表層的な統語素性を用いた チャンキングによる対訳フレーズの抽出

酒井 勇輔 †　　鈴木 寿 ‡
† 中央大学大学院 理工学研究科 情報工学専攻　‡ 中央大学 理工学部 情報工学科
{ysakai@suzuki-lab., suzuki@}ise.chuo-u.ac.jp

1 はじめに

句に基づく統計的機械翻訳 [3] では，対訳コーパスから学習された単語アライメントを基にフレーズテーブルを抽出する．現在主流の単語アライメント手法である IBM モデル [1] では，言語学的な情報を用いないので，言語構造の大きく異なる言語対において正しいアライメントが得られにくい．この問題を解決するべく，言語学的な情報を用いた様々なアライメントモデルが提案されている．

構文情報を統計的なモデルに組み込んだ手法としては，Nakazawa ら [6] の両言語の依存構造木を用いたアライメントモデルが挙げられる．このモデルでは，フレーズの依存関係確率などを用いることによって，言語間の構造の違いに適応し，アライメント性能の向上を実現している．課題点としては，木構造を用いることによる，計算量の増大と構文解析の誤りが考えられる．

また，単語単位のアライメント手法には，意味的な対応関係を持たない機能語が，不適切に対応付けされてしまうという問題が挙げられている．内容語は異なる言語間においても対応関係が明確である場合が多いが，機能語が持つ役割は各言語に固有であることも多いので，対応関係が存在しない場合がある．この性質に着目した研究として Yung ら [12] は，訓練データから対応関係が存在しない語を取り除いてアライメントを学習することにより，翻訳性能の向上を図っている．

これらの背景を踏まえて，本研究では，構文木のような深い統語情報は用いずに，品詞情報などの表層的な統語素性のみを用いて，意味的な対応関係が明確な対訳フレーズを抽出することを目指す．提案手法では，IBM モデルを内容語に限定して適用し正確度の高いアライメントを得たのちに，単語間の結びつきの強さを素性として機能語を近傍の内容語に統合することにより，対訳フレーズを抽出する．

内容語のアライメントにおいては，内容語に特化したヒューリスティックを用いて両方向のモデルを組み合わせる手法が，高いアライメント性能を実現することを確認した．また，抽出された対訳フレーズを用いて翻訳実験をおこなった結果，翻訳性能の上昇を確認した．

2 提案手法

提案する対訳フレーズ抽出法は，独立した 2 つのステップから成り立つ．まず，それぞれの対訳コーパス中から機能語を取り除き，内容語に限定した訓練データを用いてアライメントを学習する．次に，それぞれの言語ごとに機能語を近傍の内容語に統合するチャンキング処理をおこない，対訳フレーズを抽出する．

2.1 内容語のアライメント

機能語が持つ役割は各言語に固有であることも多いので，適切なアライメントを求めることは本質的に難しい．対照的に，内容語については意味的に等価な 1 対 1 の対応関係が得られやすい．図 1 は，日英新聞記事対応付けデータ (JENAAD) [11] の各文ごとに，英語と日本語の語数の差をプロットしたものである．原文では日本語の文の方が語数が多い傾向が見られるが，内容語のみの文の場合は分布の裾が絞られ言語間の偏りも軽減されている．このことからも内容語は比較的対応関係が明確であることが裏付けられる．

以上の性質に基づいて，Yung ら [12] は，対応関係が存在しない語を取り除いたコーパスから IBM モデルによってアライメントの学習をおこなっている．提案手法では，IBM モデルによって学習された 1 対多のアライメントを組み合わせる際に，内容語の特性に着目したヒューリスティックを用いる．

