Vol.2011-NL-201 No.7
Vol.2011-SLP-86 No.7
2011/5/17

インターネット上の英日統計的機械翻訳サービスの誤り分析

星野　翔[†]　　建石　由佳[†]

インターネット上の英日機械翻訳を行うサービスで統計的機械翻訳をベースとしたものが増えている．英語と日本語の間での翻訳には誤りが多く，翻訳の質に大きく影響しているが，従来用いられてきた評価手法は数値による評価尺度であり，システムのどの部分の誤りが結果に影響しているのか明らかではない．
本研究では，インターネット上の英日機械翻訳サービスを使って対訳データを翻訳し，十分な質の文を出力できているのか，またどのような誤りが影響しているのかを人手で分析し，後編集プログラムによる改善を試みた．

An Error Analysis of English-Japanese Statistical Machine Translation Services on the Internet

Sho Hoshino[†]　and Yuka Tateisi[†]

Today, English-Japanese statistical machine translation (SMT) services are available online. However, it is not yet clear how good their translation qualities are, and what types of errors affect the quality of their translation. We propose a method for machine translation evaluation using detailed error analysis. We applied the method to popular SMT systems available online, and also suggest the possibility of improving their translation quality by rule-based post-edit programs.

1. はじめに

インターネット上の機械翻訳を行うサービスで，従来のルールベース機械翻訳ではなく，統計的機械翻訳の手法を用いるものが増えている．これらのサービスは時や場所を選ばず利用でき，利便性が向上しているが，これらのサービスが十分な質の文を出力できているのかどうか，詳しく評価する必要がある．しかし，従来用いられてきた評価手法は数値による評価尺度であり 1)，システムのどの部分の誤りが結果に影響しているのかは明らかではない．本研究は，既存の英日対訳データから英日統計的機械翻訳サービスを誤り分析し，どのような誤りが影響しているのか調査することを目的とした．

まず調査のために，新聞記事の英日対訳データである日英新聞記事対応付けデータ 2)とロイター日英記事の対応付け 3)から 1 対 1 対応の文を抽出した．またそれぞれ 1 文を Google 翻訳と Bing Translator で機械翻訳した．これにより対訳文と機械翻訳文の 1 文ごとの比較が可能になり，誤りが抽出できた．抽出された誤りを，あらかじめ設定した誤りの分類方法にしたがって人手で分類し，誤り傾向の実験結果を得た．また実験結果をもとに，後編集プログラムによる改善を試みた．

2. 関連研究

本研究と同じく，統計的機械翻訳を誤り分析したものには，Vilar4)や Ukai5)の研究がある．Vilar は英西と中英，Ukai は日英について，それぞれ独自の統計的機械翻訳システムを調べているが，誤りの分類方法に関して，Vilar が大分類の下に小分類を置いているのに対し，Ukai が細かく分類を行っているという違いがある．

ルールベース機械翻訳を誤り分析したものには，渡辺 6)や成田 7)の研究がある．渡辺は，係り受け解析器を使用して正解文と機械翻訳文を比較することにより，係り受けの誤り傾向を調べている．成田はテキストを分野別に調べている．

また，機械翻訳出力の人手評価と自動評価を比較した研究には，Koehn らの研究 1)がある．人手の評価には妥当性 adequacy と流暢性 fluency を，自動評価手法に BLEU10)を用いている．

[†] 工学院大学 情報学部
Faculty of Informatics, Kogakuin University

Vol.2011-NL-201 No.7
Vol.2011-SLP-86 No.7
2011/5/17

3. 実験方法

対訳文および機械翻訳文は，後の比較に適するよう，プログラムによって全角英数字を半角英数字に，全角括弧「」を半角 2 重引用符""に変換した．

3.1 機械翻訳サービス

本研究の実験では，Vilar や Ukai らの方法 4)5)とは異なり，独自の統計的機械翻訳システムではなくインターネット上の統計的機械翻訳サービスを使用した．英日統計的機械翻訳をベースにしている Google 翻訳 8)と Bing Translator9)で，それぞれ Google Translate API と Bing API を利用し，原文を英語，訳文を日本語に設定し，出力された結果を分析に利用した．

これによりシステムの設定や利用状況に左右されず，一般的に検証可能である状態をめざした．一方で，サービスの更新によって実験結果が再現できなくなる恐れがあるが，Google 翻訳の更新があった 2011 年 12 月 16 日から 2011 年 4 月 1 日 0 時 0 分時点までは実験結果が再現できることを確認した．

