教師なし学習による文の分割*

山本 博史, 菊井 玄一郎
ATR 音声言語コミュニケーション研究所
〒619-0288 京都府相楽郡精華町光台 2-2-2

1 はじめに

日本語のように単語区切りを行わない言語を対象とした場合，文を単語単位に分割することは，構文解析の前処理として，あるいは音声認識，OCR 認識等のための言語モデルの訓練データの生成処理として必須の技術と言っても良い．単語分割のための最も一般的な方法は，あらかじめ人手で単語分割ずみのコーパスを用意しておき，それを教師として分割モデルを学習するものである．

文を単語に分割するにあたっては，当然単語そのものをあらかじめ定義しておく必要がある．単語区切りを行わない言語の場合，単語の定義そのものに曖昧性がある場合が多く，用途に応じて単語の定義が変わって来ることがありえる．このような場合，異なる単語の定義にしたがって人手で分割されたコーパスをそのまま訓練データとして用いることができなくなる．このように単語定義が一致した訓練コーパスが入手できない場合の方法として，別の単語定義で分割された文から目的の単語定義での分割に変換する方法 [1] や，人手で単語分割ずみのコーパスを用いない，すなわち教師なしの学習を行う方法 [2] が提案されている．

これらの方法の問題点は，変換，あるいは学習にあたって何らかのヒューリスティクスを用いることである．これらのヒューリスティクスは対象とする言語に依存する場合がほとんどであり，対象言語が変わった場合，特に言語的知見が得られにくいような外国語を対象とした場合には対応が難しい．

そこで，本稿ではヒューリスティクスを用いない教師なし学習による文の分割方法を提案する．本手法を用いた場合，必要となる知識は辞書エントリーのみであるため，何ら言語的知見が得られないような外国語に対しても単語分割が可能となる．

*Unspervised Sentence Segmentation Based on Minmum Entropy Criterion

教師あるいはヒューリスティクスを用いた場合，人間が正しいと思う分割基準を訓練データとして与える，あるいはヒューリスティクスに反映させることによって学習することが可能である．本稿では両者を用いない教師し学習を目的とするために，まずどのような文の分割を目的とするかを客観的基準に基づいて定めておく必要がある．本手法では，与えられた辞書エントリの組合わせの，訓練文全体のエントロピーの最小化を基準として用いることとした．この基準は人間が正しいと思う分割基準と一致している保証がない．しかしながらエントロピーの最小化は次に現れる文字列に対する予測性能を最大化することを意味しており，音声認識，OCR 認識等，目的が単語分割そのものではなく，予測性能の向上にある場合は目的に一致していると考えられる．

2 エントロピー最小化基準に基づく単語分割

あらかじめ単語分割が与えられている場合，その訓練セットに対してエントロピーを最小化するようなモデルとは，次式を最小化するものである．

$$\sum_i P(W_i|h_{i-1})log(P(W_i|h_{i-1})) \qquad (1)$$

ここで W_i は i 番目に現れる単語，h_{i-1} はその単語に至るまでの履歴である．これに対し，単語分割が与えられていない場合は文に対する全ての分割候補を考なければならないため候補単語のネットワークを考える必要がある．従って，与えられた候補単語のネットワークに対してエントロピーを最小化するようなモデルは次式で与えられることになる．

$$\sum_{p,e}\sum_j P(W_{p,e}|h_j)log(P(W_{p,e}|h_j)) \qquad (2)$$

ここで，$W_{p,e}$ は辞書エントリが e であり，開始位置が p であるような単語を，h_j は単語 $W_{p,e}$ に到達し得る候補の経路のうちの一つを表す．

従って，本手法の目的は式(2)を最小化するようなモデルを求める点にある．

しかしながら，式(2)をそのまま最小化するようなモデルを求めようとする場合，

- データスパースの問題から，式(2)を満たすモデルがオープンセットに対しても最小となるとは限らない．
- 計算量が膨大なものになる．

という理由から，式(2)におけるh_jとして，先行する1または2単語のみを考慮することとする．これを式(1)に当てはめた場合は，2または3-gramを用いることを意味している．近似後の式(2)は次の式で与えられる．

$$\sum_{p,e}\sum_{+,++} P(W_{p,e}|+,++)log(P(W_{p,e}|+,++)) \tag{3}$$

ここで$+,++$は，単語$W_{p,e}$の先行および先々行単語を表すものとする．

3 モデルの計算アルゴリズム

3.1 エントロピー最小モデルの計算

続いて，式(3)で与えられるモデルの計算手順を図1に示す．

- 手順(1)では，与えられた全ての訓練文から，与えられた辞書エントリを用いて分割可能な単語ネットワークを生成する．
- 手順(2)ではこのネットワークの各単語に対して$P(W_{p,e}|+,++)$を与えるが，この値は求まっていないため，まず初期値として単語0-gramの値を与え，全ての単語に対する生起確率を等しくおく．
- 手順(3)では各単語に対して与えられている事前確率値を，文が決まったうえでの事後確率値$P(W_{p,e},+,++,S)$に変換する．ここでSは与えられた文を表すものとする．事後確率値はForward-Backwardアルゴリズムを用いて求められる．
- 手順(4)では事後確率値に基づいて$P(W_{p,e}|+,++)$を再計算する．ここでは，各単語に対する事後確率値を，その出現回数と見なすことによって次の式に従って計算が行われる．

