統計的機械翻訳における 交差エントロピーを用いたパラメータ推定の検討

川部友大　　ルパージュ・イヴ
早稲田大学大学院 情報生産システム研究科
kawabe413@.akane.waseda.jp, yves.lepage@waseda.jp

1 はじめに

統計的機械翻訳 (SMT) における対数線形モデルは、素性スコアの対数値と寄与率により与えられる。各素性モデルは訓練過程によって学習され、各寄与率はチューニングによって推定される。しかし、開発データにおける翻訳損失最小化を目的とするバッチ学習によるチューニングでは、過学習により最適なパラメータを学習できないことがあり、特に Powell 法等の勾配を用いない手法によりパラメータを推定するエラー最小化学習法 (MERT)[3] では、パラメータ推定が局所解となる傾向がある [1] だけではなく、推定時間にもばらつきが生じる。

本論文ではパラメータ推定における過学習は、バッチ学習において各素性モデルの品質が考慮されないことにより生じると推測する。ゆえに本研究の目的は、テスト時における翻訳品質の向上とパラメータ推定時間の短縮である。本実験では、初めに Europarl コーパス[1] に収録されている 11 言語を用いて、素性モデルの品質とチューニングにおける翻訳品質の相関分析により翻訳品質に影響を与える素性を特定する。次に特定した素性モデルを用いて翻訳品質と素性モデルの品質に関するモデルを構築し、パラメータ推定を行う。素性モデルの品質はモデルにおける訓練データの学習度合を表す交差エントロピー、翻訳品質を示す指標として BLEU スコアを用いる。

2 関連研究

関連研究として、Guzman と Vogel による研究 [2] が挙げられる。当研究では SMT におけるフレーズ翻訳モデル等の素性スコアや条件付きエントロピーなどを用いて、素性モデルが翻訳品質に与える影響を評価した。また素性モデルの最適化の研究では、開発データが翻訳に与える影響を分析した研究が数多く知られている。例として、Pecina ら [4] の研究では、チューニングにおいてドメインデータが翻訳品質に与える影響を分析した。

しかしながら、モデル品質が翻訳品質に与える影響を分析することでパラメータの推定に寄与させる、すなわち素性モデルの品質データは翻訳品質を向上させる情報資源となりえるかを分析するという研究は未だかつて存在しないという点で本研究の新規性がある。そして、数多くの言語対の翻訳で用いられた素性モデルとチューニングにおける翻訳精度を解析することにより、SMT の翻訳モデルの本質を推定を行っている点で本研究の有効性がある。

3 交差エントロピーによる各素性モデルの解析

本章では、訓練過程における各素性モデルの品質が翻訳品質に与える影響を推定する。

交差エントロピー H_{cross} とは、ある事象が真の確率分布ではなく別の確率分布に従う際に、事象を特定するために必要となる平均ビット数である。このため機械学習では訓練データから学習することに得られたモデルが訓練データをどの程度学習しているかを示す指標となっており、言語モデルの品質を評価する指標であるパープレキシティ[5] にも交差エントロピーが用いられている。本実験では各素性モデルにおける交差エントロピーとチューニングにより得られた BLEU スコアを用いることにより相関分析を行い、翻訳品質に与える影響を推定する。翻訳実験を行うにあたり、コーパスは Europarl コーパス ver.3 に収録される 11 言語、翻訳エンジンは Moses.ver.3.0[2]、言語モデルは Kenlm[3]、単語アライメントには Fast-align[4]、チューニング手法は MERT を用いた。コーパスの詳細につ

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

[2] http://www.statmt.org/moses/
[3] https://kheafield.com/code/kenlm/
[4] http://www.cdec-decoder.org/guide/fast_align.html

データ	文数
トレーニング	347,614
チューニング	500
テスト	38,123

表 1: 実験で使用したデータセットの統計

いては表 1 に記載する。また、相関分析にはピアソンの積率相関係数を用いた。

Moses.ver.3.0 で用いられる 14 素性全ての実験結果を表 2 に示す。この実験結果から、SMT において必要条件とされるモデル (太字で記載) では統計的に顕著な負の相関を示すことが分かり、必要条件のモデルの品質が翻訳品質に影響を及ぼすと推測する。

