隠れマルコフモデルに基づく音声認識結果からの固有表現抽出

長谷川 隆明　　林 良彦
日本電信電話株式会社　NTTサイバースペース研究所
hasegawa.takaaki@lab.ntt.co.jp

1．はじめに

ブロードバンドネットワークの普及に伴い、映像コンテンツを検索してオンデマンドで視聴するというニーズが高まりつつある。ユーザが映像コンテンツの必要なシーンを高精度に検索できるようにするためには、映像コンテンツに予め内容に関するメタデータを付与しておかなければならない。しかしながら、コンテンツ自身の内容に関するメタデータを付与するのは人手に頼らざるを得ないため、非常にコストがかかることが問題となっている。そこで、我々は内容に関するメタデータとしてコンテンツの音声に出現する固有表現に着目した。本稿では、固有表現に関するメタデータを用いて映像コンテンツに対するアノテーションを行うことを目的として、音声認識結果から固有表現を抽出する方法を提案する。

本稿では、まず第2節で音声言語処理における固有表現抽出の課題について説明し、第3節で音声認識結果から固有表現をロバストに抽出する方法について述べる。第4節で評価実験の結果について報告し、第5節で結論を述べる。

2．音声言語処理における固有表現抽出の課題

固有表現とは、コンテンツの内容を表現する基本的な要素である人名・地名・組織名などを指す。映像コンテンツの音声には固有表現が多く出現するという前提のもとで、本稿ではTVニュース番組を処理の対象としている。適切に準備された音響・言語モデルのもとで、大語彙連続音声認識エンジン[1]により、背景雑音がないスタジオで、ニュースという分野に限定され、訓練されたアナウンサーの発声する音声は、高速かつ高精度に文字化される。しかしながら、処理速度や精度の問題から、音声認識エンジンが文字化するために利用する辞書に登録されている単語数は制限されている。特に、低頻度の単語や未知語となることが多い人名・地名・組織名などの固有表現については登録されない。このため、認識対象外となる辞書外単語(Out of Vocabulary: OOV)の問題が生じる。つまり、これらが音声に出現する場合には正しく文字化されず、辞書に登録されていて音響的にも言語的にも近いと計算された別の単語として誤って文字化されてしまう。

一方、これまでの固有表現抽出技術においては、誤りのない新聞記事等のテキストからの固有表現抽出に関する報告が主流であった。固有表現抽出の代表的な手法として、二値分類器であるSupport Vector Machine (SVM)を用いた方法[2]や、隠れマルコフモデルを用いた方法[3]があげられる。対象をテキストから音声に移すと、音声認識後のテキストである音声トランスクリプションを単純に従来と同様のテキストとみなして固有表現抽出を行うアプローチ[4]が多く、OOVの問題は解決されていない。また、音声認識エンジンの出力する情報を利用してシミュレートした認識誤りや実際の認識誤りをモデル化した報告[5]もあるが、適合率は上昇するが再現率が低下するという問題を抱えている。

このように、従来の固有表現抽出技術では音声認識におけるOOVの問題を解決することは難しい。ところが、必ずしも音声に出現する固有表現を正しく文字化できなくても、固有表現のタイプと発声される開始時刻と終了時刻が同定できれば、アプリケーションによっては十分に利用できるメタデータを提供できる。あるいは、後処理を行う

Named Entity Extraction from Speech Recognition Based on Hidden Markov Model by Takaaki Hasegawa and Yoshihiko Hayashi. NTT Cyberspace Laboratories, NTT Corporation.

ことによって正しく文字化できる可能性もある。本稿では、このような考え方に基づき、音声認識エンジンが出力する認識結果から固有表現とその発声区間を抽出する方法を提案する。

3．音声認識結果からの固有表現抽出

一般的に SVM を用いる方法の方が精度は高いが、隠れマルコフモデル(HMM)は音声認識等の誤りに対する頑健性が要求されるタスクにおいて実績があるため、本稿では NTT が開発した HMM を用いた多言語固有表現抽出エンジン[6]を採用している。

