自動獲得された言い換え表現を使った情報検索

海野 裕也[1]　　宮尾 祐介[1]　　辻井 潤一[1,2,3]
[1] 東京大学大学院情報理工学系研究科コンピュータ科学専攻
[2] 英国マンチェスター大学　[3] 英国国立テキストマイニングセンター
{unno, yusuke, tsujii}@is.s.u-tokyo.ac.jp

1　はじめに

情報検索において文書とクエリで異なる語が現れる問題は語彙の不一致として知られ，シソーラスを使ったクエリ拡張によって対処されてきた．人手で作られたシソーラスを使う場合，その規模やドメインが問題になることが多い．自動獲得されたシソーラスによるクエリ拡張は成功を納めてきたが，人手によるシソーラスとは性質が異なり，意味的に等価な表現ではなくて似たトピックへ拡張される傾向がある．一方で，近年言い換え表現をコーパスから自動獲得する手法が数多く提案されている．特に対訳コーパスから言い換え表現を得る手法は，人手で辞書を整備する必要がない上，言い換えらしさを示す言い換え確率付きで大量に言い換え表現を得ることができる．

我々はこの自動獲得された言い換え表現を従来の情報検索の枠組みに取り入れることによって，新しいクエリ拡張手法を提案する．本手法によって，ドメインに特化したクエリ拡張を行うことができるようになる．また獲得された言い換え表現は，その言い換えらしさに応じてスコア付けされるので，誤った言い換え表現による悪影響を小さく抑えることができる．

本手法の効果を確かめるために，NTCIR-1 と NTCIR-3 PATENT を使用して評価実験を行った．その結果，本手法によってクエリ拡張を行わない手法に比べて精度の高い検索結果を得られることが分かった．

2　背景

2.1　言語モデルに基づく情報検索

Ponte & Craft [5] は言語モデルを情報検索に適用する手法を提案している．彼らの手法は，文書 D から推定される言語モデルの下で，クエリ Q が生成される確率 $P(Q|D)$ を文書のランキング関数として使用するというものである．この手法では言語モデルをどのように設計するかによって検索性能が変わってくる．

Miller ら [4] は言語モデルとして，D から推定される unigram 言語モデルと，文書集合全体 C から推定される unigram 言語モデルの混合モデルを用いた：

$$P(Q|D) = \prod_{q \in Q} (\lambda P_{UL}(q|D) + (1-\lambda) P_{UL}(q|C))$$

この手法は非常に簡潔であり，また従来の TF/IDF 重みによる検索に比べて高い性能を示している．

2.2　言い換えの自動獲得

近年，言い換えの自動獲得に関する研究が盛んに行われている．我々は中でも Bannard & Callison-Burch の手法 [1] に着目した．彼らは，まずアライメントのとれた二言語対訳コーパスを用意して，同じ単語とアライメントのとれた単語を言い換え表現と見なした．例えば日本語の「二酸化炭素」と「炭酸ガス」は，両方とも英文中で「carbon dioxide」とアライメントがとられることが多い．このとき「carbon dioxide」をピボットとして,「二酸化炭素」と「炭酸ガス」が言い換え表現になっていると見なせるのである．

具体的には，以下の式によってフレーズ w_j がフレーズ w_i に言い換えられる言い換え確率を定義する：

$$P_{para}(w_i|w_j) = \sum_e P_{trans}(w_i|e) P_{trans}(e|w_j)$$

ただし P_{trans} は翻訳確率で，アライメントの頻度から $P_{trans}(w|e) = count(w,e)/count(e)$ と推定される．

2.3　関連研究

Qiu & Frei [7] や Schütze ら [9] は，語の共起関係からシソーラスを構築してクエリ拡張する手法を提案している．これらの手法は一定の成功を収めているが，同意語というよりは同一トピックの語によるクエリ拡張であり，我々の手法とは意義が異なる．実際，Mandala ら [3] はこうした共起を元にしたクエリ拡張を，WordNet などの人手で構築したシソーラスによるクエリ拡張と組み合わせることでより精度の高い検索結果を達成している．

