音声認識結果の文境界推定における識別モデルの評価

祖父江 翔　　山本 けい子　　田村 哲嗣　　速水 悟
岐阜大学 工学部

1 はじめに

音声認識を活用して，会議録や字幕などを作成するためには，「文」が明示的に示されていない音声認識結果を，後の処理のため一定の塊に整形する必要がある．後の処理の例として，「文」を入力単位としている自然言語処理があげられ，文境界推定が必要である．

これまでに，野村ら[1]は F0 が文頭で高く，文末で低いという特徴に基づいて，文境界を抽出している．福岡ら[2]は識別モデル SVM(Support Vector Machine)を用いて，書き言葉の文境界推定を書き言葉特有の改行を素性として行っている．秋田ら[3]は国会などの会議録の作成を想定した自動整形手法に，文境界推定を行っている．文境界推定を用いた応用例として，堀ら[4]は「世界メディアブラウザ」の適切な表示単位，機械翻訳を適用する単位を求めるために，文境界推定を利用している．

本研究では，文境界推定をラベリング問題として扱い，句点挿入モデルと音声認識結果から得られるポーズ情報を用いる．学習データには，新聞社 Web サイトの記事と日本語話し言葉コーパスを用い，評価には NHK ラジオニュースと NHK ニュース解説の時論公論の音声認識結果を用いた．推定には SVM[5]と CRF(Conditional Random Fields)[6,7]を用いて，それぞれの識別モデルについて評価を行った．

2 識別モデル

識別モデルには，SVM と CRF を用いる．

2.1 SVM

SVM は2値分類で用いられる識別モデルであり，各クラスのサポートベクトルと境界の距離マージンを最大化することで，分離平面を求める．本来，線形分離が不可能な場合でも，カーネル関数を用いて，データを分類することができる．学習および推定は SVM ベースのテキストチャンカである YamCha[8]を用いた．学習に用いたカーネル関数は，2 次の多項式カーネルとした．

2.2 CRF

音声認識などで使われる隠れマルコフモデル(HMM)は，特徴が互いに独立である必要がある．これに対し，CRF はその必要がなく，HMM より細かい特徴の指定が可能である．また，条件付き確率により確率が直接推定できるという特長がある．

CRF は，入力列 x に対する各出力ラベル列 y の条件付き確率 $P_\theta(y|x)$ を表現する．θは学習により求められるモデルのパラメータで，それらをベクトルにしたものが Θ である．位置 i の素性ベクトルを $f(y,x,i)$，それに基づく大域素性ベクトルを $F(y,x) = \sum_i f(y,x,i)$ とすると，$P_\theta(y|x)$ は次式で求められる．

$$P_\theta(y|x) = \frac{\exp(\Theta \cdot F(y,x))}{Z_\theta(x)} \quad (1)$$

$$Z_\theta(x) = \sum_y \exp(\Theta \cdot F(y,x)) \quad (2)$$

正しいラベル列を他のラベル列の候補と弁別するよう学習される．解析では，基本的に式(1)を最大にするラベル列 y を探索すればよい．学習および推定は CRF ベースのテキストチャンカである CRF++[9]を用いた．

3 文境界推定

文境界推定の概要を図１に示す．句点があらかじめ挿入されている，新聞記事と話し言葉のテキストデータから，機械学習により推定モデルを作成する．音声認識のテキストデータに，テキスト情報のみで作成した識別モデルを適用して文境界推定を行う．その結果と音声認識時に得られるポーズ情報を元にして，テキスト情報のみで付与されたラベルを付与し直す．

