Vol.2011-SLP-85 No.16
2011/2/5

多対多最小パターンアライメントアルゴリズムの提案と自動読み付与による評価

久保慶伍[†1]　川波弘道[†1]
猿渡洋[†1]　鹿野清宏[†1]

未知語に対する自動読み付与の重要性は高く，音声認識，音声合成, 検索クエリの予測変換などの技術において性能の改善が期待される．未知語に対する自動読み付与においては，文字などの小さい単位で表記と読みをアライメントした辞書データが必要となる．しかし，データを人手で構築するとコストが掛かるため，表記と読みの自動アライメントが研究されている．しかし，従来の研究で提案された手法では，大きい単位でのアライメントほど 1 以下の値の乗算回数が少なくなるため，大きい単位のアライメントが有利になり，小さい単位でのアライメントが困難であった．大きい単位でアライメントが行われると未知語の読み付与に対する頑健性を失われる．本報告では，学習時に各アライメントの乗算回数を表記と読みの全体の文字数にすることで，最も小さい単位で表記と読みをアライメントする手法を提案する．そして，提案手法により自動読み付与のための学習データを構築し，未知語に対する自動読み付与による評価を行った．評価の結果，提案手法が従来手法よりも最大で約 43.6%読み付与正解率を改善した．この結果から，提案法は未知語に対する自動読み付与において有効であることが実証された．

Approach of Many-to-Many Minimum Pattern Alignment Algorithm And Evaluation on Automatic Reading Annotation

Keigo KUBO,[†1] Hiromichi KAWANAMI,[†1]
Hiroshi SARUWATARI[†1] and Kiyohiro SHIKANO[†1]

Previously, a variety of automatic reading annotation to an unknown word has been researched, as improvement of the performance is expected in speech recognition, speech synthesis and predictive transform of a retrieval query, etc. Automatic reading annotation to an unknown word needs a dictionary which includes relation between a grapheme and reading on a small unit. However, it is difficult to construct manually such a dictionary due to the cost. This research addresses to obtain relation of a grapheme and reading on a small unit from a conventional word dictionary etc. automatically, and an unsupervised alignment method that uses the EM algorithm is employed. In the conventional alignment method, because the multiplication frequency decreases in the alignment by the large unit, a large unit tends to be used for alignment. In this report, we proposed a novel method that specify an alignment by the smallest unit by making the multiplication frequency of each alignment the number of characters of the grapheme and reading in training. We evaluated the proposed method on accuracy of automatic reading annotation to the unknown word. Result of evaluation show the proposed method improves the reading annotation correct about 43.6% higher than the conventional method.

1. はじめに

これまでに未知語に対する様々な自動読み付与の研究が行われている．未知語に対する自動読み付与の重要性は高く，音声認識や音声合成，検索クエリ予測変換などの技術において性能の改善が期待される．自動読み付与の研究としては，隠れマルコフモデル[1)] やオンライン識別訓練[2)] などの統計的アプローチや，括弧表現に基づく Web テキストマイニングを用いた自動読み付与といったアプローチ[3)] が提案されている．これらの研究で提案されている手法には，文字などの小さい単位で表記と読みをアライメントした辞書が必要となる．しかし，人手でこの辞書を構築するとコストが掛かるため，表記と読みの自動アライメント手法が研究されている．

これらの研究では，単語辞書などから図 1 のように表記と読みの対データを抽出し，その対データに対して，小さい単位における表記と読みのアライメントを推定することで，自動読み付与のための学習データを構築する．図 1 の「|」はアライメントの区切りを表す記号,「_」は削除文字を表す記号である．本論文における削除文字とは対応する読みがない表記を意味する．これらの研究では EM アルゴリズム[4)] を用いた教師なしアライメントが用いられている．2005 年に Damper らにより 1 対 1 アライメントが提案され[5)]，2007 年に Jiampojamarn らにより多対多アライメントが提案されている[6)]．しかし，Jiampojamarn らの多対多アライメントは，アライメントの仮説ごとに乗算回数が異なるため，大きい単位でのアライメントほど 1 以下の値の乗算回数が少なくなるため，大きい単位のアライメントが有利になり，小さい単位のアライメントが困難であった．大きい単位でアライメントが

