用例翻訳ＳＤＭＴにおける対訳情報の利用

加藤　直人
ＮＨＫ放送技術研究所
katou.n-ga@nhk.or.jp

1　はじめに

機械翻訳では一般に、対訳情報など原言語側の制約をかけた後に、言語モデルなど目的言語側の制約をかけて解（翻訳結果）を得る。しかし、用例翻訳ＳＤＭＴでは、原言語側の制約である二言語間類似度の計算コストが高いため、先に目的言語側制約である言語モデルを適用し解候補を絞った上で、二言語間類似度を計算している。これにより計算コストは抑えられるものの、解候補を絞る段階で原言語側の情報を使っていないため、本来必要な訳語が解候補に含まれない可能性があるという問題を生じる。本稿では、言語モデルを適用する際に対訳情報を用いることによって、原言語側の情報を早期に反映させる方法について述べる。

２　用例翻訳ＳＤＭＴ

2.1 概要

用例翻訳ＳＤＭＴ（Similarity-Driven Machine Translation）では、入力に類似した対訳用例を組み合わせることで翻訳を行う。ここでは、日本語から英語に翻訳する場合を例に取り、簡単に説明する。詳細については文献[1]を参照されたい。

ＳＤＭＴは図１のように大きく分けて３つのステップからなる。STEP1では、入力文が与えられると、対訳用例コーパス内の各文と単言語内類似度を計算し、入力文のすべての単語を被覆するように類似用例を収集する。ここで、単言語類似度とは入力文と用例の原言語側との類似度であり、用例翻訳一般によく使われるものである。図１の例では、3つの用例が収集されている。STEP2では、収集された用例において目的言語側の単語をマルチプルアライメントにより整列し、単語グラフを作成する。この単語グラフが解候補を構成する。STEP3では、言語モデルと二言語間類似度という２つの制約を使って翻訳結果を得る。ここで、言語モデルにはn-gramモデルを使っている。一方、二言語間類似度は原言語と目的言語の間をつなぐ類似度であり、ＳＤＭＴの特徴である。ＳＤＭＴでは、原言語側での処理と目的言語側での処理が分かれており、機械翻訳一般で行われる言語変換処理がない。

2.2 目的言語の生成（STEP3）の問題点

ＳＤＭＴにおいて、STEP3の２つの制約を適用する際は、始めに言語モデルで解候補のＮベストを求め、次にその各候補に対して二言語間類似度を計算し、両者を考慮して最適解を求めている。すなわち、言語モデルである程度解候補を絞っている。これは、言語モデルから得られる全候補に対して二言語間類似度を計算すると、STEP3の計算量が膨大となるためである。

しかしながら、言語モデルを先に適用すると、今度は必要な訳語が解候補に入らない可能性があるという問題が生じる。図１の単語グラフを使って説明する。この単語グラフでは、“Mindanao island”、“the Philippines”、“Luzon Island”が並列に並んでいる。したがって、言語モデルを適用した際にはこれらのいずれか一つが選択される。どの単語が選択されるかが問題となるが、現状では、単にn-gramの値が大きい場合を優先しているので、“the Philippines”や“Luzon Island”のn-gramの値が大きいとこれらの単語を含むパスが選択されてしまい、“Mindanao island”を含むパスがＮベストに入らない可能性がある。この問題は、目的言語（英語）側の情報だけを利用していることに起因する。

そこで、n-gramを適用する際に、原言語（日本語）側からの情報、すなわち、対訳情報を用いることが考えられる。先の例では、目的言語（英語）側単語“Mindanao island”に対応する原言語（日本語）の単語が「ミンダナオ島」であるという対訳情報を用いると、“Mindanao island”は

