非パラレルコーパスを用いた統計的機械翻訳の分野適応

岡田　大輔　　綱川　隆司　　梶　博行

静岡大学大学院情報学研究科

gs08012@s.inf.shizuoka.ac.jp, {tuna, kaji}@inf.shizuoka.ac.jp

1 はじめに

近年、インターネットの発達・普及により外国語の文書を目にする機会が増えている。機械翻訳の重要性がますます高くなり、次世代機械翻訳へのアプローチとして統計的機械翻訳(SMT)が盛んに研究されている。SMT は、対象分野のコーパスを用意するだけで、その分野の語彙や言い回しに適応した翻訳モデルを学習するという特徴をもつ(Brown et al., 1990)。しかし、大規模なパラレルコーパスが利用できる分野は限られるという問題がある。

この問題を解決するため、本稿では、ターゲット言語とソース言語の単言語コーパスを組にした非パラレルコーパスから翻訳モデルを学習する方法を提案する。さらに、対象分野の非パラレルコーパスから提案方法で学習した翻訳モデルを他の分野のパラレルコーパスから従来方法で学習した翻訳モデルを組み合わせ、パラレルコーパスが存在しない分野に SMT を適用する方法を示す。

2 非パラレルコーパスからの擬似翻訳確率の推定

2.1 名詞の擬似翻訳確率

非パラレルコーパスは文や語の対応関係を含まないので、アラインメントに基づく従来の翻訳モデル学習方法を適用することは困難である。そこで、非パラレルコーパスを用いて、ソース言語の語（以下、対象語と呼ぶ）の関連語がターゲット言語のどの訳語を支持するかを決定する方法(Kaji and Morimoto, 2002)に着目する。そして、対象語の関連語のうち各訳語を支持する関連語の比率を擬似翻訳確率と呼ぶことにする。支持する関連語が多い訳語ほど、その訳語に翻訳すべき用例がコーパス中に多く含まれている、すなわち高い翻訳確率をもつと考えられるからである。

図 1 に示すように、名詞[1]の擬似翻訳確率を以下の(1)～(3)のステップにより推定する。

(1) 関連語ペアの抽出

ソース言語とターゲット言語それぞれのコーパスから、ウィンドウ共起に基づく相互情報量が閾値θ以上である語のペアを抽出する。2 つの語 t と t'の相互情報量は次式で定義される。

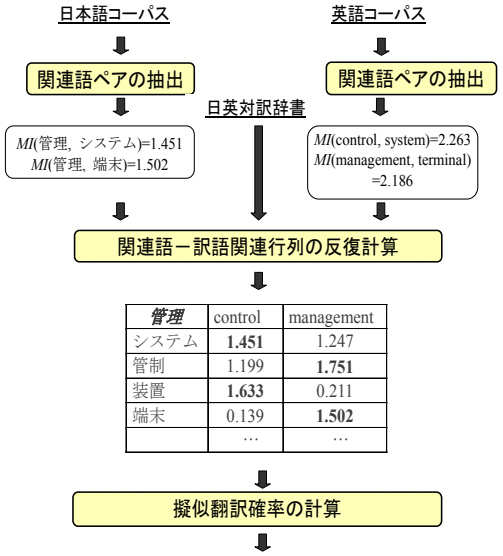

図 1: 名詞の擬似翻訳確率の推定

$$MI(t,t') = \log_2 \frac{P(t,t')}{P(t)\cdot P(t')}.$$

ここで、$P(t)$はコーパス中に t が出現する確率、$P(t, t')$はウィンドウ内で t と t' が共起する確率である。

4. の評価実験では、ウィンドウサイズを±12 語（内容語のみカウント）とした。また、相互情報量に対する閾値θは-5 とした。

(2) 関連語－訳語関連行列の反復計算

「相互に関連のある関連語は同じ訳語との関連度が高い」という仮説に基づき、対象語 f の第 i 関連語 $f'(i)$ と第 j 訳語 $e(j)$の関連度 C を次式で再帰的に定義する。

$$C_n(f'(i),e(j)) = MI(f'(i),f)\cdot \frac{\sum_{f''\in A(f,f'(i))} MI(f'(i),f'')\cdot C_{n-1}(f'',e(j))}{\max_k \sum_{f''\in A(f,f'(i))} MI(f'(i),f'')\cdot C_{n-1}(f'',e(k))}.$$

