Factored Translation Modelを用いた日英間統計的機械翻訳の調査

小田貴博　　秋葉友良
豊橋技術科学大学

1　はじめに

近年, 統計的機械翻訳における翻訳モデルにおいて, Factored Translation Model を用いる手法が提案された. この手法は単語を表層形, 原形, 品詞など複数の因子 (factor) に分解し, 因子の間, または因子の組み合わせの間で翻訳モデル郡を学習して, それらを利用する手法である. 本研究では Factored Translation Model において, 特に以下の事柄に焦点を当てる.

- Factored Translation Model を日英間翻訳に適用した場合の有効性
- 学習する前の段階での Factored Translation Model の有効性の推定方法

Factored Translation Model を利用する手法は種々の言語対に適用され有効性が示されているが,[1] 日本語の翻訳に適用した例は少ない. 本研究では Factored Translation Model を英日間の翻訳に適用し, 最も基本的な因子として考えられる日英それぞれ 3 種類の因子について, 有効な因子の組み合わせの調査を行った.

Factored Translation Model の学習において, どの因子を組み合わせるかによって翻訳性能は大きく変化する. 因子の組み合わせによって多くの種類のモデル群を考えることが出来るので全てを学習することは難しい. そこで本研究においては翻訳モデル群を学習する前の段階で, 因子の組み合わせによって翻訳モデル群がどの程度有効であるかを, 推定する方法の検討を行った.

2　関連研究

Kohen[1] らは英語とドイツ語, スペイン語, チェコ語, 中国語の間で Factored Translation Model を学習し,BLEU スコアが改善したことを示した. また, 今回実験を行わなかった生成モデルや複数の複数の翻訳モデルを使った翻訳, 生成モデルを学習する翻訳も行い, 性能の改善を確認した.

3　Factored Translation Model

本節では Factored Translation Model の学習と翻訳手法について説明する.

3.1　Factored Translation Model の学習

Factored Translation Model を学習するステップは 3 段階に分ける事が出来る.

1. 学習データの factor 化
2. 翻訳モデルの学習
3. 生成モデルの学習

翻訳モデルを学習する前に学習データの単語から表層形, 原形, 品詞なと複数の因子を得る. この操作をここでは単語の factor 化と呼ぶ. 単語を factor 化することで得た複数の因子を学習データとして用いる. 以下に「しなければ」を factor 化した場合の例を示す.

単語　→　表層 | 原形 | 品詞

し　→　し | する | 動詞
なけれ　→　なけれ | ない | 助動詞
ば　→　ば | ば | 助詞

翻訳モデルは入力側の因子からターゲット側の因子を出力するモデルである. 翻訳モデルを学習する際の学習方法は単語のみで学習するフレーズベース翻訳モデルと同様である. 単語のみで学習するフレーズベース翻訳モデルは単語列をフレーズとする一方で, Factored Translation Model における翻訳モデルは因子またはその組み合わせの列をフレーズとする点が異なる.

生成モデルはターゲット側の因子から別のターゲット側の因子を生成するモデルである. 翻訳モデルが表層以外の因子だけを出力する翻訳モデル

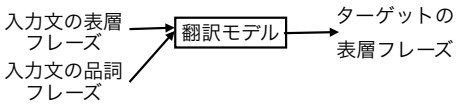

図 1: 入力側をファクタ化した場合の翻訳モデル

図 2: 出力側をファクタ化した場合の翻訳モデル

の場合には必要となるが, 今回調査した翻訳モデルの全ては表層を出力するので生成モデルの学習は行わなかった.

3.2 Factored Translation Model を用いた翻訳

Factored Translatin Model を用いた翻訳において, 学習の際の入力側と出力側の因子の組み合わせによって翻訳する場合の操作が異なる. 図 1 は入力側の因子に表層以外を考える例であり, 翻訳モデルに factor 化された文を入力する必要がある. 図 2 は出力側に表層以外の因子を考える例であり, 入力文は単語列である. 図 2 の場合には原形, 品詞の言語モデルを使って翻訳候補を評価する事が出来る. また, 入力側と出力側両方で表層以外の因子を考えることも出来る.

4 性能調査実験

本節では, ベースラインの単語の factor 化を行わない従来のフレーズベース翻訳モデルと Factored Translation Model の性能比較実験について述べる. また, 実験は次のような 3 種類の Factored Trasnlation Model について行った.

