日本語単語分割が統計的機械翻訳に与える影響の評価

星野　翔　　宮尾　祐介
総合研究大学院大学　国立情報学研究所
{hoshino, yusuke}@nii.ac.jp

1　はじめに

日本語は通常, 単語間に空白を挿入する分かち書きがされていない自然言語の一つである. このような言語で単語ごとの処理を行うためには, 適宜空白を挿入し単語間の境界を明示化する単語分割が必要である. しかし従来の統計的機械翻訳では, 形態素解析器の解析結果が日本語の単語分割に直接用いられており, 解析器ごとの分割方針や手法の違いが出力にどのような影響を及ぼしているのか明らかではなかった.

本研究は, 単語分割が統計的機械翻訳に与える影響を調査するため, 既存の形態素解析器およびヒューリスティックな手法を用いて日本語の単語分割を行い, 日本語と英語での翻訳結果を比較検討する. その上で, 統計的機械翻訳に最適な結果をもたらす単語分割手法, 単語分割の評価結果への影響, および単語分割手法の違いに左右されにくい, より客観的な評価手法の実現などの問題について議論する.

2　関連研究

単語分割が統計的機械翻訳に与える影響を評価した研究には, 英語からアラビア語での研究 [1] と, 中日での研究 [2] がある. 前者はアラビア語の複数の単語分割手法での評価結果の違いを比較し, それら単語分割手法の違いに基づいて翻訳システムを組み合わせることで, さらに良い評価結果を得ている. 後者は中日翻訳での従来の中国語単語分割に対して, より頻繁に出現する粒度の細かい部分を単語とみなして, 優先的に分割する手法を提案・比較している.

しかし両者とも評価手法として BLEU[3] を用いており, BLEU による比較が出力とテストデータでの単語分割の違いからどのように影響を受けているか明らかではなかった.

3　実験

本研究では, 単語分割手法, 分野の違うコーパス, 評価手法の 3 点それぞれの場合について最終的な評価結果を比較する実験を行うことにより, 単語分割手法の違いが統計的機械翻訳の出力にどのように影響を与えているのかを調査する. またその結果から, 現在使われている様々な単語分割手法に対して, どのような評価を行なっていくことが統計的機械翻訳にとって適切であるのか考察する.

3.1　単語分割手法

単語分割には, 形態素解析器の MeCab 0.98[*1], KyTea 0.3.2[*2], JUMAN 6.0 および 7.0[*3]と, 教師なし形態素解析を行う latticelm 0.2[*4], またヒューリスティックな単語分割手法として, 文字種による分割と, 1gram 分割を用いた.

文字種による分割では, 文字を漢字, カタカナ, ひらがな, 半角アルファベット, 全角アルファベット, 半角数字, 全角数字, その他の 8 種類に分類し, 別種の文字間である場合に空白を挿入した. 1gram 分割では, 一文字ごとに空白を挿入することによって, 全ての文字列を一文字区切りにした.

例えば, 後述の新聞記事コーパスの日本語訓練データのうち最初の 1 文では次のような結果になる:

原文　この結果、国際市場は輸出国の生産動向に大きく左右されることになる。
文字種による分割　この 結果 、国際市場 は 輸出国 の 生産動向 に 大 きく 左右 されることになる 。
1gram 分割　こ の 結 果 、国 際 市 場 は 輸 出 国 の 生 産 動 向 に 大 きく 左 右 さ れ る こ と に な る 。

[*1] http://mecab.sourceforge.net/
[*2] http://www.phontron.com/kytea/index-ja.html
[*3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[*4] http://www.phontron.com/latticelm/index-ja.html

3.2 コーパス

実験には新聞記事と Wikipedia の 2 分野のコーパスとして, 日英新聞記事対応付けデータ (JENAAD), ロイター日英記事の対応付け (以下 REUTERS) [4], Wikipedia 日英京都関連文書対訳コーパス 2.01 (以下 WIKIPEDIA) [*5]を使用した.

実験は,JENAAD と REUTERS の両方,REUTERS, WIKIPEDIA の 3 種類の組み合わせで行った.