	Features	ρ
Translation models	P($S\|T$),PTS	-0.850
	P($T\|S$),PST	-0.590
	Lex($S\|T$),LTS	-0.786
	Lex($T\|S$),LST	-0.673
Reordering models	Pb($M\|(S,T)$),PbM	-0.794
	Pb($S\|(S,T)$),PbS	0.550
	Pb($D\|(S,T)$),PbD	0.644
	Pi($M\|(S,T)$),PiM	-0.798
	Pi($S\|(S,T)$),PiS	0.552
	Pi($D\|(S,T)$),PiD	0.629
Language model	LM(T),LM	-0.755
Constraint	Dl($T\|S$),Dl	-0.680
	Wp($T\|S$)),Wp	-0.083
	Pp($T\|S$)),Pp	0.206

表 2: 各 14 素性モデルにおける交差エントロピーとチューニングにおける BLEU スコアとのピアソンの積率相関係数 ρ 相関分析の結果。本実験ではピアソンの積率相関係数を用いて相関を数値化している。この結果から言語モデル、翻訳モデルなどの SMT での翻訳の必要条件となる素性モデルにおいて顕著な相関を示すことが分かる。

4 スコアに基づく交差エントロピー

本章では、訓練過程におけるモデル全体の品質が翻訳品質に与える影響を相関分析により推定する。このため、原言語における任意のフレーズが目的言語フレーズに翻訳されるために必要となる情報量の平均とチューニングにおける翻訳品質との相関を数値化することが本章の目的である。

本実験では、任意の目的言語へ翻訳する際にチューニング $Tuning$ において各素性モデル h_i から得られた素性スコアの対数値の一単語ごとの平均、すなわち単語におけるエントロピーを重みとすることで、任意のフレーズを翻訳する際に必要となる訓練モデル全体の品質を近似している。これをスコアに基づく交差エントロピー (SBCH) と定義する。スコアに基づく交差エントロピー H_{SBCH} の定義式を以下に示す。

$$H_{SBCH}(T|S) = |Tuning| \sum_{i \in Selection} H(\hat{h}_i) \times H_{cross}(h_i(T|S)) \quad (1)$$

$$H(\hat{h}_i) = \frac{\log_2 10 \times \sum_{j \in Tuning} \log h_{ij}}{|Tuning|} \quad (2)$$

$Selection$ は SMT で用いられている全素性モデルの部分集合である。また、$|Tuning|$ は開発データの単語総数である。本実験では第 3 章と同様の条件の下で、素性モデルの組み合わせ 2^{14} 通り全てにおいて相関分析を行った。この結果 (表 3 参照) から、素性モデル毎に相関が高い組み合わせだけではなく、素性モデル単体では翻訳品質との相関が低いものも含めた素性の組み合わせが最も高い相関を示すことが分かった。この結果から、モデルの品質が翻訳品質との相関が低い素性モデルにも関わらず、翻訳品質の良し悪しを分類することができるモデルがあることが推測される。

5 交差エントロピー混合型対数線形モデル

本章ではスコアに基づく交差エントロピーと対数線形モデルを組み合わせるたモデルを用いることにより、パラメータ再推定を行う。すなわち素性モデルの品質を考慮したパラメータ推定を行うことが本章の目的である。

スコアに基づく交差エントロピーと対数線形モデルの総和である、交差エントロピー混合型対数線形モデルを以下に示す。

$$\arg\max(-|Testing| \times \sum_{i=1}^{14} (\alpha H_{cross}(h_i(T|S)) + \beta w_i) H(h_i)) \quad (3)$$

$|Testing|$ はテストデータの総単語数である。対数線形モデルは各素性におけるエントロピーの総和が最小となる翻訳文を出力するモデルであると解釈できるため、交差エントロピーを直接用いることで、素性寄与率に対するモデルの品質が与える影響を分析するモデルがこのモデルの意味である。α と β を推定するために、重回帰分析によるパラメータ推定、もしくは MERT を用いたパラメータ推定の二つを用いて実験を行う。パラメータ推定では、任意の素性モデルの組み合わせにおける α と β の最適化を行う。これを混