Riezler ら [8] は我々と同様，自動獲得した言い換え表現を使ってクエリ拡張する実験を行っているが，彼らの手法では言い換え確率をクエリ拡張する単語の選

択にしか使っていない点，また TF/IDF 重みによる検索と組み合わせている点で我々の手法と異なる．

3 手法

3.1 言語モデルとの組み合わせ

言い換え確率を言語モデルの枠組みに取り入れるために，言い換えに基づく言語モデルを定義する．本手法でも，Miller ら [4] と同様，各単語は独立に生成すると仮定して，これらのクエリ中の各単語の生成確率の積としてクエリの生成確率を定義する．文書からは，まず単語 w が生成されて，これが言い換え確率 $P_{para}(q|w)$ に従ってクエリ単語 q に書き換わると考える．我々は，このモデルを言い換え言語モデル P_{PL} として定義する．

$$P_{PL}(q|D) = \sum_{w} P_{para}(q|w) P_{UL}(w|D) \qquad (1)$$

ただし，$P_{UL}(w|D)$ は文書から推定される unigram 言語モデルで，最尤推定によって容易に推定される．この言い換えに基づく言語モデルと unigram 言語モデルとの混合分布を作り，以下のように言語モデルを作った．

$$\begin{aligned} P(Q|D) &= \prod_{q \in Q} \{\lambda(\mu P_{UL}(q|D) + (1-\mu) P_{PL(q|D)}) \\ &+ (1-\lambda) P_{UL}(q|C)\} \end{aligned} \qquad (2)$$

3.2 言い換え確率の正規化

フレーズ w_i から w_j への言い換え確率，$P_{para}(w_j|w_i)$ は w_j の出現頻度に大きく依存するため，直感に反するスコアが割り当てられることがある．我々はこのスコアを正規化することで，より適切な言い換えのスコアを割り振る方法を提案する．

例えば,「炭酸ガス」は「二酸化炭素」に比べて文書中に出現する頻度が少ないため,「carbon dioxide」とアライメントがとられる頻度も低くなる．これは,「炭酸ガス」が「carbon dioxide」の正しい訳語であるかどうかとは関係なく起こり，そのため出現頻度の少ない単語には言い換え確率が低く見積もられる．そこで，単語の出現確率 $P(w_j)$ で割ったスコアを使う．但し，頻度の低すぎるフレーズはノイズと見なし 5 回以上出現したもののみ使用した：

$$S(w_j, w_i) = P_{para}(w_j|w_i) / P(w_j)$$

このスコアを 0 から 1 に収まるように，各 w_i に対する最大値で割って，言い換えスコアとして使う：

$$S_{para}(w_j, w_i) = S(w_j, w_i) / \max_i S(w_j, w_i) \qquad (3)$$

実験では，$S_{para}(w_j, w_i)$ を式 (1) における $P_{para}(w_j|w_i)$ の代わりとして使った．ただし，このスコアは j に関して足しあわせても 1 に成らないため，確率モデルとしての解釈は持たないことには注意しなければならない．

表 1: テストデータの統計

名称	分野	文書数	サイズ	クエリ数
NTCIR-1	論文	332,918	512 MB	83
NTCIR-3	特許	697,262	22 GB	31

表 2: 対訳コーパスの統計

名称	言語	分野	文対数
NTCIR-1 titles	日英	論文	330,148
NTCIR-3 titles	日英	特許	1,701,216

4 実験

4.1 実験設定

検索用タスクとして NTCIR-1 と NTCIR-3 PATENT を用いた．それぞれの統計を表 1 に示す．NTCIR-1 のトピック 1 から 30 を開発用のテストセットに，残りを評価用に使用した．文書として各文書のタイトルと本文のみを，またクエリとして各検索課題の短い説明（Description）のみを使用した．この検索課題の説明は「〜について述べた文献」などの冗長な表現を使っていたため，これらの文末表現を手動で取り除いた文字列を使った．ただし，この前処理を行っても検索性能に大きな性能変化がないことは予備実験で確かめた．各文書とクエリは形態素解析器で単語に分解し，名詞，動詞，形容詞，副詞のみを原型に直して使用した．また，単語単体のみではなく，複合名詞を精度良く検索できるようにするために，隣接する名詞対も単語集合に含めた．

