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

田中 元貴　　南條 浩輝　　吉見 毅彦

龍谷大学 理工学部 情報メディア学科

1　はじめに

近年，機械翻訳システムの品質自動評価に関する研究が活発になってきており，これまでに様々な自動評価手法が提案されている [1]．従来の自動評価手法の中には，機械翻訳システムによる訳文が良い翻訳であるかそうでない翻訳であるかを識別する識別器を機械学習によって構築し，この識別器を利用して自動評価を行う方法がある [2, 3, 4]．これらの手法では，良い翻訳とは人間による翻訳であり，そうでない翻訳とは機械翻訳システムによる翻訳であると仮定される．このような仮定の下で，対訳コーパスにおける人間による訳文（以下，人間訳と呼ぶ）と，原文を機械翻訳システムで翻訳して得られる機械翻訳文（以下，MT 訳）を訓練事例として識別器が構築される．この識別器を用いて，評価対象の MT 訳 が良い翻訳（人間による翻訳）であるかそうでない（機械翻訳システムによる翻訳）かの二値判定が行われる．

本研究では，このような先行研究に倣い，機械翻訳システムの自動評価を行う．具体的には，識別器に入力された評価対象の MT 訳が機械翻訳システムによるものであると正しく判定されるか，人間によるものであると誤って判定されるかの識別正解率に基づいてシステムレベルの自動評価を行う．文献 [2] によると，MT 訳はその翻訳品質が向上すれば人間訳との識別が困難になるため，識別器の正解率は低下するはずであると予想されている．もし，本稿で提案する手法において MT 訳の翻訳品質が向上するにつれて識別正解率が低下することが確認されれば，識別器の正解率が低下するように機械翻訳システムを修正していけばよいという指針をシステム開発者に与えることができる．なお，翻訳品質の主な評価尺度として適切さと流暢さがある [5] が，本稿では，流暢さの自動評価を対象とする．

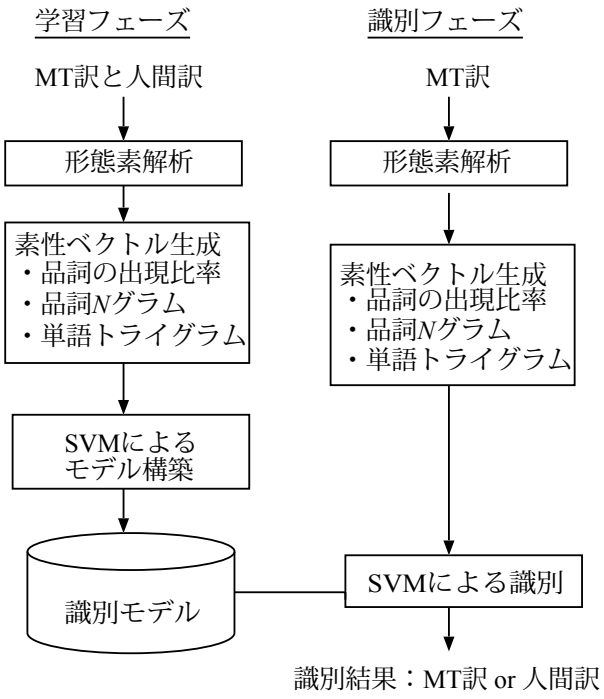

図 1: MT 訳と人間訳の識別器の概要

2　MT 訳と人間訳の識別器による自動評価

評価対象の MT 訳が良い翻訳（人間による翻訳）であるかそうでない（機械翻訳システムによる翻訳）かの二値判定を行う識別器の概要を図 1 に示す．識別器を構築する学習フェーズでは，機械学習に利用する素性を MT 訳と人間訳から抽出するために，形態素解析を行う．形態素解析には,「茶筅」[1]を利用する．形態素解析結果から，MT 訳と人間訳を識別するための素性を抽出し，それらを成分とする素性ベクトルを生成する．本研究で着目した素性については，3 節で説明する．機械学習にはサポートベクターマシンを用いる．

