日英特許翻訳における日本語単語分割の分野適応の検討

須藤 克仁 *† 永田 昌明 * 森 信介 ‡
*NTT コミュニケーション科学基礎研究所 †京都大学情報学研究科 ‡京都大学学術情報メディアセンター
sudoh.katsuhito@lab.ntt.co.jp

概 要

機械翻訳において単語分割は翻訳精度に大きく影響する要因である．日本語における単語分割は新聞等のコーパスで学習された形態素解析器によるものが一般的であるが，特許等別分野の文書に対しては文体や用語のミスマッチによる単語分割誤りが起きやすい．本報告では，まず Kytea による日本語単語分割の部分的アノテーションによる分野適応について述べ，特許データを用いた実験で単語分割精度を改善した結果を示す．そして，事後並べ替え型翻訳による日英特許翻訳実験の結果を分析し，単語分割の精度や安定性の改善と翻訳精度の改善の関係について議論する．

1 はじめに

昨今の統計翻訳技術の発展により，その対象言語対は大きく拡大しつつある．統計翻訳のアルゴリズムは基本的に言語に依存しないが，日本語や中国語等明示的に分かち書きされない言語においては，言語依存の単語分割処理を行うのが一般的である．本研究では特に日本語の翻訳とそのための単語分割について議論する．

日本語の統計翻訳における単語分割には JUMAN, 茶筌，Mecab などの形態素解析器が広く用いられている．それらは新聞記事の解析に適合するように作成・学習されていることが多く，語彙や文体が十分合致しない別の分野の文書に対しては精度が低下する傾向にある．「現代日本語書き言葉均衡コーパス (BCCWJ)」のように多数の分野の文書を利用すれば対象を広げることが可能だが，対象分野において十分な量のコーパスを構築することは容易ではない．

こうした問題に対し，ある分野に適合するように作成された既存の解析器を，新たに解析したい分野の学習データを加えていくことでその分野に適合するように変化させる「分野適応」と呼ばれる方法がある．本報告では，テキスト解析器 Kytea[1] の部分的アノテーションによる学習の枠組みを利用した日本語単語分割の特許分野への適応による単語分割精度の改善と，日本語から英語への[1]特許翻訳精度との関係について述べる．また，その結果を通じて単語分割の統計翻訳への適応についての展望について述べる．

[1]日本語単語分割と翻訳精度の関係を議論するという観点では英語から日本語への翻訳も考えられるが，目的言語側の単語分割の異なりは BLEU 等の自動評価結果に与える影響が大きいため，本報告では日本語から英語への翻訳を利用する．

2 関連研究

本報告に関連する既存研究として，中国語から英語への統計翻訳における中国語単語分割と翻訳精度の関係について調べたものがある [2]．素性の異なる 2 種類の統計的単語分割と，単純な辞書内語彙最長マッチによる単語分割との比較を通じ，下記のような分析結果が報告されている．

> 1 文字を 1 単語と考えるよりは単語分割を行う方が翻訳精度 (BLEU) が高い．
>
> 単語分割精度 (F 値) が高くても翻訳精度 (BLEU) が高くなるとは限らない．
>
> 単語分割の揺らぎ (同一文字系列に対する単語分割バリエーションのエントロピー) が小さいほど翻訳精度 (BLEU) が高くなる傾向にある．
>
> 単語分割数を若干多くする (単語長を短くする) ように調整する方が，単語分割自体の精度 (F 値) は低くなるものの，翻訳精度 (BLEU) は高くなる．

この結果は，統計翻訳での利用における単語分割器の作成・学習基準として，従来の単語分割性能 (例えば F 値) だけでは十分でないことを示唆している．特に，複合語や未知語の単語分割の揺らぎを小さくし，翻訳可能な語彙内の単語を安定して得られることが重要であると考えられる．

単語分割バリエーションのエントロピーはある文字列 w_i に対する単語分割 v_{ij} の条件付きエントロピーとして以下の式で定義される．

$$H(V|W) = -\sum_{w_i} P(w_i) \sum_{v_{ij}} P(v_{ij}|w_i) \log P(v_{ij}|w_i)$$

ある参照文字列 w_i に対する単語分割 v_{ij} のエントロピーが高い，ということは，単語分割の揺らぎが大きく，同じ文字列が様々に異なる単語列に分割される可能性があるために安定して同じ単語が得られにくくなることを意味する．

