3I4-1

潜在情報を利用したパラレルコーパス生成

Building Parallel Corpus using Latent Information

江里口 瑛子　　小林 一郎
Akiko Eriguchi　　Ichiro Kobayashi

お茶の水女子大学大学院　人間文化創成科学研究科　理学専攻
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Parallel corpora are essential for multilingual processing and statistical machine translation. Generally speaking, it costs much money and time for a human translator to build parallel corpora. More and more attention has been paid to the study in building parallel corpora from comparable corpora like Wikipedia. In this paper, we focus on the Matching Canonical Correlation Analysis (MCCA) model. It can learn bilingual translation lexicons from each monolingual corpus by means of monolingual features, such as context counts and orthographic substrings. This paper adopts a probabilistic topic model, especially a polylingual topic model, which outputs a set of multinomial distribution over words for each topic across multiple languages. We employ the latent topics estimated by the polylingual topic model to monolingual features in the MCCA model. Experimenting on Japanese-English Wikipedia corpus on Buddhism, we show how we estimate latent topics across multiple languages.

1. はじめに

機械翻訳とは，1 つの言語を他の言語へ機械的に変換する作業のことである．この機械翻訳には，大きく分けて 2 種類の手法があり，1 つは規則ベース機械翻訳手法であり，もう 1 つは統計的機械翻訳手法である．両手法に共通する問題点としては，機械翻訳が扱う対象の自然言語には曖昧性や例外が多分に含まれているということがある．前者の手法は，言語間の翻訳規則を恣意的に定める．しかし，全ての翻訳規則を網羅的に記述することが難しいという欠点がある．これに対して，後者の手法は，翻訳規則を統計的・確率的に定める．これによって，規則を網羅的に記述することが可能となり，後者の手法には，自然言語の曖昧性や例外に対応できるという利点がある．

この後者の手法は，Noisy channel モデル [Brown 93] によって，更に翻訳モデルと言語モデルに大別され，これら 2 つのモデルは対訳コーパス (パラレルコーパス) を用いて自動学習される．しかし，複数の言語でパラレルに書かれた文書は希少である．一般的に，翻訳家によるパラレルコーパスの生成が極めて高コストであるからである．他方，Web 上においては，Wikipedia やニュース記事などに見られるような，同一内容に関してそれぞれの言語で書かれた文書 (コンパラブルコーパス) は多く存在する．今日，これらコンパラブルコーパスを利用したパラレルコーパスの自動生成に対する関心が高まっている．

本研究は，データに内在する潜在的トピック，並びに，データに基づいて仮定した潜在空間に注目し，それらを用いたパラレルコーパスの自動生成の手法を提案するものである．多言語トピックモデルの手法を用いて，複数の言語で書かれた文書から潜在的トピックを推定し，得られた言語横断情報に対して正準相関分析によるマッチング (MCCA; Matching Canonical Correlation Analysis) 推定を行い，パラレルコーパスを自動生成することを目的とする．提案手法の予備実験として，多言語トピックモデルを用いて Wikipedia 京都関連文書対訳コーパス (日英コーパス) に内在する潜在的トピックの推定を行う．

連絡先: 江里口 瑛子，お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース 小林研究室，〒 112-8610 東京都文京区大塚 2-1-1，g0920506@is.ocha.ac.jp

2. 関連研究

コンパラブルコーパスを利用したパラレルコーパス生成手法には，単語の文脈情報を利用した手法 [Rapp 95, Fung 98]，翻訳用辞書と単語の出現頻度回数を利用した手法 [Vu 09]，ラベル伝播法を利用した手法 [Tamura 12]，そして Latent Semantic Indexing (LSI) [Deerwester 90] や Latent Dirichlet Allocation (LDA) [Blei 03] などの潜在的意味解析を利用した手法 [Littman 98, Tam 07, Preiss 12] がある．

