機械学習を用いた文脈自由規則の書き換えによる文圧縮

辻井 潤一
tsujii@is.s.u-tokyo.ac.jp

1 はじめに

文圧縮は短くかつ有用な要約文の生成に必要不可欠な技術であるだけでなく，ニュースのヘッドライン生成や携帯電話などの小さな表示域への表示といった応用も考えられる．Knight & Marcu [3] は，原文の構文木を書き換えることにより文圧縮を行う統計的文圧縮手法を提案している．彼らは要約文と原文の対応がとれたコーパスを用意し，それらを構文解析したあと，要約文の構文木ノードと原文の構文木ノードの対応をつけて，構文木を書き換える確率を学習する．彼らの手法は簡潔でよく定義されているが，(1) 要約文の構文木ノードと原文の構文木ノードの間で対応がとれない場合があるという問題点，および (2) 構文木の書き換え確率を生成規則にのみ依存させているため，時として重要な文節や単語を削除してしまうという問題点がある．

我々は構文木書き換え確率が生成規則だけに依存するのではなく，周辺のノードや単語にも依存していると考え，最大エントロピー法を用いる文圧縮手法を提案する．例えば，書き換え対象の生成規則だけではなく，その親や孫などの周辺ノードや，重要な単語の有無などの特徴量を最大エントロピーモデルの素性として取り入れた．また，上記 (1) の問題を解決するために，要約文の構文木と原文の構文木で対応をとるのではなく，原文の構文木の中から部分木を選び出すことで対応をつける，ボトムアップ対応法を提案する．

原文と要約文の対応のとれた Ziff-Davis コーパスを使用して，Knight らの手法と最大エントロピー法による手法と，それにボトムアップ対応法を加えた手法の性能を評価した．また，学習用コーパスのサイズを変えたときの精度も評価した．

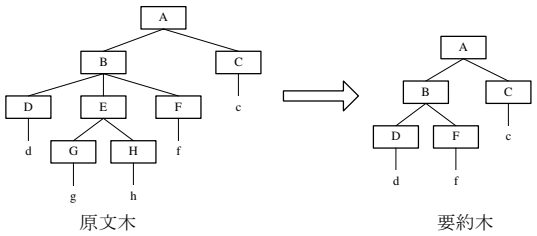

図 1: 構文木における文圧縮の例

2 背景

2.1 Noisy-channel Model に基づく文圧縮

Knight & Marcu [3] は構文木の書き換えに対して Noisy-channel Model を仮定した文圧縮手法を提案した．すなわち，短い文（要約文）にノイズが挿入され長い文（原文）が観測されたと考えるモデルを仮定した文圧縮手法を提案している．ここで，要約文は原文の部分文字列であると仮定している．つまり，単語が置き換わったり語順が変わることはないとしている．

彼らの手法では原文 l に対して，要約文 s の条件付き確率 $P(s|l)$ を最大にする要約文を求める．ベイズ則により，これは $P(l|s)P(s)$ を最大化する問題と等価である．$P(s)$ は bigram と PCFG のスコアから求められ，$P(l|s)$ は以下のように求められる．まず，原文と要約文の組を文脈自由文法の構文木に変換する．各構文木をそれぞれ原文木，要約木と呼ぶ（図 1）．つづいて，適用された生成規則を原文木と要約木で構文木のルートから比較し，適用された生成規則の対を作る．たとえば，図 1 の要約木中に $r_s = (B \rightarrow D\ F)$，原文木中に $r_l = (B \rightarrow D\ E\ F)$ の様な，右辺が部分列になっている対応関係があったとき，これを前者が後者に書き換わった（後者が前者に**縮約**された）と見なす．これに，新たに生成された木 $(E\ (G\ g)\ (H\ h))$ の生成確率を PCFG のスコアとして加える．要約文から原文

へ書き換わる確率は，対応する生成規則の書き換え確率と，対応しない木の生成確率から計算される．

$$P(l|s) = \prod_{(r_l,r_s)\in R} P_{exp}(r_l|r_s) \cdot \prod_{r\in R'} P_{cfg}(r)$$

