情報処理学会論文誌　数理モデル化と応用　Vol. 4　No. 2　13–22 (Mar. 2011)

同義語情報を用いた確率的単語アライメントモデル

進藤裕之[†1]　藤野昭典[†1]　永田昌明[†1]

2 言語間の教師なし単語アライメント問題に対して，単言語リソースである同義語辞書情報を利用して単語対応付けの精度を向上させる手法を提案する．対訳文には同じ意味を表す様々な表現が用いられるため，同義語情報を利用することでデータスパースネスの問題を解消し単語アライメントの精度向上が期待できる．しかし，単語には多義性があり，ある単語ペアが同義語であるかどうかは文脈に大きく依存する．そこで，我々はトピックモデルを利用して，同義語情報を文脈に応じて学習させる同義語の確率モデルを考案する．さらに，同義語モデルを既存の単語アライメントモデルと同時に学習させる枠組みを提案する．対訳コーパスを用いたアライメント実験の結果，同義語情報を用いない場合や，同義語情報を文脈を考慮せずに同義語情報を利用した場合に比べて，提案手法では高い精度が得られることを確認した．

Word Alignment with Synonym Information

Hiroyuki Shindo,[†1] Akinori Fujino[†1]
and Masaaki Nagata[†1]

We present a novel framework for word alignment that incorporates monolingual synonym knowledge to improve word alignment performance. We expect synonym information is helpful to overcome the data sparseness problem of word alignment since there are various lexical forms represent the same meaning in a bilingual corpus. However, synonym relations depend heavily on context or domain since a word in natural language is ambiguous. We design a synonym probabilistic model with a topic model, which uses synonym information according to the context. Moreover, we propose a word alignment framework that jointly trains our synonym model and conventional bilingual model. The experimental results show that our proposed method obtained better results compared to cases where synonym or context information is not used.

1. はじめに

単語アライメント問題は，対訳コーパスが与えられたときに，異なる言語間における単語の対応関係を推定する問題であり，現在のフレーズベースおよび文法ベースの統計的機械翻訳において最も基本的なタスクの 1 つである．単語アライメントの精度が高ければ，より良い翻訳モデルを構築することができるため，高精度な機械翻訳の実現を期待できる．

これまでに，対訳コーパスの生成モデルに基づく教師なし学習[4),12),15)–17)] や，識別学習に基づく教師あり学習[5),11),14)] など様々な単語アライメント確率モデルが提案されてきた．教師あり学習に基づくアライメント手法は，一定量の人手によりタグづけされた正解単語アライメントが必要となるが，現状では多くの言語対において正解単語アライメントデータを入手することは困難であるため，本論文では教師なし学習に基づく単語アライメント推定手法に焦点を当てる．

統計的機械翻訳で用いられる代表的な教師なし単語アライメントモデルとして，IBM model 1-5 [2)] や HMM [15)] がある．また，これらを改良したり付加情報を加えたりすることで単語アライメントの精度を向上させる手法が数多く提案されている．たとえば，単言語の知識を利用して原言語と目的言語の単語を機能語か内容語かに分類し，機能語どうしまたは内容語どうしを対応させやすくする手法がある[3)]．そのほかにも，単語間における文法的な依存関係は原言語側と目的言語側の双方で保存されている可能性が高いため，そのような文法的な知識を確率モデルに組み込むことでアライメント精度を向上させるものも存在する[8)]．このような言語の文法的知識はアライメントモデルに対して制約条件として機能し，単語対応の過学習を避けてアライメントの精度を向上させることができる．

一方，現在では自然言語処理に有用な多くの語彙的，意味的言語リソースが利用可能である．たとえば WordNet [10)] は 50 カ国以上の言語で構築されているシソーラスであり，同義語，反意語，上位語や下位語など単語の意味的な関係が記述されている．本論文ではそれらの言語資源のうち，同義語の情報を単語アライメントに利用することを考える．同義語は異なる言語の同じ単語へ対応する傾向にあるため，同義語情報を学習時に活用することで単語アライメントの推定精度の向上が期待できる．たとえば，“二酸化炭素” と “炭酸ガス” は同義語のペアであり，同じ英単語 “carbon dioxide” に対応することが期待されるた

†1 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation

め，同義語の情報は正しい単語対応関係を推定するのに有効である．

しかし，一般的に単語の同義関係は単語の表層的な形式ではなく，文脈に大きく左右される．たとえば，“head” と “forefront” は物理的な位置を表す場合はどちらも「先頭」の意味を表す名詞である．また，“head” と “chief” は会社などの話題で用いられる場合はどちらも「集団の長」の意味を表す同義語ペアである．しかし，“forefront” と “chief” はおそらくどの文脈でも同義語ペアではないであろう．以上のことから，同義語の情報を適切に利用するためには，文脈から単語の語義を推定し，複数の同義語候補の中から正しいものを選択して利用する必要がある．

我々は，教師なし単語アライメント学習のために，語彙的，意味的な言語リソースから収集された同義語情報を利用する新たな確率モデルを提案する．提案法では，文脈に応じて同義語情報を利用するために，トピック変数を導入して同義語ペアの確率モデルを構築し，それを対訳コーパスの生成モデルに基づく単語アライメントモデルと統合する．本論文では，単語アライメントモデルとして，HMM にトピックモデルを取り入れた HM-BiTAM [17] を利用する．本手法を英語とフランス語の単語アライメントタスクへ適用し，アライメント精度が向上することを示す．

全体の構成は以下のとおりである．まず 2 章で単語アライメント問題の関連研究について述べ，3 章では，2 言語の対訳コーパスを前提とした単語アライメントモデル HM-BiTAM を概説する．4 章では，我々の提案する同義語ペアの確率モデルと，提案モデルを HM-BiTAM へ利用する手法を説明する．5 章では英語とフランス語の実データを用いて本手法を適用した実験設定および結果を示し，最後に 6 章で結論および将来の展望を述べる．

