部分文字列に基づく機械翻訳

Graham Neubig†* 渡辺 太郎 ‡ 森 信介 † 河原 達也 †
† 京都大学 情報学研究科 ‡ 情報通信研究機構 * 日本学術振興会 特別研究員

1 はじめに

統計的機械翻訳 (SMT) は原言語の単語列 $\boldsymbol{f}_1^J = \{f_1, \ldots, f_J\}$ を目的言語の単語列 $\boldsymbol{e}_1^I = \{e_1, \ldots, e_I\}$ へと翻訳する問題として定式化される。しかし、「単語」の定義は決して自明ではない。

例えば、明示的な単語境界のない日本語や中国語では単語分割をしなければならず、フィンランド語のような膠着語では豊富な活用や固有表現によるスパース性の問題が存在する。機械翻訳のための単語境界最適化 [4] や活用の処理 [13]、固有表現の翻字 [1] など、各現象に個別に対応するための手法は数多く提案されてきた。

これらの問題ははすべて、「単語」を翻訳の単位としていることに起因する。一方で、[15] のように機械翻訳の最小単位を「文字」とする試みもある。活用、複合語、分かち書きされない言語などをすべて 1 つの統一した枠組みで扱える可能性があるが、スペイン語とカタロニア語のような非常に近い言語対でしか従来法に匹敵する精度を実現できていない。

本研究では、文字列に基づく機械翻訳のためのアライメント法を提案する。具体的には、多対多アライメント法を文字列に適用し、部分文字列に基づくアライメントを行う。これにより、文字、単語片、単語、フレーズのような様々な単位を必要に応じて利用でき、大きく異なる言語対でも従来法と匹敵する翻訳精度を実現する。

部分文字列に基づくアライメントの効率化と精度向上のために、[12] の多対多アライメント法に対して 2 つの改善を提案する。まず、アライメントに利用される [14] のビーム探索法に A*探索で用いられるような先読み確率を導入し、高速化を図る。次に、コーパス中の部分文字列の言語間共起頻度を利用した事前確率でモデルを初期化することでアライメントの精度向上を図る。

評価実験では、2 つの翻訳タスクで部分文字列に基づく機械翻訳が従来の単語を利用したフレーズベース翻訳と同等の精度が実現できることを示す。また、提案した先読み確率と部分文字列共起頻度に基づく事前確率は有意に翻訳精度に貢献していることも確認した。最後に、人手評価を行い、部分文字列に基づく機械翻訳は分かち書きされていないテキストや活用語、複合語、固有表現を全て 1 つの枠組みで翻訳できることが分かった。

2 アライメント法

SMT 用の翻訳モデルは原言語文 $\mathcal{F}$ と目的言語文 $\mathcal{E}$ からなる並列コーパスを用いて 2 言語間のアライメント $\mathcal{A}$ を獲得することにより学習される。学習コーパス内の対訳文の原言語側と目的言語側をそれぞれ $\boldsymbol{f}_1^J$ と $\boldsymbol{e}_1^I$ で表し、各文の要素を f_j と e_i で表す。従来のフレーズに基づく機械翻訳ではこの「要素」は単語を指すのに対し、本研究の部分文字列に基づく機械翻訳では「要素」は文字を指す。一文のアライメントを $\boldsymbol{a}_1^K$ とし、原言語の部分列 $f_u, \ldots, f_v$ と目的言語の部分列 $e_s, \ldots, e_t$ の対応をスパン $a_k = \langle s, t, u, v \rangle$ で表す。

2.1 一対多アライメント

最も広く使われているアライメント法として、IBM モデル [3] や HMM モデル [16] に代表される一対多アライメントがある。これらは翻訳方向依存のモデルであり、原言語文 $\boldsymbol{f}_1^J$ とアライメント $\boldsymbol{a}_1^K$ の条件付き確率 $P(\boldsymbol{f}_1^J, \boldsymbol{a}_1^K | \boldsymbol{e}_1^I)$ を最大化するように学習される。計算効率を考慮して、原言語の単語は目的言語の単語 1 つ以下に対応する制約が入っているため、原言語の単語が目的言語の複数の単語にアライメントすることはできない。この制約はほとんどの言語対においては明らかに強すぎるが、一対多アライメントを両方向で行い、これらをヒューリスティックスで組み合わせる対処法がある [10]。

