空間分割型 PLSI を用いた言語横断情報検索

高野 祐介[†] 村松 哲[†] 森 辰則[††]
[†] 横浜国立大学 大学院 環境情報学府 [††] 横浜国立大学 大学院 環境情報研究院
E-mail: {yusukeb,tetsu0730,mori}@forest.dnj.ynu.ac.jp

1 はじめに

近年，インターネットの普及などにより，各国の電子化文書を容易かつ大量に入手できるようになった．しかし，必要とされる情報が利用者の母国語で書かれているとは限らない．そのため，利用者が母国語により関連する他国語の文書をも検索できる言語横断情報検索 (Cross-Language Information Retrieval, CLIR) への要求が高まっている．検索要求の言語と検索対象文書の言語の差を吸収する代表的な方法として，検索要求を検索対象の言語に翻訳する検索要求翻訳型言語横断検索がある．その基本は単語翻訳であるが，一般に一つの単語に対して複数の訳語候補があるので訳語の曖昧性解消が必要である．曖昧性解消の手法としては翻訳候補に優先順位を付与する方法が代表的であり，その一つに，予め大量の対訳文書から翻訳確率情報を収集する手法がある．

本稿では，確率的 LSI(Probabilistic Latent Semantic Indexing, PLSI) に基づいて収集された確率情報を利用し検索要求翻訳を行なう方法を提案する．まず，対訳コーパスに対して PLSI を適用することにより，個々の語に対してその訳語の翻訳確率が計算できることを示す．PLSI は，行列計算の繰返しであるから，対訳コーパスが大規模になると時間 / 空間計算量が非常に大きくなり一般の計算機では処理できなくなる．そこで，対訳コーパスを分野に応じて小さく分割し，個々に PLSI を施し，得られた翻訳確率を合成することにより最終的な翻訳確率を得る方法を検討する．

2 PLSI による翻訳確率計算

PLSI は自動インデクシング法の一つであり，単語 w と文書 d の間の共起確率 $P(w,d)$ をアスペクトモデルに基づいて少数の非観測変数を媒介とする確率情報に変換し，情報検索に役立てる手法である [1]．まず，全検索対象文書について各単語の頻度を求めそれを成分とする単語 - 文書行列を作る．この行列から最尤推定により $P(w,d)$ を求めることもできるが，単語 - 文書行列は一般に疎であるため何らかのスムージングが必要である．PLSI では，単語や文書の数よりも遥かに少ない非観測変数 z を導入し，これを媒介とした確率情報を求めることにより，これに対応している．すなわち，文書 d に対して単語 w が生起する条件付確率 $P(w|d)$ は，非観測変数 $z \in Z$ を用いると $P(w|d) = \sum_{z \in Z} P(w|z)P(d|z)P(z)$ と表現することができるので，単語と文書の生起確率は非観測変数により分離され，それぞれ非観測変数の数の次元を持つベクトルにより圧縮されて表現される．

$P(w|z)$, $P(d|z)$ は単語 - 文書行列から EM アルゴリズムに基づく以下の各式を繰返し適用することにより求めることができる．

$$P(z|d,w) = \frac{P(w|z)P(d|z)P(z)}{\sum_{z' \in Z} P(w|z)P(d|z)P(z)} \quad (1)$$

$$P(w|z) = \frac{\sum_{d' \in D} n(d',w)P(z|d',w)}{\sum_{w' \in W, d' \in D} n(d',w')P(z|d',w')} \quad (2)$$

$$P(d|z) = \frac{\sum_{w' \in W} n(d,w')P(z|d,w')}{\sum_{w' \in W, d' \in D} n(d',w')P(z|d',w')} \quad (3)$$

$$P(z) = \frac{\sum_{w' \in W, d' \in D} n(d',w'P(z|d',w')}{\sum_{w' \in W, d' \in D} n(d',w')} \quad (4)$$

