単語出現頻度を考慮した事後確率制約による単語アライメント

上垣外 英剛 †　　渡辺 太郎 ‡　　高村大也 ††　　奥村学 ††

† 東京工業大学 大学院総合理工学研究科　　‡ 情報通信研究機構

†† 東京工業大学 精密工学研究所

†kamigaito@lr.pi.titech.ac.jp ††{takamura,oku}@pi.titech.ac.jp

1　はじめに

現在主流の統計的機械翻訳では大量のコーパスを用いてフレーズテーブルを作成し, デコードの際にそれらを用いることで翻訳を実現している. フレーズテーブルの作成は単語アライメントの学習結果を用いて行うため, 単語アライメントは機械翻訳において最も重要なタスクである.

IBM Model は単語アライメントを学習する際に最も多く使われている手法である. しかしアライメントが一対多に限定されているという欠点が存在する. この欠点に対する解決法として, ヒューリスティックにより, 両方向のモデルを組み合わせる手法 [4], agreement 学習により M ステップで両方向の十分統計量を結合する手法 [3], 事後確率制約を用いる手法 [2] 等が存在する. これらのモデルは内容語, 機能語の区別を行わず, アライメントも一対一な制約である. 内容語については, 日英のように言語学的に遠い言語対であったとしても容易に対応付けができるが, 機能語は対応付けが難しい.

我々は事後確率制約の枠組みにおいて, 内容語および機能語を区別した素性関数を組み合わせる手法を提案する. 内容語, 機能語の区別として, 辞書などの言語資源を用いる手法と頻度により区別する手法 [6] を試す.

提案手法に対し, 単語アライメントについての評価と, 翻訳結果についての評価の二つの実験を行った. 単語アライメントに対する評価は Hansard 英仏対訳コーパスにおいて, 提案手法による AER の低下を確認した. 翻訳結果に対する評価については, 京都大学フリー翻訳タスクと NTCIR において BLEU の上昇を確認した.

2　関連研究

Och ら [4] は, 翻訳方向によらない対称的なアライメントを生成することが精度の向上に寄与すると考え, ヒューリスティックな手法を用いて両方向のモデルを組み合わせることで, それを実現する手法を提案した. Liang ら [3] は, agreement 学習によって M ステップで両方向の十分統計量を結合することで, 対称的なアライメントを生成する手法を提案している. Ganchev ら [2] は, 事後確率制約付き EM を使用することで, 対称的なアライメントを生成する手法を提案している. 事後確率制約付き EM は E-step において, あらかじめ与えた制約を満たすように事後確率に操作を行うアルゴリズムであり, 通常の EM の事後確率に反することなく制約をより柔軟に扱える.

内容語と機能語がアライメントで結びつくことに対する問題については, Yung らの研究 [7] において検討されている. Yung らは, コーパス中から機能語を取り除いたデータを用いて IBM Model でアライメントを学習し, 学習後に機能語を復元する方法を議論している.

機能語と内容語を区別する研究としては, Setiawan らの研究 [6] が挙げられる. Yung らは予め内容語と機能語のタグ付けを必要としているが, Setiawan らは, 頻度情報を手がかりに閾値を用いて区別している. また, 区別した機能語を用いてフレーズの並び替えを行うことで, BLEU スコアを向上させている.

3　事後確率制約による単語アライメント

現在主流の IBM Model による単語アライメントは, まず片方向のアライメントを生成し, その後にそれらの結果をヒューリスティックな方法で統合して, 双方向の単語アライメントを生成する. この方法で作られる

双方向のアライメントは, 方向によって事後確率が異なる. 双方向の単語アライメントは, 事後確率が方向によらず対称であることが精度の面から望ましい [4] ことが知られている. 事後確率制約による単語アライメントでは, この双方向での同意 (agreement) を表現した素性を仮定し, 事後確率制約付き EM アルゴリズムを用いることで, 両方向で事後確率が等しくなるように学習する [2].

