MCCA モデルの日英辞書構築への適用

林 克彦 † 福西 孝章 ††　西田 昌史 † 山本 誠一 †
† 同志社大学大学院　情報工学研究科　†† 同志社大学　工学部　知識工学科

1　はじめに

機械翻訳システムや多言語情報検索システムなど様々な自然言語処理の分野において対訳辞書は重要である．対訳辞書を自動的に作成することが出来れば，辞書の整備やメンテナンスをするコストの削減に繋がる．そのため，これまでに対訳コーパスやコンパラブルコーパスなどの多言語コーパスから，辞書データを獲得する手法の研究がおこなわれてきた．

辞書データを獲得する手法は，対訳コーパスを用いたものとコンパラブルコーパスに大別できる．コンパラブルコーパスから辞書データを獲得するために，各タームの周辺の文脈の類似性を各言語で測定して訳語の推定を行う [1]．コンパラブルコーパスは Web 上の新聞記事や Wikipedia など同一の内容を扱ったサイトを利用すれば，対訳コーパスよりも入手することが容易である [2]．しかし文の対応が取れないことから，対訳辞書データの獲得が比較的難しく，精度を上げることは容易ではない．このような中で，コンパラブルコーパスのみを用いて高い精度を達成した手法 (MCCA:Matching Canonical Correlation Analysis)[3] に着目した．MCCA では字面と文脈の情報を統合し，確率的正準相関分析用いて対訳候補の共起確率を計算することで，英西や英仏などの関係が近い言語間では，対訳辞書データの構築を高い精度で行っている．

そこで本論文では MCCA モデルを利用した日英間における辞書データの構築を考えた．しかし，字面や語順が全く異なる言語対である日本語と英語では，関係の近い言語対と比べると低い精度となるので，その問題を解決するために単語間に共通した知識をヒューリスティック値として割り振ることで精度向上を試みた．以下 2 節では，MCCA モデルの手法について述べる．3 節では日英間に共通した知識について述べる．4 節では実験を通して提案手法の評価を行い，5 節では実験結果を踏まえた考察を行う．

2　MCCA モデル

2.1　概要

MCCA モデルは対訳語を単言語コーパスと呼ばれる単一言語で書かれた文章集合から抽出するために提案された手法である．各単語の情報としてコーパスから得られた文脈情報と字面情報を利用して特徴ベクトルを作成し，正準相関分析と割当問題を反復して解くことで対訳語を見つけ出す．MCCA は次のようにモデル化される．モデルの概念図は図 1 に示す．

単語集合と対応情報

$\boldsymbol{s} = (s_1, s_2, \cdots, s_{ns}) \Leftrightarrow$ 翻訳元の単語集合
$\boldsymbol{t} = (t_1, t_2, \cdots, t_{nt}) \Leftrightarrow$ 翻訳先の単語集合
$i, j \in \boldsymbol{m} \Leftrightarrow s_i, t_j$ が対訳語の関係．(対応情報)

MCCA model

$\boldsymbol{m}$ は一様分布で生成
各対応 $(i, j) \in \boldsymbol{m}$ について
(i, j) が対応するとき
　$z_{i,j} \sim N(0, I_d)$[潜在空間]
　$f_s(s_i) \sim N(W_s z_{i,j}, \Psi_s)$[$s$ 言語ベクトル空間]
　$f_t(t_j) \sim N(W_T z_{i,j}, \Psi_T)$[$t$ 言語ベクトル空間]
i が対応に含まれないとき
　$f_{s_i} \sim N(0, \sigma^2 I_d)$[$s$ 言語ベクトル空間]
j が対応に含まれないとき
　$f_{t_j} \sim N(0, \sigma^2 I_d)$[$t$ 言語ベクトル空間]

2.2　パラメータ推定

MCCA モデルで対訳の対応情報を定義したことにより，次の対数尤度関数を最尤推定で求めることができる．しかし計算量が多くなるため後述する Viterbi-EM で解く．

$$l(\theta) = \log p(\boldsymbol{s}, \boldsymbol{t}; \theta) = \log \sum_{m} p(\boldsymbol{m}, \boldsymbol{s}, \boldsymbol{t}, \theta). \qquad (1)$$

ここで $\theta = (W_S, W_T, \Psi_S, \Psi_T)$ は各言語の特徴ベクトルの正規分布モデルパラメータである．

次に EM アルゴリズムの枠組みを用いて対数尤度関数が高くなるようなパラメータ θ を求める．

- E-step:単語の結びつきが最大となるような対応 $\boldsymbol{m} \in M$ を探す．
- M-step:正準相関分析を使って，特徴ベクトルの分布を表わすパラメータ θ を探す．

