Vol.2015-NL-222 No.2
2015/7/15

中間言語モデルを用いたピボット翻訳の精度向上

三浦 明波[1,a)]　Graham Neubig[1,b)]　Sakriani Sakti[1,c)]　戸田 智基[1,d)]　中村 哲[1,e)]

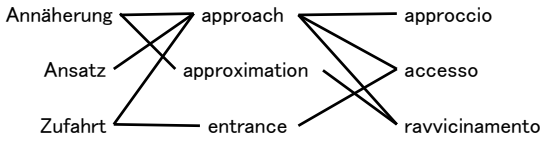
概要：統計的機械翻訳において，特定の言語対で十分な文量の対訳コーパスが得られない場合，中間言語を用いたピボット翻訳が有効な手法の一つである．複数のピボット翻訳手法が考案されている中でも，特に中間言語を介する 2 つの翻訳モデルを合成するテーブル合成手法で，高い翻訳精度を得られることが知られている．ところが，従来のテーブル合成手法では，フレーズ対応推定時に用いた中間言語の情報が「忘却」され，翻訳時には利用できない問題が発生する．本稿では，合成時に用いた中間言語の情報も「記憶」し，ピボットの言語モデルを追加の情報源として翻訳に利用する新たなテーブル合成手法を提案する．また，欧州議会議事録による多言語コーパスを用いた実験により，本手法で評価を行った全ての言語の組合せで従来手法よりも有意に高い翻訳精度が得られた．

1. はじめに

統計的機械翻訳 (Statistical Machine Translation: SMT[1]) では，学習に用いる対訳コーパスが大規模になるほど，高精度な訳出結果を得られることが知られている [2]．一方，英語を含まない言語対などを考慮すれば，多くの言語対において，大規模な対訳コーパスを取得することは困難と言える．特定の言語対で十分な文量の対訳コーパスが得られない場合，中間言語を用いたピボット翻訳が有効な解法の一つである [3]．

中間言語を用いる手法も様々なものが考案されている [4][5][6] が，特に原言語・中間言語，中間言語・目的言語の 2 つの翻訳モデルを合成し，新しく得られた原言語・目的言語翻訳モデルによって翻訳を行うテーブル合成手法で，高い翻訳精度を得られることが示されている．ところが，語義曖昧性や言語間の用語法の差異により，原言語・目的言語間のフレーズ対応を正確に推定することは困難である．

図 1 (a) はテーブル合成手法によって対応を推定するフレーズの例を示しており，図中ではドイツ語とイタリア語それぞれにおける 3 つの単語が，語義曖昧性を持つ英単語「approach」に結び付いている．このような場合，原言語・目的言語間のフレーズ対応を求め，適切な翻訳確率を推定するのは複雑な問題となってくる．その上，図 1 (b) に示すように，従来のテーブル合成手法では，合成時に原言語と目的言語の橋渡しをしていた中間言語フレーズの情報が，合成後には保存されず失われてしまう．



(a) モデル学習に用いるフレーズ対応 (独-英-伊)

Annäherung → approccio
(via: approach)
Annäherung → ravvicinamento
(via: approach, approximation)
・・・

(b) 従来手法によって得られるフレーズ対応

Annäherung → 〈approccio, approach〉
Annäherung → 〈ravvicinamento, approach〉
Annäherung → 〈ravvicinamento, approximation〉
・・・

(c) 提案手法によって得られるフレーズ対応

図 1　テーブル合成手法の例，および従来手法と提案手法の比較

この問題を克服するため，本稿では原言語と目的言語を結び付けていた中間言語フレーズの情報も翻訳モデル中に保存し，原言語から目的言語と中間言語への同時翻訳確率を推定することによって翻訳を行う新しいテーブル合成手法を提案する．図 1 (c) に，本提案手法によって得られるフレーズ対応の例を示す．本手法の利点は，英語のように中間言語として選ばれる言語は豊富な単言語資源も得られる傾向が強いため，このような追加の言語情報を翻訳システムに組み込み，翻訳の質を向上させられることにある．

[1] 奈良先端科学技術大学院大学 情報科学研究科
Nara Institute of Science and Technology
[a)] miura.akiba.lr9@is.naist.jp
[b)] neubig@is.naist.jp
[c)] ssakti@is.naist.jp
[d)] tomoki@is.naist.jp
[e)] s-nakamura@is.naist.jp

