Joint Probability SMT モデルを用いた 非直訳文書対間の表現対応づけ

熊野 正[†,‡]　田中 英輝[†]
E-mail: {kumano.t-eq, tanaka.h-ja}@nhk.or.jp
[†]NHK 放送技術研究所
〒157-8510 東京都世田谷区砧1-10-11

徳永 健伸[‡]
take@cl.cs.titech.ac.jp
[‡]東京工業大学 情報理工学研究科
〒152-8552 東京都目黒区大岡山2-12-1

概要

我々は以前より、Marcu らが提案した phrase-based joint probability SMT モデルを拡張することで、コンパラブルコーパス中の対応する表現を推定する手法を提案してきた。本手法は、あらかじめ直訳文対などを抽出することなく、コンパラブルコーパス中の表現対応対を直接推定する。また、推定時に各表現対応対の信頼性を統計検定し、信頼できる表現対のみを用いて推定を進めるため、多くの非訳出表現を含むコンパラブルコーパスに対して頑強に対応推定を行うことができる。本稿では、従来の提案手法を整理して改めて説明するとともに、再実験を行って予備的な性能評価を行った結果について報告する。

1　はじめに

直訳ではないが文書単位で内容が対応しているコンパラブルコーパスは、文などの単位で直訳されているパラレルコーパスに比べて入手しやすく、多言語で発信されるニュース記事に代表されるように、内容もバライティに富んでいる．そのため、コンパラブルコーパスは統計機械翻訳（SMT）の学習コーパスとして注目を集めており、近年このようなコーパスからの学習手法が数多く提案されている。これらの提案の多くは、コンパラブルコーパスから直訳と見なせる文対 [3, 7, 9] やより短い対応個所 [8] を発見した上で、それらから一般的な学習手法を用いて翻訳知識を獲得するものである。

しかしコンパラブルコーパスの中には、単語や表現レベルでの対訳は多数発見できるものの、文のような長さでの明確な直訳関係を見出すことが難しいものもあり、従来手法ではそのようなコーパスから多くの翻訳知識を獲得できない。また翻訳支援における対訳用例読解支援機能の実現などのために、このような文書対に対してできるだけ密に対応関係を付与したいという需要もある。

我々はこれまでに、Marcu らが提案した同時確率に基づく句ベース SMT モデル [5] を拡張することで、このようなコーパスを直接、あらかじめ直訳に近い文対などを抽出せずに、表現アラインメントの学習対象とする手法を提案した [4]。我々の拡張の中心は、表現対応が信頼できるものであるかを統計的検定で確認しながら EM 学習を進めていくところにある。本稿ではこれまでの報告を整理し直して改めて提案手法の説明を行い、また再実験を行って予備的な性能評価を行った結果について報告する。

$$
\begin{aligned}
p(E,F) = \\
& p(e_1, f_1) \cdot p(e_2\ e_3, f_2) \\
& + \\
& p(e_1, f_2) \cdot p(e_2\ e_3, f_1) \\
& + \\
& p(e_1\ e_2, f_1) \cdot p(e_3, f_2) \\
& + \\
& p(e_1\ e_2, f_2) \cdot p(e_3, f_1) \\
& + \\
& p(e_1\ e_2\ e_3, f_1\ f_2)
\end{aligned}
$$

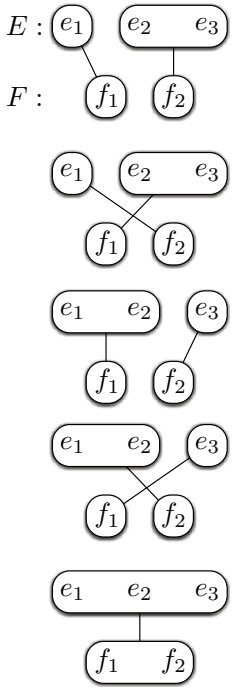

図 1　Joint Probability SMT モデル

2　Phrase-based Joint Probability SMT モデル

Marcu らの同時確率に基づく句ベース SMT モデル [5] は、2 言語文対の生起を、1 個以上の単語列からなる「表現」の 2 言語対（Marcu らは *concept* と呼んでいる）の生起の組み合わせとしてモデル化したものである。彼らの Model 1（表現の出現位置を考慮しないモデル）では、ある 2 言語表現対の集合 C があり、C の各要素の表現を言語ごとに適宜並べることである 2 言語文対 (E, F) を得ることができる時、(E, F) がこの C によって同時に生成される確率 $p(E, F, C)$ を、集合の要素である各表現対