†1 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of science and Technology

Vol.2011-SLP-85 No.16
2011/2/5

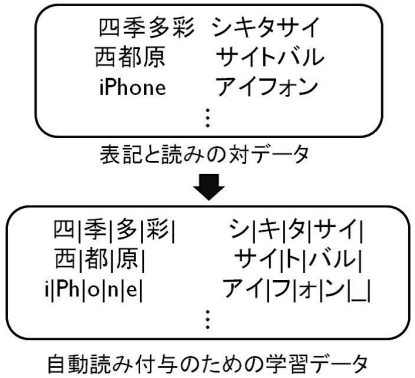

図 1 表記と読みの自動アライメント

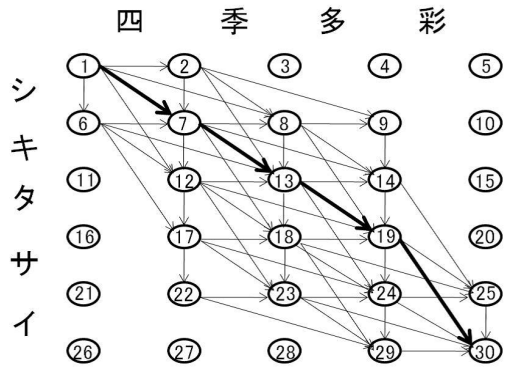

図 2 多対多アライメントの概念図

行われると未知語の読み付与に対する頑健性が失われる．この問題に対して，アライメントの大きさに制約を入れるという対策が行われているが，これは，言語依存になるばかりか，漢字などの表意文字においては，いわゆる当て字など大きい単位のアライメントも許す必要があるため，依然として大きい単位でアライメントが行われるという問題が残る．

そこで，本報告では，アライメントに制約を掛けることなく，学習時に各アライメントの乗算回数を表記と読みの全体の文字数にすることで，最も小さい単位で表記と読みをアライメントする手法を提案する．その手法を用いて，自動読み付与のための学習データを構築し，未知語に対する自動読み付与における評価を行った．

以下，第 2 節では従来手法の多対多アライメントについて説明し，第 3 節では提案法である多対多最小パターンアライメントを説明する．また，第 4 節では提案手法を用いた未知語に対する自動読み付与の評価結果を示し，第 5 節でまとめと今後の展望について述べる．

2. 従来手法 - 多対多アライメント

2.1 アルゴリズム

従来手法である多対多アライメントの基本的な考え方とアルゴリズムについて説明する．図 2 に多対多アライメントの概念図を示す．図 2 は表記「四季多彩」と読み「シキタサイ」のアライメントを表した図である．横軸が表記，縦軸が読みである．また，数字のラベルがついた○は状態を表す．

まず，表記と読みの対データを表す変数を d と定義する．多対多アライメントでは状態の遷移を表記と読みのアライメントとして考える．例えば，図 2 の状態 19 から状態 30 への遷移は「彩」が「サイ」に対応していることを表す．また，状態 1 から状態 2 への遷移は「四」が削除文字であることを表す．この表記と読みのパターンを表した個々の状態遷移を表す変数を u と定義する．また，図 2 から，表記と読みの対データ d はパターン u の系列により表されていると考えることができる．この系列をパターン系列と呼び，変数を $\boldsymbol{u}$ と定義する．また，アライメントされていない表記と読みの対データ d が与えられた時，その対データ d の正しいパターン系列 $\boldsymbol{u}$ は未知である．そのため，多対多アライメントは対データ d において考えられる全てのパターン系列 $\boldsymbol{u}$ を考慮する．その系列の集合を $\boldsymbol{U}$ と定義する．

