識別言語モデルによる機械翻訳システムの誤り分析とシステム間比較

赤部 晃一　Graham Neubig　Sakriani Sakti　戸田 智基　中村 哲
奈良先端科学技術大学院大学　情報科学研究科
{akabe.koichi.zx8, neubig, ssakti, tomoki, s-nakamura}@is.naist.jp

1　はじめに

最先端の機械翻訳システムは年々精度が向上しているが、その反面システムの内部は複雑化している。この複雑化の結果、システムの改良が翻訳結果に与える影響は必ずしも事前に推測できるわけではなく、翻訳結果を用いた評価実験によって改善すべき点を見出すことが広く行われている。この際、翻訳結果を一文ずつ見ることで誤りを見つけることもできるが、その作業は非常に労力がかかるものであり、見つけた誤りがシステム全体を大きく改善できるとも限らない。しかし、誤り箇所の検出や、誤りに対する重要度の付与を自動的に行えれば、システムの不得意な翻訳現象を容易に発見でき、分析作業の効率化を図れる。

先行研究では、この自動化の第一歩として誤りを様々な情報で分類し、それぞれの誤りを頻度順に並べる手法が提案されている [1, 2]。翻訳結果と出力文を編集距離でアライメントし、活用、並べ替え、単語脱落・挿入、語彙選択などの誤り傾向を頻度付きで抽出する仕組みである。結果一つ一つを眺める前におおよその流れを定量化できるという意味で、有用な方法である。

しかし、誤りを単純に頻度で選択すると、誤りのリストが翻訳先言語に頻繁に出現するものに支配されるという問題がある。例えば、京都フリー翻訳タスク [3] のデータで学習した翻訳モデルに対して、出力文に含まれながら参照文に含まれない誤った n-gram を頻度順に並べたものを表 1 に示す。表を見て分かるとおり、頻繁に誤って出力される n-gram は、単純に頻繁にテキストに含まれる文字列であり、これだけではシステムの特徴的な誤りを捉えることができない。

本研究では、システムの特徴的な誤りを特定する道具として、正則化された識別言語モデル [4] を利用する手法を提案する。識別言語モデルは、自然な出力言語文の特徴を捉えるように学習される通常の言語モデルとは異なり、ある特定のシステムについて、起こりやすい出力誤りを修正するように学習される。誤り分析の観点から見ると、モデルによって学習された修正パターンに目を通せば、システムの特徴的な誤りを発見できると考えられる。更に、L1 正則化 [5] を学習時に適用することで殆どの重みが 0 になるようにすれば、重要な修正パターンが更に顕著に現れる。

実験では、英日機械翻訳を題材に、識別言語モデルを用いた誤り分析の有効性について検証した。従来法と

表 1　機械翻訳で頻繁に起こる誤り

1-gram		2-gram		3-gram	
の	75	た 。	35	た 。 (終端)	35
に	60	る 。	27	る 。 (終端)	26
、	56	大学 (終端)	26	あ る 。	22
た	56	あ る	26	で あ る	18
が	54	年 （	26	さ れ た	16

して頻度の高い誤り n-gram を、提案法として識別言語モデルによって重み付けされた n-gram を評価者に提示し、それぞれの n-gram がシステムの誤りを正しく特定しているかどうかを評価した。その結果、機械翻訳システムの誤り分析において、提案法の有効性が示された。

2　識別言語モデル

本節では、実験に用いる識別言語モデルについて説明する。

ある入力文のコーパス $\mathcal{F} = \{F_1, \ldots, F_K\}$ に対して、システム出力の n-best $\mathcal{E} = \{\boldsymbol{E}_1, \ldots, \boldsymbol{E}_K\}$ が与えられたとする。識別言語モデルでは、$\boldsymbol{E}_k = \{E_1, E_2, \cdots, E_I\}$ の中の候補に対して素性関数 $\boldsymbol{\phi}(E_i)$ を定義し、素性関数と重みベクトル $\boldsymbol{w}$ の内積 $\boldsymbol{w} \cdot \boldsymbol{\phi}(E_i)$ をスコアとして定義する。