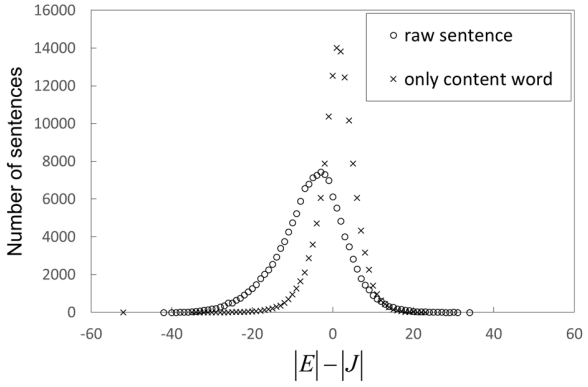

図 1: 英日対訳文の語数差

はじめに対訳コーパス中から機能語である単語を取り除く．またアライメントエラーを減らすために，言語間において内容語の語数が大きく異なる文についても訓練データから取り除く．さらにデータスパースネスに対する頑健性を高めるために，動詞などの活用する内容語については見出し語化をおこなう．次に内容語のみに限定された訓練データから IBM モデルを用いて単語アライメントを学習する．最後に IBM モデルによって得た 1 対多のアライメントを組み合わせ，対称なアライメントを獲得する．

従来のフレーズベース機械翻訳において一般的に用いられる grow-diag-final-and ヒューリスティックでは，両方向の単語アライメントの交差 (intersection) を基にして，隣接する 8 方向に結合アライメントが存在する際にそれらを貪欲に追加する [3]．このモデルは多対多の対応を獲得するためのものであるが，内容語のアライメントでは 1 対 1 の対応を前提としている．そこで，提案する limited-diag-final-and ヒューリスティックでは，8 方向ではなく，斜め 4 方向に隣接する結合アライメントのみを追加する．また，結合アライメントを貪欲に追加するのではなく，上下左右方向にすでにアライメントが存在する際には追加しないよう制限する．全くアライメントが存在しない領域の補間には既存手法と同様の final-and ヒューリスティックを用いる．

2.2 チャンキングの確率モデル

各機能語を近傍の内容語に統合し，フレーズを構成する．この過程を，アライメントが付与された内容語から次の内容語までの各単語間の中から，フレーズの区切り位置を選ぶ分類問題としてモデル化する．

入力文字列 $\boldsymbol{x} = x_1\ x_2\ ...\ x_N$ が与えられたときのフレーズの区切れ位置を，1-of-K 符号化ベクトル $\boldsymbol{y} = y_1\ y_2\ ...\ y_{N-1}$ によって表す．x_1 および x_N はアライメントが付与された内容語である．$\boldsymbol{y}$ の各要素の値は，フレーズの区切れ位置が x_i と x_{i+1} の間にあるとき，$y_j = \delta(j,i)$ となる．$P(y_i = 1 \mid \boldsymbol{x})$ を直接モデル化し，$\arg\max_i P(y_i = 1 \mid \boldsymbol{x})$ をフレーズ区切り位置として決定する．$P(y_i = 1 \mid \boldsymbol{x})$ を次式のように定式化する．

$$P(y_i = 1 \mid \boldsymbol{x}) = \sum_{k=1}^{M} w_k \frac{-\log \phi_k(x_i, x_{i+1})}{\sum_{j=1}^{N-1} -\log \phi_k(x_j, x_{j+1})} \tag{1}$$

w_k は $\sum_{k=1}^{M} w_k = 1$ の制約を満たす混合係数である．$\phi_k(x_i, x_{i+1})$ は x_i と x_{i+1} の結びつきの強さを表す任意の確率関数である．この関数の負の対数をとることにより，“x_i と x_{i+1} が結びつく” という事象の情報量を表現している．分母の正規化項を用いて，$\boldsymbol{x}$ という局所的な範囲内における，相対的な結びつきやすさにスケーリングすることによって，異なる特徴空間から得た M 個の素性値を統一的な尺度で評価している．

素性関数 $\phi_k(x_i, x_{i+1})$ には $M = 8$ として以下の 2-gram 確率を用いた．

$$\phi_1(x_i, x_{i+1}) = P(x_{i+1} \mid x_i) \tag{2}$$
$$\phi_2(x_i, x_{i+1}) = P(x_i \mid x_{i+1}) \tag{3}$$
$$\phi_3(x_i, x_{i+1}) = P(g_{i+1} \mid g_i) \tag{4}$$
$$\phi_4(x_i, x_{i+1}) = P(g_i \mid g_{i+1}) \tag{5}$$
$$\phi_5(x_i, x_{i+1}) = P(g_{i+1} \mid x_i) \tag{6}$$
$$\phi_6(x_i, x_{i+1}) = P(x_i \mid g_{i+1}) \tag{7}$$
$$\phi_7(x_i, x_{i+1}) = P(x_{i+1} \mid g_i) \tag{8}$$
$$\phi_8(x_i, x_{i+1}) = P(g_i \mid x_{i+1}) \tag{9}$$

g_i は x_i の品詞を表す．品詞情報から得られる大局的な統計量と各単語ごとの局所的な統計量を組み合わせることにより，低頻度語の推定に対する頑健性を高めている．また前向きと後ろ向きの 2-gram を用いることにより，修飾の方向に依らない推定をおこなっている．