多言語文書を対象に LDA を拡張させたモデルとして，多言語トピックモデル [Mimno 09, Ni 09, Smet 09] が提案されている．コンパラブルコーパスを提案モデルで学習することにより，文書に内在すると仮定した潜在的トピックに基づく言語横断情報の抽出を行うことができる．Vulić ら [Vulić 11] は，多言語トピックモデルに基づく潜在的トピックの観点から単語の類似度測定を行う手法を提案し，英語とイタリア語のコンパラブルコーパスに適応した．また，Zhu ら [Zhu 13] は，多言語トピックモデルによって得られた言語横断情報の比較方法を提案し，英語と中国語からなるコンパラブルコーパスに適応した．この他，多言語文書分類 [Smet 11, Ni 11] などの多言語文書処理タスクにおいても多言語トピックモデルは利用されている．

他方，Haghighi らは正準相関分析によるマッチング手法 [Haghighi 08] を提案している．Haghighi らは，単語の素性ベクトルとして文脈情報と綴り字情報を統合したものを用いており，これらに対して正準相関分析によるマッチング (MCCA) 推定を行って，訳語候補の共起確率を計算した．この結果，言語構造の関係が近しいとされる英語とスペイン語のコーパスや，英語とフランス語のコーパスに関して，彼らは，高い精度のパラレルコーパス生成に成功した．しかし，英語と中国語のコーパスなど全く異質な言語同士では高い精度は得られなかった．その理由としては，綴り字情報が単語の素性ベクトルとして適当ではなかったからだと考えられている．これに対して，林ら [林 10] は日英コーパスを対象に，特定の単語に対してヒューリスティック値を設け，最大エントロピーモデルを用

いて，素性ベクトルの重み付けに改良を加えたが，十全な結果は得られず，一部の単語ペア推定に対する改善に留まっている．

3. 正準相関分析による対訳語推定

MCCA(Matching Canonical Correlation Analysis) とは，単一言語で書かれた文書集合 (単言語コーパス) から対訳語を抽出するために提案された確率的手法である [Haghighi 08]．単語の素性ベクトルとして，その単語の文脈情報と綴り字情報を統合したものを採用し，正準相関分析と割当問題を反復して解くことで対象にしている複数言語のパラレルな単語ペア (対訳語) をそれぞれ求める．

$\mathbf{s} = (s_1, s_2, \cdots, s_{n_S})$ は翻訳元言語 (ソース言語) の単語集合を，$\mathbf{t} = (t_1, t_2, \cdots, t_{n_T})$ は翻訳先言語 (ターゲット言語) の単語集合を表し，$(i,j) \in \mathbf{m}$ は単語 s_i, t_j が対応関係にある (対訳語である) ことを表している．

MCCA
$\mathbf{m}$ は一様分布で生成
各訳語対 $(i,j) \in \mathbf{m}$ に対して
(i,j) が対訳語ペアであるなら
　$z_{i,j} \sim \mathcal{N}(0, I_d)$, [潜在空間]
　$f_S(s_i) \sim \mathcal{N}(W_S z_{i,j}, \Psi_S)$, [s のベクトル空間]
　$f_T(t_j) \sim \mathcal{N}(W_T z_{i,j}, \Psi_T)$. [t のベクトル空間]
言語 s の単語 i が対訳語に含まれない場合:
　$f_S(s_i) \sim \mathcal{N}(0, \sigma^2 I_{d_S})$.
言語 t の単語 j が対訳語に含まれない場合:
　$f_T(t_j) \sim \mathcal{N}(0, \sigma^2 I_{d_T})$.

3.1 パラメータ推定

対数尤度関数 (式 (1)) を最尤推定することによってパラメータ θ の推定を行う．ここで，$\theta = (W_S, W_T, \Psi_S, \Psi_T)$ は各言語の素性ベクトルの多変量正規分布モデルのパラメータである．パラメータ θ の推定には，EM アルゴリズムを用いる．

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

E-step では，現在のモデルパラメータから重み付き最大となる単語の関係 $\mathbf{m} \in \mathcal{M}$ を求める．M-step では，E-step で得られた $\mathbf{m}$ の下で正準相関分析を行い，各多変量正規分布モデルパラメータの更新を行う．

3.2 M-step

M-step では，正準相関分析を用いて最適パラメータ θ の推定を行う．与えられた単語の対応関係 $\mathbf{m}$ に対して対数尤度関数を最大にするパラメータを求めるため，式 (1) は式 (2) に置き換えることができる．