ここで、$A(f,f'(i))$は対象語 f と関連語 $f'(i)$に共通の関連語の集合である。すなわち、

$$A(f,f'(i)) = \{f'' \mid MI(f,f'')\ge\theta, MI(f'(i),f'')\ge\theta\}.$$

なお、C の添字 n は反復計算のサイクルを示す。

また、初期値 C_0 は、対訳辞書を介した関連語ペアのアラインメントに基づき次式で与える。

[1] 本稿では対象語を名詞に限定する。提案方法では関連語をウィンドウ共起に基づいて抽出している。このような関連語は名詞の訳語決定の手がかりとして有効であるが、動詞の訳語決定には必ずしも有効ではない。対象語が動詞の場合は、関連語を構文共起に基づいて抽出することが望ましい。

$$C_0(f'(i),e(j))=\begin{cases}\delta(f'(i),e(j))/\sum_k \delta(f'(i),e(k)) & \\ \qquad\cdots \sum_k \delta(f'(i),e(k))\neq 0 & \\ 0 \qquad \cdots otherwise & \end{cases}$$

$$\delta(f'(i),e(j))=\begin{cases}1 & \cdots \exists e'.(f,f'(i))\approx(e(j),e') \\ 0 & \cdots otherwise\end{cases}$$

ここで、≈は関連語ペアのアラインメントを表す。すなわち、$(f,f')\approx(e,e')$ は f と e、f' と e'がそれぞれ対訳関係にあることを表す。

これらの式からわかるように、曖昧性のない関連語ペアのアラインメントが種となって、関連語－訳語関連行列が反復計算される。

(3) 擬似翻訳確率の計算

対象語の各関連語はそれとの関連度が最大の訳語を支持すると考え、対象語 f の第 j 訳語 $e(j)$への擬似翻訳確率 $P_{pseudo}(e(j)|f)$を次式で定義する。

$$P_{pseudo}(e(j)\,|\,f)=\frac{|\,S(e(j))\,|+\varepsilon}{\sum_k(|\,S(e(k))\,|+\varepsilon)}.$$

ここで、$S(e(j))$は訳語 $e(j)$を支持する f の関連語の集合である。すなわち、

$$S(e(j))=\left\{f'(i)\,|\,C(f'(i),e(j))>\max_{k\neq j}C(f'(i),e(k))\right\}.$$

なお、εは、$S(e(j))$が空集合であっても擬似翻訳確率が 0 にならないようにするための微小な定数値である。4. の評価実験ではε=0.025 とした。

図 1 には、対象語“管理”に対する関連語－訳語関連行列の一部が例示されている。関連語“システム”、“装置”は訳語“control”を、関連語“管制”、“端末”は訳語“management”をそれぞれ支持し、“control”と“management”への擬似翻訳確率がそれぞれ 0.70、0.30 と推定されている。

2.2 名詞列の擬似翻訳確率

関連語－訳語関連行列は対訳辞書に含まれる対象語とその訳語について計算されるので、2.1 で提案した方法では 2 語以上からなるフレーズの擬似翻訳確率を計算することはできない。フレーズを名詞列に限り、2 語以上からなるフレーズの擬似翻訳確率を以下の(1)～(3)のステップにより推定する（図 2 参照）。

(1) 名詞列の抽出と頻度のカウント

ソース言語とターゲット言語それぞれのコーパスから、２語以上の名詞列を抽出しその頻度をカウントする。なお、より長い名詞列の部分列となっている名詞列も頻度カウントの対象とする。

(2) 名詞列のアラインメント

対訳辞書を参照しながら、構成要素の間に対訳関係が成立する名詞列を対応づける。すなわち、f_1 と e_1, f_2 と e_2, .., f_n と e_n それぞれの組の間に対訳関係が成立するとき、２つの名詞列 $F=f_1f_2\ldots f_n$ と $E=e_1e_2\ldots e_n$ を対応づける。

