単語アライメントと語順情報を用いた Earth Mover's Distanceに基づく機械翻訳のための自動評価法

越前谷 博 †
荒木 健治 ‡
† 北海学園大学大学院工学研究科 ‡ 北海道大学大学院情報科学研究科
†echi@lst.hokkai-s-u.ac.jp ‡araki@ist.hokudai.ac.jp

1 はじめに

機械翻訳研究においては，近年のニューラルネット翻訳の急速な発展により，これまでの BLEU[1] に代表されるような表層的な単語の一致に基づく自動評価法とは異なる，単語の意味を反映させた自動評価法に対するニーズが高まっている．このような背景より，単語分散表現を用いた文間類似度の計算手法が提案されている．柳本 [2] は，2 つの離散分布間の距離を輸送問題として扱う Earth Mover's Distance(EMD)[3] を用いて文書間の類似度を求めている．その際，単語には word2vec[4] による分散表現を用いている．また，Kusner ら [5] は EMD に対して単語アライメントを行ったうえで文書間の類似度を決定する Word Mover's Distance(WMD) を提案している．WMD では語のアライメントを word2vec に基づき行い，対応付けのコストが最も低い場合のコストの総和を類似度として求めている．更に，松尾ら [6] は単語の分散表現による単語アライメントを自動評価に適用し，全ての単語間のアライメントを考慮した Whole Alignment Similarity が文単位の評価に有効となることを示した．

我々はこのような単語の意味に基づく自動評価法に向けての第一段階として，確率的な手法により単語アライメントを行い，その結果に基づき得られた語順の情報を EMD に適用する新たな自動評価法を提案する．先行研究においても単語アライメントの結果を利用しているが，語順の情報を十分に反映させているとはいえない．そこで，提案手法では EMD を適用するにあたり，単語アライメントの結果と単語の位置情報を組み合わせることで翻訳文を評価する．NTCIR-7 データ [7] を用いた英語-日本語間の翻訳に対する性能評価実験の結果，単語アライメントや語順情報を用いずに EMD を適用した場合に比べ，提案手法は人手評価との間で高い相関係数を示した．

2 提案手法

提案手法では EMD を用いるにあたり，特徴に文中の単語，重みには文レベルの $tf \cdot idf$，そして，距離には Dice 係数と位置情報を組み合わせた値を利用する．

2.1 単語に対する重み

WMD では機能語などのストップワードを事前に取り除いたうえで，内容語を対象に単語アライメントを行っている．それに対して，提案手法では翻訳文と参照訳の単語の重みに $tf \cdot idf$ を文レベルに対応させた計算式を用いて，機能語と内容語を差別化する．式 (1) にその計算式を示す．

$$tf \cdot isf = (\log(tf(w,s)) + 1) \times \frac{|S|}{sf(w)} \qquad (1)$$

上式の $tf = \log(tf(w,s)) + 1$ の $tf(w,s)$ は任意の語 w における文 s 中の出現頻度である．$isf = \frac{|S|}{sf(w)}$ は任意の語 w が出現する文数 $sf(w)$ に対する全翻訳文及び全参照訳数 $|S|$ の逆数である．式 (1) では，機能語のように複数の文に出現する語については isf の値は小さくなる．その反面，出現頻度 tf は 2 以上であっても log を用いているため $tf \cdot isf$ の値はそれほど大きくはならない．それに対し，内容語は isf の値が非常に大きくなるため，tf の値が 1 でも $tf \cdot isf$ の値は大きくなる．その結果，式 (1) より機能語と内容語の差別化が可能となる．

2.2 単語アライメント

EMD を用いる際の単語間の距離として，全ての単語間の距離を単純に EMD に用いると対応関係にない単語間の影響を受け，翻訳文と参照訳との間で適切なスコアを求めることができない．そこで，提案手法で