対訳コーパスとして， NTCIR-1 と NTCIR-3 PATENT の言語横断検索テスト用の文書のタイトルを用いた．このデータを使ったのは，文アライメントをとる必要がないことと，検索課題と同一ドメインのため，似た語彙の言い換えを得られることが期待されたからである．それぞれの統計を表 2 に示した．

対訳コーパスには句アライメントがついていないため，MOSES ツールキット [2] を用いてアライメントをつけた．形態素解析器には MeCab [10] を使用した．対訳文対中の英語は PorterStemmer [6] によってステミングを施した．評価には 11 点平均精度を用いた．

言語モデル (2) における混合係数は開発用テストセットを使ってチューニングし，$\lambda = 0.2$，$\mu = 0.4$ とした．

4.2 実験結果

表 3 が，各検索課題と使用した対訳コーパス及び手法の関係である.「LM」はベースラインとして使用した Miller ら [4] の言語モデルを使った結果,「Para」は (1) の言い換え確率を使った手法,「Norm」が (3) で定義した言い換えスコアを使った手法の結果である．これらの結果から 2 つのことがいえる．1 つは，同一ドメインの対訳コーパスを使った方が，結果が良くなるとい

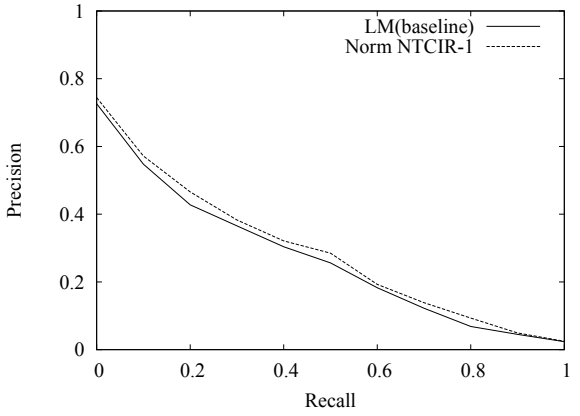

図 1: 再現率精度グラフ (NTCIR-1)

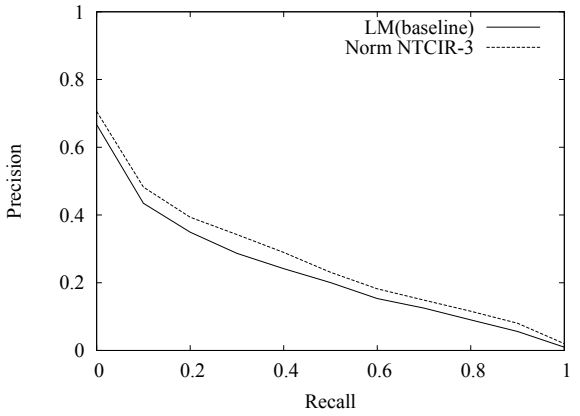

図 2: 再現率精度グラフ (NTCIR-3)

う点である．もう 1 つは正規化を施すことによっていずれの場合も性能が良くなっている点である．両テストセットで最も結果の向上した手法とベースラインの再現率精度グラフを図 1 と図 2 に示した．グラフは交差することなく，いずれの再現率においても精度が向上していることがわかる．

各検索課題ごとの性能差を調べるために，NTCIR-1 に対して，正規化した言い換えスコアと NTCIR-1 のタイトルを使った結果を，ベースラインから性能差で示したのが図 3 である．全 53 課題の内，33 課題で性能が向上した．特に 5 ポイント以上性能改善した課題は 14 件に及んだが，逆に 5 ポイント以上性能が悪化したのは 2 件のみであった．

表 3: 各手法とテストセットの評価比較

	NTCIR-1	NTCIR-3
LM (baseline)	27.90	23.76
Para NTCIR-1	28.65(+2.7%)	24.37(+2.6%)
Norm NTCIR-1	**29.71(+6.5%)**	25.28(+6.4%)
Para NTCIR-3	27.58(-1.2%)	24.57(+3.4%)
Norm NTCIR-3	28.14(+0.9%)	**27.19(+14.4%)**

