統計的機械翻訳における グラフ伝搬を用いた未知語対訳辞書構築の改善

大和田 賢一　　小町 守
首都大学東京 システムデザイン研究科
{ohwada-kenichi@ed, komachi@}.tmu.ac.jp

1 はじめに

統計的機械翻訳において，未知語 (Out-of-Vocabulary words) の問題は依然として大きなままである．大規模な対訳コーパスから翻訳ルールを学習する統計的機械翻訳では，対訳コーパス中に存在しない単語に関する翻訳ルールを学習することができず，そのような単語は未知語として扱われる．学習のための対訳コーパスが十分に無い場合や，そのために学習とテストで異なるドメインのコーパスを用いるような場合には，未知語の問題はより深刻になる．

対訳コーパスが十分な量存在しない言語ペアやドメインが多くある一方で，単言語の (monolingual) 言語データは，多くの言語またはドメインにおいて既に多く存在し，また常に生成されている．そうした状況下で，単言語コーパスまたは対照コーパスを用いて翻訳ルールを学習することを目的とした研究が，これまで多く取り組まれてきた．それらの手法は，対訳辞書構築 (Bilingual Lexicon Induction: BLI) と呼ばれ，対訳コーパスによって得られる対訳辞書を，単言語コーパスまたは対照コーパスを用いることによって拡張することを目的としている．そこでは，既に翻訳を取得している単語またはフレーズと未知語の間の分布類似度を用いることによって，既知の翻訳ルールを未知語へと伝搬させる．

近年では，対訳辞書構築にグラフベースのラベル伝搬アルゴリズムを導入することによって，単言語コーパス中の単語またはフレーズをより有効に活用する方法が提案されている．その手法では，対訳コーパスから学習された既知のフレーズテーブルにおいて，原言語側の単語またはフレーズが持っている翻訳候補とその確率からなる確率分布をラベルと見做し，そのラベルをグラフ構造を利用して未知語へと伝搬させる．

私達は，このグラフベースの手法における，各ノードのベクトル表現を比較し，それらに関する実験の結果を報告する．この論文の主要な貢献は，グラフにおけるノードのベクトル表現として，従来の語彙次元からなる疎なベクトルではなく分散表現に基づくベクトルを用いて，その有効性を検証したことである．

2 先行研究

既知の対訳辞書をシードとして，それを単言語コーパスや対照コーパス (comparable corpus) を用いることによって拡張することを目的とする対訳辞書構築 (BLI) は，Rapp (1995) によって始められた．対訳辞書構築は，ある言語において共起する 2 つの単語がある時，別の言語におけるそれらの単語の翻訳も共起するだろう，という仮定に基づいている．Marton et al. (2009) は，単言語コーパス由来の未知語の言い換えから得られた対訳辞書が，end-to-end の SMT システムを改善させうることを示した．

近年では，ラベル伝搬アルゴリズムを用いたグラフベースの手法が探究されるようになってきた．Razmara et al. (2013) は，Marton et al. (2009) の手法にグラフ伝搬アルゴリズムを適用し，また，単言語コーパス中の単語またはフレーズをラベル無しノードとしてグラフに加えた三部グラフに基づく手法を提案した．その研究では unigram，つまり未知語に対してこの手法を適用しているが，Saluja et al. (2014) は bigram にも焦点を当て，そのことの有効性を示した．

半教師あり学習手法の一種であるグラフベースのラベル伝搬アルゴリズムは，まず Zhu and Ghahramani (2002) によって提案された．それは，少数のラベル付きノードと多くのラベル無しノードによって構成されるグラフを用いて，既知のラベルの情報をラベル無しノードへと伝搬させる手法である．Razmara et al. (2013) は，その手法を改良した *modified Adsorption* (MAD) アルゴリズム (Talukdar and Crammer, 2009) を用いている．Razmara et al. (2013) が原言語側の

類似度のみを考慮してグラフを構築しているのに対して，Saluja et al. (2014) は，構造化ラベル伝搬アルゴリズム (Liu et al., 2012) を用いて，目的言語側の類似度に関する情報をも利用している．

