言語の分散表現と擬似適合性フィードバックを用いた英日言語横断検索

玉置　賢太† 　林　 佑明†† 　酒井　哲也†

† 早稲田大学基幹理工学部情報理工学科酒井研究室　〒 169–0072 東京都新宿区大久保 3–4–1
†† Language Technologies Institute, Carnegie Mellon University　5000 Forbes Ave, Pittsburgh, PA 15213 USA
E-mail: †madao@akane.waseda.jp, ††hiroakih@cs.cmu.edu, †††tetsuyasakai@acm.org

あらまし　異言語の文献にアクセスする言語横断情報検索は，翻訳の精度により性能が左右される．翻訳精度の向上について，skip-gram による分散表現を用いることで，文脈情報を持った翻訳を行なうことが提案されている．そこで本研究では，言語横断情報検索のために，異言語間にわたる言語の分散表現の類似性を利用してクエリ翻訳を行い，またクエリ翻訳の前後に擬似適合性フィードバックによりクエリ拡張を行なう方式について実験を行なう．結果としては，提案手法の検索有効性はベースラインを下回ることとなった．本論文では，この結果に対する失敗分析を行い，将来への改善点を述べる．

キーワード　言語横断情報検索, word2vec, 分散表現

1. 導　　入

言語横断情報検索は，ユーザクエリの言語と検索対象文書の言語が異なる場合の情報検索を実現する技術である [1]．その主要なアプローチは，ユーザクエリを機械翻訳などの手段により検索対象文書の言語に翻訳するものである．このため，言語横断情報検索の検索有効性は，翻訳の精度に大きく依存している．一方，単言語情報検索においては，1994 年の TREC-3 (The Third Text Retrieval Conference) あたりから，検索結果の質を高めるために擬似適合性フィードバックによるクエリ拡張が盛んに研究されてきた [3]．言語横断情報検索においても，クエリ翻訳と拡張を組み合わせることにより，質の高い検索が実現できることが知られている [4] [5]．

1 つ目の要素であるクエリ翻訳には，大きく分けて 2 種類の方法が考えられる [6]．1 つは，元言語とターゲット言語間に対応した辞書を元に翻訳を行なう，ルールベース機械翻訳である．もう 1 つは，2 言語間の対応を持つパラレルコーパスの統計的情報を元に翻訳を行なう，コーパスベースな統計的機械翻訳である．

2 つ目の要素であるクエリ拡張には，言語横断情報検索の場合，3 種類の拡張が考えられる．翻訳前にクエリと同一の言語のコーパスを用いて拡張を行う翻訳前の拡張，翻訳後に検索対象コーパス自身を用いて行う翻訳後の拡張，そして両方を行なう場合の 3 種類である [6]．

前述の辞書ベースなクエリ翻訳のアプローチとして，Mikolov らは，word2vec [7] により生成される分散表現の異言語にわたる類似性を利用した手法を提案している [8]．ここで，word2vec とは，単語のベクトル空間上に，意味が近い単語同士を近くに，意味が遠い単語同士を遠くに配置する skip-gram による分散表現手法 [9] を実装したものである．Mikolov ら [8] の手法をクエリ翻訳に用いることで，コーパス中の文脈を利用して翻訳精度を高めることが期待できる．ここでのコーパスとは，word2vec で分散表現を生成する際に用いたコーパスのことである．コーパス中の文脈を利用した翻訳精度の向上については，林らの研究 [6] で一定の効果があることが示されている．

分散表現による翻訳がクエリ翻訳に有効であるとわかると，今後の言語横断情報検索の研究に新たな選択肢を生むことができる．上で述べたとおり，word2vec により生成された分散表現には，教師データの文脈が表現されている．すなわち，教師データのジャンルにより，専門性の高い検索にも応用できる可能性がある．

本論文では，分散表現によるクエリ翻訳に加えて，クエリ翻訳の前後に擬似適合性フィードバックによりクエリ拡張を行なう手法について実験を行い，その効果を検証する．

2 章では関連研究を述べ，3 章では提案手法について述べる．4 章では実験方法について記述し，5 章でその結果について述べる．6 章で考察を行う．

2. 関 連 研 究

2.1　言語横断情報検索

言語横断情報検索においてクエリ拡張を利用する初期の研究に Ballesteros らの研究がある [4]．この文献によると，辞書ベースな言語横断情報検索の性能は単言語情報検索のそれを下回ってしまうが，その原因の 1 つに辞書ベース翻訳による語義曖昧性があると述べられている．文献では，クエリ拡張を用いることで，上記のエラーを減らす効果があるかどうか検証されている．実験では，local feedback [10] と local context analysis [11] の 2 つのクエリ拡張手法が用いられている．結果として，クエリ拡張を行うことで辞書ベースによる語義曖昧性を減らし，検索有効性を向上する効果があると示されている．

2.2　言語の分散表現

コーパス中の単語の異なり数 T を次元とするビットベクトルにより各単語を表現すると，互いに等距離なベクトルが T 個得られる．この各ベクトルを，より低い次元で表現したものがこ

こでの分散表現である [12].