多対多アライメントでは与えられた対データ d の正しいパターン系列 $\hat{\boldsymbol{u}}$（例 : 図 2 の太矢印）を推定するために，各パターン u に対応するパラメータ p_u を EM アルゴリズムにより推定する．p_u を更新前のパラメータ，$\hat{p}_u$ を更新後のパラメータとすると，E ステップでは，

$$\gamma_{\boldsymbol{u}} = \frac{\prod_{u \in \boldsymbol{u}} p_u}{\sum_{\boldsymbol{u} \in \boldsymbol{U}} \prod_{u \in \boldsymbol{u}} p_u} \tag{1}$$

を計算する．また，M ステップでは

$$\hat{p}_u = \frac{\sum_{\boldsymbol{u} \in \boldsymbol{U}_u} \gamma_{\boldsymbol{u}}}{\sum_{u \in u_{all}} \sum_{\boldsymbol{u} \in \boldsymbol{U}_u} \gamma_{\boldsymbol{u}}} \tag{2}$$

を計算する．u_{all} はパターン u の全種類の集合，$\boldsymbol{U}_u$ は u が出現するパターン系列 $\boldsymbol{u}$ の集合である．この E ステップと M ステップを Forward-Backward アルゴリズムを用いて計算し，パラメータ値が収束するまで繰り返す．そして，推定したパラメータ $\hat{p}_u$ を用いて，与えられた対データ d の正しいパターン系列 $\hat{\boldsymbol{u}}$ を推定する．パターン u の表記の文字数を i_u，読みの文字数を j_u とすると，正しいパターン系列 $\hat{\boldsymbol{u}}$ は以下の式により推定される．

$$\hat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} \prod_{u \in \boldsymbol{u}} \hat{p}_u^{n_u} \tag{3}$$

$$n_u = \begin{cases} i_u & \text{if u is deleted character} \\ j_u & \text{else if u is inserted character} \\ \max(i_u, j_u) & \text{otherwise} \end{cases} \tag{4}$$

n_u は各パターン系列 $\boldsymbol{u}$ の乗算回数の違いを調整するために用いられるスケーリング値である．また，deleted character は削除文字，inserted character は挿入文字を意味する．挿入文字とは対応する表記がない読みのことである．この計算は Viterbi アルゴリズムを用いて計算される．

2.2 多対多アライメントの問題点

多対多アライメントでは正しいパターン系列を推定する際に，各パターン系列の乗算回数の違いを調整するためにスケーリング値である n_u を導入している．しかし，学習時にはスケーリング値が考慮されていないため，パターン系列に含まれるパターン数が多いほど 1 以下の値の乗算回数が多くなる．そのため，パターン数の多いパターン系列（つまり，小さい単位でアライメントされている系列）に含まれるパターンのパラメータ値は低くなり，逆にパターン数が少ないパターン系列（つまり，大きい単位でアライメントされている系列）に含まれるパターンのパラメータ値は高くなる．結果として，大きい単位でアライメントされているパターン系列が，正しい系列として推定されやすくなる．

大きい単位でアライメントされると未知語の読み付与に対する頑健性が失われる．この問題に対して，アライメントの大きさに制約を入れるという対策が多対多アライメントでは行われているが，これは，言語依存になるばかりか，漢字などの表意文字においては，大きい単位のアライメントも許す必要があるため，依然としてこの問題が残る．そこで，学習時にもスケーリング値を考慮した多対多最小パターンアライメントを提案する．

3. 提案手法 - 多対多最小パターンアライメント

3.1 アルゴリズム

多対多最小パターンアライメントでは学習時にもスケーリング値を考慮し，全てのパターン系列において乗算回数が等しくなるようにする．これにより，大きい単位でアライメントされている系列が小さい単位でアライメントされている系列よりも優位になることを防ぐ．