本報告では上記のように統計的単語分割の方式を変えるのではなく，分野特有の語彙や文体との不整合に起因する単語分割の誤りや不安定さを分野適応によって解決する．また，単語分割の改善と統計翻訳の精度向上との関係を，日本語の表記文字種ごとに大きく異なる傾向の分析を通じて議論する．

3 単語分割器 Kytea の分野適応

Kytea の分野適応に利用するのは文中で単語分割位置を部分的にアノテーションした部分的単語分割コーパスである．部分的単語分割コーパスでは，以下の例のように文中の文字が以下の分割記号によって一文字ずつ分割されている．文字間の “|” は単語分割される位置として，“-” は単語分割されない位置としてアノテーションされていることを示し，空白となっている箇所はアノテーションがなく単語分割されるか否かが不明であることを示す．

つ ぎ に 添 付 図 面 に 従 い 、 | 本 | 発-明 | の 装 置 に つ い て 説 明 す る 。

従来の Kytea の分野適応 [1] は点予測の確信度に基づくものであったが，本報告では前節の議論を踏まえ，適応先分野において特徴的に起こる単語分割の揺らぎを小さくすることを目指す．単語分割の揺らぎは未知語の近傍で起こりやすいと予想されるため，頻出する未知語の出現箇所を選択的に追加学習するため，以下の手続きで分野適応を行った．

1. 分野適応用文書から未知語候補を抽出 [3][2]
2. 未知語候補を含む文を 3 文ずつ列挙 (未知語候補の期待頻度の降順に整列)
3. 未知語候補の出現箇所について，未知語候補を単語と認めた場合の単語分割位置を作業者が推定
4. 作業終了後の部分的単語分割コーパスを用いて Kytea を追加学習

4 事後並べ替え型日英翻訳

日英間の機械翻訳は語順の違いが大きいことから欧米言語間と比較して難しいと考えられてきた．英語から日本語への翻訳においては，主辞後置 (Head Finalization)[4] を代表とする事前並べ替え (pre-ordering) と呼ばれる技術により近年大きく統計翻訳による精度が改善している．一方で日本語から英語への翻訳では，統計翻訳の精度はルールベース翻訳と比較して大きく劣っているのが現状である [5].

本報告では日英翻訳の手法として，単語や句の翻訳と語順の入れ替えを 2 段階の統計翻訳で実現する，事後並べ替え (post-ordering)[6] 型の翻訳手法を利用する．日本語は一旦中間言語である主辞後置英語 (Head-Final English, 以下 HFE)[4] へ語順を入れ替えずに翻訳され，その後の HFE から英語への翻訳において語順の入れ替えを行う．そのための翻訳モデルは，日英の並行コーパスと，並行コーパスの英文の構文解析結果に主辞後置ルールを適用した HFE コーパスを利用して，日本語から HFE へのモデルと，HFE から英語へのモデルとして作成される．

[2]各品詞への帰属確率の推定には最大エントロピー法を用いた．

5 実験と分析

Kytea の分野適応による単語分割精度の変化と，それに伴う日英翻訳精度の変化の関係を調べるために以下の実験を行い，結果を分析した．

5.1 単語分割の分野適応実験

まず，Kytea の特許分野への適応実験を行った．分野適応のための部分アノテーションを与える学習データとして，NTCIR-7 PATMT および NTCIR-8 PATMT で学習データ・開発データとして提供された日本語特許文書を利用した．

アノテーション作業 (3 節の 3. に相当) は作業者 1 名が累計 12 時間行い，1 時間ごとの累積作業データを利用して作業時間 0〜12 時間の 13 段階の分野適応を行った Kytea の単語分割性能の比較を行った．評価データは NTCIR-9 PatentMT の英日翻訳タスクの正解訳 2,000 文のうち先頭の 500 文 (3,629 単語．以下，**単語分割評価セット**) に同じ作業者が人手で単語分割正解を付与したものである．

分野適応のアノテーション時間と単語分割精度 (F 値) の関係を図 1 に示す．部分的単語分割コーパスの追加により単語分割精度がほぼ単調に増加していることが分かり，最終的に F 値が 0.5%向上し (97.57%→98.07%)，およそ 2 割の単語分割誤りが削減できた．この結果から，KyTea が部分アノテーションを通じて特許分野に適応可能であることが確認できた．

