機械翻訳に適した短単位に基づく中国語単語分割について

王軼謳, 内元清貴, 風間淳一, Kruengkrai Canasai , 鳥澤健太郎

独立行政法人情報通信研究機構　MASTAR プロジェクト 言語基盤グループ

1. はじめに

中国語には単語と単語の間に空白を入れる「分かち書き」という習慣がないため、中国語の処理を行う際には、語境界を認定すること，すなわち単語分割は、最も基本的かつ重要な課題となる。特に、機械翻訳において性能低下を招く要因の1つとして、中国語の単語分割が困難で、十分な分割精度を得ることができないことがある。この解決を目指して多くの研究が行われてきた。しかしながら、これら既存研究の多くは、語境界が明示されている言語（例えば、英語）と中国語との対応を利用したり[6]、複数の単語分割プログラムを利用したりするが[2][5]、コーパスベースの単語分割に対して重要な要素である学習コーパスの適切な分割粒度を対象とする研究は、殆ど見あたらない。

本研究では、中国語単語分割を機械翻訳に適応させるため、簡単かつ有効な「短単位」に基づく手法を提案する。日本語の電子化辞書Unidic[8]の「短単位」を参考に、中国語に対する「短単位」を定義し、既存のコーパス（単語分割モデルの学習コーパス）を短単位で揃えるための半自動の変換手法を提案する。提案手法の有効性を評価するために、新聞記事と科学技術論文の二つのコーパスにおける中日機械翻訳実験を行った。既存のコーパスを短単位に変換したものをトレーニングに使って単語分割を行なった場合は、既存コーパスにおける単語分割基準をそのまま利用した単語分割器を用いてトレーニングした場合に比べて、最終的な翻訳結果が1.1ポイント高いBLEU値を示し、本手法が有効であることが示された。

2　中国語単語分割の短単位について

2.1　分割単位

単語分割とは、「与えられた文字列に対し、意味のある言語単位に分割する」ものである。言語単位は通常、各々のコーパスの分割基準によって定義される。

しかし、既存の注釈付き中国語コーパス（例えば、CTB コーパス(Penn Chinese Treebank)と PKU コーパス(3.1 節)）には二つの問題がある。

（1）語の単位の基準が明確でなく、統一も取れていない。トークン（コーパス中で語としてアノテートされているもの）には、形態素、語、複合語などが存在し、単位がばらばらである。

（2）トークンの形態、意味、音韻のよりも、構文的な機能を重視して分割基準を定義しているために、分割粒度が粗い傾向があり、一部の有用な形態素情報が失われている可能性がある。

例えば、ある中国語の多音節語は、多数の意味のある形態素から成り立ち、多数の英語の単語に翻訳される。中国語「艺术节」は英語に翻訳すると、「*art festival*」となる。形態素「艺术」と「节」はおのおの意味があって、それぞれ「*art*」と「 *festival*」と対応している。しかし、CTB コーパスと PKU コーパスでは、「艺术节」が 1 つのトークンとしてアノテートされている。従って、「艺术节」は既知語となるが、一方、「艺术」を共有している、「艺术团 」（*Troupe of Arts*）は学習コーパスには出現せず、未知語となって、正しく解析できないことが分かった。もし、「艺术/节」と分割するようなコーパスであったら、少なくとも「艺术」に関しては何らかの学習が行われ、「艺术团」は正しく解析できていたかもしれない。つまり、このような不適当な言語単位の

トークン化現象は、データスパース性の問題を生じさせ、単語分割の過程において未知語を増やし、単語アライメントの性能を悪化させる傾向がある。未知語の存在は単語分割を難しくする要因であり、単語アライメントの性能は統計翻訳の性能と密接に結びついているのである。

上述の問題に対処するため、主として言語の形態的な側面に着目して、中国語の「短単位」という概念を提案する。短単位（既存の分割基準より短い単位）とは、独立の意味を持つ最小の文字列のことである。例えば、上記の「艺术」、「团」、「节」である。一般に単位を短くすればするほど、取り出した単位は基本的な語となって、データスパースネスを軽減することができる。短単位は既存の分割単位よりゆれが少ない、つまりコーパスの精度が高いため、用例を収集して分析を行う上では便利で有効な単位である。

2.2 短単位の規準

既存のコーパスから短単位を検出し、処理するため、操作主義的な立場から、以下の規準を設けた。

（1）1文字と2文字のトークンについて、既存のコーパスにある1文字或いは2文字のトークンは短単位だと考える。文字レベルまで分解しないように、2文字のトークンはそのまま残す。

（2）3文字以上のトークンについては、トークンの意味が構成要素の集合とすると、このトークンは短単位の集合と考えられるので、短単位に分割する。この場合、短単位の間にオーバーラップがなく、少なくとも１つの短単位が多文字列であることを前提とする。例えば、3文字のトークンABCがあるとする、もし、これが意味的にAB+CあるいはA+BCであれば、AB/C、A/BCのように分割する。また、A+B+C、AC+BC、ABCであれば、そのままとする。（例えば、上海市=（上海+市）＝＞上海/市;「大中小」=（大+中+小）と「中小学」=（中学+小学）はそのままである。）