3.2 対訳データ

英日対訳データには，日英新聞記事対応付けデータ（JENAAD）2) と，ロイター日英記事の対応付け（REUTER）3)で，一対一文対応となっているものを，文字コードを Unicode に変換して使用した. 実験には, 対訳データ中の 15 万文対と 5 万文対から，先頭の 200 文対を選んで使用した．

以下の例では，原文に JANAAD を，機械翻訳に Google 翻訳を使用した．

3.3 誤りの分類

誤りの分類作業は人手で行った．誤りの分類には Vilar4)のように大分類から小分類に分けていく方法を用いた．大分類には Missing words, Word order, Incorrect words, Unknown words の 4 つを，小分類には，日本語では Incorrect words がよく出現し大分類のままでは曖昧であると考えて Incorrect word senses, Incorrect word forms, Incorrect words (others) の 3 つを使用した．また 1 文に複数の誤りがみられる場合は，当てはまる誤りすべてに属するものとした．

3.3.1 Missing words

原文にある語がどのような形態でも出力されていない場合に分類した．

The other stores sell only semitransparent blue or white garbage bags.
他の店は、半透明の青や白のごみ袋を販売しています。

3.3.2 Word order

機械翻訳と参照訳で語の位置が変わり，係り受け関係が異なる場合に分類した．

While fundamental change entails risk, we place our trust in the creativity, effort and dedication of people as the true sources of economic and social progress.
根本的な変化は、リスクを伴うが、我々は創造性、努力と経済的および社会的進歩の真の源泉としての人々の献身に我々の信頼を置く。

3.3.3 Incorrect words

以下 3 つの小分類で分類した．

Incorrect word senses

原文の語と意味が異なっている場合に分類した．

2. The international community is at the threshold of a new era, freed from the burden of the East-West conflict.
2。国際社会は東西対立の重荷から解放された新しい時代のしきい値にあります。

Incorrect word forms

日本語として語の形態が不自然な場合に分類した．

He must teach the younger generations the need for the freedom.
彼は若い世代の自由の必要性を教えなければならない。

Incorrect words (others)

その他，不要語が含まれるなどの場合に分類した．

The Government of Japan confirms that it is prepared to share such experiences with the Russian Federation in various areas including macroeconomic policy, reform of fiscal and financial systems and industrial structure, and promotion of small and medium sized enterprises.
日本政府は、それがマクロ経済政策、財政上及び金融システムや産業構造の改革、中小企業の振興を含む様々な分野でロシアとのこのような経験を共有する準備ができていることを確認します。

3.3.4 Unknown words

原文の語がそのまま出力された場合に分類した．

They agreed on the importance of restoring stable and longlasting exchange rate relationships.
彼らは安定して longlasting 為替レートの関係を復元することの重要性について合意した。

3.4 評価

対訳文に対する機械翻訳の質を測定するために, 自動と手動の両方で評価を下した．

自動評価手法には，BLEU と Word Error Rate (WER) を用いた．どちらの手法も，人間による参照訳と機械翻訳の候補訳の差を数値による評価尺度によって評価する．そのため日本語文の自動評価にはトークン化の前処理が必要であり，形態素解析エンジン MeCab の「わかち書き」オプションを使用した．

手動評価では，人手による 5-1 の 5 段階評価を用いた．適合性と流暢性を用いなかったのは，Koehn らの研究 1)によれば，評価方法として問題があるとされているためである． 評価基準は次の通りである．

Vol.2011-NL-201 No.7
Vol.2011-SLP-86 No.7
2011/5/17

5. 特に優れているもの
Unlike the map of Europe, the map of Czechoslovakia has not changed.
ヨーロッパの地図とは異なり、チェコスロバキアのマップが変更されていません。
4. 多少の誤りがあるが，十分なもの
2. The international community is at the threshold of a new era, freed from the burden of the East-West conflict.
2。国際社会は東西対立の重荷から解放された新しい時代のしきい値にあります。
3. 誤りが許容可能なもの
We welcome the recent release of two hostages in Lebanon.
我々は、レバノンでの 2 人の人質の最新のリリースを歓迎する。
2. 誤りによって原文の意味が大きく損なわれているもの
Economic issues have assumed new prominence.
経済問題は、新しい隆起を想定している。
1. 意味を成さない，支離滅裂，あるいは日本語として不適切なもの
Why is the world entering a new historic epoch?
なぜ新しい歴史的なエポックを入力して、世界のですか？