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

式 (2) によって新たに示された最尤推定問題は，正準相関分析によって解くことができる．言語の特徴ベクトルをそれぞれ射影し，射影先の各特徴ベクトルを比較した際，相関が最大となるように固有値ベクトル U_S, U_T を固有値問題として求めることで，パラメータ θ は式 (3–6) より求まる．

$$W_S = C_{SS} U_S P^{\frac{1}{2}}, \tag{3}$$

$$W_T = C_{TT} U_T P^{\frac{1}{2}}, \tag{4}$$

$$\Psi_S = C_{SS} - W_S W_S^{\mathrm{T}}, \tag{5}$$

$$\Psi_T = C_{TT} - W_T W_T^{\mathrm{T}}, \tag{6}$$

$$C_{SS} = \frac{1}{|\mathbf{m}|} \sum_{(i,j)\in\mathbf{m}} f_S(s_i) f_S(s_i)^{\mathrm{T}}. \tag{7}$$

ここで，P は $d \times d$ の正準相関係数行列を表す．C_{TT} は，C_{SS} と同様に共分散行列の計算で求めることができる．

3.3 E-step

E-step では，単語間の重み付き最大マッチング $\mathbf{m} \in \mathcal{M}$ を求める．M-step で求めた θ と式 (8) を用いることで，ソース言語の単語とターゲット言語の単語の対応関係情報を求めることができる．

$$\mathbf{m} = \arg\max_{\mathbf{m}'} \log p(\mathbf{m}', \mathbf{s}, \mathbf{t}; \theta). \tag{8}$$

ただし，計算量を抑えるために，式 (8) をそのまま解くのではなく，単語のマッチング最大化問題 (式 (9)) に置き換えて解く．ここで，式 (10) は，ソース言語の単語 i とターゲット言語の単語 j 間のマッチング辺の重み (対訳語となる確率) を表す．

$$\log p(\mathbf{m}, \mathbf{s}, \mathbf{t}; \theta) = \sum_{(i,j)\in\mathbf{m}} w_{i,j} + C, \tag{9}$$

$$w_{i,j} = \log p(s_i, t_j; \theta) - \log p(s_i; \theta) - \log p(t_j; \theta). \tag{10}$$

4. 提案手法: 潜在的トピックによるパラレルコーパス生成

本研究では，林ら [林 10] と同様に日英コーパスを対象に，MCCA の抱える素性ベクトルの綴り字情報の問題点を改善する手法の提案を行う．具体的には，多言語トピックモデルによって得た言語横断情報 (単語のもつ潜在的トピック分布 ϕ^l) を単語の素性ベクトルに採用し，MCCA 推定を行い，パラレルコーパスを生成し，精度の確定を行う．

4.1 多言語トピックモデル

PLDA(Polylingual Latent Dirichlet Allocation) [Mimno 09] とは，複数言語で書かれた文書を文書組とみなし，この文書組を同時に分析するため，トピックモデルの枠組みに基づいて提案された手法である．我々は，PLDA を日英コーパスに対して用いる．

パラレルでない多言語文書を対象にした処理手法では,「同一内容に関して書かれた文書であれば同じ意味の単語が同じ頻度で出てきやすい」という仮定の下で，単語の共起情報や文脈情報などに着目した研究がなされてきた [Rapp 95, Fung 98, Vu 09]．Mimno ら [Mimno 09] は，この仮定を「同一内容に関して複数言語で書かれた文書組であれば，各文書組内に含まれる話題 (潜在的トピック) の比率 (θ) は等しい」という仮定の下，多言語トピックモデルを提案した．

図 1 は PLDA のグラフィカルモデルを表す．背景が白色の変数は潜在変数を表し，背景が灰色の変数は観測変数を表す．各言語 $l = 1, \cdots, L$ に対して，言語毎のトピック分布集合 $\Phi^1, \cdots, \Phi^L$ ($\Phi^l = \{\phi_1^l, \cdots, \phi_K^l\}$) が存在する．