変数は 2.1 節で定義した変数と同じ変数を扱う．パラメータの更新はまず，

$$\gamma_{\boldsymbol{u}} = \frac{\prod_{u \in \boldsymbol{u}} p_u^{n_u}}{\sum_{\boldsymbol{u} \in \boldsymbol{U}} \prod_{u \in \boldsymbol{u}} p_u^{n_u}} \tag{5}$$

を計算する．ただし，n_u だけ再定義し，

$$n_u = i_u + j_u \tag{6}$$

とする．n_u を式 6 のように定義することで，乗算回数はどの系列においても対データの表記と読みの文字数を足し合わせたものに一致する．次に，計算した $\gamma_{\boldsymbol{u}}$ を用いて

$$\hat{p}_u = \frac{\sum_{\boldsymbol{u} \in \boldsymbol{U}_u} \gamma_{\boldsymbol{u}}}{\sum_{u \in u_{all}} \sum_{\boldsymbol{u} \in \boldsymbol{U}_u} \gamma_{\boldsymbol{u}}} \tag{7}$$

を計算する．この式 5 と式 7 をパラメータ値が収束するまで繰り返し計算することでスケーリングを考慮したパラメータ $\hat{p}_u$ を推定する．

これにより，最も小さい単位でのアライメントが可能になる．しかし，学習時にこのようなスケーリングを考慮することで起こる問題もある．本節の残りではスケーリングを考慮することで起こる問題とその対処方法について説明する．

3.2 削除文字と挿入文字の問題

多対多アライメントでは削除文字と挿入文字のパターンが出現すると乗算回数が増加するため，それらのパターンが持つパラメータ値が小さくなり，結果として，削除文字や挿入文字の出現を抑制することが可能であった．しかし，多対多最小パターンアライメントでは各パターン系列ごとの乗算回数が同じになるため，削除文字や挿入文字の出現を抑制するこ

とができず，結果として，削除文字や挿入文字を正確に特定することができない．

そこで，学習時には削除文字・挿入文字を含むパターン系列を考慮せずに学習させ，アライメント時に初めて削除文字・挿入文字を含むパターン系列を考慮する．そして，アライメントの際，削除文字または挿入文字と仮定された文字はそのパターン系列の計算に反映させないことを考える．つまり，削除文字または挿入文字と仮定された文字以外のパターンだけでパターン系列の計算を行い，もしそのパターン系列のスコアが他のパターン系列よりも高かった場合は，その文字が削除文字または挿入文字であったと考える．これにより，削除文字や挿入文字を特定する．この考えをアライメントに導入した式を以下に示す．

$$\hat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} \prod_{u \in \boldsymbol{u}} \hat{p}_u^{\frac{\tilde{n}_u}{N - n_{\boldsymbol{u}di}}} \tag{8}$$

$$\tilde{n}_u = \begin{cases} 0 & \text{if u is deleted character or inserted character} \\ i_u + j_u & \text{otherwise} \end{cases} \tag{9}$$

N は対データの表記と読みを足し合わせた全体の文字数，$n_{\boldsymbol{u}di}$ はパターン系列 $\boldsymbol{u}$ における全削除文字の表記の文字数と全挿入文字の読みの文字数を足し合わせた数である．削除文字と挿入文字のスケーリング値 n_u を 0 にすることで削除文字と挿入文字のパラメータを 1 にする．さらに，全体の文字数から削除文字の文字数と挿入文字の文字数を引き，その値で相乗平均を取ることで，削除文字と挿入文字をパターン系列の計算に反映しないようにしている．もし，スコアが等しいパターン系列が存在する場合は，どちらのパターン系列が正しいかわからないため，パターン数が少ないパターン系列を選択する．