2. 関 連 研 究

確率的単語アライメントモデルには，対訳コーパスの生成モデルに基づく教師なし学習と，識別学習に基づく教師あり学習によるものがある．教師なし学習による単語アライメントモデルの代表例として，統計翻訳で広く利用されている IBM model [2] や HMM [15] がある．IBM model や HMM 単語アライメントモデルでは，雑音のある通信路モデルを仮定しているため，原言語と目的言語間の予測単語アライメントは 1 対多となる．これを原言語と目的言語を入れ替えることにより双方向で単語アライメントを推定し，単純なヒューリスティクスによって組み合わせることで多対多の予測単語アライメントを得る．LEAF [4] では，モデルをあらかじめ多対多に拡張することで従来の IBM model と比較して精度の高い単語アライメントを獲得している．また，Liang ら [7] は，双方向の HMM アライメントモデルを結合して同時に学習させることで，従来の HMM アライメントモデルよりも高い精度を達成している．

単語アライメントモデルに付加情報を加えて精度を向上させる手法も多く存在する．たとえば，Deng ら [3] は単語のエントロピーを基準にして各単語を機能語か内容語に分類し，機能語どうしや内容語どうしが高い確率で対応するように制約を加えた単語アライメントモデルを提案している．また，Ma ら [8] は，単語間の文法的な依存関係に着目し，それらを特徴量として単語アライメントの学習に利用している．

単語の多義性を解消してアライメントの精度を向上させる手法として，Zhao らの手法 [16),17)] がある．彼らの手法は，対訳文に隠れたトピックがあると仮定し，従来の IBM model や HMM をトピックごとの混合分布に拡張することで高精度な単語アライメントを獲得するものである．

3. 単語アライメントモデル

本章では，トピックモデルと HMM に基づく単語アライメントモデルである HM-BiTAM を概説する．

HM-BiTAM は，2 言語対訳コーパスの生成モデルであり，潜在変数としてトピック z，アライメント a およびトピック分布ベクトル θ を有する．トピックとは，たとえば “科学”，“ニュース”，または “医療” など文の話題を表す変数であり，各対訳文に対して 1 つ割り当てられる．対訳文は K 個のトピックのいずれかに属し，同じトピックに属する対訳文は同じ単語出現確率分布から生成されると仮定する．実際には各対訳文がどのトピックに属するかは不明で，かつ「科学」や「スポーツ」などのラベルも不明である．そこで，トピックモデルでは文の話題の種類に相当する潜在トピックを考え，類似の話題を持つ対訳文が同じ潜在トピックに属するようにモデルを学習させる．潜在トピック変数は $1 \sim K$ までのいずれかの値をとり，各対訳文の話題を特定するのに寄与するため単語の語義曖昧性を解消するのに有効である．以下，潜在トピックを単にトピックと表記する．アライメント変数は，目的言語の各単語がどの原言語の単語と対応関係にあるかを表す変数である．トピック分布ベクトルとは，各トピックがどれくらい出現しやすいかという確率をベクトル形式で表現したものである．以下に，HM-BiTAM の生成過程を示す．

(1)　$\boldsymbol{\theta} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$：　ディリクレ分布に従ってトピック分布ベクトルを生成する．

(2)　各対訳文ペア (E_n, F_n) について

(a)　$z_n \sim \mathrm{Multinomial}(\boldsymbol{\theta})$：　多項分布に従ってトピック z_n を生成する．

(b) 原言語の各単語位置 $i_n = 1, \ldots, I_n$ について
  (i) $e_{i_n} \sim p(e_n | z_n; \boldsymbol{\beta})$：　トピックに依存した原言語のユニグラムモデルに従って原言語の単語 e_{i_n} を生成する．
(c) 目的言語の各単語位置 $j_n = 1, \ldots, J_n$ について
  (i) $a_{j_n} \sim p(a_{j_n} | a_{j_n - 1}; \mathbf{T})$：　一次マルコフモデルに従ってアライメント変数 a_{j_n} を生成する．
  (ii) $f_{j_n} \sim p\left(f_{j_n} \middle| e_{a_{j_n}}, z_n; \mathbf{B}\right)$：　ポジション j_n に対応する原言語の単語 $e_{a_{j_n}}$ とトピック z_n に依存した単語翻訳モデルに従って目的言語の単語 f_{j_n} を生成する．

ただし，アライメント変数 $a_{j_n} = i_n$ は原言語の単語 e_{i_n} と目的言語の単語 f_{j_n} が対応関係にあることを表す．$\boldsymbol{\alpha}$ はトピック分布ベクトル $\boldsymbol{\theta}$ の確率モデルのパラメータ，$\boldsymbol{\beta} = \{\beta_{k,e}\}$ は原言語の単語出現確率分布である．原言語の単語出現確率分布はトピック k に依存し，$\beta_{k,e}$ は $p(e|z=k)$ に相当する．$\mathbf{B} = \{B_{k,e,f}\}$ は，トピック k の下で原言語の単語 e から目的言語の単語 f への単語翻訳確率を表す．すなわち，$B_{k,e,f}$ は $p(f|e, z=k)$ に相当する．$\mathbf{T} = \{T_{i,i'}\}$ は単語の位置 i から i' への遷移確率である．HM-BiTAM では，従来の HMM 単語アライメントモデルと同様に，アライメント変数が一次マルコフモデルに従うと仮定している．**図 1** に HM-BiTAM のグラフィカルモデルを示す．

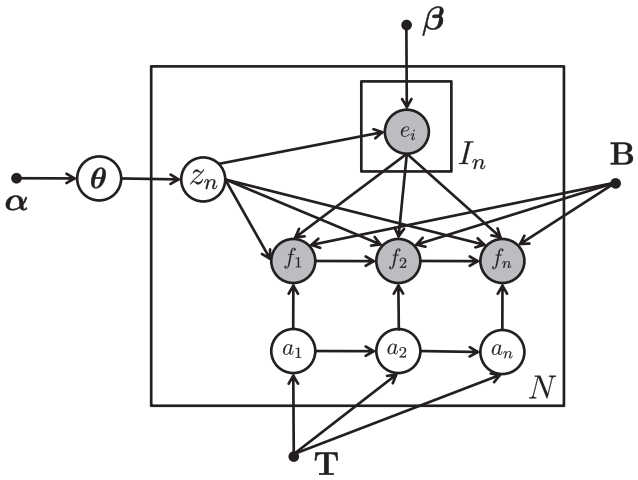

図 1　HM-BiTAM のグラフィカルモデル
Fig. 1　Graphical model of HM-BiTAM.