E-step では現在のモデルパラメータから単語の対応 $\boldsymbol{m}$ を求め，M-step ではこの $\boldsymbol{m}$ の下で正準相関分析を行い，仮定した正規分布モデルパラメータを更新する．つまり単語対応の取得，正準相関分析，最尤推定

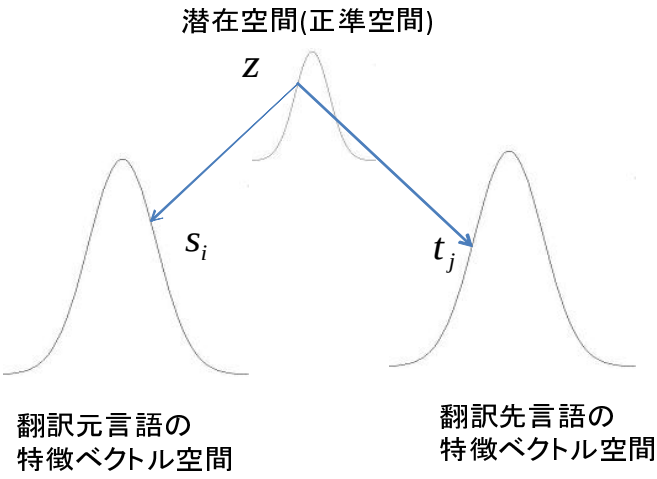

図 1: MCCA のイメージ図

という手順を繰り返し行っている．以下 2.3，2.4 節にて E-step，M-Step の処理を詳細に説明する．

2.3 確率的正準相関分析 (M-step)

与えられた対応 $\boldsymbol{m}$ に対して対数尤度関数を最大にするパラメータを探す問題なので次の式に書き換えることができる．

$$\max_{\theta} \sum_{(i,j)\in\boldsymbol{m}} \log p(s_i, t_j; \theta). \tag{2}$$

この最尤推定は正準相関分析によって求めることができ，各言語の特徴ベクトルの射影した先の相関が最大となるよう固有値ベクトル U_S, U_T を固有値問題として求めることができる．
パラメータ θ は次式で求めることができる．

$$W_S = C_{SS}U_S P^{1/2}, W_T = C_{TT}U_T P^{1/2} \tag{3}$$

$$\Psi_S = C_{SS} - W_S W_S^t, \Psi_T = C_{TT} - W_T W_T^t. \tag{4}$$

$C_{SS} = \frac{1}{|\boldsymbol{m}|} \sum_{(i,j)\in\boldsymbol{m}} f_S(s_i) f_S(s_i)^t$ であり，C_{TT} も C_{SS} と同様に共分散行列の計算で求められる．P は正準相関係数の行列である．

2.4 対応 (E-step)

M-step で求まった θ を使い対応情報を求める．

$$\boldsymbol{m} = \operatorname*{argmax}_{\boldsymbol{m}'} \log p(\boldsymbol{m}', \boldsymbol{s}, \boldsymbol{t}; \theta). \tag{5}$$

この問題は計算量の観点から解くのが難しいので次のように近似する．単語 s_i と t_j の重み (対訳ペアとなる確率) を

$$w_{i,j} = \log p(s_i, t_j; \theta) - \log p(s_i; \theta) - \log p(t_j; \theta) \tag{6}$$

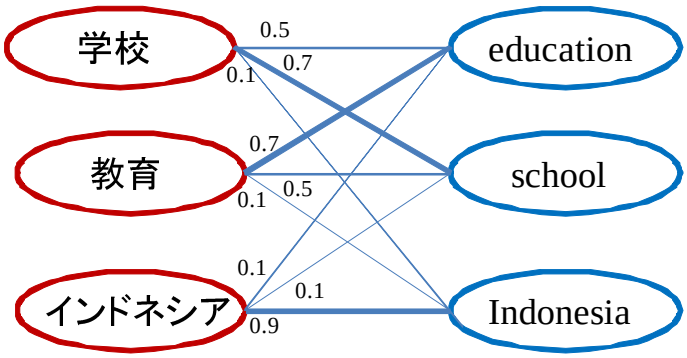

図 2: 割当問題の概念図

とし

$$\log p(\boldsymbol{m}, \boldsymbol{s}, \boldsymbol{t}; \theta) = \sum_{(i,j)\in\boldsymbol{m}} w_{i,j} + C \tag{7}$$

を最大化する割当問題と考える．図 2 にその概念図を示した．

2.5 特徴ベクトル

特徴ベクトルは以下の 2 つから成り立つ．部分文字列の出現頻度に基づいて文字列 (単語) の類似度を計算する文字列カーネルと，ある単語の周辺に共起する単語を記憶する方法がある．関係の近い言語間には同じ語源の言葉が多数存在し，字面が似ていると単語の対応が取れることが多い．そのため文字列カーネルが役立つと考えられる．単語の共起を記憶する手法は，どの言語でもある単語の周辺に出現する単語は同じような物であるという仮説に基づいている．

