目的言語の構文解析器を用いた機械翻訳のプレオーダリング

Preordering using a Target-Language Parser for Statistical Machine Translation

日本放送協会　放送技術研究所ヒューマンインターフェース研究部専任研究員　**後藤　功雄**

PROFILE　2014 年京都大学大学院情報学研究科博士課程修了。博士（情報学）。1997 年 NHK 入局。2008 年から 2013 年まで情報通信研究機構に出向。自然言語処理の研究に従事。

1 はじめに

筆者らは統計的機械翻訳（SMT）において語順推定を改良する研究を情報通信研究機構および NHK にて実施した [1,2]。本稿では、この研究成果について紹介する。語順が大きく異なる言語間の機械翻訳では、目的言語の語順を推定する必要がある。語順を推定するために統計的機械翻訳（SMT）では、語彙化語順推定モデル [3]、階層フレーズベース [4]、構文ベース [5]、プレオーダリング [6] などの手法が提案されてきた。プレオーダリングは、原言語文のみに対する処理であるために、長距離の語順並べ替えに有用な原言語の構文構造をシンプルに利用できるという特徴がある。英日翻訳で高性能な英語の構文解析器を用いたプレオーダリングは有効性が高いことが確認されている [7,8]。構文構造を用いる既存のプレオーダリング手法は原言語の構文解析器を必要とする。しかし、多くの言語では高性能な構文解析器は利用できない。機械翻訳が必要とされる状況として、原言語では高性能な構文解析器が利用できないが目的言語では利用でき、原言語と目的言語の語順が大きく異なる場合が考えられる。本稿はこの状況で利用できるプレオーダリング手法を提案する。提案手法は、目的言語の構文解析器で獲得した目的言語文の構文構造を原言語文に射影して同期率の高い原言語の構文構造を構築することで、構文構造を利用するプレオーダリングモデルを構築する。

2 プレオーダリング手法

機械翻訳は原言語文 F を目的言語文 E へ変換する処理と定義できる。この処理で語順が異なる言語間では語順の変更が必要である。プレオーダリングによる翻訳は語順並べ替えと訳語選択の処理を 2 段階に分けて、次のように翻訳する。はじめに、F を、ほぼ目的言語の語順である原言語の単語列 F' に並べ替え（プレオーダリング）、次に、F' を E に翻訳する。

プレオーダリング手法として多くの手法が提案されている。ほとんどのプレオーダリング手法は、原言語の構文解析器と並べ替えルールを用いる [6,7]。これらの手法は、原言語の構文解析器が利用できない場合は適用できない。この場合でも利用できる、構文解析器を必要としない手法も提案されている [9]。この手法は対訳コーパスと単語アラインメントを用いてシンタックスに基づかない構造（非構文の構造）の解析器を構築する。そして、この解析器で原言語文の構造を解析して BTG[10] に基づいて並べ替える。

構文構造は、非構文の構造に比べて語順の推定で次の点で優れていると考えられる。

- 構文構造は意味表現と部分構造が一致していると考えられる。例えば、節は 1 つの意味表現になっておりかつ構文構造の部分構造になっている。それに対して、非構文の構造は必ずしも意味表現と部分構造が一致するとは限らない。
- 構文構造は非構文の構造より情報量が多い。構文構造は多くのフレーズラベルを用いるが、非構文の構造は

1 種類のフレーズラベルしか用いない。

3 提案手法の概要

提案手法は、原言語の構文解析器が利用できない場合でも、目的言語の構文解析器を用いることで、構文構造に基づいたプレオーダリングができる。対訳文では、原言語と目的言語の構文構造は類似していることが期待される [11]。この期待に基づいて対訳文中の原言語の構文構造を構築し、ITG[10] に基づくプレオーダリングモデルを学習する。

ITG/BTG の効果的な学習には、対訳構造の同期率が高いことが重要である。なぜなら、ITG/BTG は同期している部分から学習されるためである。そこで、言語間の射影によって構文構造の同期率が高い対訳文を選択し、さらに射影に基づいて同期率が高い構文構造を構築することによって、ITG/BTG の効果的な学習を促進する。

