社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＮＬ－185 (2)
2008－SLP－ 71 (2)
2008／5／22

単語類似度ネットワークを通じた自動同義語獲得

王 玉馨　清水 伸幸　吉田 稔　中川 裕志

東京大学 情報基盤センター
{mini_wang, shimizu, minoru, nakagawa}@r.dl.itc.u-tokyo.ac.jp

コーパスから同義語の対を抽出するための一般的な方法では、通常二つ単語間の類似度(例えば、*cosine* 類似度)が必要である。類似度を使用することで、特定のクエリ単語に対しての類似語ランキングが可能になり、同義語候補リストから正しい同義語が認定できる。この論文では、それに加えて、単語類似度ネットワークを分析する新しい方法を提案する。単語類似度ネットワークでは閾値以上の類似度をアークとして、単語をノードとして定義する。提案する自動同義語候補選択ためのランク閾値(Rank Threshold for synonym candidate Selection method, RTS)によって類似度の順位が閾値以内のアークが構成される単語類似度ネットワークはスケールフリーグラフである。この性質に基づいて、我々は新しい同義語候補のリランキング手法を提案する。これを相互リランキング法(Mutual Re-ranking Method, MRM)と呼ぶ。同義語獲得における提案手法の有効性を示すためにMRM方法をReuters-21578に適用した。実験結果によって、RTSとMRMが同義語抽出の品質の向上させることが示された。

Automatic Synonym Acquisition through Word Similarity Network

Yuxin WANG, Nobuyuki SHIMIZU, Minoru YOSHIDA, and Hiroshi NAKAGAWA

Information Technology Center, The University of Tokyo
{mini_wang, shimizu, minoru, nakagawa}@r.dl.itc.u-tokyo.ac.jp

Popular methods for acquiring synonymous word pairs from a corpus usually require a similarity metric between two words, such as *cosine* similarity. This metric enables us to retrieve words similar to a query word, and we identify true synonyms from the list of synonym candidates. Instead of stopping at this point, we propose to go further by analyzing word similarity network that are induced by the similarity metric for the edges with the similarities that are ranked as top threshold number. By introducing the rank threshold for synonym candidate selection method (RTS), our analysis shows that the network exhibits a scale-free property. This insight obtained from the network leads us to a method for re-ranking the synonym candidates -- a mutual re-ranking method (MRM). We apply our methods to Reuters-21578 to show the generality of the methods on synonym acquisition. The results show that RTS and MRM boosts the quality of acquired synonyms.

1. はじめに

シソーラスはテキスト処理たとえばテキスト分類における効率と性能を向上させるための最も重要なリソースの一つである。したがって、シソーラスの根幹をなす同義語の獲得は重要なタスクである。特定のドメインにおける高品質のシソーラスがなければ、例えば事故報告コーパス中の重大事件を過大か過小に評価してしまうかもしれない。ただし、対象にしているドメインにおいて、自動的に同義語を取得するのは簡単ではない。

同義語獲得するには様々な方法(Hindle 1990; Lin 1998; 萩原ら 2005; 萩原ら 2006)を提案されている。それらは、同義語名詞なら似ている文脈情報を持つという分布仮説(Harris 1985)に基づいている。同義語獲得では、一般的に以下のステップの通りにこの仮説は実行される。まず、第 1 段階の処理では、コーパスから抽出された重要度の高い単語の文脈特徴における統計情報を抽出する。したがって、各単語はこれらの文脈特徴のベクトルによって表される。第 2 段階の処理では、*cosine* 類似度などの類似性量度を選び、それをクエリ単語と同義語候補の単語対に適用して、類似度を計算する。類似度の降順で各クエリ単語の同義語候補リストを作る。最後に同義語リストからトップ候補を選んで、クエリ単語の同義語と認定する。

本論文では第 2 段階の処理の後に新たに第 3 段階の処理を加えることを提案する。第 2 段階の類似度によって形成されるネットワーク、すなわち単語をノード、類似度の順位が閾値以内の単語間にアークを持つとしてネットワークの構造を調べてみると、スケールフリーの性質を持っているのが分かった。したがって、クエリ単語の同義語である可能性が比較的高い単語だけを対象にする自動同義語候補選択ためのランク閾値を決める手法 RTS(Rank Threshold for synonym candidate Selection method)と単語類似度ネットワークの構造を活用する同義語候補の相互リランキング法 MRM(Mutual Re-ranking Method, MRM)を提案する。MRM でリランキングする際、スケールフリーネットワークにある類似度の降順でランクされた同義語候補はハブ単語と非ハブ単語を分けて扱う。同義語関係は対称だが、我々の MRM は対称ではない。以前の研究では単語類似度ネットワークの構造的な特性を使用していない。本論文で提案した RTS と MRM 手法は 2 節で詳しく説明する。