原言語側の文 $\boldsymbol{x}^s$ と, 目的言語側の文 $\boldsymbol{x}^t$ を, 簡単のために合わせて $\boldsymbol{x} = (\boldsymbol{x}^s, \boldsymbol{x}^t)$ と表す. アライメント $\boldsymbol{y}$ に対して, E-step において制約を適用する前の期待値を $\overrightarrow{p}_\theta(\overrightarrow{\boldsymbol{y}}|\boldsymbol{x})$, $\overleftarrow{p}_\theta(\overleftarrow{\boldsymbol{y}}|\boldsymbol{x})$ のように表す. 矢印はアライメントの方向を表しており, 例えば $\overrightarrow{\boldsymbol{y}}$ は $\boldsymbol{x}^s$ から $\boldsymbol{y}^t$ の方向のアライメントであり, x^s の各単語から, x^t の単語への一対多の関係を表す. $\boldsymbol{x}^t$, $\boldsymbol{x}^s$ に対するインデックスを $i, j (1 \leq i \leq |\boldsymbol{x}^t|, 1 \leq j \leq |\boldsymbol{x}^s|)$ と定義する. $\overrightarrow{\boldsymbol{y}}$ の例として, 単語アライメントモデルに IBM Model 1 を用いた場合,

$$p(\boldsymbol{x}^t, \overrightarrow{\boldsymbol{y}}|\boldsymbol{x}^s) = \prod_i \frac{1}{|\boldsymbol{x}^s|+1} p_t(\boldsymbol{x}^t_i|\boldsymbol{x}^s_{\overrightarrow{y}_i}) \tag{1}$$

のように表せる. $\boldsymbol{y}$ を一対多の制約を取り除いた, 多対多の単語アライメントとすると, agreement 制約の素性表現は次式で表される.

$$\phi_{i,j}(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} +1 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j) \\ -1 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i) \\ 0 & otherwise \end{cases} \tag{2}$$

λ を ϕ を重み付けるパラメータとする. 学習の際に, 順方向のアライメントの事後確率が, 逆方向よりも高い時, 逆方向の事後確率が高くなるよう, ϕ の事後確率である $E_{q_\lambda}[\phi_{ij}(x, y)] = 0$ の制約のもとで λ を更新していく. agreement 制約によって求められる双方向のアライメントの事後確率は次式で表される.

$$q_\lambda(\boldsymbol{y}|\boldsymbol{x}) = \frac{\overrightarrow{p}_\theta(\overrightarrow{\boldsymbol{y}}|\boldsymbol{x}) + \overleftarrow{p}_\theta(\overleftarrow{\boldsymbol{y}}|\boldsymbol{x})}{2} \cdot \frac{exp\{-\lambda \cdot \phi(\boldsymbol{x}, \boldsymbol{y})\}}{Z} \tag{3}$$

$$Z = \frac{1}{2}\left(\frac{Z_{\overrightarrow{q}}}{\overrightarrow{p}_\theta} + \frac{Z_{\overleftarrow{q}}}{\overleftarrow{p\theta}}\right) \tag{4}$$

片方向のアライメントの事後確率は次式で表される.

$$\overrightarrow{q}(\overrightarrow{\boldsymbol{y}}|\boldsymbol{x}) = \frac{1}{Z_{\overrightarrow{q}}} \overrightarrow{p}_\theta(\overrightarrow{\boldsymbol{y}}, \boldsymbol{x}) exp\{-\lambda \cdot \phi(\boldsymbol{x}, \boldsymbol{y})\} \tag{5}$$

$$Z_{\overrightarrow{q}} = \sum_{\overrightarrow{y}} \overrightarrow{p}_\theta(\overrightarrow{\boldsymbol{y}}, \boldsymbol{x}) exp\{-\lambda \cdot \phi(\boldsymbol{x}, \boldsymbol{y})\} \tag{6}$$

式とは逆の方向のアライメントについても同様である.