$(\vec{e}, \vec{f})$ の同時生起確率 $p(\vec{e}, \vec{f})$ の積と考える。そして、文対 (E, F) を生成可能な全ての C にわたって $p(E, F, C)$ を足し合わせた結果として、文対 (E, F) の同時生起確率 $p(E, F)$ を定義する（図 1)。これを式で表すと (1) 式となる（$\{C|L(E, F, C)\}$: 言語ごとに表現を適宜並べることで (E, F) を過不足なく生成できるような C の集合)。

$$p(E,F) = \sum_{\{C|L(E,F,C)\}} p(E,F,C) = \sum_{\{C|L(E,F,C)\}} \prod_{(\vec{e},\vec{f})\in C} p(\vec{e},\vec{f}) \quad (1)$$

このモデルにおける $p(\vec{e}, \vec{f})$ の学習は、EM 法を用いて以下の手順で行うことができる。

0. コーパス中の各文対 (E, F) について、(E, F) を生成可能な C による $p(E, F, C)$ が全て等しいと考え、C 中の各要素 $(\vec{e}, \vec{f})$ が (E, F) の生成に使われる回数の期待値を計算。コーパス全体で集計することで、学習コーパスにおける表現対 $(\vec{e}, \vec{f})$ の使用回数の期待値 $t(\vec{e}, \vec{f})$ を初期化する。
1. $t(\vec{e}, \vec{f})$ に応じて $p(\vec{e}, \vec{f})$ を配分する。
2. 学習コーパス中の各文対 (E, F) について、
   (a) (E, F) を生成可能な全ての C を列挙し、それぞれ $p(E, F, C)$ を求める（ただし全ての C を列挙するのは計算量的に困難なため、実装では $p(E, F, C)$ が大きい C を発見的に探索することで列挙に代える)。
   (b) $p(E, F, C)$ に応じて、各 C 中の各要素 $(\vec{e}, \vec{f})$ が (E, F) の生成に使われる回数の期待値を与える。
3. コーパス全体にわたって各 $(\vec{e}, \vec{f})$ の使用回数の期待値を集計し、$t(\vec{e}, \vec{f})$ を更新する。
4. 1.〜3. を繰り返す。

3 非直訳文書対への適用のための拡張

Marcu らの手法は直訳文対の生起をモデル化したものだが、(1) 式を以下のように読み替えるだけで容易に、我々の課題である「非直訳」「文書対」の生起モデルと見なすことができる。既存の多くの SMT 手法で採用されている条件付確率に基づくモデルと異なり、翻訳元と翻訳先の非対訳表現の存在を別々の方法でモデル化する必要はない。

- E, F は各々 1 文以上の連鎖（各文は 1 語以上の連鎖）である文書。
- 表現対 $(\vec{e}, \vec{f})$ はどちらか一方が単語を持たない空表現 (ϕ) であってもよい。また、表現 $\vec{e}, \vec{f}$ は複数の文にまたがらないこととする。

しかし、非直訳文書対コーパスからのこのモデルの学習は、Marcu らの提案手法どおりではうまくいかない。なぜならこのモデルでは、より少ない数の表現対によって学習コーパス中の各文書対を生成できる方が全体として文書対生起確率を高めることができるためである。もし対訳を持たない部分があらかじめ分かっていれば問題ないが、そうでない限り、本来対訳を持たない部分も含めて、できるだけ多くの表現を対応づけるよう学習が進んでしまう。結果として、アラインメント学習で周知の問題である、低頻度表現が “garbage collector” となってしまう現象 [1] が無視できなくなる。

この問題を解決するために、我々は、Moore の提案 [6] を参考に、統計的に有意に相関して生起すると判断できる表現対のみを対訳表現対候補としながら学習を進める手法を提案する。以下では、Marcu らの学習手法の各手順に対して我々が行った拡張を説明する。

[手順 0] $t(\vec{e}, \vec{f})$ の初期化

Marcu らの手法と同様に、コーパス中の各文書対 (E, F) について、(E, F) を生成可能な表現対集合 C が何通りあり、またそれらのうちある表現対 $(\vec{e}, \vec{f})$ を含むものが何通りあるかを計算で求めることにより、$t(\vec{e}, \vec{f})$ の初期化を行う。