以下で先行研究について説明する。分布仮説について、萩原ら(2006) は役に立つ文脈情報を調べた。彼らは三つ英語のコーパスを使って、同義語抽出に違う文脈情報:係り受け関係(動詞の主語と目的語)および名詞の修飾語などの有効性を比較した。実験結果から英語の同義語抽出に対しては以上すべて使った文脈情報の組み合せは一番有効であることが示された。寺田ら(2006)は名詞の隣接語を文脈語として用い、日本語の航空安全報告レポートから日本語の同義語を自動的に得た。窓サイズ 0 から 4 で隣接語を用いた実験の結果、専門分野の同義語獲得においては窓のサイズは 2 の場合が最も効果的であることを示した。

この論文の第二の目的は、萩原ら(2006)と寺田ら(2006)が同義語獲得するため使った文脈情報を利用して、提案した RTS と MRM を英語のコーパスに適用し評価することである。

この論文の構成は以下の通りである。2 節では、提案する単語類似度ネットワークに基づいた自動同義語候補選択ためのランク閾値の決める手法 RTS と同義語候補の相互リランキング法 MRM を導入する。そこでは、指定のコーパスからの単語類似度ネットワークを分析する。2 節で想定した類似度と文脈特徴については 3 節で詳しく説明する。4 節で実験結果について報告する。5 節はこの論文のまとめである。

2. 単語類似度ネットワーク

あらかじめ決められた数よりも多く出現した名詞を目標名詞と呼ぶ。目標名詞はノードとして、二つノード間で付随する類似度がアークとして、単語類似度ネットワークを描ける。今のタスクはあるクエリノードに隣接しているアークからの真の同義語アーク(クエリ単語に対しての正解同義語)を選ぶことである。この論文の中で、自動同義語獲得の対象はコーパスの名詞に限る。この節で調べる単語類似度ネットワークは Reuters-21578 コーパスから得たものである。

以下では単語類似度ネットワークのスケールフリーの性質について説明する。

2.1 スケールフリーネットワークとRTS手法

スケールフリーネットワークでは、ほとんどのノードは次数が少ないが、いくつかのノードは次数が大きなハブノードである。より具体的には、スケールフリーネットワークはノードからの出次数 k が式(1)で表わされるようにベキ乗則に従う分布 P(k)である接続グラフである。

$$P(k) \propto k^{-\gamma} \quad (1)$$

図 1 はスケールフリーネットワークの例である。図 1 の中で強調している黒いノードはハブノードである。

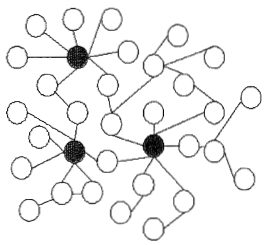

図 1. スケールフリーネットワーク.

実験対象としているコーパスから得られた単語が形成するグラフがスケールフリーネットワークであることを示すために、単語類似度ネットワークにおけるノード間の類似度の順位が閾値 *RT* 以内のアークだけを取り出して、ノード間の関係を図 2 で表した。

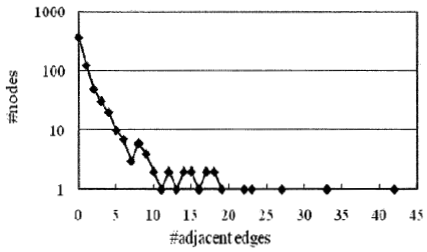

図 2. 隣接ノード数とその頻度の関係.