PLDA の生成過程は以下の通りである．$\boldsymbol{w} = (\boldsymbol{w}^1, \cdots, \boldsymbol{w}^L)$ は L 種類全ての言語の文書集合を表す．ここで，$Dir(\cdot)$ はディリクレ分布を表し，w_n^l は言語 l の n 番目の単語，z_n^l は言語 l の n 番目の単語の潜在的トピック，ϕ_k^l は言語 l のトピック k の単語分布，そして θ_k はトピック k の文書分布を表す．ただし，本研究で用いる多言語トピックモデルは，$L = 2$ のときの PLDA とする．

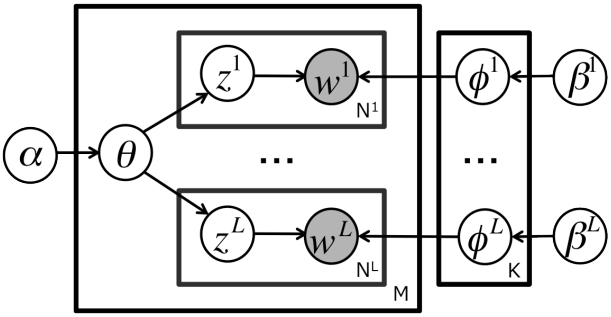

図 1: PLDA のグラフィカルモデル

1. 言語 l の各トピック $k = 1, \cdots, K$ について:

$$\phi^l \sim Dir(\beta^l). \tag{11}$$

2. 言語 l の各文書 $d^l = 1, \cdots, M$ について:

$$\boldsymbol{\theta} \sim Dir(\alpha) \tag{12}$$

(a) 言語 l の各単語 $w^l_n = 1, \cdots, N^l$ について:

$$\boldsymbol{z}^l \sim P(\boldsymbol{z}^l|\boldsymbol{\theta}), \tag{13}$$
$$\boldsymbol{w}^l \sim P(\boldsymbol{w}^l|\boldsymbol{z}^l, \Phi^l). \tag{14}$$

5. 予備実験: PLDA によるトピック推定

5.1 実験仕様

Wikipedia 日英京都関連文書対訳コーパス[*1]を対象データとして用い，PLDA による多言語トピック推定を行った．Wikipedia 日英京都関連文書対訳コーパスは，京都に関する約 50 万文書の Wikipedia の日本語記事を人手によって英語に翻訳したものであり，多言語翻訳などを目的に生成された日英対訳コーパスである．英語翻訳文は 3 種類用意されており，それぞれ，一次翻訳文，二次翻訳文，最終翻訳文である．このうち，本予備実験で用いる英語コーパスは最終翻訳文とする．

Wikipedia 日英京都関連文書対訳コーパスは，記事の内容によって 15 のカテゴリによって分けられており，本予備実験では，このうち，仏教カテゴリに含まれる 1,061 文書 ($M = 1061$) を用いる．トピックモデルによって推定される各潜在的トピックは，対象文書中に含まれる名詞によって特徴付けられるため [Griffiths 05]，本予備実験においては，各日英コーパスから名詞のみを抽出し，これらに対してトピック推定を行う．日英コーパスから名詞を抽出するため，日本語コーパス，英語コーパスの形態素解析器として，それぞれ，MeCab[Kudo 04]，TreeTagger[Schmid 94] を用いた．抽出された名詞数は，日本語コーパス，英語コーパス，それぞれにおいて，21,172 個と 19,824 個であった．また，トピック推定におけるストップワードの影響を調査するため，更に名詞のストップワードを除いたデータセットも用意した．このとき，抽出された名詞数は，日本語コーパス，英語コーパス，それぞれにおいて，21,090 個と 21090 個であった．PLDA におけるハイパーパラメータ α, β^l は，それぞれ，$\alpha = 50/K$, $\beta^l = 0.01$ とする．トピック数 K はパラメータとし，$K \in \{500, 800, 1200\}$ の範囲を動かす．トピック推定には周辺化ギブスサンプリングを用い，反復回数は 200 回とする．

*1 http://alaginrc.nict.go.jp/WikiCorpus/

5.2 実験結果

PLDA の最適トピック数の決定にはパープレキシティ値を用いた．式 (15) は，PLDA によって推定された言語 l の言語モデルにおけるパープレキシティ値の算出式を表す．ここで，θ_{d^l,z^l} は，言語 l の d 番目の文書に対して割り当てられた潜在的トピック z^l の値を表し，$\phi_{z^l,w^l_{d^l,i}}$ は，言語 l の d 番目の文書中に出現する i 番目の単語 $w^l_{d^l,i}$ に割り当てられた潜在的トピック z^l の値を表す．

$$P(\boldsymbol{w}^l) = exp(-\frac{1}{N^l}\sum_{d^l,i} log(\sum_{z^l} \theta_{d^l,z^l}\phi_{z^l,w^l_{d^l,i}})). \tag{15}$$

