고속 음성 문서 검색을 위한 Expected Matching Score 기반의 문서 확장 기법*)

서민구, 정규준, 오영환
한국과학기술원 전자전산학과 전산학전공 음성인터페이스 연구실

Expected Matching Score Based Document Expansion for Fast Spoken Document Retrieval

Minkoo Seo, Gue Jun Jung, and Yung-Hwan Oh
Voice Interface Lab., Div. of Computer Science, EECS Dept., KAIST
E-mail : mkseo@bulsai.kaist.ac.kr, sylph@bulsai.kaist.ac.kr, yhoh@bulsai.kaist.ac.kr

Abstract

Many works have been done in the field of retrieving audio segments that contain human speeches without captions. To retrieve newly coined words and proper nouns, subwords were commonly used as indexing units in conjunction with query or document expansion. Among them, document expansion with subwords has serious drawback of large computation overhead. Therefore, in this paper, we propose *Expected Matching Score* based document expansion that effectively reduces computational overhead without much loss in retrieval precisions. Experiments have shown 13.9 times of speed up at the loss of 0.2% in the retrieval precision.

I. 서론

음성 문서 검색(Spoken Document Retrieva; SDR)을 효율적으로 수행하기 위해서는 크게 두 가지 문제를 해결해야한다. 첫째는 음성 문서 내 미등록 어휘(Out-of-Vocabulary; OOV)가 포함되어있다는 것이고, 둘째는 음성 인식의 결과가 100% 정확하지 않다는 점이다. 이 중 첫 번째 문제를 해결하기 위해서 음소 n-gram과 같은 서브 워드를 인덱싱 단위로 사용하는 기법이 주로 사용되어 왔으며, 두번째 문제의 해결을 위해서는 질의 확장이나 문서 확장 기법이 널리 사용되어왔다[1,2,3,4,5,6].

질의 확장 기법에서는 검색 정확도의 향상을 위해 질의에 새로운 질의어를 추가하는 기법이 이용된다. 이러한 검색 기법의 대표적인 방법으로 음성 인식기가 인식과정에서 혼동하기 쉬운 단어를 질의에 추가하는 방식이 이용되고 있다[3,4]. 이 방식은 검색 정확도는 향상시키나, 높은 정확도를 위해 많은 수의 질의어를 추가시켜야하는 한계가 있다. 이 외에도 검색 대상과 내용이 유사한 이차 코퍼스(secondary corpus)로부터 질의어와 관련된 단어를 추출하여 이를 질의어에 추가하는 기법 역시 제안되었다[7,8]. 하지만 이 기법은 질의어와 관련성이 낮은 단어가 추가될 가능성이 있으며, 질의어와 관련된 단어를 찾는 문제 역시 쉽지 않다는 한계가 있다[8].

이차 코퍼스는 또한 문서 확장 기법에서도 사용될 수 있다. Amit Singhal 과 Fernando Pereira[9]는 음성 문서와 관련성이 높은 문서로부터 단어를 추출하고 이를 음성 문서에 추가하는 기법을 사용하였다. 그러나 이 방법에서도 이차 코퍼스 확보 문제와 문서간의 유사성 평가의 문제가 남게 된다.

혼동 행렬(Confusion Matrix)을 사용한 음소 n-gram 간의 유사 검색[2,4,6]은 음성 문서 검색을 위한 또 다른 문서 확장 기법이다. 이 방법에서는 정확

*) 본 연구는 방위사업청과 국과연의 지원을 받아 2006년도 국방 소프트웨어 설계 특화센터를 통해 수행되었음.

히 일치되는 음소 n-gram 을 찾는 대신, 동적 계획법(Dynamic Programming; DP)을 사용하여 음소 n-gram 간의 유사성을 평가하여, 음소 인식 단계에서 발생하는 삽입, 삭제, 치환 오류를 보정한다. 또한 검색 정확도를 향상시키기 위해 음소 2-gram, 3-gram, ..., n-gram 으로 검색을 수행한 뒤 그 결과의 가중치 합을 질의와 음성 문서간의 관련도로 이용하기도 한다. 그러나 이 기법은 많은 수의 DP를 유발하여, 검색 속도가 느리게 되는 한계가 있다[4].