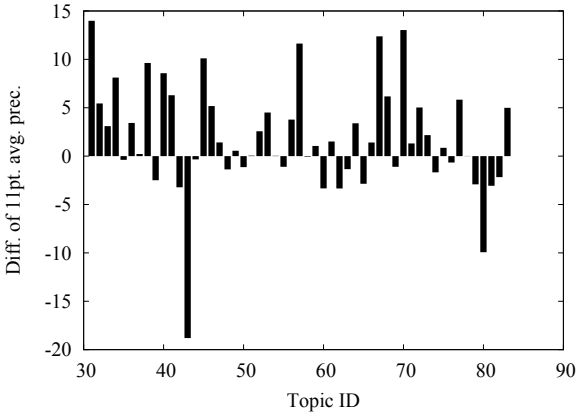

図 3: 検索課題ごとの性能差 (NTCIR-1)

表 4: クエリ拡張の影響を受けた文書

クエリ	文書	ランク
··· における特徴次元**リダクション**	··· 特徴空間の次元**縮小**の ···	1609 → 3
デジタル著作物の改変および無断の ···	··· を通した**ディジタル**映像の ···	45 → 6
日本語文における**カタカナ**外来語	**片仮名**表記の揺れ誤りや ···	223 → 21

4.3 改善した例

我々の手法によって改善した検索例を表 4 に示した．「クエリ」と「文書」はそれぞれ改善例のクエリと文書の抜粋であり,「ランク」にはランクの変化を示した．

最初の例は同義語への拡張の例である.「縮小」や「低減」といった単語が「リダクション」の同義語として認識されるため，こうした単語を含む文書も検索できるようになった．2 番目の例は異表記の例である.「デジタル」という単語は「ディジタル」とも表記されるが，いずれも英語では「digital」になるため，言い換えとして認識することができる．こうしたカタカナ語の異表記は非常に多い上，人手で管理するには非常にコストがかかる．3 番目の例も異表記の例だが，こちらは漢字とカナの違いである．これらの表記揺れに対しても，本手法は効率的に働くことが分かる．

4.4 悪化した例

結果の悪くなった課題を調べたところ，期待通りのクエリ拡張が行われていることが多かった．NTCIR-1 で最も結果が悪くなったクエリは「動画像圧縮を行う知能化イメージセンサ」である．クエリ拡張によって,「知的」あるいは「インテリジェント」などが「知能化」の言い換えとして認識された．これは期待される挙動であったものの，結果としてこれらの単語を含むが，クエリ中の重要単語である「イメージセンサ」を含まない文書が上位にランクしてしまったのが問題である．

これはクエリ拡張が正しく働かないためではなく，ベースとなる重み関数が不適切なため，クエリ中でさ

表 5: 言い換えの例

	既存手法	正規化後
デジタル	ディジタル: 0.77, デジタル: 0.15, 数値: 0.01, 電子: 0.01, ディジタル-・: 0.01	デジタル: 1.0, Digital: 0.89, ディジタル: 0.87, ディジタル-· 0.79, の-ディジタル: 0.59
本	**主: 0.11**, 本: 0.10, **単: 0.10**, **永久: 0.06**, 図書: 0.06	ブック: 1.0, 図書: 0.54, 本: 0.44, 帳: 0.43, 書: 0.2
体	**の: 0.50**, 体: 0.10, **材料: 0.06**, 物体: 0.02, **中: 0.01**	Body: 1.0, ボディ: 1.0, 胴体: 0.67, 身体: 0.41, 体内: 0.25

ほど重要でない「知能化」に大きな重みが与えられたためと考えられる．特に，重み付けは基本的に単語頻度のみに基づいており，周辺単語とは独立に決定される．このクエリにおいては,「イメージセンサ」が重要語であったにもかかわらず,「知能化」に大きな重みが与えられてしまった．

