統語ラベルクラスタリングを用いた構文拡張機械翻訳

美野 秀弥　　渡辺 太郎　　隅田　英一郎

情報通信研究機構 ユニバーサルコミュニケーション研究所

{hideya.mino , taro.watanabe , eiichiro.sumita}@nict.go.jp

1　はじめに

近年、構文木の木構造や統語ラベルなどの統語情報を用いた統計的機械翻訳が盛んに行われている。Zollmann ら (2006) は構文木から得られる統語ラベルを用いた機械翻訳システム、構文拡張機械翻訳 (SAMT) を提案している。また、須藤ら (2013) や Wang ら (2010) は翻訳の精度を上げるために、言語学的知見や詳細な言語素性を用いて統語ラベルの細分化を行っている。統語情報を利用することで詳細な翻訳規則の記述が可能となるため、統語的に正しい翻訳を行うことができると期待できる。しかし、統語ラベルの過度な増加は処理時間の増加やデータスパースネスの問題を伴う。Hanneman ら (2013) はこの問題に対応するために、増加した統語ラベルの情報を持った同期文法の翻訳規則を用いて、翻訳の精度を下げることなく増加した統語ラベルを粗くする手法を提案している。そして、中国語から英語への翻訳において、両言語の構文情報を用いた SAMT に適用し、統語ラベル数が減少しているにも関わらず翻訳の質が向上したと報告している。

本稿では、Hanneman らのアプローチを参考にして、統語ラベルの尤度が近いものをクラスタリングするのではなく、同期文法の翻訳規則の尤度を最大化するように統語ラベルのクラスタリングを行う手法を提案する。また、交換アルゴリズム (Martin et al., 1998) を用いることで高速なクラスタリングを実現する。そして、目的言語の統語ラベルを用いた SAMT を利用して評価実験を行い、その効果を確認する。

2　構文拡張機械翻訳

構文拡張機械翻訳 (SAMT) は、構文解析で獲得した統語ラベルを非終端記号として用いて翻訳を行う手法である。文法を仮定せずに句単位で原言語から目的言語の変換と並び替えをして翻訳を行うフレーズベース翻訳 (Koehn et al., 2003) や、2 種類の非終端記号を用いて記述された翻訳規則を利用して翻訳を行う階層的フレーズベース翻訳 (Chiang, 2007) と比較すると、SAMT は、詳細に分類した非終端記号を用いて翻訳規則を記述できるため、並び替えの精度が高くなることが期待できる。Zollmann ら (2006) は、構文解析で獲得した統語ラベルと、組み合わせ範疇文法 (CCG) に基づいた次の 3 種類の細分化した統語ラベルを非終端記号として用い、SAMT の精度を高めている。

X/Y ： ラベル X のうち右端のラベル Y を除く
$X\backslash Y$ ： ラベル X のうち左端のラベル Y を除く
$X+Y$ ： 隣接したラベル X,Y を合わせる

本稿では、Moses[1] で実装されている統語ラベルの細分化 (SAMT1,2,4) を用いる。また、構文解析で付与される統語ラベルのみを利用した SAMT0 と、SAMT2 を拡張した SAMT5 を加える。数字が大きくになるにつれて細分化の条件が緩和され、ラベル数は増加する。SAMT5 で細分化した統語ラベルの例を表 1 に示す。

SAMT0 ： 構文解析で付与された統語ラベル
SAMT1 ： 隣接し、かつ同じ親を持つ場合に X+Y を加える (例 DT+NN)。子ノードの右端、或は左端のラベルを除いた X/Y,X\Y を加える (例 S\DT)。
SAMT2 ： SAMT1 に加え、同じ親を持たない場合でも X+Y を加える (例 NN+VBD)。
SAMT4 ： SAMT2 に加え、ラベルが付与されないノードにラベル FAIL を与える。
SAMT5 ： SAMT2 において 3 ノード間の細分化を認める (例 DT+NN+VP)。ラベルが付与されないノードにラベル FAIL を与える。

[1] http://www.statmt.org/moses/

表 1: 細分化した統語ラベルの例 (SAMT5)

S, TOP, NPB+VP, DT+NN+VP			
	NN+VP,NN+VBD+ADJP, S\DT		
NPB+VBD, S/ADJP			
	NN+VBD		
NPB, DT+NN, S/VP		VP, VBD+ADJP	
DT,NPB/NN	NN,NPB\DT	VBD	JJ,ADJP
the	**light**	**was**	**red**

