開発・運用コストの低い機械翻訳評価システム

山本晋平　　南條 浩輝　　吉見 毅彦

龍谷大学 理工学部 情報メディア学科

e-mail: syamamoto@nlp.i.ryukoku.ac.jp

1　はじめに

機械翻訳の開発は盛んに行われていが，その性能はまだ人間の翻訳に比べ劣っている．今後もさらなる研究開発が必要である．機械翻訳の開発過程において機械翻訳文の評価は重要なプロセスであるが，これを人手で行うのは，評価コストが高い．自動評価システムが実現できれば評価コストを軽減することができ，機械翻訳の開発がスムーズに行えるようになる．

人手による機械翻訳文の品質評価では，機械翻訳文に対して適切さと流暢さの二つの側面から評価値が付与される．適切さは，原文によって読者に伝わる情報のうちどの程度が翻訳文によって伝わるかを測る尺度である．一方，流暢さは，翻訳文が目的言語の文としてどの程度流暢 (自然) であるかを原文とは独立に測る尺度である．本研究で扱う評価システムは，流暢さの側面から評価を行う．

これまでに様々な評価システムが示されているが，２章で述べるように，これらを開発・運用するためには一定のコストがかかる．このため，開発・運用コストが低い評価システムが求められている．本研究では，開発・運用コストを抑えた評価システムを提案し，その評価精度を検証する．

2　機械翻訳の自動評価

既存の自動評価手法のなかには，機械学習によって評価システムを構築するものがある．機械学習を利用する手法は，機械学習によって回帰モデルを構築する手法と識別モデルを構築する手法に分けることができる．以下で，2 種類のモデルの利点と欠点について述べる．

回帰モデルは，機械翻訳文から抽出された素性に基づいてその機械翻訳文の良さを表わす評価値を予測する．このため，回帰モデルを構築する手法 [1, 2, 3, 4] では，システム評価値によるきめ細かい評価が可能である．しかし，モデルを構築するために大量の学習データに人手で評価値を付与する必要があり，学習データの作成に要するコストが比較的大きい．

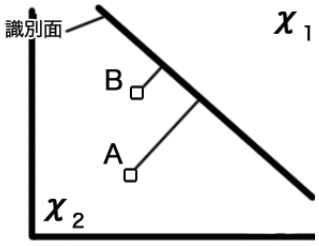

図 1: 事例と識別面との距離を評価値とするモデル

識別モデルを構築する手法 [5, 6, 7] では，評価対象の機械翻訳文を人間による翻訳とみなすことができるかどうかの二値判定を行なう．この手法では，対訳コーパスにおける訳文 (人間翻訳文) と，原文を機械翻訳で翻訳して得られる機械翻訳文を学習データとすることができるため，大量の学習データに人手評価値を付与する必要がなく，コストは比較的小さく抑えられる．しかし，二値判定であるため，きめ細かい評価とはいえない．

回帰モデルと識別モデルの利点を活かすために，人手評価値が付与されていない学習データを用いて回帰モデルを構築する方法が示されている [8]．この方法では，識別面と評価対象事例 (機械翻訳文) との距離を評価値として出力する．例えば，図 1 では入力された機械翻訳文を，人間による翻訳とみなせるかどうかを判定する識別モデルについて考えている．χ_1 を人間翻訳文のクラス，χ_2 を機械翻訳文のクラスとした場合，入力された機械翻訳文 A と B はどちらも機械翻訳文であると識別される．しかし，B は A より識別面に近いため，B の方が人間翻訳文に近く，A より良い機械翻訳文であるとみなせる．この方法によって，入力された機械翻訳文に対して評価値を与えることができるため，きめ細かい評価が行え，かつ，開発コスト

が低い評価システムを開発することが可能になった．

3 流暢さの評価のための素性の検討

本研究では，Kulesza の方法 [8] に倣い，人手評価値の付与されていない学習データを用いて，回帰モデルを作成する．機械学習はサポートベクターマシンを用いる．Kulesza の評価システムでは，機械学習のための素性として，BLEU，NIST，WER，PER が使われている．このため，これらの素性値を得るために複数の参照訳が必要になり，運用に関してはコストがかかってしまう．この運用コストを抑えるために，本研究では，参照訳を必要としない言語的特徴を機械学習のための素性とする．ここでは，機械翻訳文の流暢さを評価対象としているため，文献 [7] に示されている品詞出現比率，品詞 Ngram($N = 1, 2, 3, 4, 5$)，単語 3gram の 3 種類の素性を用いた．

品詞出現比率 人間が英語を日本語に翻訳する場合，より自然な日本語にするために様々な工夫をしている．そのうちの一つとして品詞転換がある．品詞転換とは，例えば英語の名詞を和訳する際，それを日本語の名詞として翻訳せずに他の品詞を使うことである．品詞転換を行うことによって，より自然な日本語の文に翻訳できることがある．人間は，このような工夫を行って翻訳しているが，機械翻訳では適切な品詞転換が行われるとは限らない．このため，人間にとって不自然な表現になることがある．