3.2.1 JENAAD+REUTERS

JENAAD と REUTERS の両方を使用する場合には, JENAAD から先頭の 147109 文対, REUTERS から先頭の 53891 文対を抽出し, 先頭の 500 文対ずつを合計した 1000 文対をテストデータ, 残りの合計 20 万文対を訓練データとした.

3.2.2 REUTERS

REUTERS のみの場合は, 先頭の 500 文を除いた 56282 文対のうち, 始めの 1000 文対をテストデータ, 残りの 5 万文対を訓練データとした.

3.2.3 WIKIPEDIA

WIKIPEDIA には京都関連の複数分野文書が XML 形式で含まれている. まずコーパス内の文書を LTT, EPR, FML, BDS, CLT, BLD, GNM, SCL, ROD, SNT, PNM, HST, RLW, SAT の順で各分野ごとに通し番号で並べて結合し, 単一文書とした. その上で Python 2.7.2 の xml.etree.ElementTree.parse で解析できた 477036 文対のうち, Moses で問題のある「|」の文字が日英のどちらかに含まれている文を取り除いた 477012 文対を原データとした. 原データから 477 行ごとに 1 文対をとりだした合計 1000 文対をテストデータ, 残りの 476012 文対を訓練データとした. なお, 英文には訳文修正済のものを使用した.

3.3 評価手法

英語と日本語での統計的機械翻訳では, 機械翻訳の出力とテストデータをそれぞれ単語分割した後, 主に単語の差異から文の違いを評価する評価手法を用いている. このような既存の評価手法には, 出力とテストデータが事前に単語分割されていなければ比較できず, またテストデータの単語区切りと一致する単語区切りを行った出力が高く評価されてしまうという問題があった.

そこで本研究では従来型の評価手法である BLEU, RIBES [5, 6] に加えて, 単語分割の影響をうけにくい評価手法とされている BLEU in Characters [7] を空白を全て取り除いた上で用いた. なお, BLEU と BLEU in Characters では共に 4-gram までの値を用いた.

3.4 実験手順

実験では英日・日英の両方について統計的機械翻訳を行った.

統計的機械翻訳のベースラインには, SRILM 1.5.12[*6],Giza++ 1.0.5 (2011 年 9 月 24 日版)[*7],Moses (2010-08-13)[*8] を用いた. また英日翻訳のベースラインでは質をより高めるため, 前処理に Head-Finalization [8] を加えた. その際, 英文の解析には Enju 2.4.2[*9]を用い, NTCIR9 で用いられた Head-Finalization ルール [9] のうち下記のものを採用した:

- Head が句の最後尾に無ければ, 句を逆順にする.
- 等位接続 (coordination) では逆順にしない.
- 名詞の複数形を単数に変換する.
- 冠詞 (a, an, the) は出力せず, 削除する.

実験はデータの種類に関係なく以下の手順で行った. 前処理には単語分割の後, 文字を正規化するための文字変換の処理を加えた. 日本語の文字変換には NTCIR-9 PATMT タスクで用いられたスクリプト[*10]を使用した. 英語では Moses に含まれるスクリプトの tokenizer と lowercase を用いた.

1. 英日翻訳の場合,英文データを Head-Finalization スクリプトで事前に変換する.
2. 日本語の前処理として, 各単語分割手法で単語分割し, 文字変換する.
3. 英語の前処理として, tokenizer で単語分割し, lowercase で文字変換する.
4. 前処理済みデータを訓練データとテストデータに分割する.
5. 訓練データから言語モデルを作成する.
6. 英日双方の訓練データを用いて, 英日・日英の 2 種類の翻訳モデルを作成する.
7. 言語モデルと翻訳モデルを用いてデコードを行い, テストデータを翻訳する.

[*5] http://alaginrc.nict.go.jp/WikiCorpus/
独立行政法人情報通信研究機構作成

[*6] http://www.speech.sri.com/projects/srilm/
[*7] http://code.google.com/p/giza-pp/
[*8] http://sourceforge.net/projects/mosesdecoder/
[*9] http://www-tsujii.is.s.u-tokyo.ac.jp/enju/index.ja.html
[*10] http://ntcir.nii.ac.jp/PatentMT/dataPreparationJE

8. 各評価手法でデコード後の出力とテストデータを比較する.