(3) 擬似翻訳確率の計算

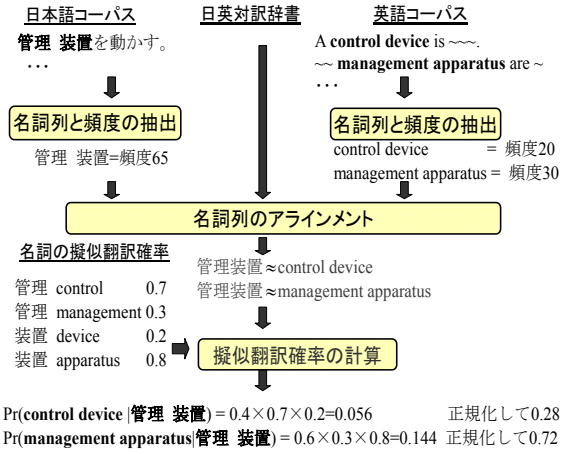

図 2:名詞列の擬似翻訳確率の推定

(2)の結果、名詞列 $F=f_1f_2\ldots f_n$ と m 個の名詞列 $E(j)=e_1(j)e_2(j)\ldots e_n(j)$（$j$=1, 2, .., m）のアラインメントが得られたとする。このとき、擬似翻訳確率を次式で定義する。

$$P_{pseudo}(E(j)\,|\,F)=\frac{\#(E(j))\cdot\prod_i P_{pseudo}(e_i(j)\,|\,f_i)}{\sum_k \#(E(k))\cdot\prod_i P_{pseudo}(e_i(k)\,|\,f_i)}.$$

ここで、$\#(E(j))$は名詞列 $E(j)$の出現頻度を表す。

この式は、名詞列 $E(j)$の出現頻度に構成要素間の擬似翻訳確率 $P_{pseudo}(e_i(j)|f_i)$（i=1, 2, …, n）の積を乗算した値を求め、正規化することを表している。名詞列のアラインメントがすべて正しければ、$E(j)$の出現頻度の比率を翻訳確率とするのがよい。構成要素間の擬似翻訳確率の積を乗算する理由は、名詞列のアラインメントの誤りの影響を小さくするためである。対訳辞書はさまざまな文脈で成立する可能性のある対訳関係を含んでいるため、対訳辞書を介した名詞列のアラインメントでは誤ったアラインメントが得られることがある。しかし、誤ったアラインメントの場合、構成要素間の擬似翻訳確率もすべてが大きな値をもつ可能性は小さい。上の式によれば、訳語として正しくない名詞列に大きな翻訳確率を与えることを防ぐことができると考える。

3 擬似翻訳確率を用いた SMT の分野適応

2. で提案した方法で推定した擬似翻訳確率だけでは SMT を実行することはできない。擬似翻訳確率を求めることができるのは名詞と名詞列のみである。また、相互情報量に基づく関連語ペアの抽出を基本としているため、コーパス中の出現頻度が低い語に対しては擬似翻訳確率を計算することができないからである。

そこで、分野外のパラレルコーパスから学習した翻訳モデルを分野内の非パラレルコーパスを用いて分野に適応させるアプローチをとる。すなわち、図 3 に示すように、GIZA++ (Och and Ney, 2002)と train-factored-phrase-model.perl (Koehn et al., 2003)により分野外のパ

ラレルコーパスから学習した翻訳確率と提案方法により分野内の非パラレルコーパスから推定した擬似翻訳確率の平均をとる。一方の方法で翻訳確率が推定できないフレーズ対については、他方の方法で推定された値をそのまま採用する。なお、ターゲット言語の言語モデル（N グラム確率）は、SRILM (Stolcke, 2002)を用いて分野内のターゲット言語コーパスから学習する。また、SMT システム（デコーダ）として Moses を利用する。

4 評価実験

4.1 実験方法

4.1.1 実験 1

提案方法により分野適応させた SMT と分野外パラレルコーパスから学習した翻訳モデルをそのまま用いる SMT（以下、従来方法と呼ぶ）の比較実験を行った。なお、英単語の大文字の小文字化は行わなかった。

実験に使用したコーパスと対訳辞書は次のとおりである。

(1) トレーニングコーパス

(a) 分野外パラレルコーパス：Japio の 2003 年の日英特許抄録の物理分野（20,000 抄録(日 5.32MB, 英 4.54MB)）。

(b) 分野内非パラレルコーパス：JSTの 1981 年から 2005 年までの日英科学技術文献抄録[2]の基礎化学分野（日 151,958 抄録(90.8MB), 英 102,730 抄録(64.9MB)）。ただし、(2)のテストコーパスとして抽出した部分を除外。