但し，R は上記手続きにより対応づけられた生成規則の対の集合であり，R' は l の構文木中に含まれるが対応づけられなかった生成規則の集合である．

要約文を作成する際は，原文をまず構文木に変換し，その中の各生成規則 r_l と，縮約される候補 r_s に対して，$P(r_l|r_s)$ を求める．縮約の候補は，r_l の右辺の非終端記号の数を n とすると，少なくとも一つ残すとして，2^n-1 通り存在する．すべての書き換えの確率が求まれば，$P(s|l)$ が求まる．要約文の長さで正規化した $log(P(s|l))/|s|$ を最大とする要約文を結果として返す．

2.2 最大エントロピーモデル

最大エントロピーモデル [1] とは，観測事象 x_i と履歴事象 y_i との組からなる学習データ $\{(x_1,y_1),\cdots,(x_n,y_n)\}$ から，条件付き確率 $P(x|y)$ を推定するモデルである．最大エントロピー法では，観測事象と履歴事象の組 (x,y) の特徴を素性関数という二値関数を使って表す．素性関数 f_i は (x,y) がある特徴 i をもつときにのみ 1 となり，それ以外で 0 となるように定める．このとき，学習データにおける f_i の期待値とモデルにおける f_i の期待値が一致すると仮定し，これを制約と呼ぶ．すべての素性関数に関して制約を満たし，分布のエントロピー $H(P) = -\sum_{x,y} P(x,y)\log(P(x,y))$ を最大化させる分布 $P^*(x,y)$ が求める分布となる．

3 アルゴリズム

3.1 最大エントロピー法による圧縮

Knight らの手法では，原文木と要約木における生成規則の対から書き換え確率が計算されたが，親の非終端記号や，単語などの周辺の情報も圧縮するかしないかに関わっていることは予想できる．しかし，これらの情報すべてを彼らの枠組みで扱うのは難しい．

本論文では Knight らのモデルの自然な拡張として，CFG の生成規則を縮約する確率を最大エントロピー法を使って計算する手法を提案する．最大エントロピー法では，先のような特徴も素性関数という形で容易に取り入れることができる．

親ノード/現在のノード/子ノードと削除するか否か/残った子ノード/ 子ノードの数/ルートノードからの深さ/削除する子ノード/ 削除する子ノードの終端記号/子ノードは否定の副詞か/ 子ノード列の 3 グラム/ 子ノードが一つだけ残りかつ親ノードと同じか/ 子ノードが一つだけ残りかつ自分のノードと同じか/ 子ノード列を何カ所に分けるか/ 主辞の子ノードを削除するか/ 最も左・右の子ノード/ 左右の兄弟ノード

表 1: 使用した素性

CFG の構文木中の各生成規則を縮約させるため，各生成規則に関して，その右辺の部分列を選択する問題を考える．たとえば，入力文の構文木中に $(A \rightarrow B\ C\ D)$ という生成規則があったとき，右辺の $B\ C\ D$ に対する部分列の候補は，少なくとも一つの非終端記号を残すものとして以下の 7 通りの候補がある．

$$\mathcal{Y} = \{(B),(C),(D),(B\ C),(B\ D),(C\ D),(B\ C\ D)\}$$

この中から 1 つの候補を選択する問題として，各候補 $y \in \mathcal{Y}$ に対する条件付き確率 $P(y|\mathcal{Y})$ を最大エントロピー法により求めることができる．使用した素性は，表 1 の通りである．

要約木中の生成規則 r_s と対応する原文木の生成規則 r_l の対から，要約木の候補の確率値が計算できる．

$$P(s|l) = \prod_{(r_s,r_l)\in R} P(r_s|r_l)$$

但し，R は対応する生成規則の対の集合である．また，特定の長さの要約を得る場合，その長さの候補の中で $P(s|l)$ を最大にする候補を選べばよい．

3.2 ボトムアップ対応法

コーパス中の 2 つの構文木を対応付けるだけでは Knight らの様に対処できない要約例が存在する．特に複文，例えば図 2 の “I don’t think so, I said.” という文と，その要約 “I don’t think so.” は，ルートの S が対応していない．

我々はこの問題に対処するために，ボトムアップに対応をとる方法を提案する．図 3 の様な “dghfc” という文に対して，“dgc” という文を対応づけるとする．まず，構文木の終端記号の内，要約文に含まれる記号に丸でマークを付ける．次にマークの付いた終端記号を子として持つ非終端記号に，再帰的に太線でマークを付ける．こうしてマークの付いた記号およびそれを結ぶ枝のみを抽出すると図の右側の木が得られ，これを