クエリ単語と第 *i* 候補単語の類似度が s_i であるとすると、クエリ単語に対する第 *i* 候補単語の出現確率をここでは式(2)で近似する。

$$p_i = \frac{e^{s_i}}{\sum_i e^{s_i}} \quad (2)$$

このように近似すると、パープレキシティ *RT* は式(3)のようになる。

$$RT = 2^H \quad (3)$$

$$H = \sum p_i \log p_i \quad (4)$$

コーパスから得たグラフの定義によれば、*RT* はあるクエリ単語に対して情報理論上の同義語の数と見なされるので、ランク閾値として使用するのは妥当と考えられる。ここで、提案した式(3)で定義された自動的にそれぞれのクエリ単語の同義語候補のランク閾値 *RT* を決める方法 RTS(Rank Threshold for synonym candidate Selection method)と呼ぶ。

単語類似度ネットワークにおける各クエリ単語から同義語候補まで二つ単語間の類似度が最上位の *RT* 個をとって、隣接ノード数とその頻度の関係は図 2 で示す。予想通り、単語類似度ネットワークにおいて、次数と与えられた次数を持つノード数の関係はほぼベキ乗則に従っている。

2.2 相互リランキング手法

クエリ単語 *x* は同義語候補 *y* があると仮定する。もし *y* が *x* に対して真の同義語であるとすると、逆にクエリ単語 *y* に対して *x* が真の同義語であることも成立する。この考えに基づいて、我々は相互リランキング手法(MRM)を提案する。

リランキングする前に、とりあえずクエリ単語 *x* と *x* の各同義語候補(クエリ単語以外の目標単語)間の類似度を計算する。そうして、各同義語候補は類似度の降順で同義語候補リストを作る。このリストはリジナル同義語候補リストと呼ぶ。

続いて、MRM は以下の式(5)で計算されたランクスコア(*RS*)を用いて、このオリジナル同義語候補リストをリランクする。まず、オリジナル同義語候補リストから一つの同義語候補 *y* を選んで、*y* がクエリ *x* のオリジナル同義語候補リストの順位は *rank*(*x*, *y*)と表す。反対に、*y* はクエリ単語として、*y* の同義語候補リストも得られる。もし、*x* はクエリ *y* の同義語候補リストで現れたら、*x* がクエリ *y* のオリジナル同義語候補リストの順位は *rank*(*y*, *x*)と表す。*x* が現れないなら、*x* をクエリ *y* の同義語候補と見なさない。

言い換えれば、クエリ *x* に連続しているアークの中で *rank*(*x*, *y*)番目に重いのは *x* から *y* までのアークである。反対に、クエリ *y* に連続しているアー

クの中で *rank*(*y*, *x*)番目に重いのは *y* から *x* までのアークであるともいえる。

x から *y* までランクスコア(*RS*(*x*, *y*))は式(5)で定義する:

$$RS(x,y) = A * \log(rank(x,y)) + \log(rank(y,x)) \quad (5)$$

式(5)の *A* は *x* から *y* までの同義語候補リスト順位と *y* から *x* までの同義語候補リスト順位を相互考えるための定数である。

式(5)を用いて、クエリ *x* のオリジナル同義語候補リストにある各候補語 y_i の *RS*(*x*, y_i)計算してから、*RS*(*x*, y_i)の昇順で *x* の全候補語をもう一度ランク付けする。

四種類の文脈情報特徴の組み合せ(文脈情報特徴については詳細を節 4.1 と 4.2 で述べる)のとき、MRM での実験結果によって、定数 *A* が 1/2 の際、ほとんどの場合同義語獲得の性能はピークだった。定数 *A* を変数とした時の式(5)の性能を図 3 で示す。変数 *A* が最適設定 1/2 の時、同義語獲得の性能は 29.3%から 31.4%まで向上した。

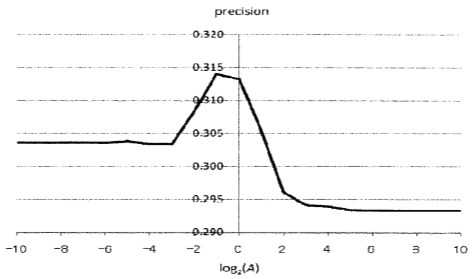

図 3. 定数 *A* の性能.