各 n-best の中の候補を参照文と比較し、自動評価尺度のスコアを計算する。このスコアが最大となる文をオラクル E^* として選択し、E^* のスコアが n-best 中で高くなるように素性の重み $\boldsymbol{w}$ を学習する。

2.1　構造化パーセプトロンによる識別言語モデル

識別言語モデルの学習は構造学習の一種である。先行研究では、構造学習の最も単純な手法である構造化パーセプトロンを、識別言語モデルの学習において有用な手法であると示している [6]。構造化パーセプトロンでは、候補中の参照文 E^* と、モデルによって最も大きなスコアが振られる仮説 $\hat{E}$ の素性列を比較して更新を行う。

1 回の更新において、E^* と $\hat{E}$ の差分を用いて $\boldsymbol{w}$ を更新する。$\hat{E}$ と E^* が等しいときは差分が $\mathbf{0}$ のため更新を行わない。重みの更新は $\mathcal{F}$ 全体に対して一文ごとに逐次的に行い、反復回数や重みの収束といった終了条件が満たされるまで反復する [4]。

2.2 L1 正則化による素性選択

本研究では、機械翻訳システムの誤り傾向をより明確にするため、重みの学習時に L1 正則化を行う。L1 正則化は、重みベクトルに対して L1 ノルム $\|\boldsymbol{w}\|_1 = \sum_i |w_i|$ に比例するペナルティを与える。L1 正則化を用いる時に、重み列 $\boldsymbol{w}$ の中で多くの素性が 0 となるため、識別能力に大きな影響を与えない素性をモデルから削除することが可能となる。

L1 正則化された識別モデルを学習する簡単かつ効率的な方法として、前向き後ろ向き分割 (forward-backward splitting; FOBOS) アルゴリズムがある [7]。一般的なパーセプトロンでは正則化を重みの更新時に行うが、FOBOS では重みの更新と正則化の処理を分割し、重みの利用時に前回からの正則化分をまとめて計算し、効率化を図る。

2.3 識別言語モデルの素性

識別言語モデルの素性として様々な情報を利用できるが、本研究では簡単のため、以下の 3 種類の素性を利用する[*1]：

1. 翻訳仮説を生成したシステムのスコア ϕ_s：システム出力を修正するように学習するため、学習の初期においてシステムスコアによる順位付けが必要である。
2. 翻訳仮説に含まれる n-gram の頻度 ϕ_n：n-gram に対して重み付けをすることで、システムが出力する誤った n-gram を捉える。
3. 翻訳仮説の単語数 ϕ_l：翻訳システムが利用する評価尺度に単語数に対する罰則がある場合、単語数を調整するのに用いられる。

3 識別言語モデルを用いた誤り分析

本節では、提案手法である識別言語モデルを用いた誤り分析について述べる。まず、単一システムの誤り箇所の特定における手法を説明し、次に複数のシステムを比較した際の、それぞれのシステムの不得意分野を分析する手法について述べる。

3.1 識別言語モデルを用いた誤り箇所の特定・分析

識別言語モデルの中で、大きな重みが学習されている n-gram 素性は、「システムが十分に生成できていない n-gram」と捉えられる。逆に、低い重みが学習されている n-gram 素性は「システムが誤って生成する傾向にある n-gram」と捉えられる。従って、n-gram 素性を重みの順に並べ替えて、特に大きい、または小さい重みが学習されたものを人手で検証することで、システムの誤り傾向を把握できると考えられる。

実際に評価を行う場合、注目対象となる n-gram がシステム出力に含まれていた方が分析しやすい。このため本研究では、特に低い重みが学習された n-gram を評価対象とする。システムの特徴的な誤りを捉えている手法では、低い重みが学習された n-gram を見た際、それが実際の誤り箇所を捉えている割合が大きくなる。

代表的な n-gram の選び方として、頻度に基づく従来法と提案法をそれぞれ用いる:

従来法： システム出力に含まれながら参照文に含まれない n-gram を頻度順に並べたものを代表的な誤り n-gram とする。

提案法： 各 n-gram を識別言語モデルによって学習された重みの順に並べ替えて、その重みが著しく低いものを代表的な誤り n-gram とする。

実際に誤り分析を行う際、この代表的な誤り n-gram が含まれるシステム出力を中心に調査していく。詳細は分析の目標によるが、本研究の評価に用いた一例を 4.3 節で詳しく述べる。

3.2 システム間の比較

システム開発において誤り分析を行う際、単一のシステムの傾向をつかむだけではなく、システム間の比較が用いられることも多い。

本研究では、識別言語モデルによって重み付けされた代表的な n-gram が、実際の誤り傾向を適切に捉えているかどうかを確認するために、まず複数の翻訳システムが出力した n-best から識別言語モデルを学習する。その際に特に低い重みが付いた代表的な誤り n-gram を誤りの種類ごとに分類し、各システムの誤り傾向を比較する。この結果を先行研究で報告されている誤り傾向と比較し、一致するかどうかを検証する。

4 評価実験

提案手法の有効性を検証するために、機械翻訳における実験を行った。

4.1 実験設定

京都フリー翻訳タスク (KFTT) のデータ (表 2) を用いて英日翻訳システムを構築し、それらが出力した 500-best リストを対象とした。

単一システムにおける実験では、システムとして Travatar[8] に基づいて構築された forest-to-string モデル (F2S) を用いた。単語アラインメントは、構文情報を用いる Nile[*2]で行い、英日それぞれの構文解析

表 2　KFTT のデータサイズ

	記事数	文数	単語数	
			英語	日本語
学習	14126	330k	5.91M	6.09M
dev セット	15	1166	24.3k	26.8k
test セット	15	1160	26.7k	28.5k

[*1] 本研究の対象にしないが、単語 n-gram 以外に品詞や統語情報を用いた素性を加えれば、[1] で述べられているような統語的な誤りパターンも捉えることができる。

[*2] http://code.google.com/p/nile/

表 3　各システムにおける、識別言語モデルを用いる場合と用いない場合の翻訳精度

システム	RIBES(dev)		RIBES(test)	
	学習前	学習後	学習前	学習後
PBMT	0.6703	0.7821	0.6853	0.6796
HIERO	0.6867	0.7744	0.6907	0.6973
F2S	0.7229	0.7699	0.7379	0.7394

に Egret[*3]を利用した。システム間比較では F2S に加え、Moses[9] に基づいて構築されたフレーズベースモデル (PBMT) と階層的フレーズベースモデル (HIERO) を用いた。Moses の学習では、単語アラインメントに GIZA++[10] を利用した。各モデルとも、英日翻訳で人間の評価と高い相関を示す RIBES[11] を評価尺度として最適化されたものを用いた。3 種類のシステムにおいて、断りのない限りデフォルトの設定を用いた。

識別言語モデルの学習に 2.2 節で述べた FOBOS を利用した。反復回数を 100 回とし、n-gram 素性の最大長を 3 とした。評価尺度として、RIBES を利用した。学習には、KFTT の dev セットを用い、正則化係数が 5.0×10^{-6} から 1.0×10^{-3} の範囲で test セットの RIBES が最大となるようにモデルを学習した。

4.2　識別言語モデルの修正能力

まず、識別言語モデルが実際に有用な修正パターンを学習できているかどうかを定量的に測るために、dev セット、及び test セットのモデル適用後の RIBES を測った。3 システムにおける結果を表 3 に示す。正則化係数は、PBMT は 1.5×10^{-5}、HIERO は 5×10^{-5}、F2S は 1.5×10^{-4} とした。

表を見ると、dev セットは精度が大きく向上しているが、test セットは各システムとも精度が大きく変化していないことが分かる。この原因として、学習の主な素性を n-gram に限定しており、学習データも比較的少ないことが挙げられる。しかし、本研究の目標である誤り分析において、未知データに対する修正能力は本質的には重要ではなく、次節以降、誤り分析における有用性を検証する。