以上のようなことから，本研究では，人間翻訳文の流暢さと機械翻訳文の流暢さの違いを適切に表現できる素性として品詞の出現比率に着目した．ある品詞の出現比率は，訳文中に出現した全品詞に対するある品詞の割合であるとする．サポートベクターマシンによる機械学習のための素性ベクトルは，形態素解析システム「茶筅」[1]の品詞名を素性名とし，その品詞の文中での出現比率を素性値とする成分で構成する．

品詞 Ngram 品詞の出現比率という素性は，ある品詞単独の情報を表すことはできるが，複数の品詞間の共起関係を表現することはできない．人間が自然な文に翻訳する場合，全体の表現や前後の単語や句を考慮しながら，英文を翻訳している．しかし，機械翻訳では，人間のように十分に考慮することができると限らない．

そこで，共起関係（品詞列）を意識し，この品詞 Ngram を素性として利用することにした．素性ベクトルは，品詞 Ngram を素性名とし，素性値を 1 とする成分で構成する．訳文中に同じ品詞 Ngram が複数回出現した場合でも素性値は 1 とする．

単語 3gram 機械翻訳文の流暢さをより直接的に表現するために，単語 3gram を利用する．単語 3gram は，単語の組合せであるため，どのような単語がどのように組み合わされているかまで考慮できる．このため，機械翻訳文と人間翻訳文の流暢さを識別するのに有効な素性であると考えられる．

素性ベクトルは，機械翻訳文に出現した単語 3gram を素性名とし，素性値を 1 とする成分で構成する．素性値は，品詞 Ngram の場合と同様に，機械翻訳文中に同じ単語 3gram が複数回出現した場合でも素性値は 1 とする．

素性の組み合わせ 上記の言語的特徴を組み合わせた素性についても検証する．具体的には，品詞出現比率，品詞 Ngram，単語 3gram の 3 種類を組み合わせた場合と，品詞 Ngram と単語 3gram の 2 種類を組み合わせた場合について実験を行う．

4 実験と考察

4.1 実験方法

評価システムの有効性を検証する実験には，ロイター日英対訳コーパス [9] から抽出した英文と和文それぞれ 12900 文を用いた．英文を市販の翻訳ソフトで翻訳した結果を機械翻訳文，対訳コーパスの和文を人間翻訳文とした．評価用データとして機械翻訳文から 500 文を抽出し，残りの 25300 文 (機械翻訳 12400 文，人間翻訳 12900 文) を学習データとした．評価用データとした機械翻訳文を流暢さの観点から 100 点満点で 3 名の評価者で採点し，その点数に該当する評価値（表 1）の平均を付与した．