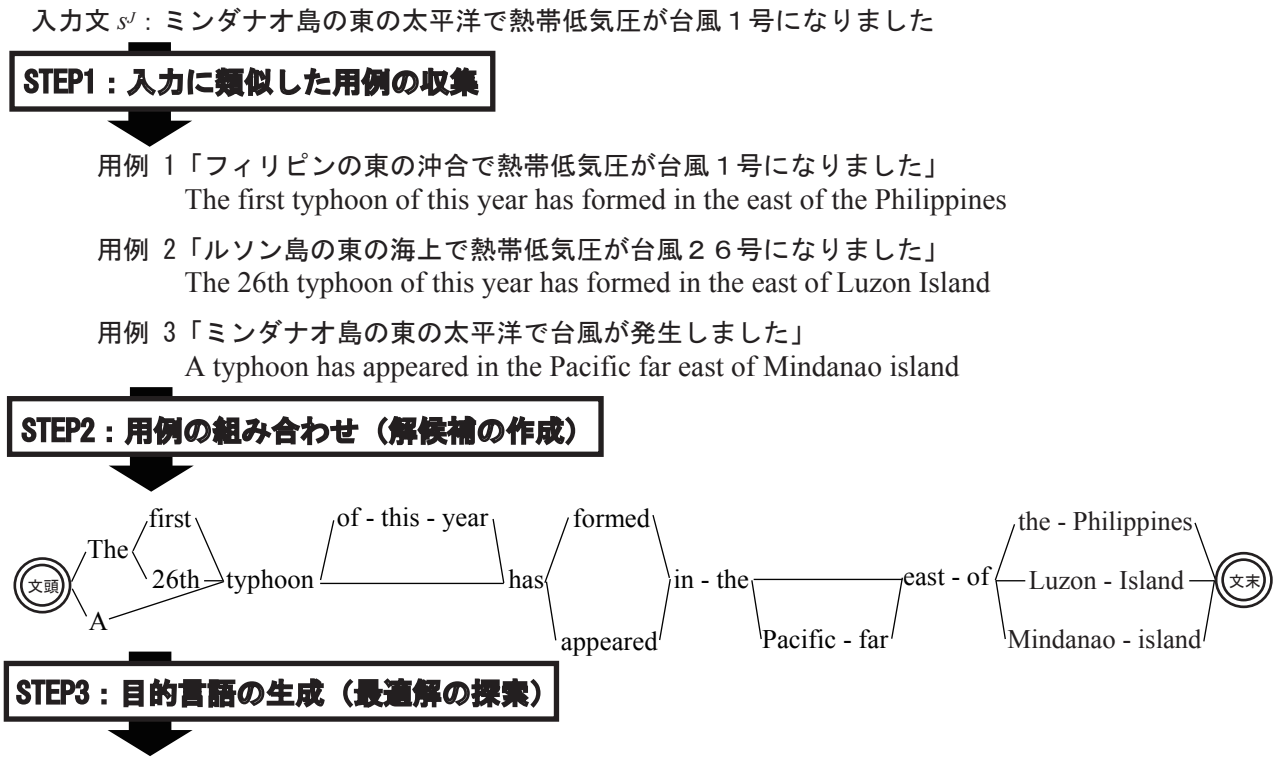


図１ 用例翻訳ＳＤＭＴの概要

必須な単語であることがわかる。さらに、“the Philippines”や“Luzon Island”の訳語が入力文には出現しないので、これらは不要な単語であることがわかる。

そこで、本稿では言語モデルを適用する際に二言語間の一つの制約として、対訳情報を利用する手法について述べる。提案手法では、n-gram モデルの段階で対訳情報を用いることにより、必須な単語ではその優先度を上げるとともに不要な単語では下げ、より適確な解を得られることが期待できる。

３　対訳情報の利用

3.1 定式化

提案手法では、n-gram モデルのほかに対訳情報も用いて、言語モデルによるスコアは次のように定式化した。

$$(1)\quad p(e_1 \cdots e_I) = \prod_{i=1}^{I} p(e_i \mid e_{i-n+1} \cdots e_{i-1}) \prod_{i=1}^{I} p'(e_i)$$

第二因子が対訳情報による追加項である。原言語側の入力文 s^J 中の単語 j_i に対応する訳語 e_i が単語グラフ中に存在するならば、$p'(e_i)$ を加点する。

実際の計算では対数値を使って、式 (2) のように定義している。

$$(2)\quad \log p'(e_i) = \begin{cases} q_1 & if\ j_i \in s^J \\ q_2 & if\ j_i \notin s^J \\ 0 & otherwise \end{cases}$$

式 (2) において、$\log p'(e_i) = q_1\ (>0)$ が加点である。同式ではさらに、j_i が入力文 s^J 中に存在しない場合には、$\log p'(e_i) = q_2\ (\leqq 0)$ と減点している。

今回は対訳情報として、固有名詞や数量表現などの固有表現のみに限定した。一般の対訳辞書を利用することも考えられるが、誤った対訳を使ってしまう可能性もある。そこで、気象災害ニュースに頻出し、誤った対訳を使う可能性が低い固有表現のみを利用した。さらに、対訳用例コーパスから自動アライメント [3] をして得られる対訳情報を利用することも考えられる .

