Confusion Network Based Vector Space Model を用いた音声ドキュメントの検索*

○胡　新輝，呉　友政，柏岡　秀紀（NiCT/ATR）

1　はじめに

本論文では、Confusion Network（CN）をベースにしたベクトル空間モデルを用いて、音声ドキュメントの検索手法を提案する。Confidence に依存するtfの計算と出現頻度に依存するidfの計算を結合する事によって、良い検索性能を得られる事が我々の研究によって確認された[1]。本論文では、これらの計算方法を使って、異なる音声認識率の結果に対して検索実験を行う。日本語話し言葉コーパス（CSJ）を対象にした実験で、本手法は、高い認識誤り率（WER）の場合でも、1-bestの認識結果を用いる場合と比べて、6%以上の性能改善が得られた。

2　Confusion Network をベースにしたベクトル空間モデル

2.1 CN の導入

近年、目覚ましく進歩した音声認識技術のおかげで、1-best の結果でも、ニュース放送など比較的に整った音声の検索は、満足度の高い検索性能が得られる[2]。しかし、自由発話の音声では、認識性能がなかなか改善できなく、単語の誤り率（WER）が 30%より大きい場合が多い。このような場合、失われた情報が多いため、1-best のみで検索を行うと、十分な検索性能が得られない。

この問題に対処するために、多数の認識の仮説を活用することを考える。音声認識の Lattice を使えば、Lattice 中に含まれる大量の不確定な内容を利用することができ、1-best で失われた有用な情報を補うことができると考えられる。我々は Mangu [3] が提案した CN を採用して、自由発話の音声データに対して、情報検索の研究を試みている。

音声認識器が出力する認識結果の Lattice に変換をかけて[2], 図１のような単語 CN を得る。このように CN は、多数の alignment（図のノードに相当）と、alignment 間に複数の競合する単語候補を含んでいる。これらの候補は confusion set という。Confusion set の各単語は、それぞれの事後確率を持つ。以下、CN に対してベクトル空間モデルを定義する。

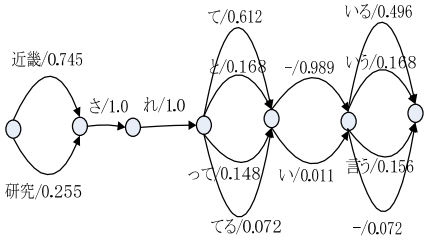

Fig. 1 CN の例

2.2 ベクトル空間モデルの定義

(1) 基本定義

D ― CNが表現するドキュメント。

P(w|o, D) ― ドキュメントDの位置oに単語wが発生する事後確率。

(2) tf･idf のベクトル空間モデルの計算

我々は、テキスト検索によく使われているベクトル空間モデルである tf・idf 方式を CN に適用して、音声ドキュメントのキーワードの検索を行った。実験の結果として、以下の結論を得た[1]。

- tfの計算の際、confidenceが単語のランクより有効である。
- idf計算の際、単純な単語カウントの方がconfidenceの使用より有効である。

従って、以下の様に、tf と idf を計算する。

$$tf(w,D) = \sum_{i=1}^{|occ(w,D)|} P(w \mid o_i, D) \quad (1)$$

$$idf(w) = \log(N / \sum_{D \in C} O(w,D)) \quad (2)$$

ここで、$o(w,D) = \begin{cases} 1 & if \quad P(w|o,D)>0 \\ 0 & otherwise \end{cases}$

Cは、すべてのドキュメントの集合である。

(3) 検索 Query とドキュメント間の類似度の計算は、下式で行う。

* Spoken Document Retrieval Using Vector Space Model Based on Confusion Network, Xinhui Hu, Youzheng Wu, Hideki Kashioka　(NiCT/ATR).

$$sim(d_i,q)=\frac{\sum_{w_q\in V} tf(w_q\mid d_i)\cdot C(w\mid q)}{\sqrt{0.8\cdot avdl+0.2\cdot|d|}} \quad (3)$$