一方、統語ラベルの細分化は翻訳処理時間の増加やデータスパースネスの問題を引き起こす。SAMT は翻訳規則で翻訳を行う際に CKY+アルゴリズムを用いており、翻訳時の最悪計算時間は翻訳対象文のトークン数を N、翻訳規則数を G とすると $O(N^3G)$ である。統語ラベルを細分化すると翻訳規則数が増えるため、1 文の翻訳処理にかかる時間は増加する。

SAMT の翻訳精度を落とさずに細分化した統語ラベル数を少なくする手法に統語ラベルのクラスタリングが挙げられる。Hanneman ら (2013) は、原言語側と目的言語側の統語ラベルを用いた SAMT で獲得した翻訳規則を用い、次の手順でクラスタリングを行う手法を提案している。

1. 構文解析結果を利用して統語ラベルを細分化する
2. 細分化した統語ラベルを用いて翻訳規則を学習する
3. 学習で獲得した翻訳規則を用いて統語ラベルのクラスタリングを行う
4. クラスタリングの結果を用いて再度翻訳規則を学習する

クラスタリングには、翻訳規則内の原言語側の統語ラベルの頻度分布を利用し、分布の距離が最も近い目的言語側の統語ラベル同士をクラスタ化する凝集型クラスタリングを用いている。そのため、クラスタリングの計算量は、統語ラベル数を T とすると $O(T^2logT)$ となり、統語ラベルの増加により計算量も大きく増加する。

3 統語ラベルクラスタリング

2 節で述べたように、細分化により増加した統語ラベルをそのまま利用して SAMT を行うと、翻訳処理時の計算量の大幅な増加やデータスパースネスの問題が発生することがある。また、Hanneman ら (2013) の手法では統語ラベル数が大きいとクラスタリングの計算量が大きく増加する問題がある。

そこで、本稿では、計算量を抑えることができる Exchange Clustering(Martin et al., 1998; Uszkoreit and Brants, 2008) を利用して統語ラベルクラスタリングを行う。Hanneman らは凝集型クラスタリングを用いているため、全ての要素の組み合わせについて計算する必要があり計算量が増加する傾向にあるが、Exchange Clustering は、まず全分類対象を適当なクラスタに割り振り、目的関数が最大になるようにクラスタ間で要素を移動させてクラスタリングを行うため、全ての組み合わせを計算する必要がない。

以下、Exchange Clustering に必要な目的関数について説明した後、Exchange Clustering の概要について説明する。

3.1 目的関数

Exchange Clustering に用いる目的関数には、翻訳規則の生成確率モデルの対数尤度関数を用いる。

統語ラベル $t_{i,j}^{par}, t_i, t_j$ からなる翻訳規則 $t_{i,j}^{par} \rightarrow t_i t_j$ において、$c(t)$ を統語ラベル t のクラスとすると、統語ラベルをクラスタリングすることにより得られる生成確率は

$$\begin{aligned} p(t_i, t_j | t_{i,j}^{par}) &\approx p_0(t_i|c(t_i)) \cdot p_0(t_j|c(t_j)) \\ &\quad \cdot p_1(c(t_i), c(t_j)|c(t_{i,j}^{par})) \quad (1) \end{aligned}$$

となる。

この式 1 に対し、子ノードのラベルが互いに独立であると仮定し、さらに Uszkoreit ら (2008) による予測的なモデルによる近似を導入することで、

$$\begin{aligned} p(t_i, t_j | t_{i,j}^{par}) &\approx p_0(t_i|c(t_i)) \cdot p_0(t_j|c(t_j)) \\ &\quad \cdot p_1(c(t_i))|c(t_{i,j}^{par})) \\ &\quad \cdot p_1(c(t_j)|c(t_{i,j}^{par})) \quad (2) \\ &\approx p_0(t_i|c(t_i)) \cdot p_0(t_j|c(t_j)) \\ &\quad \cdot p_1(c(t_i)|t_{i,j}^{par}) \\ &\quad \cdot p_1(c(t_j)|t_{i,j}^{par}) \quad (3) \end{aligned}$$

となる。

そして、式 3 の生起確率を相対頻度で表し、

$$p(t_i, t_j | t_{i,j}^{par}) = \frac{N(t_i)}{N(c(t_i))} \cdot \frac{N(t_i^{par}, c(t_i))}{N(t_i^{par})} \cdot \frac{N(t_j)}{N(c(t_j))} \cdot \frac{N(t_{i,j}^{par}, c(t_j))}{N(t_{i,j}^{par})} \quad (4)$$