文書 E が w_e 単語からなり、かつ s_e 文（各文は 1 単語以上）に分割されている（同様に文書 F は w_f 単語 / s_f 文）時、E, F をそれぞれいくつかの空でない単語列に分割して表現を構成し、さらにそれら表現のいくつかを言語間で対応づける（残った未対応表現は各々 ϕ と対応づけたものと見なす）ような表現間アラインメントのやり方は、(2) 式の $A(w_e, s_e, w_f, s_f)$ 通りだけある。

$$A(w_e,s_e,w_f,s_f) = \sum_{k=0}^{\min(w_e,w_f)} k! \sum_{i=\max(k,s_e)}^{w_e} \sum_{j=\max(k,s_f)}^{w_f} \binom{w_e-s_e}{i-s_e}\binom{i}{k}\binom{w_f-s_f}{j-s_f}\binom{j}{k} \quad (2)$$

また、E, F 中のあるそれぞれ l_e, l_f 語の区間を考えた時、これらがいずれもちょうど 1 つの表現 $\vec{e}, \vec{f}$ になるように E, F が分割され、かつ両表現が対応づけられているようなアラインメントのやり方は、これらの単語列を各々 E, F から取り去った残りをアラインメントするやり方の半分（あとの半分は両表現がいずれも ϕ と対応づけられている場合)、すなわち $(A(w_e - l_e, s_e + \delta_e, w_f - l_f, s_f + \delta_f) / 2)$ 通り（δ_e, δ_f は単語列を取り去ることで文が分割されうることによる補正項) だけある。従って、$A(w_e, s_e, w_f, s_f)$ 通りのアラインメントのうち、あるそれぞれ l_e, l_f 語の区間がちょうど表現対として対応づけられる確率は (3) 式で計算できる。

$$\frac{A(w_e - l_e, s_e + \delta_e, w_f - l_f, s_f + \delta_f)}{2\,A(w_e, s_e, w_f, s_f)} \quad (3)$$

もし E, F のどちらもその中に重複した単語列がないならば、(2) 式で求めた $A(w_e, s_e, w_f, s_f)$ 種類の各アラ

インメントでできる表現対集合は全て異なるものとなるからこれを C の種類数と見なし、(3) 式を文書対の生起にこの表現対が使われる回数の期待値と見なすことができる。しかしそうでない場合、$A(w_e, s_e, w_f, s_f)$ 種類のアラインメントの中には同じ表現対集合が重複して現れることになる。モデルの定義より、同じ表現対集合から生成可能な文対はアラインメントが異なっていても区別されないため、この重複を取り除いたものが C の種類数となる。重複数を厳密に計算するのは非常に計算量が大きいため、概算として (4) 式による近似を C の種類数と見なす（$c(E, \vec{e})$ は E 中に $\vec{e}$ と同じ単語列が現れる回数、$e_1, e2, \ldots$ はその各単語列を指す。F 側も同様）。

$$C(E, w_e, s_e, F, w_f, s_f) \simeq A(w_e, s_e, w_f, s_f) \times \prod_{\{(\vec{e},\vec{f})|c(E,\vec{e})c(F,\vec{f})>1\}} \left(1 - \left(1 - \frac{1}{c(E,\vec{e})c(F,\vec{f})}\right) \times \sum_{i=1}^{c(E,\vec{e})} \sum_{j=1}^{c(F,\vec{f})} \frac{A(w_e - l_e, s_e + \delta_{e_i}, w_f - l_f, s_f + \delta_{f_j})}{2\,A(w_e, s_e, w_f, s_f)}\right) \tag{4}$$

以上より、文書対 (E, F) の生成に表現対 $(\vec{e}, \vec{f})$ が使われる回数の期待値 $t(\vec{e}, \vec{f}|(E, F))$ は (5) 式のように初期化でき、これを表現対ごとにコーパス全体で集計することで、各 $t(\vec{e}, \vec{f})$ の初期値を得る。

$$t(\vec{e}, \vec{f}|(E, F)) = \frac{C(E - \vec{e}, w_e - l_e, s_e + \delta_e, F - \vec{f}, w_f - l_f, s_f + \delta_f)}{2\,C(E, w_e, s_e, F, w_f, s_f)} \tag{5}$$

