.1%で最も高かった．

### 3.2 window 幅による文脈語の重み付けの変更

3.1 節では，window 幅の中で文脈語の位置にかかわらず，文脈語の重み付けは同一であったが，window 幅の中で同義語・同義語候補語に近いほど類似度に対する文脈語の寄与度が大きいと仮定し，window 幅の中での文脈語の重み付けを変化させて実験を行った．window 幅 α の場合の同義語候補語から j 番目の位置の文脈語の重み付けを $\frac{\alpha-j+1}{\alpha}$ 倍し，$\frac{\alpha-j+1}{\alpha}\log(w_i+1)$ とした．window[3,3] で実験した結果，平均精度は 31.4%で，補正をしない 39.2%より低い結果となった．window 幅の中で同義語・同義語候補語に近いほど文脈語の影響が大きいと仮定し，そのような重み付けを考慮したが，平均精度は低下した．

window 幅 [2,2] の平均精度が最も高かったので，window 幅 [2,2] までは，文脈語の重み付けを変化させないで，それ以降の window 幅での文脈語の重み付けを $1/2^{(j-2)}$ 倍したもので実験を行った．window[4,4] でテストした結果，平均精度は 21.1%で，補正をしない 33.3%よりかなり低い結果となった．

### 3.3 高頻度の文脈語の選択

これまで文脈語は，文脈語全体での頻度が最小頻度 50，最大頻度 600 という制限はあるもののその範囲のものを全て取得している．しかしながら，同義語を判別するには個々の同義語に対しては，それ程多くない文脈語が関係しているのではないかとの仮定のもと，それぞれの同義語候補語に対して，頻度が大きい方から N 番目までの文脈語のみを取得して window 幅 [2,2] で実験を行った．その結果，平均精度は，N=10 で 30.8%，N=20 で 36.4%，N=30 で 40.1%となり，個々の同義語に対して頻度制限をしない 43.1%よりも低い値であった．

### 3.4 文脈語の正規化

同義語同士の周辺に出現する文脈語を観察すると，文脈語の中に同義語が多く存在する．例えば,「Cargo」と「貨物」という同義語には,「Cargo Loading」と「貨物搭載」というように「Loading」と「搭載」という同義語が出現する．これらの文脈語の同義語を正規化することによる平均精度への影響を調べた．28 対の同義語について，58 個の文脈語の正規化を行った．同義語について，平均精度の変化を調べたところ，平均精度が極めて向上したものが 4 個，その他の同義語については，若干，平均精度が向上または低下した．その結果，全体での平均精度は 42.7%であった．表 2 に 4 個の同義語についての

表 2: 文脈語の正規化による平均精度 (%) の比較

| 同義語 | 正規化した文脈語 | 平均精度の変化 |
|---|---|---|
| BRFG = ブリーフィング | DISP = ディスパッチ<br>Dispatch = ディスパッチ<br>Cabin = キャビン | 0.35 ↓ 3.03 |
| Bird = 鳥 | Strike = 衝突 | 1.75 ↓ 0.83 |
| Cargo = 貨物 | Loading = 搭載 | 0.98 ↓ 3.57 |
| Handling = ハンドリング | PAX = 旅客<br>GND = グランド<br>Ground = グランド | 10 ↓ 100 |

表 3: 同義語候補語の頻度による精度の高低の分類基準

| | |
|---|---|
| 基準 1 | 頻度 50 未満で，平均精度が 50%以上 |
| 基準 2 | 頻度 100 未満で，平均精度が 50%以上 |
| 基準 3 | 頻度を大きくすると，平均精度が 50%以上 |
| 基準 4 | 頻度を大きくすると，平均精度が 10%以上 |
| 基準 5 | 頻度を大きくしても，平均精度が 10%未満 |