図 2 は，PLDA で学習された各言語モデルのパープレキシティ値をトピック毎にプロットしたものである．with はストップワードを含めたコーパスを用いた場合の結果であり，without はストップワード除いたコーパスを用いた場合の結果である．PLDA による日本語モデル，英語モデルの最適トピック数は，with の場合が $K = 500$，without の場合が $K = 800$ となった．

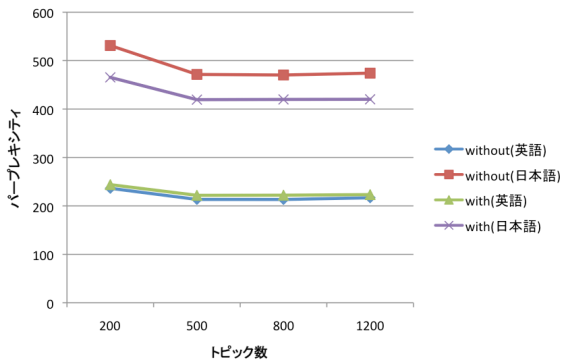

図 2: トピック毎の各言語モデルのパープレキシティ値

表 1, 2 は，最小パープレキシティ値によって定めた最適トピック数を用いた際の，with コーパス，without コーパスそれぞれにおいて推定された潜在的トピックとそのトピック内に出現する単語をまとめたものである．with コーパスと without コーパスそれぞれにおいて，潜在的トピック毎に求まった英語および日本語の各単語を，出現確率が高いものから順に 10 個ずつ表記した．

5.3 考察

表 1 から，ストップワードを除去しなかった with コーパスでは，各トピックにおいて,「ため」や「こと」などの日本語のストップワードが高い確率で出現していることが見て取れる．また，アルファベット 1 文字あるいは平仮名 1 文字が名詞として抽出され，トピック推定か行われていることもまた分かる．これについては，各言語で使用した形態素解析器が名詞抽出に失敗していることが理由に挙げられる．他方，without コーパスでは，このような，アルファベット 1 文字あるい平仮名 1 文字はストップワードリストに含められており，既に除去されている．両者のコーパスからトピック推定した結果 (表 1, 2) を比較してみると，without コーパスによる結果の方が，各言語の形態素解析によるノイズが少なく，各トピックを構成する単語のまとまりが良い．

6. おわりに

MCCA の抱える素性ベクトルの綴り字情報の問題点を改善するため，MCCA 推定の際に，多言語トピックモデルを用い

表 1: トピック毎の英日単語表の一部 (with)

Topic 9	Topic 104	Topic 495
temple: 年	temple: 年	temple: 年
sect: ため	buddha: よう	period 三
kyoto: よう	sutra: 県	buddhist: 国
people: こと	priest: こと	imperial: の
who: 日	kukai: 日本	priest: もの
buddhist: 市	buddhist: 仏教	sect: 像
buddhism 論	ritual: 寺院	kyoto: ため
zen: 禅	time: 市	school: これ
city: 経	s: 世	year: 本尊

表 2: トピック毎の英日単語表の一部 (without)

Topic 34	Topic 69	Topic701
temple: 最澄	temple: 菩薩	buddhist: 仏教
period: 仏	statue: 法	temple: 寺
kannon: 像	school: 時代	kyoto: 色
city: 日本	buddhism: 像	sutra: 下賜
father: 僧	scripture: 坐禅	priest: 無量
buddhism: 相	kyoto: 不動明王	school: 姿
kukai: 文庫	nenbutsu: 経	sect: 如来
keisaku: 歳	ceremony: 院	period: 善信
bosatsu: 寺	enlightenment: 法華宗	nichiren: 集
fudo: 衆	age: 日蓮	region: 経