Mikolov らの発表した文献 [9] では，Skip-gram モデルにより言語の分散表現を生成する方式が示されている．Skip-gram モデルとは，図 1 のように，ある単語に着目したときに，その周辺に生起する単語を予測するモデルである．その値は，以下

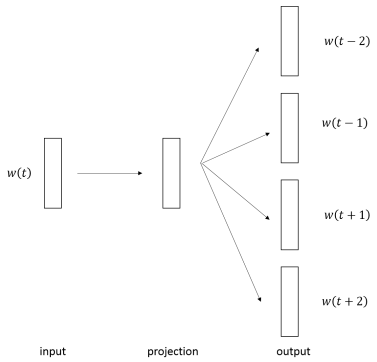

図 1　Skip-gram のイメージ

の式のように平均対数尤度を最大化することにより求められる．

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

ここでは，T をコーパス内の単語数とし，コーパス内の単語を $(w_1, w_2, w_3, \ldots, w_T)$ としている．また，ウィンドウサイズを c とする．

第 1 章で述べたとおり，word2vec はこの手法を実装したものである．このツールにより生成されたベクトル空間は，コーパスに含まれる各単語の意味的な遠近関係，つまりコーパスの文脈を表現している．また，ある単語のベクトルを別の単語同士の関係から，線形演算により求めることができる場合がある．例としては，$vec('Berlin') - vec('Germany') + vec('France')$ から，$vec('Paris')$ に近いベクトルが求められる場合などである．このことから，word2vec により得られるベクトル空間は，コーパス内の単語の意味的な遷移も表現できていると考えられる．

2.3　異言語間の分散表現の類似性

word2vec により得られたベクトル空間について，もう 1 つ興味深い性質が知られている [8]．それは，ベクトル空間上に存在する各単語の位置関係に，異言語にわたって類似性が見られる，というものである．図 2 に Mikolov ら [8] の論文から転載した具体例を示す．この性質を利用し，両言語ベクトル空間内の単語ベクトル同士のマッピングを考えれば，精度のよい単語翻訳が実現できると考えられる．

Mikolov ら [8] は，2 言語間の単語ベクトル変換規則を取得するため，2 つのベクトル空間にわたる線形遷移を学習する方法を示している．ここで得られる変換規則は翻訳行列と呼称され，以下のような確率的勾配降下法 [13] の問題を解くことで求められる．

$$\min_{W}\sum_{i=1}^{n} \| Wx_i - z_i \|^2 \tag{2}$$

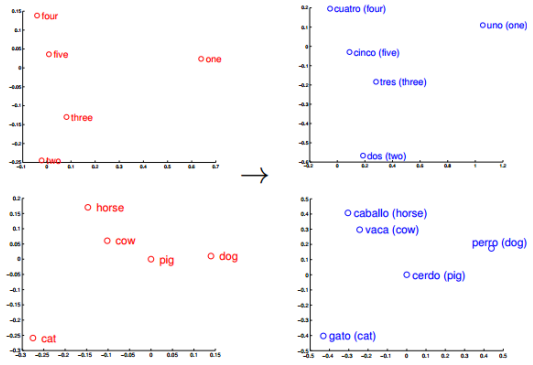

図 2　英語とスペイン語それぞれのコーパスによる分散表現間の類似性 (左が英語の分散表現，右がスペイン語の分散表現．Mikolov ら [8] より転載．)

ここで，W は求める翻訳行列であり，x_i は一方のコーパス内に存在する単語 i の単語ベクトル，z_i は x_i と同じ意味を持ち，他方のコーパス内に存在する単語 i の単語ベクトルである．n は学習に用いる単語の対の数である．

Mikolov ら [8] の研究では，英語・スペイン語間，英語・チェコ語間，また英語・ベトナム語間での翻訳について実験と評価を行なっている．この実験により，翻訳行列による翻訳が，言語的に関連している英語・スペイン語間だけでなく，関連が無いチェコ語やベトナム語などとの翻訳に対しても機能することが示された．