ここで，上記 PLSI を利用して単語の翻訳確率を求め，CLIR に利用する方法を提案する．上述の方法によれば，z を媒介変数とした各単語の生起確率 $P(w|z)$ を求めることができるので，z を対象とするすべての言語に跨る変数として扱い確率推定が行なえれば翻訳確率を求めることができるはずである．例えば，日本語の単語 w_j が英語の単語 e_j に翻訳される確率 $P(w_e|w_j)$ は次式となる．

$$P(w_e|w_j) = \frac{P(w_e,w_j)}{P(w_j)} = \frac{\sum_{z \in Z} P(w_e|z)P(w_j|z)P(z)}{\sum_{z \in Z} P(w_j|z)P(z)} \quad (5)$$

非観測変数を言語に依存しないようにするには，言語横断型 LSI(CL-LSI) で行なわれていた手法 [2] と同様に，対訳文書組を 1 つの文書と見立て，単語 - 文書行列を得ればよい．文書が言語非依存になっているために対訳となる単語は文書出現分布が似通っていると期待される．そのため，上記単語 - 文書行列に PLSI を施せば言語に依存しない非観測変数を媒介とした単語の条件付確率が得られると考えられる．本稿では上記手法により翻訳情報を得る手法を言語横断 PLSI(Cross Language PLSI, CL-PLSI) と呼ぶ．

3 空間分割型 CL-PLSI

PLSI は EM アルゴリズムに基づく行列の繰返し演算である．文書集合，単語集合，非観測変数の集合を D,W,Z とすると，$(|D|\times|Z|+|W|\times|Z|+|Z|)\times 2$ に比例する空

間計算量が必要となる[1]．また，各確率値が収束するために，ある程度の繰返し回数が必要である．そのため，対訳コーパスの規模が大きくなるほどその処理にかかる時間 / 空間計算量が増大し，一般の計算機では扱えなくなる．

そこで本稿では，対訳コーパスを分野に応じて分割し，各部分コーパスに対して PLSI を行ない，得られた各翻訳確率を合成することで，最終的な翻訳確率を求める手法を提案する．これを空間分割型 CL-PLSI と呼ぶ．

3.1 各部分コーパスにおける確率情報

分野に応じて分割された各部分コーパスから得られる確率情報は，対応する分野における翻訳情報を表現していると考えられる．本稿では，コーパス分割により得られた分野の各々を空間と呼ぶ．PLSI における非観測変数については各空間毎に別の変数が独立して導入される．よって，ある空間 dom における単語の生起確率ならびに文書の生起確率は，dom を明示して，それぞれ，$P(w|z,dom)$, $P(d|z,dom)$ と表すことができる．なお，以下ではすべての空間における非観測変数の集合を改めて Z とし，ある媒介変数の生起確率を $P(z)$ と記すが，z 自身が分野 dom に依存していることに注意されたい．

3.2 分割空間と検索要求の関係

各分割コーパスから得られた翻訳確率を合成するために，各空間と，検索要求との関係を導く．

まず，空間集合 DOM 内のある 1 つの空間 dom は，文書集合であり，検索要求 Q は，分野に依らない 1 つの文書であるとする．Q が与えられた時にある空間 dom が選択される確率 $P(dom|Q)$ は次式で計算される．

$$P(dom|Q) = \frac{P(Q|dom)P(dom)}{\sum_{dom' \in DOM} P(Q|dom')P(dom')} = \frac{P(Q|dom)}{\sum_{dom' \in DOM} P(Q|dom')} \quad (6)$$

$$P(Q|dom) \cong P(q_1 \ldots q_m|dom) \cong \prod_{q_i \in Q} P(q_i|dom) \quad (7)$$

$$P(q_i|dom) = \frac{\sum_{d \in dom} n(d,q_i)}{\sum_{d \in dom} \sum_{w' \in d} n(d,w')} \quad (8)$$

ただし，Q 内の単語を $q_1 \ldots q_m$ とし,$P(dom)$ は，空間 dom に依らず一定であるとする．

ここで，各空間での頻度 $n(d,q_i)$ が 0 となる単語が存在する場合，値 $P(dom|Q)$ が 0 となってしまい，その空間の情報が無視されてしまう．しかし，コーパスの表現する標本空間で 0 であっても実際の確率が 0 でないこともある．そのため，本稿ではグッド・チューリングの推定 [3] を用いて，スムージングを行った．

