劣モジュラ関数最大化によるフレーズテーブル削減

西野 正彬　　鈴木 潤　　永田 昌明

NTT コミュニケーション科学基礎研究所
{nishino.masaaki,suzuki.jun,nagata.masaaki}@lab.ntt.co.jp

概 要

本稿ではフレーズに基づく機械翻訳において用いられるフレーズテーブルから不要なフレーズ対を削減する方法を提案する．提案法はこの問題を組合せ最適化問題の一種である，要素数制約下での劣モジュラ関数最大化問題として定式化し，貪欲法による近似解法によって解く．実験では提案法が既存のエントロピーを用いたフレーズテーブル削減法と同等あるいはより高い BLEU 値を示すことを確認した．

1　はじめに

フレーズに基づく統計的機械翻訳で用いられるフレーズテーブルとは，対訳関係にある原言語・目的言語のフレーズ対の集合のことである．現実的な機械翻訳システムにおいては，フレーズテーブルは 1 億を超える膨大な量のフレーズ対を含むことが多い．このため，フレーズに基づく機械翻訳システムを動作させるために，計算機は十分な記憶容量をもつストレージを備えていなければならないという課題があった．また，フレーズテーブルが大きくなるにつれ，デコードにおける探索空間が増大するため，探索により時間がかかるようになるという課題もあった．

フレーズテーブル削減法 (Phrase Table Pruning Method) は，フレーズテーブルから不要なフレーズ対を取り除くことによって，フレーズテーブルの大きさを削減する方法である．もちろん，無作為にフレーズ対を取り除くことによってもフレーズテーブルのサイズを削減できるが，翻訳システムの性能は当然ながら低下する．そのため，フレーズテーブル削減法は何らかの基準に基づいて削除するフレーズ対を選択することによって，性能低下を最小限に抑えつつフレーズテーブル削減を実現することが求められる．

既存のフレーズテーブル削減法は，何らかの評価尺度に従ってフレーズ対に対するスコアを計算してフレーズ対をランキングし，上位のフレーズ対のみを取り出すことによってフレーズテーブル削減を行っていた．フレーズ対を評価するための尺度としては，頻度や条件付き確率を用いる素朴な方法や，フィッシャーの正確確率検定などの統計量を用いることで，フレーズ対質を評価する方法などが用いられてきた [Johnson 07]．文献 [Zens 12] ではこうしたフレーズテーブル削減のための各種スコア算出法を網羅的に評価し，エントロピーに基づく削減法 [Ling 12, Zens 12] が最もよい性能を示したと報告している．エントロピーに基づく削減法はフレーズ対の冗長性，すなわちあるフレーズ対を別のフレーズ対群によって置き換え可能であるかを評価することによって，フレーズ対の価値を評価する方法である．実現する方法である．エントロピーに基づく方法はフレーズテーブルの大きさに対しては線形に動作する一方で，フレーズ対に含まれるフレーズ長に対しては指数時間の計算が必要となる．そのため，長いフレーズが大量に含まれる場合には計算が困難であるという課題があった．

本稿では既存のランキングに基づくフレーズテーブル削減法とは異なるアプローチとして，フレーズテーブル削減を組合せ最適化問題の一種である要素数制約下での劣モジュラ関数最大化問題として定式化して解く方法を提案する．劣モジュラ関数とは集合を引数にとって実数値を出力する集合関数の一種である．要素数制約下での劣モジュラ関数最大化問題は一般的には NP 困難であるが，貪欲法を用いて得られる近似解のスコアが最適解の $(1-1/e)$ よりも大きくなることが理論的に保証できるという性質がある [Nemhauser 78]．

本稿ではフレーズテーブルがいかに訓練データ中の語彙を被覆しているかを評価するシンプルな劣モジュラ関数を目的関数として採用する．検証では 4 種類の異なる訓練データセットに対して提案手法を適用し，削減後のフレーズテーブルを用いた翻訳システムの BLEU 値が，エントロピーを用いたフレーズテーブル削減法と同等あるいはより高い値を示したことを確認した．