となる。$N(t), N(c(t))$ は全翻訳規則中のラベル、クラスタの頻度を表し、$N(t, c)$ は t を親ノードのラベルとして持つクラスタ c の頻度を表す。頻度には翻訳規則に付与されているフレーズの頻度を足した値を用い、クラスタの頻度はそのクラスタ内のラベルの総頻度を表す。

この式 4 を対数尤度関数に置き換えて足し上げた次の式 5 が目的関数となる。

$$\begin{aligned} F(C) &= \sum_{t \in T} N(t) \cdot log \frac{N(t)}{N(c(t))} \\ &\quad + \sum_{u \in T, c \in C} N(u, c) \cdot log \frac{N(u, c)}{N(u)} \\ &= \sum_{t \in T, c \in c(t)} N(t, c) \cdot log N(t, c) \\ &\quad - \sum_{c \in C} N(c) \cdot log N(c) \end{aligned} \quad (5)$$

C は全クラスタを、T は全統語ラベルを表す。

3.2 Exchange Clustering

Exchange Clustering の概要 (表 2) を説明する。

まず、あらかじめクラスタ数を決めておき、全ての統語ラベルを適当なクラスタに割り振る。提案手法では、頻度の少ないラベルがクラスタ化されるように割り振った。

そして、それぞれのラベルを他のクラスタに移動した際の $F(C)$ の変化を求め、$F(C)$ が最も増加しているクラスタに移動する。増加する移動先がない場合は移動しない。

この操作を、一定の条件を満たすまで行う。提案手法では、操作回数の上限を 50 回とし、全てのラベルの移動がなくなった時点で操作は終了する。

4 評価実験

提案手法の有効性を確認するために日英翻訳タスクにおいて評価実験を行った。本稿で扱う構文拡張機械翻訳 (SAMT) は、目的言語の統語ラベルを用いて行った。SAMT の実装として Moses を利用した。

表 2: Exchange Clustering の概要

start with the initial mapping (label $t \rightarrow c(t)$)
compute objective function F(C)
for each label t do
remove label t from $c(t)$
for each class k do
move label t tentatively to class k
compute F(C) for this exchange
move label t to class with maximum F(C)
do until the class mapping does not change

4.1 実験設定

本稿では、旅行会話に関する対訳コーパスである IWSLT2007(Fordyce, 2007) のデータセットの中の日英のデータを用いて実験を行った (訓練データ: 約 4 万文対、開発データ、テストデータ:約 500 文対)。

日本語の単語分割には KyTea(Neubig et al., 2011) を利用した。英語のトークン化には Moses に付属しているスクリプトを利用した。英語の構文解析器には、Collins parser(Collins, 1999) を利用した。言語モデルは、IRSTLM[2] を用いて英語の単語 5-gram で学習した。アライメント学習には MGIZA++[3] を、チューニングには MERT(Och, 2003) を利用した。デコーダには、Moses 0.91 を使用した。

4.2 比較手法

次の手法を従来法として比較を行った。

PB-SMT フレーズベース翻訳

Hiero 階層的フレーズベース翻訳

SAMT0 構文解析器が付与する統語ラベルを用いた SAMT

SAMT1,2,4,5 細分化した統語ラベルを用いた SAMT

BL_SAMT1,2,4,5 凝集型クラスタリングでクラスタ化した統語ラベルを用いた SAMT

BL_SAMT1,2,4,5 は、Hanneman ら (2013) の手法を目的言語側の統語ラベルのみを用いて行うようにアレンジしたものを、SAMT1,2,4,5 に適用した結果である。クラスタ数は 80 とした。

[2]http://sourceforge.net/projects/irstlm/
[3]http://sourceforge.net/projects/mgizapp/