プレオーダリングモデルは次のステップで構築する。

1. 目的言語の構文解析器を用いて、対訳コーパスの目的言語文の 2 分木構文構造を獲得する。
2. 目的言語文の部分的な構文構造を、単語アラインメントを用いて原言語文に射影する。(4.1 節)
3. 射影された部分構造を用いて同期率の高い対訳文を選択する。(4.2 節)
4. 射影された部分構造を用いて確率的 CFG と教師無しの確率的品詞推定モデルを構築する。(4.3 節)
5. 射影された部分構造を制約として用いて、構築した確率モデルで訓練データの原言語文を構文解析し、同期率の高い構文構造を構築する。(4.4 節)
6. 構築した原言語の構文構造と単語アラインメントを用いて ITG に基づくプレオーダリングモデルを学習して構築する。(4.5 節)

プレオーダリングモデルを構築した後、このモデルを用いて対訳コーパスの原言語文をプレオーダリングして、F' と E のパラレルコーパスを構築する。このコーパスを用いて SMT のモデルを学習する。

入力文の翻訳は、プレオーダリングモデルを用いて入力文 F を F' に変換してから SMT で翻訳する。

本研究のメインの貢献は、目的言語の構文解析器を用いたプレオーダリングの枠組みである。これに加えて、射影による新しい句構造構築手法を提案する。提案手法は既存の射影による句構造構築手法 [12] と比べて次の 2 つの違いがある。(1) CFG の確率推定において、既存手法では射影から得られる曖昧性のある候補の確率に一様分布を仮定しているが、この仮定は正しくない。それに対して、提案手法は全ての候補の確率を計算する。(2) 既存手法は原言語の品詞タグを必要とするが、提案手法は必要としない。ただし、原言語文の単語分割は必要である。

以下、プレオーダリングモデルの構築の詳細について説明する。

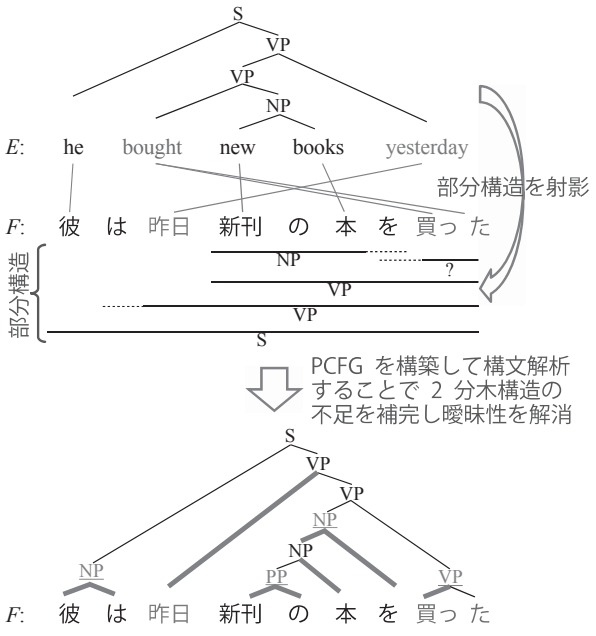

図 1 部分構造の射影と 2 分木構造の構築例

4 モデルの訓練

本節では、前記のステップ 2 以降を説明する。

4.1 部分構造の射影

まず、自動単語アラインメント手法を用いて対訳文間の単語アラインメントを獲得する。このアラインメントを介して目的言語文の部分的な構文構造を原言語文に射影する。これによって原言語文の部分的な構文構造が得られる。射影の例を図 1 の上部に示す。

射影は次のように行う。単語アラインメントを介し

て E の部分木のスパンに対応する F の部分木のスパンを特定し、E の部分木の根のフレーズラベルを F のスパンに付与する。F の部分木のスパンは、E のスパン中の語にアラインメントされた語の左端から右端までとする。この F のスパンを最小射影スパンと呼ぶ。最小射影スパンに隣接していてアラインメントされていない語は、このスパンに含まれる可能性があり、スパンの範囲に曖昧性がある状態とする。図 1 では、最小射影スパンを水平の実線、アラインメントされていない語の部分を水平の破線で示している。