4 単語出現頻度を考慮した素性関数

Ganchev ら [2] は, 事後確率制約における素性関数として, 一対一の対応付けを表現した agreement 制約を用いた. 内容語は, たとえ言語学的に遠い言語対であっても一対一の対応付けが可能であるが, 機能語は一対多, あるいは全く対応しないことが多くなる. そこで, 内容語と機能語とを明示的に区別する素性関数を提案する. Setiawan ら [6] に従い, コーパス中の出現頻度がある閾値より大きなものを機能語, 小さなものを内容語とそれぞれ仮定して扱う.

ミスマッチ制約 冒頭で述べたように, 機能語と内容語が結びつくアライメントは誤りが多いため, そのようなアライメントに対して, ペナルティを与えることで, アライメントの精度は向上すると考えられる. そのために, 機能語と内容語とが対応付けられた場合に事後確率を減らす制約 f2c を考える. 双方向のアライメントの事後確率を求める際に, 常に低い方向の事後確率を agreement 後の事後確率として用いることで, f2c を実現する. 以下簡単のために, アライメントの方向による事後確率の差を $\delta_{i,j}(\boldsymbol{x}, \boldsymbol{y}) = \overrightarrow{p}_\theta(i, j|\boldsymbol{x}, \boldsymbol{y}) - \overleftarrow{p}_\theta(i, j|\boldsymbol{x}, \boldsymbol{y})$ とする. 制約 f2c の素性条件は以下の条件で表される.

$$\phi^{\mathrm{f2c}}_{i,j}(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} +1 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j) \cap ((x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cup (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{C}^s)) \cap (\delta_{i,j}(\boldsymbol{x}, \boldsymbol{y}) > 0) \\ 0 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i) \cap ((x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cup (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{C}^s)) \cap (\delta_{i,j}(\boldsymbol{x}, \boldsymbol{y}) > 0) \\ 0 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j) \cap ((x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cup (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{C}^s)) \cap (\delta_{i,j}(\boldsymbol{x}, \boldsymbol{y}) < 0) \\ -1 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i) \cap ((x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cup (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{C}^s)) \cap (\delta_{i,j}(\boldsymbol{x}, \boldsymbol{y}) < 0) \end{cases} \tag{7}$$

$\mathcal{C}^s$, $\mathcal{C}^t$ および $\mathcal{F}^s$, $\mathcal{F}^t$ は, 原言語および目的言語の内容語および機能語の集合をそれぞれ表している. この条件に当てはまらない場合は agreement 制約の素性条件 (2) を用いる.

マッチング制約 ミスマッチ制約では, 内容語と機能語との対応付けに対して, 直接ペナルティを与えるのに対し, 機能語は機能語と, 内容語は内容語と結びつくようなアライメントの事後確率を, その他のアライメントの事後確率よりも高くすることで, 機能語と内容語が結びつくようなアライメントは減少すると考えられる. ミスマッチ制約とは逆に, 双方向のアライメントの事後確率を求める際に, 常に高い方向のアライメントの事後確率を agreement 後の事後確率として用いることで, これを実現した. 機能語と機能語との

表 1: 単語アライメントに対する評価

	京都大学フリー翻訳タスク					Hansard (仏英)				
制約	precision	recall	AER	$F_1(0.3)$	$F_1(0.5)$	precision	recall	AER	F1(0.3)	F1 (0.5)
agreement	54.60	48.88	48.42	50.46	51.58	95.45	84.82	9.36	87.75	89.82
f2f	53.97	48.93	48.67	50.34	51.33	95.52	**84.99**	9.35	87.75	89.84
c2c	54.65	48.86	48.41	50.46	51.59	**95.56**	84.79	**9.22**	**87.91**	**89.97**
f2c	**54.76**	48.82	48.38	50.46	51.62	95.45	84.87	9.34	87.79	89.85
c2null	54.56	48.96	48.39	50.52	51.61	95.45	84.84	9.35	87.77	89.84
dic	54.63	**49.06**	**48.30**	**50.61**	**51.70**					