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
PbM										○	○	○	○	○
PbS							○	○	○	○	○	○	○	○
PbD					○	○	○	○	○	○	○	○	○	○
PiM						○	○	○	○	○	○	○	○	○
PiS											○	○	○	○
PiD						○	○	○	○	○	○	○	○	○
LM			○	○	○	○	○	○	○	○	○	○	○	○
PST	○	○	○	○	○	○	○	○	○	○	○	○	○	○
LST													○	○
PTS		○	○	○	○	○	○	○	○	○	○	○	○	○
LTS													○	○
Dl									○	○	○	○	○	○
Wp												○		○
Pp				○	○	○	○	○	○	○	○	○	○	○
$-\rho$	0.795	0.847	0.900	0.903	0.903	**0.904**	0.902	0.899	0.896	0.893	0.887	0.865	0.798	0.800

表 3: 各 14 素性の組み合わせによる SBCH とチューニングにおける BLEU スコアとの相関分析の結果。表には各要素数における組み合わせにおいて最も相関が高かったものを記載している。横列は用いた素性モデルの数、縦列は素性モデルを示す。相関係数はピアソンの積率相関係数 ρ を用いた。いずれの結果も負の相関を示した。この結果から、素性モデル毎に相関が高い組み合わせだけではなく、素性モデル単体では翻訳品質との相関が低いものも含めた素性の組み合わせが最も高い相関を示すことが分かった。

合型交差エントロピー H_{CSBCH} として以下に定義する。

$$H_{CSBCH} = -|Tuning| \times \sum_{i \in \widehat{Selection}} (\alpha H_{cross}(h_i(T|S)) + \beta w_i) H(h_i) \quad (4)$$

また、再推定後のパラメータ $\hat{w}_i$ を以下に示す。

$$\hat{w}_i = \begin{cases} \alpha \times H_{cross}(h_i(T|S)) + \beta \times w_i & (i \in \widehat{Selection}) \\ w_i & (otherwise) \end{cases} \quad (5)$$

今後の実験でパラメータ推定を行う際にも、素性サイズごとの素性の組み合わせはスコアに基づく交差エントロピーにおいて最も相関を示した素性組み合わせ $\widehat{Selection}$ を用いる。

6 翻訳実験

重回帰分析	MERT	初期	チューニング	
		○		I
			○	T
○		○		IC
○			○	TC
	○	○		IM
	○		○	TM

表 4: 翻訳実験の内訳。初期は初期パラメータを用いた実験、チューニングはチューニング後のパラメータを用いた実験であることを示す。

本章では交差エントロピー混合型対数線形モデルを用いたフランス語からフィンランド語 (fr-fi)、ポルトガル語からスペイン語 (pt-es) への翻訳実験を行う。表 5 に示す通り、MERT を用いたフランス語からフィンランド語への翻訳実験ではチューニングを行ったにも関わらず、BLEU スコアが低下する。この原因は、モデルの品質を考慮しないことによるパラメータの過学習であると解釈できる。本実験では、1. 重回帰分析によるパラメータ推定 (C) と 2.MERT を用いたパラメータ推定 (M) の二つの場合における翻訳実験を行う。また、初期パラメータとチューニング後のパラメータの二つの条件の下でパラメータ再推定を行う。パラメータ再推定後、本実験では原言語 1 文につき 100-best をデコーディングし、交差エントロピー混合型対数線形モデルを用いてリランキングを行う。翻訳品質は BLEU を用いて数値化した。

表 4 に実験の内訳を示す。実験で用いられるモデルやデータセットは相関分析の際に用いた条件に従う (表 1 参照)。

翻訳実験の結果を表 6 に示す。$\dagger^1$、$\dagger^2$、$\dagger^{1,2}$ はそれぞれ I、T、I・T いずれもと比較して統計的に有意な翻訳品質の向上が見られた（p 値が 0.05 以下となった）ことを示す。$\ddagger^1$、$\ddagger^2$、$\ddagger^{1,2}$ はそれぞれ I、T、I・T いずれもの信頼区間の幅よりも高い翻訳品質の向上が見られたことを示す。本実験では、SBCH と BLEU との相関分析最も相関が高かった 6 素性での組み合わせを用いて翻訳実験を行った。本実験の結果より、交差エントロピー混合型対数線形モデルを用いることで、