5.2 翻訳実験

続いて，単語分割の分野適応の各段階において，日英翻訳の精度がどう変化するかを調べるための実験を行った．学習およびテスト用のデータは NTCIR-9 PatentMT[5] の日英翻訳タスクのものを利用した．テストセット (以下 **NTCIR 日英テストセット**) は 2,000 文，67,210 単語 (Enju による解析の結果) であった．

事後並べ替え型翻訳 (以下 PoMT) は，日本語から HFE への phrase-based 翻訳と，HFE から英語への string-to-tree 翻訳によって実現した．PoMT で利用する英語の構文木は Enju (ver. 2.4.1) の構文解析結果を利用して作成した．翻訳デコーダは moses および moses_chart (revision 3717) を利用し，GIZA++, MERT などを用いた標準的なモデル学習を行った．評価には大文字・小文字は区別しない BLEU[3] と RIBES(ver. 1.01) を利用した．

なお，PoMT は NTCIR 日英テストセットにおいて BLEU 31.66%の性能であり，moses の通常の phrase-based 翻訳 (29.53%)，moses_chart による日英 string-to-tree 翻訳 (30.97%) の性能を上回った．

[3]NTCIR-9 における BLEU の公式評価は大文字・小文字を区別して `mteval-v13a.pl` で行われたが，本報告では `mteval` の単語再分割の影響を排除した単純な比較のために，`multi-bleu.perl` による結果を用いた．

5.2.1　単語分割評価セット

5.1 の実験で利用した単語分割評価セットを PoMT で翻訳した実験の結果を図 2 に示す．図 2(a) には英語の BLEU に加え，PoMT の途中段階である HFE での BLEU (HFE の参照訳は英語の参照訳から自動生成) を併記している．図 2(b) は英語の PER (Position-independent Word Error Rate) の結果である．

図 2 に示したいずれの評価においても，単語分割の適応後は適応前よりも安定して高い翻訳性能が得られている．特に，最初の 1 時間作業後の向上の度合いが大きかった．しかしながら，その後は図 1 の単語分割精度のような単調な性能向上は得られなかった．

5.2.2　NTCIR 日英テストセット

比較のため，NTCIR 日英テストセットでの実験を行った．翻訳モデルは前記実験と同一である．結果を図 3 に示す．BLEU や PER の大まかな変動傾向は図 2 と似ているものの，数値の変動幅は小さく，適応前と 12 時間作業後の間で BLEU 0.4 ポイント程度の改善が見られるものの，途中では適応前を下回る結果が出ているところもある．本実験からは，単語分割の適応による安定した翻訳性能改善は確認できなかった．

5.3　分析: 単語分割と翻訳精度の関係

5.3.1　単語分割精度

図 4 に単語を構成する文字種ごとの単語分割精度を，また，図 5 にアノテーションされた未知語候補数の推移を示す．NTCIR 日英テストセットについては先頭の 1,000 文について単語分割の正解を別途付与して評価した．英数字を除いて概ねアノテーション数の増加に伴って単語分割精度も向上した．漢字では “本 | 発明” のような特許文書特有の熟語，“ポリゴン | ミラー” のようなカタカナ複合語がアノテーションされ，正しく分割できるようになった．

分野適応のための未知語候補の選択は期待頻度に基づくため，文字種による出現頻度の違いや文字組み合わせの分散の大きさの影響が反映されている．特に数字については文字組み合わせの分散が非常に大きく期待頻度が小さくなるため未知語候補として選択されず，英字も非常に少ないアノテーションしか得られなかったことから英数字で構成される単語の分割精度を安定して向上させるには至らなかったと考えられる．

5.3.2　単語分割エントロピー

続いて，単語分割バリエーションのエントロピーを参照文字列の文字種ごとに計算し，文字種による単語分割の揺らぎをそれぞれ評価した．参照文字列 (w_i) には Mecab の単語分割結果を利用した[4]．その結果特に変化が認められた，数字，英字，記号，漢字，カタカナと，全単語についてのエントロピーを図 6 に示す．