以下、式(5)と定数 *A*=1/2 について直観的に説明する。この式のユニークな特性は類似性が本質的には対称であるにもかかわらず、相互から見たランクが非対称ということである。また、図 2 で示したランク閾値を超えるアークによって構成される単語類似度ネットワークはスケールフリーグラフである。

スケールフリーの単語類似度ネットワークの中でクエリ単語はハブの場合、同義語候補の数は大きいので、正しく同義語を選ぶのは難しく、ランクの信頼性は低くなる。反対に、同義語候補の数は少なければ、同義語の選ぶのは相対的に簡単である。式(5)で、もしクエリ単語 *x* がハブノードである場合、定数 *A*=1/2 はハブでないノードから獲得された情報 *rank*(*y*, *x*)に大きい重みを与えて、正しい同義語 *y* のリランク結果を高くする。逆に、もしクエリ単語 *x* が非ハブノードで、*y* がハブである場合、定数 *A*=1/2 はハブ *y* からの *rank*(*y*, *x*)は通常非常に低いため、リラング後の結果に影響はあまりないと見られる。以上は式(5)が同義語を獲得するに有効の理由である。

以下でこの非対称特性を利用し有効性を得た二つノードの例を示す。 同義語ネットワークは重み付きの完全グラフだが、各クエリの類似度閾値は式(3)によって自動的に選択された閾値を設定し、重要ではない候補語を捨てている。

名詞「total」をクエリとすると、単語ネットワークにおける同義語候補は以下のようになる。

total:

stake, rest, fed, *number, money, share, capital,*amount, stock, decision, …

名詞「amount」がクエリの場合の同義語候補は以下の通りである。

amount:

time, *number, *total, stake, chance, money, volume, value, fund, share, …

星印(*)は正しい同義語を示す。この例では、リスト中の単語は類似度の降順でランクしており、ノード「total」はハブとして「amount」は非ハブノードとして考えられる。この例によると、「amount」が「total」の同義語候補として順位が低く、「total」が「amount」の候補として順位が高いため、ハブノードの同義語獲得は困難だと分かる。反対に、非ハブノードの同義語を得るのは相対的に簡単である。MRM は、ハブでないノードを利用することで高い性能を出す手法である。

3. 実験設定

3.1 文脈情報

同義語抽出の実験は名詞を目標対象とする。名詞は頻度があらかじめ決めた閾値より少ないものを無視することとする。目標名詞の文脈情報を抽出するのは萩原 (2006) などが提案した方法と同様で、名詞を主語あるいは目的語として取る動詞、名詞につく形容詞、さらに、名詞と前後連続している隣接語、の四種類である。隣接語の場合は、前後一つから三つまで連続語を使って比較する。

3.2 類似度

この研究における目標名詞の類似度を計算するために、ベクトル空間モテル VSM(Vector Space Model)を採用する。VSM は比較的簡単でありながら、同義語獲得での有効性が知られている(寺田ら、2006; 萩原ら、2006)。VSM でそれぞれの目標名詞はベクトルとして表される。ベクトルの次元は文脈情報の特徴語で、値は重みを付けた対応する特徴語の頻度である。

ここで、文脈情報特徴語の頻度は *tf* と定義する。類似度を計算するため、萩原らは *tf*・*idf* という重みを付ける方法を用いたが、寺田らは *log*(*tf*+1)は *tf*・*idf* より効果的であることを示した。本研究では寺田らの *log*(*tf*+1)という重みを採用する。

類似度の計算は、式(6)の *cosine* 法を用いる。

$$\cos(x, y) = \frac{x \cdot y}{\| x \| \cdot \| y \|} \qquad (6)$$

3.3 実験コーパスと結果の評価基準

本研究に関する実験では英語の Rueters-21578 コーパス(28MB)[1]と英語の構文解析器 Enju[2]を使用する。目標名詞の選択は Enju の出力における POS タグを基づく。POS タグは"NN"と"NNP"だけで名詞として扱われる。Rueters-21578 コーパスが相対的に小さいので、目標名詞を選択するための頻度閾値は 50 に設定されている。この頻度閾値で、選択された目標名詞の総数は 385 個である。

同義語の候補リストは以下の通りに作る: すべての目標名詞対の類似度を計算したのち、各目標名詞はクエリとして与え、他のすべての目標名詞を同義語候補として類似度の降順でランクして、目標名詞の同義語候補リストとして扱う。

本研究で提案した自動的に同義語獲得する方法の性能を評価するために、Longman 同義語辞書を使用する。選択された385個の目標名詞の中、154 個が Longman 同義語辞書の中に含まれている。評価には、情報検索分野でよく使われている平均精度という評価測定を使用する。あるクエリに関して、*k* 位にランクされた同義語候補の精度は以下の通りである。

$$precision\ (k) = \frac{1}{k} \sum_{1 \le i \le k} r_i \qquad (7)$$

ここで、候補語は正しいなら r_i=1 が、なければ r_i =0 である。そして、クエリ語に対して与えられる全部同義語候補語の平均精度は

$$Ave\,\mathrm{Pr}\,e = \frac{1}{|D_q|} \sum_{1 \le j \le N} (r_j * precision(j)) \qquad (8)$$