マッチング, 内容語と内容語とのマッチングを表現した制約を, それぞれ f2f および c2c として表記する. 制約 f2f の素性条件は以下の条件で表される.

$$\phi^{\mathrm{f2f}}_{i,j}(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} +1 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j) \cap (x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cap(\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) > 0) \\ 0 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i) \cap (x^t_i \notin \mathcal{C}^t \cap x^s_j \notin \mathcal{F}^s) \\ & \cap(\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) > 0) \\ 0 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j) \cap (x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cap(\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) < 0) \\ -1 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i) \cap (x^t_i \notin \mathcal{C}^t \cap x^s_j \notin \mathcal{F}^s) \\ & \cap(\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) < 0) \end{cases} \tag{8}$$

この式の $\mathcal{F}^s$ および $\mathcal{F}^t$ を, それぞれ $\mathcal{C}^s$ および $\mathcal{C}^t$ に変更した条件が, 内容語と内容語のアライメントのマッチングを表現した制約となる. この条件に当てはまらない場合は, agreement 制約の素性条件 (2) を用いる.

その他の制約　上記の内容語や機能語とのマッチングに基づく制約以外に, 内容語の削除に対して, ペナルティを与える制約 c2null, および辞書に基づく制約 dic を提案する. c2null は, 内容語が NULL シンボルと対応付けられて削除されることが問題になるであろうという直感に基づいている. 内容語と NULL が結びつくようなアライメントの場合, E ステップで事後確率を 0 に近づけている. dic は辞書に含まれているアライメントを, 必ず正解として扱う制約で, 辞書に含まれるアライメントの場合, E ステップで事後確率を 1 に近づけている. 辞書としては EDR 日英対訳辞書を使用した.

5　実験

提案手法に対する単語アライメントの精度, またそれらのアライメントを翻訳時に使用した際の精度を評価した. 今回は, 実験データとして, 対訳コーパスである京都大学フリー翻訳タスク (日英), Hansard(仏英), NTCIR(日英) を使用した. Hansard, NTCIR については, 計算時間の問題から, 訓練データとして一部のコーパスのみを使用した. 訓練データのサイズは京都大学フリー翻訳タスクが 325,347 文, Hansard が 100,000 文, NTCIR が 300,000 文となっている.

5.1　単語アライメントの評価

提案手法を用いた場合の単語アライメントの精度を評価した. テストデータのサイズは Hansard が 477 文, 京都大学フリー翻訳タスクが 1170 文である. 京都大学フリー翻訳タスクのテストデータはアライメントに sure, possible の表記がないため, 全てのアライメントを sure として扱った. また Hansard に関しては, 今回仏英対訳辞書を入手できなかったため, 辞書を用いた場合のアライメントは評価していない. 訓練時には, まず IBM Model 1 を用いてアライメントを学習した. 次に, その結果を利用して HMM を初期化して EM を実行し, この結果を評価に用いた. また, 最後に双方向のアライメントを生成する際には, 京都大学フリー翻訳タスクの場合は grow-diag-final を, Hansard の場合は intersect を用いた. 実験の結果を表 1 に示す. 評価は, precision/recall/F_1 および AER[4] で行った. F_1 の括弧内の数字は precision の重みを表している. また, 本実験では, 内容語と機能語とを区別する頻度の閾値を復数試し, 最も良い結果を示している閾値を用いた.

5.2　翻訳結果の評価

生成されたアライメントを用いて, 実際に翻訳した際の結果を評価した. デコードには mosesdecoder を使用した. また, パラメータチューニングは k-best MIRA [1] で行った. 使用した開発データのサイズは, 京都大学フリー翻訳タスク (KFTT) が 1166 文, NTCIR が 2000 文となっている. パラメータチューニングは, それぞれの制約について 5 回行い, BLEU[5] の平均値, および最大値を用いて評価を行った. KFTT における評価の結果を表 2 に, NTCIR における評価の結果を表 3 にそれぞれ示す.