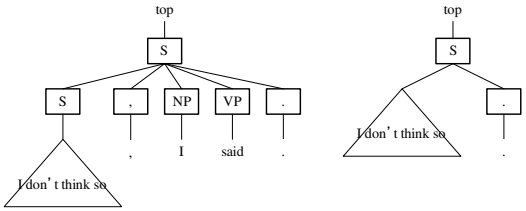

図 2: 対応のつかない複文の例

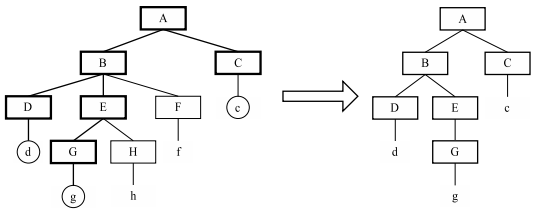

図 3: ボトムアップに対応付ける例

要約木と見なす．要約文に対して構文解析器は使用しない．

3.3 パラメタによる長さの調整

Knight らは長さの違う要約候補中から 1 つの候補を選択するのに，対数確率を文長で割ったスコアを用いた．我々はこれの自然な拡張として，パラメタ α を用意し，要約 s のスコア S を，$S_\alpha(s) = length(s)^\alpha \cdot \log(P(s|l))$ とした．適切な α の値はアルゴリズムによって異なり，これを上下させることで出力結果の文長を調整できる．

4 実験

学習と評価には Knight らが Ziff-Davis コーパスから抽出した，文対応のとれたコーパスを使用した．527 文をトレーニングデータに，263 文を開発用のテストセットに，残りの 264 文を最終的なテストセットとして使用した．CFG の構文解析器として，Charniak & Johnson の reranking parser [2] を使用した．評価尺度には単語単位の F 値，bigram 単位の F 値，BLEU スコア [4] を用いた．

実験は 2 種類行った．一方は要求する要約文の文長と原文を与え，決められた長さの要約を出力する．もう一方は原文のみを渡して要約文を出力する．前者の実験では長さによるスコアの影響がないため，手法間の比較に適しているが，文長が与えられるためにより簡単なタスクになる．後者の方が自然なタスクだが，

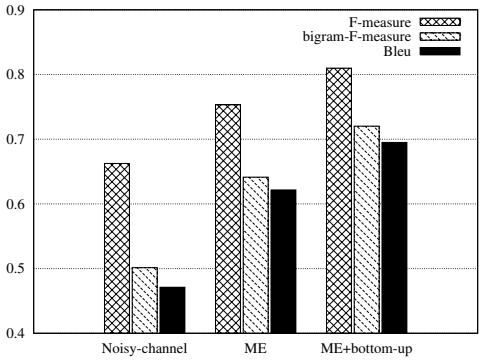

図 4: 出力長を指定したときの実験結果

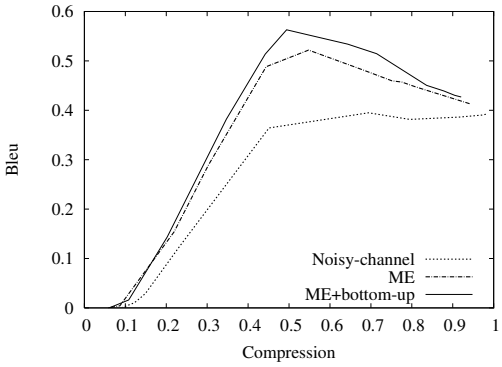

図 5: 要約率と BLEU スコアの関係

各スコアは出力文の長さに強い影響を受けるので，手法間での結果の比較は難しい．我々の実装では，長さによる重みをつけることで，圧縮率を調整できるため，圧縮率とスコアの関係を調べた．

図 4 は，要約文長が与えられたときの各アルゴリズムの実行結果の比較である．以下，Noisy-channel は Noisy-channel モデルの，ME は最大エントロピー法による，ME+bottom-up は最大エントロピー法にボトムアップ対応法を使ったときの結果である．文長は正解文の文長を与える．最大エントロピー法にボトムアップ対応法を用いた実験結果がどのスコアでも一番よい精度を達成した．

もう一方の実験は，長さによるスコアの重み α を変えて行った．これを調整して平均長の異なる要約結果を得ることが可能である．このときの圧縮率とそれに対する BLEU スコアをアルゴリズム間で比較したのが，図 5 である．これより，我々の提案手法は同じ圧縮率

原文	this $ 549 package is a feature-rich , true relational database that uses dbase-compatible file formats .
正解	this relational database uses dbase-compatible file formats .
Noisy-channel	is database that uses dbase-compatible file formats .
ME	this package is a relational database that .
bottom-up	package is that uses dbase-compatible file formats .