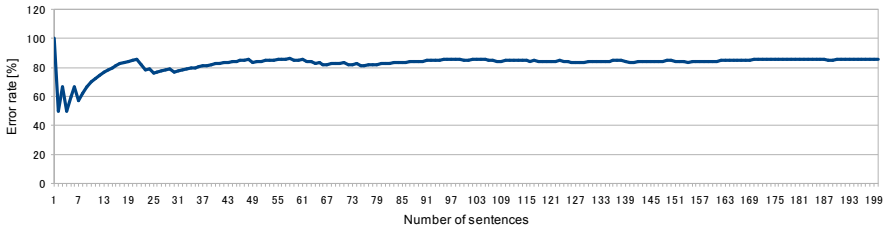

図 1　JENAAD と Google 翻訳での文全体の誤り率

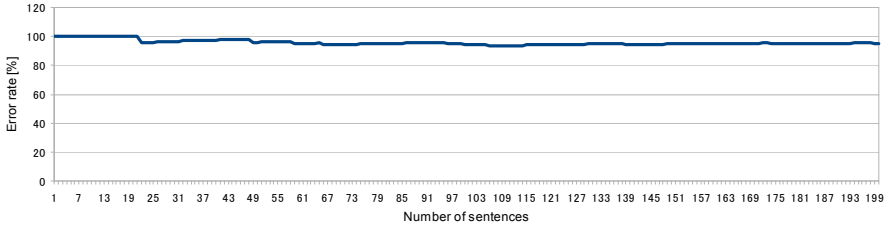

図 2　JENAAD と Bing Translator での文全体の誤り率

4. 実験結果

4.1 文全体の誤り率

すべての文について，誤りが 1 つでも含まれている割合を調べたのが**図 1～4** である．JENAAD と REUTER の双方で，Google 翻訳の誤り率が Bing Translator を下回る結果となった．一方，Google 翻訳と Bing Translator の両方とも，REUTER での誤り率が JENAAD での誤り率を上回っている．

また多少のばらつきはあるものの，4 つとも 200 文の時点で変化が収束しはじめており，誤りが 1 つでも含まれている割合の対訳データ全体の近似値が測定できたものと推測できる．

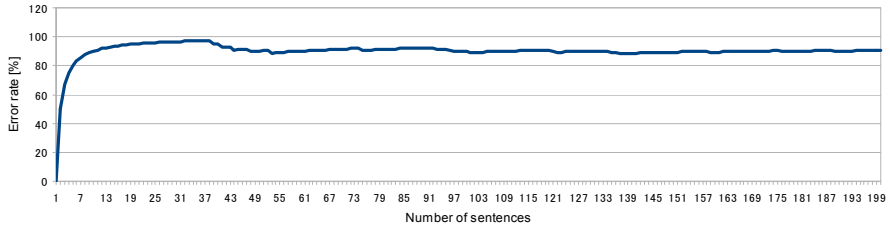

図 3　REUTER と Google 翻訳での文全体の誤り率

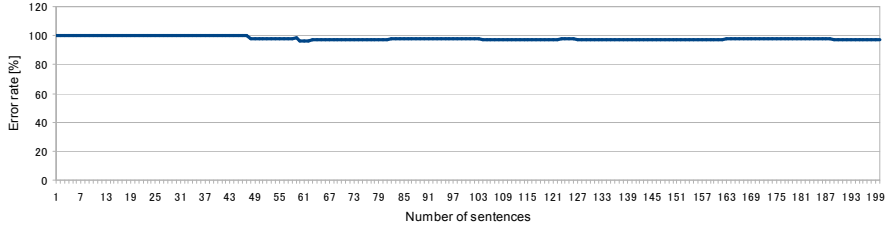

図 4　REUTER と Bing Translator での文全体の誤り率

Vol.2011-NL-201 No.7
Vol.2011-SLP-86 No.7
2011/5/17

表 1　分類別の誤り個数

	Missing words	Word order	Incorrect word senses	Incorrect word forms	Incorrect words (others)	Unknown words
JENAAD と Google翻訳	35	70	106	105	34	1
JENAAD と Bing Translator	89	95	141	75	24	2
REUTER と Google翻訳	15	36	75	26	13	1
REUTER と Bing Translator	35	53	83	30	5	0