본 논문에서는, *Expected Matching Score 기반의 음소 n-gram 유사 검색*과 벡터 공간 모델(Vector Space Model)을 사용한 음성 문서 검색 기법을 제안하여, 혼동 행렬을 사용한 음소 n-gram 유사 검색의 검색 속도 문제를 해결한다. 제안된 기법은 질의에서 나타난 음소 n-gram과 음성 문서 내 음소 n-gram 간에 공통적으로 포함된 음소 개수를 기준으로 하여 DP의 수행여부를 결정한다. 실험 결과, 검색 정확도의 저하가 0.2%~0.1%에 불과하면서도 검색 속도는 13.9~4.4배 향상 되었다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 연구에서 제안된 벡터 공간 모델과 음소 n-gram간의 유사 검색에 대해 설명한다. 3장에서는 본 논문에서 제안하는 *Expected Matching Score 기반 유사 검색*을 제안한다. 마지막으로 4장에서는 실험 결과를 보이며, 5장에서는 본 논문의 결론과 향후 연구 방향에 대해 기술한다.

II. 유사 검색 기법

2.1. 검색 모델

본 논문에서는 가장 널리 알려진 검색 모델 중 하나인 벡터 공간 모델을 사용하였다. 제안된 시스템에서는, 사용자의 질의 q 가 주어지면 이를 음소열로 변환한 후 이로부터 가능한 모든 음소 n-gram을 추출한다. 즉, 음성 인식기가 t 개의 음소를 사용한다고 하면 사용자의 질의는 t^n 차원 벡터 $\vec{q}$ 로 변환된다. 이 때, 각 차원에는 해당 음소 n-gram이 q에 나타난 회수가 저장된다. 음성 문서 d 역시 동일한 방법으로 음성 문서 벡터 $\vec{d}$ 로 변환된다. 이후 $\vec{q}$ 와 $\vec{d}$ 간의 관련성은 코사인 유사도(Cosine Similarity)를 사용하여 다음과 같이 평가된다.

$$S_n(q, d) = \frac{\vec{q} \bullet \vec{d}}{\|\vec{q}\| \|\vec{d}\|} = \sum_{i \in q} \frac{q[i]\, d[i]}{\|\vec{q}\| \|\vec{d}\|}$$

Kenny Ng[2], Nicolas Moreau et al.[4]에서는 검색 정확도를 더욱 향상시키기 위해 음소 n-gram의 병합을 제안하였다. 이 기법에서는 서로 다른 길이의 n-gram을 질의와 음성 문서에서 추출한 뒤, $S_n(q, d)$ 의 가중치 합을 최종 관련도로 평가하였다. 이를 식으로 보이면 다음과 같다.

$$S_f(q, d) = \sum_n w_n S_n(q, d) \quad \text{(식 1)}$$

이 때, w_n 은 특정 길이의 n-gram에 대한 선호도를 반영하기 위한 가중치 값이며 [4]에서는 3-gram의 검색 정확도가 다른 n-gram에 비해 우수하다는 점을 반영하기 위해 $w_3 = 0.5, w_4 = 0.2, w_2 = w_5 = w_6 = 0.1$의 값을 사용하였다. 따라서 본 논문에서 역시 $S_f(q, d)$ 및 [4]와 같은 가중치 값을 사용하였다.

2.2. 음소 n-gram 간의 유사 검색

음소 n-gram 의 음성 문서내 출현 회수를 평가할 때, 유사 검색을 통해 잘못 인식된 음소를 보정할 수 있다. 이 방법에서는, 음성 문서 d 에서 음소 n-gram i를 다른 음소 n-gram j 로 오인식할 확률을 이용한다. 이를 이용하여 Kenny Ng[3,4]는 다음과 같은 음소 n-gram 유사 검색을 제안하였다.

$$S_a(q, d) = \sum_{i \in q} \sum_{j \in d} p(j|i) \frac{q[i]\, d[j]}{\|\vec{q}\| \|\vec{d^*}\|} \quad \text{(식 2)}$$

식에서 $p(j|i)$는 i를 j로 오인식 할 확률이며 $\vec{d^*}$는 $p(j|i) \times d[j]$ 값에 따라 새로 계산된 문서 벡터이다.