また、手順 1 で必要となる、表現対 $(\vec{e}, \vec{f})$ の「共起回数」 $o(\vec{e}, \vec{f})$ を求めておく。我々は共起を「対応づけ可能な形で共に出現する」現象と定義する。従って、文書対 (E, F) 中の表現対 $(\vec{e}, \vec{f})$ の共起回数 $o(\vec{e}, \vec{f}|(E, F))$ は、「実際に表現対として (E, F) の生起に使われる回数の期待値」と「$\vec{e}$, $\vec{f}$ がともに ϕ と対応づけられる形で出現する回数の期待値」の和として計算できる。これらは初期状態では等しく $t(\vec{e}, \vec{f}|(E, F))$ となるので、$o(\vec{e}, \vec{f}|(E, F)) = 2\,t(\vec{e}, \vec{f}|(E, F))$ となり、これを表現対ごとにコーパス全体で集計することで、各 $o(\vec{e}, \vec{f})$ の初期値を得る。

[手順 1] LLR を用いた表現対の検定と $p(\vec{e}, \vec{f})$ の計算

各表現対 $(\vec{e}, \vec{f})$ について共起回数に基づく独立性検定を行い、ある有意水準にて出現に相関が認められた表現対についてのみ、$p(\vec{e}, \vec{f}) = t(\vec{e}, \vec{f}) \,/\, \sum_{(\vec{e},\vec{f})} t(\vec{e}, \vec{f})$ を信頼できる確率値として採用する。検定の尺度には、対数尤度比（LLR）を用いる。LLR の計算方法については Dunning の論文 [2] を参照されたい。

[手順 2〜3] $t(\vec{e}, \vec{f})$ の更新

Marcu らの手法と同様に、学習コーパス中の各文対 (E, F) について、(E, F) を生成可能な全ての C を列挙し、それぞれ $p(E, F, C)$ を求める（ただし実装では、全ての C を列挙する代わりに $p(E, F, C)$ が大きい C を発見的に探索する。なぜなら $p(E, F, C)$ の小さい C は (E, F) の生起への寄与が小さく、無視可能だからである）。

ただし、手順 1 で行った検定の結果、「信頼できない」とされた表現対の $p(\vec{e}, \vec{f})$ は $p(E, F, C)$ の計算に使用しない。その代わりに、C 中の信頼できない表現対が覆っている E, F 中の単語の 1-gram 確率を用いて、(6) 式のように $p(E, F, C)$ を見積もる（$R(\vec{e}, \vec{f})$ は $(\vec{e}, \vec{f})$ が信頼できるときに真、$p(e)$, $p(f)$ は各言語の単語 1-gram 確率）。

$$p(E, F, C) \simeq \prod_{\{(\vec{e},\vec{f})\in C|R(\vec{e},\vec{f})\}} p(\vec{e}, \vec{f}) \times \prod_{\{(\vec{e},\vec{f})\in C|\neg R(\vec{e},\vec{f})\}} \left(\prod_{e\in\vec{e}} p(e) \cdot \prod_{f\in\vec{f}} p(f)\right) \tag{6}$$

(E, F) を生成可能な各 C が (E, F) の生成に使われる確率は $p(E, F, C) \,/\, \sum_C p(E, F, C)$ であるから、C の要素である各 $(\vec{e}, \vec{f})$ が (E, F) の生成に使われる回数の期待値を見積もることができ、これを全ての C について、さらに全ての (E, F) について集計することで、更新された $t(\vec{e}, \vec{f})$ を得ることができる。また、同様の手順で更新された $o(\vec{e}, \vec{f})$ も得ることができる。

$p(E, F, C)$ の計算にあたっては、C のうち信頼できる表現対の部分が同じものは等確率となることから、手順 0 の初期化時の計算と類似の方法を用いることで、これら同一確率となる C を全て列挙することなしに更新の計算を行うことができる。

4 実験

提案手法の予備的な評価として、NHK 日英ニュースコーパスを用いて提案手法によるモデル学習を行い、あらかじめ定めた回数の繰り返し後得られた、各記事対に対する最尤のアラインメントの精度を人手で評価した。NHK 日英ニュースコーパスは数文程度からなる日英の記事対の集まりで、各記事対は同一話題について書かれたものであるが、各々相手言語側に存在しない内容を含むことが多く、また明確な文対応も認定しにくい。

