機能語相当の日本語MWEの段階的まとめあげによる統計的機械翻訳の精度向上

坂本 明子　園尾 聡　田中 浩之　釜谷 聡史
株式会社東芝 研究開発センター
akiko7.sakamoto@toshiba.co.jp

1　はじめに

近年，統計ベースの機械翻訳手法 (Statistical Machine Translation = SMT) が盛んに研究されている．SMT では，対訳コーパスにおける訳文の対応関係を自動的に付与（対訳アライメント）し，これをモデル化して翻訳している．現在，対訳アライメントの多くは，単語間の対応関係を推定する手法となっている．

対応関係を付与する対象である単語は，形態素解析によって与えるが，形態素解析に適した単位が，対訳関係を付与するアライメントに適した単位であるとは限らない．多くのアライメント手法では，1対多，多対多の対応関係を付与するが，これらも形態素の対応関係を元に推定する．それゆえ，複数の形態素がまとまり，一つの非構成的な意味を持つ Multiword Expression (MWE) は，従来のアライメント手法では適切な対応関係を付与することが難しい．

Ma ら [6] は，中国語の MWE に注目し，中国語ツリーバンクから，その構文構造に基づいて，特に内容語相当の MWE を段階的に獲得し，アライメント時に段階的にまとめ上げる手法を提案した．これにより，1対多，多対多のアライメント関係を，1対1のアライメント関係に集約することが可能となり，結果，翻訳精度も向上すると報告している．

日本語からの翻訳では，機能語相当の MWE を適切に扱うことが翻訳精度の向上に寄与する [11]．しかし，日本語の機能語相当の MWE は，検出や意味判定が難しく，辞書を用いた一括登録が難しいという問題がある．そこで我々は，対訳コーパスの単語アライメント結果から得られるフレーズ対から，日本語の機能語相当の MWE を目視により抽出し，順次単語としてまとめ上げを行うことで，統計翻訳精度の改善を試みる．

以降，本稿では，第2章に日本語機能表現について説明し，第3章に提案手法を示す．第4章にその実験と結果にについて報告して，第5章にまとめを行う．

2　機能語相当の日本語MWE

本稿で扱う，機能語相当の MWE とは，2形態素以上から構成され，全体で1つの機能語 (助動詞，助詞，接続詞) として振る舞う形態素列を指し，坂本 (2009)[11] で指摘した通り，2種類の曖昧性を持つ．

1つは，内容的用法と機能的用法の間の曖昧性である．以下に文例を示す．

機能的用法：(1) 夢 について 話す.
内容的用法：(2) 絵の具が壁に ついて いる.

「について」という表記は,「について」という≪対象≫の意味を表す [9] 助詞としても使うことができ，この用法を「機能的用法」と呼ぶ．一方,「つく（「付く」,「着く」など)」といった動詞 (内容語) を表すこともでき，この用法を「内容的用法」と呼ぶ．

もう1つの曖昧性は，複数の機能的用法間の曖昧性である．例えば,「によって」という表記は，例文3，4において機能的用法を持つ．しかし，例文3では≪手段≫という意味を持つ一方で，例文4では≪場合≫という意味を持つ [8]．従って,「によって」という表記は複数の機能的用法間の曖昧性を持つと言える．

≪手段≫：(3) 調査 によって 解明した.
≪場合≫：(4) 場合 によって 対策が違う.

更に，同じ「機能的用法」であっても，格助詞相当の働きを持っていたり，接続詞相当の働きを持っていたりする場合があり，訳語も異なることがある．例えば，例文5と7において「では」は例文6の”at”と例文8の”then”のように訳し分けが必要である．

- **助詞相当の「では」**

日本語: (5) ホテル で は 両替 は でき ます か ？
英　語: (6) Can I exchange money at the hotel?

- **接続詞相当の「では」**

日本語: (7) では， その セット に コーヒー を 付け て ください .

英　語: (8) Well , please add coffee to that set then.

以上に示した通り，日本語の機能語相当の MWE は，用法・意味の曖昧性，および翻訳時の訳し分けの必要性から，機械翻訳において注意して扱うべき対象である．

3　提案手法

3.1　手法の概要

日本語の機能語相当の MWE は，ある程度種類が限られるものの，表現が多様であるため，網羅的に辞書に登録することが難しい．また，辞書に登録できたとしても，前章で示したような用法・意味の曖昧性から，正しく検知することが難しい．

そこで我々は，翻訳精度を向上するという観点から，対訳アライメントの実例に基づいて，機能語相当の MWE を抽出できないかと考えた．そこで，翻訳の精度向上に寄与する範囲で，コーパス中の機能語相当の振る舞いをする最小の形態素系列を MWE としてまとめ上げ，1 形態素と見なす．