中間言語フレーズの情報を翻訳時に役立てるため，同期文脈自由文法 (Synchronous Context-free Grammar: SCFG[7]) を複数の目的言語文の同時生成に対応できるよう拡張した複数同期文脈自由文法 (Multi-Synchronous CFG: MSCFG[8]) を用いて翻訳モデルの学習を行う．MSCFG による翻訳モデルを構築するために，原言語・中間言語および中間言語・目的言語の SCFG 翻訳規則が格納されたルールテーブルを元に，SCFG ルールテーブルとしてではなく，原言語・目的言語・中間言語の MSCFG ルールテーブルとして合成し，これによってピボットを記憶する．訳出候補の探索時には，生成文の自然性を評価し，適切な語彙選択を促すために言語モデルを用いるが，目的言語モデルのみでなく，中間言語モデルも同時に用いた探索を行う．本手法の有効性を調査するため，欧州議会議事録を元にした Europarl 多言語コーパスから英語を中間言語とし，異なる 4 つの言語を用いて実験を行ったが，すべての組合せにおいて従来手法よりも有意に高い翻訳精度が得られた．

2. 機械翻訳方式

2.1 同期文脈自由文法

本節では先ず，階層的フレーズベース翻訳 (Hierarchical Phrase-Based Translation: Hiero[7]) を代表とする様々な翻訳方式で用いられる SCFG について紹介する．SCFG は，以下のような置換規則によって構成される．

$$X \rightarrow \langle \bar{s},\ \bar{t} \rangle \tag{1}$$

ここで， X は置換規則の親記号であり，$\bar{s}$ と $\bar{t}$ はそれぞれ原言語と目的言語における終端記号と非終端記号からなる記号列である．$\bar{s}$ と $\bar{t}$ にはそれぞれ同じ数の非終端記号が含まれ，対応する記号に対して同じインデックスが付与される．以下に置換規則の例を示す．

$$X \rightarrow \langle X_0 \ \text{ of } \ X_1,\ X_1 \ \text{の}\ X_0 \rangle \tag{2}$$

Chiang による SCFG の学習手法では，対訳文と単語アラインメントを元に自動的に SCFG ルールが抽出される．抽出された各々のルールには，双方向のフレーズ翻訳確率 $\phi(\bar{s}|\bar{t})$, $\phi(\bar{t}|\bar{s})$，双方向の語彙翻訳確率 $\phi_{lex}(\bar{s}|\bar{t})$, $\phi_{lex}(\bar{t}|\bar{s})$，ワードペナルティ ($\bar{t}$ の終端記号数)，フレーズペナルティ (定数 1) の計 6 つのスコアが付与される．

翻訳時には，導出に用いられるルールのスコアと，生成される目的言語文の言語モデルスコアの和を導出確率として最大化するよう探索を行う．言語モデルを考慮しない場合，CKY+法 [9] によって効率的な探索を行ってスコア最大の導出を得ることが可能である．言語モデルを考慮する場合には，キューブ枝刈り [7] などの近似法により探索空間を抑えつつ，目的言語モデルを考慮した探索が可能である．

2.2 複数同期文脈自由文法

MSCFG[8] は，SCFG を複数の目的言語文を同時に生成できるように拡張されている．SCFG では生成規則中の目的言語記号列 $\bar{t}$ が単一であったが，MSCFG では以下のように N 個の目的言語記号列を有する．

$$X \rightarrow \langle \bar{s},\ \overline{t_1},\ ...,\ \overline{t_N} \rangle \tag{3}$$

通常の MSCFG 学習手法では，SCFG ルール抽出手法を一般化し，行アラインメントの取れた多言語コーパスから多言語置換規則が抽出され，複数の目的言語を考慮したスコアが付与される．

MSCFG で複数の目的言語モデルを考慮した探索を行う場合，各言語毎に導出中の単語列を記憶し，組合せ毎に状態を区別する必要があるため，探索手順にも複数の手法が考えられる．SCFG の探索における単一の目的言語文を単純に複数の目的言語文に拡張して同時に展開する同時探索では，探索幅の制限により主要な目的言語文の多様性が失われてしまう可能性がある．そこで先ず，第一の目的言語文のみを考慮した組合せで探索して多様性を確保し，続いてその他の目的言語文との組合せに展開する逐次探索により，主要な目的言語を重視した効率的な探索が行える．