ここで、N がすべての同義語候補の数で、D_q は正しい同義語の数である。

ある単語ノードに関して、式(3)で計算された *RT*が理論上の連続アークの数ため、*RT*は同義語数の閾値として用いるのが妥当だと考えられる。ある単語ノードに関して実際の同義語数はゼロになるかもしれないと見なされるので、さらに平均精度と再現率の調和平均 *F-measure* を使用して、提案した同義語抽出方法の総合的な性能を評価する。

再現率は以下の通りに与えます。

$$recall = \frac{\#\,SynonymExt\ ractedTerm}{\#\,TotalSynon\ ymExistTer\ m} \qquad (9)$$

さらに、*F-measure* は以下の通り計算する。

$$F - measure = \frac{2 * precision\ * recall}{precision\ + recall} \qquad (10)$$

4. 実験結果

この節では、それぞれ四種類の文脈情報特徴を利用した実験結果とそれぞれ組み合せの実験結果を報告する。それに、提案した RTS と MRM 手法を適用する前後の性能比較および関連する議論をする。

4.1 文脈情報の重要性

この実験では四種類の文脈情報特徴の有効性を評価する: 主語と目的語の動詞、名詞の形容詞、および前後連続している隣接語である。隣接語の場合は、最も役に立つ文脈情報特徴を得られるために、前後 1 から 3 までの窓サイズをテストした。これから、それぞれ隣接語の窓サイズは隣接語 1、隣接語 2、および隣接語 3 と呼ぶ。四種類文脈情報特徴を別々に用いた同義語抽出の実験結果は表1で示している。

---

1 http://www.daviddlewis.com/resources/testcollections/reuters21578/.

2 http://www-tsujii.is.s.u-tokyo.ac.jp/enju/manual/index.html

別々に各文脈情報特徴を用いた実験では、目的名詞を目的語として取る動詞が同義語獲得に最も効果的であった(13.9%)。次に効果的な文脈情報特徴は、名詞の形容詞が 2 番目で、隣接語1が3番目である。主語の動詞の有効性は最低で、5.5%しかなく、目的語の動詞と名詞の形容詞の性能の半分未満である。隣接語の場合、隣接語 1 の性能は一番高い。隣接語 2 は隣接語 1 よりわずかに下がったが、隣接語 3 はさらに低くなった。つまり、本研究の実験結果によって、隣接語を用いて同義語獲得する場合、隣接語 1 は一番役に立つ文脈情報特徴で、より大きい窓がより役に立つ特徴を提供するとは言えないことを示した。

表 1 各文脈情報特徴の性能と有効性の順番

文脈情報特徴	平均精度	有効性の順番
主語の動詞	0.055	4
目的語の動詞	0.139	1
名詞の形容詞	0.113	2
隣接語 1	0.105	3
隣接語 2	0.104	
隣接語 3	0.091	

隣接語 1 が異なった窓サイズの中で最もよく効いたため、文脈情報特徴を組み合せるときは目的語と主語の動詞、名詞の形容詞、および隣接語 1 の四種類の文脈情報特徴で実験を行った。しかしながら、四種類すべての文脈情報特徴の組み合せは目的語の動詞だけを用いるときより性能が低い。ここから、同義語獲得では各文脈情報特徴の重要性が異なっているため、組み合せる時はそれぞれの文脈情報特徴に異なる重みを付けるべきだと考えられる。

続いて、文脈情報特徴を異なる重みで組み合せ実験した。この実験では、性能を目的語の動詞だけを用いるときより高めるため、各文脈情報特徴を表 1 に示されている有効性順に一つずつ加え、重みを調整して、実験をした。実験結果は表 2 に示されている。

表 2 異なる重みを付けた各文脈情報特徴の組み合せの性能

文脈情報特徴	重み	平均精度
目的語の動詞	1	0.1387
目的語の動詞+名詞の形容詞	2+1	0.1717
目的語の動詞+名詞の形容詞 +隣接語 1	4+2+0.3	0.1739
目的語の動詞+名詞の形容詞 +隣接語 1+主語の動詞	4+2+0.3+0.01	0.1741

