文書文脈を用いた翻訳精度、速度の改善

大西 貴士　　内山 将夫　　隅田 英一郎
情報通信研究機構　MASTARプロジェクト　言語翻訳グループ

{takashi.onishi,mutiyama,eiichiro.sumita}@nict.go.jp

1　はじめに

フレーズベース統計翻訳は、語順の変化が小さい言語間では良い性能が得られているが、日英のように語順の変化が大きい言語間ではフレーズの並べ替えが失敗することにより、翻訳精度と速度が低下する問題がある。特に、特許文書の翻訳は1文が長いこともあり、並べ替えの失敗による精度低下が大きな問題となっている。このようなフレーズ並べ替え問題に対して、これまで様々な手法が提案されている。例えば、名詞句や節の単位で文を分割し、2段階で翻訳を行う手法 [1, 2] や、原言語側の構文情報を用いて並べ替えを制限する手法 [3] などがある。しかし、これらの先行研究で文書レベルの文脈情報を用いたものはなかった。

本研究では、文書レベルの文脈情報を用いてフレーズの並べ替えを制限する手法を提案する。文脈情報として、翻訳対象文が出現した文書（文脈文書）に特徴的に出現するフレーズ（特徴フレーズ）を利用する。提案手法では、あらかじめ文脈文書から特徴フレーズを抽出しておき、各文の翻訳時に特徴フレーズが1ブロックに翻訳されるように並べ替えを制限する。特徴フレーズを1ブロックとしてみなせるため、その分だけ仮想的に文長が短くなり翻訳精度と速度が向上する。

日英特許翻訳タスクで実験を行ったところ、提案手法はベースライン手法に対し、1.14BLEUポイントの有意な精度向上が得られた。また、翻訳速度も84%向上した。

2　特許文書の翻訳

本研究では、翻訳対象として特許文書を扱う。特許文書は、新出語が多く1文も長いという特徴があるため翻訳が難しい。なぜなら、特許文書は新しく発明したものについて説明した文書であるから、その中ではその特許にしか出現しないフレーズ、主に部材名、が多い。こうしたフレーズはトレーニングコーパスでは出現しないため、そのままの形ではフレーズ翻訳は得られない。そのため複数のフレーズを組み合わせて翻訳されることになり、1文の長いこともあいまってフレーズ並べ替えの組合せの爆発を招き、翻訳精度と速度が低下する。例えば、表1のベースラインの訳文では、フレーズ並べ替えが失敗しているため全体の翻訳が失敗している。特に、“第 1 の 絶縁 膜 で ある 層間 絶縁 膜 1　2”のフレーズが、1ブロックに翻訳されるべきところを *“an interlayer insulating film 12”* と *“a first insulating film”* の2つのブロックに別れてしまっている。

このような場合、特許の部材名のような名詞句は異なる言語間でも1つの名詞句として翻訳されると考えられるので、その名詞句が1ブロックに翻訳されるように並べ替えの制約を加える方法が考えられる。しかし、どの範囲が1ブロックに翻訳されるのかを決定することは自明ではない。名詞句は様々な修飾によってより大きな名詞句になることができる。特に特許文書では、入れ子になった巨大な名詞句が多く見られる。このとき、並べ替えの制約の範囲が大きすぎても、小さすぎても、制約としては弱いものになってしまうため、これらの名詞句の中からどの名詞句を制約として使うかを決定するのは難しい。したがって、制約の範囲としてどのような名詞句を選択するかが大変重要な問題であり、本論文では、その問題に対して、特許文書の特質を利用した解決策を提案する。

3　文脈を用いた並べ替え制約

以上のような問題を解決するため、我々は翻訳対象文が出現した文書（文脈文書）を用いて、フレーズ並べ替えの制約範囲を決定する手法を提案する。特許文書を翻訳する場合、まず、文脈文書であるその特許文書から特徴的に出現しているフレーズ（特徴フレーズ）を抽出する。ここで得られた特徴フレーズは、その特許の部材名であることが多い。部材名は目的言語でも1つの名詞句として翻訳されると考えられるので、特徴フレーズの範囲をフレーズ並べ替えの制約範囲として使えると考えた。そこで、特許文書の各文を翻訳する際に、重要フレーズが含まれていれば、その範囲の翻訳を1ブロックになるように並べ替えの制約を加えてデコードする。以下で提案手法の手順を説明する。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