(2) テストコーパス

(1)の(b)のコーパスから抽出した対訳文 1000 文。対訳辞書を参照して抄録対に含まれる日英の文の類似度を計算し(Utiyama and Isahara, 2007)、類似度の高いペアを抽出した。

(3) 対訳辞書

EDR 対訳辞書、英辞郎、EDICT から名詞のみを抽出してマージした辞書を使用した。日本語が 163,247 語、英語が 93,727 語、日英の対訳関係が 333,656 対含まれる。

翻訳方向は日本語から英語とし、テストコーパス中の日本語文を翻訳した。テストコーパ中の英文はレファレンス訳として利用した。従来方法として次の２つのケースを実行した。

- 従来方法：分野外パラレルコーパスから学習した翻訳モデルをそのまま使用する。
- 従来方法+辞書：対訳辞書の全ての訳語に一様な擬似翻訳確率を与え、分野外パラレルコーパスから学習した翻訳確率との平均をとる。

[2] 科学技術文献抄録データは基本的には同一の文献に対する日英の抄録対であるが、対訳であるとは限らず、一方の言語の抄録が欠けているものも多いので、全体としては非パラレルコーパスといえる。

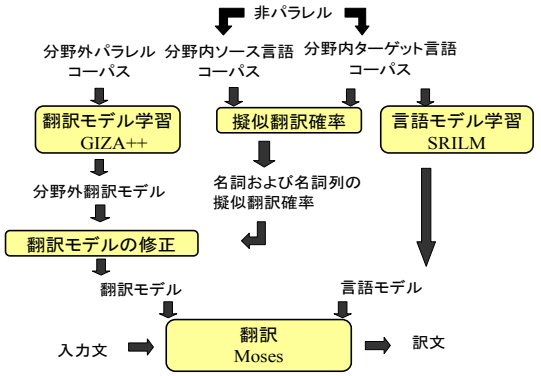

図 3:擬似翻訳確率を用いた SMT の分野適応

また、提案方法については、トレーニングコーパスとして使用する分野内非パラレルコーパスの量が異なる４つのケースを実行した。

(i) (1)の(b)のコーパス全体を使用。

(ii) (1)の(b)のうち日本語抄録は半分を使用。

(iii) (1)の(b)のうち英語抄録は半分を使用。

(iv) (1)の(b)のうち日本語抄録、英語抄録とも半分を使用。

4.1.2 実験 2

実験 1 の分野外パラレルコーパスを分野内パラレルコーパスに置き換えた実験を行った。その目的は、パラレルコーパスは小規模なものしか利用できないが非パラレルコーパスは大規模なものが利用できる分野においても提案方法が有効であることを示すためである。分野内パラレルコーパスとして、テストコーパスの作成と同様の方法で(1)の(b)のコーパスから抽出した対

表 2:実験 1 の結果

	トレーニングコーパス	BLEU
従来方法	分野外パラレルコーパス（物理）	**0.1142**
従来方法+辞書		**0.1294**
提案方法(i)	分野内非パラレルコーパス（基礎化学）	**0.1313**
提案方法(ii)		**0.1307**
提案方法(iii)		**0.1318**
提案方法(iv)		**0.1307**

表 3:実験 2 の結果

	トレーニングコーパス	BLEU
従来方法	分野内パラレルコーパス（基礎化学）	**0.1637**
従来方法+辞書		**0.1632**
提案方法(i)	分野内非パラレルコーパス（基礎化学）	**0.1690**
提案方法(ii)		**0.1675**
提案方法(iii)		**0.1677**
提案方法(iv)		**0.1665**

訳文 20,000 文(日 3.61MB,英 3.17MB)を使用した。当然のことながら、テストコーパスとして抽出した対訳文とは重複しないようにした。

4.2 実験結果

各方法による翻訳結果に対して BLEU スコア(Papineni et al., 2002)を算出した。BLEU スコアは翻訳結果とレファレンス訳との n-gram 適合率を示すもので、本実験では 1〜4 -gram の適合率を用いた。実験 1 と実験 2 の結果をそれぞれ表 2 と表 3 にまとめた。