固有名詞では、対訳辞書とともに、簡単な translitelation での処理も行っている。数量表現は、対訳用例コーパスを観察して作成した日英数量表現変換ルールで処理している。

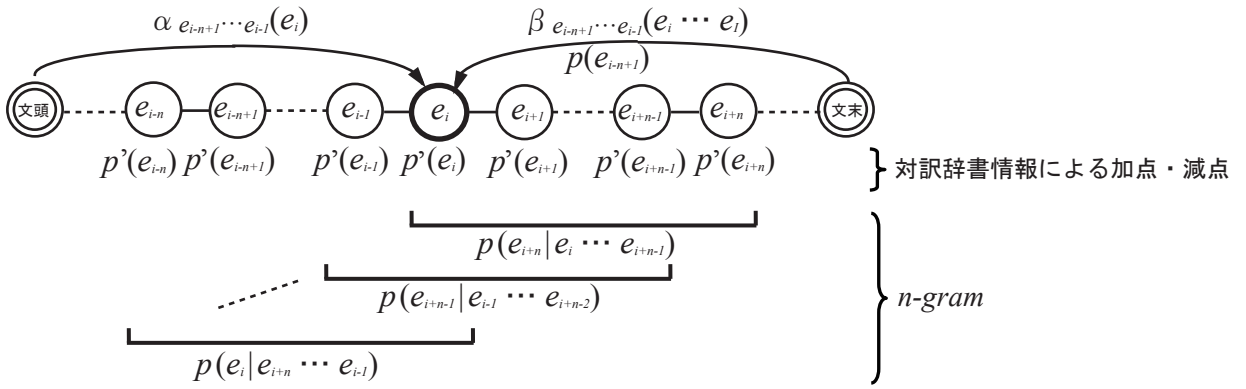


図 2 探索アルゴリズム

3.2 探索アルゴリズム

探索アルゴリズムは、対訳辞書情報による加点（減点）が扱えるように、前向き DP 後向き A*[2] を拡張した。図 2 を使って説明する。

図 2 において､各ノードは目的言語側の単語（英単語）を表す。ノードには、そのノードの優先する度合いを表すスコア $p'()$ を付与している。ある単語 e_i が翻訳結果に必須であるならば、そのスコア $p'(e_i)$ を高くすればよい。

前向き DP では、文頭ノードからノード e_i までのパスの中で、パス $e_{i-n+1}\cdots e_{i-1}$ を通る場合のスコア $\alpha_{e_{i-n+1}\cdots e_{i-1}}(e_i)$ **を求める。** $\alpha_{e_{i-n+1}\cdots e_{i-1}}(e_i)$ に対して式 (3) **が成り立つ。**

(3) $$\alpha_{e_{i-n+1}\cdots e_{i-1}}(e_i) = \max_{e_{i-n}}\left[p'(e_{i-1}) \times \alpha_{e_{i-n}\cdots e_{i-2}}(e_{i-1}) p(e_i \mid e_{i-n+1}\cdots e_{i-1}) \right]$$

式 (3) の中で、$p'(e_{i-1})$ が提案手法により加えられた項である。

一方、後向き A*では、ノード e_i から文末ノードまでのパス $e_i\cdots e_I$ の中で、パス $e_{i-n+1}\cdots e_{i-1}$ を通るスコア $\beta_{e_{i-n+1}\cdots e_{i-1}}(e_i\cdots e_I)$ を求める。$\beta_{e_{i-n+1}\cdots e_{i-1}}(e_i\cdots e_I)$ に対して式 (4) **が成り立つ。**

(4) $$\beta_{e_{i-n+1}\cdots e_{i-1}}(e_i\cdots e_I) = p'(e_i) \times \beta_{e_{i-n}\cdots e_{i-2}}(e_{i+1}\cdots e_I) \prod_{k=1}^{n} p(e_{i+k-1} \mid e_i \cdots e_{i+k-n-1})$$