HM-BiTAM では，対訳文の確率 $p(\mathbf{E}, \mathbf{F})$ を以下のようにモデル化する．

$$p(\mathbf{E}, \mathbf{F}; \boldsymbol{\Phi}) = \sum_{\mathbf{z}} \sum_{\mathbf{a}} \int p(\{E_n\}, \{F_n\}, \mathbf{z}, \mathbf{a}, \boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{1}$$

ただし，$\boldsymbol{\Phi} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{T}, \mathbf{B}\}$ は HM-BiTAM のパラメータセットである．

4. 提案モデル

4.1 同義語データ確率モデル

本節では，提案法で用いる目的言語の同義語ペア $\{f, f'\} = \{(f_m, f'_m)\}_{m=1}^{M}$ の確率モデル $p(\{f, f'\})$ について述べる．自然言語の単語には多義性があるため，言語リソースなどから収集された同義語ペアの同義関係はつねに成立するものではなく，文脈に大きく依存する．

ここで，同義語ペアはある共通の “意味” s という条件の下で独立に生成されると仮定すると，同義語ペアの生成確率を

$$p\left(f, f'\right) \propto \sum_{m}^{M} p(f|s)\, p\left(f'|s\right) p(s) \tag{2}$$

とモデル化できる．

本研究では，同義語情報を単語アライメントの学習に利用するために，原言語の単語を利用して目的言語の同義語をモデル化することを考える．たとえば，目的言語（日本語）の “金星” と “明星” はどちらも原言語（英語）の “Venus” に対応するので，原言語の単語を用いて同義語の語義を表現することができる．しかし，Venus には “女神” の意味もあるように単語には多義性があり，文脈に応じて単語の意味が変わる．そこで，提案手法では，単語の多義性に対処するためにトピック z を導入し，原言語の単語 e とトピック z の組合せ (e, z) で目的言語の同義語の語義が定まると考える．そして，その語義に対して同義語のペアが生成されると仮定する．この仮定の下では，同義語のペア集合の生成確率を以下のようにモデル化できる．

$$p\left(\left\{f, f'\right\}\right) \propto \prod_{(f,f')} \sum_{e,z} p(f|e,z)\, p\left(f'|e,z\right) p(e,z) \tag{3}$$

図 2 に同義語データモデルのグラフィカルモデルを示す．

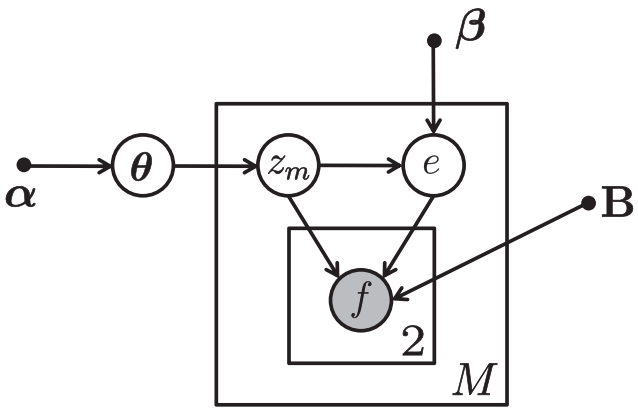

図 2　同義語モデルのグラフィカルモデル
Fig. 2　Graphical model of synonym pair model.

4.2　同義語モデルと統合された単語アライメントモデル

提案手法では，上述の HM-BiTAM と同義語モデルを統合したモデルを用いて単語アライメントを学習させる．この統合は，HM-BiTAM と同義語モデルに含まれる目的言語と原言語の確率モデルおよびトピックの確率モデル $p(f|e,z)$，$p(e|z)$，$p(z)$ のパラメータを互いに共有することにより実現する．具体的には，式 (3) の同義語モデル $p(\{f,f'\})$ を，HM-BiTAM と同じパラメータセット $\mathbf{\Phi}=\{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{T},\mathbf{B}\}$ を用いて以下のようにパラメータ化すればよい．

$$p(f|e,z=k) \equiv p(f|e,z=k;\mathbf{B}) = B_{k,e,f}, \tag{4}$$

$$\begin{aligned} p(e,z=k) &\equiv p(e|z=k;\boldsymbol{\beta})\,p(z=k;\boldsymbol{\alpha}) \\ &= \beta_{k,e}\int \theta_k p(\boldsymbol{\theta};\boldsymbol{\alpha})\,d\boldsymbol{\theta}. \end{aligned} \tag{5}$$

したがって，同義語ペアの確率は以下のように書ける．

$$\begin{aligned} p\left(f,f';\mathbf{\Phi}\right) &\propto \sum_{e,z}\int p(f|e,z)\,p\left(f'|e,z\right)p(e|z)\,p(z|\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta} \\ &= \sum_{k,e}\beta_{k,e}B_{k,e,f}B_{k,e,f'}\int p(z=k|\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta} \end{aligned}$$

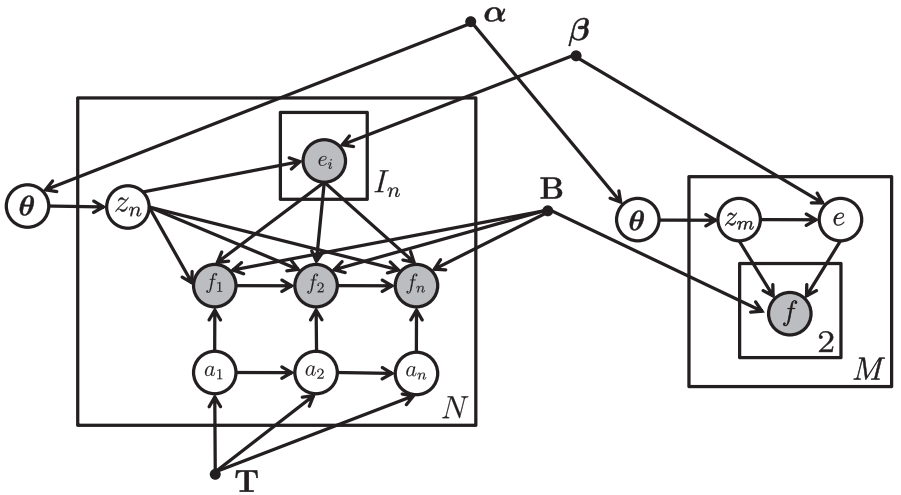

図 3　パラメータを共有した提案手法のグラフィカルモデル
Fig. 3　Graphical model of proposed method.