4 実験結果と考察

表 1, 2 は, それぞれ英日・日英翻訳の実験結果である. 各表には順に JENAAD+REUTERS, REUTERS, WIKIPEDIA のコーパスでの評価結果を記載している. また表中の表記のうち,「文字種」は文字種による単語分割,「1gram」は 1gram 分割のことである.

MeCab, KyTea, JUMAN の 3 つの形態素解析器による単語分割結果は, 全ての表で latticelm と「文字種」のスコアより高かった. 例えば表 1 の REUTERS の結果を見ると, 形態素解析器による BLEU スコアが 24.02 〜 23.04 なのに対し, 他の手法では 17.87 〜 11.06 といずれも低かった.

「文字種」は, 表 2 の JENAAD+REUTERS と REUTERS, および表 1, 2 の WIKIPEDIA での結果を見ると, BLEU in Characters の評価では最低の値であり, また他の評価手法でも低い値となった. これは文字の種別のみで単語区切り位置を判断しているため, 本来単語分割されるべき位置で分割され, また分割されるべきでない位置でされ, 適切でない単語分割となってしまったためだと考えられる.

「1gram」の結果は, 表 1, 2 の JENAAD+REUTERS の結果を見ると BLEU と RIBES で最も高いスコアであり,また表 1 の WIKIPEDIA では BLEU in Characters が最高だった. しかし表 1 の REUTERS では BLEU と BLEU in Characters で最低で, 表 2 の REUTERS と, 表 1, 2 の WIKIPEDIA でも概ね形態素解析器より低かった. これには次のような説明が考えられる:

1. 本研究の英日翻訳では Head-Finalization を行い, 英日の語順が近づいていた. そのため日英翻訳の場合に比べて句のアライメントが容易だった.
2. 20 万文以上の訓練データでは, 5 万文の訓練データより最適な句アライメントが学習しやすかった.
3. 「1gram」の単語分割では, 1 単語が 1 文字であるため, 最適な句アライメントが機械翻訳にとって最適な単語の組み合わせになりやすい.

実際の出力結果を見ると, 短い文で翻訳が上手くいっており, 全体的に質の高い結果だった. 一方で長い文や出現頻度の低い固有名詞がある場合には不自然な翻訳となっていた.

結果として, 統計的機械翻訳の単語分割では既存の形態素解析器が十分有効に働いた. また Head-Finalization, 20 万文以上のコーパス,「1gram」の 3 つの組み合わせで, 形態素解析器での評価結果を超えるような結果を達成することも可能であるとわかった.

一方で, 表 1 の JENAAD+REUTERS を見ると,「1gram」では BLEU, RIBES が最も高いスコアを示したにも関わらず, BLEU in Characters のスコアは形態素解析器のスコアより低い値であった. また同じく表 1 の REUTERS では,「文字種」の BLEU と BLEU in Characters の値が最も低かったにも関わらず, RIBES では latticelm が最も低い値だった.

つまり, いずれ評価手法でも, 単語分割手法とコーパスの組み合わせにより異なる結果となり, 単語分割が統計的機械翻訳に与える影響を評価する上で適切な, 一貫的な結果は得らないことがわかった.

今後はより統計的機械翻訳に適した単語分割として, 1gram の単語分割で固有名詞の分割を除外した場合や, 新聞記事や Wikipedia 以外の分野のコーパスについて調べることが課題である.

5 おわりに

本研究は, 日本語単語分割が統計的機械翻訳に与える影響を評価するため, 異なる単語分割と 2 分野のコーパスを使って, 最終的な評価結果を比較する実験を行った. これにより単語分割手法の違いが統計的機械翻訳の出力に大きく影響を与えており, また既存の形態素解析器は単語分割手法として十分有効であることがわかった.

しかしながら, いずれの評価手法も, 単語分割が統計的機械翻訳に与える影響を評価する上で適切な評価手法とは必ずしも言えないことも明らかになった. 今後の統計的機械翻訳の評価手法には, 単語分割手法からも影響されにくい, 一貫的な評価手法が必要である.