3.3 翻訳確率の合成

ある空間 dom において語 w_j が語 w_e に翻訳される確率は，次のように表せる．

$$P(w_e|w_j,dom) = \frac{\sum_{z \in Z} P(w_j|z,dom)P(w_e|z,dom)P(z)}{\sum_{z \in Z} P(w_j|z,dom)P(z)} \quad (9)$$

検索要求に従って，この確率を DOM 中の全ての空間に亙り合成すると次式となる．

$$P(w_e|w_j) = \sum_{dom \in DOM} P(w_e|w_j,dom)P(dom|Q) \quad (10)$$

上記合成によって得られた翻訳確率を用いて，翻訳辞書に現れる訳語候補の優先順位を決定する．この方法では，検索要求に依存して空間選択の確率が変わるため，大規模対訳コーパスに PLSI が適用可能となるだけではなく，検索要求の表現する文脈に適した訳語の決定選択ができると考える．

4 評価実験

翻訳精度を直接評価することは難しいので，翻訳された検索要求による検索精度により外的評価を行なう．検索精度の評価には NTCIR2 テストコレクションを用いた [4]．NTCIR2 言語横断タスク用テストコレクションは NTCIR1 で利用されたテストコレクション (学会論文要約集) に対し科学技術文献 (科学研究費に関する文書) を加えて作成されている．NTCIR2 の言語横断タスクにおいては，NTCIR1 で用いられた資源を学習用に利用してよいことになっているため，以下の対訳辞書の抽出，翻訳確率計算は NTCIR2 テストコレクションのサブセットである NTCIR1 テストコレクションを用いて行なっている．

4.1 対訳辞書の作成と文書のインデキシング

NTCIR1 対訳コーパス内で各文書のキーワード部分は，表 1に示すように単語単位でほぼ対訳になっているため，この対応を利用して本実験用の対訳辞書を作成した．

このとき，「トリーイング」，「画像処理」に対する訳語候補にそれぞれ「treeing」「image prossing」を追加する．キーワードを取り出す際には，stemming 処理を行い，誤字に対しては，Levenstein 距離を用いて正しい綴の表現を見つけ修正した．作成した対訳辞書の項目数は日英対訳辞書 243150 件, 英日対訳辞書 245014 件となった．

検索課題及び検索対象文書に対する検索用索引の作成に際しては，上記方法によって収集された対訳辞書に含まれる単語，句にのみ注目し，他の語はすべて無視した．

[1] 2 倍となるのは，EM アルゴリズムによる繰返し計算において，一つ前の値が必要となるためである．

表 1: 対訳コーパスにおけるキーワードの対応例

<KYWDTYPE = "kannji"> トリーイング // 画像処理 // ポリエチレン </KYWD>
<KYWETYPE = "alpha"> treeing // image prossing // polyethylene </KYWE>

4.2 翻訳確率の作成

翻訳確率の計算には，NTCIR1 テストコレクションの日英対訳文書 339483 件のうち，タイトル，キーワード，アブストラクトが対訳となっている 180802 件を用いた．これを文書中の学会フィールドに基づいて 10 の分野に分割した．また，コーパスの規模による翻訳精度の比較のために各々の分野において 20%(35883 件) の量の対訳コーパスも用意した．分割数による精度の比較のために，各分野を更に無作為に分割し，20，30，40 分割のものも用意した．各分割における 1 部分コーパス当たりの文書数を表 2に示す．非観測変数は，各空間毎に 100，200，500 個用意した場合を比較した．