4.3 結果の検討

実験の結果から、以下の結論を得た。

(1) 実験 1 の結果から、提案方法が分野適応に有効であるといえる。なお、単純に従来方法の翻訳モデルに対訳辞書を追加しただけの場合と比較しても BLEU スコアは向上している。

(2) 実験 2 の結果から、パラレルコーパスが分野外の場合だけでなく、パラレルコーパスが分野内の場合でも、提案方法により翻訳精度が向上するといえる。分野内パラレルコーパスを利用する場合は、分野外パラレルコーパスを利用する場合と異なり、対訳辞書を組合せただけでは翻訳精度が向上しないこともわかった。

(3) 非パラレルコーパス全体を利用した場合と日英とも半分にした場合を比較すると、前者の BLEU スコアのほうが高い。提案方法の効果は一般的にコーパスの量に応じて大きくなると考えられる。しかし、実験 1 では、英語抄録のみを半分にした場合の BLEU スコアがさらに高くなっており、より多くのケースについて比較してみる必要がある。

5 今後の課題

(1) 擬似翻訳確率の推定の計算パラメータ最適化

関連語ペアの抽出におけるウィンドウサイズ、共起頻度の閾値、相互情報量の閾値などの値を変更して実験し、最適値を探す。

(2) フレーズテーブルの修正の重み最適化

提案方法で推定された擬似翻訳確率と従来方法で推定された翻訳確率の平均をとっているが、擬似翻訳確率の重みを変えて実験し、最適値を探す。

(3) 動詞に対する擬似翻訳確率の推定

動詞の訳語を決定の手がかりとしては目的語などが有効であり、構文共起に基づいて関連語を抽出することが望ましい。関連語－訳語関連行列の反復計算のアルゴリズムもそれに合わせて変形することが必要である。

6 関連研究

Koehn and Knight(2000)は、EM アルゴリズムを用いて非パラレルコーパスから翻訳確率を求める方法を提案している。この方法ではターゲット言語コーパス中の訳語の出現頻度に強く影響された結果が得られる。これに対し、提案方法ではソース言語コーパス中の対象語の語義の分布を反映した結果が得られる。

また、Wu, et al.(2008)は、対象分野の対訳辞書を用いて、分野外のパラレルコーパスから学習した翻訳モデルを対象分野に適応させる方法を提案している。しかし、対訳辞書は翻訳確率の分野適応に十分な情報を含んでいないと思われる。

7 むすび

非パラレルコーパスから名詞や名詞列の擬似翻訳確率を推定する方法を提案した。提案方法を用いて分野内非パラレルコーパスから推定した擬似翻訳確率を分野外パラレルコーパス（または分野内小規模パラレルコーパス）から推定した翻訳確率に重ね合わせることにより、SMT の BLEU スコアが向上することを実験により確認した。今後の課題として、擬似翻訳確率の推定におけるパラメータの最適化、動詞の擬似翻訳確率推定方法の開発などがあげられる。

謝辞：科学技術文献抄録をご提供いただいた（独）科学技術振興機構と特許抄録をご提供いただいた（財）日本特許情報機構に感謝申し上げます。

参考文献

Brown, Peter F., John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics,* Vol. 16. No. 2, pp. 79–85.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Unsupervised Word Sense Disambiguation Using Bilingual Comparable Corpora. *Proc. COLING 2002,* pp. 411–417.

Koehn, Philipp and Kevin Knight. 2000. Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm. *Proc. AAAI 2000,* pp. 711-715.

Koehn, Philipp, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *Proc. HLT-NAACL 2003,* pp. 127–133.

Och, Franz J. and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics,* Vol. 29. No. 1, pp. 19-51.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proc. ACL 2002,* pp. 311-318.

Stolcke, Andreas. 2002. SRILM - An Extensible Language Modeling Toolkit. *Proc. Intl. Conf. Spoken Language Processing,* pp. 901-904.

Utiyama, Masao and Hitoshi Isahara. 2007. A Japanese-English Patent Parallel Corpus. *Proc. Machine Translation Summit XI,* pp. 475-482.

Wu, Hua, Haifeng Wang and Chengqing Zong. 2008. Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Corpora. *Proc. COLING 2008,* pp. 993–1000.