表 2　評価結果

	手動評価	自動評価	
	スコア [1-5]	BLEU [0.0-100.0]	WER [%]
JENAAD と Google翻訳	3.145	29.65	35.5
JENAAD と Bing Translator	2.205	20.04	30.74
REUTER と Google翻訳	2.86	26.23	29.86
REUTER と Bing Translator	2.03	17.84	26.94

4.2 分類別の誤り個数と誤り率

分類別の誤り個数は**表 1** のようになった．Missing words と Incorrect words (others) の出現数は比較的少ない．Unknown words は 200 文に対して 1,2 個以下で，ほとんど出現しないことがわかった．

また分類別に誤りが含まれている割合を調べてたのが**図 5〜8** である．対訳データにより違いがあるが，Incorrect word senses が最も出現しやすく，次に Incorrect word forms と Word order が出現しやすかった．さらに，JENAAD と REUTER の 2 つデータで誤りの傾向が異なり，JENAAD では Incorrect words (others)と Unknown words を除きそれぞれがよく出現するのに対して，REUTER では Incorrect word senses がよく出現し，その他の誤りはあまり認められなかった．

Google 翻訳と Bing Translator との間では，Word order と Missing words の誤り率に違いがあった．Google 翻訳では Word order が 40%以下，Missing words が 20%以下なのに対し，Bing Translator では両方とも 40%を超えるときがあり，誤りやすい．

4.3 評価結果

人手評価による 5-1 の 5 段階評価と，自動評価手法の BLEU と WER によるスコアは**表 2** のようになった．誤り個数や誤り率と同じく，どの評価手法においても Google 翻訳のスコアが Bing Translator を上回り，また JENAAD のスコアは REUTER のスコアより高かった．