実験は表 1 に示すように、コーパス規模（大規模コーパスは小規模コーパスを完全に含む）、LLR 閾値、繰り返し回数の組み合わせを変えて 6 種類行った。評価は、各実験共通の 10 記事対に対して、1 人の評価者が各表現対応の正しさを 3 値（完全に対訳、部分的に対訳、対訳でない）で判定した。表 1 に正解率（完全もしくは部分的に対訳と判定された表現対の割合）、および実験の結果得られた表現対の平均長（単語数）と被覆率（記事対中の単語のうち、実験の結果得られた表現対によって相手言語側との対応を持っているものの割合）を示す。

ここに示す評価結果は暫定的なものであり、現在構築

実験		1	2	3	4	5	6
コーパス規模（記事対数）		1,000					2,000
（延べ/異なり 単語数）	日	287,597 / 10,855					578,374 / 18,182
	英	161,976 / 10,521					312,353 / 17,905
LLR 閾値（対応する $\chi^2(1)$ 確率[*1]）		3.841 (.95)			2.706 (.80)	.4549 (.50)	3.841 (.95)
繰り返し回数		1	3	5			
完全 or 部分正解率（延べ/異なり）		.520/.468	.767/.709	.837/.743	.624/.495	.410/.309	.848/.761
表現の平均長	日	1.14	1.25	1.26	1.34	1.26	1.29
（単語数）	英	1.08	1.23	1.21	1.20	1.19	1.23
被覆率	日	.177	.198	.211	.293	.440	.251
（単語の割合）	英	.190	.226	.247	.365	.536	.284

表 1　実験条件と結果

中の評価用正解データを用いて後日精緻な評価を行う予定であるが、少なくとも現時点の実験結果から以下のような傾向が読み取れる。

- 実験 3〜5 の比較より、LLR 閾値を上げると対訳表現の正解率が向上することから、LLR による表現対の検定は “garbage collector” の抑制に役立っているものと考えられる。
- 実験 1〜3 の比較より、繰り返し回数を増やすと正解率、被覆率とも向上することから、表現対の検定を組み込んだ EM 法による学習は正しく機能していると考えられる。
- 実験 3 と 6 の比較より、LLR 閾値を変えずにコーパス規模を大きくすると被覆率は向上するが正解率はあまり変化しないことから、LLR による表現対の検定はコーパス規模によらず表現対応の正解率を一定に保つよう機能していると考えられる。

5　おわりに

我々が以前より提案していた、phrase-based joint probability SMT モデルを拡張することでコンパラブルコーパスから直接表現アラインメントを学習する手法を整理し直し、改めて報告した。予備的な性能評価の結果から、数文程度の日英非直訳文書対からなるコーパスに対して安定した性能で Marcu らの Model 1 相当の表現対応推定が実現可能であることを示した。計算コストなどの問題を解決していくことで、表現の出現位置や構文など、より高次の制約を取り入れたモデルをコンパラブルコーパスから学習するための土台としたい。

また、現在本研究の評価用に人手で表現対応を付与したコーパスを構築中である。本稿で報告した予備実験では対訳推定の再現率を評価することができないため、後日改めて精緻な評価を行った結果を報告する機会を持ちたい。

[*1] LLR 統計量は漸近的に $\chi^2(1)$ 分布に近似できる。

参考文献

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer, and Surya Mohanty. But dictionaries are data too. In *Proceedings of the ARPA Workshop on Human Language Technology (HLT '93)*, pp. 202–205, 1993.

[2] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, Vol. 19, No. 1, pp. 61–74, 1993.

[3] Pascale Fung and Percy Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pp. 1051–1057, 2004.

[4] Tadashi Kumano, Hideki Tanaka, and Takenobu Tokunaga. Extracting phrasal alignments from comparable corpora by using joint probability SMT model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pp. 95–103, 2007.

[5] Daniel Marcu and William Wong. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 133–139, 2002.

[6] Robert C. Moore. Improving IBM word alignment model 1. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL '04)*, pp. 518–525, 2004.

[7] Dragos Stefan Munteanu and Daniel Marcu. Processing comparable corpora with bilingual suffix trees. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 289–295, 2002.

[8] Dragos Stefan Munteanu and Daniel Marcu. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pp. 81–88, 2006.

[9] Bing Zhao and Stephan Vogel. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the IEEE International Conference on Data Mining (ICDM 2002)*, pp. 745–748, 2002.