しかし、一対多アライメントは f_j という 1 つの要素にアライメントを行うために十分な情報が含まれていることが前提となっている。各要素を単語とする場合はこの前提が成立するが、アライメント単位が文字となると各文字に十分な情報が含まれておらず、アライメントが失敗することがしばしばある。

2.2 多対多アライメント

一対多の制限を設けず、任意の部分文字列 e_s^t と f_u^v の対応を許す多対多アライメント法も近年注目されている [2, 6, 12]。このため、1 つの文字に十分な情報量が含まれない場合でも、部分文字列を利用し、正確なアライメントを実現することができる。また、利用する部分文字列の長さを自動調整することで、必要に応じて、文字、単語片、単語、フレーズなどの様々な対応が学習でき、低頻度語から高頻度語の正確なアライメントを実現できることが見込まれる。

本研究では、ベイズ学習と Inversion Transduction Grammar (ITG) を用いる手法に注目する。ITG は同

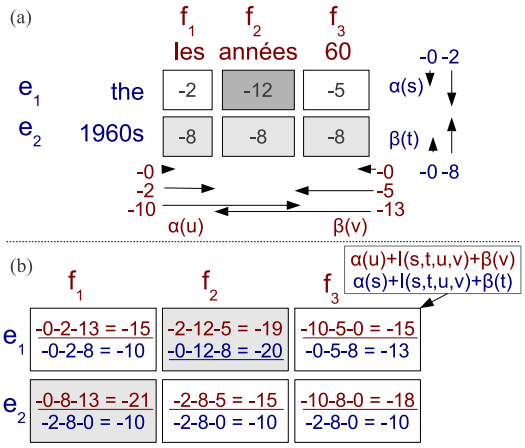

図 1: (a) 確率チャート。セルの中の数字は内側対数確率であり、矢印に付いている数字は外側確率である。(b) 先読み確率付きチャート。下線がついているのは 2 つの確率の最小値。薄い灰色と濃い灰色はそれぞれ、ビーム幅が $\log(P) \geq -3$ と $\log(P) \geq -6$ の場合に枝刈りされるスパンを示す。

期文脈自由文法の一種で、非終端記号を生成する時に単語の並べ換えを行うことが特徴である [17]。

3 先読み確率を用いた ITG 構文解析

本研究では、簡潔なモデルを学習しながら従来法と同等の精度を実現できる [12] の ITG モデルを利用する[1]。[12] の枠組みでは、コーパス中の文を 1 文ごとにボトムアップ構文解析を行い、確率チャートを作成してから、トップダウンで ITG の構文木とフレーズ対応をサンプリングする。

この確率チャートの一例を図 1(a) で示す。各言語の $\boldsymbol{e}_s^t$ と $\boldsymbol{f}_u^v$ が対応している確率を表すスパンに対して、内側確率 $I(a_{s,t,u,v})$ が付与される。この内側確率は、終端記号から $\boldsymbol{e}_s^t$ と $\boldsymbol{f}_u^v$ の組を生成する確率 $P_t(\boldsymbol{e}_s^t, \boldsymbol{f}_u^v)$ と、普通 (str) または倒置 (inv) の非終端記号を記号確率分布 $P_x(\cdot)$ から生成し、より短いスパンを再帰的に生成する確率の和からなっている。

$$
\begin{aligned}
I(a_{s,t,u,v}) = P_t(e_s^t, f_u^v) \\
+ \sum_{s \leq S \leq t} \sum_{u \leq U \leq v} P_x(\mathrm{str}) I(a_{s,S,u,U}) I(a_{S,t,U,v}) \\
+ \sum_{s \leq S \leq t} \sum_{u \leq U \leq v} P_x(\mathrm{inv}) I(a_{s,S,U,v}) I(a_{S,t,u,U})
\end{aligned}
$$

P_t と P_x は [12] の通り定義する。

この確率は文長が n の場合、$O(n^6)$ 時間の構文解析アルゴリズムですべてのスパンに対して計算可能であるが、n が長く（約 10 以上）となれば構文解析に非常に時間がかかり、学習が不可能となる。このため、ビーム探索 [14] などを用いた近似が必要となる。