3.　提 案 手 法

本研究では，第 2 章で述べた異言語間の分散表現が持つ類似性を利用した翻訳を，言語横断情報検索におけるクエリ翻訳に適用する．さらに，分散表現によるクエリ翻訳に翻訳前，翻訳後，そして翻訳前後の 3 種類の擬似適合性フィードバックを用いたクエリ拡張を行う．

以上の手法により，英日言語横断情報検索を行い，その効果を検証する．

3.1　クエリ翻訳

クエリの翻訳は，文献 [8] で述べられた，翻訳行列により翻訳を行なう方式に従う．翻訳行列は前述の通り，式 2 により生成する．

翻訳行列の生成に用いる教師データは，2 言語にそれぞれ存在する同じ意味を持つ単語の対である．教師データの生成については，文献 [6] に従う．この論文では，分散表現生成に用いたコーパス内から 5000 語を取り出し，Google 翻訳 [14] により翻訳を行い，教師データを生成している．

3.2　擬似適合性フィードバックを用いたクエリ拡張

クエリ拡張には，前述のとおり，翻訳前，翻訳後，翻訳前後の 3 種類について拡張を行い，その効果の違いを検証する．ここで，各拡張に用いるコーパスは，拡張するクエリの言語と同じ言語のものを用いる．

本論文では，拡張前の検索結果における上位の文書を適合文書とみなし，Robertson らによる Offer Weight [15] に基づき擬似適合性フィードバックを行うことにより，クエリ拡張を実

現する．以下に，その計算式を示す．

$$OW = r * \log\left(\frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)}\right) \quad (3)$$

ここで，N はコーパス内の文書数，n はそのうち当該クエリタームを含む文書数，R は全適合文書数，r はそのうち当該クエリタームを含む文書数である．

実験では，適合文書とみなす文書数 R を 10 とし，拡張クエリタームの個数も 10 とした．

3.3 ベースライン

本研究では，言語横断情報検索システムのベースラインとして，クエリ翻訳に Google 翻訳 [14] を用いた場合の実験も行い，提案手法との検索有効性を比較する．提案手法との違いは，クエリ翻訳に翻訳行列を用いるか，Google 翻訳を用いるかというだけである．つまり，クエリ拡張については提案手法と同様に，3 種類の拡張を行い，提案手法との検索有効性を比較する．

4. 実験方法

実験方法をまとめた図 3 を示す．

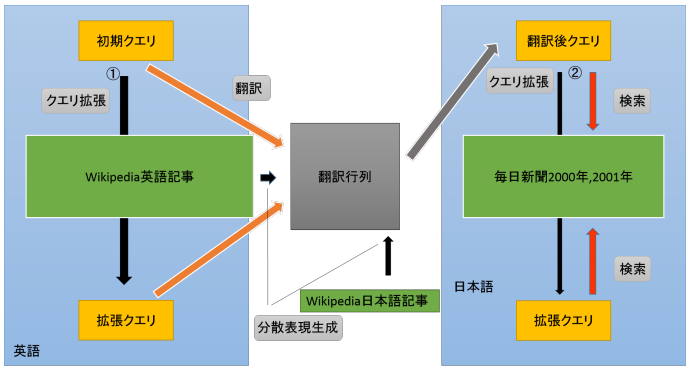

図 3 実験の概要

4.1 評価データ

本論文の検索対象記事には，NTCIR-5,6 CLIR タスク [16] で用いられた，毎日新聞の 2000 年と 2001 年の新聞記事 (約 20 万記事) を用いる．記事の情報のうち，実験に用いる部分は，林らの研究 [6] に倣っている．C0 タグで表される索引記事番号を文書 ID とし，文書として扱うのは，T2 タグで表される本文のみとしている．また，評価するトピックは NTCIR-5,6 CLIR で用いる 97 トピック及びこれらの多値適合性判定データを用いる．この時，クエリとして扱うのは，各トピックのタイトルフィールドとしている．

4.2 実験データ

精度の高い分散表現生成には，極めて大きいサイズのコーパスが必要となる．文献 [8] では，単語数にして数百億単語と，巨大なコーパスを用いていた．本論文では，コーパスとして Wikipedia の英・日記事 [17] を与えている．Wikipedia 英語記事は約 460 万記事であり，Wikipedia 日本語記事は約 90 万記事である．また，word2vec に与える次元数のパラメータについては，Mikolov らの文献 [8] に倣っている．これによると，ソース言語の英語を 800 次元，ターゲット言語のスペイン語を 200 次元としていることから，本研究では，英語を 800 次元，日本語を 200 次元としている．