翻訳の精度向上に寄与するような，すなわち，対訳アライメントに適した機能語相当の MWE を段階的に検出するために，短い形態素系列から順にまとめ上げる手法を採る．このとき，まとめ上げる形態素列の決定は，対訳アライメントによって獲得したフレーズ対の目視評価によって行い，言語学的見地を加える．

3.2　Minimal Phrase Pair

対訳コーパスの単語アライメント結果から，対称化アライメントを構築したとき，そのフレーズ対に含まれる全てのアライメント端点が自身のフレーズ中に存在し，かつ上記の条件を満たすような二つ以上のフレーズ対に分割不可能なものを，Minimal Phrase Pair と呼ぶ [3]．図 1 に，Minimal Phrase Pair を示す．図において,「地元 の」と “local” ,「料理 を」と “food” ,「食べ」と “eat” ,「たい ん です が」と “I ’ d like to” ,「。」と “.” が，Minimal Phrase Pair に相当する．

Minimal Phrase Pair は，統計翻訳のフレーズテーブルを構成する最小単位となる．そこで，フレーズテーブル全体ではなく，Minimal Phrase Pair に限って評価することで，対訳としての対応関係の付く，より短い単位から評価できるようになり，段階的に日本語 MWE としてまとめ上げることが可能となる．

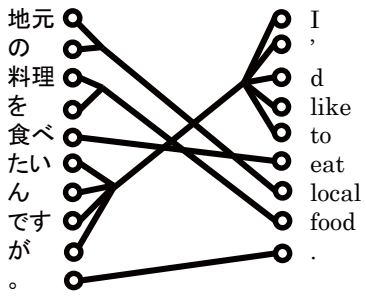

図 1: Minimal Phrase の図解

3.3　日本語 MWE のまとめ上げ

以下に，まとめ上げるべき機能語相当の日本語 MWE を選定する手順を示す．

(手順 1) 対訳コーパスを単語アライメントし，対称化アライメントを構築する

(手順 2) Minimal Phrase Pair を求める

(手順 3) 日本語が 2 形態素以上から構成され，出現回数が 2 回以上の Minimal Pharse を抜き出す

(手順 4) 目視で日本語 MWE を選定し，これを形態素解析辞書に登録する

これらの手順をブートストラップ的に繰り返し，形態素解析辞書に追加する MWE を段階的に獲得する．ここで，手順 4 において登録するべき MWE は，下記の観点でを選定する．

- **日本語 MWE の観点**：

  内容的/機能的用法間の曖昧性，および複数の機能的用法間の曖昧性がない．

- **対訳の観点**：

  同一の日本語フレーズ表記に対応する，全ての英語フレーズが下記の条件を満たす．

  A. 置き換え可能
  B. 倒置が起こっていない
  C. 主語を 2 種類以上含んでいない
  D. 記号のみで構成されない

A–D の基準は，以下の仮定に基づいて設計した．

まず，A の基準では，当該の日本語 MWE を目的言語に翻訳する際に訳し分け不要，すなわち，意味検出が不要であることを判別する．

英語の疑問形が持つ倒置は，日本語では疑問詞の役割を持つ形態素に相当すると考えることができる．英語対訳に倒置を含む場合は，日本語において疑問詞が付加される可能性がある．そこで，言語間で意味が保存できないと仮定し，B の基準を設けた．

日本語では人称主語や形式主語の省略は良く起こる．フレーズ対を見た時に，日本語にはない上記の主語が英語側にわき出しているように見えることがある．このようなフレーズ対は，翻訳時に誤訳を招く原因になると仮定し，C の基準を設けた．

記号は，文脈により異なる役割を持つことが多い．例えば「・」は，並列，同格，区切りなどに表す時に用いられ，用途が広い．そこで，D の基準として，記号を含む日本語フレーズはまとめ上げないこととした．

以上の基準に基づく判定例を，表 1 に示す．この様にすることで，意味が確定的となる形態素系列のみを，MWE としてまとめることができ，短い単位から段階的に構成することができる．

4 実験

4.1 実験条件

対訳コーパスには，独自に開発した旅行ドメインの日英対訳データ 118176 文を用いた．このうち，116676 文を学習に，500 文をパラメータ調整に，1000 文をテストにそれぞれ用いた．

日本語の形態素解析には，MeCab[5] を用いた．形態素辞書には，より短い単位からの MWE 検出を実現するため，形態素が短く設計されている，Unidic[10] を使用した．一方，英語は分かち書きされた単語を表層のまま用いた．また，対訳アライメントには GIZA++[7] を，統計翻訳エンジンとして Moses[4] を使用した．

4.2 実験結果

まず，ベースラインとして，MWE のまとめ上げなしで，そのまま学習させ，翻訳品質を BLEU，RIBES[2]，IMPACT[1] を用いて評価した．結果を表 2 に示す．