- 入力側を factor 化した翻訳モデル
- 出力側を factor 化した翻訳モデル
- 入力側と出力側を factor 化した翻訳モデル

入力側と出力側の factor を組み合わせを考える際に, それぞれの因子が持っている情報量を考慮して, 出力側の情報量よりも入力側の情報量が極端に少ないモデルについては学習しなかった. 例えば, 入力の因子が品詞で出力側が表層形とった組み合わせは入力側の情報量が出力側の情報量に比べてあまりにも小さくなりすぎてしまうので, 有効なモデルを学習出来ないと考えた.

4.1 実験条件

翻訳モデルの学習と翻訳に Moses[3], 言語モデルの学習に SRILM[4], 日本語側の tokenize と factor 化には mecab[5], 英語側の factor 化には tree-tagger[6], 英語側の tokenize と lowercase には wmt09[7] で用いられるツールを使用した. 対象な単語アライメントを求めるヒューリスティックには grow-diag-final-and を用いて,reordering モデルの学習には msd-bidirectional-fe を用いた. 表層形と原形の言語モデルには, 5-gram を Kneser-Ney smoothing でスムージングしたものを使用した. 品詞の言語モデルには, 5gram を Witten-Bell smoothing でスムージングしたものを使用した.

学習データ,開発セット,テストセットは NTCIR-7 特許翻訳タスク [2] のデータを用いた. 学習データには PSD 1798571 文のうち, 日英の文ペアが共に 120 単語以下になっている文ペア 1686320 文のみを使用した. 開発セットには最初の 100 文を用いてテストセットには同タスクの formal-run テストデータ 1381 文を使用した.

4.2 実験結果

実験は翻訳モデルが一つの場合について行い, 翻訳性能を NTCIR-7 の BLEU キットを利用して評価した. 入力側を factor 化した場合の翻訳性能を表 1 に示す. 入力側を factor 化した場合の翻訳性能を表 2 に示す. 入力側と出力側の両方を factor 化した場合の翻訳性能を表 4 に示す. 英日方向の翻訳における性能は factor 化を行うことでベースラインよりも性能が改善しているが, 日英方向の翻訳に関しては性能の改善をすることが出来なかった.

5 分析

今回実験を行った中で英日方向については入力側 factor を表層, 品詞として, 出力側 factor を表層原形とした場合に最も BLEU が高くなった. また, 日英方向はベースラインの BLEU が最も高かった. 表 1 において factor が表層, 品詞であるときに, 両方向については BLEU が比較的高くなっているが, 表 2 においては表層, 品詞の場合には英日方向の

表 1: 入力側を factor 化した場合の BLEU

factor	英日	日英
ベースライン	19.30	24.37
表層, 原形	20.28	22.12
表層, 品詞	21.55	24.00
表層, 原形, 品詞	20.37	21.04

表 2: 出力側を factor 化した場合の BLEU

factor	英日	日英
ベースライン	19.30	24.37
表層, 原形	20.11	20.80
表層, 品詞	18.35	22.49
表層, 原形, 品詞	21.63	21.51

BLEU が低くなっている．これは日本語の言語モデルの性能が原因であると考えられる．表 3 から日本語側の品詞の種類は 14 種類である．日本語品詞の言語モデルの perplexity が大きくなり, 英日方向の出力 factor が表層, 品詞の場合には品詞の言語モデルが悪影響を与えていることが考えられる．また表 4 の英日方向では入力側を表層, 品詞とした場合の 3 種類のモデルと出力側の factor が同じになっているモデルの BLEU を比較すると, 例外が一つあるのを除けば高くなっている．表 1 においても同様な傾向が見られ, 英語の品詞を入力 factor に入れることが有効であると考えられる．

6 有効な factor の組み合わせの推定

Factored Translation Model を用いた翻訳は多数の組み合わせ方が考えられる一方で, 有効な組み合わせはその中の一部である．学習する前の段階で有効な factor の組み合わせを推定する方法が必要になり, 本節では方法の検討について説明する．

同じ単語に対応づけられる factor の頻度は英語と日本語側で同程度であると仮定して, factor の頻度で並び替えた頻度グラフをその factor を表す関数とした．関数を式 (1) に示す．

$$f(i) = \mathrm{freq}(factor_i)$$
$$factor_i : \text{頻度で並び替えた } i \text{ 番目の } factor \tag{1}$$

関数同士の相関係数を求め, 相関係数が高いほど良い組み合わせであると推定した．二つの関数を f_i と g_i とすると相関係数は式 (2) で表される．

$$\frac{\sum_i (f_i - \hat{f}_i)(g_i - \hat{g}_i)}{\sqrt{\sum_i (f_i - \hat{f}_i)^2} \cdot \sqrt{\sum_i (g_i - \hat{g}_i)^2}} \tag{2}$$