$p(j|i)$의 계산에서는 길이가 각각 l_i, l_j 인 음소 n-gram i와 j가 주어졌을 때, 음소 혼동행렬을 사용하여 크기 $l_i \times l_j$인 행렬 A를 다음과 같이 DP를 사용해 계산하였다.

$$A(m, n) = \begin{cases} \text{if } m = 0, n = 0, & 1. \\ \text{if } m = 0, n > 0, & A(0, n-1) \times C(\lambda, j[n-1]) \\ \text{if } m > 0, n = 0, & A(m-1, 0) \times C(i[m-1], \lambda) \\ otherwise, & \\ \quad \max \begin{cases} A(m-1, n) \times C(i[m-1], \lambda) \\ A(m-1, n-1) \times C(i[m-1], j[n-1]) \\ A(m, n-1) \times C(\lambda, j[n-1]) \end{cases} & \end{cases}$$

이 식에서 $C(r, h)$는 참조 음소(reference phoneme) r이 주어졌을 때, 음성 인식 수행 뒤 가설 음소(hypothesis phoneme) h를 관찰할 확률을, λ는 참조 또는 가설 음소가 없을 때를 의미한다. 행렬 A의 계산이 끝나면 $A(l_i, l_j)$ 값이 $p(j|i)$에 해당하게 된다.

이와 같은 유사 검색이 특히 음소 n-gram의 병합과 함께 사용되었을 경우 검색 정확도를 높여주지만, 계산량이 급격히 증가하게 된다. 이 문제를 해결하기 위해 [4]에서는 $p(j|i)$를 오프라인으로 계산 후, 음소 n-gram 간의 비교 시 사용하는 방법을 제안하였다. 그러나 이렇게 계산된 정보의 양은 메모리에 저장할 수 없을 정도로 매우 크다.

III. Expected Matching Score 기반 유사 검색

식 2에 기반한 유사 검색은 서로 연관이 없는 두 음소 n-gram간의 유사도가 질의와 음성 문서간의 관련도에 포함된다는 단점과, 음소 오인식을 보상하는 장점의 두 가지 상반된 효과를 갖는다. 이 중, 상호 연관이 없는 두개의 음소 n-gram간의 유사도는 시스템의 검색 속도를 저하 시킬 뿐만 아니라, 불필요한 DP를 유발시킨다.

따라서 본 장에서는 *Expected Matching Score*를 사용해 $p(j|i)$의 계산이 필요한지의 여부를 결정짓는 기법을 제안한다. 제안하는 방법의 전체적인 프레임워크는 다음과 같다.

1. 질의가 주어지면, 질의 내 음소 n-gram i에 대해, i가 음소 인식기로 인식되었을 때 몇 개의 음소가 제대로 인식되는가의 기대값인 Expected Matching Score, 즉 $EM(i)$를 계산 한다.
2. 검색 단계에서, 음성 문서 내 음소 n-gram j가 주어지면, DP 수행 시 i와 j간의 매치 개수의 상한을 구하고 이를 $UM(i, j)$라 한다.
3. 만약 $EM(i) \leq UM(i, j)$가 성립하면 j를 i가 오인식된 것인 가능성이 있는 것으로 보고 $p(j|i)$를 계산한다. 그렇지 않으면, j와 i를 서로 완전히 다른 음소 n-gram으로 보고 $p(j|i)$를 0으로 놓는다.