3 ラベル伝搬を用いたグラフベースの対訳辞書構築

この節では，私達が今回ベースラインとして比較対象にした，Razmara et al. (2013) の単言語コーパスを用いたグラフベースの対訳辞書構築手法について詳述する．Razmara et al. (2013) におけるグラフは，*i*) 各未知語，*ii*) フレーズテーブルにおける原言語側の単語またはフレーズ，そして *iii*) 単言語コーパスにおけるその他の単語またはフレーズ，に対応する 3 種類のノードを持ち，各ノードに対応する単語またはフレーズ間の類似度をノード間のエッジの重みとしている．フレーズテーブルの原言語側に存在する単語またはフレーズは翻訳候補と翻訳確率をラベルとして保持しており，そのラベルの情報をグラフ構造を利用して未知語へと伝搬させる．

3.1 文脈ベクトルの作成

グラフベースの手法に限らず，対訳辞書構築では一般的に単語またはフレーズを文脈ベクトルで表現する．Marton et al. (2009) や Razmara et al. (2013) では，規模の大きい原言語側の単言語コーパスを利用してその文脈ベクトルを作成している．まず，各ノードに対応する単語またはフレーズの単言語コーパスにおける出現に対して，それらの左右に決められた窓サイズ内で出現する単語を文脈語とし，その文脈語との共起カウントを各成分とする共起ベクトルを作成する．

文脈ベクトルの作成方法には，文脈語の位置を相対位置によって区別するもの (Rapp, 1995) と区別しないもの (Marton et al., 2009; Razmara et al., 2013) がある．Saluja et al. (2014) では左右のみが区別され，それぞれの中での位置では区別されない．

集計された共起カウントを用いて，単語と文脈語との間の関連性尺度 (association measure) が計算でき，一般的にはその値で共起カウントを置き換えたベクトルを用いる．関連性尺度としては，相互情報量 (PMI)，条件付き確率，尤度比等の値が用いられる．

3.2 類似度の計算

ベクトルとして表現された各単語またはフレーズの間で類似度を計算し，その類似度を対応するノード間のエッジの重みとして用いる．類似度の尺度としては，コサイン類似度，L1-ノルム，Jensen-Shannon Divergence 等が用いられうる．Razmara et al. (2013) では，予備実験の結果に基づき PMI を用いたベクトルとコサイン類似度の組み合わせを用いている．

3.3 グラフ構築

前述のように，グラフは単語またはフレーズをノードとし，それらの間の類似度をエッジの重みとする．Razmara et al. (2013) で提案されるグラフは，フレーズテーブルの原言語側に存在する単語またはフレーズにあたるラベル有りノードと，各未知語そして単言語コーパス中のその他の単語に対応する 2 種類のラベル無しノードの，合わせて 3 種類のノードがあり，同じ種類のノード間にはエッジが存在しない三部グラフとなっている．各ノードがエッジによって連結されるノードの数は，類似度が上位の k 個に限定される．あるノードに関して，類似度の上位 k 個を計算するためには，種類が異なる全てのノードとの間の類似度を計算しなければならないが，Razmara et al. (2013) ではある文脈語を窓内に持つ単語集合を返す転置インデックスを構築し，文脈語を一語も共有しないノードを候補から外すことで類似度計算の量を削減している．

3.4 ラベル伝搬

グラフベースのラベル伝搬アルゴリズム (Zhu and Ghahramani, 2002) は，少数のラベル有りノードから，多くのラベル無しノードへとラベル情報を伝搬させる，半教師有り学習のアルゴリズムであり，近い位置にあるデータ点は類似したクラスラベルを持つという仮定に基づくアルゴリズムである．対訳辞書構築の問題設定では，ラベルは目的言語の翻訳候補の集合に対する確率分布である．