式 (4) の中で、$p'(e_i)$ が提案手法により加えられた項である。

式 (3) と式 (4) を使って、あるパス（翻訳結果）$p(e_1\cdots e_I)$ のスコアは式 (5) で表される。

(5) $$p(e_1\cdots e_I) = \alpha_{e_{i-n+1}\cdots e_{i-1}}(e_i) \times \beta_{e_{i-n+1}\cdots e_{i-1}}(e_i\cdots e_I)$$

後向き A*では、式 (5) を計算し、スコアが高い上位 N 個を得ることによって、言語モデルによる解候補を求めることができる。

４　実験

対訳情報を利用してＳＤＭＴの日英翻訳実験を行った。対訳用例コーパスは、ＮＨＫの日本語ニュースの中から気象・災害に関する記事を選択し、翻訳者に英訳してもらった文から構成される。対訳用例は 24,429 文であり、評価文は 60 文である。最もよく使われている自動評価指標 BLEU[3] で評価した。BLEU で利用した正解訳は 1 文のみである。また、言語モデルには 5-gram を用い、その際の N ベストは N = 5,000 とした。

パラメータ q_1、q_2 **を変化させて実験を行った。**まず、それぞれのパラメータが単独で BLEU の改善に寄与するかを見るために、$q_2 = 0$ として q_1 を変化させた場合と、$q_1 = 0$ として q_2 **を変化させた**場合を行った。**図３に実験結果を示す。これを見**ると $q_2 = 0$ の場合には $q_1 = 5.0$ のとき BLEU = 0.2677 と最も高く、q_1 を大きくすると逆に値が低くなる傾向がある。一方、$q_1 = 0$ の場合には、q_2 を小さくすると BLEU は向上する傾向があり、$q_2 =$ -1000.0 のときに BLEU = 0.2634 と最も高く、それよりも小さい値ではほぼ一定となっている。

次に、前の実験で BLEU の最高値が得られたパラメータ $q_1 = 5.0$ と $q_2 = -1000.0$ に対して、今度はそれぞれパラメータ q_2、q_1 **を変化させて実験を行っ**

た。図 4 に結果を示す。これを見ると、それぞれ q_1 = 4.0、q_2 = -1000.0 のとき BLEU = 0.2758、q_1 = 5.0、q_2 = -500.0 のとき BLEU = 0.2752 と最高値をとっており、このあたりに最適なパラメータがあると推測される。ベースライン（q_1 = 0、q_2 = 0）のときは BLEU = 0.2467 であるので、いずれの場合も若干 BLEU 値が改善されている。

５　おわりに

用例翻訳ＳＤＭＴの改良の一つとして、対訳情報の利用について述べた。対訳情報は、n-gram 言語モデルにおいて単語に加点・減点することで利用している。実験の結果、BLEU の値が若干改善した。すなわち、対訳用例を利用した効果があったと考えられる。

今回は対訳情報として固有名詞に限定したが、今後はさらに一般の単語へ拡大していく。また、単語に限らず、フレーズ間の対訳情報も利用する予定である。現在は、言語モデルと二言語間類似度という２つの制約を別々に適用しているが、これらの制約を統合することも考えていきたい。