4.3 クエリ拡張用コーパス

クエリ拡張については，翻訳前のクエリ拡張と，翻訳後のクエリ拡張に，それぞれの言語のコーパスを与えている．翻訳前のクエリ拡張の言語は，本論文においては英語である．ここでは図 3 の矢印 1 のとおり，Wikipedia 英語記事を元にクエリ拡張を行っている．翻訳後のクエリ拡張の言語は日本語であり，図 3 の矢印 2 のとおり，検索対象コーパスである NTCIR-5,6 CLIR [16] の毎日新聞の記事そのものを用いる．

4.4 検索エンジン

本論文では，検索対象コーパスの indexing，及び検索に Indri [19] を用いている．Indri の検索モデルは Ponte らによる combination of the language modeling [20] と，Turtle らによる inference network retrieval frameworks [21] がベースとなっている．

実験では，Indri の機能を利用し，初期クエリタームの重みを 1.0 に，拡張クエリタームの重みを 0.5 とした．

4.5 評価指標

評価には，NTCIR におけるいくつかのタスクの評価に利用されてきた NTCIREVAL [22] を利用している．NTCIREVAL により，様々な評価指標から実験結果を評価した値が算出されるが，本論文では特に MSnDCG@1000 の値を用いる．

nDCG とは，normalized discounted cumulative gain のことで，検索された適合文書のランクが下位であるほど，その価値が下がるとした評価指標である [23]．本来の価値から，ランクにより価値を下げたものを減損利得と呼び，DCG は減損利得の和であり，DCG を正規化したものが nDCG である．このため，nDCG の最大値は 1 となる．本論文で用いる MSnDCG とは，Burges ら [24] により再定義された nDCG のことである [23]．

5. 実験結果

5.1 手法ごとの平均 MSnDCG 値

提案手法及びベースラインのそれぞれについて，3 種類の拡張を行い検索をした結果を表 1 に示す．

表 1 各手法における平均 MSnDCG 値の比較

	拡張無し	翻訳前	翻訳後	翻訳前&翻訳後
提案手法	0.1577	0.2240	0.1627	0.1926
ベースライン	0.3363	0.2053	0.3269	0.3000

結果としては，翻訳行列を用いた各手法は，対応する Google 翻訳によるベースラインの検索有効性を概ね下回った．提案手法のうち，最も平均値が高くなったのは翻訳前拡張を行った場合であるが，この値もベースラインの拡張無しの場合を下回っている．なお表 1 の結果は現在最新版ではないので，後日正誤表を作成し，http://www.f.waseda.jp/tetsuya/publications.html にリンクを添付する．

5.2 辞書ベース言語横断情報検索に対するクエリ拡張の有効性

表 1 より，提案手法の全ての拡張において拡張無しの MSnDCG 値を上回っている．このことから，辞書ベース言語横断情報検索に対するクエリ拡張の有効性が確認できる．以下に，提案手法におけるトピックごとの拡張無しに対する拡張有りの MSnDCG 値の分布を示す．