表 6 に相関係数を示す．相関係数の値が太字になっている項は相関係数の値が上位 5 位だった factor の組み合わせである．相関係数,BLEU スコアで並び替えた上位 5 件の factor の組み合わせを表 5 に示す．英日方向に関してはあまり推定出来ていないが, 日英方向については 5 件のうち 3 件は推定した組み合わせと同じものが含まれていた．

表 3: 学習データに存在する factor ユニグラムの種類

factor	英語	日本語
表層, 原形, 品詞	217683	138259
表層, 品詞	217682	138017
表層, 原形	179992	137765
表層	178454	136566
品詞	21204	48089
原形	46	14

表 4: 入力側と出力側を factor 化した場合の BLEU

入力側 factor	出力側 factor	英日	日英
ベースライン		19.30	24.37
表層, 原形	表層, 原形	19.66	19.25
表層, 原形	表層, 品詞	18.40	17.08
表層, 原形	表層, 原形, 品詞	19.21	17.08
表層, 品詞	表層, 原形	21.85	22.01
表層, 品詞	表層, 品詞	20.34	21.57
表層, 品詞	表層, 原形, 品詞	19.82	21.48
表層, 原形, 品詞	表層, 原形	19.24	18.87
表層, 原形, 品詞	表層, 品詞	18.66	18.24
表層, 原形, 品詞	表層, 原形, 品詞	20.10	16.18

7 おわりに

本論文では日本語と英語の間での Factored Translation を使った翻訳モデルの性能の調査と学習前

表 5: 相関係数,BLEU で並び替えた factor の組み合わせ

相関係数		英日 BLEU		日英 BLEU	
英語側	日本語側	英語側	日本語側	英語側	日本語側
表層	表層, 原形, 品詞	表層, 品詞	表層, 原形	表層	表層
表層	表層, 品詞	表層	表層, 原形, 品詞	表層	表層, 品詞
表層	表層, 原形	表層, 品詞	表層	表層, 品詞	表層
表層, 原形	表層, 原形, 品詞	表層, 原形, 品詞	表層	表層	表層, 原形
表層, 原形	表層, 品詞	表層	表層, 原形	表層	表層, 原形, 品詞

の段階での翻訳モデルの性能を推定する方法の検討を行った. 今回提案した推定方法は翻訳方向に無関係の値をで評価するので, 今後の課題として翻訳方向にも関係する値で評価する手法を検討する予定である. また, 今回実験を行わなかった複数の翻訳モデルや生成モデルを使った翻訳の性能の評価も行っていきたい.

表 6: factor 間の相関係数

英語 factor	日本語 factor	相関係数
表層	表層	0.003377
表層	表層, 原形	**0.003395**
表層	表層, 品詞	**0.003397**
表層	表層, 原形, 品詞	**0.003402**
表層, 原形	表層	0.003355
表層, 原形	表層, 原形	0.003374
表層, 原形	表層, 品詞	**0.003376**
表層, 原形	表層, 品詞, 原形	**0.003381**
表層, 品詞	表層	0.002970
表層, 品詞	表層, 原形	0.002986
表層, 品詞	表層, 品詞	0.002986
表層, 品詞	表層, 原形, 品詞	0.002990
表層, 原形, 品詞	表層	0.002969
表層, 原形, 品詞	表層, 原形	0.002986
表層, 原形, 品詞	表層, 品詞	0.002986
表層, 原形, 品詞	表層, 原形, 品詞	0.002990

参考文献

[1] P.Koehn and H.Hoang, "Factored Translation Models", In Proceeding of Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868-876, 2007.

[2] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, "Overview of the Patent Translation Task at the NTCIR-7 Workshop", In Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

[3] 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", In Proceeding of ACL 2007 demonstration session, pp.177-180, 2007. http://www.statmt.org/moses/

[4] A. Stolcke, "SRILM - an extensible language modeling toolkit", ICSLP, pp.901-904, 2002.

[5] Taku Kudo, "MeCab: Yet Another Part-of-Speech and Morphological Analyzer", http://mecab.sourceforge.net/

[6] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees", InProceeding of ECML-98, pp.25-36, 1998 .

[7] "EACL 2009 FOURTH WORKSHOP ON STATISTICAL MACHINE TRANSLATION", http://www.statmt.org/wmt09/