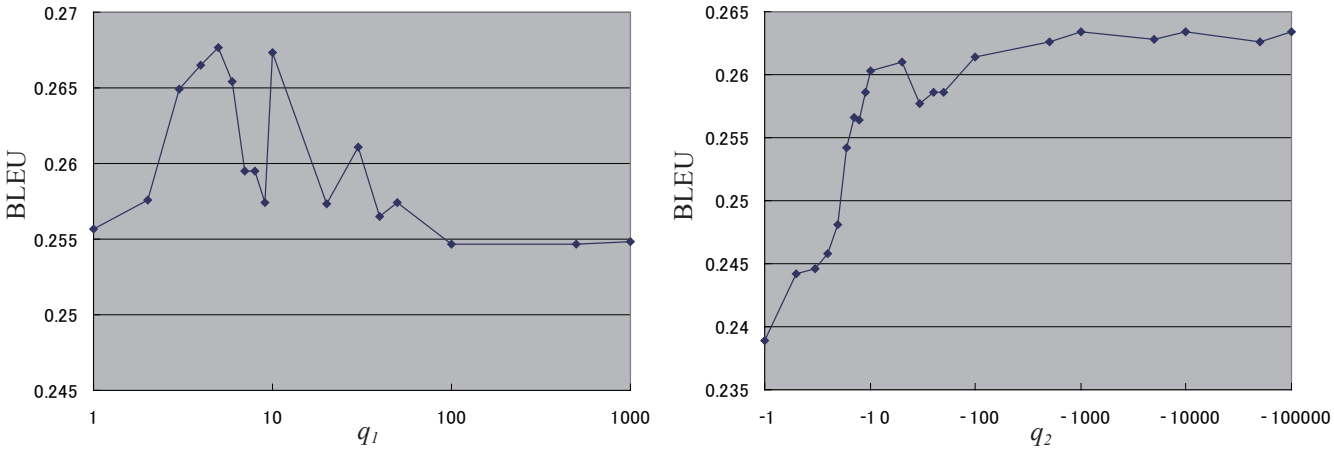
参考文献

[1] 加藤直人．ＳＤＭＴ：用例翻訳への新しいアプローチ．情報処理学会自然言語処理研究会，NL-170，pp. 151-156，2005.

[2] 永田昌明．統計的言語モデルと N-best 探索を用いた日本語形態素解析法．情報処理学会論文誌，Vol. 40，No. 9，pp. 3420-3431，1999.

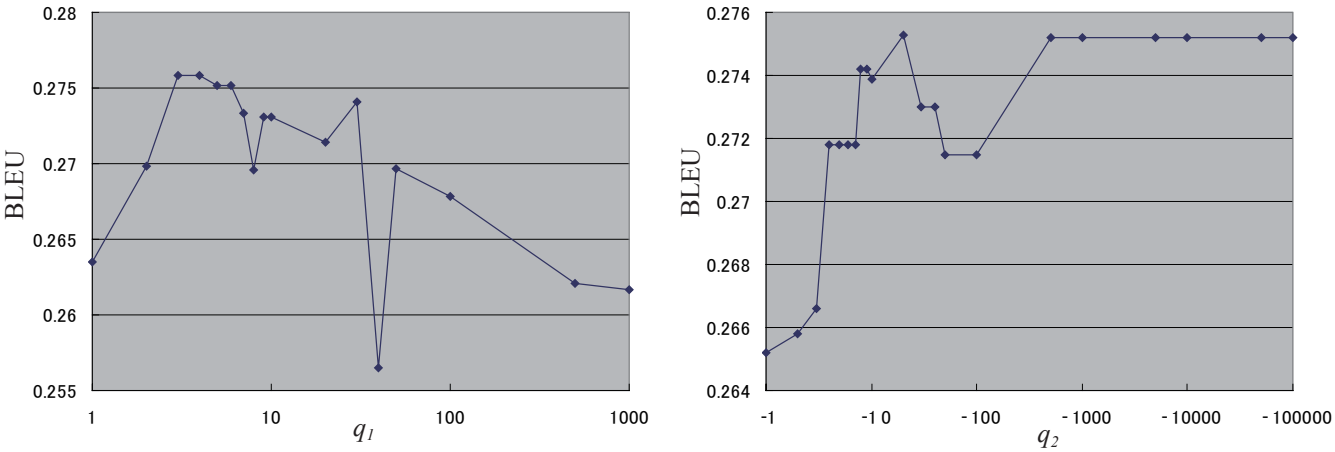
[3] 加藤直人．日本語・英語ニュースを対象とした文・単語の同時アライメント．言語処理学会第９回年次大会，pp. 887-890，2008.

[4] Kishore Papineni et al. BLEU: a Method for Automatic Evaluation of Machine Translation. Procs. of ACL2002, pp. 311-318, 2002.



(a) q_2 = 0.0 と固定し、q_1 を変化させたとき　　(b) q_1 = 0.0 と固定し、q_2 を変化させたとき

図３　片方のパラメータを０としたときの結果



(a) q_2 = 1000.0 と固定し、q_1 を変化させたとき　　(a) q_1 = 5.0 と固定し、q_2 を変化させたとき

図４　片方のパラメータを図３の最高値としたときの結果