最小射影スパンが互いに部分的に重複する（不整合と呼ぶ）と、射影された部分構造を補完しても木構造を構成できないので、不整合な部分構造は削除する。

4.2 同期率の高い対訳文の選択

射影した部分構造を用いて構造の同期率が高い対訳文を選択する。各対訳文対での同期率とは、「原言語文でのスパンの数」を「原言語文中の語数−1」で割った値で計算する。選択した対訳文は 4.3 ～ 4.5 節で用いる。

4.3 構文解析のための確率モデルの構築

射影された原言語の部分構造からプレオーダリングモデルの学習に用いる 2 分木構造を獲得するために、構文解析用の確率モデルを構築する。F とその部分構造を入力として用いて、原言語に対する確率的文脈自由文法（PCFG）および教師無しの確率的品詞推定モデルを構築する。これらのモデルを Pitman-Yor 過程 (PY)[13] を用いて構築する。なぜなら、その “rich-get-richer” の特性が部分的に付与された構造を持つデータからモデルを学習するのに適しているためである。

ここで用いる CFG ルール $x \rightarrow \alpha$ は、非終端記号 $x \in V$ と 2 つの非終端記号で構成される順序対 α からなる。非終端記号の集合 V は $V = \mathcal{L} \cup \mathcal{T}$ で、$\mathcal{L}$ はフレーズラベルの集合である。$\mathcal{T}=\{1,2,\cdots,|\mathcal{T}|\}$ は原言語の教師無しの品詞タグを表す数字の集合で、$|\mathcal{T}|$ は品詞タグの種類数を表す。訓練データ中の原言語の単語集合を $\mathcal{F}$ とし、$F = f_1\ f_2 \cdots f_m,\ f \in \mathcal{F}$ とする。木構造 D の確率は、その構成要素である CFG ルールと単語の確率の積により式 (1) で計算する。

$$P(D) = \prod_{x \rightarrow \alpha \in \mathcal{R}} P(\alpha|x)^{c(x \rightarrow \alpha, D)} \prod_{i=1}^{m} P(f_i|t_i) \quad (1)$$

ここで、$\mathcal{R}$ は CFG ルールの集合を表し、$c(x \rightarrow \alpha, D)$ は D を構成する CFG ルール $x \rightarrow \alpha$ の頻度を表し、$t \in \mathcal{T}$ は品詞タグを表し、t の添え字 i は F での単語位置を表す。木構造の根のフレーズラベルには、フレーズラベル S を用いる。

PY モデルは CFG ルールまたは原言語の単語の確率分布として次式で表される。

$P(\alpha|x) \sim PY_x\ (d_{\text{cfg}},\ \theta_{\text{cfg}},\ P_{\text{base}}\ (\alpha|x))$

$P(f|t) \sim PY_x\ (d_{\text{tag}},\ \theta_{\text{tag}},\ P_{\text{base}}\ (f|t))$

ここで、d_{cfg}, θ_{cfg}, d_{tag}, θ_{tag} は、PY モデルのハイパーパラメータであり、文献 [14] の手法で最適化する。バックオフの確率分布には一様分布、すなわち、$P_{\text{base}}\ (\alpha|x) = 1/|V|^2$ および $P_{\text{base}}\ (f|t) = 1/|\mathcal{F}|$ を用いる。ここで、$|V|$ は非終端記号の種類数、$|\mathcal{F}|$ は訓練データ中の原言語の単語の種類数である。

式 (1) および次の制約に基づいてサンプリングすることでモデルを構築する。最小射影スパンが存在する場合は、アラインメントされていない語の部分を除いたスパンが最小射影スパンと不整合にならないスパンをサンプリングする。そして、フレーズラベルが射影されているスパンでは、射影されているフレーズラベルをサンプリングする。

サンプリングは、動的計画法に基づいて文構造単位でギブスサンプリングにより行う [15]。各文において、CYK アルゴリズムでボトムアップに内側確率を計算し、次に各 CFG ルールを頂点とするサブツリーの内側確率を用いてトップダウンで部分木構造をサンプリングする。計算コストを削減するために、内側確率を計算する際には文中の各語に対して確率が上位の品詞タグのみを用いる。後の実験では、上位 5 位以内の品詞タグを利用した。[1]