は翻訳文と参照訳間で単語アライメントを行い，対応関係が明確に得られる単語間についてのみ距離を求める．単語アライメントは単語間の Dice 係数と表層情報を用いて行う．以下に翻訳文と参照訳中の単語間の対応関係における信頼度の計算式を示す．

$$\text{信頼度} = \begin{cases} \frac{Dice\ 係数+1.0}{2.0} & (w_c = w_r のとき) \\ \frac{Dice\ 係数}{2.0} & (w_c \neq w_r のとき) \end{cases} \quad (2)$$

$$Dice\ 係数 = \frac{2 \times f_{cr}}{f_c + f_r} \quad (3)$$

式 (2) の信頼度では Dice 係数のみではなく，単語の表層が一致しているかどうかも考慮する．Dice 係数だけでは単語間の対応関係を一意に決定できない場合に表層情報を利用することは有効である．具体的には翻訳文中の単語 w_c と参照訳中の単語 w_f が表層レベルで一致する場合には，Dice 係数に 1.0，一致しない場合には何も加えずに 2.0 で割った値を信頼度とする．Dice 係数は式 (3) より得る．ここで f_c は翻訳文中の単語 w_c の全翻訳文に対する出現頻度である．f_r は参照訳中の単語 w_r の全参照訳文に対する出現頻度である．そして，f_{cr} は翻訳文の単語 w_c と参照訳 w_r の単語が対応関係にある全翻訳文と参照訳の組に同時に出現する頻度である．

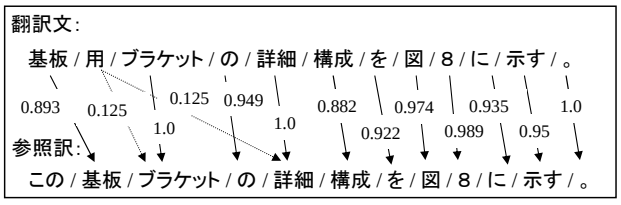

図 1: 単語アライメントの具体例

図 1 に上式の信頼度により決定された単語アライメントの具体例を示す．単語間の対応関係は翻訳文の任意の単語を基準として参照訳の全単語との間で求め，最も信頼度の高い参照訳の単語を選択することで決定する．図 1 の数値は対応関係が得られた単語間の信頼度である．信頼度が同じ単語が複数存在した場合には一意に決定できないとして対応関係は得られないものとする．対応関係に多義性を残したまま EMD を適用することは適切なスコアを得る際の妨げとなる可能性が高くなるためである．図 1 においては翻訳文中の単語「用」は参照訳中の単語「ブラケット」と「詳細」との間の信頼度が共に 0.125 となり一意に対応付けができないと同時に，参照訳中の「ブラケット」と「詳細」は 0.125 よりも高い信頼度 1.0 で翻訳文中の単語「ブラケット」と「詳細」との間でそれぞれ対応関係が得られる．したがって，翻訳文中の単語「用」は参照訳中のいずれの単語とも対応付けは行われない．提案手法では Dice 係数と表層レベルの情報の両方を用いることで対応関係を決定する際の多義性の解消を図っている．

2.3 単語アライメントに基づく距離計算

提案手法では，単語アライメントの結果に基づいて単語間の距離を決定する．計算式を以下の式 (4) に示す．

$$d = \begin{cases} 1.0 - 信頼度 \times pos\_diff & (対応関係ありのとき) \\ 1.0 & (対応関係なしのとき) \end{cases} \quad (4)$$

$$pos\_diff = 1.0 - \left| \frac{pos(w_c)}{len(c)} - \frac{pos(w_r)}{len(r)} \right| \quad (5)$$

式 (4) の信頼度は式 (2) を意味し，pos_diff は対応関係にある翻訳文中の単語と参照訳中の単語の文中における相対位置のずれを示している．すなわち，提案手法では語順の違いを距離計算に反映させている．翻訳文と参照訳の間で対応関係にある単語であっても語順が大きく異なれば pos_diff の値は小さくなり信頼度に対して負の重みとなる．したがって，距離 d は値が大きくなる．逆に，相対位置が近いほど語順に差がないため pos_diff の値は大きくなり信頼度に対する重みは小さくなる．その結果，距離 d は値が小さくなる．また，単語アライメントの結果，対応関係の存在しない単語間の距離 d は全て最大値の 1.0 とする．