識別フェーズでは，評価対象の MT 訳に対して，学

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

習フェーズでの処理と同様の処理行って素性を抽出し，その素性に基づいて，その MT 訳が人間による翻訳（人間訳）であるか，機械翻訳システムによる翻訳（MT 訳）であるかを判定する．

3　識別器の構築に用いる素性

MT 訳の流暢さの自動評価を行う場合，人間訳の流暢さと MT 訳の流暢さの違いを適切に表現できる手がかりを機械学習で用いる素性として選ぶ必要がある．本研究では，品詞の出現比率，品詞 N グラム（$N = 1, 2, 3$），単語トライグラムに着目し，これらの素性を個別に用いて識別器を構築する．

3.1　品詞の出現比率

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 [6]．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，機械翻訳システム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．具体的な例を以下に示す．“failure” は,「失敗」や「〜をしないこと」などに訳すことができる．下記の英文を和訳するとき “failure” の品詞転換をしないと MT 訳のように「失敗」となり不自然な翻訳になる．“failure” を名詞のままでなく動詞に転換して翻訳することにより，人間訳のように品詞転換前に比べてより自然な翻訳ができる．

英文： His failure to have contact with the other side was fatal to him.

MT 訳： 他のサイドと連絡をとっていることについての彼の 失敗 は彼にとって致命的でした．

人間訳： あいつが相手方と接触 しそこなった のが，結局は奴の命取りとなった．

このような品詞転換は，上記の例のような名詞と動詞の間だけでなく，形容詞と名詞の間などでも行われることがある．例えば，“very good” は「非常によい」という意味である．品詞転換を行わずに翻訳すると下記の MT 訳のようになり，自然な表現ではない．しかし，“very good” を形容詞ではなく名詞に転換することにより，下記の人間訳のように自然な表現で翻訳することができる．

英文： very good angler

MT 訳： 非常に良い 釣り師

人間訳： 魚釣りの 名人

以上のようなことから，本研究では，人間訳の流暢さ（自然さ）と MT 訳の流暢さ 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,「茶筅」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．

3.2　品詞 N グラム

3.1 節で述べた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，機械翻訳システムでは，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 N グラム（$N = 1, 2, 3$）を素性として利用することにした．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，品詞 N グラムを素性名とし，素性値を 1 とする成分で構成する．訳文中に同じ品詞 N グラムが複数回出現した場合でも素性値は 1 とする．

3.3　単語トライグラム

3.2 節では品詞 N グラムに着目したが，ここでは，訳文の流暢さをより直接的に表現するために，単語トライグラムを利用する．単語のトライグラムは，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，MT

訳と人間訳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，訳文に出現した単語トライグラムを素性名とし，素性値を 1 とする成分で構成する．素性値は，品詞 N グラムの場合と同様に，訳文中に同じ単語トライグラムが複数回出現した場合でも素性値は 1 とする．

4 実験と考察

本節では，次の二点について検証するために行った実験の結果について述べる．

1. 本研究の識別器は，人間訳と MT 訳の識別をどの程度の正解率で行うことができるのか．
2. MT 訳の流暢さが上がるにつれて，本研究の識別器の正解率が低下するかどうか．

実験には，ロイター英日対訳コーパス [7] から抽出した 12900 文を使用した．この対訳コーパスの和文を人間訳とした．また，この対訳コーパスの英文を 3 つの市販の翻訳ソフトによって翻訳した結果を MT 訳とした．以下，これらの MT 訳を MT 訳 A，MT 訳 B，MT 訳 C とする．

サポートベクターマシンによる機械学習には TinySVM [2]を利用した．カーネル関数は 1 次の多項式とした．カーネル関数以外のパラメータは，標準設定されている値を使用した．

4.1 素性の種類と機械翻訳システムと識別器の正解率

識別器の構築に使用した素性と機械翻訳システムの違いによって識別正解率がどのように変化するかを検証する．各 MT 訳と人間訳を合わせた合計 25800 件を事例集合とした．この事例集合を 5 分割し，交差検定を行った．試験事例の MT 訳に対する識別正解率を表 1 に示す．数値は 5 分割交差検定の平均値である．