表 2: 京都大学フリー翻訳タスク.

制約	BLEU (平均)	BLEU (最大)
agreement	15.88	15.99
f2c	**16.14**	**16.36**
f2f	16.11	16.33
c2c	16.03	16.25
c2null	16.12	16.24
dic	15.90	16.06

表 3: NTCIR.

制約	BLEU (平均)	BLEU (最大)
agreement	24.42	24.53
f2c	24.38	24.53
f2f	24.47	24.56
c2c	**24.49**	24.60
c2null	24.32	24.48
dic	24.48	**24.62**

6 考察

表 1 の Hansard データにおける評価では, 提案した制約 c2c が最も高い値を示している. これは, 低頻度であればあるほど内容語の割合が高くなることから, 高めの閾値を使用することで, 機能語と内容語のアライメントを内容語と内容語のアライメントとして扱う誤りを減らすことができるためであると考えられる. 一方 f2c については, 適度な閾値を設定することが難しいため, あまり高い値を示していないと考えられる. KFTT における評価も Hansard と同様の傾向を示している. さらに KFTT における評価からは, dic が有効であることが分かる.

表 2 の翻訳実験結果から, KFTT では提案した制約を用いた場合 BLEU が大きく上昇しているのに対し, 表 3 の NTCIR では BLEU の上昇は小幅に留まっている. KFTT では単語アライメントのテストデータがあるため閾値を設定しやすいのに対し, NTCIR は存在しないため, ということが考えられる. さらに, コーパスの傾向が違うということが挙げられる. KFTT は京都観光に関するデータであることから, 人物名や地名といった固有名詞が多く, 本研究で提案している, 低頻度語を内容語と考える方法が適していると考えられる.

提案している制約の比較については最終的な翻訳に対する評価を考えると f2c が最も有効である. これは, 機能語と内容語でアライメントを作りにくくするという目的に, 最も直接的な制約であることが理由だと考えられる. dic については, 単語アライメントに対する評価では高い値を示しているのに対して, 実際に翻訳を行なった際は, BLEU の向上にあまり効果的ではないことが分かる. これは, 問題となっている低頻度語が, ジップの法則に従いロングテールに大量に存在しているために, 辞書ではカバーしきれないためであると考えられる.

7 結論

本研究では, 事後確率制約による単語アライメントに対して内容語と機能語とを明示的に区別する制約を導入した. また, 内容語と機能語との区別として, コーパス中の頻度情報を用い, 例えば内容語と機能語との対応付けに対してペナルティを与えることでより単語アライメントの精度を向上させる手法を提案した. また提案した手法に対して, いくつかの実験を実施した. 単語アライメントの精度については KFTT と Hansard を用いた実験で提案手法を用いた場合の AER の減少を確認した. また翻訳結果の精度については KFTT において提案手法を用いた場合の BLEU の上昇を確認した. ただし翻訳結果については NTCIR をデータとして用いた場合には大きな BLEU の上昇は認められなかった. これは現在閾値を人手で決定しているためであると考えられる. 今後の課題としてはこのような問題を解決するために, 閾値を自動決定する方法の提案や, 閾値を用いない手法の提案を考えている.

参考文献

[1] Colin Cherry and George Foster. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 427–436. Association for Computational Linguistics, 2012.

[2] Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, Vol. 99, pp. 2001–2049, 2010.

[3] Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 104–111, New York City, USA, June 2006. Association for Computational Linguistics.

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

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

[6] Hendra Setiawan, Min-Yen Kan, and Haizhou Li. Ordering phrases with function words. In *Proceedings of the 45th annual meeting on association for computational linguistics*, pp. 712–719. Association for Computational Linguistics, 2007.

[7] Frances Yung, Kevin Duh, and Yuji Matsumoto. Learning core word alignments for statistical machine translation.