4.3　誤りの特定・分析における識別言語モデルの効果

本節では、低い重みが学習された素性が、システムの特徴的な誤りを示しているかどうかを検証する。検証の手順を次に示す。

1. 従来法と提案法によって、代表的な n-gram を順に 100 個選択する。
2. 選択された n-gram が含まれているシステム出力文を、dev セットから最大で 10 文ずつランダムに選択する。
3. 選択された文を、選択要因となった n-gram とともに評価者に提示する。

[*3] http://code.google.com/p/egret-parser/

En	rinzai school in china
Ja(Ref)	中国 に おける 臨済 宗
Ja(MT)	中国 に [お い て] 臨済 学校
Rules	pp (x0:in x1:np) → x1 x0 in (”in”) → ”に” ”お” ”い” ”て”
Eval	文脈依存置換誤り
En	belonged to the tenryu-ji sect until 1905 .
Ja(Ref)	明治 38 年 （ 1905 年 ） まで は 天竜 寺 派 に 属 し た 。
Ja(MT)	1905 年 （ 明治 38 年 まで 天龍 寺 宗 に [属 し て] い た 。
Rules	s (vp (x0:vbd x1:vp’) . (”.”)) → x1 x0 ”。” vbd (”belonged”) → ”属” ”し” ”て” ”い” ”た”
Eval	活用誤り

図 1　実際の評価例

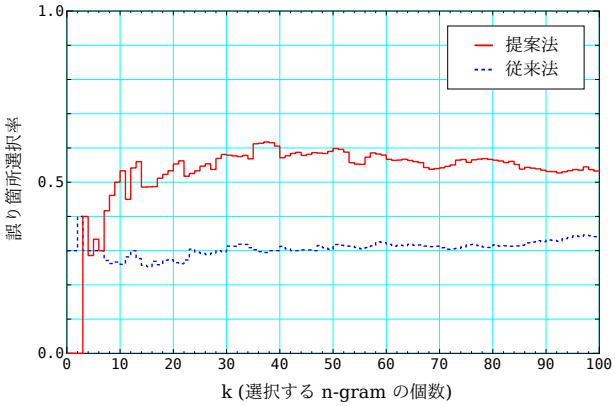

図 2　代表的な n-gram の誤り箇所選択率

4. 評価者は、文中に示された n-gram が誤っているかどうか、誤りの種類とともに記録する。

評価は英語と日本語に熟達している機械翻訳の専門家 1 名に行ってもらった。この際、両手法を公平に分析できるようにするため、評価文の順序を手法に依存しないように並べ替えた。実際にシートに記録された誤りの例を図 1 に示す。

各手法における代表的な n-gram の内、システム出力に含まれるものを上位 k 個選択した場合に、それらの n-gram が実際に誤り箇所を捉える割合を図 2 に示す。図を見ると、従来法では代表的な n-gram の多くが問題のない箇所を選択していることが分かる。一方提案法では、上位 30 個程度の代表的な n-gram を選択した際に、従来法に比べて誤り箇所を高い比率で捉えることが分かる。

次に、代表的な n-gram を各手法により上位 30 個選択した時、n-gram が選択する箇所の誤りの内訳を表 4 に示す。この結果から、殆どの種類の誤りに対して、提案法が従来法に比べて高精度で捉えていることが確認できる。特に並べ換え誤りの差が大きいが、この理由として並べ換え誤りに着目する RIBES を評価尺度として利用していることが挙げられる。

表 4　上位 30 個の代表的な n-gram の内訳

誤りの種類	従来法	提案法
誤りでない	0.69	0.42
活用誤り	0.04	0.01
否定・肯定誤り	0.00	0.04
未知語	0.00	0.00
並べ換え誤り	0.01	0.09
挿入誤り	0.02	0.03
削除誤り	0.08	0.22
文脈依存置換誤り	0.07	0.11
文脈非依存置換誤り	0.08	0.08