$$P(W_{p,e}|+,++) = \frac{\sum P(W_{p,e},+,++,S)}{\sum P(+,++,S)} \tag{4}$$

教師あり学習の場合，$P(W_{p,e}|+,++)$は単語列$W_{p,e},+,++$の出現回数に基づいて計算される．本手順に対して教師の場合にあてはめると，生成されたネットワークには分岐が存在しないため，全ての単語に対する事後確率値は1となり，出現回数と一致するため，教師あり学習の拡張と見なすことができる．

- 手順(5)では，手順(4)で再計算された$P(W_{p,e}|+,++)$を各単語に対して割り当てなおす．

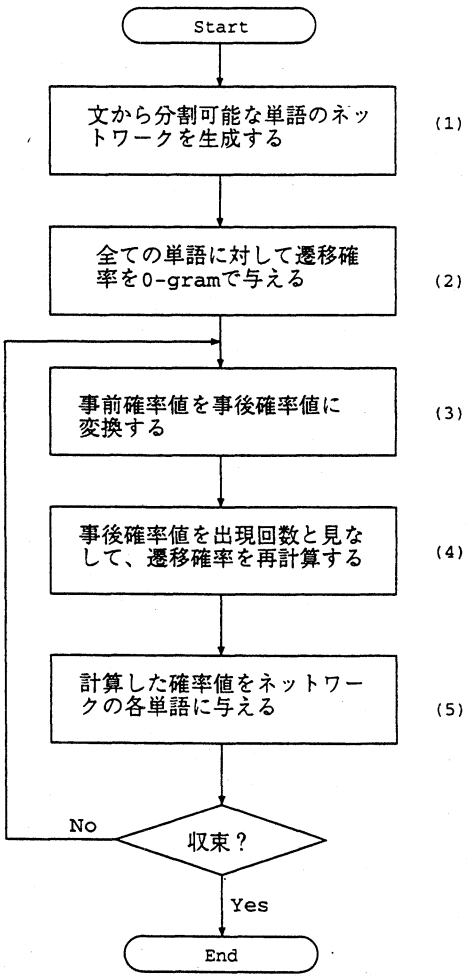

図 *1.* モデルの計算のフロー

3.2 確率値の平滑化

前節の手順で求められたモデルでは，訓練セットに現れなかった単語の組合わせに対しては確率0を返す．訓練セットは有限であるため，この

ままではオープンなセットに現れる単語の組合わせに対して確率0を割り当てることが生じるのを避けるためである．平滑化の方法としてはBack-off平滑化を用いる．ただし通常のBack-off平滑化とは異なり，出現回数として小数を取ることを考慮しなければならない．Back-offとしてGood-Turing[3]を用いた場合は次のような方法をとる．

- 本手法では，教師あり学習の場合，単語列の出現回数が整数を取るのに対し，小数を取る．従って出現回数が1未満の場合もあり得る．このような出現回数を取るような単語列に対する$P(W_{p,e}|+,++)$の値は信頼性が低いと考えられる．そこでまず，出現回数が1未満の単語列のエントリをすべてCut-offする．
- Good-Turingでは，各整数回出現した単語列ごとにディスカウント係数を求める．そこで，小数である各出現回数をいったん整数に丸めたのち，各回出現ごとのディスカウント係数を求める．
- 出現回数がi以上，$i+1$未満の単語列に対して，上で求めた整数回数に対するディスカウント係数を1次補間した値でディスカウントを行う．
- ディスカウントされた確率値からBack-off係数を求め，Back-offを行う．

4 評価実験

4.1 実験条件

提案法に対する評価実験を行う．対象の訓練文は日本語旅行対話約14万7千文であり，市販の会話用例集に一般的に現れるような内容からなっている．辞書は茶筅2.0b6[4]の辞書に含まれる異なり表記約31万6千エントリ(表記のみを用い，品詞等の属性情報は用いていない)を用いた．また，評価文は訓練文はとは別の日本語旅行対話1377文を用いた．比較の対象は同一の訓練文を用いて，教師ありの学習を行った場合である．教師データの生成は，訓練文を茶筅2.0b6を用いて形態素解析を行った結果を用いた．モデルの次元は提案法，教師あり共に2，すなわち$P(W_{p,e}|+)$，$P(W_i|P(W_{i-1})$とした．

4.2 提案法における単語分割例

次に，提案法を用いて評価文を単語分割した結果と，同一文に対して教師ありの学習を行って得られた2-gramを用いて得られた結果との比較を行う．評価文を単語分割は提案法，教師あり共に，モデルが最も高い確率値を示す分割を出力結果とした．以下に両者の出力例を示す．上段が教師あり(S)，下段が提案法(U)の出力例である．