$$= \frac{1}{\sum_{k'}\alpha_{k'}}\sum_{k,e}\alpha_k\beta_{k,e}B_{k,e,f}B_{k,e,f'} \tag{6}$$

同義語ペア $\{f,f'\}$ の生成過程は，以下のようになる．

(1)　$\boldsymbol{\theta}\sim\mathrm{Dirichlet}(\boldsymbol{\alpha})$：　ディリクレ分布に従ってトピック分布ベクトルを生成する．

(2)　各同義語ペア (f_m,f'_m) について

- (a)　$z_m\sim\mathrm{Multinomial}(\boldsymbol{\theta})$：　多項分布に従ってトピック z_m を生成する．
- (b)　$e_m\sim p(e_m|z_m;\boldsymbol{\beta})$：　トピックに依存した原言語のユニグラムモデルに従って原言語の単語 e_m を生成する．
- (c)　$f_m\sim p(f_m|e_m,z_m;\mathbf{B})$：　トピックと原言語の単語に依存した単語翻訳モデルに従って目的言語の単語 f_m を生成する．
- (d)　$f'_m\sim p(f_m|e_m,z_m;\mathbf{B})$：　トピックと原言語の単語に依存した単語翻訳モデルに従って目的言語の単語 f'_m を生成する．

図 3 に HM-BiTAM と統合した同義語モデルのグラフィカルモデルを示す．提案手法では，HM-BiTAM の対数周辺尤度と同義語モデルの対数周辺尤度を同時に最大化させる $\mathbf{\Phi}$ を推定値 $\hat{\mathbf{\Phi}}$ とする．

$$\hat{\mathbf{\Phi}} = \underset{\mathbf{\Phi}}{\operatorname{argmax}} \left\{ \log p\left(\mathbf{E}, \mathbf{F}; \mathbf{\Phi}\right) + \zeta \log p\left(\left\{f, f'\right\}; \mathbf{\Phi}\right) \right\} \tag{7}$$

ただし，ζ は同義語モデルの重みを調節するハイパーパラメータである．

4.3 学　　習

本節では，提案手法の学習法について説明する．式 (7) は解析的に解くことができないため，変分 EM アルゴリズム[1] を用いて数値的に $\hat{\mathbf{\Phi}}$ の学習を行う．変分 EM アルゴリズムでは，変分近似によって対数周辺尤度の下限を計算し，繰返し計算によってパラメータの最適解を探索する．まず，式 (7) の第 1 項に新たな分布 q を導入し，Jensen の不等式を利用すると周辺尤度の下限 l_b を得る．

$$\begin{aligned} \log p\left(\mathbf{E}, \mathbf{F}; \mathbf{\Phi}\right) &= \log \sum_{\mathbf{z}} \sum_{\mathbf{a}} \int p\left(\mathbf{E}, \mathbf{F}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{a}; \mathbf{\Phi}\right) d\boldsymbol{\theta} \\ &\geq \sum_{\mathbf{z}} \sum_{\mathbf{a}} \int q\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a}\right) \log \frac{p\left(\mathbf{E}, \mathbf{F}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{a}; \mathbf{\Phi}\right)}{q\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a}\right)} d\boldsymbol{\theta} \\ &\equiv l_b \end{aligned} \tag{8}$$

$\langle \cdot \rangle_q$ は関数 q の下での期待値を表す．変分 EM アルゴリズムでは，下限 l_b を最大化させるモデルパラメータ $\mathbf{\Phi}$ の値と $q\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a}\right)$ の分布を一方を固定して交互に繰返し計算することで $\mathbf{\Phi}$ の推定値を求める．このとき，$\log p\left(\mathbf{E}, \mathbf{F}; \mathbf{\Phi}\right)$ と l_b には以下の関係がある．

$$\log p\left(\mathbf{F}, \mathbf{E}; \mathbf{\Phi}\right) = l_b + \mathrm{KL}\left(q\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a}\right) \| p\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a} \,|\mathbf{E}, \mathbf{F}; \mathbf{\Phi}\right)\right) \tag{9}$$

ここで，$\mathrm{KL}\left(q \| p\right)$ は分布 q と p の Kullback-Leibler 距離である．

$\mathbf{\Phi}$ 固定の下で l_b を最大化させる q は $p\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a} \,|\mathbf{E}, \mathbf{F}\right)$ である．しかし，$p\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a} \,|\mathbf{E}, \mathbf{F}\right)$ を直接計算することは困難であるため，関数 q を新たなパラメータ $\boldsymbol{\phi}$，$\boldsymbol{\lambda}$，$\boldsymbol{\gamma}$ を用いて $q\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a}\right) = q\left(\mathbf{z} \,|\boldsymbol{\phi}\right) q\left(\mathbf{a} \,|\boldsymbol{\lambda}\right) q\left(\boldsymbol{\theta} \,|\boldsymbol{\gamma}\right)$ と変分近似する．$q\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a}\right)$ は真の事後分布 $p\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a} \,|\mathbf{E}, \mathbf{F}\right)$ をできるだけよく近似するものが望ましいので，$q\left(\mathbf{z}; \boldsymbol{\phi}\right)$，$q\left(\mathbf{a}; \boldsymbol{\lambda}\right)$，$q\left(\boldsymbol{\theta}; \boldsymbol{\gamma}\right)$ をそれぞれ，元の $\mathbf{z}$, $\mathbf{a}$，$\boldsymbol{\theta}$ の分布と同じ多項分布，一次マルコフモデル，ディリクレ分布で与える．すなわち，

$$q\left(\mathbf{z}; \boldsymbol{\phi}\right) = \prod_{n=1}^{N} p\left(z_n; \boldsymbol{\phi}\right) \tag{10}$$

$$q\left(\mathbf{a}; \boldsymbol{\lambda}\right) = \prod_{n=1}^{N} \prod_{j_n=1}^{J_n} p\left(a_{j_n} \,|a_{j_n - 1}; \boldsymbol{\lambda}\right) \tag{11}$$

$$q\left(\boldsymbol{\theta}; \boldsymbol{\gamma}\right) \propto \prod_{k=1}^{K} \theta_k^{\gamma_k - 1} \tag{12}$$