原文	パッド電極 1 1 は 、 **第 1 の 絶縁 膜 で ある 層間 絶縁 膜 1 2** を 介し て 半導体 基板 1 0 の 表面 に 形成 され て いる 。
参照訳	the pad electrode 11 is formed on the top surface of the semiconductor substrate 10 through **an interlayer insulation film 12 that is a first insulation film** .
訳文（ベースライン）	**an interlayer insulating film 12** is formed on the surface of a semiconductor substrate 10 , a pad electrode 11 via **a first insulating film** .
原文 + 並べ替え制約	パッド電極 1 1 は 、 **<zone> 第 1 の 絶縁 膜 で ある 層間 絶縁 膜 1 2 </zone>** を 介し て 半導体 基板 1 0 の 表面 に 形成 され て いる 。
訳文（提案手法）	pad electrode 11 is formed on the surface of the semiconductor substrate 10 through **the interlayer insulating film 12 of the first insulating film** .

表 1: 翻訳例

3.1 特徴フレーズ候補の抽出

まず、文脈文書から特徴フレーズの候補となるフレーズを抽出する。最も単純な方法としては、全ての部分フレーズを特徴フレーズの候補とすることが考えられるが、より精度のよい特徴フレーズを獲得するため構文解析を行い、文脈文書中のすべての名詞句を候補として抽出する。さらに、フレーズテーブルに含まれるフレーズは候補から除外する。これは、フレーズテーブルに含まれるフレーズは 1 フレーズとして翻訳されると考え、このフレーズの範囲を並べ替えの制約範囲とする必要がないと考えたからである。

3.2 C-value によるランキング

上で抽出された名詞句は様々な長さで、ネストした関係になっている。そのためこれらの中から特徴フレーズとして用語性の高いものを選ぶ必要がある。そこで、我々は C-value[4] を利用し、C-value が閾値（TH）以上のものを特徴フレーズとして並べ替えの制約に用いた。C-value は、自動用語抽出のための尺度で、以下の式で表される。

$$C\text{-}value(p) = \begin{cases} (l(p)-1)\, n(p) & (c(p)=0) \\ (l(p)-1)\left(n(p) - \frac{t(p)}{c(p)}\right) & (c(p)>0) \end{cases}$$

ここで、$l(p)$ はフレーズ p の長さ、$n(p)$ は文書中の p の出現回数、$t(p)$ は p を部分として含むフレーズの出現回数、$c(p)$ はそのようなフレーズの異なり数である。C-value が大きな名詞句は、文脈文書中で頻出するフレーズであるためその特許での意味のあるまとまりであると考えられ、1 ブロックに翻訳されると考えられる。

3.3 並べ替え制約を加えたデコード

各文の翻訳の際には、C-value の大きい順に重要フレーズを調べていき、文中に重要フレーズがあれば、その範囲を並べ替え制約として採用する。デコーダは Moses[5] を用いた。Moses は、<zone> タグで指定した範囲を 1 ブロックとして翻訳する機能がある [6]。そこで、表 1 のように重要フレーズを <zone> タグで囲み、Moses に入力する。これによって、重要フレーズが 1 ブロックに翻訳され、正しい翻訳結果が得られる。

Set	# sentences	# words
Training	1.8M	J 70M E 59M
Dev	2000	J 79K E 69K
Test	1251	J 49K E 42K

表 2: 実験に用いたデータセット

4 実験

提案手法の有効性を評価するために NTCIR-8 特許翻訳タスク [7] のデータセットを用いて実験を行った。データセットの詳細は表 2 のようになっている。Dev セットと Test セットは、各文が出現した特許明細書も一緒に配布されており、提案手法の文脈文書は、この特許明細書を用いた。

4.1 ベースライン手法

ベースライン手法として、Moses[5] を用い、以下のような標準的な設定を用いた。

- grow-diag-final-and
- 5-gram language model
- msd-bidirectional-fe reordering model
- distortion-limit = -1 (unlimited)

パラメータチューニングは、Dev セットを用いて行い、BLEU 値が最大になるように Minimum Error Rate Training[8] を行った。

4.2 提案手法

重要フレーズ候補が違う 4 通りの手法を比較した。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

System	BLEU(%)	Gain(%)	TH
Baseline	27.75	–	–
ALL	28.06	+0.31 *	70
ALL-PT	28.50	+0.75 **	50
NP	28.70	+0.95 **	7
NP-PT	**28.89**	**+1.14** **	3