表 1 より，品詞の出現比率を素性とした場合，MT 訳 A に対する識別正解率は約 65%であり，MT 訳 B と MT 訳 C と比べると低いことがわかる．MT 訳 A に対する識別正解率が低い原因については，今のところ明らかではない．今後分析を進めていく必要がある．

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

表 1: 識別正解率

	MT 訳 A	MT 訳 B	MT 訳 C
品詞の出現比率	64.88%	84.53%	86.43%
品詞 N グラム	91.08%	96.14%	97.41%
単語トライグラム	97.11%	97.90%	98.18%

品詞 N グラムを素性として用いると，91%〜97%程度の識別正解率が得られる．単語トライグラムを用いた場合，識別正解率は 97%〜98%と非常に高くなる．これらのことから，品詞 N グラムと単語トライグラムが MT 訳と人間訳を識別するのに特に有効な素性であるといえる．

4.2 流暢さの人手評価値と識別正解率

MT 訳の流暢さと識別器の正解率を比較し，MT 訳が流暢な表現であればあるほど識別正解率が低下するかどうかを検証する．この検証を行うために，MT 訳 A に対して，流暢さの観点から日本人評価者 2 名による人手評価を行った．評価対象文として，MT 訳 A の 12900 文から 500 文を無作為に抽出した．MT 訳の日本語としての流暢さを 100 点満点で採点しその点数に該当する評価値（表 2）を付与するよう評価者に指示した．評価者には，英文は示さず MT 訳だけを示した．

表 2: 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

MT 訳 A の 500 文を人手評価値の上位群（評価値 3，評価値 4）と下位群（評価値 1，評価値 2）の 2 群に分け，それぞれの群での識別正解率を求めた．2 名の評価者ごとに，人手評価値と識別正解率の関係をそれぞれ表 3 と表 4 に示す．

表 3 から，下位群での識別正解率より上位群での識別正解率のほうが 7〜12 程度低いことがわかる．また，表 4 から，下位群より上位群のほうが 5〜13 程度低くなっていることがわかる．この識別正解率の低下

表 3: 評価者 a による評価値と識別正解率の関係

	下位群	上位群	差
品詞の出現比率	66.23%	56.82%	9.41
品詞 N グラム	86.69%	77.27%	12.42
単語トライグラム	98.03%	90.91%	7.12

表 4: 評価者 b による評価値と識別正解率の関係

	下位群	上位群	差
品詞の出現比率	69.41%	56.88%	12.53
品詞 N グラム	90.29%	85.00%	5.29
単語トライグラム	99.12%	93.75%	5.37

傾向から，提案手法をシステムレベルでの自動評価に利用できそうであると考える．

5 結論

本稿では，人間訳と MT 訳を識別する識別器の構築に有効な素性を提案した．さらに，MT 訳に人手評価値を付加して上位群と下位群の二つの群に分割し，群と識別正解率の関係を検証した．その結果，これらの素性を用いた識別器の正解率は，評価者による評価値が上がるにつれて低下するという仮説を実証することができた．したがって，機械翻訳システム開発では，提案した素性による識別器の正解率が低下するように機械翻訳システムを修正していけばよいと言える．

評価者による人手評価は 1 種類の MT 訳に対してしか行うことができなかったため，他の MT 訳でも検証することが今後の課題である．

参考文献

[1] 安田圭志, 隅田英一郎. 機械翻訳の研究・開発における翻訳自動評価技術とその応用. 人工知能学会誌, Vol. 23, No. 1, pp. 2–9, 2008.

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

[3] 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.

[4] A. Kulesza and S. M. Shieber. A Learning Approach to Improving Sentence-Level MT Evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 75–84, 2004.

[5] 隅田英一郎, 佐々木裕, 山本誠一. 機械翻訳システム評価法の最前線. 情報処理, Vol. 46, No. 5, pp. 552–557, 2005.

[6] 中村保男. 英和翻訳の原理・技法. 日外アソシエーツ, 2003.

[7] M. Utiyama and H Isahara. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 72–79, 2003.