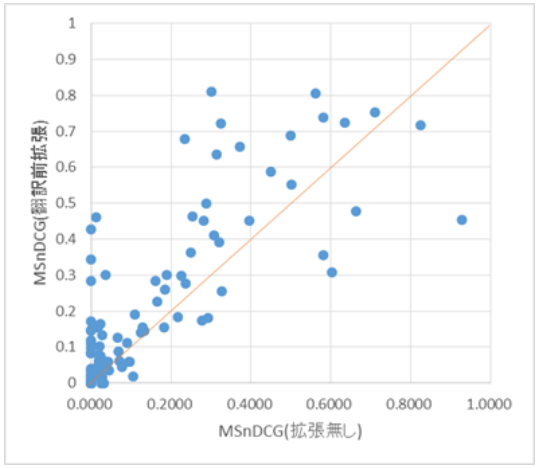

図 4　提案手法における拡張無しに対する翻訳前拡張有りの MSnDCG 値の分布

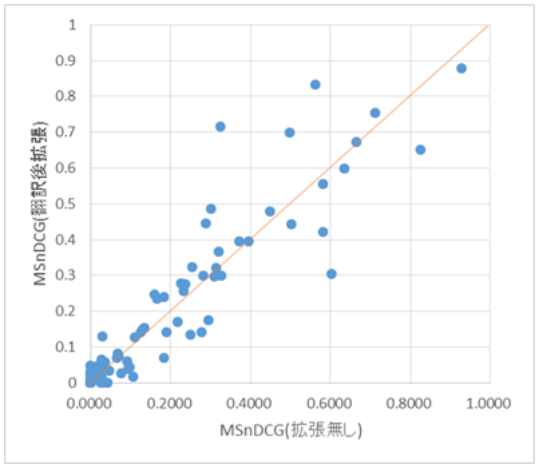

図 5　提案手法における拡張無しに対する翻訳後拡張有りの MSnDCG 値の分布

図 4 に表される翻訳前拡張の分布を見ると，拡張無しに比べて大きく低い値が出ているのは 1 点であり，残りは $x = y$ に沿っているか高い値を取っている．特に，拡張無しにおいて低い値が出ているトピックに関して性能の向上が確認できる．これに対し，図 5 によると，値の分布が $x = y$ に沿っていることがわかる．このことから，翻訳後拡張においては大きな性能向上が少ないことがわかる．特に，拡張無しにおいて低い値が出ていたトピックに性能の向上が見られず，0 付近に多くの分布が固まってしまっていることが確認できる．また，翻訳前&翻訳後の平均値は，翻訳後拡張ありの結果に引っ張られてしまい相乗効果でさらに性能が向上する結果にはならなかった．

結論としては，提案手法には翻訳前拡張が有効であり，翻訳後拡張の効果は限定的なものであった．トピックごとの分析は後述する．

5.3 ランダム化 Tukey HSD 検定による統計的有意性

本論文では統計的有意性を確かめるために，ランダム化 Tukey HSD 検定 [25] による実験システムの任意の 2 値の p 値を算出した．以下にその値を示し，比較・分析する．

表 2　ベースライン手法同士の各拡張に対する p 値

システム対	拡張無し	翻訳前	翻訳後	翻訳前&翻訳後
拡張無し	-	0	0.9802	0.4656
翻訳前	-	-	0	0.0006
翻訳後	-	-	-	0.6996
翻訳前&翻訳後	-	-	-	-

表 3　提案手法同士の各拡張に対する p 値

システム対	拡張無し	翻訳前	翻訳後	翻訳前&翻訳後
拡張無し	-	0.1706	1	0.8828
翻訳前	-	-	0.2552	0.9290
翻訳後	-	-	-	0.9460
翻訳前&翻訳後	-	-	-	-

表 4　提案手法とベースライン手法の任意の 2 システムの p 値

システム対		ベースライン			
		拡張無し	翻訳前	翻訳後	翻訳前&翻訳後
提案手法	拡張無し	0	0.6422	0	0
	翻訳前	0.0004	0.9946	0.0022	0.0710
	翻訳後	0	0.7726	0	0
	翻訳前&翻訳後	0	0.9998	0	0.0010

表 4 より，提案手法とベースライン手法を比較してみると概ね有意水準 5%において有意差がある．ただし，ベースラインの翻訳前拡張と提案手法の各拡張とを比べると，軒並み p 値が高く有意差が得られていない．また，提案手法同士を比べた場合も有意水準 5%において有意差が得られていない．よって，これらのシステムについてはより多くのデータを用いた検証が必要であると考えられる．また表 3 より，特に拡張無しに対して翻訳後拡張を行っても検索に大差が無いことがわかる．

5.4 トピックごとの分析

5.1 節ではシステムごとの平均 MSnDCG 値を比較・分析した．ここでは，特徴的なトピックを抽出し，比較・分析を行う．

NTCIR-5 : 002 ‘President of Peru, Alberto Fujimori, scandal, bride’

このトピックでは，翻訳後拡張において MSnDCG 値の向上が見られた．ここでは，「Peru」を正しく翻訳することに成功しているが，重要な固有名詞である「Fujimori」が「パボン」という意味が通らない単語に翻訳されてしまっていた．しかし，正しく翻訳された単語から翻訳後拡張によって「フジモリ」を得ることができたため，検索有効性の向上につながったと考えられる．また，提案手法による翻訳では「of」が「鷹木恵子編

著」と翻訳されるため，このトピック以外でも「鷹木恵子編著」という単語が頻出した．

NTCIR-5 : 011 ‘Ichiro, Rookie of the Year, Major League’

今回の実験で，提案手法において拡張無しの状態で MSnDCG 値が 0 の場合，翻訳前・翻訳後のいずれの拡張を行っても検索有効性が大きく向上しないことがわかった．特に，拡張無しで MSnDCG が 0 だった 26 件の中で 11 件ものトピックにおいて，クエリ拡張による検索有効性の向上が一切見られなかった．原因としては，元クエリ翻訳の失敗が考えられる．拡張無しでの検索有効性が低いということは，すなわち翻訳に失敗しているということである．つまり，元クエリの翻訳に失敗しているため，いくらクエリ拡張を行っても検索有効性が向上しなかったと考えられる．トピック 011 はその 1 例である．

