機械翻訳のための自動評価法における文分割を用いた大局的評価の利用

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

統計翻訳やニューラルネット翻訳の研究の進展に伴い，より精度の高い機械翻訳のための自動評価法が求められている．自動評価法は BLEU[1] に代表される n-gram 一致率に基づく手法，RIBES[2] のように語順に基づく手法，そして，METEOR[3] や IMPACT[4] のように共通チャンクに基づく手法など様々なものが提案されている．しかし，これらの手法は単語を最小単位とした局所的な評価法であり，大局的な観点からの評価という点で不十分と考えられる．

従来より大局的な情報を用いた自動評価法は提案されている．我々は，チャンキングツールより得た名詞句を最小単位として翻訳文を評価し，その評価スコアと単語を最小単位とした評価スコアの間の相加平均を最終的な評価スコアとする手法を提案している [5]．また，Liangyou ら [6] は，チャンキングツールを用いて名詞句だけでなく，動詞句の抽出も行っている．しかし，これらの手法では名詞句や動詞句を抽出する際に大規模なコーパスが必要となり多言語への適応は容易ではない．これらの先行研究 [5, 6] では，共に評価対象の言語は英語のみとなっている．

そこで本稿では，様々な言語に適用可能であり，かつ，大局的な観点での評価を考慮した，新たな自動評価法を提案する．提案手法では，文中に存在する機能語に相当する語をストップワード，即ち，文中の区切り単語と位置付け，それらを参照訳のみから自動的に抽出する．次いで，ストップワードを用いて翻訳文と参照訳をそれぞれいくつかの部分に分割し，その部分を最小単位とすることにより翻訳文に対する大局的な評価を行う．そして，得られた大局的な観点からの評価スコアを，単語を最小単位とした局所的な評価スコアに対する重みとして用いる．WMT2014 データと NTCIR-7 データを用いた評価実験の結果，提案手法の有効性を確認した．

2 文分割に基づく大局的な評価

2.1 ストップワードの決定

提案手法では，翻訳文に対する大局的な観点からの評価スコアを求める．始めに，全参照訳[1]に対する統計的なアプローチに基づき，機能語に相当する語を抽出する．具体的には，全参照訳に出現する全単語に対して式 (1) を用いて数値化する．

$$tf \cdot idf = \log(tf(w, |R|)) \times \frac{|R|}{df(w)} \qquad (1)$$

$tf = \log(tf(w, |R|))$ は任意の語 w における全参照訳 R 中の出現頻度である．$idf = \frac{|R|}{df(w)}$ は任意の語 w における全参照訳数 $|R|$ に対する出現参照訳数の逆数である．この式 (1) より，機能語のように複数の参照訳に出現する語については idf の値は非常に小さくなる．また，出現頻度 $tf(w, |R|)$ は高いが log を用いているため $tf \cdot idf$ の値は過度に大きくはならない．それに対し，内容語は idf の値が非常に大きくなるため，tf の値が小さくても $tf \cdot idf$ の値は大きくなる．したがって，式 (1) より機能語に相当する語の $tf \cdot idf$ の値は基本的には小さくなる．次いで，以下の式 (2) より閾値を求め，その閾値以下の $tf \cdot idf$ の値を持つ語を機能語に相当する語として抽出する．本稿では，このように抽出された語を文分割のために利用するためストップワードと呼ぶこととする．

$$\text{閾値} = \log\left(\frac{|R|}{\mu}\right) \times \mu \qquad (2)$$

式 (2) の μ はパラメータである．全参照訳にそれぞれ 1 度出現することを前提とし，μ 分の 1 の割合で全参照訳に出現した語をストップワードとして採用するということを示している．例えば，全参照訳の 10 分の 1 の割合で出現した語をストップワードとする場合