Neubig らによる MSCFG の学習手法では，原言語と複数の目的言語とで行アラインメントの取れた大規模な多言語コーパスが必要となるため，直接の対訳コーパスの取得が限られている際には用いることができない．そこで次節からは，2 つの SCFG ルールテーブルを合成することによって MSCFG ルールテーブルを得る手法について述べる．

3. ピボット翻訳手法

SMT において中間言語を用いて翻訳を行う手法は複数考案されており，以下の 3 種類に大別することができる．

逐次的ピボット翻訳 (Cascade):
入力文を原言語・中間言語の翻訳モデルで翻訳し，その出力を中間言語・目的言語文で翻訳する [3]

コーパス翻訳方式 (Synthetic):
原言語・中間言語あるいは目的言語・中間言語の対訳コーパスの中間言語側を翻訳することで擬似的な対訳コーパスを得る [3]

テーブル合成方式 (Triangulation):
原言語・中間言語および中間言語・目的言語の翻訳モデルを合成することで原言語・目的言語の翻訳モデルを得る [4]

とりわけ，テーブル合成手法によるピボット翻訳で高品質な訳出結果が得られることが知られているため [5]，以降ではテーブル合成手法に的を絞って議論を行う．

3.1 従来のテーブル合成手法

Cohn らによるテーブル合成手法 [4] では，先ず原言語・中間言語および中間言語・目的言語の翻訳モデルを対訳データによって学習し，それぞれをフレーズ (ルール) テーブル T_{SP}, T_{PT} として格納する．そして，テーブル T_{SP}, T_{PT} のそれぞれにフレーズ対 $\langle \overline{s}, \overline{p} \rangle$, $\langle \overline{p}, \overline{t} \rangle$ が含まれるような中間言語フレーズ $\overline{p}$ が存在する場合，以下のような規則を作成する．

$$X \rightarrow \langle \overline{s},\ \overline{t} \rangle \tag{4}$$

作成されたすべての規則について，フレーズ翻訳確率 $\phi(\cdot)$ と語彙翻訳確率 $\phi_{lex}(\cdot)$ を以下の計算によって推定することで原言語・目的言語のテーブル T_{ST} を合成する．

$$\phi\left(\overline{t}|\overline{s}\right) = \sum_{\overline{p} \in T_{SP} \cap T_{PT}} \phi\left(\overline{t}|\overline{p}\right) \phi\left(\overline{p}|\overline{s}\right) \tag{5}$$

$$\phi\left(\overline{s}|\overline{t}\right) = \sum_{\overline{p} \in T_{SP} \cap T_{PT}} \phi\left(\overline{s}|\overline{p}\right) \phi\left(\overline{p}|\overline{t}\right) \tag{6}$$

$$\phi_{lex}\left(\overline{t}|\overline{s}\right) = \sum_{\overline{p} \in T_{SP} \cap T_{PT}} \phi_{lex}\left(\overline{t}|\overline{p}\right) \phi_{lex}\left(\overline{p}|\overline{s}\right) \tag{7}$$

$$\phi_{lex}\left(\overline{s}|\overline{t}\right) = \sum_{\overline{p} \in T_{SP} \cap T_{PT}} \phi_{lex}\left(\overline{s}|\overline{p}\right) \phi_{lex}\left(\overline{p}|\overline{t}\right) \tag{8}$$

式 (5)-(8) は，以下のような条件を満たす無記憶通信路モデルに基いている．

$$\phi\left(\overline{t}|\overline{p}, \overline{s}\right) = \phi\left(\overline{t}|\overline{p}\right) \tag{9}$$

$$\phi\left(\overline{s}|\overline{p}, \overline{t}\right) = \phi\left(\overline{s}|\overline{p}\right) \tag{10}$$

ところが，現実には語の多義性や言語間の文法の不一致によって，これらの数式は正確ではない．結果として，ピボット翻訳は通常の機械翻訳よりも遥かに大きな曖昧性の問題を抱えている．