グラフに用いるノードの数を $|V|$，翻訳候補の数 m とし，$|V|\times|V|$ のノード間の遷移確率行列 T と，$|V|\times(m+1)$ のラベル行列 Y を考える [1]．その時 T は，

$$T_{ij} = \frac{w_{ij}}{\sum_{k=1}^{|V|} w_{kj}} \tag{1}$$

となる．ここで，w_{ij} は，i 番目と j 番目のノードの間のエッジの重み（類似度）である．この式では全て

[1] 列数が $m+1$ になるのは，翻訳候補数 m にどの候補にも翻訳されない場合のラベルを加えるため．

の語彙に関する重みの総和が分母になっているが，今回の問題では類似度が上位の k 個の近傍に関する重みのみを用いる．

ここから，ラベル伝搬アルゴリズムは，

$$Y \leftarrow TY \tag{2}$$

のように行列演算を行うことによる伝搬を繰り返し，各反復時に Y の各列を翻訳候補に対する確率分布になるように正規化する．また，アルゴリズム開始時のラベル有りノードに対して，各反復時に初期のラベル分布へと復元させ，ラベル有りノードのラベル分布が変化しないようにする．

4 ノードのベクトル表現への単語分散表現の適用

これまでの対訳辞書構築手法では，文脈語タイプの数が次元数になっている疎なベクトル表現が用いられてきた．

私達は，低次元で密なベクトルを類似度計算のためのノードの表現として用いることの効果を検討するために，以下の 2 つのベクトル表現を適用した．

SVD: 単言語コーパスから得られるカウントによって構築される共起行列の各セルを positive-PMI の値で置き換えた行列に SVD (singular value decomposition)[2] を適用し次元削減を行う．後述する実験では，次元数は 300 を用いた．結果として得られる行列の各行を対応する単語のベクトル表現として用いる．

word2vec: 単言語コーパスを学習データとして word2vec (Mikolov et al., 2013) を学習し，結果として得られる単語ベクトルをノードの表現として適用した．実験では，CBOW モデル，窓サイズ 3，次元数 300 を用いた [3]．

5 実験

5.1 実験設定

私達は，科学論文ドメインの ASPEC コーパスにおける日英翻訳のタスクにおいて，フレーズベース SMT システム (PBSMT) に対して今回の手法を適用した．ベースラインの PBSMT の訓練に用いる対訳コーパスの文数は 5 万文と 40 万文を用い，グラフ伝搬手法に用いる単言語コーパスの文数は 50 万文と 300 万文を用いて実験を行った．

end-to-end の PBSMT システムとして，私達は Moses 2.1.1 (Koehn et al., 2007) をデフォルトの素性とパラメータで用いた．単語アライメントの取得には GIZA++ 1.0.7 (Och and Ney, 2003) を利用し，言語モデルの構築には KenLM (Heafield, 2011) により 5-gram のモデルを学習した．また，素性の重みの最適化には MERT (Och, 2003) を用いた．

ラベル伝搬には Junto (Talukdar and Crammer, 2009) において実装されているラベル伝搬アルゴリズム (Zhu and Ghahramani, 2002) を用いた [4]．

比較対象として Razmara et al. (2013) を再実装した．その際，文脈語の窓サイズは左右 3 単語ずつ，グラフにおいてエッジを張る近傍の数 k には 20 を用いた [5]．また，数字とアルファベットを含む語はノードにしないようにした．ノードのベクトル表現の各成分の値には positive-PMI を，類似度の計算にはコサイン類似度を使用した．

私達は，三部グラフに加えて，ラベル有りノードと未知語に対応するラベル無しノードのみを用いた二部グラフにおいても実験を行った．

5.2 評価尺度

Razmara et al. (2013) と同様に，intrinsic な評価尺度として MRR（平均逆順位）と Recall を用いた．以下に MRR と Recall の計算方法を示す．

MRR: ラベル伝搬によって獲得された未知語の翻訳候補リスト [6] を翻訳確率の降順にソートし，そのリストにおける未知語に対する正解の翻訳 [7] のランクを得る．そして，そのランクの逆数の平均値を MRR の値として用いる [8]．

Recall: ランクに関係の無い評価を行うために，Recall を計算する．未知語の正解の翻訳が，手法によって得られたリストの上位 20 個中に存在すれば 1，そうでなければ 0 としてカウントし，未知語の数で割ることによって値を得る．

[2] scikit-learn において実装されている truncated SVD を用いた．

[3] 単語の最小カウントはデフォルトの値とし，それにより単言語コーパス中の全ての語に対してはベクトルが作られないため，そのような語はグラフ構築における近傍の候補から外れる．