2.6 ブートストラッピング

EM アルゴリズムのステップを増やすごとに，正準相関分析に使う単語数を逐次的に増やした．

3 MCCA モデルの日英への適用

3.1 予備実験環境

データは日英新聞記事対応データ (5 万文の対訳コーパス)[4] を用い，形態素解析には Chasen，TreeTagger を使用した．またコーパスから評価用の辞書を作成するため GIZA++を使用した．

従来，英語と関係の近い言語，たとえばスペイン語などではまったく同じ字面の言葉が多く存在し，それらの文脈情報などを頼りに対訳語を増やすのが可能であるが，日英間では同じ字面の言葉がごく少数しか存在しない．そのため，GIZA++で学習した正解対訳語データ 300 語を初期ペアとして使用した．

また日英間では文法構造が大きく異なるため，カウントする共起単語を周囲 4 語 (前後 4 単語ずつ)，7 語，10 語，30 語で実験した．なおこの実験では品詞は名詞しか用いていない．また日英間では使用されている

表 1: 従来法による単語の評価値

窓幅	$p_{0.1}$	$p_{0.25}$	BestF 値
4	0.325	0.228	0.238
7	0.415	0.137	0.177
10	0.432	0.184	0.212
30	0.489	0.270	0.259

表 2: 窓幅 30 の良い例と悪い例の対訳単語リスト

正否	英語	日本語
Y	Hiroshima	広島
Y	Pakistan	パキスタン
Y	ASEAN	ASEAN
Y	Vietnam	ベトナム
Y	AIDS	エイズ
N	HIV	感染
N	Shizuoka	千葉
N	Tomiichi	演説
N	Kuwait	侵略
N	Asia-Pacific	太平洋

文字が違うため部分文字列の類似度は対応の情報に含んでいない．予備実験では対訳語の候補となる単語はコーパスの中に現れる 2000 単語とした．評価尺度として再現率，適合率，F 値を用いた．

3.2 従来法の予備実験結果

従来法を日英データで実験した．その結果を表 1 に示す．また改善点を調べるために興味深い 5 単語を表 2 にリストアップした．p_x は再現率が x の時の適合率とする．

表 1 から日英間では部分文字列の情報が使えないため，訳語の抽出精度が低い，一方，日英間の文法の違いを窓幅を増やすことで解消できていることがわかる．表 2 では同じような文脈に登場する単語を捉えてることが分かる．

3.3 提案法

本稿では対訳ペアの共起確率だけを E-step でのスコアとして使うのではなく，ヒューリスティック値として次の 3 つの情報を加えることを考えた．最初の情報として地名，組織名，人名といった固有表現タグの使用した．2 番目の情報として外来語は本来の言葉と似た発音であることに着目して，日本語の読みをアルファベットに変換した文字列と，英単語の文字列との編集距離を利用した．3 番目の情報として意味的な情報を近づけるために接頭辞と語根の情報を使用することにした．そしてこれらの数値を経験的に定め，数式 (6) を以下のように変更した．

$$w_{i,j} = \log p(s_i, t_j; \theta) - \log p(s_i; \theta) - \log p(t_j; \theta) + h(s_i, t_j). \quad (8)$$

$h(s_i, t_j)$ はヒューリスティック値であり，それぞれのヒューリスティック関数が機能するときにスコアが付与される．

本実験で具体的にどのようにヒューリスティック値を扱ったか説明する．まず，固有値表現 (PERSON, LOCATION, ORGANIZATION) が単語間で等しい時，次に，接頭辞接尾辞辞書に各単語の部分文字列が含まれている時，最後に編集距離が閾値よりも小さい時にヒューリスティック関数が 0.05 の値を取りそれ以外の時は 0 の値を取る．

表 3: 3 種類のヒューリスティクス値による実験結果

使用した情報	$p_{0.1}$	$p_{0.25}$	$p_{0.33}$	BestF 値
読み情報	0.567	0.290	0.212	0.268
固有表現	0.541	0.280	0.216	0.264
接頭辞語根辞書	0.554	0.296	0.238	0.276

4 実験

4.1 実験環境

提案法の実験では日英新聞記事対応データを対訳コーパス，非対訳コーパス，コンパラブルコーパスとして 5 万文ずつ使用した．なお非対訳コーパス，コンパラブルコーパスはそれぞれ日英新聞記事データの 5 万文，1 万文が対応していない．固有表現を抽出するため英語は Stanford Named Entity Recognizer[5]，日本語は Chasen を使用した．接頭辞，語根情報を集めるために初期対訳ペア 300 語の中から，接頭辞，接尾辞として考えられるペアを人手で選別し接頭辞，語根辞書として登録した．ここでは接頭辞は 33 組，語根は 13 組使用した．