原文	the user can then abort the transmission , he said .
正解	the user can then abort the transmission .
Noisy-channel	the user can abort the transmission said .
ME	the user can abort the transmission said .
bottom-up	the user can then abort the transmission .

原文	it is likely that both companies will work on integrating multimedia with database technologies .
正解	both companies will work on integrating multimedia with database technologies .
Noisy-channel	it is likely that both companies will work on integrating .
ME	it is likely that both companies will work on integrating .
bottom-up	it is will work on integrating multimedia with database technologies .

表 2: 出力結果の例

でも旧来のアルゴリズムより良い要約結果を得ていることがわかる．

表 2 にいくつかの要約結果を示した．それぞれ上から，原文，人間による正解，Noisy-channel Model による出力，最大エントロピー法を用いた時の出力，最大エントロピー法にボトムアップ対応法を用いた時の出力例である．1 つ目は，最大エントロピー法を用いたことによって改善された例である．これに限らず，Noisy-channel モデルは主語を落としてしまうことが何度か観察された．これは to 不定詞や分詞構文といった主語を持たない S が多数存在するためと考えられる．生成規則のみから判断する Noisy-channel モデルにとって，これを区別するのは難しい．しかし，最大エントロピー法であれば，例えばルートの S は親ノードを持たないため，これを素性に加えることで学習することができる．2 つ目は，ボトムアップ対応法によって改善された複文の例である．これは先の説明でも示したような典型的な複文の例である．ボトムアップ対応法をとっていない 2 つは，“said” を除去できていない．3 つ目はボトムアップ対応法でも解決できない例である．この変換を我々の方法の上で正しく行うためには，$(S\ (VP\ (SBAR\ (S\ (...)))))$ という，unary 規則が多数適用された形の木を生成しなければならないが，このような木が生成される確率は非常に低い．

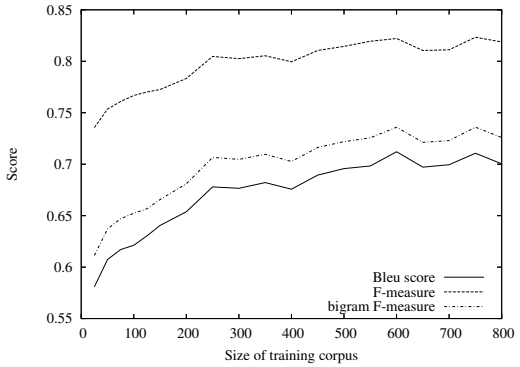

図 6: コーパスサイズと精度

使用コーパスの量と精度の関係を調べるために，学習時に使う文の数を変えて実験を行った．図 6 は，もっとも精度の高かった，最大エントロピー法とボトムアップ対応付けを行うアルゴリズムに対する，学習コーパス量と各スコアの対応である．コーパス量が増えるにつれて精度の向上が見られるが，600 文あたりをピークにかなり緩やかになっている．

5 結論

我々は構文木を変換させる圧縮手法に対して機械学習を適用し，2 つの実験と 3 つの評価基準において，従来より精度の高い要約文を生成することを示した．また，従来の方法では難しかった形の文に対しても，高い精度で圧縮できるようになった．我々のモデルでは正しく圧縮できない可能性のある問題も残っているため，これを改善させることで精度の向上が期待できる．また，評価基準にもまだ改善の余地があると考えられる．

参考文献

[1] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, Vol. 22, No. 1, pp. 39–71, 1996.

[2] E. Charniak and M. Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proc. of ACL'05*, pp. 173–180, June 2005.

[3] K. Knight and D. Marcu. Statistics-based summarization - step one: Sentence compression. In *Proc. of AAAI/IAAI'00*, pp. 703–710. AAAI Press / The MIT Press, 2000.

[4] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL'02*, pp. 311–318, 2002.