4.4 個別の誤り例

全体の結果とは別に，人手で評価した際にとくに目立つ誤りがあった．いずれの組み合わせでも，『no』『not』が訳されず文全体の意味が逆になってしまうという重大な

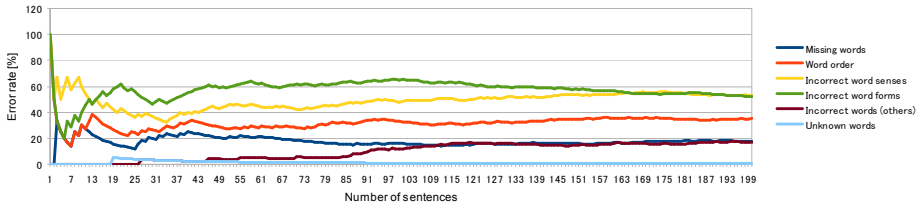

図 5　JENAAD と Google 翻訳での分類別の誤り率

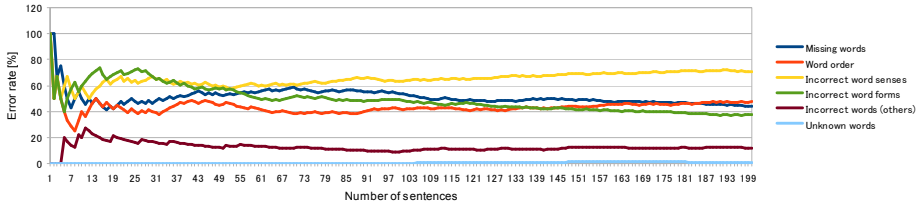

図 6　JENAAD と Bing Translator での分類別の誤り率

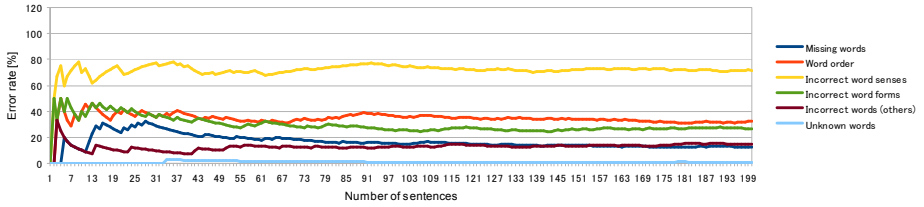

図 7　REUTER と Google 翻訳での分類別の誤り率

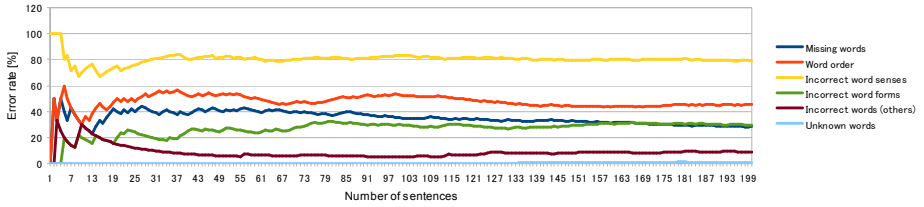

図 8　REUTER と Bing Translator での分類別の誤り率

Vol.2011-NL-201 No.7
Vol.2011-SLP-86 No.7
2011/5/17

表 3 編集前後での評価結果の比較（REUTER）

	Google翻訳		Bing Translator	
	BLEU	WER	BLEU	WER
後編集(1) 助詞	26.34	29.79		
後編集(2) 単位換算	26.27	29.88		
後編集(3) 曜日情報	26.08	29.57	17.85	29.57
後編集(2)+(3)	26.38	29.81		
編集前	26.23	29.86	17.84	26.94

誤りがあった．また『It is ～ for』や『It is ～ that』のような it を主語とする文の質が非常に悪かった．さらに REUTER のデータを使用した場合において,『$10.0 billion』を『$ 10.0 億ドル』,『840 million』を『840 万』と誤訳する単位換算の誤りが目立った．

5. 後編集プログラムと実験結果

5.1 後編集プログラム

以上の実験とは別に，ルールベースの後編集プログラムで機械翻訳文を編集することによって，単純な方法で機械翻訳文が改善可能かどうか調査した．ルールの部分には REUTER と Google 翻訳の組み合わせの実験結果から得られた，次のような 3 つのヒューリスティックを使用した．

(1) 『プライベートカラチコットン協会が（KCA は)』のような不要な助詞を『プライベートカラチコットン協会（KCA）は』のように削除する．
(2) 『95000000000 ドル』のような単位換算の誤りを『950 億ドル』と訂正する．
(3) 『The Times reported on Wednesday』を『タイムズ紙が報じた。』と翻訳するような曜日の翻訳抜けが発生している文の先頭に,『水曜日に（中略）タイムズ紙が報じた。』と曜日情報を追加する．

後編集プログラムの評価には,3.4 と同じく自動評価手法の BLEU と WER を用いた．

5.2 後編集の結果

表 3 は 3.4 の後編集プログラムを使用して後編集をした結果である．表中の（1）～（3）はそれぞれ 3.4 の（1）～（3）の誤りに対応する．これらはすべて REUTER の対訳データを使用した場合で，JENAAD では同様の後編集可能な誤りが見つからなかった．また Bing Translator では（3）曜日情報の誤りを除いて，同様の誤りが見つからなかった．

（1）～（3）の後編集を個別に行った場合，Google 翻訳と（3）曜日情報の誤りの組み合わせ以外では，編集前と比べてスコアが上昇した．一方で，スコアが上昇した後編集の（2）と（3）を Google 翻訳を組み合わせた場合，BLEU では最も高いスコアを得たが，WER では編集前と比べてスコアが下がってしまった．

6. 考察

今回の実験により, いくつかのことが明らかとなった. まず Google 翻訳の質は Bing Translator の質より高かった. Google 翻訳の質の平均は, ほとんど 3.4 の手動評価で「3. 誤りが許容可能なもの」と設定したものだった．しかし同じ基準で「4. 多少の誤りがあるが，十分なもの」と設定した評価からは遠く，まだまだ十分な質の文を出力しているとは言えない．Bing Translator の質の平均は，同様の基準で「2. 誤りによって原文の意味が大きく損なわれているもの」と設定したもので，十分な質には程遠い．

また誤り率について, 1 つでも誤りが含まれている文について調べる場合に限れば, 200 文を調べることによってある程度有効なデータを得られることがわかった．もちろん，さらにデータ量を増やした場合にきちんと収束していくのかどうか，また分類別の誤り率でも収束するのかなど，さらなる検証が必要である．

誤りの分類に関しては, Google 翻訳と Bing Translator で, Word order と Missing words の誤りの出現率が大きく異なることがわかった．日本語と英語の間では語順が大きく異なるため，Word order の誤り率は翻訳の質に大きく影響する．また翻訳されない Missing words が有れば，翻訳文全体の意味を損なってしまうため，これも質に大きく影響する．この 2 つの誤り率の低さが Google 翻訳の質の高さに大いに関係している．一方で Google 翻訳と Bing Translator でほぼ同一の訳文を出力することもある．これらのことから，Google 翻訳は統計的機械翻訳をベースとしているが，この 2 つの誤りを減らすために，翻訳の仮定で何らかの特別な措置をとっていると推測できる．