式 (7) の第 2 項も同様に，Jensen の不等式を利用して下限 l_m を求めると以下のようになる．

$$\begin{aligned} \log p\left(\left\{f, f'\right\}; \mathbf{\Phi}\right) &= \sum_{m}^{M} \log p\left(f_m, f'_m; \mathbf{\Phi}\right) \\ &\geq \sum_{m}^{M} \left(\sum_{k,e} p\left(k, e \,|\mathbf{f}_m\right) \log \frac{p\left(\mathbf{f}_m, k, e\right)}{p\left(k, e \,|\mathbf{f}_m\right)} - \log \sum_{k'} \alpha_{k'} \right) \\ &\equiv l_m \end{aligned} \tag{13}$$

ただし，$\mathbf{f}_m = \left(f_m, f'_m\right)$ である．

第 1 項と第 2 項の重み付き和 $l_b + \zeta l_m$ を各パラメータで偏微分することにより，式 (7) の下限を最大化させるパラメータを求めることができる．ただし，ディリクレ分布のパラメータ α は解析的に解くことができないため，gradient descent 法[6] を用いて更新する．結局，変分 EM アルゴリズムの E ステップおよび M ステップは以下のようになる．詳細は付録参照のこと．

4.3.1 E-step

$$\widehat{\gamma}_k = \alpha_k + \sum_{n} \phi_{n,k} \tag{14}$$

$$\begin{aligned} \widehat{\phi}_{n,k} \propto \exp\left(\Psi\left(\gamma_k\right) - \Psi\left(\sum_{k'} \gamma_{k'} \right) \right) \cdot \exp\left(\sum_{j,i} \lambda_{n,j,i} \log \beta_{k, e_{i_n}} \right) \\ \cdot \ \exp\left(\sum_{i,j} \lambda_{n,j,i} \log B_{k, e_{i_n}, f_{j_n}} \right) \end{aligned} \tag{15}$$

$$\begin{aligned} \widehat{\lambda}_{n,j,i} \propto \exp\left(\sum_{i'} \lambda_{n,j-1,i'} \log T_{i,i'} \right) \cdot \exp\left(\sum_{i''} \lambda_{n,j+1,i''} \log T_{i'',i} \right) \\ \cdot \ \exp\left(\sum_{k} \phi_{n,k} \log B_{k, e_{i_n}, f_{j_n}} \right) \cdot \exp\left(\sum_{k} \phi_{n,k} \log \beta_{k, e_{i_n}} \right) \end{aligned} \tag{16}$$

4.3.2 M-step

$$\hat{\beta}_{k,e} \propto \sum_n \sum_j \sum_i \delta\left(e_{i_n}, e\right) \lambda_{n,j,i} \phi_{n,k} + \zeta \sum_m^M p\left(k, e \left| f_m, f'_m \right.\right) \tag{17}$$

$$\hat{B}_{k,e,f} \propto \sum_n \sum_j \sum_i \delta\left(f_{j_n}, f\right) \delta\left(e_{i_n}, e\right) \lambda_{n,j,i} \phi_{n,k} + \zeta \sum_m^M \left(\delta\left(f_m, f\right) + \delta\left(f'_m, f\right)\right) p\left(k, e \left| f_m, f'_m \right.\right) \tag{18}$$

$$\hat{T}_{i',i} \propto \sum_n \sum_j \lambda_{n,j,i'} \lambda_{n,j-1,i} \tag{19}$$

ただし，

$$p\left(k, e \left| f_m, f'_m \right.\right) = \frac{\alpha_k \beta_{k,e} B_{k,e,f_m} B_{k,e,f'_m}}{\sum_k \sum_e \alpha_k \beta_{k,e} B_{k,e,f_m} B_{k,e,f'_m}} \tag{20}$$

また，$\delta(a,b)$ は $a = b$ のとき 1，それ以外は 0 を取る指示関数であり，Ψ はディガンマ関数である．

E ステップでは，確率分布 q の各パラメータ $\boldsymbol{\phi}$，$\boldsymbol{\lambda}$，$\boldsymbol{\gamma}$ を更新することにより，真の事後分布 $p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{a} | \mathbf{E}, \mathbf{F})$ を近似する．次に，M ステップでは近似された事後分布 q の下で残りのパラメータを更新する．以上の 2 つのステップを収束するまで繰り返すことで，逐次的に周辺尤度の最大化を実行する．

学習された最適なパラメータを用いて，各対訳文における最終的な予測アライメントは以下のように計算される．

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \; q\left(\mathbf{a} \left| \hat{\boldsymbol{\lambda}} \right.\right) \tag{21}$$

5. 実　　験

5.1 実験設定

提案手法の評価を行うため，Hansards データセット[9] を用いて単語アライメントの実験を行った．Hansards データセットは，英語とフランス語の 2 言語対訳コーパスで，規模は約 2000 万文である．本データセットのうち，447 文は人手による正解単語アライメント情報が付与されている．我々は，この 447 文の中からランダムに選択した 100 文を開発用データセット，残りの 347 文を評価用テストデータとした．開発用データセットは重み ζ を最適化するために用いた．トレーニングデータは以下のように構成した．まず，評価用テストデータから正解単語アライメント情報を削除した 347 文の対訳文と，人手で正解の付与されていない残りの対訳文の中からランダムに 10 k，50 k，100 k の文を選択し，これを混合したものをトレーニングデータとした．したがって，トレーニングデータには必ず評価用テストデータの対訳文が含まれている．このトレーニングデータを用いて単語アライメントの教師なし学習を行い，347 文の評価用対訳文の推定結果と正解を比較して単語アライメントの精度を評価した．したがって，トレーニングに正解単語アライメントの情報はいっさい使用していない．

英語およびフランス語の同義語辞書は，それぞれ WordNet 2.1[10] および Wolf 0.1.4[13] から収集した．WordNet は英語の意味的な概念を扱う言語リソースで，単語が synset と呼ばれる同義語のグループに分類されており．同義語ペアのデータを得ることができる．WOLF は WordNet やその他の各種言語リソースから構築されたフランス語の WordNet である．我々は，これらのリソースから得られた同義語ペアのうち，いずれの単語もトレーニングデータ中に含まれる場合のみ学習に使用した．