このトピックを翻訳行列により翻訳すると「レギュラー級 キングス時代 鷹木恵子編著 する なる スプレイク スコットランドリーグ」となった．このように，クエリ翻訳によって全く意味の違う単語になってしまっていることがわかる．文脈を考慮した翻訳は,「Ichiro」のような一般的な人名とも有名人の愛称とも取れる単語の翻訳で曖昧性を回避することが期待されていた．しかし，ここでは全く関係ない「レギュラー級」という単語に翻訳されてしまっている．

NTCIR-5 : 017 ‘India and Pakistan territorial conflict, nuclear weapons’

トピック 017 は拡張無しと比べて翻訳前拡張の結果が最も悪くなったトピックである．また，翻訳後拡張の MSnDCG 値も拡張無しと比べて低くなった．このトピックの特徴として，拡張無しの状態で MSnDCG 値が 0.9286 と非常に高いことが挙げられる．翻訳前拡張されたクエリを提案手法により翻訳すると,「security」や「test」などの関係の薄い単語が拡張語として含まれている．このようなノイズによって検索有効性が下がってしまったと考えられる．

翻訳後の拡張は翻訳前の拡張に比べて MSnDCG 値の差が少ない．この原因は，翻訳前拡張には拡張時と翻訳時で 2 重にノイズがかかってしまったためであると考えられる．翻訳後拡張でも「春日」「考之」といったノイズが含まれている．しかし，翻訳前拡張では前述の拡張時のノイズに加え，翻訳時に「weapon」が「グレネード」になり,「indian」が「カリブ諸国」になるなど，元は関連のあった単語が翻訳により関連が薄くなってしまっている．以上のように 2 重にノイズが含まれることにより，MSnDCG 値が大幅に下がる結果になったと考えられる．

NTCIR-6 : 003 ‘Embryonic Stem Cells’

このトピックでは，提案手法 (拡張無し) に対してベースライン (拡張無し) の MSnDCG 値が最も高くなった．提案手法によるこのトピックの翻訳結果は「王冠状 脂質二重膜」となっており，ベースラインを用いた場合「胚の 幹 細胞」となっている．提案手法を用いた場合，近くの単語を発見できずクエリが 1 つ減ってしまっている上，翻訳できている単語も意味が遠くなってしまっている．それに対し，ベースラインでは意味の近い翻訳を実現できている．この様に，翻訳精度の差がそのまま検索有効性の差となってしまっている．

NTCIR-6:059 ‘Television Broadcasting, Digitalization’

このトピックは翻訳前拡張により，拡張無しよりも MSnDCG 値が向上した例である．もともとのクエリが 3 単語であり，かつ拡張により「digital」や「signal」,「uhf」など関連の高い単語が拡張されている．加えて，元クエリの「digitalization」が「電子化」と翻訳され若干意味が遠い単語になってしまったのを,「digital」が「デジタル」と翻訳されたことでフォローすることができている．以上のような理由から翻訳前拡張によって検索有効性を向上することができたと考えられる．

トピックごとの分析から，提案手法はいくつかのトピックにおいて極端に結果が悪かったために平均値でベースラインを下回ったわけではなく，万遍無く多くのトピックで評価値が下回っていたことがわかった．また，この原因が翻訳精度の低さにあることがわかった．翻訳精度ではベースラインを大きく下回ったが，翻訳前拡張と翻訳後拡張がそれぞれ提案手法に対し有効であり，それぞれの特性を確認することができた．

5.5 元クエリから stopwords を除去した場合の結果

本論文では，追加実験として元クエリから stopwords を除去した場合についての MSnDCG 値を算出した．以下にその値と，ランダム化 Tukey HSD 検定による p 値を示す．

表 5 stopwords を除去した場合の拡張無しシステムにおける平均 MSnDCG 値

	元クエリ	stopwords 除去
提案手法	0.1675	0.1682
ベースライン	0.3636	0.3635

表 6 元クエリを用いたシステムと stopwords 除去を行ったシステムの任意の 2 値の p 値

システム対		stopwords 除去	
		ベースライン	提案手法
元クエリ	ベースライン	0.9796	-
	提案手法	-	0.2958

表 5 より，提案手法において stopwords 除去を行った場合，平均値として若干の向上が見られた．トピックごとに注目してみると，stopwords を含んだトピックは 97 トピック中 16 トピックであり，そのうち 10 トピックについて MSnDCG 値の変化が見られた．ここで変化が無かったトピックは全て，元クエリによる MSnDCG 値が 0 であった．これにより検索有効性の向上の余地が無かったと考えられる．また表 6 より，提案手法の元クエリを用いたシステムと stopwords を除去したシステムとの p 値には，有意水準 5%において有意差がない．これは前述したとおり，stopwords を含むトピックが 97 トピック中 16 トピックしかないことが大きな原因であると考えられる．