4.4 同期率の高い構造の獲得

構築した確率モデルを用いて、射影されたスパンと

1　品詞タグの確率は初期状態では全て等確率とした。

フレーズラベルの制約の下で訓練データの原言語文を構文解析することで、射影されたスパンやラベルの不足を補完し、スパンの曖昧性を解消する。これによって、目的言語文の構造と同期率が高い2分木構造を獲得する。例を図1の下部に示す。

4.5 プレオーダリングモデルの学習

前節で獲得した原言語の2分木構文構造と単語アラインメントからプレオーダリングモデルを学習する。このプレオーダリングモデルは、PCFGを用いた構文解析とITGを組み合わせたモデル（ITG構文解析モデル）として構築する。

プレオーダリングモデルの訓練データは次のようにして構築する。獲得したFの2分木構造の任意の子ノードの順番を替えることで、Eの語順に最も近くなるものをF'とし、その構造を特定する。語順の近さの基準にはKendallのτを用いる。そして、Fの構造に対して、FとF'とで子ノードの順番が異なるノードのフレーズラベルには"_SW"を付与し、順番が同じノードのフレーズラベルには"_ST"を付与する。この構造はITGからの導出と考えることができる。次に、この2分木構造を用いてPCFGの学習アルゴリズムでITG構文解析モデルを学習する。学習したモデルが提案手法のプレオーダリングモデルである。この学習アルゴリズムには、隠れクラスを使う手法[16]を用いる。

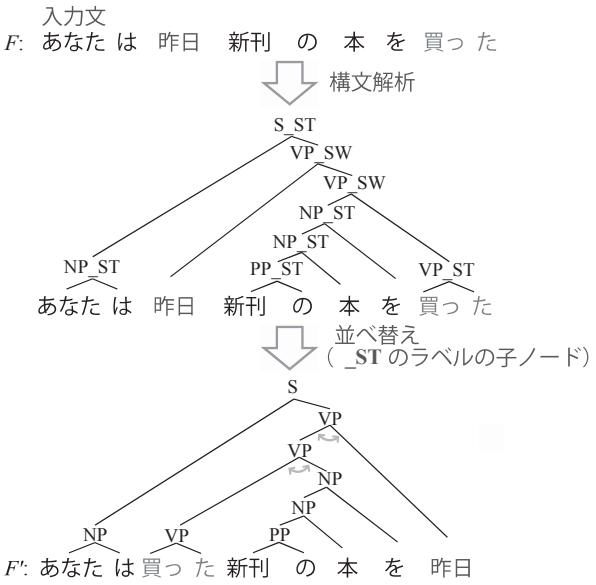

図2 入力文をプレオーダリングする例

5 プレオーダリング

入力文はITG構文解析モデルを用いてプレオーダリングする。入力文をプレオーダリングする例を図2に示す。はじめに、ITG構文解析モデルで入力文を構文解析する。このとき2分木構造と並べ替えを特定する"_SW"と"_ST"を含むフレーズラベルを決定する。そして、2分木で"_SW"が付与されたノードの子ノードの順番を変更する。

訓練データは、4.1節で射影されたスパンの制約のもとでITG構文解析モデルで原言語文を構文解析して並べ替える。

6 実験

日英と中英の特許翻訳の実験をNTCIR-9とNTCIR-10の特許機械翻訳タスク[8,17]のデータを用いて行った。

6.1 設定

NTCIR-9のデータとNTCIR-10のデータとでは、訓練データと開発データは同じで、テストデータは異なる。訓練データは日英が約318万文対で中英が100万文対である。開発データは、日英・日中それぞれ2,000文対である。テストデータは、NTCIR-9で2,000文、NTCIR-10で2,300文である。英語の構文解析器にEnju、日本語の単語分割にMeCab、中国語の単語分割にStanford segmenterを用いた。日本語の英数字は、英単語の単位に合わせて単語分割した。翻訳モデルの学習には、40単語以下の文で英語側の文が構文解析できたものを訓練データとして用いた。これは日英で約206万文対、中英で約40万文対であった。単語アラインメントはGIZA++とgrow-diag-final-andヒューリスティックおよび誤りを低減させる前後処理（英語の冠詞と日本語助詞「が」「を」「は」をアラインメント推定時に削除）[18]により獲得した。訓練データの目的言語文を用いて5-gramの言語モデルを学習した。