3.2 提案するテーブル合成手法

前述の曖昧性の問題に対処するため，本稿で提案するテーブル合成手法では，関連する中間言語フレーズを追加の言語情報として記憶することで，曖昧性の解消に利用する．特に，中間言語フレーズ $\overline{p}$ で確率周辺化を行い SCFG ルールを合成する代わりに，原言語・目的言語・中間言語のフレーズ対応を MSCFG ルールとして以下のように合成する．

$$X \rightarrow \langle \overline{s},\ \overline{t},\ \overline{p} \rangle \tag{11}$$

このような規則を用いて翻訳を行うことによって，同時生成される中間言語文を通じて中間言語モデルなどのような追加の素性を取り入れることが可能となる．式 (5)-(8) に加えて，目的言語と中間言語を同時に考慮した翻訳確率 $\phi(\overline{t}, \overline{p}|\overline{s})$, $\phi(\overline{s}|\overline{p}, \overline{t})$ を以下のように推定する．

$$\phi\left(\overline{t}, \overline{p}|\overline{s}\right) = \phi\left(\overline{t}|\overline{p}\right) \phi\left(\overline{p}|\overline{s}\right) \tag{12}$$

$$\phi\left(\overline{s}|\overline{p}, \overline{t}\right) = \phi\left(\overline{s}|\overline{p}\right) \tag{13}$$

原言語・中間言語間の翻訳確率 $\phi(\overline{p}|\overline{s})$, $\phi(\overline{s}|\overline{p})$, $\phi_{lex}(\overline{p}|\overline{s})$, $\phi_{lex}(\overline{s}|\overline{p})$ はテーブル T_{SP} のスコアをそのまま用いることが可能である．これら 10 個の翻訳確率に加えて，$\overline{t}$ と $\overline{p}$ に含まれる非終端記号数を 2 つのワードペナルティとし，定数 1 のフレーズペナルティの合わせて 13 個のスコアが MSCFG ルールにおける素性となる．

このように中間言語を記憶するテーブル合成手法では，$\langle \overline{s}, \overline{t} \rangle$ ではなく，$\langle \overline{s}, \overline{t}, \overline{p} \rangle$ の全組み合わせを記録するため，従来より大きなルールテーブルが合成されてしまう．計算資源を節約するためには，幾つかのフィルタリング手法が考えられる．Neubig らによると，主要な目的言語 T_1 と補助的な目的言語 T_2 で翻訳を行う際には，T_1-フィルタリング手法 [8] が効果的である．この手法では，原言語フレーズ $\overline{s}$ に対して，先ず言語 T_1 において $\phi(\overline{t_1}|\overline{s})$ が上位 L 個までの $\overline{t_1}$ を残し，それぞれの t_1 に対して $\phi(\overline{t_1}, \overline{t_2}|\overline{s})$ が最大となるような $\overline{t_2}$ を残す．

4. 実験的評価

4.1 実験設定

本提案手法の有用性を評価するため，欧州諸言語を広くカバーし，ピボット翻訳のような多言語翻訳タスクで広く用いられる Europarl コーパス [10] を用いて実験を行った．本実験では，英語 (en) を中間言語として固定し，欧州の中でも特に話者数の多いドイツ語 (de), スペイン語 (es), フランス語 (fr), イタリア語 (it) の 4 言語の組合せでピボット翻訳を行い，手法毎の翻訳精度を比較した．5 言語間の対訳コーパスを得るため，先ず Gale-Church アラインメント法 [11] によって行アラインメントの取れた多言語コーパス約 90 万文を取得し，そこから 1,500 文ずつを最適化と評価用に取り出した．それぞれの翻訳モデルの学習には 10 万文，目的言語モデルの学習にも 10 万文を用いた．また，多くの場合，英語においては大規模な単言語資源が取得可能であるため，最大 200 万文までのデータを用いて段階的に学習を行った中間言語モデルを用意した．デコーダには Travatar [12] を用い，付属の Hiero ルール抽出コードを用いて SCFG 翻訳モデルの学習を行った．翻訳結果の評価には，自動評価尺度 BLEU [13] を用い，各翻訳モデルは MERT[14] により，開発データセットに対して BLEU スコアが最大となるようにパラメータを調整を行った．提案手法のテーブル合成によって得られた MSCFG ルールテーブルは，$L = 20$ の T_1 フィルタリング手法によって枝刈りを行った．また，本実験の MSCFG を用いた探索では，目的言語文，中間言語文の順に組合せを展開する逐次探索を行った．