[1] 文字ベース機械翻訳の場合では、高い精度を実現する上で数十文字を含む長いフレーズを利用する必要があるため、長いフレーズを含んでもメモリに格納できる簡潔なモデルを学習することがさらに重要となる。

本節では、ITG の構文解析を効率化するための先読み確率を提案する。[14] のビーム探索法では、各スパンを長さ $l = t - s + v - u$ に応じて優先度付き待ち行列に挿入し、各待ち行列を長さの昇順に処理していく。各待ち行列を処理する際に、スパンの個数 [14]、または各待ち行列の最も確率の高いスパン $\hat{a}$ からの確率幅 [12] でビームを設ける。確率幅を用いたビームの場合は、内側確率 $I(a_k)$ でビームの処理順を決め、$I(a_k) < cI(\hat{a})$ のスパン a_k を枝刈りする。c はビーム幅を表す定数であり、探索の正確さとスピードのバランスを調整できる。

$I(a_k)$ は競合しているスパンの存在を考慮しないため、待ち行列を内側確率順に処理すると探索誤りの可能性が高い。例えば図 1(a) では、“les/the” と “les/1960s” は両方とも “les” を含むため競合している。その中で比較的内側確率の低い “les/1960s” は枝刈りしても害はないと考えられる。しかし、同等の内側確率を持つ “années/1960s,” は高確率のスパンと競合していないため、枝刈りするとアライメント結果に弊害を及ぼす可能性が高い。

こういった関係を探索時に考慮するために、a_k に含まれないスパンの確率を表す外側確率 $O(a_k)$ と内側確率を組み合わせた数値で待ち行列の処理順を決めることができる。しかし、実際の外側確率 $O(a_k)$ を算出することは構文解析自体と同等の計算量を必要とするため、効率的に計算できるヒューリスティック $O^*(a_k)$ で近似する。このような手法は確率的文脈自由文法の A*探索 [7] や、単語に基づく ITG の tic-tac-toe 枝刈り [18] でも利用されるが、ここではフレーズに基づく ITG で利用可能なヒューリスティックを導入する。

具体的に、フレーズ生成確率 P_t を計算する時に、両言語のスパン e_s^t と f_u^v の最大内側確率 I^* を独立に保持する。

$$
\begin{aligned}
I_e^*(s,t) &= \max_{\{\tilde{a}=\langle \tilde{s},\tilde{t},\tilde{u},\tilde{v}\rangle; \tilde{s}=s, \tilde{t}=t\}} P_t(\tilde{a}) \\
I_f^*(u,v) &= \max_{\{\tilde{a}=\langle \tilde{s},\tilde{t},\tilde{u},\tilde{v}\rangle; \tilde{u}=u, \tilde{v}=v\}} P_t(\tilde{a})
\end{aligned}
$$

次に各言語ごとに、前向きビタビ確率 α と後ろ向きビタビ確率 β を計算する。例えば、$\alpha(s)$ は e のスパン $(0, s)$ に対して、$(0, s)$ の部分スパンを組み合わせることで得られる最大の確率と定義する。

$$
\alpha(s) = \max_{\{S_1,\ldots,S_x\}} I_e^*(0, S_1) I_e^*(S_1, S_2) \ldots I_e^*(S_x, s)
$$

後ろ向き確率 $\beta(t)$ と f_u^v に対する確率も同じように定義でき、それぞれの確率は前向き後ろ向きアルゴリズムを用いて $O(n^2)$ 時間で計算可能である。最後に、各スパンに対して、外側確率のヒューリスティックを以下のように定義する。

$$
O^*(a_{s,t,u,v}) = \min(\alpha_e(s) * \beta_e(t), \alpha_f(u) * \beta_f(v))
$$

図 1(b) は外側確率を利用したチャートを示してあり、内側確率が最大の “les/the” は “années/1960s” や

“60/1960s” との差が縮まったが、“les/1960s” はまだ比較的に確率が低い。このため、ビーム幅を狭くしても優良なスパンを枝刈りせずに、競合するスパンを枝刈りできるようになっている。