1. • S 濃い / コーヒー / が / 飲み / たい
   • U 濃い / コーヒー / が / 飲 / みたい
2. • S この / あたり / に / デパート / は / あり / ます / か
   • U この / あたり / に / デパート / はあ / り / ま / すか
3. • S どの / 便 / なら / 乗れ / ます / か
   • U どの / 便 / なら / 乗れ / ま / すか
4. • S この / 小包 / を / 日本 / に / 送り / たい / の / です / が
   • U この / 小包 / を / 日本 / に / 送り / たい / ので / すが
5. • S 私 / は / ギター / を / 弾き / ます
   • U 私 / は / ギター / を / 弾 / きま / す
6. • S タクシー / で / いくら / かかり / ます / か
   • U タクシー / で / いくらか / かり / ま / すか
7. • S 電気 / が / つか / ない / ん / です
   • U 電 / 気がつか / ない / んで / す

両者の一致度は，教師ありの場合を正解とみなした時の，提案法の単語正解精度で約68%であった．両者の比較において，名詞に関しては比較的一致度が高いように見受けられるが，実際の提案法の名詞の再現率は約85.8%であった．機能語に関しては全ての例に見られるように，教師ありの場合と一致しない場合が多く，文全体の分割が一致している文は約25%であった．また，動詞に関しては，1,4のように語幹部分で分割される例が見受けられ，その場合，語尾部分は機能語と同じような振る舞いをすることが多い．その他の例としては，6のように，教師ありの場合起こり得ると思われる例や，7のように教師ありと違う分割を行うことによって，頻度の高い単語が出現するような例が見受けられた．

4.3 エントロピーによる評価

続いて，提案法と教師ありの場合のエントロピーによる評価を行った．

表1に示すように，提案法は単語あたりのエントロピーでは教師ありよりも大きいものの，総単語数が少ないため，総エントロピーでは教師ありよりも低い値を示している．このことから，提案法による単語分割は，人間の直感とは必ずしも一

表 1. 提案法と教師ありのエントロピーによる比較

手法	単語数	Entropy/Word	総 Entropy
教師あり	12,425	4.23	52,574
提案法	12,101	4.26	51,580

致しないが，次の文字列に対する予測精度の点からは，教師ありより高い性能を示していることがわかる．

5 今後の課題

提案法は，教師なし学習よりも低いエントロピーを示し，次の文字列に対する予測精度の点からは優れていることがわかった．しかし一方では，教師なし学習との一致で約 68% とかなりの食い違いが見られ，名詞のみに限って見ても約 85.8% にとどまった．このことは，出力結果が文字列として求められる音声認識，OCR 認識等，では問題とならないが，翻訳等，文中の単語の意味を保持しなければならないような場合には非常に大きな問題点となる．「はじめに」で述べたように，教師なし学習をヒューリスティクスなしで行う場合，文の分割の目的となる客観的基準を定めておく必要がある．この意味で，単語の意味を保持するならば，それに対する客観的基準を定めておく必要がある．

今後は，バイリンガルコーパスを用いて，両言語間の単語の共起関係の保持を意味の保持と定義することによって，翻訳等に用いる場合でも問題のない学習方法を検討していく予定である．

6 おわりに

本稿では，文を単語に分割するための手法として，ヒューリスティクスを用いない教師なし学習法を提案した．本手法では，知識は辞書エントリのみでるため，言語の依存せず，言語に対する知見が得づらいような外国語に対しても適用可能である．学習では，文に対して可能なあらゆる分割を表現したネットワークに対して，エントロピーを最小化する方法が取られる．このため，次文字列に対する予測性能の点では期待が持て，出力結果として文字列として求められる音声認識，OCR 認識等には有効と考えられる．

本手法に有効性を検証するために行った日本語旅行対話約 14 万 7 千文，約 31 万 6 千エントリの辞書を用いた実験では，教師なし学習との一致度で約 68%，そのうち名詞のみを対象とした一致度で約 85.8% にとどまったものの，学習の基準としたエントロピーにおいては 51,580 と，教師あり学習の 52,574 を上回る性能を示し，有効性を確認することができた．

REFERENCES

[1] 下畑 光夫, 隅田 英一郎: “語境界の差異を伴う形態素情報変換,” 第 7 回言語処理学会年次大会, pp. 18-21 (2001).

[2] 飯塚 泰樹: “接続確率最小法による教師なし単語分割,” 情処学自然言語処理研報, 2000-NL-139, pp. 33-40 (2000).

[3] I. J. Good: “The Population Frequencies of Species and Estimation of Population Parameters,” *Biometrika*, Vol.40, no.3 and 4, pp. 237-264, (1953).

[4] 日本語形態素解析システム茶筅 Ver2.0b6 http://chasen.aist-nara.ac.jp/