削除文字と挿入文字を抑制したい場合は式 8 に以下のような削除文字と挿入文字を抑制する penalty 変数を導入する．

$$\hat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} \prod_{u \in \boldsymbol{u}} \hat{p}_u^{\frac{\tilde{n}_u}{N - n_{\boldsymbol{u}di}}(1+\text{penalty})} \tag{10}$$

この式 10 のアライメントの計算式により削除文字と挿入文字の問題に対処する．

3.3 誤りパターンが引き起こす問題

多対多最小パターンアライメントは，対データの全てのパターン系列を考慮する．そのため，表記と読みの対応が誤っているパターン（誤りパターン）も学習される．誤りパターンが学習されることにより，当て字などの単語に対して強制的に誤ったアライメントが行われる可能性がある．

表 1 誤りパターンの例

例：「紙鳶　イカノボリ」,「大平紙業　タイヘイシギョウ」
紙鳶　イカノボリ
→ {…, **(紙，イ)**(鳶，カノボリ) ,…}
大平紙業　タイヘイシギョウ
→ {…, (大平，タイヘ)**(紙，イ)**(業，シギョウ) ,…}

その例を表 1 に示す．表 1 は「紙鳶　イカノボリ」と「大平紙業　タイヘイシギョウ」という 2 つの対データのパターン系列の一部を示している．表 1 のパターン系列の一部から「大平紙業　タイヘイシギョウ」の誤りパターンである (紙，イ) が,「紙鳶　イカノボリ」のパターン系列にも出現していることが分かる．これにより，誤りパターンである (紙，イ) のパラメータが学習により高い値を取り，当て字である「紙鳶」の「紙」に対して「イ」がアライメントされる問題が起こる．

そこで，全ての誤りパターンに対して，表記と読みの対応が正しいパターンよりも低いパラメータを一律に与えることで，この問題を回避することを考える．しかし，実際には誤りパターンは未知である．そこで，leave-one-out のアプローチを用いて，パラメータの学習終了後，アライメントを行う対象の対データ以外の全ての対データに対して一度アライメントを行い，そのアライメントに使用されなかったパターンは誤りパターンと仮定する．もし，対象の対データに表れる全てのパターンが誤りパターンの場合は，全ての誤りパターンに対して一律に小さい値が与えられているため，全てのパターン系列のスコアが等しくなり，小節 3.2 より，パターン数が少ないパターン系列（つまり，大きい単位でアライメントされた系列）が選択される．つまり，表 1 を例に説明すると，対象の対データが「紙鳶　イカノボリ」の場合，他の対データである「大平紙業　タイヘイシギョウ」についてまずアライメントが行われる．もし,「大平紙業　タイヘイシギョウ」が適切に (大，タイ)(平，ヘイ)(紙，シ)(業，ギョウ) とアライメントされた場合は，(紙，イ) を含め,「紙鳶　イカノボリ」に出現する全ての分割されたパターンは誤りパターンとなり,「紙鳶　イカノボリ」は (紙鳶，イカノボリ) としてアライメントされることになる．この方法により，誤りパターンが引き起こす問題を回避する．

4. 自動読み付与による評価実験

4.1 実験内容

多対多最小パターンアライメントの自動読み付与における有効性を示すために，未知語に対する自動読み付与の評価実験を行った．実験内容は，まず従来手法の多対多アライメントと提案手法の多対多最小パターンアライメントで単語辞書から抽出した対データをアライメントする．次に，そのアライメントを行ったデータを学習データとして自動読み付与のためのモデルを構築する．そして，未知語が収録されたテストデータに対して自動読み付与を行い，その読み付与正解率（推定した読みが正解の読みと完全に一致した割合）を算出する．