ここで、c(w|q)は、検索クエリ（Query）に現われた単語ｗの頻度である。|d|は、ドキュメントｄに含まれるCNの数である。

$avdl\left(=\frac{1}{|C|}\sum_{d'\in C}|d'|\right)$はドキュメント長の平均である。

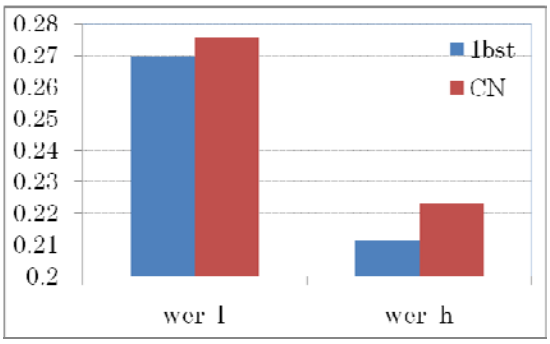

Fig. 2 異なる認識率時の検索性能

3　実験

3.1 データと実験条件

日本語話し言葉コーパス（ＣＳＪ）の「学会講演」と「模擬講演」の２７０２音声ファイルを使って、検索実験を行った。両方の講演を合わせて６００時間を超える。上述の音声データに二種類の認識条件で認識を行い、それぞれ単語誤り率(WER)がwer_h=42．13とwer_l=30．25の認識結果のCNを得た。

3.2 評価尺度

検索性能の評価尺度には、0.0から1.0まで0.1刻みの各再現率における精度、それらの11点の再現率レベルにおける補間精度IPを平均した11点の平均精度AP, 及びAPを全検索クエリで平均したMAP(Mean Average Precision)を用いた。

3.3 Queryと適合性判定

Query文は、CSJテストコレクション[4]の中から解答が存在する３９個の質問文に対して、形態素解析をした後、人手でストップワード、名詞以外のものを排除して作った。また、今回の検索は、既知語の検索であるため、Queryのキーワード列に、未知語が含まれないようにした。適合性の判定作業は、TREC の評価ツールの trec_eval で行った。検索対象の単位としては、講演単位である。実際の判定は、適合（Relevant）と不適合（Irrelevant）の二種類で行い、テストコレクションの“部分適合”は適合と見なした。

3.4 実験結果

図2には、二つ認識率の条件での検索結果を示す。それぞれ1-bestに対する検索結果と比較している。どちらの認識率の場合でも、ＣＮの使用により、検索性能が改善している。また、wer_lとwer_hのそれぞれの改善は2.2%と6.2%であり、認識誤りが高い時の改善が大きい。

4　まとめ

本研究では、テキスト検索によく使われているベクトル空間モデルであるtf・idf方式をCNに適用する方法を提案し、それを用いて異なる音声認識率の音声ドキュメントの検索を作った。実験の結果として、以下の結論を得た。

- CNをベースにしたベクトル空間モデルは、1-bestの結果より、検索性能が良い。
- 認識率が悪い時、CNは効果を発揮する。

今後、より多くの認識率の異なる音声データを使って、実験を行う予定である。また、単語以外の検索単位も考慮して、認識誤りにより全体の性能の悪化に対する対策を研究する。

謝辞

本研究は、文部科学省科研補助金「特定領域研究」の“情報爆発時代に向けた新しいIT基盤技術の研究”－「音声ドキュメントの検索に関する研究」により実施したものである。

参考文献

[1] 胡, et al. “Latticeをベースにした音声ドキュメントの検索に関する研究”, 音講論集, 1-Q-32, Mar. 2008.

[2] Garofolo, et al. ,“The TREC Spoken Document Retrieval Track: A Success Story.” Proc. RIAO Conference 2000: pp 1-20, 2000.

[3] L. Mangu, et al., “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks,” Computer Speech and Language, vol. 14, no.4, pp.373-400, Oct, 2000.

[4] 秋葉友良、et al. “音声ドキュメント検索テストコレクションの試作と検索性能評価、“第1回音声ドキュメント処理ワークショップ, 2007.