提案手法（Proposed）は次のように学習した。4.2

節の同期率の高い対訳文の選択では、上位 10 万文を選択した。4.3 節の確率モデルの学習では、$|T|$=50 とし、サンプリングをデータ全体に対して 100 回行った。Berkeley parser [16] をプレオーダリングモデルの学習および構文解析に用いた。翻訳にはフレーズベース SMT の Moses[19] を用い、distortion limit の設定値を標準設定の 6 とした。

比較手法として、次の 6 つの手法を用いた。

- フレーズベース SMT ＋語彙化語順推定モデル（$PBMT_L$）[19]
- 階層フレーズベース SMT（HPBMT）[4]
- String-to-tree 構文ベース SMT（SBMT）[5]
- フレーズベース SMT ＋単語列ラベリングに基づく語順推定モデル（$PBMT_D$）[18]
- 原言語の依存構造解析器を用いたプレオーダリング（SRCDEP）[20]
- 構文解析器不要のプレオーダリング（LADER）[9]

$PBMT_D$ は Moses 互換のデコーダーを用い、他は Moses を用いて翻訳した。$PBMT_L$ の語順推定モデルの学習には翻訳モデルの訓練データを全て用いた。$PBMT_D$ の語順推定モデルの学習には 20 万文を用いた。SRCDEP で利用する依存構造解析には CaboCha（日本語）と Stanford parser & tagger（中国語）を用いた。CaboCha の出力は単語の依存構造に変換して利用した。SRCDEP の並べ替えルールの学習には翻訳モデルの訓練データを全て用いた。LADER のプレオーダリングモデルの学習には 4.2 節の手法で選択した 10 万文の訓練データ（すなわち Proposed と同じ訓練データ）を用いて 100 回の繰り返し計算を行った。HPBMT と SBMT の max-chart-span の設定は無制限とした。他の手法のフレーズベース SMT での distortion limitの設定値はシステム名の添え字で示す。

表 1 日英翻訳の評価結果

	NTCIR-9		NTCIR-10	
	RIBES	BLEU	RIBES	BLEU
$PBMT_{L-4}$	65.48	26.73	65.53	27.44
$PBMT_{L-20}$	68.79	30.92	68.30	31.07
HPBMT	70.11	30.29	69.69	30.77
SBMT	72.54	31.94	71.32	32.40
$PBMT_{D-20}$	73.54	33.14	72.23	33.87
$SRCDEP_{-6}$	71.88	29.23	71.20	29.40
$LADER_{-6}$	74.31	32.98	73.98	33.90
$Proposed_{-6}$	76.35	33.83	75.81	34.90

表 2 中英翻訳の評価結果

	NTCIR-9		NTCIR-10	
	RIBES	BLEU	RIBES	BLEU
$PBMT_{L-4}$	75.02	29.22	74.24	30.65
$PBMT_{L-10}$	76.11	31.20	75.41	32.34
HPBMT	77.68	32.39	77.45	33.61
SBMT	78.44	32.47	77.68	33.90
$PBMT_{D-10}$	77.98	33.03	77.48	34.28
$SRCDEP_{-6}$	76.88	28.85	76.14	29.36
$LADER_{-6}$	78.18	30.80	77.06	31.12
$Proposed_{-6}$	81.61	35.16	81.05	36.22

表 3 提案手法の対訳文選択の効果

	NTCIR-9		NTCIR-10	
	RIBES	BLEU	RIBES	BLEU
LADER（提案法不適用）	72.33	32.30	70.96	33.07
LADER（提案法適用）	74.31	32.98	73.98	33.90

6.2 結果

評価は、自動評価の BLEU-4 と RIBES v1.01 で行った。評価結果を表 1 と表 2 に示す。日英翻訳および中英翻訳のいずれにおいても、NTCIR-9 と NTCIR-10 のデータで、提案手法は比較した手法より高い RIBES および BLEU のスコアが得られた。これによって、提案手法の有効性が確認された。SRCDEP の並べ替えルールは構造の一部のみを考慮するが、提案手法は文全体を考慮して並べ替える。提案手法は LADER では用いない構文を用いる。これらの違いの有効性が確認された。