pos_diff は式 (5) より得る．式 (5) は我々が従来より提案している表層情報に基づく自動評価法 IMPACT[8] において，翻訳文と参照訳間の一致チャンクを一意に決定するために用いられている．$pos(w_c)$ と $pos(w_r)$ はそれぞれ翻訳文中と参照訳中の位置を示し，$len(c)$ と $len(r)$ はそれぞれ翻訳文と参照訳の長さ，すなわち，構成単語数を示す．例えば，図 1 にある翻訳文と参照訳の単語「基板」の相対位置はそれぞれ $0.083(=\frac{1}{12})$ と $0.167(=\frac{2}{12})$ となるため，pos_diff の値は $0.916(=1.0-|0.083-0.167|)$ となる．その結果，翻訳文と参照訳の単語「基板」の距離は $0.182(=1.0-0.893\times0.916)$ となる．このような距離計算により得られる，図 1 の翻訳文と参照訳との間の距離行列を図 2 に示す．図 2 では行が参照訳，列が翻訳文に対応する．

	この	基板	ブラケット	の	詳細	構成	を	図	8	に	示す	。
基板	1.	0.182	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.
用	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.
ブラケット	1.	1.	0.	1.	1.	1.	1.	1.	1.	1.	1.	1.
の	1.	1.	1.	0.051	1.	1.	1.	1.	1.	1.	1.	1.
詳細	1.	1.	1.	1.	0.	1.	1.	1.	1.	1.	1.	1.
構成	1.	1.	1.	1.	1.	0.118	1.	1.	1.	1.	1.	1.
を	1.	1.	1.	1.	1.	1.	0.078	1.	1.	1.	1.	1.
図	1.	1.	1.	1.	1.	1.	1.	0.026	1.	1.	1.	1.
8	1.	1.	1.	1.	1.	1.	1.	1.	0.011	1.	1.	1.
に	1.	1.	1.	1.	1.	1.	1.	1.	1.	0.065	1.	1.
示す	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	0.05	1.
。	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	0.

図 2: 距離行列の具体例

提案手法ではEMDを用いてスコアを得るために，特徴には単語，重みには式 (1) による文レベルの $tf \cdot idf$，距離には単語アライメントと語順を考慮した式 (4) を用いる．EMDを適用する際には，重みは翻訳文と参照訳でそれぞれ総和を1.0とし，EMDの値を0.0〜1.0の範囲にする．また，EMDは値が小さいほど類似度が高くなるため，以下の式 (6) により，スコアと類似度を比例させる．

$$score = 1.0 - EMD\text{ 値} \tag{6}$$

3 性能評価実験

3.1 実験方法

評価実験はNTCIR-7データを用い，提案手法により得られるスコアと人手評価との相関を求めることで行う．NTCIR-7データの内訳は英語の翻訳文と日本語の翻訳文がそれぞれ100ずつあり，全翻訳文にはそれぞれ参照訳が1つずつ用意されている．英日の機械翻訳システム数は5，日英の機械翻訳システム数は15である．また，提案手法の有効性を検証するために，距離行列に全単語間のDice係数の値を用いた場合，すなわち，単語アライメントを行わずにEMDを用いてスコアを求めた場合のシステム（ベースライン）と距離計算に語順情報を用いずに信頼度のみを用いた場合のシステムを使用する．人手評価との相関はPearson coeeficientとKendall's τ correlationをAdequacyとFluencyの観点から求めた．

3.2 実験結果

表1に英語から日本語への翻訳におけるシステムレベルの実験結果，表2に日本語から英語への翻訳におけるシステムレベルの実験結果をそれぞれ示す．また，表3に英語から日本語への翻訳における文レベルの実験結果，表4に日本語から英語への翻訳における文レベルの実験結果をそれぞれ示す．全ての実験結果には表層レベルに基づく自動評価法であるBLEUとIMPACTの相関係数を付与している．