我々は，GIZA++ 1.0.3[12]，HM-BiTAM および提案手法で単語アライメント精度の比較を行った．ただし，HM-BiTAM は我々が独自に実装したものを用いている．GIZA++ は，IBM model-4 による単語アライメントであり，HM-BiTAM は式 (7) で $\zeta = 0$ に相当する．

IBM model，HM-BiTAM や提案手法のような雑音のある通信路モデルに基づく単語アライメントモデルでは，原言語と目的言語を入れ替えることにより 2 方向の単語アライメント結果が得られる．本実験では，英語を原言語，フランス語を目的言語とした場合と，原言語をフランス語，目的言語を英語とした場合の 2 方向の結果を “GROW” ヒューリスティクス[12),16)] を用いて統合し，1 方向よりも高精度かつロバストな予測単語アライメントを得た．

本データセットには，S（sure）または P（probable）の 2 種類の正解単語アライメントのラベルが人手によって付与されている．S アライメントは，確実に対応関係である単語ペアに対して付与されたアライメント情報であり，P アライメントはそれ以外の（不確実な）アライメントである．予測した単語アライメントの精度は，Precision，Recall，F-measure，AER を用いて評価した．これらの尺度は，単語アライメント問題で標準的に用いられる評価基準である[12]．

$$\text{Precision} = \frac{|\text{A} \cap \text{P}|}{|\text{A}|} \tag{22}$$

$$\text{Recall} = \frac{|\text{A} \cap \text{S}|}{|\text{S}|} \tag{23}$$

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{24}$$

$$\text{AER} = 1 - \frac{|\text{A} \cap \text{S}| + |\text{A} \cap \text{P}|}{|\text{A}| + |\text{S}|} \tag{25}$$

ただし，|A|，|S| はそれぞれ予測アライメントの数，S アライメントの数を表す．また，|A ∩ S| は，予測アライメントと S アライメントに共通に含まれているアライメントの数であり，|A ∩ P| も同様である．

5.2 結　　果

表 1 は，10 k，50 k，100 k のトレーニングデータで学習された単語アライメントの精度を示している．提案手法が F 値と AER で最も良い性能を示している．この結果から，我々の同義語データを利用した単語アライメント手法は，精度を向上させることに効果的であることが分かる．

前述のように，我々の主なアイディアは，トピック変数と，異なる言語の単語を用いて同義語ペアの精密なモデル化を行うというものである．言語リソースから収集された同義語ペアが同義語であるかは，文脈に大きく依存する．この問題に対処するため，我々はトピック変数を導入した同義語ペアのモデル化を行い，単語の語義曖昧性を解消しつつ対訳文の文脈に応じた同義語ペアを単語アライメント学習に利用可能となった．本モデルの効果を検証するために，我々は同義語データを単語アライメント学習に利用する単純なヒューリスティクス（SRH: Synonym Replacement Heuristics）を用いてテストを行った．SRH は，同義語データ中に含まれる同義語ペアを利用し，トレーニングデータ中の単語を片方の同義語に置き換えるというヒューリスティクスである．たとえば，単語 A と単語 B が同義語ペアである場合，トレーニングデータ中のすべての単語 B は単語 A に置換される．SRH では，“head” のように複数の同義語ペアを持つ単語は，どの同義語ペアに置き換えられるかはランダムに決定される．したがって，文脈に応じて正しく単語が同義語に置換された場合，単語アライメントの精度向上が期待できるが，そうでない場合は逆に精度を悪化させる恐れがある．**表 2** に示すように，SRH によりトレーニングセットにおける英語とフランス語の語彙数は期待どおり大きく減少した．この際に用いた各トレーニングデータの同義語数を**表 3** に示す．

SRH を実行した後，GIZA++および HM-BiTAM の単語アライメント精度を検証した．

表 1　単語アライメント精度の比較．トレーニングデータのサイズはそれぞれ (a) 10 k，(b) 50 k，(c) 100 k
Table 1　Comparison of word alignment accuracy. The best results are indicated in bold type. The training data set sizes are (a) 10 k, (b) 50 k, (c) 100 k.

10 k		Precision	Recall	F-measure	AER
GIZA++	standard	0.856	0.718	0.781	0.207
	with SRH	0.874	0.720	0.789	0.198
HM-BiTAM	standard	0.869	0.788	0.826	0.169
	with SRH	0.884	0.790	0.834	0.160
Proposed		**0.941**	**0.808**	**0.870**	**0.123**

(a)

50 k		Precision	Recall	F-measure	AER
GIZA++	standard	0.905	0.770	0.832	0.156
	with SRH	0.903	0.759	0.825	0.164
HM-BiTAM	standard	0.901	0.814	0.855	0.140
	with SRH	0.899	0.808	0.853	0.145
Proposed		**0.947**	**0.824**	**0.881**	**0.112**

(b)

100 k		Precision	Recall	F-measure	AER
GIZA++	standard	0.925	0.791	0.853	0.136
	with SRH	**0.934**	0.803	0.864	0.126
HM-BiTAM	standard	0.898	0.851	0.874	0.124
	with SRH	0.909	0.860	0.879	0.114
Proposed		0.927	**0.862**	**0.893**	**0.103**

(c)

表 2　10 k，50 k，100 k のトレーニングデータの語彙数
Table 2　The number of vocabularies in the 10 k, 50 k and 100 k data sets.

# vocabularies		10 k	50 k	100 k
English	standard	8,578	16,924	22,817
	with SRH	5,435	7,235	13,978
French	standard	10,791	21,872	30,294
	with SRH	9,737	20,077	27,970

SRH は，10 k と 100 k のデータセットでは精度が若干向上したが，50 k のデータセットでは精度が悪化した．これは，同義語ペアの情報を誤った文脈で利用してしまったことにより，単語アライメントの精度が悪化してしまった影響であると考えられる．

表 3　10 k，50 k，100 k のトレーニングデータの同義語ペア数
Table 3　The number of synonym pairs in the 10 k, 50 k and 100 k data sets.