実験条件は表 2 の通りである．単語辞書には NAIST Japanese Dictionary[⋆1]（約 29 万語）を使用した．NAIST Japanese Dictionary には複合語といった形態素よりも大きい単位が存在するため，後段の自動読み付与において重要な特徴量となるコンテキストを十分な長さで持っている．後段の自動読み付与の学習データには従来手法である多対多アライメント (m2m; many-to-many alignment) と提案手法である多対多最小パターンアライメント (mp; minimum pattern alignment) で NAIST Japanese Dictionary をアライメントしたデータを用いた．テストデータには，はてなキーワード[⋆2]から取得した約 11 万 7 千のキーワードの内，mecab[⋆3] で未知語扱いにされた漢字を含むキーワードのみ (2985 語) を使用した．mecab の辞書には NAIST Japanese Dictionary を用いた．各手法のアライメントの条件として，学習回数はパラメータ値の収束が見られた 5 回で，削除文字の出現は許可した．また， mp において削除文字と挿入文字の出現を抑制する変数である penalty の値は 0(mp0)，0.5(mp0.5)，1(mp1) とした．さらに，アライメントの大きさの制約に関して，mp は全て制約なし，m2m は制約なし（m2m without constraints）と制約あり（3 対 3 未満のアライメントのみ[⋆4]，m2m with constraints）の 2 つを実験した．自動読み付与手法にはオンライン識別訓練[2] を採用した．従来手法の多対多アライメントには m2m-aligner[⋆5]，オンライン識別訓練には DirecTL+[⋆6]を使用した．

⋆1 http://sourceforge.jp/projects/naist-jdic/
⋆2 http://d.hatena.ne.jp/keyword/
⋆3 http://mecab.sourceforge.net/
⋆4 つまり，1 対 1，1 対 2，2 対 1，3 対 1，3 対 2，1 対 3，2 対 3 のアライメントのみ有効．3 対 3 未満で実験を行った理由は，日本語において（南，ミナミ）や（豆腐皮，ユバ）といったアライメントが存在するためである．
⋆5 http://code.google.com/p/m2m-aligner/
⋆6 http://code.google.com/p/directl-p/

表 2 実験条件

単語辞書	NAIST Japanese Dictionary (約 29 万語)
学習データ	多対多アライメント (m2m) で単語辞書をアライメントしたデータ 多対多最小パターンアライメント (mp) で単語辞書をアライメントしたデータ
テストデータ	はてなキーワード (2009 年取得，約 11 万 7 千のキーワードの内，mecab で未知語扱いにされた漢字を含むキーワードのみ（2958 語）を使用)
アライメント条件	学習回数は 5 回 削除文字の出現を許可（挿入文字は禁止） mp において penalty は 0(mp0)，0.5(mp0.5)，1(mp1) を試行 mp は全てアライメントの制約なし m2m はアライメントの制約なし（m2m without constraints）と制約あり（3 対 3 未満のアライメントのみ，m2m with constraints）の 2 つ
自動読み付与手法	オンライン識別訓練[2]

表 3 mp0，m2m with constraints，m2m without constraints のアライメントの違い（一部）

mp0	m2m with constraints	m2m without constraints
蔵 | 良 | クラ | ラ |	蔵 | 良 | クラ | ラ |	蔵良 | クララ |
南 | 川 | 原 | ミナミ | カワ | ラ |	南 | 川 | 原 | ミナミ | カワ | ラ |	南 | 川原 | ミナミ | カワラ |
桜 | 見 | サクラ | ミ |	桜 | 見 | サクラ | ミ |	桜見 | サクラミ |
邦 | 郎 | クニ | オ |	邦郎 | クニオ |	邦郎 | クニオ |
飯 | 淵 | ハ | ブチ |	飯淵 | ハブチ |	飯淵 | ハブチ |