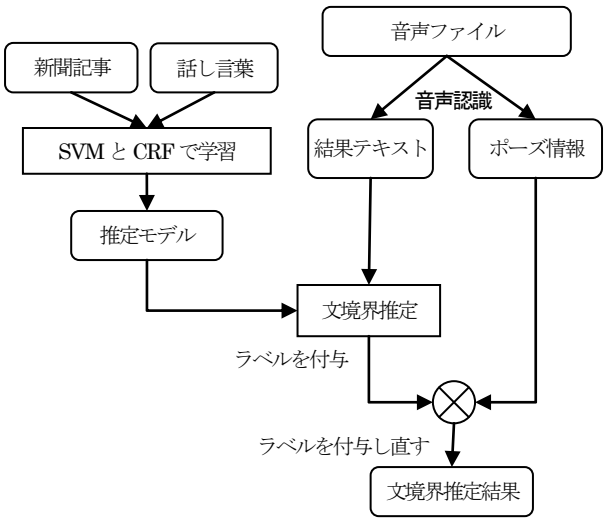

図 1 文境界推定の概要

推定はラベリング問題とする．句点が挿入される単語に対し「文末」ラベルを付与し，それ以外の単語に対し「非文末」ラベルを付与する．学習においても上記 2 つのラベルを用いる．

4 学習のための素性の決定

学習に用いる適切な素性を決定するため，前後の形態素数，使用する単語情報について，CRF による学習で予備実験を行った．予備実験の学習には 5.1 に示すデータ，評価には 6 に示すデータの書き起こしのみを用い，音響的な情報は使用していない．

4.1 前後の形態素数

学習素性の形態素を，前 0~4 形態素，後ろ 0~4 形態素で，変化させて文境界推定の評価を行った．前後形態素の変化による結果を表 1 に示す．

表 1 前後の形態素数による文境界推定結果(F 値)

前＼後	0	1	2	3	4
0	41.1	74.3	86.9	74.9	74.4
1	82.5	87.3	87.6	87.8	88.4
2	82.9	86.7	**89.5**	87.3	87.2
3	81.9	87.8	88.6	88.8	88.9
4	80.9	87.4	87.5	88.3	86.8

表 1 より，前後 2 形態素を学習時の素性にした場合が一番良い結果が得られた．

4.2 単語情報

MeCab[10]の出力(表層形，品詞，品詞細分類 1，活用形)の中で，使用する素性を組み合わせて文境界推定の評価を行った．素性は①表層形のみ，②表層形＋品詞，③表層形＋品詞＋品詞細分類 1，④表層形＋品詞＋活用形，⑤表層形＋品詞＋品詞細分類 1＋活用形の 5 パターンで比較した．5 パターンについての結果を表 2 に示す．

表 2 単語情報の変化による文境界推定結果

パターン	F 値
①表層形のみ	88.3
②表層形＋品詞	**89.5**
③表層形＋品詞＋品詞細分類 1	**89.5**
④表層形＋品詞＋活用形	80.4
⑤表層形＋品詞＋品詞細分類 1＋活用形	**89.5**

表 2 より，②③⑤で良い結果が得られた．素性数が少ない場合，計算時間が短くなるため，「表層形＋品詞」を素性に用いることとした．

5 モデルの作成・ポーズ上を用いた推定

5.1 文境界の学習

学習データから文境界推定モデルを作成する．学習データには，新聞社(5 社)の Web サイトの記事(2008 年 5 月~2008 年 11 月に掲載された 96,956 件のうちランダムに抽出した 5 万文)と，日本語話し言葉コーパス(CSJ)[11]の「文末」タグ付き講演と RWCP 会議音声データベース[12]の 2 つ(1.1 万文)を用いる．素性には，予備実験より，学習する前後 2 形態素と品詞情報(表層形，品詞)を用いる．品詞情報は形態素解析エンジン MeCab により自動付与する．

5.2 ポーズ情報

発話区間の終わり 3 形態素は，音声認識率が全体と比べて低い(10%程度)．識別モデルによるテキスト情報のみの推定だけでは，音声認識誤りに対して推定が困難であるため，音声認識の際に検出されるポーズ情報を，文境界推定の情報に加える．