4.2 RTS と MRM の有効性

4.1 節での実験では同義語獲得に各文脈情報特徴とそれらの組み合せの有効性を示した。4.2 節では本研究で提案した自動的に同義語候補を絞り込むためのランク閾値を決める手法 RTS の有効性を示すために、4.1 節実験結果で性能が一番高い組み合せ(表 2 中最後のケース、これから「組み合せ」と呼ぶ)で類似度の閾値と比較した。RTS によって自動的に選択された候補ランク閾値と、類似度の閾値を変化させながら閾値以上の性能を比較した結果は図 4 で示されている。図 4 の横軸は全コーパスに適用する人手で決めた類似度の閾値であるが、縦軸は様々な類似度の閾値で同義語候補を閾値以上のみに絞り込んだ際の *F-measure* である。凡例の「RTS」は提案したランク閾値の決める手法 RTS で、つまり 2.1 節の式(3)によって自動的に選択された *RT* で各クエリ単語から同義語候補まで二つ単語間の類似度が最上位の *RT* 個を取ることで、同義語候補を絞り込んだ場合の *F-measure* である。

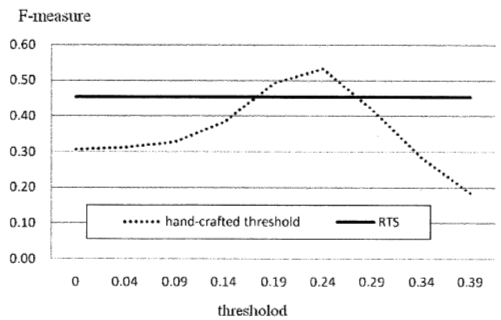

図 4. 自動同義語候補選択のためのランク閾値の決める手法 RTS と人手決めの類似度閾値を適用された性能比較。

実験結果によって、自動同義語候補選択ためのランク閾値の決める手法(RTS)が比較的によく寄与することを示している。RTS での性能は一番高いわけではないが、RTS を用いない場合よりは RTS を用いるほうが同義語獲得に有効的な傾向になることが分かった。特筆すべきは、閾値を人手で決めるのは難しいが、RTS はクエリ単語によって自動的に決定されることである。教師データなしでは各クエリ単語に対して同義語候補を絞り込むための閾値が分からないため、本研究の実験結果は実用的に重要であると考えられる。

次に、提案した同義語候補の相互リランキング法(MRM)の有効性を知るため、*cosine* の類似度を基づく各クエリ単語に RTS を適用した上さらに全体に MRM を適用して、実験した。RTS の *RT* と MRM の *RS* は式(3)と式(5)の通りに計算する。実験結果は表 3 に示されている。

表 3 提案した RTS と MRM を用いる前後と各文脈情報特徴と組み合せの性能比較。

文脈情報特徴	平均精度			再現率	F-measure		
	---	RTS	MRM		--	RTS	MRM
主語の動詞	.055			.935	.104		
		.112	.168	.714		.193	.272
目的語の動詞	.139			.909	.241		
		.235	.268	.818		.365	.404
名詞の形容詞	.113			.916	.202		
		.225	.251	.727		.344	.373
隣接語 1	.105			.851	.187		
		.131	.178	.520		.209	.265
組み合せ	.174		.184	.935	.294		.308
		.293	.314	.818		.432	.454

表 3 の中、「組み合せ」という行は表 2 の中最後のケース場合で、つまり異なる重みを付けた四種類文脈情報特徴の組み合せを意味する。「--」の列は提案した RTS と MRM のいずれも用いない場合の性能で、「RTS」列は RTS 方法での性能で、「MRM」は RTS を用いた上にさらに MRM を加えて適用した性能を意味する。「--」列の場合は、全部クエリに対して類似度閾値は 0 に設定されている。つまり、類似度が 0 でなければ、絞り込んだ同義語候補に含まれると仮定している。

表 3 からは、類似度の閾値が低い時、同義語獲得の再現率は高いが、クエリ単語に対応する比較的に低い類似度をもっている同義語候補は同義語獲得の精度は低いことが分かる。逆に、表 3 にはないが、類似度の閾値は高ければ、比較的に低い類似度をもっている同義語候補は無視されるため、精度は向上するが、再現率は劣化する。そこで、式(3)に基づく各クエリ単語対して自動的に同義語候補の数を決める *RT* を適用することによって、クエリ単語に対応する比較的に高い類似度をもっている同義語候補だけが同義語として扱われる。したがって、表 3 の通り、ベースラインである閾値 0 の場合と比べ平均精度は全体的に改善した。結論として、RTS を用いて再現率は少し劣化したが、総合的な性能指標 *F-measure* は改善した。