さらに、4.2 節で提案した、構造の射影に基づく構造の同期率が高い対訳文選択の効果について検証する。提案手法は同期率が高い対訳文の選択が必要不可欠であるため、LADER を用いて日英翻訳の比較実験を行った。次の 2 つの条件で学習した LADER の結果を比較する。（1）提案手法の対訳文選択を適用しなかった訓練データ 10 万文で学習した LADER(提案法不適用) と（2）提案手法の対訳文選択を適用して選択した訓練データ 10 万文で学習した LADER(提案法適用) である。結果を表 3 に示す。RIBES および BLEU のスコアは、LADER(提案法適用) のほうが LADER(提案法不適用) より高い。これにより、提案手法である構造の同期率が高い対訳文の選択は、BTG の効果的な学習に効果があることが確認された。

7 まとめ

目的言語の構文解析器を用いた機械翻訳のプレオーダリング手法を提案した。提案手法は、原言語の構文解析器を必要とせずに構文構造を利用してプレオーダリングすることができる。また、提案手法は、言語間の部分構造の射影により、文構造の同期率が高い対訳文対の選択と同期率が高い原言語の構文構造の構築を行う。これによって得られた同期率が高い構文構造を用いることで、ITG の学習を促進して効果的な学習を実現した。日英・日中の特許翻訳で有効性を確認した。

参考文献

[1] Isao Goto, Masao Utiyama, Eiichiro Sumita, and Sadao Kurohashi. Preordering using a target-language parser via cross-language syntactic projection for statistical machine translation, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 14(13):13:1–13:23, June 2015. DOI: http://dx.doi.org/10.1145/2699925.

[2] 後藤功雄 , 内山将夫 , 隅田英一郎 , 黒橋禎夫 . 目的言語の構文解析器を用いた機械翻訳のプレオーダリング . 言語処理学会第 21 回年次大会 (NLP2015), pages 429–432, 2015.

[3] Christoph Tillman. A unigram orientation model for statistical machine translation. HLT-NAACL 2004.

[4] David Chiang. Hierarchical phrase-based translation. CL, 33(2):201–228, 2007.

[5] Hieu Hoang, Philipp Koehn, and Adam Lopez. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. IWSLT 2009.

[6] Fei Xia and Michael McCord. Improving a statistical MT system with automatically learned rewrite patterns. Coling 2004.

[7] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. HPSG-based preprocessing for English-to-Japanese translation. ACM TALIP, 11(3):8, 2012.

[8] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the patent machine translation task at the NTCIR-9 workshop. NTCIR-9, 2011.

[9] Graham Neubig, Taro Watanabe, and Shinsuke Mori. Inducing a discriminative parser to optimize machine translation reordering. EMNLP 2012.

[10] Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. CL, 23(3):377–403, 1997.

[11] Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. Bootstrapping parsers via syntactic projection across parallel texts. Nat. Lang. Eng., 11(3):311–325, 2005.

[12] Wenbin Jiang, Qun Liu, and Yajuan Lv. Relaxed cross-lingual projection of constituent syntax. EMNLP 2011.

[13] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. Ann. Prob., 25(2), 1997.

[14] Yee Whye Teh. A bayesian interpretation of interpolated Kneser-Ney. NUS School of Computing Technical Report TRA2/06, 2006.

[15] Mark Johnson, Thomas Griffiths, and Sharon Goldwater. Bayesian inference for PCFGs via Markov chain Monte Carlo. NAACL 2007.

[16] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. ACL 2006.

[17] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the patent machine translation task at the NTCIR-10 workshop. NTCIR-10, 2013.

[18] Isao Goto, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. Distortion model based on word sequence labeling for statistical machine translation. ACM TALIP, 13(1):2, 2014.

[19] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. ACL 2007.

[20] Dmitriy Genzel. Automatically learning source-side reordering rules for large scale machine translation. Coling 2010.