次に，3.3 節に述べた手順に従って日本語機能語相当の MWE を段階的にまとめ上げ，それを日本語原文データとして学習させ，翻訳品質を評価した．結果を表 2 に示す．

表 2: 各ループの統計

ループ	baseline	1 回目	2 回目
MPP	-	11,171,067	11,176,392
目視数	-	2115	2053
WP	-	41(17)	9(7)
PT	165,159,570	163,983,168	163,994,614
RIBES	0.691491	0.701933	0.705920
IMPACT	0.4671	0.4696	0.4716
BLEU	26.12	25.30	25.59

MPP:Minimal Phrase Pair 数，目視数:目視評価対象のフレーズ対数，WP:まとめ上げたフレーズ対数（括弧内は表層の異なり数），PT:フレーズテーブルサイズ

1 回目のループでは，抽出した 41 のフレーズ対に含まれる，17 の日本語表現を，下記のようにまとめた．

(17) もう おなか が いっぱい なので デザート は 結構 です .

(18) この 素材 は 色あせ し ませんか ？

まとめ上げを行った結果，RIBES の値は 0.701993 となり，ベースラインの RIBES 値よりも改善した．

2 回目のループでは，抽出した 9 フレーズ対に含まれる 7 つの日本語表現を抽出した．

(19) これ だけ しかない の です か ？

(20) グレープフルーツ の ような 感じ です ね .

1 回目で抽出した 17 表現と 2 回目で抽出した 7 表現を合わせた 24 表現について，全ての実験用コーパスでまとめ上げを行った結果，RIBES 値は 0.705920 となり，2 回目のループ，および，ベースラインからの改善を確認した．

5 おわりに

本稿では，対訳アライメントによって得たフレーズ対を評価し，機能語相当の MWE としてまとめる手順を繰り返すことで，段階的に翻訳精度を改善した．

機能語相当の MWE は，用法，意味の判定が難しく，一括して登録すると翻訳精度に悪影響を与えることがある．提案手法を用いれば，対訳コーパスに頻出する表現を段階的にまとめることができることから，副作用が小さいという利点がある．

参考文献

[1] Hiroshi Echizen'ya, Kenji Araki, and Eduard Hovy. Optimization for efficient determination

表 1: 同一の機能語相当の日本語 MWE 表記に対応する英語表現の分析基準の種類

条件	例
A	(9) (良い例) 日本語:「それ なら」, 英語: “then”, “and then”
	(10) (悪い例) 日本語:「しか あり ませ ん」, 英語: “there are only”, “we only have”
B	(11) (良い例) 日本語:「て い ませ ん」, 英語: “I haven ’ t”, “I didn ’ t”, ”haven ’ t”
	(12) (悪い例) 日本語:「でき ます」, 英語: “I can”, “can I”
C	(13) (良い例) 日本語:「たい です が」, 英語: “I want to”, “I want”, “want to”
	(14) (悪い例) 日本語:「でき ませ ん」, 英語: “I cannot”, “we cannot”
D	(15) (良い例) 日本語:「な の で」, 英語: “so”, “,so”
	(16) (悪い例) 日本語:「か ？」, 英語: “?”

表 3: 機能表現まとめ上げの前処理により次のループで翻訳結果が改善した例

ループ回数	原文 (日本語)	翻訳結果 (英語)
ベースライン	(21) こちら の 中医 治療 室 は 有名 だ と 聞 き ま し た .	(22) this is the Chinese medicine .
1 回目	(23) こちら の 中医 治療 室 は 有名 だ と 聞 き ました .	(24) this is the Chinese medicine clinic I heard that it is famous .

of chunk in automatic evaluation for machine translation. In *Proc. of the OPTHLT-2012/ COLING 2012*, pp. 17–30, 2012.

[2] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proc. of the EMNLP-2010*, pp. 944–952, 2010.

[3] Abraham Ittycheriah and Salim Roukos. Direct translation model 2. In *Proc. of the HLT-NAACL 2007*, pp. 57–64, 2007.

[4] Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proc. of the Second Workshop on Statistical Machine Translation*, pp. 224–227, 2007.

[5] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proc. of the EMNLP-2004*, pp. 230–237, 2004.

[6] Yanjun Ma, Nicolas Stroppa, and Andy Way. Bootstrapping word alignment via word packing. In *Proc. of ACL-2007*, pp. 304–311, 2007.

[7] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[8] グループ・ジャマシィ. 日本語文型辞典. くろしお出版, 1998.

[9] 友松悦子, 宮本淳, 和栗雅子. 日本語表現文型辞典. アルク, 2007.

[10] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源 : 形態素解析用電子化辞書の開発とその応用. 日本語科学, pp. 101–122, 2007.

[11] 坂本明子, 宇津呂武仁, 松吉俊. 日本語機能表現の集約的英訳. 言語処理学会第 15 回年次大会論文集, pp. 654–657, 2009.