4.5 正規化の効果

言い換え確率の正規化の効果を示すために，獲得された言い換え表現の例を表 5 に示した．それぞれ言い換え確率の高い 5 つを，スコア付きで示している．

最初の例は頻度の少なすぎる語の例である.「Digital」という単語は出現頻度が低すぎるため，正しい言い換え表現にもかかわらず，既存手法では上位に現れない．2 番目と 3 番目の例は頻度の高い単語に誤ってアライメントがとられた例である．特に「の」は出現頻度が高い上，英語に直接対応する単語がないことが多いため，間違ったアライメントができて言い換え確率が高くなりやすい．こうしたアライメントの間違いの影響も，正規化によって軽減できることがわかる．

4.6 曖昧な訳語による問題

二言語対訳コーパスを使った言い換え表現の獲得は高い精度で言い換えを抽出できる一方で，ピボットの言語の性質を引き継いでしまう．特に，元の言語で曖昧性のない語が，ピボット言語において意味に曖昧性があると，その悪影響が言い換え確率に伝搬してしまう．

表 6 に例を示した．英語の「bank」には,「銀行」と「堤防」の 2 つの意味がある．そのため，P_{trans}(bank|銀行) も P_{trans}(堤防 |bank) も高い値をもち，結果として「銀行」と「堤防」が言い換えと認識されてしまう．特に英語には全く異なる意味を持った多義語が多いため，こうした現象が多く発生する．

今回の実験では，クエリ中にこうした単語がなかったため，この問題による悪影響は確認されなかった．しかし，対象文書によっては問題になることが予想される．改善策としては，別言語の対訳コーパスの結果を組み合わせたり，周辺単語の類似度から曖昧性を解消して，こうした間違いを減らす工夫が考えられる．

表 6: 曖昧な英単語による間違いの例

日本語	英語	言い換え
銀行	bank	**河岸**: 1.0, 銀行: 1.0, バンク: 1.0, **堤防**: 0.14, **護岸**: 0.12
演奏	play	演奏: 1.0, **遊び**: 0.89, 鳴り: 0.14, 音楽: 0.14, MUSIC: 0.10
粒子	particle	**助詞**: 1.0, 粒子: 0.89, **態**: 0.68, 粒子-の: 0.61, 微粒子: 0.44

5 結論

本研究は言い換え確率に基づく言語モデルを構築し，対訳コーパスから自動獲得した言い換え表現を使って情報検索に応用した．この際，検索課題のドメインと同じ対訳コーパスから得た言い換え表現を用いた方が，高い精度で検索できることが分かった．また，従来の言い換え確率ではアライメントの間違いに弱かったが，これを正規化することによってより精度の高い言い換えを得ることが可能になり，また検索性能も向上した．

本手法で精度の下がった検索課題では，クエリは適切に拡張されたものの，ベースとなる重み関数が不適切なため精度向上に結びつかなかった．また，ピボットとする言語中での意味の曖昧性が言い換え表現獲得に悪影響を及ぼすことが分かった．今後の課題としては，係受け関係などのクエリ文字列中の他の単語との関係を使った重み関数の開発，及び複数言語の対訳コーパスや共起単語から，曖昧な語義に影響されない頑健な言い換え表現の獲得を行う必要がある．

参考文献

[1] C. Bannard and C. Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL '05*, 2005.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of. ACL '07, Demo Sessions*, 2007.

[3] R. Mandala, T. Tokunaga, and H. Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *Proc of SIGIR '99*, 1999.

[4] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In *Proc. of SIGIR '99*, 1999.

[5] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. of SIGIR '98*, 1998.

[6] M. F. Porter. An algorithm for suffix stripping. *Program*, Vol. 14, No. 3, 1980.

[7] Y. Qiu and H. P. Frei. Concept based query expansion. In *Proc. of SIGIR '93*, 1993.

[8] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu. Statistical machine translation for query expansion in answer retrieval. In *Proc. of ACL '07*, 2007.

[9] H. Schütze and J.O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, Vol. 33, No. 3, 1997.

[10] 工藤拓. MeCab: Yet Another Part-of-Speech and Morphological Analyzer (http://mecab.sourceforge.net).