[4] イテレーション回数は，全ての設定で 3 回とした．

[5] 三部グラフにおいては，各ノードは 15 個のラベル有りノードと 5 個のラベル無しノードへと繋がれる．

[6] リストとしてはシステムの出力のうち上位 100 位までを用いた．

[7] GIZA++ によって与えられる単語アライメントを正解とした．dev セットと test セットを 100 万文対の対訳文と結合したデータを GIZA++ に処理させ，アライメントを得た．

[8] 一つの未知語に対して複数の正解が与えられる場合には，全ての正解に関するランクを用いて平均を計算した．

表 1: 三部グラフにおける MRR と Recall (%)

サイズ	単言語 50 万文		単言語 200 万文	
ノード表現	MRR	Recall	MRR	Recall
ベースライン	1.61	3.86	2.30	4.84
+min5 [1]	2.17	4.72	2.65	5.46
SVD	1.18	3.26	1.58	4.16
word2vec	1.50	4.40	2.05	5.37

[1] 単語最小カウント 5 で学習された word2vec のモデル中にある単語のみでグラフを構築したベースライン手法.

表 2: 二部グラフにおける MRR と Recall (%)

サイズ	単言語 50 万文		単言語 200 万文	
ノード表現	MRR	Recall	MRR	Recall
ベースライン	1.16	3.82	1.50	4.97
SVD	0.65	3.00	0.64	3.43
word2vec	0.89	3.38	1.39	4.28

5.3 実験結果と考察

各提案手法について，対訳コーパスとして 5 万文対 [9] を用いた時の，MRR，Recall による評価結果を表 1，表 2 に示す.

当然予想されるように，用いる単言語コーパスのサイズを大きくすることで，結果の大きな改善が見られる．また，Razmara et al. (2013) の実験結果と同様に，単言語コーパスを用いて三部グラフを構築することで MRR，Recall ともに二部グラフからの向上が見られ，特に MRR でより大きく向上している.

ノード表現の効果において SVD を用いることはどの設定においても良い影響をもたらさなかった．word2vec を三部グラフにおいて用いることは MRR をわずかに下げる一方で，Recall に向上をもたらす．二部グラフにおいては，予想に反して word2vec がベースラインを下回った．また，三部グラフでのベースラインと word2vec の結果における違いは，ベクトルの学習時に低頻度語を無視していることの効果であるかもしれないため，三部グラフにおいて word2vec の語彙中に存在する語のみを用いてベースライン手法による実験を行った．その結果は word2vec を上回ったため，低頻度語を用いないことの有効性が示されたと考えられる.

6 おわりに

グラフベースのラベル伝搬アルゴリズムを用いた未知語に関する対訳辞書構築において，先行研究における語彙次元からなる疎なベクトルに変えて分散表現に基づくベクトルをノードの表現として用いた．分散表現を用いる効果は確認されなかったが，低頻度語を省くことで MRR と Recall に改善が見られた.

今後については，異なるグラフ構造とラベル伝搬アルゴリズムを用いること，未知語だけでなく未知語を含むフレーズにも手法を適用することを計画している.

[9] このサイズの対訳コーパスでは，dev，test セットにおける未知語の数は，合わせて 2,038 個であった.

References

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of WMT*, pages 187–197.

Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of NAACL-HLT*, pages 518–523.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL (demo)*, pages 177–180.

Shujie Liu, Chi-Ho Li, Mu Li, and Ming Zhou. 2012. Learning translation consensus with structured label propagation. In *Proceedings of ACL*, pages 302–310.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of EMNLP*, pages 381–390.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of ACL*, pages 320–322.

Majid Razmara, Maryam Siahbani, Reza Haffari, and Anoop Sarkar. 2013. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of ACL*, pages 1105–1115.

Avneesh Saluja, Hany Hassan, Kristina Toutanova, and Chris Quirk. 2014. Graph-based semi-supervised learning of translation models from monolingual data. In *Proceedings of ACL*, pages 676–686.

Partha Pratim Talukdar and Koby Crammer. 2009. New regularized algorithms for transductive learning. In *Proceedings of ECML/PKDD*, pages 442–457.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of EMNLP-CoNLL*, pages 24–36.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University.