HMM を用いた固有表現抽出では、HMM のそれぞれの状態がひとつの固有表現クラスに相当し、自身の状態も含めすべての他の状態に遷移することが可能である。状態間の遷移および各状態内における単語選択を含めて、文頭から文末まで到達可能なすべての経路についての遷移確率が最大となる経路、すなわち固有表現クラス付き単語列が決定される。このため、事前に必要な言語モデルは、固有表現クラス付き bi-gram となる。

音声認識結果からの固有表現抽出を行うために、まず固有表現クラス付き bi-gram を学習する。このために、ニュース音声とそのトランスクリプションの対を用意する。トランスクリプションは、人手で作成され、固有表現のタグが付与されるものとする。一方、ニュース音声は音声認識エンジンにより単語列の N-best 候補が出力される。モデルの学習時には、ニュース音声の音声認識結果とこれに対応するトランスクリプションとを対比させることにより、固有表現クラス付き単語 bi-gram を得る。具体的なイメージとしては、図1に示すようになる。音声認識結果は単語列の一位候補のみを対象として、トランスクリプションは形態素解析により得られる単語列を対象とする。音声認識結果とトランスクリプションの DP マッチングを行い、両者の編集距離が最小となるように対応付ける。このとき、単語の表記や読みの重なり度合いを基準として用いる。対応付けの結果、トランスクリプションに付与されている固有表現クラスを音声認識結果の単語列に付与する。

実行時には、音声認識エンジンから得られるニュース音声の単語列の N-best 候補から変換される単語グラフを入力とし、学習された固有表現クラス付き単語 bi-gram に基づいて、HMM により

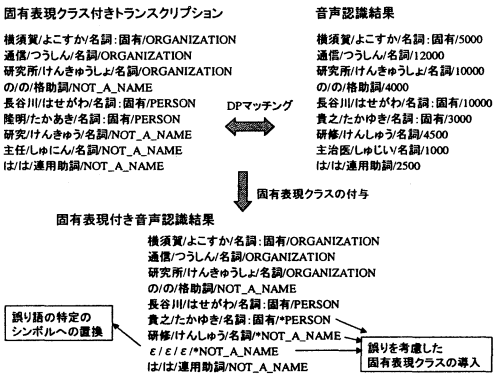

図 1　音声認識結果への固有表現クラスの付与例

最尤の固有表現クラスを付与する。音声認識エンジンは単語の開始時刻と終了時刻を出力するが、単語グラフに変換する際に、音声認識エンジンが出力する単語の区間と次の単語の区間との間に僅かな時間の隙間が生じることがあるので、このような場合にはトランスクリプションとの親和性の高い読点を入れることにより、単語グラフ内のすべての単語は途切れないようにしている。

ロバスト性を高めるために、我々は固有表現クラス付き bi-gram モデルを学習するときに、以下に示すような2つのアイデアの導入を提案する。

3．1　誤り語の特定のシンボルへの置換

ロバスト性を高めるためのひとつの方法として、我々は認識誤り語を特定のシンボルへ置換してから固有表現クラス付き単語 bi-gram を学習する方法を提案する。認識誤り語を共通のシンボルで表しておくことにより、認識誤りに見られる共通の特徴をモデル化できると考えている。

認識誤り語を検出する方法はいくつか考えられるが、最も単純な方法は、DP マッチングで得られた音声認識結果とトランスクリプションの単語列の対応付けにより、各単語の認識誤りを検出する方法である。もうひとつの方法は、音声認識エンジンによって出力される各単語の信頼度スコアを用いる方法である。信頼度スコアは、音声認識エンジンが出力する N-best 候補間のスコア差に基づく信頼度尺度であり、値が大きいほど信頼できる。そこで、信頼度スコアがある閾値よりも低い単語を認識誤り語と見なす。さらにこれらの二

つの方法を組み合わせる方法もある。

また、いずれの方法でも、実行時においては、信頼度スコアを手がかりにして擬似的に認識誤りと見なして共通のシンボルへの置換を考慮する必要がある。固有表現クラス付き単語 bi-gram を参照するときに、元々の単語と認識誤りを表す共通のシンボルを用いたときの確率の大きい方の固有表現クラス付き単語 bi-gram を採用する。