2 劣モジュラ関数最大化

台集合 Ω の部分集合 $X \subseteq \Omega$ を受け取り，実数値を返す集合関数 $g : 2^\Omega \mapsto \mathbb{R}$ が，任意の集合 $X, Y \in 2^\Omega$ に対して

$$g(X) + g(Y) \geq g(X \cup Y) + g(X \cap Y)$$

を満たすとき，g を劣モジュラ関数とよぶ．さらに，$X \subseteq Y$ であるような任意の $X, Y \in 2^\Omega$ に対して，常に $g(X) \leq g(Y)$ を満たすような劣モジュラ関数 g を単調劣モジュラ関数とよぶ．要素数制約下での単調劣モジュラ関数の最大化問題とは，$|X| = K$ を満たす $X \in 2^\Omega$ のうち，目的関数 $g(X)$ を最大とするものを見つける問題である．ここで K は解に含まれる要素数を定めるパラメータである．この問題は NP 困難ではあるが，貪欲法を用いて解くことによって，解のスコアが厳密解の $(1-1/e)$ 倍以上となることが保証されている近似解が得られるという性質がある [Nemhauser 78]．貪欲法によって近似解を求めるアルゴリズムを Algorithm 1 に示す．このアルゴリズムでは，まず空の集合 $X = \emptyset$ を用意し，$|X| = K$ となるまで，$g(X \cup \{x\}) - g(X)$ を最大とするような要素 $x \in 2^\Omega \setminus X$ を集合に加えていくことで近似解を求める．

フレーズテーブル削減問題は，要素数が K であるようなフレーズテーブルの部分集合のうち，スコアを最大とするものを選択する問題と見なすことができる．すなわち，部分集合 X の良さを評価する適切な評価関数 $g(X)$ を単調な劣モジュラ関数として定義することができれば，貪欲法によって精度保証つきの近似解を得ることが可能となる．

関数の評価が定数時間で行えると仮定すると，貪欲法の実行時間は台集合のサイズ N に対し $O(NK)$ となる．これは，アルゴリズムの 3 行目の処理で各 x について関数を評価するために $O(N)$ 時間かかるためである．フレーズテーブル削減のように N が大きな値をとる場合には，貪欲法を素朴に実装すると現実的な実行時間で動作しない可能性がある．しかし，遅延評価テクニックを用いることで，N が大きな値をとる場合であっても実用的には高速に貪欲法を実行できることが知られている [Leskovec 07]．

Algorithm 1 貪欲法による劣モジュラ関数最大化法

入力: 台集合 Ω, 要素数 K
出力: $|X| = K$ を満たす集合 $X \in 2^\Omega$

1: $X \leftarrow \emptyset$
2: **for** t in 1 to K **do**
3: $\quad g(X \cup \{x\}) - g(X)$ を最大化する $x \in \Omega \setminus X$ を選択．
4: $\quad X \leftarrow X \cup \{x\}$
5: **return** X

3 目的関数の設計

はじめに以下で用いる概念を導入する．$\Omega = \{z_1, \ldots, z_M\}$ を M 個のフレーズ対からなるフレーズテーブルとする．各フレーズ対 z_i は，原言語のフレーズ p_i と目的言語のフレーズ q_i のペアとして，$z_i = (p_i, q_i)$ として定義されるものとする．各フレーズはそれぞれ単語の列として，$p_i = (p_{i1}, \ldots, p_{i|p_i|})$, $q_i = (q_{i1}, \ldots, q_{i|q_i|})$ として表されるものとする．ここで p_{ij} は p_i 中の j 番目の語を，q_{ij} は q_i 中の j 番目の語をそれぞれ表すとする．次に訓練データである対訳コーパスにおける原言語の文の集合を $F = \{f_1, \ldots, f_N\}$, 目的言語の文の集合を $E = \{e_1, \ldots, e_N\}$ とする．N を対訳コーパスに含まれる対訳文対の総数とする．e_i, f_i はそれぞれの集合における i 番目の文を表し，(f_i, e_i) は対訳関係にあるとする．各文は単語の列として，それぞれ $f_i = (f_{i1}, \ldots, f_{i|f_i|})$, $e_i = (e_{i1}, \ldots, e_{i|e_i|})$ として表されているものとする．ここで e_{ij}, f_{ij} は e_i の j 番目の語，f_i の j 番目の語をそれぞれ表すものとする．

あるフレーズ対 $z_j = (p_j, q_j)$ と対訳文対 (f_i, e_i) に対して，p_j が f_i に部分列と一致し，かつ q_j が e_i の部分列と一致しているとき，z_j がペア (f_i, e_i) に出現すると定義する．また，フレーズ対 z_j が (f_i, e_i) に出現しているとき，f_i 中の語 f_{ik} が，p_j に一致する部分列に含まれるのであれば，f_{ik} は z_j によって被覆されていると定義する．目的言語の単語 e_{ik} についても同様である．

