E-013

日本語話者の中国語運用能力との比較による日中翻訳評価の検討

A Preliminary Study on Evaluating Japanese-to-Chinese Translations Based on Chinese Language Skills of Native Japanese Speakers

胡 新輝† 安田 圭志† 竹澤 寿幸† 菊井 玄一郎†
Xinhui Hu Keiji Yasuda Toshiyuki Takezawa Genichiro Kikui

1．まえがき

機械翻訳システムの評価は、時間と労力が掛かる作業である。自動的且つ信頼性が高い評価手法が求められている。ATR では、音声翻訳システムの翻訳能力を分かりやすく数量化するものとして、人間の言語運用能力（TOEIC スコア）と比較して数量化する手法を提案し、日英方向についてその有効性を確認した［１、２］。本稿では、TOEIC の代わりに HSK（漢語水平考試）を使うことで同じ方法が日中翻訳にも適用可能か検討した。その試験を受けた複数の日本人に旅行会話文を漢字文とピンインで中国語に訳させたもの集め、訳文の評価スコアと HSK 級の関係を調べた。その結果、日中方向に相関が高く、適用できる見込みが得られた。

2．HSK とは

近年、中国の経済発展と共に中国語のブームが世界中で巻き起こっている。以前中国語を学ぶ者はほとんど中国の言語と文化に対する趣味や関心からであったが、いまでは、中国語を学ぶ動機は生計を立てるための実際の必要から出ている。このような背景から、中国内外に各種の中国語検定試験が設けられ、目的に合わせて試験を受けることが可能である。その中に、中国の教育部（日本の文部科学省に相当）が設けた中国語（漢語）を母語としない中国語学習者のための唯一公認の中国語能力認定標準化国家試験“漢語水平考試（HSK）”がある「3」。これは、１９８０年代から始まり、中国内外で実施されている。年間４０万人の受験者がいるという。試験は、中国国内の２７の主要都市で毎年５月、７月、１２月に年３回実施しているほか、海外では日本を含む２７ヵ国６７都市で実施されている。現在、HSK の検定証明は、数多くの中国の大学への留学、及び沢山の外資系会社が中国人以外の人を採用する際に重要な根拠となっている。

試験は、ヒアリング、文法、読解の三つの部分に分けられている。HSK 等級は、基礎（１-２級）、初等と中等（３-８級）、高等（９-11 級）の３種からなる。高等は、中国国内で年１回（５月）のみ実施され、試験問題も他と異なる。そのほかの級は、受験者が同一の試験問題に取組み、受験者の得点（スコア）をそれぞれの等級に換算し（数字の大きい方が上級）、規定の等級に達すれば、中国国家漢語水平考試委員会より「漢語水平証書（ＨＳＫ証書）」が授与される。

3. 翻訳評価方法

3.1 データ収集と準備

本文では、異なる HSK 等級を持ち、日本語を母語とする被験者を集めた。彼らに音声で日本語の文を聞かせながら、中国語の漢字文及びピンイン文に訳して貰う。その後、この訳された漢字文書に形態素解析をして、セグメンテーションされたデータを得た（以後 word と記す）。ピンイン文は、一文字毎に書かれる。ここで、ピンイン文と対応させるために、漢字文を文字ごとに分解した実験も行った（以後、syllable と記す）。

一方、スコアリングするために複数の中国語正解訳（リファレンスデータ）が必要である。今回は１文につき、正解訳を１３文作成した。これらも、形態素解析プログラムで解析し、セグメンテーションされた漢字文とピンイン文を得た。

各受験者のセグメンテーションされた漢字文とピンイン文をそれぞれ、スコアリングルーチンに通して、訳文のスコアを得る。

図１に、この流れを示す。

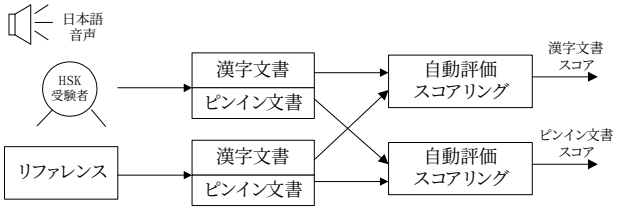

図１　データ収集と準備

3.2 自動評価する方式

安田らは、TOEIC（英語検定試験）の受験者が作成した日―英方向の訳文に対して、DP 及び n-gram 方式の自動評価方法を行い、それらの方式の有効性を確かめた[２]。そこで提案されている二つの方法を今回の予備検討で用いることにした。

3.2.1 DP-based 法

$$S_{DP} = \frac{1}{N_{total}} \sum_{j=1}^{N_{total}} \max_{i} \left(\frac{T_{ij} - S_{ij} - I_{ij} - D_{ij}}{T_{ij}}, 0 \right) \quad (1)$$

但し、N_{total} は、テストセットにある文の総数、T_{ij} は、リファレンスファイル i の j 番目の文の単語数。S_{ij} は、j 番目のテスト文とリファレンスファイル i を比較する時に置き換えられた数、I_{ij} は同様に比較する時に挿入された数、D_{ij} は比較する時に削除された数である。