次節では以下の 6 つの翻訳手法を比較評価する．

Source	Target	BLEU Score [%]					
		Direct	Cascade	Tri. SCFG (baseline)	Tri. MSCFG -PivotLM	Tri. MSCFG +PivotLM 100k	Tri. MSCFG +PivotLM 2M
de	es	27.10	25.05	25.31	25.38	25.52	† **25.75**
	fr	25.65	23.86	24.12	24.16	24.25	† **24.58**
	it	23.04	20.76	21.27	21.42	† 21.65	‡ **22.29**
es	de	20.11	18.52	18.77	18.97	19.08	† **19.40**
	fr	33.48	27.00	29.54	† 29.87	† 29.91	† **29.95**
	it	27.82	22.57	25.11	25.01	25.18	‡ **25.64**
fr	de	19.69	18.01	18.73	18.77	18.87	† **19.19**
	es	34.36	27.26	30.31	30.53	† 30.73	‡ **31.00**
	it	28.48	22.73	25.31	25.50	† 25.72	‡ **26.22**
it	de	19.09	14.03	17.35	† 17.99	‡ 18.17	‡ **18.52**
	es	31.99	25.64	28.85	28.83	29.01	† **29.31**
	fr	31.39	25.87	28.48	28.40	28.63	† **29.02**

表 1　各手法による翻訳精度．太字はそれぞれの言語対において最も BLEU スコアが高いことを示し，短剣符は提案手法の翻訳精度が従来手法よりも統計的に有意であることを示す (†: $p < 0.05$, ‡: $p < 0.01$)

Direct:
比較のため，中間言語を用いず原言語・目的言語の直接対訳コーパスで学習した SCFG で翻訳．

Cascade:
原言語・中間言語および中間言語・目的言語の SCFG モデルで逐次的ピボット翻訳．

Tri. SCFG:
原言語・中間言語および中間言語・目的言語の SCFG モデルを合成し，原言語・目的言語の SCFG モデルによって翻訳．

Tri. MSCFG:
原言語・中間言語および中間言語・目的言語の SCFG モデルを合成し，原言語・目的言語・中間言語の MSCFG によって翻訳．「-Pivot」は中間言語モデルを用いないことを示し，「+PivotLM 100k/2M」はそれぞれ 10 万文，200 万文で学習した中間言語を用いることを示す．

4.2　実験結果

表 1 に，英語を介したすべての言語対におけるピボット翻訳の結果を示す．評価値から，提案したテーブル合成手法で中間言語モデルを考慮した翻訳を行った場合，すべての言語対において従来のテーブル合成手法よりも BLEU スコアが上昇していることが確認できる．すべての組合せにおいて，テーブル合成手法で中間言語情報を記憶し，200 万文の言語モデルを考慮して翻訳を行った場合に最も高いスコアを達成しており，従来法に比べ 0.4 から 1.2 ほどの BLEU 値の向上が見られる．このことから，中間言語情報を記憶し，これを翻訳に利用することが曖昧性の解消に繋がり，安定して翻訳精度を改善できると言えるだろう．

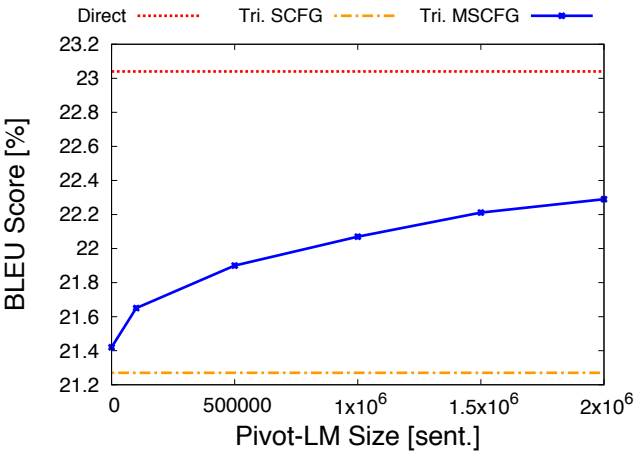


図 2　中間言語モデル規模がピボット翻訳精度に与える影響 (独-伊)