以上の定義をふまえ，フレーズテーブル削減のための目的関数を以下のように定義する．

$$g(X) = \sum_{i=1}^{N} \sum_{k=1}^{|f_i|} \log\left[c(X, f_{ik}) + 1\right] + \sum_{i=1}^{N} \sum_{k=1}^{|e_i|} \log\left[c(X, e_{ik}) + 1\right]. \quad (1)$$

ここで $c(X, f_{ik})$ はフレーズ対の集合 X に含まれるフレーズ対のうち，原言語 F に含まれる i 番目の文の k 番目の単語 f_{ik} を被覆するものの数を表す．この目的関数は，X に含まれるフレーズ対が対訳コーパス F, E 中の語を被覆する回数が多いほど高いスコアを与える．ただし，$c(X, f_{ik})$ の対数をとっているため，コーパス中のより広範囲の語を被覆するような集合 X に対し

表 1: 訓練データの語数（単位は億個）.

	単語数	
言語対	原言語	目的言語
アラビア語-英語	1.86	1.59
中国語-英語	1.27	1.39
スペイン語-英語	2.00	1.81
ドイツ語-英語	0.34	0.36

表 2: フレーズテーブルのサイズ（単位は億個）.

言語対	フレーズ対数
アラビア語-英語	2.34
中国語-英語	1.69
スペイン語-英語	2.70
ドイツ語-英語	0.64

てより高いスコア与える．このような目的関数を用いることによって，多くの語を被覆するような訓練コーパス中で頻出のフレーズ対を含みつつ，かつ語彙が特定の語に偏らない，冗長性の少ないフレーズ対の集合を求めることができる．$g(X)$ は単調な劣モジュラ関数となっている．

4 検証

4.1 設定

検証では，アラビア語-英語，中国語-英語，スペイン語-英語，ドイツ語-英語の 4 種類の訓練データセットを用いた．アラビア語-英語，中国語-英語のデータセットは NIST OpenMT 2012 のものを，スペイン語-英語，ドイツ語-英語のデータセットは WMT 2012 のものを用いた．表 1 に各訓練データに含まれる単語数を示す．いずれのデータセットに対しても，英語を目的言語として実験を行った．

フレーズに基づく翻訳システムとして Moses [Koehn 07] を用いた．言語モデルとして，English GigaWord V5 コーパス (LDC2011T07), WMT2012 で配布された単言語データ，Google Web 1T 5-gram V1 data (LDC2006T13) の 3 種類のデータセットから個別に学習された 5 グラム Kneser-Ney 言語モデルを併せて利用した．

単語アラインメントは Moses に組み込まれている giza++ [Och 03b] をデフォルト設定で実行して 得られたものを用いた．素性重みは，Moses に組み込まれているエラー最小化学習法 (MERT) [Och 03a] に開発データセットを与えて実行することで学習した．素性重みはフレーズテーブル削減によって新しいフレーズテーブルが得られるごとに再度学習しなおした．また，フレーズ抽出法によって抽出されるフレーズの最大長さは 7 とした．フレーズテーブルのサイズを表 2 に示す．

ベースラインとしてエントロピーを用いたフレーズテーブル削減法 [Ling 12, Zens 12] を採用する．エントロピーを用いた方法は文献 [Zens 12] において，他の手法よりも高い BLEU スコアを示すことが確かめられている．なお，エントロピーベースの削減法のパラメータは先行研究における推奨パラメータを利用した．

4.2 結果

フレーズテーブル削減を行い，テストデータにおける BLEU 値を比較した結果を図に示す．図の横軸は元のフレーズテーブルのサイズに対する，圧縮後のフレーズテーブルサイズの大きさの比率，縦軸はテストデータにおける BLEU の値である．この結果より，スペイン語-英語，アラビア語-英語データセットにおいては提案法がエントロピー法よりも常に高い BLEU 値を示している事がわかる．中国語-英語データセットにおいても概ねエントロピー法よりも高い BLEU 値を示しているが，フレーズテーブルのサイズによってはエントロピー法のほうが高い BLEU 値をとることもあった．ドイツ語-英語データセットでは，低圧縮な場合ではエントロピー法のスコアが高くなる結果が得られた．

今回の検証では，フレーズテーブルのサイズが大きなデータセットの場合に，提案法がより高い性能を示す傾向が見られた．このような傾向を示す要因については今後より詳細な分析が求められる．