4 部分文字列共起による事前確率

前節で述べた ITG モデルはフレーズ生成確率 P_t が重要な要素となっている。ベイズ法を使った ITG アライメントでは、P_t に対して事前確率 $P_{prior}(\boldsymbol{e}_s^t, \boldsymbol{f}_u^v)$ を定義し、対応している可能性の高いフレーズ対に高い確率を与える事前分布を利用すれば学習がより少ない反復回数でより正確なアライメントに収束することが知られている [2]。本節では、単語翻訳確率を利用する従来法と部分文字列共起に基づく提案手法の 2 通りの事前確率を紹介する。

4.1 単語翻訳確率に基づく事前確率

多対多アライメントを扱う先行研究では、IBM Model 1 に基づく単語翻訳確率をフレーズアライメントの事前確率として用いる。この確率 P_{prior} を DeNero ら [6] と同じく以下のように定義する。

$$P_{prior}(\boldsymbol{e}, \boldsymbol{f}) = M_0(\boldsymbol{e}, \boldsymbol{f}) P_{pois}(|\boldsymbol{e}|; \lambda) P_{pois}(|\boldsymbol{f}|; \lambda)$$

$$M_0(\boldsymbol{e}, \boldsymbol{f}) = (P_{m1}(\boldsymbol{f}|\boldsymbol{e}) P_{uni}(\boldsymbol{e}) P_{m1}(\boldsymbol{e}|\boldsymbol{f}) P_{uni}(\boldsymbol{f}))^{\frac{1}{2}}.$$

P_{pois} は平均長パラメータ $\lambda = 0.01$ を持つポアソン分布である。P_{m1} は単語確率（文字に基づく翻訳の場合は文字確率）に基づく IBM Model 1 確率である [3]。しかし、第 2 節で述べたように、IBM Model の確率は文字に基づく翻訳の場合は情報量が不足し、正確に翻訳確率を推定することができない。

4.2 部分文字列共起に基づく事前確率

単語確率の代わりに、コーパスのすべての部分文字列を用いた言語間共起頻度に基づく事前確率を提案する。これは [5] が提案した手法と類似しているが、直接ヒューリスティックなアライメントに利用する代わりに、文内のアライメントの関係を総合的に考慮する確率モデルの初期値として利用する。

この事前確率を、$c(\boldsymbol{e})$ と $c(\boldsymbol{f})$、$c(\boldsymbol{e}, \boldsymbol{f})$ という 3 つの部分文字列頻度を用いて定義する。$c(\boldsymbol{e})$ と $c(\boldsymbol{f})$ はそれぞれ部分文字列 $\boldsymbol{e}$ と $\boldsymbol{f}$ が存在する文の数であり、$c(\boldsymbol{e}, \boldsymbol{f})$ は目的言語側に $\boldsymbol{e}$、原言語側に $\boldsymbol{f}$ が存在する文の数である。これらの数を効率的に計算・格納するために、拡張接尾辞配列を利用し、2 回以上現れる極大部分文字列のみを対象とする[2]。

これらの統計量を効率的に計算できるものの、各部分文字列対に対して $c(\boldsymbol{e}, \boldsymbol{f})$ をメモリで格納することは実用上不可能である。本手法では、まず、各頻度から定数 d=5 を引く。これには 2 つの効果がある。まず、$c(\boldsymbol{e}, \boldsymbol{f}) < d$ の共起頻度を格納する必要がなくなり、必要なメモリを節約できる。また、頻度を割引くことで、学習コーパスに対する過学習を防ぐこともできる。次に、両側の条件付き確率 $P(\boldsymbol{e}|\boldsymbol{f})$ と $P(\boldsymbol{f}|\boldsymbol{e})$ がある定数確率（ここでは 0.1）以下のものを枝刈りし、さらにメモリ領域を節約する。