謝辞

本研究の設定と関連研究についてご助言くださいました, 京都大学黒橋・河原研究室の Ben Humphreys 氏, latticelm のより効率的な使用方法をご教示くださいました, 京都大学メディアアーカイブ研究室の Graham Neubig 氏の両氏に深く感謝いたします.

表1　英日翻訳

	KyTea	MeCab	JUMAN6	JUMAN7	文字種	latticelm	1gram
JENAAD+REUTERS (20 万文対)							
BLEU	25.42	25.61	24.86	24.53	19.33	11.82	33.78
RIBES	76.42	77.07	76.21	75.81	73.48	63.93	77.3
BLEU in Characters	35.86	36.07	35.94	35.54	30.53	28.67	34.04
REUTERS (5 万文対)							
BLEU	24.02	23.16	23.04		17.87	11.25	11.06
RIBES	66.4	67.37	66.48		67.02	55.3	63.2
BLEU in Characters	34.65	34.34	34.23		29.25	26.48	12.43
WIKIPEDIA (476012 文対)							
BLEU	21.07	19.82	22	20.72	12.8	10.46	27.8
RIBES	69.19	69.45	70.05	70.15	63.39	58.92	71.98
BLEU in Characters	29.48	29.02	28.89	29.13	24.8	25.62	28.07

表2　日英翻訳

	KyTea	MeCab	JUMAN6	JUMAN7	文字種	latticelm	1gram
JENAAD+REUTERS (20 万文対)							
BLEU	16.57	16.85	17.22	17.26	13.8	14.61	15.31
RIBES	64.99	65	64.94	64.78	64.87	64.5	63.28
BLEU in Characters	52.91	53.13	53.32	52.99	41.03	49.23	54.77
REUTERS (5 万文対)							
BLEU	16.91	17.42	17.52		15.01	14.01	15.66
RIBES	64.86	64.96	65.12		64.74	63.59	63.1
BLEU in Characters	51.19	50.96	51.15		39.8	45.18	53.09
WIKIPEDIA (476012 文対)							
BLEU	16.61	16.88	16.58	16.61	12.81	15.2	15.42
RIBES	65.16	65.77	65.4	65.42	62.54	64.89	64.74
BLEU in Characters	44.91	45.55	45.81	45.15	32.21	41.91	45.03

参考文献

[1] A. Lavie and H. Al-Haj. The Impact of Arabic Morphological Segmentation on Broad-Scale Phrase-based SMT. *Machine Translation and Morphologically-rich Languages: Research Workshop of the Israel Science Foundation*, University of Haifa, Israel, 26 January, 2011.

[2] Y. Wang, K. Uchimoto, J. Kazama, C. Kruengkrai, and K. Torisawa. Adapting Chinese Word Segmentation for Machine Translation Based on Short Units. *Proc. of LREC*, pp.1758-1764, 2010

[3] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: a method for automatic evaluation of machine translation. *Proc. of ACL*, pp.311-318, 2002.

[4] M. Utiyama and H. Isahara. Reliable Measures for Aligning Japanese-English News Articles and Sentences. *Proc. of ACL*, pp.72-79, 2003.

[5] 平尾努, 磯崎秀樹, Kevin Duh, 須藤克仁, 塚田元, 永田昌明. RIBES: 順位相関に基づく翻訳の自動評価法. 言語処理学会第 17 回年次大会, pp.1115-1118, 2011.

[6] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. Automatic Evaluation of Translation Quality for Distant Language Pairs. *Proc. of EMNLP*, pp.944-952, 2010.

[7] Y. Lepage and E. Denoual. BLEU in Characters: Towards Automatic MT Evaluation in Languages without Word Delimiters. *Companion Volume to the Proc. of IJCNLP*, pp.81-86, 2005.

[8] H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh. Head Finalization: A Simple Reordering Rule for SOV Languages. *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pp.244-251, 2010.

[9] K. Sudoh, K. Duh, H. Tsukada, M. Nagata, X. Wu, T. Matsuzaki and J. Tsujii. NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT. *Proc. of the 9th NTCIR Workshop Meeting*, pp.585-592, 2011.