[4]Mecab 自体の揺らぎの影響は類似するシステム間の比較では無視できる程度と考えられる．

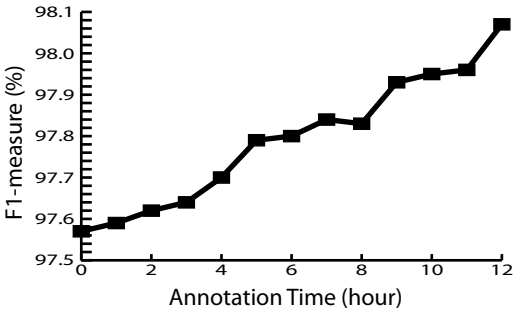


図 1: 単語分割精度 (F 値) の変化

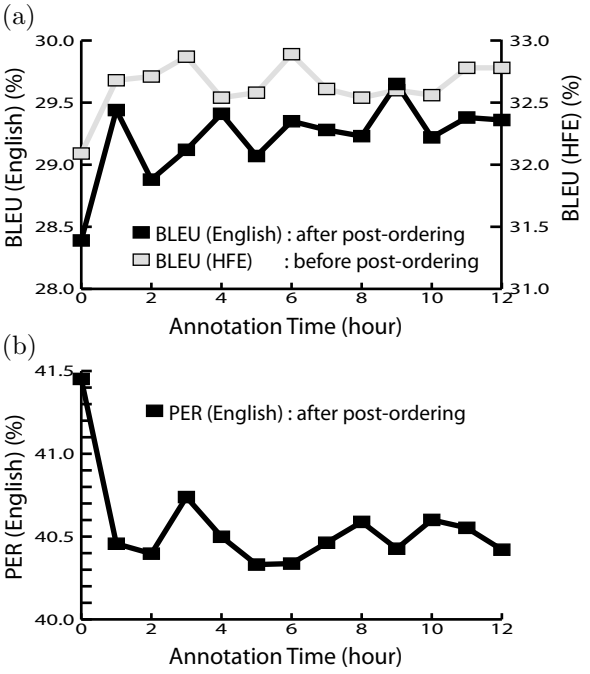


図 2: 単語分割評価セットにおける翻訳評価尺度の変化 (英語/HFE の BLEU および英語の PER)

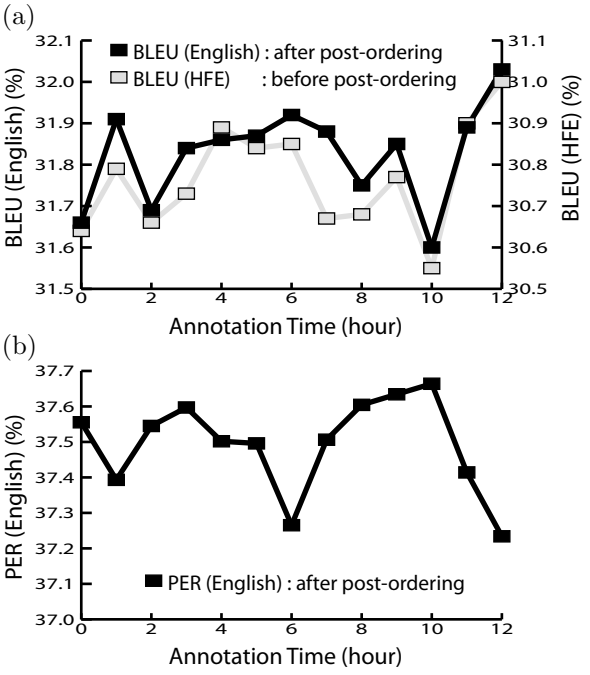
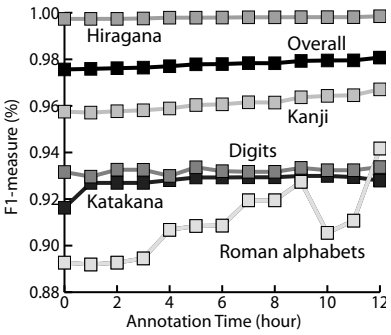
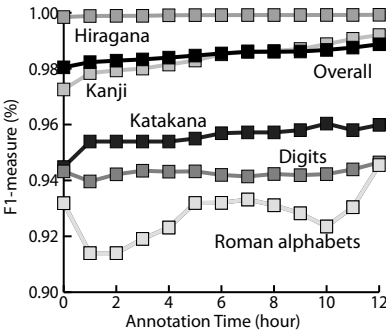


図 3: NTCIR 日英テストセットにおける翻訳評価尺度の変化 (英語/HFE の BLEU および英語の PER)



(a) 単語分割評価セット



(b) NTCIR 日英テストセット (1,000 文)

図 4: 文字種ごとの単語分割精度 (F 値)