（3）必要に応じて、出現率の高い短い文字列(3文字以下の文字列）は短単位とする(2.3節)。

（4）明らかに分割符号（句読点、記号など）で分けている文字列は分割する。

2.3 短単位への変換アプローチ

以上を念頭におき、次に述べるような方法で、既存のコーパスから短単位の基準へ変換を試みた。

（1）日本語の短単位を参考に、アライメント情報を利用して、変換ルールを整理する。

基本的な考え方としては、中日対訳コーパスに対する単語アライメント情報を利用して、中国語側の長単位(より長いトークン)を分解する。PKUコーパスの中国語文を日本語に訳した対訳コーパスが構築されており、本研究ではこれを利用する。日本語においては、短単位に従った電子化辞書であるUnidicが開発されており、その辞書を用いた形態素解析器 MeCab[7]を利用することができる。まず、Unidic を用いた Mecab を用いて日本語を分割し、GIZA++を用い、中日の単語アライメントを得る。そして、長単位情報を獲得するために（1-to-n） アライメント結果（中国語側の１語が複数の日本語の語と対応）を抽出する。この（1-to-n） アライメント情報を利用して自動的に長単位を分解することを試みた。しかし、この方法ではごく一部の長単位を短単位に正しく分割することしかできなかった。そこで、これらの情報を利用し、品詞によって長単位を分類し、カテゴリーごとに変換ルールを作る。その後、プログラムに変換ルールを実装し、長単位を短単位へ変換する。現時点では、28 個の変換ルールがある。例えば、数詞について、一、十、百、千の桁ごとに１短単位とするというルールがあり、長単位 「八百二十四万二千一百二十一」を「八百/二十/四/万/二千/一百/二十/一」に分解する。

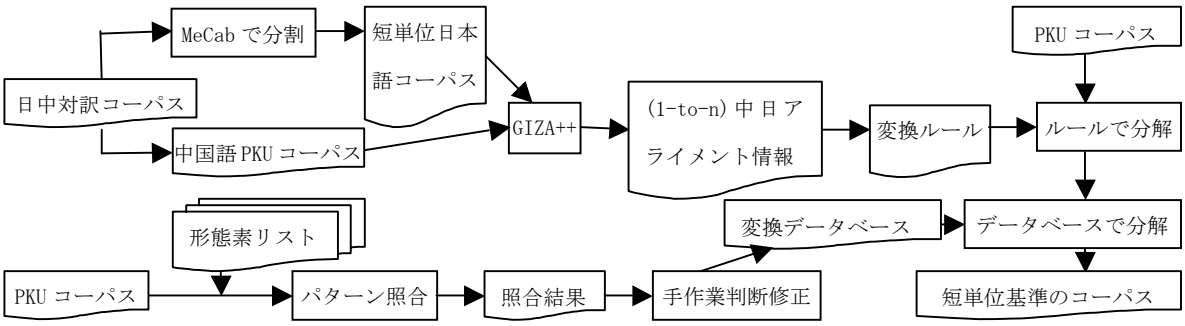

図 1　短単位変換への流れ

(2) 外部辞書を用いて変換データベースを構築する。

基本的な言語知識は、既存の辞書に含まれていることが多い。それゆえに外部辞書を利用することにより、短単位へ変換する情報を得、データベースを構築することができる。具体的には、下記の処理を行い、データベースを構築した。

ステップ1:次のような形態素リストを作成する。

（i）lemmaリスト（157, 293個）には、次のような情報が含まれている。

①固有名詞：文字長に制限がない。

②2～4文字からなる辞書の見出し語

辞書としては、いくつかの電子化辞書と既存のコーパスのトークンをマージして「単語 品詞」という形で作成したものを用いる。

③2～3 文字の短単位語

2.2 節の短単位規準(3) に基づいて、既知語(辞書の見出し語) と組み合わせて既知語となる未知語(辞書にない語)は、短単位であると仮定する。この仮定に立脚して、辞書の見出し語を「既知語＋未知語」のように分割する。そして、3 語以上の辞書の見出し語についてそのような分割を行なうことができた未知語について、それが 1 文字である場合は、morpheme として抽出し、2-3 文字である場合は、lemma として抽出する。例えば、「一団和气」、「一团乱麻」、「一团漆黑」、「漆黑」、「乱麻」、「和气」が辞書の見出し語である場合、「一团」 は短単位として抽出する。

（ii）morphemeリスト（549個）には、上記の③で生成された、接頭語、接尾辞、動詞・形容詞・副詞に対する形態素（例えば、前置詞：于；否定形態素：未)、普通名詞の形態素などがある。