表 5　学習によって低い重みが付与された n-gram

1-gram		2-gram		3-gram	
化	−5.565	、 幕府	−7.604	氏 は 「	−8.615
伝え	−5.541	職 を	−7.596	は 鎌倉 時代	−8.565
者	−4.598	皇太子 と	−6.554	よう に 。	−8.512
作家	−4.542	に 『	−6.551	職 を 失	−7.596
慶長	−4.531	し な	−6.528	、 その 後	−7.546

識別言語モデルの学習によって低い重みが付与された n-gram を表 5 に示す。表を見ると、単純に頻度順に並べた場合 (表 1) に比べ、具体的な誤り箇所の特定に役立つ名詞や動詞なども多く含まれていることが分かる。

4.4　システム間の比較結果

PBMT と HIERO に対して前節と同様に代表的な n-gram を上位 30 個選択し、誤り分析を行った。この結果を F2S と合わせて表 6 に示す。表を見ると、PBMT と HIERO では、F2S に比べて活用誤りや並べ換え誤りが頻繁に発生していることが分かる。この理由として、フレーズベースの翻訳システムが大域的な文の構造を翻訳に利用しないシステムであるのに対し、F2S は品詞情報や構文情報といった統語情報を利用した翻訳システムであり、並べ替えや活用の変化などに頑健であることが挙げられる。このような傾向は [12] などの先行研究でも指摘されており、今回の提案手法で選択された誤り箇所が、機械翻訳システムの誤り傾向をつかむのに有用であると考えられる。

5　まとめ・今後の課題

本研究で、識別言語モデルの誤り分析の道具としての可能性について調べた。その結果、単純に頻度の高い誤り n-gram を調べるよりも、識別言語モデルによって学習された重みを用いた方が、システムの特徴を捉える上で有効であると示した。

今後、提案法に基づいて翻訳モデルを修正した場合に、翻訳精度にどのように影響を与えるか検証する。

表 6　3 つのシステムにおける誤りの内訳。太字は最も良いシステムより 0.05 以上悪い精度を示す。

誤りの種類	PBMT	HIERO	F2S
誤りでない	0.25	0.44	0.42
活用誤り	**0.07**	**0.09**	0.01
否定・肯定誤り	0.00	0.02	0.04
未知語	0.00	0.00	0.00
並べ換え誤り	**0.19**	**0.24**	0.09
挿入誤り	0.02	0.01	0.03
削除誤り	0.15	0.11	**0.22**
文脈依存置換誤り	0.09	0.08	0.11
文脈非依存置換誤り	**0.23**	0.02	**0.08**

参考文献

[1] David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error analysis of statistical machine translation output. In *Proc. LREC*, pages 697–702, 2006.

[2] Maja Popovic and Hermann Ney. Towards automatic error analysis of machine translation output. In *Computational Linguistics*, pages 657–688, 2011.

[3] Graham Neubig. The Kyoto free translation task. `http://www.phontron.com/kftt`, 2011.

[4] Brian Roark, Murat Saraclar, and Michael Collins. Discriminative n-gram language modeling. *Computer Speech & Language*, 21(2):373–392, 2007.

[5] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, pages 267–288, 1996.

[6] Michael Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proc. EMNLP*, pages 1–8, 2002.

[7] John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. In *Journal of Machine Learning Research*, volume 10, 2009.

[8] Graham Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proceedings of the ACL Demonstration Track*, Sofia, Bulgaria, August 2013.

[9] 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. In *Proc. ACL*, pages 177–180, Prague, Czech Republic, 2007.

[10] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[11] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proc. EMNLP*, pages 944–952, 2010.

[12] 丹生 伊左夫, Graham Neubig, 小林 和也, Sakriani Sakti, 戸田 智基, and 中村 哲. 構文情報が機械翻訳に及ぼす影響の分析. In **情報処理学会 第** *212* **回自然言語処理研究会** *(SIG-NL)*, 北海道, 7 2013.