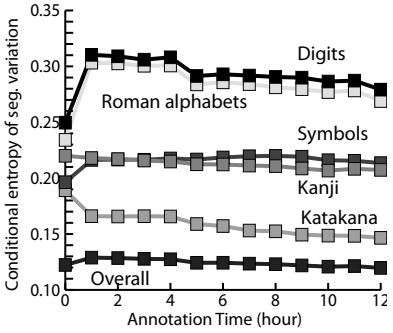


図 6: 単語分割エントロピーの変化

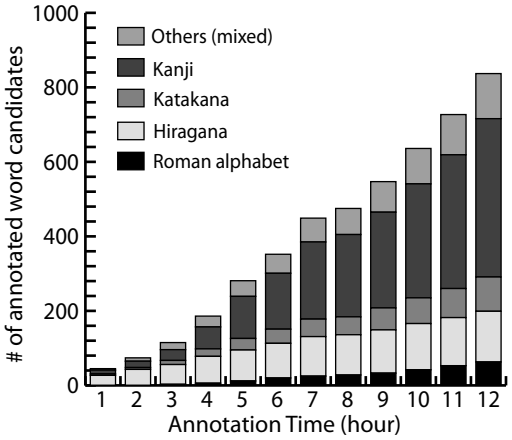


図 5: アノテーションした未知語候補数

分野適応によって顕著に単語分割が安定したのはカタカナ (全単語中の割合:7.8%) であり，適応の進捗に合わせて単調にエントロピーが減少している．漢字 (28.3%) も変化量は多くないが単調減少である．一方で，数字 (7.3%)，英字 (3.2%)，記号 (2.1%) は適応開始時に大きく不安定さが増し，適応が進むと若干改善する傾向を示している．前述の通り数字はアノテーションが行われていないため，この変化は英数字が混在する単語における改善によるものと考えられる．なお，適応開始時には英数字・カタカナで平均単語長が減少しており，このときに単語分割数が増加し安定性が低下していると考えられる．ここから，分野適応によってまず大きく分割数の傾向が変化し，徐々に分割バリエーションが収斂し安定化していることが分かる．

5.3.3 既存研究との比較

既存研究 [2] では単語分割の方法を大きく変化させており，その結果単語分割の比較における精度 (F 値) やエントロピーの差が大きく，その結果翻訳精度の差も大きくないながらも概ね安定している．一方本報告における実験では，単語分割の精度は分野適応により安定して向上しているものの，その絶対的な精度の差は限定的であったために，翻訳に有意に差を与えるまで至らなかったと考えられる．

5.4 議論

本報告の単語分割の分野適応によって，特許文書に特有の熟語やカタカナ語の分割を改善できることを確認した．一方で英数字周辺はアノテーションが不足し小幅の改善に留まった．また，全体的な単語分割精度や安定性の改善に比して，翻訳への効果は限定的であった．

これは，統計翻訳という応用においては “全体的な” 単語分割改善とは異なる指針を必要とする可能性を示唆している．統計翻訳においては単語分割誤りが句対応を通じて隠蔽される可能性があるため，分割精度が高いことに加えて，分割を誤ってしまう箇所についても翻訳への影響の大きい内容語，低頻度語などを “安定して” 分割できることが重要であると考えられる．

6 おわりに

本報告では特許翻訳のための単語分割の分野適応について述べた．未知語候補への単語分割アノテーションが単語分割精度の向上に貢献することを実験によって確認したが，翻訳精度については精度改善が不安定で有意な差を得るには至らなかった．実験結果の分析を通じ，全体的な単語分割誤りを減らすだけでなく，翻訳に影響を与えるような種類の単語についての分割の精度や安定性を高めるための分野・タスク適応が必要であるという知見を得た．今後はそれを踏まえ，統計翻訳に適した単語分割適応方法について検討を進める．

参考文献

[1] G. Neubig *et al.* 点推定と能動学習を用いた自動単語分割器の分野適応. 第 16 回言語処理学会年次大会, 2010.

[2] P.-C. Chang *et al.* Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proc. of WMT*, pp. 224-232. 2008.

[3] 森, 長尾. n グラム統計によるコーパスからの未知語抽出. 情報処理学会研究報告 1995-NL-108, 1995.

[4] H. Isozaki *et al.* Head Finalization: A Simple Reordering Rule for SOV Languages. In *Proc. of WMT-MetricsMATR*, pp. 244-251. 2010.

[5] I. Goto *et al.* Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proc. of NTCIR-9*, pp. 559–578. 2011.

[6] K. Sudoh *et al.* Post-ordering in Statistical Machine Translation. In *Proc. of MT Summit XIII*, pp. 316-323. 2011.