混合パラメータ w_k の値は言語依存である．例えば日本語の場合は，名詞などの内容語の後に助詞などの機能語が結びついて一つの文節を構成する場合が多い

表 1: 内容語と機能語の概略

言語	内容語	機能語
英語	名詞，代名詞，動詞，形容詞，副詞，間投詞	助動詞，冠詞，前置詞，接続詞，記号
日本語	名詞，動詞，形容詞，副詞，連体詞，感動詞	接頭詞，接続詞，助詞，助動詞，記号，フィラー

ので，前向きの素性に対する重みを大きくすると適切な区切り位置が得られやすい．最適な w_k の値を決定するために，人手によって区切り位置を与えた少量のコーパス $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)}), ..., (\boldsymbol{x}^{(|\mathcal{D}|)}, \boldsymbol{y}^{(|\mathcal{D}|)})\}$ を用いてパラメータチューニングをおこなう．EM アルゴリズムの枠組みに基づいたパラメータチューニングの手順は以下の通りである．

Algorithm 1 Parameter tuning

initialize the mixing parameters w_k
repeat
  for all $(\boldsymbol{x}^{(d)}, \boldsymbol{y}^{(d)})$ such that $1 \leq d \leq |\mathcal{D}|$ **do**
    for all k such that $1 \leq k \leq M$ **do**
      for all i such that $1 \leq i \leq N-1$ **do**
        if $y_i = 1$ **then**
          $\Phi_k(x_i, x_{i+1}) \leftarrow \frac{-\log \phi_k(x_i, x_{i+1})}{\sum_{j=1}^{N-1} -\log \phi_k(x_j, x_{j+1})}$
        else
          $\Phi_k(x_i, x_{i+1}) \leftarrow \frac{-\log(1-\phi_k(x_i, x_{i+1}))}{\sum_{j=1}^{N-1} -\log(1-\phi_k(x_j, x_{j+1}))}$
        end if
      end for
      $\gamma_{d,k} \leftarrow \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{w_k \Phi_k(x_i, x_{i+1})}{\sum_{k'=1}^{M} w_{k'} \Phi_{k'}(x_i, x_{i+1})}$
    end for
  end for
  $w_k^{\text{new}} \leftarrow \frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} \gamma_{d,k}$
until the parameters converge

$\gamma_{d,k}$ は，それぞれのデータ $\boldsymbol{x}^{(d)}$ に対して，混合要素 k が $\boldsymbol{y}^{(d)}$ の観測を説明する，負担の度合いを表している．

3 評価実験

内容語アライメントの性能およびチャンキングによって抽出した対訳フレーズを用いた翻訳の性能の評価をおこなった．

表 2: 内容語アライメントの評価

Method	precision	recall	AER
intersection	97.98	83.03	10.11
grow-diag-final-and	85.11	93.92	10.71
limited-diag-final-and	97.27	91.14	5.90

3.1 内容語アライメントの評価

日英新聞記事対応付けデータ (JENAAD) を用いて内容語のアライメントの性能評価をおこなった．訓練データのサイズは，内容語の語数差が 10 語以上異なる文を除いた 120,000 文，テストデータのサイズは 100 文である．形態素解析器として，英語には Enju のスーパータガー [5]，日本語には MeCab [4] を用いた．それぞれの形態素解析器が出力する品詞情報を基に，内容語と機能語を表 1 のように規定した [1]．IBM モデルの学習には GIZA++ [8] の標準的な設定を用いた．アライメントの評価指標として precision/recall および AER [8] を用いた．

intersection, grow-diag-final-and および limited-diag-final-and によって獲得された，それぞれの内容語アライメントに対する評価結果を表 2 に示す．limited-diag-final-and では precision の低下を最小限に抑えながらも recall を大幅に上昇させることにより，AER の低下を実現している．