[1]全参照訳とは，1 つの翻訳文に対する参照訳ではなく，全ての評価対象の翻訳文に対する参照訳を意味する．

には，μ の値を 10 に設定する．式 (1) において，全参照訳の 10 分の 1 の割合で 1 度のみ出現する場合，$tf(w, |R|)$ の値は $\frac{|R|}{10}$ となり，tf は $\log\left(\frac{|R|}{10}\right)$ より得られる．また，idf の値は $10(=\frac{|R|}{|R|/10})$ となる．したがって，$\log\left(\frac{|R|}{10}\right) \times 10$ の値が閾値となる．このような方法により，ストップワードの個数を指定するのではなく，参照訳数に応じてストップワードを動的に決めることが可能となる．

2.2 ストップワードに基づく文分割

ストップワードが抽出されるとそれらを用いて翻訳文と参照訳を分割する．始めに最長共通部分列 (LCS) に基づき共通単語を決定する．そして，共通単語がストップワードの場合，翻訳文と参照訳両方の分割の区切り位置として使用される．図 1 を用いて提案手法による文分割処理の詳細を述べる．始めに，翻訳文と参照訳間における共通部分を LCS に基づき決定する．ここで共通部分とは一致単語が連続している箇所を一つの単位としたものである．したがって，図 1 の (1) では共通部分は “I was” と “to see the gnashing of teeth and argument”，“.” の 3 つとなる．また，この共通部分に含まれるストップワードは “*to*”，“*of*”，“*and*”，そして，“.” である．その結果，翻訳文，参照訳共に 4 つに分割される．

次いで，分割部分の対応付けを行う．対応付けは分割された部分毎に類似度を求めることで行う．類似度は式 (3) より得る．$length(P)$ は部分の単語数を，LCS は部分間の LCS を示している．例えば，図 1 の (2) では翻訳文の部分 “I was waiting at the gate” を基準に参照訳の 5 つの部分との類似度を求めるとそれぞれ “I was excepting” との類似度は 0.333(=2/6)，“see gnashing” との類似度は 0.0，“teeth” との類似度は 0.0，“a fight breaking out at the gate” との類似度は 0.5(=3/6) となる．この結果，翻訳文の部分 “I was waiting at the gate” から見た場合に最も類似度が高い “a fight breaking out at the gate” が選択される．同様に，参照訳の部分 “a fight breaking out at the gate” については類似度が最も高い翻訳文中の部分 “I was waiting at the gate” が選択される．このように翻訳文側の部分と参照訳側の部分が互いに選択される場合，対応する部分であると見なす．

$$sim = \frac{LCS}{length(P)} \tag{3}$$

更に，対応付けされた分割部分に対しては同じ番号を付与し，一般化する．対応する部分が存在しない場合には “P@” とする．図 1 では翻訳文の “I was waiting at the gate” と参照訳の “a fight breaking out at the gate” が対応部分が存在しないとして “P@” に一般化される．その結果，翻訳文は “P1 P2 P3 P@”，参照訳は “P@ P2 P3 P1” となる．この一般化された翻訳文と参照訳の間で ROUGE-L[7] を求める．ROUGE-L は，出現順に厳しい評価基準であり，部分の出現順を評価するために適している．また，評価スコアが 0.0 から 1.0 に正規化されるため，評価スコアの値を直感的に捉えることができる．図 1 の例では “at the gate” の位置が翻訳文と参照訳の間で大きく異なっていることが評価スコアを低下させる原因となる．このように提案手法ではストップワードを用いた文分割により大局的な観点での評価を行っている．