続いて、クエリ単語と同義語候補対の関係を考慮し、さらに提案した同義語候補の相互リランキング法(MRM)を適用することによって、平均精度と *F-measure* の両方は全体的に改善した。表 3 の平均精度は定数 *A* が最適化された式(5)で計算された *RS* によってのものである(図 1 は組み合せの文脈情報特徴で定数 *A* 最適化後の結果である)。実験結果によって、提案した RTS の上さらに MRM を応用したところ、定数 *A* が 1/2 のとき同義語獲得の性能を最も向上した。四種類の文脈情報特性の組み合せの場合、同義語獲得の平均精度は 29.3%から 31.4%まで改善することが示された。

5. まとめ

本論文では、各クエリ単語と候補単語対間の類似度に基づき自動的に同義語候補を絞り込むためのランク閾値の決める手法(RTS)と同義語獲得のために単語類似度ネットワークを通じて各クエリ単語と候補単語を相互にリランキングする手法(MRM)を提案した。提案した方法での同義語獲得の優位性を示すために、Reuters-21578 コーパスに適用して様々な実験をした。

この研究の主な貢献は二つある。まず、提案した RTS(自動同義語候補選択のためランク閾値を決める手法)で選択された単語ノードの単語類似度ネットワークがスケールフリーの特性を持っていることを示した。更に、同義語候補リストを改良するため MRM(相互リランキング法)を提案した。詳

細な実験により、ベースラインの類似度の閾値よりRTSで自動的に選択したランクの閾値が有効であり、MRMを加えてさらに有効性を示した:RTSとMRMで平均精度は29.3%から31.4%まで、*F-measure*は43.2%から45.4%まで改善した。

もう一つの貢献は四種類の文脈情報特徴(主語の動詞、目的語の動詞、名詞の形容詞および前後連続している隣接語)の組み合せに異なる重みを付ける調査である。この研究における実験コーパスは小さく、文脈情報特徴はまばらだが、実験結果は四種類の文脈情報特徴を組み合わせる手法の有効性を示した。

Longman同義語辞書での実験結果によって、四種類の文脈情報特徴の中で、目的語の動詞が最も効果的であることを示した。二番目は名詞の形容詞で、三番目は隣接語である。主語の動詞の貢献は最も低く、目的語の動詞の半分未満である。隣接語の場合は、窓サイズ1が最も効果的で、より大きい窓サイズはより役に立つ文脈情報特徴を提供するとは言えないとわかった。したがって、異なる重みを付けて四種類の文脈情報特徴を組み合せする時、各文脈情報特徴の有効性に対応する重みでの組み合せが最も有効で、中でも目的語の動詞がもっている文脈情報特徴は一番重要だと考えられる。

同義語獲得はテキスト分析、テキストマイニング、およびシソーラス生成など様々な自然言語アプリケーションに非常に役に立つ。今後の計画としては、日本語の助詞が提供している文脈情報など他種類の文脈情報の有効性を調べ、同義語獲得の更なる性能の向上を目指す。

参考文献

1. Dekang Lin. 1998. Automatic retrieval and clustering of similar words. *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational linguistics (COLING-ACL'98)*:786-774.
2. Donal Hindle. 1990. Noun classification from predicate-argument structures. *Proc. of the 28th Annual Meeting of the ACL*:268-275.
3. Macro Baroni and Sabrina Bisi. 2004. Using cooccurrence statistics and the web to discover synonyms in a technical language. *Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC 2004).*
4. Masato Hagiwara, Yasuhiro Ogawa, Katsuhiko Toyama. 2005. PLSI Utilization for Automatic Thesaurus Construction. *Proc. of the Second International Joint Conference on Natural Language Processing (IJCNLP-05)*:334-345.
5. Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiro Toyama. 2006. Selection of Effective Contextual Information for Automatic Synonym Acquisition. *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*:353-360.
6. Zellig Harris. 1985. Distributional Structure. *The Philosophy of Linguistics.* Oxford University Press:26-47.
7. 寺田昭、吉田稔、中川裕志。2006。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・自然言語処理研究会報告、IPSJ SIG Notes. 2006-NL-176-(13), pp. 87-94.