	Source	Target	BLEU of tuning	Time (min.)	BLEU of testing
I	fr	fi	14.55	0	14.67
T	fr	fi	15.03 †	49	14.67
I	pt	es	373.6	0	37.56
T	pt	es	38.90 †	76	39.05 †

表 5: フランス語 (fr) からフィンランド語 (fi)、ポルトガル語 (pt) からスペイン語 (es) への翻訳結果。実験で用いられるモデルやデータセットは相関分析の際に用いた条件に従う（表 1 参照)。† はチューニングにより統計的に有意な向上が見られた (p-value が 0.05 以下であった) ことを示す。Time はチューニングの推定時間を示し、本研究では壁時計時間を採用した。この結果から、チューニングにおいては統計的に有意な翻訳品質の向上が得られているにも関わらず、フランス語からフィンランド語への翻訳ではテストでは有意な結果を得られていないことが分かる。

	BLEU		Time (min.)	
	fr-fi	pt-es	fr-fi	pt-es
I	14.67	37.56	0	0
T	14.67	39.05	49	76
IC	14.67	**37.90** $\dagger^{2}$ $\ddagger^{2}$	17	17
TC	**14.76** $\dagger^{1,2}$	**39.24** $\dagger^{1,2}$ $\ddagger^{1}$	54	81
IM	**14.71** $\dagger^{1}$	**37.89** $\dagger^{1}$ $\ddagger^{1}$	**29**	18
TM	**14.79** $\dagger^{1,2}$	**39.26** $\dagger^{1,2}$ $\ddagger^{1,2}$	69	83

表 6: fr-fi,pt-es における翻訳実験の結果。実験名は表 4 を参照。Time はパラメータ再推定時間を示しており、本研究では壁時計時間を用いた。本実験の結果から、全ての提案手法において統計的に有意な翻訳品質の向上が見られたことが明らかとなった。

両言語対とも統計的に有意な翻訳品質の向上を得ることができた。また、統計的手法を用いた場合に比べ、MERT を用いた場合のほうがよりよい結果を得ることが分かる。そして、fr-fi において初期パラメータの結果を用いることにより翻訳品質の向上を行うことができたことからも、言語族が異なる翻訳において既存の最適化手法より有効であることを明かにした。

7 今後の課題

本研究では、素性モデルの品質に着目したパラメータ付けを行うことにより、翻訳品質の向上に貢献しただけでなくチューニング時間も短縮することにも成功した。今後の課題として、チューニングによる統計的に有意な BLEU スコアの向上が見られる言語対の翻訳においても、翻訳品質の向上だけではなく、チューニング時間の短縮ができるパラメータ推定手法を考察していきたい。

参考文献

[1] David Chiang, Yuval Marton, and Philip Resnik. Online large-margin training of syntactic and structural translation features. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2008)*, pages 224–233. Association for Computational Linguistics, 2008.

[2] Francisco Guzman and Stephan Vogel. Understanding the performance of statistical MT systems: A linear regression framework. In *Proceedings of the 24rd International Conference on Computational Linguistics (COLING 2012)*, volume Technical Papers, pages 1029–1044, Mumbai, India, December 2012.

[3] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, volume 5, pages 160–167. Association for Computational Linguistics, 2003.

[4] Pavel Pecina, Antonio Toral, Andy Way, Vassilis Papavassiliou, Prokopis Prokopidis, and Maria Giagkou. Towards using web-crawled data for domain adaptation in statistical machine translation. *In Proceedings of European Association for Machine Translation*, pages 297–304, 2011.

[5] Taro Watanabe, Kenji Imamura, Hideto Kazawa, Graham Neubig, and Satoshi Nakamura. *Machine Translation (in Japanese)*, volume 4 of *Natural Language Processing*. Corona, 2 edition, July 2015.