また，MSCFG に合成するが中間言語モデルを用いず翻訳を行う場合を見てみると，SCFG に合成する場合よりも多くの言語対で僅かに精度が向上している．これは，追加の翻訳確率などのスコアが有効な素性として働き，適切な語彙選択に繋がったことなどが原因として考えられる．

中間言語モデルのサイズがピボット翻訳精度に与える影響の大きさは言語対によって異なってはいるが，中間言語モデルの規模が大きくなるほど精度が向上することも確認できる．図 2 はドイツ語・イタリア語のピボット翻訳において異なるデータサイズで学習した言語モデルが翻訳精度に与える影響を示す．

最後に，本提案手法によって中間言語側で曖昧性が解消されて精度向上に繋がったと考えられる訳出の例を示す．

入力文 (ドイツ語):

ich bedaure , daß es keine gemeinsame annäherung gegeben hat .

参照文 (イタリア語):

sono spiacente del mancato approccio comune .

Tri. SCFG:

mi rammarico per il fatto che non si ravvicinamento comune . (BLEU+1: 13.84)

Tri. MSCFG+PivotLM 2M:

mi dispiace che non esiste un approccio comune . (BLEU+1: 25.10)

i regret that there is no common approach .

(Generated English Sentence)

上記の Tri. MSCFG+PivotLM 2M で導出に用いられた MSCFG ルールでは，イタリア語「approccio」と英語「approach」が結び付いており，生成される英文中の単語の前後関係から適切な語彙選択を促し，精度向上に繋がったものと考えられる．

5. おわりに

本稿ではピボット翻訳において，中間言語を介する 2 つの SCFG ルールテーブルを 1 つの MSCFG ルールテーブルに合成し中間言語情報を記憶して翻訳を行う新しいテーブル合成手法を提案した．そして実験結果から，本手法で得られる MSCFG ルールテーブルと，大規模な言語モデルを用いて翻訳を行うことで高いピボット翻訳精度が得られることが分かった．今後の計画として，中間言語における表現を工夫し，合成されるルールの翻訳確率を高精度に推定する手法の提案などを行っていきたい．

謝辞： 本研究の一部は，Microsoft CORE プロジェクトの助成を受け実施したものである．

参考文献

[1] Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 19, pp. 263–312, 1993.

[2] Christopher Dyer, Aaron Cordova, Alex Mont, and Jimmy Lin. Fast, easy, and cheap: construction of statistical machine translation models with MapReduce. In *Proc. WMT*, pp. 199–207, 2008.

[3] Adrià de Gispert and José B. Mariño. Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish. In *Proc. of LREC 5th Workshop on Strategies for developing machine translation for minority languages*, 2006.

[4] Trevor Cohn and Mirella Lapata. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proc. ACL*, pp. 728–735, June 2007.

[5] Masao Utiyama and Hitoshi Isahara. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Proc. NAACL*, pp. 484–491, 2007.

[6] Xiaoning Zhu, Zhongjun He, Hua Wu, Conghui Zhu, Haifeng Wang, and Tiejun Zhao. Improving Pivot-Based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs. In *Proc. EMNLP*, 2014.

[7] David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, Vol. 33, No. 2, pp. 201–228, 2007.

[8] Graham Neubig, Philip Arthur, and Kevin Duh. Multi-Target Machine Translation with Multi-Synchronous Context-free Grammars. In *Proc. NAACL*, 2015.

[9] Jean-Cédric Chappelier, Martin Rajman, et al. A Generalized CYK Algorithm for Parsing Stochastic CFG. *TAPD*, Vol. 98, No. 133-137, p. 5, 1998.

[10] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5, pp. 79–86, 2005.

[11] William A Gale and Kenneth W Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, Vol. 19, No. 1, pp. 75–102, 1993.

[12] Graham Neubig. Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers. In *Proc. ACL Demo Track*, pp. 91–96, 2013.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pp. 311–318, 2002.

[14] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. ACL*, pp. 160–167, 2003.