表 3: 実験結果

Type	Label	Rule	F(C)	BLEU
PB-SMT	-	40k	-	45.01
Hiero	1	2.0M	-4.7 e+08	49.31
SAMT0	62	193k	-2.3e+06	31.65
SAMT1	1506	580k	-7.8 e+06	33.74
BL_SAMT1	80	350k	-4.9 e+07	40.45
EC_SAMT1	80	455k	-7.8 e+06	41.76
SAMT2	3589	1.9M	-3.4 e+07	47.66
BL_SAMT2	80	1.1M	-2.6 e+08	50.15
EC_SAMT2	80	1.4M	-3.5 e+07	50.46
SAMT4	3619	3.7M	-5.7 e+07	46.0
BL_SAMT4	80	3.3M	-4.6 e+08	49.1
EC_SAMT4	80	2.7M	-6.4 e+07	49.61
SAMT5	17k	4.1M	-4.4 e+07	47.66
BL_SAMT5	80	2.1M	-4.7 e+08	49.19
EC_SAMT5	80	3.8M	-5.0 e+07	49.75

4.3 実験結果と分析

実験結果を表 3 に示す。提案手法は EC_SAMT 1,2,4,5 で表す。クラスタ数は 80 とした。

翻訳精度の評価尺度には BLEU(Papineni et al., 2002) を用い、統語ラベル数と翻訳規則数、最適化基準である $F(C)$ の値も調べた。

表 3 より、提案手法の EC_SAMT5 が最も BLEU 値が高く、統語ラベルクラスタリングが有効に動作していると考えられる。

ラベルの合成条件を変えた SAMT0,1,2,4,5 を見ると、SAMT1 から 2 への効果は高いが、それ以降の細かい細分化は大幅な精度の向上には繋がっていない。翻訳規則数が増えているにも関わらず精度が向上していない理由としては、誤った規則が生成されている、適用されない規則が生成されているなどが考えられる。

5 まとめ

本稿では、目的言語側の統語ラベルを用いた構文拡張機械翻訳 (SAMT) において、細分化した統語ラベルを Exchange Clustering でクラスタ化する手法を提案した。そして、日英翻訳タスクで実験を行い、提案手法が翻訳精度の向上に繋がることを示した。今後は、本稿で利用した統語ラベルの細分化に加え、言語学的知見や詳細な言語素性に基づく統語ラベルの細分化を行い、評価したい。また、大規模かつ長文を含んだコーパスで評価を行いたい。

参考文献

David Chiang. *Hierarchical phrase-based translation*. Computational Linguistics, Vol. 33, No. 2, pages 201-228, June 2007.

Micheal Collins. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis,University of Pennsylvania 1999.

Cameron Shaw Fordyce. *Overview of the 4th International Workshop on Spoken Language Translation IWSLT 2007 evaluation campaign*. Proceedings of IWSLT 2007, pages 112, Trento, Italy, October 2007.

Greg Hanneman and Alon Lavie. *Improving syntax-augmented machine translation by coarsening the label set*. Proceedings of AACL-HLT 2013, pages 288-297, Atlanta, Georgia, USA, June 2013.

Phillip Koehn, Franz Josef Och, and Daniel Marcu. *Statistical phrase-based translation*. Proceedings of HLT-NAACL 2003, pages 4854, Edmonton, Canada, May/June 2003.

Sven Martin, Jorg Liermann, and Hermann Ney. *Algorithms for bigram and trigram word clustering* . Speech Communication, Vol.24, pages 19-37, 1998.

Graham Neubig, Yosuke Nakata, Shinsuke Mori. *Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis*. Proceedings of ACL-HLT 2011, pages 529533, Portland, Oregon, USA, June 2011.

Franz Josef Och. *Minimum Error Rate Training in Statistical Machine Translation*. Proceedings of ACL 2003, pages 160-167, Sapporo, Japan, July 2003.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. *BLEU: a method for automatic evaluation of machine translation*. Proceedings of ACL 2002, pages 311318, Philadelphia, PA, USA, July 2002.

Jakob Uszkoreit and Thorsten Brants. *Distributed Word Clustering for Large Scale Class-Based Language Modeling in Machine Translation*. Proceedings of ACL-HLT 2008, pp. 755-762, Columbus, Ohio, USA, June 2008.

Wei Wang, Jonathan May, Kevin Knight, and Daniel Marcu. *Re-structuring, re-labeling, and re-aligning for syntax-based machine translation*. Computational Linguistics, Vol. 36, No. 2, pages 247277, June 2010.

Andreas Zollmann and Ashish Venugopal. *Syntax augmented machine translation via chart parsing*. Proceedings of the SMT Workshop, HLT-NAACL 2006, pages 138-141, New York, NY, USA, June 2006.

須藤克仁, 進藤裕之, 塚田元, 永田昌明. **統計翻訳における統語的ラベル細分化の検討**. 言語処理学会第 19 回年次大会, pages 390-393, March 2013.