各頻度に基づいて、確率 P_{cooc} を算出する。予備実験で先行研究 [5] が提案した共起頻度、ダイス係数、χ^2 係数等の組み合わせ方を試したが、本研究で新たに提案する両側の条件付き確率を組み合わせた事前確率が最も高い精度を実現することが分かった。

$$\begin{aligned} P_{cooc}(\boldsymbol{e}, \boldsymbol{f}) &= P_{cooc}(\boldsymbol{e}|\boldsymbol{f}) P_{cooc}(\boldsymbol{f}|\boldsymbol{e}) / Z \\ &= \left(\frac{c(\boldsymbol{e}, \boldsymbol{f}) - d}{c(\boldsymbol{f}) - d} \right) \left(\frac{c(\boldsymbol{e}, \boldsymbol{f}) - d}{c(\boldsymbol{e}) - d} \right) / Z \end{aligned}$$

これは $c(\boldsymbol{e}, \boldsymbol{f}) > d$ のフレーズ対のみに対して計算され、Z は以下のような正規化項である。

$$Z = \sum_{\{\boldsymbol{e}, \boldsymbol{f}; c(\boldsymbol{e}, \boldsymbol{f}) > d\}} P_{cooc}(\boldsymbol{e}|\boldsymbol{f}) P_{cooc}(\boldsymbol{f}|\boldsymbol{e})$$

d を引いた頻度を利用しているため、学習コーパス内の部分文字列対であっても $P_{cooc}(\boldsymbol{e}, \boldsymbol{f}) = 0$ になることが多い。事前確率の役割は主にモデルの初期化に利用され、アライメント仮説を完全に除外するとむしろアライメント精度が大きく下がる可能性が高いため、部分文字列に基づく事前確率を Model 1 に基づく翻訳確率との線形補間を行い、すべての部分文字列対に確率を与えることを保証する。

$$P_{prior}(\boldsymbol{e}, \boldsymbol{f}) = \lambda P_{cooc}(\boldsymbol{e}, \boldsymbol{f}) + (1 - \lambda) P_{m1}(\boldsymbol{e}, \boldsymbol{f})$$

なお、補間係数 λ にディリクレ分布 ($\alpha = 1$) に基づく事前確率を与え、学習の段階で推定する。

5 実験評価

5.1 実験設定

部分文字列に基づく機械翻訳の実現可能性を検証するために、日本語・英語 (ja-en)、フィンランド語・英語 (fi-en) の 2 言語対を用いた実験評価を行う。日本語のデータとして、京都フリー翻訳タスクの Wikipedia 翻訳データ [11] を利用し、フィンランド語のデータとして 2005 ACL Shared Task の EuroParl データ [9] を利用した。明示的な単語境界のない言語の代表として日本語を選択し、単語翻訳モデル構築にはタスク付属の Wikipedia に適応された単語分割モデルを利用した。活用が豊富な膠着語としてフィンランド語を選択し、ACL Shared Task のデータは予めトークンに分割してあるため、これ以上の処理を行わずに使用した。両タスクでは、学習データとして 100 文字以下の文を使用し、その諸元は表 1 で示す。

一対多アライメントには GIZA++[3]、多対多アライメントには pialign[4]を提案法で改良したバージョンを利用した。GIZA++の設定は主にデフォルトを利用

[2]拡張接尾辞配列の計算にオープンソースライブラリ esaxx を利用した http://code.google.com/p/esaxx/ (2011 年 6 月)。

[3]http://code.google.com/p/giza-pp/ (2011 年 6 月)。

[4]http://phontron.com/pialign/ (2011 年 6 月)。

	fi	en	ja	en
TM 学習	2.23M	3.10M	2.34M	2.13M
LM 学習	-	15.5M	-	11.5M
重み学習	42.0k	58.7k	34.4k	30.8k
テスト	41.4k	58.0k	28.5k	26.6k

表 1: 各コーパスの単語数。

	fi-en	ja-en
GIZA-word	20.41/ 60.01/ 27.89	**17.95**/ 56.47/**24.70**
ITG-word	**20.83**/ 61.04/ 28.46	17.14/ 56.60/**24.89**
GIZA-char	6.91/ 41.62/ 14.39	9.46/ 49.02/ 18.34
ITG-char	18.38/**62.44**/**28.94**	15.84/**58.41**/ 24.58

表 2: GIZA++と ITG モデルの単語 BLEU/文字 BLEU/METEOR による評価。太字は最も精度の高いシステムと統計的に有意でない差を示す ($p = 0.05$ [8])。