# synonyms	10 k	50 k	100 k
English	7,756	17,273	23,187
French	1,677	2,524	2,980

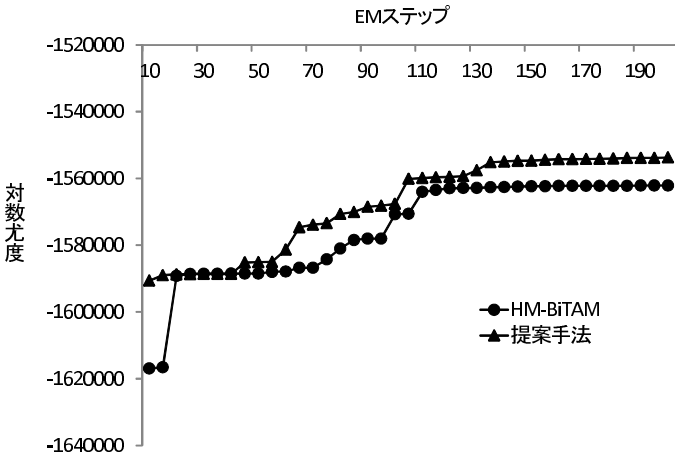

図 4　対訳コーパスの対数尤度の比較
Fig. 4　Comparison of log likelihood in the bilingual corpus.

図 4 は，EM ステップと対訳コーパスの周辺尤度との関係を，HM-BiTAM と提案手法とで比較したものである．トレーニングデータは 10 k を用いて，フランス語を原言語，英語を目的言語とした．また，HM-BiTAM と提案手法は同じ初期値に設定し，5 回試行したときの平均尤度である．

提案手法は，HM-BiTAM よりも少ないステップで尤度の高い解へ到達できることが分かる．これは，提案手法が対訳コーパスだけでなく同義語の情報も利用して，より良い単語翻訳確率やトピックの生成確率を推定できるためである．提案手法は対訳コーパスと同義語データの同時周辺尤度を最大化させているが，対訳コーパスの周辺尤度のみに着目しても提案手法は HM-BiTAM より尤度の高い解へ収束する傾向がある．

参考までに，文献 9) によれば，提案法で 10 万の対訳文を学習に用いた場合（AER: 0.103）よりも低い AER を達成していたのは，約 500 万，1000 万の対訳文を学習に用いた GIZA++ のシステム（AER：0.0893，0.0853）と，英語の構文解析情報および英単語の分布類似度を付加情報として利用したシステム（AER: 0.0571）であった．表 1 より，1 万〜10 万文の対訳文データを用いた実験において本手法は GIZA++より高精度であり，本手法が大規模データを扱うことができるようになれば，さらなるアライメントの精度向上が期待できる．したがって，提案手法のパラメータ学習を効率化させることが今後の課題となる．また，英語の構文解析情報と英単語の分布類似度を利用する方法と，同義語情報を利用する本手法では異なる情報を用いており，両手法を組み合わせることでさらなるアラインメントの精度向上が期待できる．この組合せを実現することも今後の課題である．

6. 結　　論

我々は，同義語の情報を教師なし単語アライメントモデルへ利用する枠組みを提案した．ある単語ペアが同義語であるかどうかは文脈に大きく依存するため，我々はトピックモデルを利用して文脈を特定し，単語の語義曖昧性を解消しながら同義語ペアの確率モデルを考案した．また，同義語モデルのパラメータを 2 言語の単語アライメントモデルと同時に学習することで，同義語の情報を単語アライメントへ利用する枠組みを提案した．我々の手法は，2 言語の対訳情報と単言語の同義語情報を効率的に利用し，教師なし単語アライメントの精度を向上させた．今後は，本手法を異なる言語間での単語アライメント問題へ適用することや，統計的機械翻訳へ応用することが考えられる．

参 考 文 献

1) Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M. and West, M.: The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures, *Bayesian Statistics 7: Proc. 7th Valencia International Meeting*, June 2-6, 2002, p.453, Oxford University Press, USA (2003).
2) Brown, P.F., Della Pietra, V.J., Della Pietra, S.A. and Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation, *Computational linguistics*, Vol.19, No.2, pp.263–311 (1993).
3) Deng, Y. and Gao, Y.: Guiding Statistical Word Alignment Models With Prior Knowledge, *Proc. 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp.1–8, Association for Computational Linguistics (2007).
4) Fraser, A. and Marcu, D.: Getting the structure right for word alignment: LEAF,

Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (*EMNLP-CoNLL*), Prague, Czech Republic, pp.51–60, Association for Computational Linguistics (2007).

5) Haghighi, A., Blitzer, J., DeNero, J. and Klein, D.: Better Word Alignments with Supervised ITG Models, *Proc. Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore, pp.923–931, Association for Computational Linguistics (2009).

6) Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. and Haussler, D.: Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology, *Computer Applications in the Biosciences*, Vol.12 (1996).

7) Liang, P., Taskar, B. and Klein, D.: Alignment by agreement, *Proc. main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp.104–111, Association for Computational Linguistics (2006).

8) Ma, Y., Ozdowska, S., Sun, Y. and Way, A.: Improving Word Alignment Using Syntactic Dependencies, *Proc. ACL-08: HLT 2nd Workshop on Syntax and Structure in Statistical Translation* (*SSST-2*), Columbus, Ohio, pp.69–77, Association for Computational Linguistics (2008).

9) Mihalcea, R. and Pedersen, T.: An evaluation exercise for word alignment, *Proc. HLT-NAACL 2003 Workshop on building and using parallel texts: Data driven machine translation and beyond-Volume 3*, p.10, Association for Computational Linguistics (2003).

10) Miller, G.A.: WordNet: A lexical database for English, *Comm. ACM*, Vol.38, No.11, p.41 (1995).

11) Moore, R.C., Yih, W.-t. and Bode, A.: Improved Discriminative Bilingual Word Alignment, *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.513–520, Association for Computational Linguistics (2006).

12) Och, F.J. and Ney, H.: A systematic comparison of various statistical alignment models, *Computational Linguistics*, Vol.29, No.1, pp.19–51 (2003).

13) Sagot, B. and Fiser, D.: Building a free French wordnet from multilingual resources, *Proc. Ontolex* (2008).