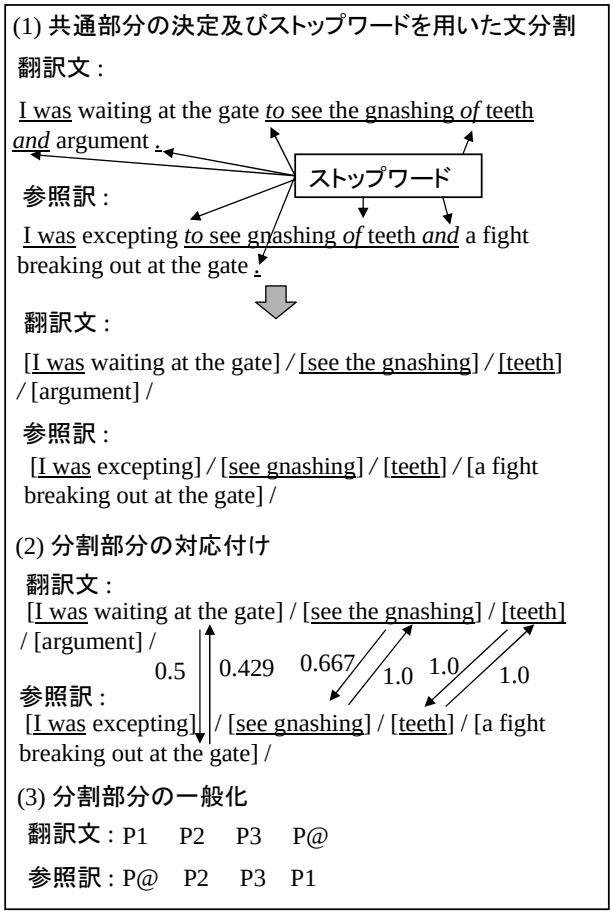

図 1: ストップワードを用いた文分割の例

3 IMPACT を用いた局所的な評価

翻訳文と参照訳間における単語を最小単位とした局所的な評価には，我々が従来より提案している自動評

価法 IMPACT[4] を使用する．IMPACT は共通チャンクを再帰的に決定することで，出現順が異なる共通チャンクであっても評価スコアに反映することができる．その際には出現順の異なる共通チャンクに対して重みパラメータを用いて制御する．IMPACT では以下の式 (4) から (7) より評価スコアを求める．

$$Ch\_score = \sum_{ch \in ch\_num} length(ch)^{\beta} \qquad (4)$$

式 (4) の Ch_score は共通チャンク毎に構成単語数に対してパラメータ β を用いて重みづけを行った値の総和を示している．例えば，図 1 の翻訳文と参照訳間ではパラメータ β が 2 の場合，共通チャンク “I was” においては $4(=2^2)$，“to see gnashinh of teeth and” において $36(=6^2)$，“.” においては $1(=1^2)$ となるため，Ch_score の値は 41 となる．

$$R = \frac{\left(\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score\right)\right)^{\frac{1}{\beta}}}{m} \qquad (5)$$

$$P = \frac{\left(\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score\right)\right)^{\frac{1}{\beta}}}{n} \qquad (6)$$

式 (5) と (6) は式 (4) より求めた値に対して，それぞれ参照訳の語数と翻訳文の語数を用いて 0.0 から 1.0 の範囲となるように正規化を行っている．ただし，出現順が異なる共通チャンクが存在する場合には，パラメータ α を用いて重みを変化させている．例えば，図 1 の例では，出現順の異なる共通チャンクとして “at the gate” が存在する．この共通チャンに対しては Ch_score は $9(=3^2)$ となる．しかし，他の共通チャンクと出現順が異なるため，パラメータ α が 0.5 の場合には，$4.5(=0.5^1 \times 9)$ となり，小さな値となる．このように IMPACT では出現順の異なる共通チャンクに対しては，パラメータ α を用いて制御している．

$$local\_score = \frac{(1+\gamma^2)RP}{R+\gamma^2 P} \qquad (7)$$

式 (7) は式 (5) と (6) より得られた R と P の調和平均を示している．γ は $\frac{P_{ph}}{R_{ph}}$ より得られる値である．このように提案手法では IMPACT を用いて局所的な評価スコアを得ることができる．

4 大局的な評価の利用

提案手法では，大局的な評価スコアを局所的な評価スコアの重みとして用いることで最終的な評価スコアを得る．以下の式 (8) にその計算式を示す．式 (8) の $global\_score$ は 2 章で述べた大局的な評価スコアである．δ は $global\_score$ に対する重みパラメータである．$global\_score$ の値は 1.0 以下であるため，δ を 0.0 から 1.0 とすることで最終的なスコアも 0.0 から 1.0 となるようにしている．

$$score = (global\_score)^{\delta} \times local\_score \qquad (8)$$

5 性能評価実験

5.1 実験データ

性能評価は WMT2014 の Metrics Task データ [8] と NTCIR-7[9] の PATMT データを用いて行った．WMT2014 では，5 つの言語と英語における双方向での翻訳，NTCIR-7 では日本語と英語間の翻訳に関して翻訳文と参照訳，そして，人手評価が提供されている．また，評価は相関係数を求めることで行った．システム単位の評価においてはピアソンの相関係数，セグメント単位の評価においてはケンドールの順位相関係数を求めた．