4.2 実験結果

実験として，提案した 3 種類の情報のうちどれが効果的か，対訳語候補数に応じて適合率がどれほど変化するか，コーパスの状況に応じての性能比較を確認できるようにした．それぞれ表 3，表 4，図 3 に示した．なお本実験の訳語候補数は断りがなければ 1000 単語，窓幅は 30 としている．

表 4: コーパス別性能

コーパスの種類	$p_{0.1}$	$p_{0.25}$	$p_{0.33}$	BestF 値
対訳コーパス	0.718	0.315	0.254	0.287
コンパラブルコーパス	0.647	0.287	0.191	0.267
ノンパラレルコーパス	0.323	0.126	-	0.167

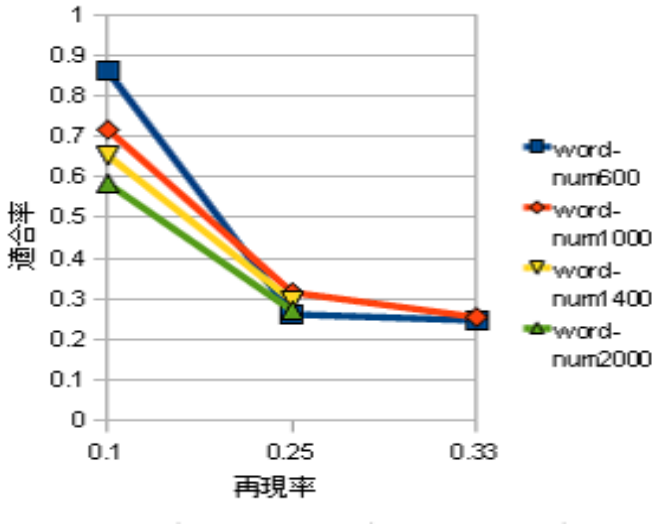

図 3: 訳語候補数，再現率，適合率の関係

表 4，図 3 を見る限り再現値が上がるにつれ適合率が低くなるのは，今回の実験データに対していえる共通の特徴である．スコア上位の単語のペアの適合率が高い傾向にあり，スコアの再現率がある値を超えると一気に適合率が下がる傾向にある．これは上位の単語のペアが固有名詞で，今回提案した情報が対訳ペアを作る際に非常に有効であるということを表している．それと同時に固有名詞ではない名詞の適合率を上げることは難しいということも表している．

また，表 4 より対訳コーパスの適合率とノンパラレルコーパスの適合率が倍以上差がついているのは，日英間の文法の違いをカバーするために窓幅を 30 語取ってしまったことでノイズとなるデータが多くなってしまったということが考えられる．表 3 ではヒューリスティックの種類による結果の差はあまり見られなかった．ちなみに MCCA の枠組みを用いずヒューリスティック値のみを用いてマッチングの実験を行ったが，再現率が 0.1 のときに適合率が 10%を下回る明らかに悪い結果となった．つまりヒューリスティック値のみでは単語の対応を決定できないことが分かった．

5 まとめ

本稿では MCCA モデルを日英間に適用したときに，精度を上げるための手法を提案した．提案手法は一部の単語に有効であり従来法よりも優れた精度を示した．広範囲に渡る固有表現以外の単語に対してはあまり効果がなかった．ヒューリスティック値を最大エントロピーモデルでモデル化し各素性に対する重みを調節することで，自動的に良いパラメータを探し精度を上げられると思うが，今回の実験では初期対訳ペア 300 語以外は使用しない条件だったので，訓練データが少なく実験できなかった．今後は本稿での実験によって抽出した対訳辞書を機械翻訳実験に用いることを検討している．また，フレーズに対して素性を割り当てることで，MCCA を用いたフレーズマッチングに拡張することでフレーズベース統計翻訳の非ドメイン実験による精度向上も検討している．

参考文献

[1] Rapp, R. ., “Identifying Word Translations in Non-Parallel Texts.”, In ACL 1995 ,pp. 320-322.

[2] 宇津呂 武仁, 日野 浩平, 堀内 貴司, 中川 聖一, “日英関連報道記事を用いた訳語対応推定”, In Journal of natural language processing 12(5) pp.43-69 2005/10 言語処理学会.

[3] A.Haghighi, P.Liang, T.B. Kirkpartrick, D.Klein., “Learning Bilingual Lexicons from Monolingual Corpora”, In ACL 2008.

[4] Masao Utiyama and Hitoshi Isahara. “Reliable Measures for Aligning Japanese-English News Articles and Sentences.”, In ACL 2003.

[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning, “ Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”, In ACL 2005, pp. 363-370.