14) Taskar, B., Simon, L.-J. and Dan, K.: A Discriminative Matching Approach to Word Alignment, *Proc. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, pp.73–80, Association for Computational Linguistics (2005).

15) Vogel, S., Ney, H. and Tillmann, C.: HMM-based word alignment in statistical translation, *Proc. 16th Conference on Computational Linguistics-Volume 2*, pp.836–841, Association for Computational Linguistics Morristown, NJ, USA (1996).

16) Zhao, B. and Xing, E.P.: BiTAM: Bilingual topic admixture models for word alignment, *Proc. COLING/ACL on Main Conference Poster Sessions*, p.976, Association for Computational Linguistics (2006).

17) Zhao, B. and Xing, E.P.: HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation, *Advances in Neural Information Processing Systems 20*, Cambridge, MA, pp.1689–1696, MIT Press (2008).

付　　録

対訳コーパスの周辺尤度の下限は，変分近似を用いると以下のように計算される．

$$
\begin{aligned}
\log p(\mathbf{F},\mathbf{E}) &= \log \sum_{\mathbf{z}} \sum_{\mathbf{a}} \int p(\mathbf{F},\mathbf{E},\boldsymbol{\theta},\mathbf{z},\mathbf{a})\, d\boldsymbol{\theta} \\
&\geq \sum_{\mathbf{z}} \sum_{\mathbf{a}} \int q(\boldsymbol{\theta},\mathbf{z},\mathbf{a}) \log \frac{p(\mathbf{F},\mathbf{E},\boldsymbol{\theta},\mathbf{z},\mathbf{a})}{q(\boldsymbol{\theta},\mathbf{z},\mathbf{a})} d\boldsymbol{\theta} \\
&\approx \sum_{\mathbf{z}} \sum_{\mathbf{a}} q(\mathbf{z};\boldsymbol{\phi})\, q(\mathbf{a};\boldsymbol{\lambda}) \int q(\boldsymbol{\theta};\boldsymbol{\gamma}) \left(\log \frac{p(\mathbf{F},\mathbf{E},\boldsymbol{\theta},\mathbf{z},\mathbf{a})}{q(\mathbf{z};\boldsymbol{\phi})\, q(\mathbf{a};\boldsymbol{\lambda})\, q(\boldsymbol{\theta};\boldsymbol{\gamma})} \right) d\boldsymbol{\theta} \\
&= \langle \log p(\boldsymbol{\theta};\boldsymbol{\alpha}) \rangle_q + \langle \log p(\mathbf{z}\,|\boldsymbol{\theta}) \rangle_q + \langle \log p(\mathbf{E}\,|\mathbf{z};\boldsymbol{\beta}) \rangle_q \\
&+ \langle \log p(\mathbf{a};\mathbf{T}) \rangle_q + \langle \log p(\mathbf{F}\,|\mathbf{a},\mathbf{z},\mathbf{E};\mathbf{B}) \rangle_q - \langle \log q(\boldsymbol{\theta};\boldsymbol{\gamma}) \rangle_q \\
&- \langle \log q(\mathbf{z};\boldsymbol{\phi}) \rangle_q - \langle \log q(\mathbf{a};\boldsymbol{\lambda}) \rangle_q \\
&= \log \Gamma\left(\sum_k \alpha_k \right) - \sum_k \log \Gamma(\alpha_k) \\
&+ \sum_k (\alpha_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{k'} \gamma_{k'} \right) \right) \\
&+ \sum_n \sum_k \phi_{n,k} \left(\Psi(\gamma_k) - \Psi\left(\sum_{k'} \gamma_{k'} \right) \right) \\
&+ \sum_n \sum_k \sum_j \sum_i \lambda_{n,j,i} \phi_{n,k} \log \beta_{k,e_{i_n}} \\
&+ \sum_n \sum_j \sum_i \lambda_{n,j,i} \sum_{i'} \lambda_{n,j-1,i'} \log T_{i,i'} \\
&+ \sum_n \sum_j \sum_i \lambda_{n,j,i} \sum_{f \in V_f} \sum_{e \in V_e} \delta(f_{j_n}, f)\, \delta(e_{i_n}, e) \sum_k \phi_{n,k} \log B_{k,e,f} \\
&- \log \Gamma\left(\sum_k \gamma_k \right) + \sum_k \log \Gamma(\gamma_k) - \sum_k (\gamma_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{k'} \gamma_{k'} \right) \right)
\end{aligned}
$$

$$-\sum_n \sum_k \phi_{n,k} \log \phi_{n,k} - \sum_n \sum_j \sum_i \lambda_{n,j,i} \log \lambda_{n,j,i}$$

対訳コーパスと同義語データの周辺尤度の下限を各パラメータで偏微分することにより，変分 EM アルゴリズムの更新式が導出される．

(平成 22 年 8 月 31 日受付)
(平成 22 年 10 月 19 日再受付)
(平成 22 年 11 月 11 日採録)

進藤　裕之（正会員）

2009 年早稲田大学大学院先進理工学研究科修士課程修了．同年 NTT 入社．統計的自然言語処理の研究に従事．現在，NTT コミュニケーション科学基礎研究所研究員．ACL 会員．

藤野　昭典（正会員）

1995 年京都大学工学部精密工学科卒業．1997 年同大学大学院修士課程修了．2009 年同大学院博士課程修了．博士（情報学）．1997 年 NTT 入社．機械学習，テキスト処理等の研究に従事．現在，NTT コミュニケーション科学基礎研究所研究主任．電子情報通信学会 PRMU 研究奨励賞（2004 年度），FIT 論文賞（2005 年）等受賞．電子情報通信学会，IEEE 各会員．

永田　昌明（正会員）

1987 年京都大学大学院工学研究科修士課程修了．工学博士．同年 NTT 入社．1989 年から 4 年間 ATR 自動翻訳電話研究所へ出向．1999 年から 1 年間 AT&T 研究所客員研究員．統計的自然言語処理の研究に従事．現在，NTT コミュニケーション科学基礎研究所主幹研究員．情報処理学会奨励賞（1991 年），情報処理学会論文賞（1995 年），人工知能学会研究奨励賞（1995 年）等受賞．電子情報通信学会，人工知能学会，言語処理学会，ACL 各会員．