3．2　誤りを考慮した固有表現クラスの導入

ロバスト性を高めるもうひとつの方法として、認識誤りを考慮した固有表現クラスの導入を提案する。従来の固有表現クラスに加え、認識誤り時の専用の固有表現クラスを設けることにより、認識の正誤まで考慮したより精密な文脈をモデル化できると考えている。

我々が提案する方法は、通常本来発声されている単語の持っている固有表現クラスを単純に認識誤り語に付与するのに対し、固有表現クラスのクラス数を倍にして、正しく認識された語に付与する固有表現クラスと認識誤り語に付与する固有表現クラスを区別することである。例えば、人名を表す PERSON という固有表現クラスに対して、認識誤りだが人名を表す*PERSON という固有表現クラスを新たに設ける。

4．評価

我々は提案手法を評価するために、TV ニュース番組を対象とした実験を行った。TV ニュースの同一番組一ヶ月分の映像と、映像の音声に対応した人手による固有表現タグ付きトランスクリプションを用意した。固有表現タグは IREX ワークショップ [7] の仕様に準拠し、PERSON・LOCATION・ORGANIZATION・DATE・TIME・MONEY・PERCENT・ARTIFACT の 8 種類としている。学習データは、29 日分の 9,955 文、固有表現 23,400 個である。評価データは、学習データの日付より後の 2 日分の 716 文、固有表現 1,767 個のうち開始時刻と終了時刻が判別できる 1,334 個とし、各固有表現の開始時刻と終了時刻も人手で付与した。音声認識における学習データの単語正解率は、アナウンサーの音声部分 4,732 文については高精度だが、レポーターの音声や BGM の存在する部分が含まれるので、全体としては 55.1%であった。

誤り語の特定のシンボルへの置換と認識誤りを考慮した固有表現クラスの導入について、これらの組み合わせも含めて実験を行った結果を、F 値、再現率、適合率の順でそれぞれを表 1 に示す。評価基準は、完全一致（固有表現の種別、表記、読み、開始時刻、終了時刻のすべてが一致）、表記以外一致、区間一致（表記と読み以外が一致）とした。ただし、時刻の一致については、音声認識エンジンの出力に幅が生じているため、前後 50ms の幅を持たせている。実験では、学習データを 6,788 文（20 日分）とこれを含む 9,955 文（29 日分）の二つを用意して、学習データの量についてどの程度の差が出るかの検証も試みた。なお、ベースラインには、固有表現タグ付きトランスクリプションからそのまま固有表現クラス付き bi-gram を学習したものを使用している。また、いずれも固有表現クラス付き bi-gram の学習時にはトランスクリプションから得られる固有表現クラス付き単語 bi-gram も合わせて使用している。学習時と実行時とも、信頼度スコアは単語の読みの長さで正規化したものを使用し、学習時に置換する低信頼語および実行時に誤り語とする閾値はすべて同じに設定した。実験の結果から、誤りクラスの導入により再現率・適合率とも精度が上がることと、誤り語の特定のシンボルへの置換により適合率が上がることが確認できた。この傾向は、特に区間一致について顕著に表れた。また、いずれも学習データの量が増えると、ベースラインに比べて精度向上の勾配が大きいことも検証できた。

誤り語の置換では、実行時に信頼度スコアが低いため置換される単語が実際に認識誤りである場合には、誤って固有表現に同定される危険性を回避できる一方、正しく認識された単語でも信頼度スコアが低いと認識誤りと見なすため再現率は僅かに低下する。信頼度スコアによる置換が健闘しているのは、実行時において音声認識エンジンの認識誤りの特徴をうまくシミュレートできたからだと考えられる。ただし、低信頼度とする閾値によってかなり結果が変わってくるので、閾値は実験を通して慎重に決定する必要がある。また、誤りを考慮したクラスの導入は、認識誤りでかつ固有表現であるクラスが実行時に誤認識された固有表現を捉えることにより、再現率が高くなり、同時に、認識誤りでかつ固有表現以外であるクラスが実行時にご認識された固有表現でない単語を吸