サポートベクターマシンには TinySVM[2]を用いた．

[1] http://chasen-legacy.sourceforge.jp/

[2] http://chasen.org/ taku/software/TinySVM/

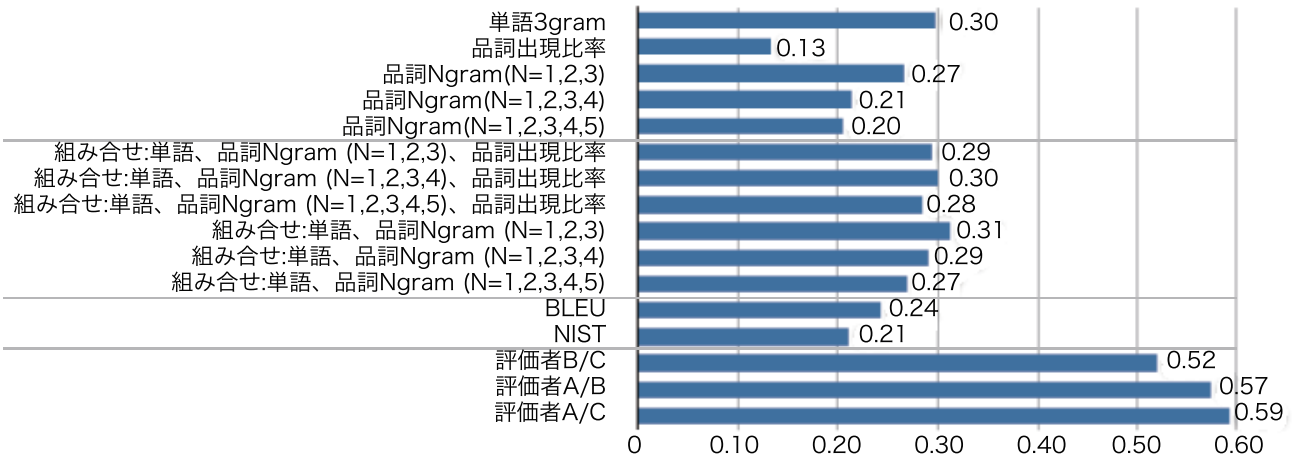

図 2: システム評価値と人手評価値との相関係数の比較

表 1: 流暢さの人手評価基準

評価値	基準
1	0 点〜24 点程度
2	25 点〜49 点程度
3	50 点〜74 点程度
4	75 点〜100 点程度

カーネル関数は 1 次多項式とし，他のパラメータは TinySVM の標準設定値を用いた．

評価システムの有効性の検証は，評価システムが出力するシステム評価値と人手評価値のスピアマンの順位相関係数を見ることによって行う．比較のため，BLEU，NIST と人手評価値との相関係数も求める．BLEU，NIST への入力は Chasen[3]による形態素解析結果とし，参照訳としては，ロイター日英対訳コーパスの和文を用いる．また，3 名の人手評価値間の相関係数を求め，そのうち最も高い相関係数を手法の上限とみなす．

4.2 実験結果と考察

図 2 に，システム評価値と人手評価値 (3 名の平均値) の相関係数，既存の評価システムと人手評価値 (3 名の平均値) の相関係数，人手評価値間の相関係数を示す．

[3]http://chasen-legacy.sourceforge.jp/

まず，提案手法で単語 3gram，品詞出現比率，品詞 Ngram の各素性を単独で用いて評価システムを構築した場合について見る．なお，品詞 Ngram を素性とする場合には，$N = 1, 2, 3$ の 3 通りの組み合せ，$N = 1, 2, 3, 4$ の 4 通りの組み合せ，$N = 1, 2, 3, 4, 5$ の 5 通りの組み合せを用いた．各素性を単独で用いた場合，単語 3gram のときに 0.30 と最も高い相関係数となり，品詞出現比率のときに 0.13 と最も低い相関係数となった．

次に，提案手法で素性を組み合わせて評価システムを構築した場合について見る．素性の組み合わせ方として，(1) 単語 3gram，品詞出現比率，品詞 Ngram を組み合わせた場合と，(2) 単独で用いたときに最も相関係数が低かった品詞出現比率を除いて，単語 3gram と品詞 Ngram だけを組み合わせた場合について相関係数を求めた．(1) の場合，品詞 Ngram の N の違いによる相関係数への影響はほとんど見られない．(2) の場合についても，品詞 Ngram の N による影響は大きくない．(1) の場合と (2) の場合とを比べても，相関係数の差は大きくない．また，(1) の場合も (2) の場合も，素性を単独で用いた場合と比べると相関係数の変動幅は小さい．これらのことから，素性の組み合せが有効であることが示唆される．

既存の評価システムと人手評価値との相関係数については，BLEU は 0.24，NIST は 0.21 であった．提案手法による評価システムと人手評価値との相関係数は，品詞出現比率を素性として構築した場合を除い

て，BLEU，NIST よりも高い．なお，人手評価値間の相関係数で最も高い値は 0.59 であった．

5 おわりに

本稿では，開発コストの低い回帰モデル [8] において，機械学習に用いる素性を変えて運用コストを抑えた評価システムを構築し，その評価精度の有効性を検証した．実験ではシステム評価値と人手評価値との相関係数を求め，BLEU，NIST と人手評価値との相関係数と比べた．実験の結果，品詞出現比率を素性として構築した評価システムを除いた他のすべての評価システムと人手評価値との相関係数は，BLEU，NIST よりも高かった．これにより，運用コストを低く抑えた場合でも有効な評価精度を保てることが分かった．

今回の実験では，評価システムの有効性を 1 つの機械翻訳でしか検証することができなかった．他の複数の機械翻訳を対象として有効性を検証することが今後の課題である．

参考文献

[1] C. B. Quirk. Training a Sentence-Level Machine Translation Confidence Measure. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 825–828, 2004.

[2] J. S. Albrecht and R. Hwa. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 880–887, 2007.

[3] J. S. Albrecht and R. Hwa. Regression for Sentence-Level MT Evaluation with Pseudo References. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 296–303, 2007.

[4] M. Paul, A. Finch, and E. Sumita. Translation quality prediction using multiple automatic evaluation metrics. 言語処理学会第 13 回年次大会発表論文集, pp. 95–98, 2007.

[5] S. Corston-Oliver, M. Gamon, and C. Brockett. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 148–155, 2001.

[6] M. Gamon, A. Aue, and M. Smets. Sentence-level MT evaluation without reference translations: Beyond language modeling. In *Proceedings of EAMT 10th Annual Conference*, pp. 103–111, 2005.

[7] 田中元貴, 南條浩輝, 吉見毅彦. 機械翻訳文と人間による翻訳文で構築した識別器による機械翻訳システムの自動評価. 言語処理学会第 14 回年次大会発表論文集, pp. 865–868, 2008.

[8] Alex Kulesza. A learning approach to improving sentence-level mt evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, 2004.

[9] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning japanese-english news articles and sentences. In *ACL-2003*, 72-79, 2003.