6. 結　　論

6.1 擬似適合性フィードバックを用いたクエリ拡張の役割

本論文では，5.2 節で述べた比較の通り，辞書ベース言語横断情報検索に対する擬似適合性フィードバックを用いたクエリ拡張に一定の有効性が確認できた．今回の実験では，平均値だけ見ると翻訳後拡張の MSnDCG 値は拡張無しと大きく変わることが無かった．しかし，5.4 節におけるトピックごとの分析により翻訳前拡張だけでなく，翻訳後拡張の役割も確認できた．

まず翻訳前拡張には，翻訳前そのままの意味を持つ単語によってクエリ拡張を行うことができるという利点があることがわかった．このため，翻訳前拡張の拡張語はトピックに関連した単語を多く得ることができた．加えて，関連した単語を多く得ることができるため，元クエリの翻訳が上手くいかなかった時に，近い意味を持つ別の拡張クエリタームによって翻訳後のクエリの意味を補完しているケースも確認できた．ただし，拡張無しの時点で高い検索有効性を持つトピックに対しては，拡張クエリを得るタイミングと拡張クエリを翻訳するタイミングの 2 度のタイミングでノイズが混ざる可能性があることがわかった．このため，元々高い検索有効性を持つトピックに対しては，検索有効性が大きく下がる可能性があるという知見が得られた．

翻訳後拡張には，拡張クエリタームにノイズが混じった場合であってもその後は検索するだけであるので，ノイズが混ざる可能性が低いことがわかった．このため，元々高い検索有効性を持つトピックであっても，その有効性を大きく下げることが無いと考えられる．しかし，翻訳後拡張の精度は拡張無しの場合の検索有効性と同じく翻訳自体の精度によるところが大きいことが確認できている．翻訳前拡張のように，翻訳が上手く行かなかった場合の意味の補完ができている例も確認できたが，特に元クエリタームが少ない場合において検索有効性の向上が少ないことがわかった．元のクエリタームが少なく，その全てにおいて翻訳が上手くいかなかったため意味の補完もできず，拡張クエリタームと元クエリの関連性が薄くなってしまったことが原因であると考えられる．

以上のように，翻訳前拡張と翻訳後拡張はともに利点と欠点を持つことがわかった．今後はこれらの特徴を分析し有効にクエリ拡張を行うことで，辞書ベース言語横断情報検索において，さらなる検索有効性の向上が望めると考えられる．

6.2 翻訳行列を利用した言語横断情報検索への課題

本論文は，提案手法である翻訳行列を利用した言語横断情報検索に，いくつかの課題が残る結果となった．まず，5.4 節で挙げた翻訳精度の問題である．表 1 より，提案手法は全ての拡張パターンにおいて Google 翻訳を用いたベースラインより低い平均値が算出されている．また，以下にトピックごとの提案手法 (拡張無し) とベースライン (拡張無し) の MSnDCG 値の差を示す．

以上の 2 つの図を見ると，いくつかのトピックで大きな差をつけられているのではなく，満遍なく MSnDCG 値がベースラインを下回っていることがわかる．このため，どれかのトピックが提案手法と相性が悪かったために平均値が下回ったわけではなく，単純に翻訳精度の差が平均値に表れたと考えられる．よって，翻訳行列を用いた言語横断情報検索の検索有効性を向上させるには，翻訳精度の向上が不可欠であると考えられる．

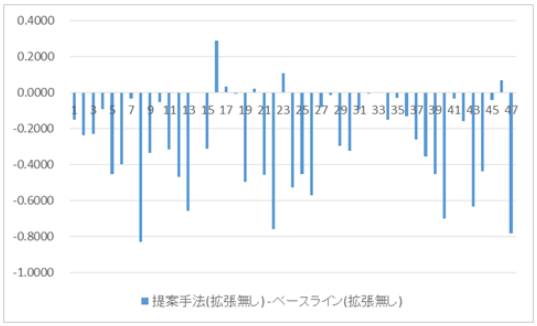

図 6　NTCIR-5 に含まれるトピック集合における MSnDCG の差 (拡張無し)