文境界推定は，識別モデルの推定結果とポーズ情報の論理積によって行った．

6 評価

音声認識エンジンは，Julius[13]を用いる．音声区間ごとに，ファイルを分割して音声認識を行った．分割されたファイルの平均長は，8.74 秒であった．今回は，この分割情報をポーズ情報として使用した．文境界推定性能を評価するデータとして，比較的新聞記事に近い発話内容の NHK ラジオニュース(Web で配信)3 回分(10 分 1 回，15 分 2 回)と，ニュース解説の時論公論，「補正成立 解散政局の行方」と「国内排出権取引 実施へ」の 2 回分(各 10 分)を用いた．音声認識率は，ラジオニュース 3 回の平均 73.3%で，ニュース解説 2 回の平均 69.5%であった．

6.1 評価方法

評価には，再現率，精度，F 値を用いる．文境界の正解データは人手で句点を挿入したものを使用する．

$$\text{精度} = \frac{\text{人手で挿入した句点と本手法で挿入した句点との一致箇所数}}{\text{本手法で挿入した句点数}} \quad (3)$$

$$\text{再現率} = \frac{\text{人手で挿入した句点と本手法で挿入した句点との一致箇所数}}{\text{人手で挿入した句点数}} \quad (4)$$

$$\text{F 値} = \frac{2\times\text{精度}\times\text{再現率}}{\text{精度}+\text{再現率}} \quad (5)$$

6.2 結果

表 3,4 に書き起こし，表 5,6 に音声認識結果における文境界推定の再現率，精度，F 値の平均値を示す．

表 3 文境界推定結果(書き起こし:ラジオニュース)

ポーズ情報	モデル	精度	再現率	F 値
なし	SVM	79.0	97.8	87.3
	CRF	82.6	97.8	89.5
あり	SVM	98.4	97.8	98.1
	CRF	98.8	97.8	98.3

表 4 文境界推定結果(書き起こし:ニュース解説)

ポーズ情報	モデル	精度	再現率	F 値
なし	SVM	65.8	97.2	78.5
	CRF	69.9	97.2	81.2
あり	SVM	100.0	88.2	93.8
	CRF	100.0	88.2	93.8

表 5 文境界推定結果(音声認識:ラジオニュース)

ポーズ情報	モデル	精度	再現率	F 値
なし	SVM	73.3	89.4	80.5
	CRF	78.7	88.1	83.1
あり	SVM	93.3	89.4	91.3
	CRF	94.0	88.1	90.9

表 6 文境界推定結果(音声認識:ニュース解説)

ポーズ情報	モデル	精度	再現率	F 値
なし	SVM	65.9	72.0	68.5
	CRF	67.4	70.4	68.7
あり	SVM	96.3	64.6	77.3
	CRF	98.1	63.0	76.7

ポーズ情報を用いた場合，どちらの識別器においても文境界推定の性能が向上した．F 値に関しては識別モデル間で，ほとんど差が見られなかった．

6.3 推定誤り

推定誤りを起こすパターンは，両方の識別モデルで同じような傾向であった．推定誤りを起こしたパターンの例をいくつか示す．

(1) 音声認識の誤りにおける誤り

推定誤り例:

① ○したいとしています → ×市内の態度

② ○大崎市 → ×恐ろしい

例①は，本来は文末であるが，音声認識誤りによって文末でない言葉に変わり，文末と推定できなかった．例②は，本来文末ではないが，音声認識誤りで文末のような言葉に変わり，文末と推定された．書き起こすと文の途中に当たるため，ポーズ情報を用いて「文末」と推定されないが，例②の場合はポーズがあったため推定を誤った．

(2) 音声認識結果は正しいが推定の誤り

推定誤り例:

① ことになります

② ものがありますね

例①や例②は，音声認識は誤っていないが，推定を誤った場合である．これは，ポーズ情報によって「文末」と判定されなかったもので，テキスト情報での推定は正しかった．

6.4 考察

認識誤りの傾向が, SVM と CRF で同じような傾向となった．どちらのモデルも，テキスト情報のみの場合に, 再現率はほとんど差がなく, 精度は CRF が良いというものであった．ラジオニュースとニュース解説で比較すると，ラジオニュースの原稿は, 新聞記事と近く，学習データに新聞社 Web サイトの記事を用いたため, 書き起こしで 80%を超える精度となった．ニュース解説は，ラジオニュースとやや内容が異なり，学習データに話し言葉コーパスを用いたが，(CSJ は講演による話し言葉，RWCP は会議音声のため)コーパスとの内容の違いが，精度の低下を起こしたと考えられる．また，ポーズの傾向の違いが差を生じさせたと考えられる．ラジオニュースでは, ニュース原稿の 1 文を途中で止めずに話されることがほとんどである．ニュース解説では, 話し言葉特有の言い淀みがあり，その部分がポーズとして検出され，推定誤りとなった場合や，文境界の部分で間を置かずにしゃべったことで，ポーズが検出されず推定誤りとなった場合もあった．

7 おわりに

文境界推定を SVM と CRF の 2 つの識別モデルで評価を行った．予備実験で，前後の単語数や単語情報といった学習素性を決定した上で評価を行った．テキスト情報のみでは, SVM より CRF の精度が高く，ポーズ情報を加えるとほとんど差がなかった．

今回の評価では，ニュースなど新聞記事に近いデータを扱ったため，講義や会議のような話し言葉など他のデータの種類で評価をしていく必要がある．用いた音声は比較的雑音が少ないデータのため，文境界推定においてポーズ情報が有用であったが，雑音下における推定では，他の韻律情報を用いた手法を検討する必要がある．識別モデルだけでは，推定が難しい音声認識誤りについて，誤り傾向を分析した上で誤り訂正を行って，識別モデルを適用する方法が考えられる．推定結果を用いて，映像コンテンツの字幕およびキーワード抽出[14]に，適用することも考えられる．

参考文献

[1] 野村和弘, 河原達也, 堂下修司. “講義の自動アーカイブ化のための韻律情報を用いた講義音声の文境界抽出.” 電子情報通信学会技術研究報告, SP98-80, 1998.

[2] 福岡健太, 松本裕治. “Support Vector Machine を用いた日本語書き言葉の文境界推定.” 言語処理学会第 11 年次大会論文集, 2005.

[3] 秋田裕哉, 河原達也. “会議事録作成のための話し言葉音声認識結果の自動整形.” 秋季音響学会講演論文集, pp.103-104., 2008.

[4] 堀貴明, ほか. “「世界メディアブラウザ」ー音声認識と統計翻訳に基づく多言語動画コンテンツ検索／閲覧システム.” 第 2 回音声ドキュメントワークショップ講演論文集, pp.59-66., 2008.

[5] 荒木雅弘, “フリーソフトで作る音声認識システム”, 森北出版株式会社, 2007.

[6] John, Lafferty, McCallum Andrew, and Pereira Fernando. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.” *In Proceeding of the 18th International Conference on Machine Learning*, 2001.

[7] 坪井祐太, 鹿島久嗣, 工藤拓. “言語モデルにおける識別モデルの発展ーHMM から CRF まで.” 言語処理学会第 12 回年次大会チュートリアル, 2006.

[8] YamCha: http://chasen.org/~taku/software/yamcha/

[9] CRF++: http://crfpp.sourceforge.net/

[10] 形態素解析エンジン MeCab, http://mecab.sourceforge.net/

[11] 古井貞熙, 前川喜久雄, 井佐原均. “科学技術振興調整費開放的融合研究推進制度-大規模コーパスに基づく『話し言葉工学』の構築.”音響学会誌, Vol.56-11, pp.725-755, 2000

[12] 田中和世, 伊藤克亘, 岡隆一, 松村博. “RWCP 会議音声データベース 2001.” 第 16 回人工知能学会全国大会論文. 2002

[13] 大語彙音声認識エンジン Julius, http://julius.sourceforge.jp/

[14] 岡本昌直, 上地春奈, 山本けい子, 田村哲嗣, 速水悟. “キーワード抽出による映像コンテンツの理解支援とその心理的評価.” 言語処理学会第 15 年次大会論文集, 2009.