表 1　誤り語の置換と誤りを考慮したクラスの導入の実験結果（F 値／再現率／適合率）

	ベースライン		誤り＆低信頼語置換		低信頼度語置換	
学習データ量	6,788 文	9,955 文	6,788 文	9,955 文	6,788 文	9,955 文
完全一致	56.7/59.9/53.8	59.8/64.5/55.7	56.6/56.1/57.1	60.3/60.0/60.6	56.6/56.0/57.2	60.0/59.7/60.4
表記以外一致	58.2/61.5/55.2	61.3/66.2/57.2	59.9/59.4/60.5	63.7/63.4/64.0	60.0/59.4/60.6	63.6/63.2/64.0
区間一致	59.3/62.7/56.3	62.5/67.4/58.2	63.4/62.8/64.0	67.5/67.2/67.8	63.5/62.8/64.2	67.3/66.9/67.8
	誤りクラス		誤りクラス＋誤り＆低信頼語置換		誤りクラス＋低信頼度語置換	
学習データ量	6,788 文	9,955 文	6,788 文	9,955 文	6,788 文	9,955 文
完全一致	56.9/59.8/54.3	61.5/65.4/58.0	56.0/57.3/54.8	60.9/63.3/58.7	56.7/57.7/55.6	61.4/63.5/59.5
表記以外一致	60.3/63.3/57.5	65.1/69.3/61.4	59.5/60.9/58.2	64.8/67.3/62.4	60.3/61.4/59.2	65.3/67.5/63.2
区間一致	63.3/66.5/60.3	67.7/72.0/63.8	62.4/63.8/61.0	67.7/70.3/65.2	63.4/64.5/62.2	68.5/70.8/66.3

収することにより、適合率が高くなったと考えられる。また、誤りを考慮したクラスと低信頼度語の置換との組み合わせのF値が最高だったのは、誤認識と信頼度スコアというそれぞれ別の尺度を独立に適用したことによる効果だと考えられる。さらに、いずれの方法でも、学習データ量の増加に伴う精度の上昇幅がベースラインに比べて大きかったことから、実験に用いたデータ量ではまだ飽和していないと考えられる。誤り語の特定のシンボルへの置換では、元々制限されている語彙の中で多くの単語が特定のシンボルに集約され、一方、誤りを考慮した固有表現クラスの導入では、データ量はそのままでクラス数が倍増する。このため、いずれもデータスパースネスの問題が懸念される。しかし、さらに多くのデータが用意できれば、さらなる精度の向上が期待できる。

5．おわりに

本稿では、映像コンテンツに対するメタデータを付与するという観点から音声に出現する固有表現に注目し、音声内容を音声認識エンジンにより認識させ、認識誤りを含む音声認識結果からロバストに固有表現を抽出する方法を提案した。キーとなるアイデアは、誤り語の特定のシンボルへの置換と、誤りを考慮した固有表現クラスの導入である。TV ニュースを対象とした実験結果から、ベースラインを上回る精度が得られ、特に固有表現の音声区間の同定に対しての有効性が確認できた。映像コンテンツの音声に出現する固有表現にメタデータを付与するという目的に対して、提案手法はより有効であると言える。

今後は、さらにデータを増やし、精度を高めるための分析を進めるとともに、同定された固有表現の音声区間を映像コンテンツの検索において生かすための有用な方法について検討していく。

参考文献

[1] 野田，山口，大附，小川，中川，今村： 音声認識エンジン VoiceRex の開発，音響学会 1999 年秋季研究発表会.

[2] 山田，工藤，松本： Support Vector Machines を用いた日本語固有表現抽出，情処研報 NL142-17，pp. 121-128，2001.

[3] Bikel, D. M., Miller, S. Schwartz, R. and Weischedel, R.: Nymble: A High Performance Learning Name-Finder, ANLP' 97 pp. 194-201, 1997.

[4] Kubala, F., Schwartz, R., Stone, R. and Weischedel, R.: Named Entity Extraction from Speech, 1998 Darpa Broadcast News Transcription and Understanding Workshop, 1998.

[5] Palmer, D. D., and Ostendorf, M., Burger, J. D.: Robust Information Extraction From Spoken Language Data, Eurospeech99, pp. 1035-1038, 1999.

[6] 齋藤，永田: HMM に基づく多言語固有表現抽出システムの開発，言語処理学会第 9 回年次大会，2003.

[7] IREX 実行委員会:IREX ワークショップ予稿集，1999.