例を示す．平均精度の変化は，上の数字が文脈語を正規化しない場合，下の数字が文脈語を正規化した場合を示す．このことから，人間の感覚の同義語同士の周辺に出現する文脈語の同義語よりも，より多くの語が類似度に関係しており，少数の文脈語の同義語の正規化により平均精度が向上する同義語は少なかったといえる．

# 4 同義語候補語の種類による精度比較

## 4.1 同義語候補語の分類

これまでは，主に文脈語について様々な観点から平均精度を調べてきたが，本節では，同義語候補語についての性質を調べる．同義語について「Dispatch」と「DISP」のようなアルファベット同士，カタカナとアルファベット,「座席」と「席」のような漢字同士,「Check」と「検査」のようなそれ以外のものに分類して，表 3 のような基準で平均精度を調べた．

表 4 にその結果を示すが，横軸の基準の数字は，各分類毎の基準 1〜5 での比率を示す．各分類の括弧の中の数字は，各分類の全体での比率を示す．アルファベット同士は，基準 1 と基準 2 の合計で 81%以上の平均精度が得られた．カタカナとアルファベットでは，基準 1 から基準 3 までの合計でも平均精度は 15% 程度であった．この理由としては，カタカナとアルファベットでは，周辺に出現する語が異なるためである．漢字同士の場合には，基準 1 から基準 3 までの合計で平均精度は約 63%得られた．それ以外の場合は，全体の 76%を占めるが，基準 1 から基準 3 までの合計で平均精度は約 27%であった．

表 4: 同義語候補語の種類による平均精度 (%) の比較

| 分類 \ 基準 | 基準 1 | 基準 2 | 基準 3 | 基準 4 | 基準 5 |
|---|---|---|---|---|---|
| アルファベット同士 (13.4) | 69.2 | 12.5 | 6.7 | 8.7 | 2.9 |
| カタカナアルファベット (8.5) | 4.5 | 3.0 | 7.6 | 18.2 | 66.7 |
| 漢字同士 (2.1) | 18.8 | 12.5 | 31.3 | 12.5 | 25.0 |
| それ以外 (76) | 10.7 | 3.9 | 12.7 | 22.8 | 49.9 |

## 4.2 日本語の略語

日本語の略語の平均精度について調査した．日本語の略語とは，例えば,「整備作業」と「整備」のようなものであり，完全に包含されるものである．したがって,「整備作業」と「整備点検」のようなものは含まれない．単名詞と複合語を合わせた同義語候補語 1,693 個について日本語の省略語の辞書を作成したところ，エントリー数：92，項目数：123，平均項目数：1.34 であった．この辞書を使用して実験したところ，平均精度で 52.3%という高い精度が得られた．これは，日本語の原型語と一部省略されている略語では，その周辺には同じような文脈語が出現しやすいと考えられ，本手法の得意な分野だといえる．

複合名詞の選択については，専門用語抽出システムが抽出したものの中で，重要度評価値が 3,000 以上の用語の中の複合名詞を使用した．

# 5 特異値分解による精度変化

同義語候補語は，行 ($t$) が文脈語，列 ($d$) が同義語候補語の $t \times d$ 行列 $A_{t \times d}$ と考えられる．特異値分解により，$A_{t \times d}$ は，(2) 式のように 3 つの行列 $U_{t \times n}$，$S_{n \times n}$，

表 5: 特異値分解による次元の縮約による平均精度 (%) の変化

| 縮約次元 | 平均精度 |
|---|---|
| 1,000 | 43.2 |
| 750 | 43.4 |
| 500 | 43.3 |
| 200 | 42.7 |
| 100 | 40.0 |
| 50 | 35.7 |

$\boldsymbol{V_{d \times n}}$ の積に分解できる [1]．ここで，$\boldsymbol{U}$ は $\boldsymbol{t}$ 次の直交行列，$\boldsymbol{V}$ は $\boldsymbol{d}$ 次の直交行列，$\boldsymbol{S}$ は $\boldsymbol{n}$ 次の対角行列である．

$$\boldsymbol{A_{t \times d}} = \boldsymbol{U_{t \times n}} \times \boldsymbol{S_{n \times n}} \times (\boldsymbol{V_{d \times n}})^{\boldsymbol{T}} \quad (2)$$