ステップ2：形態素シーケンスを定義する。

パターン：(morpheme)*(lemma)＋(morpheme)*

ステップ3：既存のコーパスから、上記のパターンと照合できるトークンを抽出して、変換データベースを作成する。

ステップ4：コーパスの精度を保証し、分割の曖昧さを解消するために、手作業で変換データベースを検証し、修正する。

ステップ5：手作業により修正された変換データベースを用いて、長単位を短単位に分解する。現時点では、変換データベースには3617個のルールがある。いくつかの例を表1に示す。表 1 の長単位の品詞はコーパスの品詞であり、短単位の品詞は形態素リストの品詞である (手作業修正後)。

表 1.　変換データベースの例

長単位	品詞	短単位 1	品詞 1	短単位 2	品詞 2
意大利语	n	意大利	ns	语	n
全球化	v	全球	n	化	sv
决定于	v	决定	v	于	p

短単位へ変換する流れは図 1 に示す。この処理は既に情報付けされたコーパスに基づいて行うため、手作業による修正にもあまり時間がかからない。3 万文のコーパスに対して、変換データベースの修正は一人で５～7 日で済む。

3　実験

3.1 実験セット

(1)　中国語単語分割モデルとして、単語・文字ハイブリッドによる識別型の形態素解析と品詞タグ付けモデルMMA[1]を使う。

(2)　SMTモデルとして、Mosesツールキット[4]を利用する。実験では，devset（表2）で、MERT[3]を用いてパラメータチューニングを行う．

(3)　中国語のコーパス：上記(1)のモデルは北京大学が作成したPKUコーパス（PKU Treebank、30,686 の文）を用いて学習する。

(4)　対訳コーパス:新聞記事（NIKKEI_BP: 日経BPの科学新聞記事における中日対訳コーパス）と科学技術論文（NICT_JC_SP: NICTで開発された科学技術論文における中日対訳コーパス）の二つの対訳コーパスを用い、中日機械翻訳実験を行う。コーパスは、表2に示すように、trainset、devset、testset1 とtestset2に分割して用いる。

表2.　二つの対訳コーパスの情報

		Trainset		Devset		Ttestset1		Ttestset2	
		中	日	中	日	中	日	中	日
NIKKEI_BP	文数	245,554		1,000		500		500	
	単語	7,019,359	8,238,960	28019	33025	14477	16893	13886	16042
NICT_JC_SP	文数	371,868		1,603		500		500	
	単語	11,054,040	13,617,437	49855	60440	15230	18298	15464	18534

3.2 実験結果と評価

提案手法を評価するために、元のPKUコーパスで学習された中国語単語分割モデルMMAによる分割を用いた翻訳モデルをベースラインモデルとする。次に、短単位に分解されたPKUコーパスで学習された中国語単語分割モデルMMAによる分割を用いた翻訳モデルを短単位モデルとする。翻訳実験の結果は表３のとおりである。両方の分野における実験に対して、短単位モデルはベースラインモデルよりBLEU値において、1.1～1.8ポイント高く、中日統計翻訳における短単位基準の従来の分割基準に対する優位性が示されている。

表3.　中日SMTの実験結果

BLEU	科学技術論文		科学新聞記事	
	Testset1	Testset2	Testset1	Testset2
ベースライン	0.2701	0.2755	0.3155	0.3297
短単位	0.2829	0.2870	0.3335	0.3410

4　おわりに

本稿では、中国語単語分割において短単位という概念を導入して、既存のコーパスを自動で短単位へと変換する手法を提案した。提案手法は学習コーパスから形態素情報を効果的・的確に捉え、統計翻訳の性能を向上させることができる。翻訳実験より、本手法が日中機械翻訳の精度の向上を実現できることが示され、SMTに対して短単位がより適切な分割粒度であることが確認された。さらに、本手法は処理が簡単なことから、低コストで大規模なコーパスを処理するのにも適当である。

今後は短単位の基準をより厳密に設計すること、本手法をさらに多様なコーパスに適用すること、より多くの分野と他の言語対でこのアプローチを検証することが課題である。

参考文献

[1]Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa and Hitoshi Isahara.2009. An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging. ACL-IJCNLP.

[2] Christopher Dyer, Smaranda Muresan, and PhilipResnik. 2008. Generalizing word lattice translation. ACL-HLT,

[3] Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. ACL.

[4]Philipp Koehn, Franz Josef Och, and Daniel Marcu.2003. Statistical phrase-based translation. In Proc.of NAACL-HLT.

[5]Ruiqiang Zhang, Keiji Yasuda and Eiichiro Sumita.2008. Improved statistical machine translation by multiple Chinese word segmentation. In Proceedings of the Third Workshop on Statistical Machine Translation,

[6] Yanjun Ma and Andy Way. 2009. Bilingually Motivated Domain-Adapted Word Segmentation for Statistical Machine Translation. In Proceedings of the EACL,

[7] http://mecab.sourceforge.net/

[8] http://www.tokuteicorpus.jp/dist/