表 1: EtoJ におけるシステムレベルの実験結果

	Pearson		Kendall	
	adequacy	fluency	adequacy	fluency
提案手法	-0.313	0.365	0.0	0.2
ベースライン	-0.246	0.351	0.0	0.2
語順情報なし	-0.597	0.066	0.0	0.2
BLEU	-0.199	0.184	0.0	0.2
IMPACT	0.254	0.489	0.2	0.4

表 2: JtoE におけるシステムレベルの実験結果

	Pearson		Kendall	
	adequacy	fluency	adequacy	fluency
提案手法	0.815	0.938	0.785	0.612
ベースライン	0.752	0.864	0.632	0.536
語順情報なし	0.701	0.877	0.287	0.191
BLEU	0.730	0.881	0.498	0.440
IMPACT	0.814	0.935	0.689	0.555

表 3: EtoJ における文レベルの実験結果

	Pearson		Kendall	
	adequacy	fluency	adequacy	fluency
提案手法	0.537	0.503	0.346	0.354
ベースライン	-0.124	-0.070	-0.022	0.006
語順情報なし	0.415	0.424	0.260	0.298
sentBLEU	0.440	0.447	0.267	0.306
IMPACT	0.657	0.583	0.461	0.419

表 4: JtoE における文レベルの実験結果

	Pearson		Kendall	
	adequacy	fluency	adequacy	fluency
提案手法	0.521	0.539	0.367	0.386
ベースライン	0.067	0.155	0.090	0.110
語順情報なし	0.464	0.519	0.331	0.359
sentBLEU	0.463	0.478	0.333	0.354
IMPACT	0.631	0.646	0.466	0.467

3.3 考察

表 1 から表 4 の実験結果は提案手法の有効性を示している．提案手法は単語アライメントなしのベースラインと語順情報なしのシステムに比べ，表 1 の Pearson の adequacy を除いて同じ相関係数もしくは高い相関係数を示している．したがって，単語アライメントと語順の情報に基づいた距離計算を EMD に適用することは有効であることが明らかとなった．表 1 の相関係数はいずれの自動評価法においても低い値となっているが機械翻訳システムの数が 5 と少ないことから，一部の評価の誤りが相関係数の大幅な低下をもたらしたと考えられる．

また，単語の表層レベルでの一致に基づく従来の自動評価法との比較では BLEU に対して提案手法の相関係数は高い．しかし，IMPACT との比較では表 2 よりシステムレベルでは提案手法の方が高い相関係数を示しているが，文レベルでは IMPACT の方が提案手法より高い相関係数を示している．したがって，提案手法において文レベルでの評価精度の向上が課題である．

4 おわりに

本稿では，単語アライメントと語順情報を用いた EMD に基づく新たな自動評価法を提案し，提案手法の有効性を性能評価実験に基づき示した．今後は，EMD を適用する際に，特徴量として word2vec を用いる予定である．word2vec を用いた場合でも，提案手法の単語アライメントと語順情報は適用可能であり，更なる精度向上が期待できると考えられる．

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318.

[2] 柳本豪一. 2015. 単語の分散表現を利用した文書類似度. In Proceedings of the 29th Annual Conference of the Japanese Society for Artificial Intelligence, 4K1-1.

[3] Yossi Rubner, Carlo Tomasi and Leonidas J. Guibas. 2000. *The Earth Mover's Distance as a Metric for Image Retrieval.* International Journal of Computer Vision, 40(2), 99–121.

[4] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space.* In Proceedings of Workshop at ICLR.

[5] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin and Kilian Q. Weinberger. 2015. *FromWord Embeddings To Document Distances.* In Proceedings of the 32nd International Conference on Machine Learning, 957–966.

[6] 松尾潤樹，小町守，須藤克仁. 2016. 単語分散表現を用いた単語アライメントによる日英機械翻訳の自動評価尺度. 情報処理学会第 227 回自然言語処理研究会, Vol.2016-NL-229 No.20, 1-7.

[7] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro. 2008. *Overview of the Patent Translation Task at the NTCIR-7 Workshop.* In Proceedings of NTCIR-7 Workshop Meeting, 389–400.

[8] Hiroshi Echizen-ya and Kenji Araki. 2007. *Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum.* In Proceedings of the Eleventh Machine Translation Summit, 151–158.