し、文字に基づくアライメントの高速化のために最終モデルとして HMM モデルを用いた。pialign もデフォルトを利用し、文字に基づくアライメントの場合だけビーム幅を 10^{-10} ではなく 10^{-4} とした。デコーダとして Moses[5]を利用し、スタック幅を 200 から 1000 へ変更した。重みは単語 BLEU スコアを最適化するように学習した。言語モデルとして Kneser-Ney 平滑化を利用し、単語の場合は単語 5-gram、文字の場合は文字 12-gram を用いた。

5.2 評価

各手法の単語 BLEU、文字 BLEU、METEOR による評価結果を表 2 で示す。提案手法は両タスクにおいて、文字列を翻訳する先行研究を大きく上回ることが分かる。また、単語に基づく機械翻訳より文字 BLEU で上回り、METEOR で同等の精度となり、単語 BLEU で下回り、全体的に匹敵する精度である[6]。

また、2 名の評価者が 100 文に対して 0〜5 の意味的妥当性評価を評価基準とした主観評価を行った。その結果、ITG-word と ITG-char はそれぞれ ja-en で 2.085 と 2.154、fi-en で 2.851 と 2.826 となり、いずれも有意差とならなかった。ITG-char の評価が ITG-word の評価を 2 点以上上回った原因を詳しく調べると、26 文中の 13 文は原言語側の未知語の扱いによる改善、5 文は目的言語側の未知語生成の改善、5 文は未知語でない低頻度語のアライメントの改善によるものであった。逆に ITG-word が ITG-char の評価を上回った例の多くは単語並べ換え、または中・高頻度語の語彙選択によるものであった。

[5] `http://statmt.org/moses/` (2011 年 6 月)。

[6] 仏英・独英翻訳では単語に基づく手法は提案法を上回ったため、提案法は膠着語や単語境界が曖昧な言語では効果的であると考えられる。

6 おわりに

本論文では、部分文字列のアライメントを利用することで、「単語」という概念を用いずに従来のモデルと同等の精度を実現しながら、未知語と低頻度語によるスパース性に対応できることを示した。ここでは、文中の空白を通常の文字と同様に扱ったが、これからの課題として空白の存在を利用し、アライメントやデコーディングの効率化を目指す。

参考文献

[1] Y. Al-Onaizan and K. Knight. Translating named entities using monolingual and bilingual resources. In *Proc. ACL*, 2002.

[2] P. Blunsom, T. Cohn, C. Dyer, and M. Osborne. A Gibbs sampler for phrasal synchronous grammar induction. In *Proc. ACL*, 2009.

[3] P. F. Brown, V. J. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19, 1993.

[4] P.-C. Chang, M. Galley, and C. D. Manning. Optimizing Chinese word segmentation for machine translation performance. In *Proc. WMT*, 2008.

[5] F. Cromieres. Sub-sentential alignment using substring co-occurrence counts. In *Proc. COLING/ACL 2006 Student Research Workshop*, 2006.

[6] J. DeNero, A. Bouchard-Côté, and D. Klein. Sampling alignment structure under a Bayesian translation model. In *Proc. EMNLP*, 2008.

[7] D. Klein and C. D. Manning. A* parsing: fast exact viterbi parse selection. In *Proc. HLT*, 2003.

[8] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*, 2004.

[9] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, 2005.

[10] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. HLT*, 2003.

[11] G. Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[12] G. Neubig, T. Watanabe, E. Sumita, S. Mori, and T. Kawahara. An unsupervised model for joint phrase alignment and extraction. In *Proc. ACL*, pp. 632–641, Portland, USA, June 2011.

[13] S. Nießen and H. Ney. Improving SMT quality with morpho-syntactic analysis. In *Proc. COLING*, 2000.

[14] M. Saers, J. Nivre, and D. Wu. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *Proc. IWPT*, 2009.

[15] D. Vilar, J.-T. Peter, and H. Ney. Can we translate letters. In *Proc. WMT*, 2007.

[16] S. Vogel, H. Ney, and C. Tillmann. HMM-based word alignment in statistical translation. In *Proc. COLING*, 1996.

[17] D. Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3), 1997.

[18] H. Zhang and D. Gildea. Stochastic lexicalized inversion transduction grammar for alignment. In *Proc. ACL*, 2005.