しかし，どのような誤りが影響しているのかを調べるという目的は上手く達成できたものの, 小分類において Incorrect word senses と Incorrect word forms の区別が曖昧であるなど，人手の分類基準をいかにして正確に保つかという課題が残った．分類に必要な作業量が多かったことも含めて，将来的には，自動化された誤り分類と人手による誤り分析によってシステムを改善していくことが望ましいと考える．

さらに翻訳サービスの違いと同様に，対訳データによって翻訳の難しさや話題が大きく異なり，誤りの分類結果に影響するということがわかった．今回は 2 つの対訳データを使って実験したが，一般的な誤りの傾向を求めるには不十分だった．今後，この結果がこの 2 つのデータにのみ限定されるのかどうか，別の対訳データを使って調べたい．

最後に後編集プログラムによる改善では，自動評価手法の BLEU と WER の双方においてスコアの改善がほとんどみられなかった．この原因は，個別の事例に対しては上手く編集できたが，全体の誤り率に対する出現率を考慮していなかったためだと考

Vol.2011-NL-201 No.7
Vol.2011-SLP-86 No.7
2011/5/17

えられる．誤りの分類結果から，Incorrect word senses や Incorrect word forms に当たる誤りを優先的に後編集することにより，翻訳結果を大幅に改善できることがわかった．しかし，そのような誤りを単純なルールベースの後編集プログラムで上手く編集するためには，もっと誤りの事例を増やして一般化していくことが必要である．

7. おわりに

本研究はインターネット上の英日統計的機械翻訳サービスを誤り分析して，十分な質の文を出力できているのか，またどのような誤りが影響しているのか調査した．その結果，Google 翻訳は十分な質の文を出力できているとは言えず，Bing Translator は十分な質には程遠かった．また誤りの分類別の詳細なデータを得ることが出来た．それにより，翻訳サービスごとに誤り方も異なることがわかったが，調査に使用した日英新聞記事対応付けデータとロイター日英記事の対応付けの違いも影響した．

誤り分析から得られたルールによって，後編集プログラムによる翻訳文の改善を試みたが失敗した．英日統計的機械翻訳を改善していくためには，自動化された誤り分類による，大規模な誤り分析が必要である．

謝辞 本研究について，東京大学辻井研究室の呉先超特任研究員と国立情報学研究所の宮尾祐介准教授から，多大なご助言をいただいた．

参考文献

1) Koehn, P. and Monz, C.: Manual and Automatic Evaluation of Machine Translation between European Languages, in *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT) 2006 Workshop on Statistical Machine Translation*, pp.102-121 (2006).
2) 日英新聞記事対応付けデータ (JENAAD)
http://mastarpj.nict.go.jp/~mutiyama/jea/index-ja.html
3) Alignment of Reuters Corpora
http://mastarpj.nict.go.jp/~mutiyama/jea/reuters/index.html
4) Vilar, D., Xu, J., D'Haro, L. F., and Ney, H.: Error Analysis of Machine Translation Output, in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp.697-702 (2006).
5) Ukai, I.: Error Analysis of the English-Japanese Statistical Machine Translation System, Bachelor's degree thesis with honours, School of Informatics, University of Edinburgh (2008).
6) 渡辺桂子: 渡辺桂子: 英日機械翻訳における誤りの傾向, 工学院大学卒業論文 (2009).
7) 成田一: 翻訳ソフトの実力評価, 情報処理学会研究報告, 第 125 回自然言語処理研究会 NL125, pp.123-130 (1998).
8) Google 翻訳
http://translate.google.co.jp/
9) Bing Translator
http://www.microsofttranslator.com/
10) Papineni, K., Roukos, R., Ward, T., and Zhu, W. J.: BLEU: a Method for Automatic Evaluation of Machine Translation, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.311-318 (2002).
11) Leusch, G., Ueffing, N., and Ney, H.: A Novel String-to-String Distance Measure With Applications to Machine Translation Evaluation, in *Proceedings of MT Summit IX*, pp.240-247 (2003).
12) MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.sourceforge.net/
13) Utiyama, M. and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences, in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics Conference (ACL)*, pp.72-79 (2003).
14) Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J.: Further meta-evaluation of machine translation, in *Proceedings of the Third Workshop on Statistical Machine Translation (WMT08)*, pp.70-106 (2008).
15) Koehn, P.: Statistical Machine Translation, Cambridge University Press, UK (2010).