表 3: 実験結果

ALL

重要フレーズ候補として、文脈文書中の 15 語以下のすべての部分単語列を用いたもの。構文解析やフレーズテーブルによるフィルタリングは行わない。

ALL-PT

ALL の重要フレーズ候補のうち、フレーズテーブルにフレーズ翻訳が登録されていないフレーズのみを重要フレーズ候補として用いたもの。

NP

文脈文書を構文解析し、長さが 15 語以下でヘッドが名詞相当となる連続部分木を重要フレーズ候補として用いたもの。構文解析器は、CaboCha[9] を用いた。

NP-PT

NP の重要フレーズ候補のうち、フレーズテーブルにフレーズ翻訳が登録されていないフレーズのみを重要フレーズ候補として用いたもの。

ここで、重要フレーズ候補の長さを 15 語以下に限定したのは、候補の数を抑えるためと長すぎることにより並べ替えの制約が弱くなることを防ぐためである。C-value の閾値（TH）は、1〜100 の範囲でいくつかの値で実験を行い、Dev セットの BLEU 値が最も大きくなる TH の値を使用した。デコード時のパラメータはベースライン手法のものと同じ値を使用した。

4.3 結果と考察

評価は、Test セットの BLEU スコアで行った。結果は、表 3 の通りである。提案手法は、4 つすべての方法でベースラインを上回る精度を示した。最も精度が良かったのは NP-PT で、1.14BLEU ポイントの向上が見られた。また、ブートストラップ法 [10] で検定を行ったところ、ベースラインと比較して、ALL-PT、NP、NP-PT が $p < 0.01$、ALL が $p < 0.05$ で有意に優れていることが分かった。このことから、文脈文書を用いて並べ替え制約の範囲を決定する方法が有効であるといえる。

4.3.1 構文情報による効果

NP-PT と ALL-PT、NP と ALL のどちらも有意（$p < 0.01$）な差がついており、構文情報を並べ替え制約に使うことが有効であることを示している。また、図 1 は、C-value の閾値（TH）と各手法の BLEU 値をプロットしたものである。これによると、NP や NP-PT は TH が小さいほうが精度がよくなっており、一方、ALL や ALL-PT は TH が大きいほうが精度が良い。このことから、構文情報を使用しない場合でも C-value を用いた足きりによって精度向上に有効な並べ替え制約を設定できることを示している。

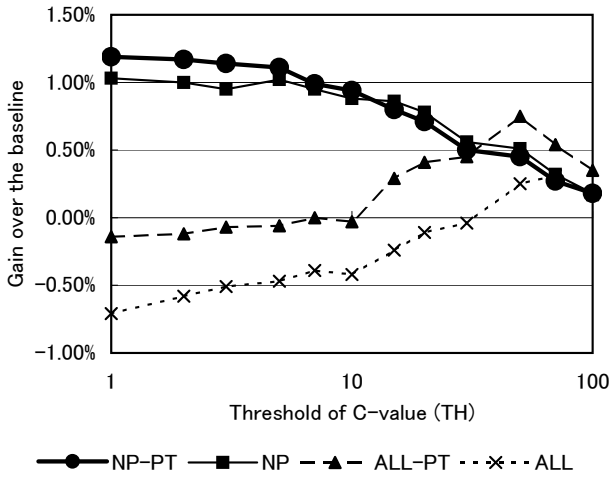

図 1: C-value の閾値（TH）と精度向上

4.3.2 フレーズテーブルによる効果

NP-PT と NP の差は有意ではなかったが、ALL-PT と ALL では有意（$p < 0.01$）な差がついた。図 1 を見ても NP-PT と NP の差は ALL-PT と ALL より小さい。NP-PT の場合は、構文情報を用いることで十分質の良い重要フレーズ候補が得られているためフレーズテーブルを用いたフィルタリングがあまり効かなかったと考えられる。一方、ALL-PT の場合は、重要フレーズ候補の質が良くないためフィルタリングの効果が大きくなったと考えられる。

4.3.3 制約 1 つあたりの精度向上

表 4 は、NP-PT（TH=3）について、1 文中の <zone> の数ごとのゲインを計算したものである。<zone> が 0 の場合は、並べ替え制約がない場合なのでベースラインと結果は同じになる。制約がある場合のベースラインとの差は、制約の数が多いほど大きくなり、平均して制約 1 個あたり約 0.4%のゲインがあった。