5.2 実験結果及び考察

システム単位とシステム単位の実験結果を表 1 と表 2 にそれぞれ示す．提案手法においては，式 (2) のパラメータ μ の値は WMT2013 データを用いて行ったチューニングより 6.0 を用いた．また，式 (8) のパラメータ δ の値は $global\_score$ の影響を過度に受けないように 0.1 を用いた．表 1 と表 2 において，WMT2014 の “fromEn” は英語から 5 つの言語への翻訳，“toEn” は 5 つの言語から英語への翻訳におけるそれぞれの相関係数の平均である．また，NTCIR-7 の “fromEn” は英語から日本語への翻訳，“toEn” は日本語から英語への翻訳における adequacy と fluency の相関係数の平均である．

表 1: システム単位の実験結果

	WMT2014		NTCIR-7	
	fromEn	toEn	fromEn	toEn
提案手法	0.815	0.898	0.332	0.901
IMPACT	0.816	0.896	0.316	0.889
METEOR	0.815	0.829	0.025	0.800
BLEU	0.803	0.888	-0.008	0.805

表 2: セグメント単位の実験結果

	WMT2014		NTCIR-7	
	fromEn	toEn	fromEn	toEn
提案手法	0.283	0.309	0.456	0.481
IMPACT	0.282	0.310	0.448	0.479
METEOR	0.306	0.354	0.256	0.375

表 1 のシステム単位では，大局的な評価を用いない IMPACT に比べ，提案手法の相関係数は WMT2014 の “fromEn” を除き，全て高い値を示した．NTCIR-7 の “fromEn” においてはいずれの自動評価法も低い相関係数を示しているが，これはシステムの数が 5 つと非常に少なかったため，一つのシステムの評価が与える影響が強く，相関係数の大幅な低下を招いたと考えられる．表 2 のセグメント単位では，WMT2014 の “toEn” 以外の全てで IMPACT よりも高い相関係数を示した．システム単位とセグメント単位共に NTCIR-7 の “fromEn” において比較的大きな改善が見られた．

しかし，METEOR との比較において，WMT2014 のシステム単位では提案手法と IMPACT 共に高い相関係数を示しているが，セグメント単位では下回っている．この点は今後に向けた大きな課題である．

6 おわりに

本稿では，様々な言語に適用可能であり，かつ，大局的な観点からの評価を考慮した，新たな自動評価法を提案した．性能評価実験を通して，提案手法の有効性を確認した．今後はセグメント単位での評価精度の向上のための研究を行う予定である．

謝辞

この研究は国立情報学研究所との共同研究に関連して行われた．

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] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh and Hajime Tsukada. 2010. *Automatic Evaluation of Translation Quality for Distant Language Pairs.* In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 944–952

[3] Michael Denkowski and Alon Lavie. 2014. *Meteor Universal: Language Specific Translation Evaluation for Any Target Language.* In Proceedings of Ninth Workshop on Statistical Machine Translation, 376–380

[4] 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.

[5] Hiroshi Echizen-ya and Kenji Araki. 2010. *Automatic Evaluation Method for Machine Translation using Noun-Phrase Chunking.* In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 108–117.

[6] Li Liangyou, Gong Zhengxian and Zhou Guodong. 2012. *Phrase-Based Evaluation for Machine Translation.* In Proceedings of the 19th International Conference on Computational Linguistics, 663–672

[7] Chin-Yew Lin and Franz Josef Och. 2004. *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics.* In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 606–613

[8] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia and Aleš Tamchyna. 2014. *Findings of the 2014 Workshop on Statistical Machine Translation.* In Proceedings of the Ninth Workshop on Statistical Machine Translation, 12–58.

[9] 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.