5 おわりに

本稿ではフレーズテーブル削減問題を，組合せ最適化問題の一種である要素数制約下での劣モジュラ関数最大化問題として定式化して貪欲法で解く方法を提案した．提案法は単純であり，実装も容易であるにもかかわらず，実験では既存のエントロピーを用いたフレーズテーブル削減法よりも高い性能を示すことが確認できた．さらに，提案法にはフレーズ長に依存せずに高速に動作する，調整が必要なパラメータを含まないといった，システムの実運用上有用な特徴があることも，エントロピー法に対するアドバンテージになると考える．

提案法は目的関数をカスタマイズすることで容易に拡張可能である．本稿では手法のシンプルさを重視して，式 (1) のような比較的素朴な目的関数を利用したが，例えばフィッシャー検定のスコアを目的関数で使うなどの発展形を考えることもできる．より効果的な目的関数の考案は今後の課題である．

参考文献

[Johnson 07] Johnson, H., Martin, J., Foster, G., and Kuhn, R.: Improving Translation Quality by Discarding Most of the Phrasetable, in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 967–975, Prague, Czech Republic (2007), Association for Computational Linguistics

[Koehn 07] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180 (2007)

[Leskovec 07] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N.: Cost-effective Outbreak Detection in Networks, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 420–429 (2007)

[Ling 12] Ling, W., Graça, J. a., Trancoso, I., and Black, A.: Entropy-based Pruning for Phrase-based Machine Translation, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 962–971, Jeju Island, Korea (2012), Association for Computational Linguistics

[Nemhauser 78] Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L.: An analysis of approximations for maximizing submodular set functions—I, *Mathematical Programming*, Vol. 14, No. 1, pp. 265–294 (1978)

[Och 03a] Och, F. J.: Minimum Error Rate Training in Statistical Machine Translation, in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 160–167 (2003)

[Och 03b] Och, F. J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51 (2003)

[Zens 12] Zens, R., Stanton, D., and Xu, P.: A Systematic Comparison of Phrase Table Pruning Techniques, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 972–983, Jeju Island, Korea (2012), Association for Computational Linguistics

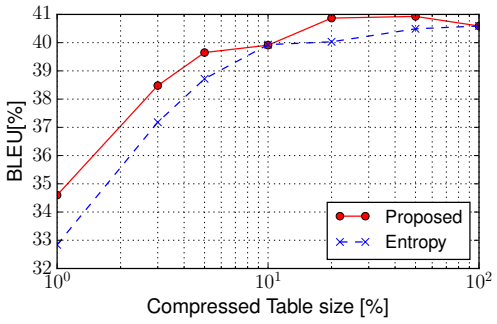

図 1: 実験結果（アラビア語-英語）

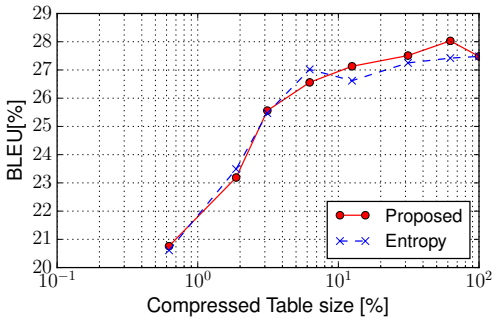

図 2: 実験結果（中国語-英語）

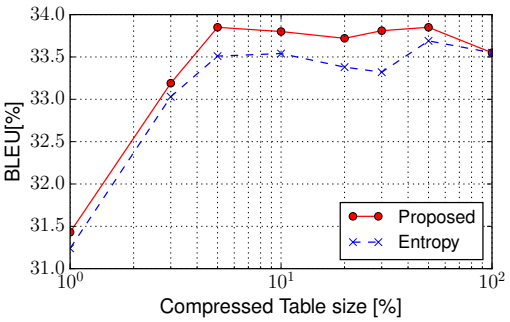

図 3: 実験結果（スペイン語-英語）

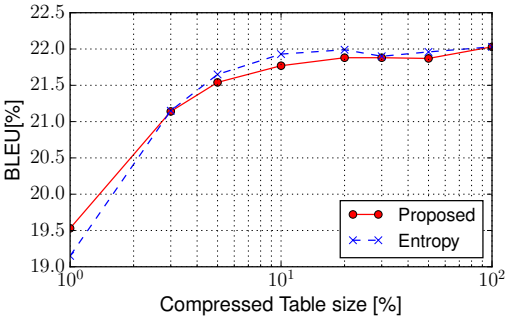

図 4: 実験結果（ドイツ語-英語）