4.3.4 文短縮率と速度向上

提案手法は、重要フレーズの範囲を 1 ブロックとして扱い、制約の範囲を超える並べ替えを制限される。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

# zones	# sentences	Baseline	NP-PT(TH=3)	Gain
0	75	33.33	33.33	+0.00
1	349	29.78	29.94	+0.16
2	344	28.05	28.96	+0.91
3	239	27.21	28.53	+1.32
4+	244	25.87	27.90	+2.03

表 4: 並べ替え制約（<zone>）の数と翻訳精度

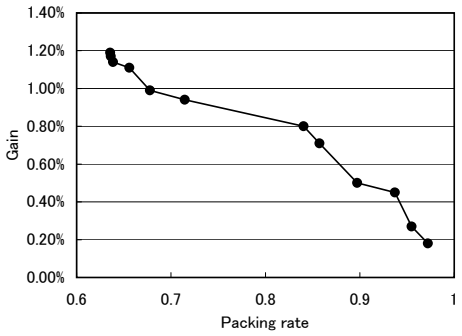

図 2: 短縮率と精度向上（NP-PT）

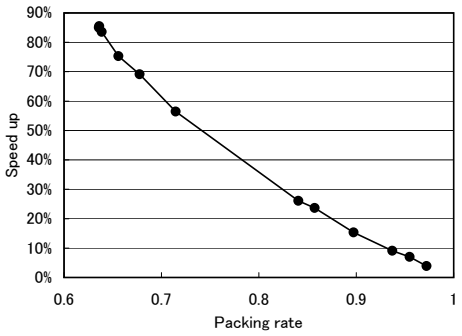

図 3: 短縮率と速度向上（NP-PT）

これは、文全体的な並べ替えで考えると、重要フレーズを仮想的に 1 つの単語として扱っていることに相当する。そのため、考慮すべき探索範囲か減るため翻訳にかかる時間は短くなる。図 2、3 は、NP-PT において、重要フレーズを 1 単語として数えた場合の文長の短縮率と精度向上、速度向上の関係をプロットしたものである。TH=3 の場合は、短縮率が 0.64、速度向上が 84% だった。これらをみると、短縮率が小さくなるにしたがって、翻訳精度と速度が向上していることが分かる。

5 おわりに

本研究では、フレーズベース統計翻訳において、文書レベルの文脈情報を用いてフレーズ並べ替えを制限する手法を提案した。提案手法では、文脈文書に特徴的に出現する特徴フレーズを利用し、特徴フレーズが 1 ブロックに翻訳されるように並べ替えを制限する。日英特許翻訳タスクで実験を行ったところ、提案手法はベースライン手法に対し、1.14BLEU ポイントの有意な精度向上が得られ、翻訳速度も 84%向上した。提案手法は基本的には言語に依存しない手法であるので、今後は、日英以外の言語ペアでも評価を行いたい。

参考文献

[1] P. Koehn and K. Knight: “Feature-rich statistical translation of noun phrases”, Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2003).

[2] K. Sudoh, K. Duh, H. Tsukada, T. Hirao and M. Nagata: “Divide and translate: Improving long distance reordering in statistical machine translation”, Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 418–427 (2010).

[3] C. Cherry: “Cohesive phrase-based decoding for statistical machine translation”, Proceedings of ACL-08: HLT, pp. 72–80 (2008).

[4] K. T. Frantzi and S. Ananiadou: “Extracting nested collocations”, Proceedings of COLING 1996, pp. 41–46 (1996).

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst: “Moses: Open source toolkit for statistical machine translation”, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177–180 (2007).

[6] P. Koehn and B. Haddow: “Edinburgh’s submission to all tracks of the WMT 2009 shared task with reordering and speed improvements to Moses”, Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 160–164 (2009).

[7] A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, T. Ehara, H. Echizen-ya and S. Shimohata: “Overview of the patent translation task at the ntcir-8 workshop”, Proceedings of NTCIR-8 Workshop Meeting, pp. 371–376 (2010).

[8] F. J. Och: “Minimum error rate training in statistical machine translation”, Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 160–167 (2003).

[9] T. Kudo and Y. Matsumoto: “Japanese dependency analysis using cascaded chunking”, Proceedings of CoNLL-2002, pp. 63–69 (2002).

[10] P. Koehn: “Statistical significance tests for machine translation evaluation”, Proceedings of EMNLP 2004, pp. 388–395 (2004).

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.