3.2.2 N-gram-based 法

N-gram-based 法として BLEU［４］を採用する。この方法のスコアリング式は、以下のようになる。

$$S_{BLUE} = \exp\left\{ \sum_{n=1}^{N} w_n \log(p_n) - \max\left(\frac{L^*_{ref}}{L_{sys}} - 1, 0 \right) \right\} \quad (2)$$

但し、N は n-gram の最大長、w_n はその逆数である。L_{sys} はテストデータにあるスコアリングすべき訳文の単

† ATR 音声言語コミュニケーション研究所

語数、L^*_{ref} はテスト文の長さと一番近いリファレンス文の長さである。p_n は、以下の式で計算する。

$$P_n = \frac{\sum_i \text{訳文}i\text{とリファレンス}i\text{で一致した}n\text{-}gram\text{数}}{\sum_i \text{訳文}i\text{中の全部}n\text{-}gram\text{数}} \quad （3）$$

4．予備実験と結果

本試験では、BTEC（Basic Travel Expression Corpus）というコーパスのテストセットから１ファイル（５１０文）を選び、テスト文として使った。また、HSK の受験者１０人を集めた。内訳は、４級者が２人、５，６級は、それぞれ１人、７級は２人、８級は３人、９級は１人であった。

スコア計算は、漢字文とピンインに分けて行う。複数人の HSK 級のスコアは、それらの平均値とした。

表 1は、実験用データと HSK 級との相関係数である。これによって、N-gram-based 方式のとき、N-gram=４を選択し、実験を行った。

表 1　実験データの相関係数

N-gram	1	2	3	4	5
N-gram,word	0.8758	0.9126	0.9219	0.9232	0.9007
N-gram,syllable	0.8448	0.8983	0.9166	0.9243	0.9253
N-gram,pinyin	0.6437	0.6779	0.7114	0.7249	0.7275
DP, word	0.5748				
DP, syllable	0.7722				
DP, pinyin	0.9032				

図２、３に、上述の二つの方法による、それぞれ漢字文とピンイン文のスコアリング結果を示す。これらの結果から得られた知見を次に記す。

- 漢字文とピンイン文の何れのスコアも、HSK 級が上がると上昇する傾向がある。スコアの絶対値は漢字文の方がピンイン文より大きい。
- ピンイン文に対しては DP 法の方が相関が高く、漢字文に対しては N-gram-based 法の方が相関が高い。

5．結論

本文では、異なる中国語能力検定試験（HSK）のレベルを持つネイティブの日本人が訳した中国語文書に対して、N-gram、DP などの分析方法によってスコアリングを試みた。実験の結果から、これらのスコアは、受験者の実際に持つ HSK レベルを反映している傾向があることが分かった。従って、今後、機械翻訳システムの自動評価にこのデータを活用することが有効と期待できる

今後、この方式を機械翻訳システムの評価に導入し、その有効性を確かめていく。また、受験者の数を増やし、各レベルでのばらつきや信頼性の議論を行う予定である。

６．謝辞　本研究は総務省の研究委託「携帯電話等を用いた多言語自動翻訳システム」により実施したものである。

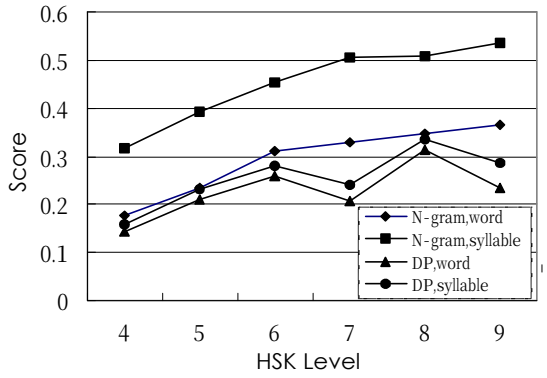

図２　漢字文のスコア

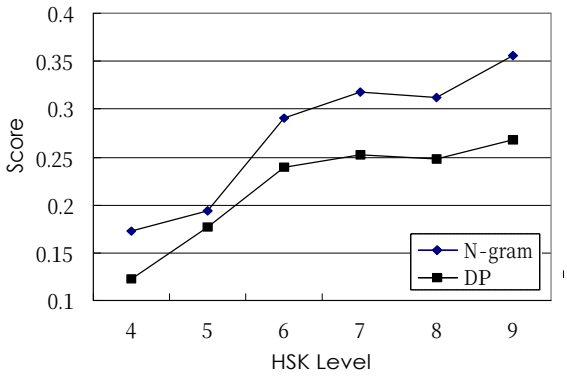

図３　ピンイン文のスコア

7．参考文献

[1] 菅谷史昭、竹澤寿幸、横尾昭男、山本誠一、“音声翻訳システムと人間との比較による音声翻訳能力評価手法の提案と比較実験、”信学論、Vol. J84-D-II, No.11, pp. 2362-2370, 2001

[2] Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S., and Yanagida, M., “Application of Automatic Evaluation Methods to Measuring a Capability of Speech Translation System,” Proc. 10th Conference of the European Chapter of the Association for Computational Linguistics, pp. 371-378, 2003.

[3] HSK: http://www.hsk.org.cn/、http://www.jyda-ie.or.jp/hsk/top.htm

[4] Papineni, K., Roukos, S., Ward, T., Zhu, W-J., “Bleu: a Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.