表 2: 1 ドメインあたりの平均文書数

分割数	10	20	30	40	10(20%)
平均文書数	18080	9040	6027	4520	3569

項目の 10(20%) はサイズが 20% での 10 分割である

4.3 検索方法

検索トピック中の多くの語を検索要求とするために TITLE，DESCRIPTION，NARRATIVE フィールドを合わせたものを検索要求とした．その検索要求中の各単語 / 句に対し，対訳辞書を用いて得た対訳候補のうち，翻訳確率が最上位の候補を訳語として翻訳を行なった．そして，TFIDF による語の重み付けならびにベクトル空間法による類似度計算に基づく標準的な情報検索システムを用いて検索を行った．同システムにおいては，フィードバック処理は行なっていない．

4.4 各種ベースライン

提案手法における翻訳が情報検索においてどのように精度向上に寄与するかを調べるために，「空間分割を行なわない PLSI による翻訳手法での検索」，「単語間の相互情報量による共起判定に基づく翻訳手法での検索」，「χ^2 分布による共起判定に基づく翻訳手法での検索」と比較する実験を行った．また，本手法と同様にコーパスを分割する手法である空間分割型 CL-LSI[2] とも比較を行なった．ただし，実験条件の詳細は必ずしも一致していないので注意されたい[2]．上記 4 種類の方法をベースラインとする．

実験で用いた検索システムでの検索精度の上限として「人手によって翻訳した検索要求での検索」の精度も調べた．NTCIR2 のトピックは日本語と英語で用意されているので，人手による検索要求には対応する対訳文書を用いた．

[2] トピック内の TITLE フィールドを検索要求に用いておらず，単語の抽出方法としては C-value に基づく手法を用いている

5 結果

各手法における検索精度を図 1に示す．「日英」は日本語の検索質問を英語に翻訳し，英語の文書を検索する場合を示し，「英日」はその逆である．項目の「同一言語」は、人手での翻訳での検索結果である．

なお，トピック中の DESCRIPTION フィールドをのみを用いた実験 (NTCIR2) も行っているが，すべての条件において他の手法よりも低い値になった．これは DESCRIPTION に現れる単語数が少ないことが原因と考えられる．本実験では索引作成においてコーパス中のキーワードフィールドから作成した対訳辞書を用いているため，検索要求が短いと翻訳可能な語の絶対数が少なくり，検索精度が低下するためであろう．

6 考察

6.1 PLSI に基づく翻訳手法

まず，空間分割を行なわない CL-PLSI の検索精度を調べると，非観測変数の数を 500 程度にすれば，他の空間分割を行なわないベースラインと同等の精度を示すことがわかる．非観測変数の数と精度の関係を見ると，その数を増加させれば更なる精度の向上が期待できるが，より高性能な計算機が必要である．いずれにせよ，PLSI に基づく翻訳確率計算の基本的な有効性が確認された．

6.2 空間分割の効用

表 2によれば，1 部分コーパス当たりの文書数は，20% コーパス全体 (35883 文書) を利用する場合に比べて，コーパス 100% を 40 分割した場合には，1/45 程度であることが分かる．したがって，大量のコーパスを低い空間計算量で扱えることが確認できた．

20% コーパスを用いた実験によれば，空間分割を行なうことで，分割を行なわない場合に比べて大幅に精度が向上することが確認できた．さらに，100% コーパスを用いた実験によれば，各ベースラインと比べ，空間分割を用いた CL-PLSI は，日英，英日ともに各条件で検索精度が高くなった．

以上より，空間分割の効果は計算量の減少に留まらず，検索精度の向上にも寄与することがわかる．その理由を考えてみる．まず，非分割の手法では，検索要求の持つ

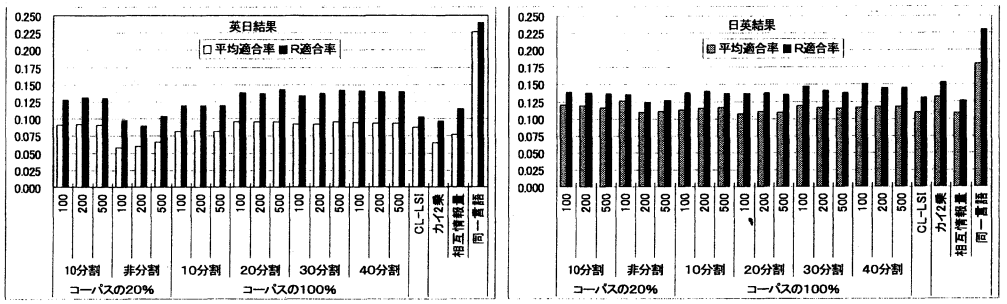

図 1: 各手法による検索精度 (左: 英日翻訳 (日本語文書検索), 右: 日英翻訳 (英語文書検索))