4.2 実験結果

実験結果を図 3 に示す．図 3 から，mp でアライメントしたデータを学習データとして用いた自動読み付与は，m2m の制約なし（m2m without constraints）よりも約 43.6%，m2m の制約あり（m2m with constraints）よりも約 3.5%読み付与正解率を改善した．これは，mp によるアライメントが m2m のアライメントよりも小さい単位でアライメントされており，未知語の読み付与に対する頑健性が高いことが理由として挙げられる．表 3 に mp0，m2m with constraints，m2m without constraints のアライメントの違いを示す．表 3 から，どちらも正確にアライメントが取れていることが分かるが，mp の方は最も小さい単位でアライメントが行われている．これは，学習時にもスケーリングを考慮することで，大きい単位でのアライメントが優位になることを防いでいるからである．これにより，多対多最小パターンアライメントは未知語に対する自動読み付与において有効であると言える．

また，今回 mp において penalty の変化による性能の違いは見られなかった．これは，今回のテストデータが漢字を含むキーワードに限定されており，削除文字が少なかったためだ

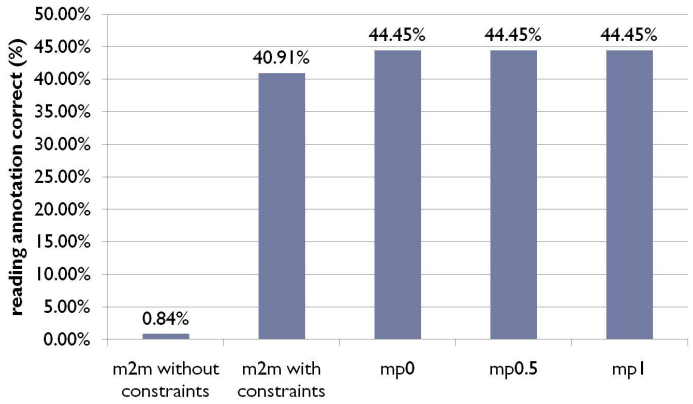

図 3 実験結果

と考えられる．英語などの削除文字が頻出するテストデータにおいては penalty の変化による性能の違いが現れると思われる．

5. まとめと今後の展望

本報告では，アライメントの制約なしで，単語辞書から最も小さい単位の表記と読みの対応関係を求める多対多最小パターンアライメントを提案した．そして，多対多最小パターンアライメントにより読み付与のための学習データを構築し，未知語に対する自動読み付与の評価を行った．評価の結果，多対多最小パターンアライメントが従来手法の多対多アライメントのアライメントの制約なしよりも約 43.6%，制約ありよりも約 3.5%読み付与正解率を改善した．これにより，多対多最初パターンアライメントは未知語に対する自動読み付与において有効であると言える．

今後の展望として，他の言語における評価を行いたいと考えている．また，現実的な問題として対象の言語の先見的知識が全く使えない場合は少なく，多くの場合は先見的知識を使うことができると考えられる．そのため，完全な教師なし学習ではなく，対象の言語の先見的知識も考慮に入れたアライメント手法について研究を行う．

参 考 文 献

1) Paul Taylor, “Hidden Markov Models for Grapheme to Phoneme Conversion,” Proceedings of the 9th, European Conference on Speech Communication and Technology 2005.
2) Sittichai Jiampojamarn, Colin Cherry and Grzegorz Kondrak, “Integrating joint n-gram features into a discriminative training framework,” Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pp.697-700, June 2010.
3) Jumpei Miyake, Shota Takeuchi, Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano, “Automatic Reading Annotation to Japanese Trendy Words based on Parentheses Expression,” Proceedings of Oriental COCOSDA 2008, November, 2008.
4) Arthur Dempster, Nan Laird and Donald Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” In Journal of the Royal Statistical Society, Series B, pp.1-38, 1977.
5) Robert I. Damper, Yannick Marchand, John DS. Marsters and Alexander I. Bazin, “Aligning text and phonemes for speech technology applications using an EM-like algorithm,” International Journal of speech Technology, Vol.8, No.2, pp.147-160, June, 2005.
6) Sittichai Jiampojamarn, Grzegorz Kondrak and Tarek Sherif, “Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion,” Proceedings of NAACL HLT 2007, pp.372-379, April, 2007.