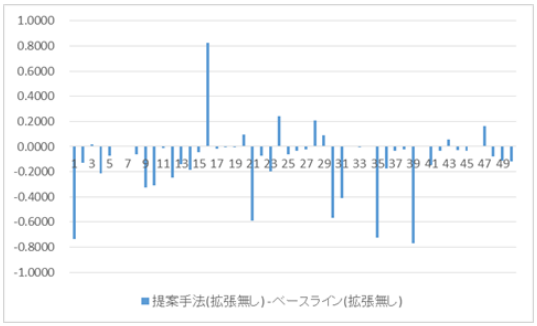

図 7　NTCIR-6 に含まれるトピック集合における MSnDCG の差 (拡張無し)

翻訳精度の向上のためにできることとしては，分散表現の調整が挙げられる．本論文では，分散表現生成の際に設定する次元数を，Mikolov ら [8] によって示された 800 次元と 200 次元とした．ただし，Mikolov らの実験に用いられた言語は英語・スペイン語・チェコ語であったため，必ずしも英日翻訳に適しているとは限らない．よって，次元数に関してはさらなる吟味が必要であると考えられる．

また，追加実験により stopwords を除去することにより有意に検索有効性が上昇しないことがわかった．このため，翻訳行列の絶対的な翻訳精度を向上するためには，他のアプローチが必要である．

文　　献

[1] Hull, David A., and Grefenstette, Gregory. Querying across languages: a dictionary-based approach to multilingual information retrieval. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1996.

[2] Grefenstette, Gregory (ed.): Cross-Language Information Retrieval, Kluwer Academic Publishers, 1998.

[3] Voorhees, Ellen, M. and Harman, Donna, K.: TREC: Experiment and Evaluation in Information Retrieval, The MIT Press, 2005.

[4] Ballesteros, Lisa, and Croft, W. Bruce. Phrasal translation and query expansion techniques for cross-language informa-

tion retrieval. ACM SIGIR Forum. Vol. 31. No. SI. ACM, 1997.
[5] Sakai, T., Koyama, M., Izuha, T., Kumano, A., Manabe, T. and Kokubu, T.: Toshiba BRIDJE at NTCIR-6 CLIR: The Head/Lead Method and Graded Relevance Feedback, NTCIR-6 Proceedings, pp.36-43, May 2007.
[6] 林 佑明, 酒井 哲也. 言語の分散表現による文脈情報を利用した言語横断情報検索 DEIM Forum, 2015.
[7] word2vec, http://code.google.com/p/word2vec
[8] Mikolov, Tomas, Le, Quoc V. and Sutskever, Ilya. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013).
[9] Mikolov, Tomas, et al. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. 2013.
[10] Attar, Rony, and Fraenkel, Aviezri S. Local feedback in full-text retrieval systems. Journal of the ACM (JACM) 24.3 (1977): 397-417.
[11] Xu, Jinxi, and Croft, W. Bruce. Query expansion using local and global document analysis. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1996.
[12] 西尾 泰和, word2vec による自然言語処理, オライリー・ジャパン, 2014.
[13] Gardner, William A. Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique. Signal Processing 6.2 (1984): 113-133.
[14] Google 翻訳, http://google.com/translate
[15] Robertson, Stephen E., and Sparck Jones, Karen. Simple, proven approaches to text retrieval. University of Cambridge. Computer Laboratory, 1994.
[16] Kishida, Kazuaki, et al. Overview of CLIR task at the fifth NTCIR workshop. Proc. Fifth NTCIR Workshop. 2005.
[17] Wikipedia 記事コーパス, http://dumps.wikimedia.org/
[18] Pre-trained word and phrase vectors, https://code.google.com/p/word2vec/.
[19] Indri, http://sourceforge.net/p/lemur/wiki/Home/.
[20] Ponte, Jay M., and Croft, W. Bruce. A language modeling approach to information retrieval. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998.
[21] Turtle, Howard, and Croft, W. Bruce. Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems (TOIS) 9.3 (1991): 187-222.
[22] Sakai, T. NTCIREVAL: A Generic Toolkit for Information Access Evaluation. FIT 2011. Volume 2. RD-004. pp.22-30. 2011.
[23] 酒井哲也, 情報アクセス評価方法論 検索エンジンの進歩のために, コロナ社, 2015.
[24] Burges, Chris, et al. Learning to rank using gradient descent. Proceedings of the 22nd international conference on Machine learning. ACM, 2005.
[25] Sakai, T., Metrics, Statistics, Tests , PROMISE Winter School 2013:Bridging between Information Retrieval and Databases (LNCS 8173). pp.116-163. Springer. 2014.