文脈と翻訳確率との関連は考慮されないので, 訳語選択に際して対訳コーパス全体の表す分野への依存性が高くなる. 対訳コーパスが様々な分野を包含する場合には'無難な'訳語候補が選択されることになろう. 一方, 提案手法では, 分割された空間に対して検索要求の持つ文脈に応じて重みづけがなされるので, コーパス全体ではなく, そのうちの一部を選択的に選んでいる. これにより分野に応じた検索に有用な訳語決定が可能であると考えられる.

6.3 非観測変数の数と精度

各条件においての z の数は検索精度にあまり変化が見られず, ほぼ同じ精度となった. 今回の検索実験の規模においては, 100 次元程度以下に圧縮してもよいことが示されたので, 空間分割を行なえば比較的小規模な計算機システムにおいても, 翻訳確率計算が可能であることが予測される.

6.4 各部分コーパスの規模と精度

英日での結果では, 10 分割において, 20% コーパスを用いた場合の方が, 100% コーパスを用いた場合よりも検索精度が高くなった. 一方, 20 分割以上の値で比べると, 100% コーパスを用いた場合の方が検索精度が高くなっている. これは, 1 部分コーパス当たりの文書数が関係していると考えられる. つまり, 各分割文書集合から翻訳確率を生成する際に, 単語数が多すぎると, 非観測変数に対して確率分布が一様なってしまう. それにより翻訳語の決定において翻訳確率の差がなくなり精度が落ちると考えられる.

分割数に関しては, 日英, 英日共に, 大きくなるほど精度が高くなる傾向が見られる. この理由は次のように考えることができよう. まず, 分割数が大きいと, 翻訳確率を求める空間が小さくなるので偏りをもった確率情報が生成される. つまり分野依存性がより高い翻訳確率の計算となる. そして, これらを合成する際に, 検索要求に基づいて重みづけがなされるために, 検索要求の表す文脈に特化した訳語候補を適切に選べるのであろう. つまり, コーパス全体から得た分野にあまり依存しない翻訳確率を使うよりも, 分野依存の局所的な翻訳確率を適切に組み合わせるほうが, 訳語の決定に有効に働くと考えられる.

7 おわりに

本稿では検索要求翻訳型言語横断検索における訳語候補の自動決定において, 大量の対訳コーパスから翻訳情報を収集する手法として空間分割型 CL-PLSI を提案した. そして, NTCIR2 テストコレクションによる評価により, 同手法の有効性を示した. しかし, 単言語検索に比べると検索精度にはまだ開きがある. これについては, 自動生成された翻訳辞書の精度の問題も原因の一つである. 一般の対訳辞書を用いた実験も行なう必要がある. また, 今回の実験では検索対象文書に付随する分野情報を用いているが, そのような情報が利用できない場合の検討も必要である. 今後の課題としたい.

参考文献

[1] Thomas Hofmann. Probabilistic Latent Semantic Indexing. *SIGIR'99*, pp. 50–57, 1999.

[2] 森辰則, 国分智晴, 田中崇. 空間分割型 CL-LSI による大規模言語横断情報検索. 情報処理学会論文：データベース, vol.43 No.SIG 2(TOD 13), pp. 27–36, 3 月 2002.

[3] 北研二, 中村哲, 永田昌明. 音声言語処理 (コーパスに基づくアプローチ). 森北出版株式会社, 1996.

[4] NTCIR Project. NTCIR (NII-NACSIS test collection for IR systems) project web page. http://research.nii.ac.jp/ ntcadm/index-en.html, 2003.