3.2 翻訳結果の評価

チャンキングによって抽出された対訳フレーズを用いて翻訳実験をおこなった．ベースライン手法には grow-diag-final-and によるアライメントから抽出されたフレーズテーブルを用いている．提案手法ではチャンキングによって抽出された対訳フレーズをベースラインのフレーズテーブルに追加する．

[1] 実際にはさらに細分された品詞体系に基づいて規定されるが，紙面の都合上概略のみを示す．

表 3: 翻訳結果の評価

Method	JENAAD					KFTT				
	1-gram	2-gram	3-gram	4-gram	BLEU	1-gram	2-gram	3-gram	4-gram	BLEU
Baseline	47.2	15.9	6.7	3.0	11.01	54.5	26.4	14.2	8.2	20.22
Proposed	47.2	16.2	7.0	3.2	11.34	54.4	26.3	14.2	8.3	20.23

対訳コーパスには日英新聞記事対応付けデータ (JENAAD) および京都フリー翻訳タスク (KFTT)[2] を用いた．訓練データのサイズは JENAAD が 120,000 文，KFTT が 329,169 文，開発データのサイズは JENAAD が 3,000 文，KFTT が 1,235 文，テストデータのサイズは JENAAD が 4,213 文，KFTT が 2,326 文となっている．チャンキングのパラメータチューニングには，それぞれの訓練データごとに最初の 100 文を用いた．言語モデルには SRILM [10] によって学習された 5-gram モデルを用いた．デコーダには Moses [2] を用いた．デコーダのパラメータチューニングは MERT [7] によっておこなった．翻訳評価指標として，BLEU [9] を用いた．

翻訳結果に対する BLEU の値とその内訳を表 3 に示す．JENAAD については，大局的な対応関係を用いることにより，2-gram, 3-gram, 4-gram の適合率が上昇している．一方 KFTT については，BLEU の上昇は僅かであった．これは，専門用語に対する形態素解析が適切におこなわれなかったことが理由として考えられる．

4 おわりに

本研究では，既存のアライメント手法を内容語に限定して適用したのちに，機能語を近傍の内容語に統合することにより，対訳フレーズを抽出する手法を提案した．内容語のアライメントにおいては，内容語に特化したヒューリスティックを用いて両方向のモデルを組み合わせる手法が，高いアライメント性能を実現することを確認した．また，抽出された対訳フレーズを用いて翻訳実験をおこなった結果，翻訳性能の上昇を確認した．

提案した手法では，内容語と機能語の区別を人手によって実験的に規定したが，各品詞をどちらに分類すべきかは自明ではない．また，チャンキングの重みパラメータを決定するために，人手によって区切り位置を与えたコーパスを用いた．これらの人手による処理を自動化し，各言語に対する前提知識を必要としないモデルに改善することが，今後の課題である．

[2]Graham Neubig, "The Kyoto Free Translation Task," http://www.phontron.com/kftt/, 2011.

参考文献

[1] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[2] P. Koehn, et al. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. ACL*, 2007.

[3] P. Koehn, F. J. Och, and D. Marcu. Statistical Phrase-based Translation. In *Proc. NAACL*, pp. 48–54, 2003.

[4] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proc. EMNLP*, pp.230–237, 2004.

[5] Y. Miyao, and J. Tsujii. Feature Forest Models for Probabilistic HPSG Parsing. *Computational Linguistics*, 34(1):35–80, 2008.

[6] T. Nakazawa, and S. Kurohashi. Alignment by Bilingual Generation and Monolingual Derivation. In *Proc. COLING*, pp. 1963–1978, 2012.

[7] F. J. Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. ACL*, pp. 160–167, 2003.

[8] F. J. Och, and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

[9] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*, pp. 311–318, 2002.

[10] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. ICSLP*, pp. 901–904, 2002.

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

[12] F. Yung, K. Duh, and Y. Matsumoto. Analysis and Prediction of Unalignable Words in Parallel Text. In *Proc. EACL*, pp. 190–194, 2014.