て得た言語横断情報を単語の素性ベクトルに採用し，対訳語推定を行う手法の提案を行った．

予備実験として，Wikipedia 日英京都関連文書対訳コーパスを用いて PLDA による多言語文書への潜在的トピック情報の推定を行った．最適トピック数を決定し，ストップワードを除去したコーパスを用いた場合と除去しなかった場合とでトピック推定を行い，比較を行った結果，ストップワードを除去した方がトピックのまとまりが良いことが分かった．今後，MCCA による訳語対マッチングを行い，提案手法の検証を行う．

参考文献

[Blei 03] Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research* (2003)

[Brown 93] Brown, P. F. , Pietra, V. J. D., Pietra, S. A. D. and Mercer, R. L.: The mathematics of statistical machine translation: parameter estimation, *Journal of Computational Linguistics – Special issue on using large corpora: II* (1993)

[Rapp 95] Rapp, R.: Identifying word translations in nonparallel texts, In *Proceedings of the ACL* (1995)

[Fung 98] Fung, P. and Yee, L. Y.: An IR approach for translating new words from nonparallel, comparable texts, In *Proceedings of COLING and ACL* (1998)

[Vu 09] Vu, T., Aw, A. T. and Zhang, M.: Feature-based method for document alignment in comparable news corpora, In *Proceedings of EACL* (2009)

[Deerwester 90] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman R.: Indexing by latent semantic analysis, *Journal of the American Society for Information Science* (1990)

[Littman 98] Littman, M., Dumais, S. T. and Landauer, T.K.: Automatic cross-language information retrieval using latent semantic indexing, *Cross-Language Information Retrieval* (1998)

[Tam 07] Tam, Y., Lane, I. and Schultz, T.: Bilingual-LSA based LM adaption for spoken language translation, In *Proceedings of ACL* (2007)

[Preiss 12] Preiss, J.: Identifying comparable corpora using LDA, In *Proceedings of the NAACL: Human Language Technologies* (2012)

[Mimno 09] Mimno D., Wallach, H. M., Naradowsky, J., Smith, D. A. and McCallum, A.: Polylingual topic models, In *Proceedings of EMNLP* (2009)

[Ni 09] Ni, X., Sun, J., Hu, J. and Chen, Z.: Mining multilingual topics from Wikipedia, In *Proceedings of the 18th International Conference on WWW* (2009)

[Smet 09] De Smet, W. and Moens, M.: Cross-language linking of news stories on the Web using interlingual topic modeling, In *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining* (2009)

[Vulić 11] Vulić, I., De Smet, W. and Moens, M.: Identifying words translations from comparable corpora using latent topic models, In *Proceedings of ACL* (2011)

[Griffiths 05] Griffiths, T. L., Steyvers, M., Blei, D. M. and Tenenbaum, J. B.: Integrating topics and syntax, In *Advances in NIPS 17* (2005)

[Kudo 04] Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, In *Proceedings of EMNLP* (2004)

[Schmid 94] Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees, In *Proceedings of International Conference on New Methods in Language Processing* (1994)

[Zhu 13] Zhu, Z., Li, M., Chen, L. and Yang, Z.: Building comparable corpora based on bilingual LDA model, In *Proceedings of ACL* (2013)

[Smet 11] De Smet, W., Tang, J. and Moens, M.: Knowledge transfer across multilingual corpora via latent topics, In *Proceedings of the 15th PAKDD* (2011)

[Ni 11] Ni, X., Sun, J., Hu, J. and Chen, Z.: Cross lingual text classification by mining multilingual topics from Wikipedia, In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (2011)

[Haghighi 08] Haghighi, A., Liang, P., Berg-Kirkpatrick, T. and Klein, D.: Learning bilingual lexicons from monolingual corpora, In *Proceedings of the ACL* (2008)

[Tamura 12] Tamura, A., Watanabe, T. and Sumita, E.: Bilingual Lexicon Extraction from Comparable Corpora Using Label Propagation, In *Proceedings of EMNLP and CNLL* (2012)

[林 10] 林 克彦, 福西 孝章, 西田 昌史, 山本 誠一. MCCA モデルの日英辞書構築への適用, 言語処理学会第 *16* 回年次大会発表論文集, *pp. 982–985* (2010)