なお，$\boldsymbol{n}$ は，$\boldsymbol{t}$ と $\boldsymbol{d}$ の内の小さい方の値である．

類似度の計算は，$\boldsymbol{A^T A}$ を計算しているが，

$$\begin{aligned} \boldsymbol{A^T A} &= (\boldsymbol{USV^T})^{\boldsymbol{T}} \boldsymbol{USV^T} \\ &= \boldsymbol{VS^T U^T USV^T} \\ &= (\boldsymbol{SV^T})^{\boldsymbol{T}} (\boldsymbol{SV^T}) \end{aligned} \quad (3)$$

であるから，$(\boldsymbol{SV^T})^{\boldsymbol{T}}(\boldsymbol{SV^T})$ を計算することと等価である．したがって，$\boldsymbol{n}$ の次元を下げて，$(\boldsymbol{SV^T})^{\boldsymbol{T}}(\boldsymbol{SV^T})$ を計算することにより，縮約した次元での類似度を計算することができる．window[2,2] での元の次元 2,186 に対して，縮約した次元での平均精度の計算結果を表5 に示す．この表から縮約次元が 500 までは，縮約しない次元での平均精度 (43.1%) を上回るがその効果は僅かである．また，縮約次元を 100 にしても平均精度は，40%であった．この結果，特異値分解により平均精度の向上はあまり期待できないが，計算コストとスペースを節約できることが分かる．

## 6 結論および今後の課題

本論文では，特定分野における同義語の文脈語，同義語の種類・頻度について平均精度に対する影響を調査した．その結果，以下の知見が得られた：

- 文脈語の window 幅による重み付けを変化させたが，平均精度は低下した．
- 個別の同義語候補語に対する文脈語として頻度の大きなもののみを使用したが，平均精度は低下した．
- 同義語候補語に種類別での平均精度は，アルファベット同士が最も高かった．カタカナとアルファベットの平均精度は，最も低かったが，この手法の限界を示しているとも云える．これらについては，文脈語の正規化をするか，あるいは，文字情報を用いて同義語候補語の絞込みをすることにより，平均精度の向上が見込まれる．日本語の略語については，この手法の得意な分野であり，平均精度で約 52%の結果が得られた．
- 特異値分解により平均精度はほとんど向上しなかったが，平均精度の低下なしに計算コストとスペースを節約できることが分かった．
- 同義語の類似度を文脈情報から計算する手法を提案したが，副次的にこの手法により関連語を探索できることが分かった．例えば，「引き返し」という語をシステムに入力すると，「RTO(Rejected Takeoff)」,「GTB(Ground Turnback)」,「ATB(Air Turnback)」という関連語が得られる．

今後の課題としては，以下が挙げられる：

- 今回は，出現頻度 100 以上の同義語候補語について実験を行ったが，実用では，出現頻度をもう少し低くする必要がある．最小頻度 100 での同義語候補数が 1,343 であるのに対して，最小頻度を 50 にすると 2,192，最小頻度を 10 にすると 5,936 になった．
- 航空分野だけでなく他の分野の同義語でもこの手法をテストして有効性を確認する必要がある．

## 7 謝辞

専門用語自動抽出システムは，東京大学中川研究室・横浜国立大学森研究室で開発された用語抽出システムを使用させて頂きました．ここに感謝いたします．

## 参考文献

[1] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[2] 寺田昭，吉田稔，中川裕志. 文脈情報による同義語辞書作成支援ツール. **情報処理学会研究報告**, 2006-NL-176:87–94, 2006.

[3] 中川裕志，森辰則，湯本紘彰. 出現頻度と連接頻度に基づく専門用語抽出. **自然言語処理**, 10(1):27–45, 2003.