이러한 방법은 DP 수행 전에 음소의 삽입, 삭제, 치환 오류의 개수를 제한하는 것으로 볼 수 있다. $EM(i)$는 다음과 같이 계산된다.

$$EM(i) = \lfloor \sum_{1 \leq k \leq t} C(i[k], i[k]) \rfloor \quad \text{(식 3)}$$

식에서 $i[k]$는 음소 n-gram i의 k번째 음소를 의미한다. 결국 $EM(i)$는 주어진 음소 n-gram내에서 몇 개의 음소가 정확하게 인식 될 것인가의 기대값을 계산하는 것으로 볼 수 있다.

또한 본 논문에서는 $EM(i)$의 정확성을 더욱 높이기 위해 훈련 데이터를 사용하여 질의 내 음소열 p_1, p_2가 주어졌을 때 p_1뒤의 p_2가 음소 인식 수행 시 정확히 인식될 확률을 표현한 음소 인식 정확도 행렬 $M(p_2|p_1)$을 작성 하였다. 그 뒤, Bigram을 사용한 Expected Matching Score 를 식 4와 같이 계산하였다.

$$EM_{BI}(i) = \lfloor \sum_{2 \leq k \leq t} M(i[k]|i[k-1]) \rfloor \quad \text{(식 4)}$$

다음, $UM(i, j)$를 계산하기 위해 음소 인식기에서 사용되는 음소의 개수를 t라 하고, 각 음소마다 1부터 t까지의 색인이 부여 되었다고 가정한다. 그러면 질의 내 음소 n-gram i는 각 차원에 해당 음소가 발생한 회수를 기록한 t차원 벡터 $FV(i)$ 표현할 수 있다. 마찬가지 방법으로 음성 문서 내 음소 n-gram j를 $FV(j)$로 변환 하면, $UM(i, j)$는 다음과 같이 계산된다.

$$UM(i, j) = \sum_{1 \leq k \leq t} \min(FV(i)[k], FV(j)[k]) \text{(식 5)}$$

이 식에서 $FV(i)[k]$와 $FV(j)[k]$는 각각 $FV(i)$, $FV(j)$의 k번째 차원의 값을 의미한다.

본 절에서 제안한 방법은 $UM(i, j)$를 계산 후 DP 계산량을 효율적으로 감소시켜, 기존의 방식에 비해 빠른 수행이 가능해진다.

IV. 실험 결과

4.1. 실험 환경

본 연구에서 사용한 시스템에서는 음성 인식기로 Sphinx4[10]과 그와 함께 제공되는 음향 모델을 사용하였다. 음소 n-gram 언어 모델은 Hub4의 1996/97 English Broadcast News Transcripts와 FreeTTS[11]를 사용하여 생성하였다.

평가 데이터로는 3시간 분량의 1999 Hub4 Broadcast News Evaluation English Test Material을 사용하였으며, 해당 데이터의 스크립트에 표시되어있는대로 웨이브 파일을 세그먼트 하고, 각 세그먼트를 별개의 음성 문서로 취급하였다. 이들 중 최초 30분은 시스템 개발 및 혼동 행렬 계산에 이용하였으며, 나머지 2시간 30분을 사용해 평가하였다. 30분 분량에 대한 실험결과 음소 인식기의 정확도는 53%이다.

4.2. 질의 선택

질의어는 단일 단어로서 테스트 데이터 내 20회 이상 출현하고, 그 길이가 5자 이상인 문자로 제한하였다. 5자 이상의 제한을 둔 이유는 그 단어 내 음소의 수가 충분하여 음소 n-gram의 퓨전이 잘 활용될 수 있도록 하기 위함이다. 이렇게 뽑은 질의어의 수는 총 20개였다.

4.3. 평가 기준

질의어에 대한 정답 셋이 존재하지 않으므로, 스크립트에 대해 벡터 공간 모델을 사용한 검색을 별도로 수행하여, 검색결과의 상위 5, 10, 20개 문서를 정답으로 간주하였다[12].

검색 정확도 평가는 정보 검색에서 널리 사용되는 기준인 11-point[12] 및 MAP(Mean Average Precision)[13]를 사용하였다. 이 중 11-point는 d_1, d_2, d_3 ,..., d_n 의 n개 문서가 최종 검색 결과로 주어졌을 때 이 중 실제 정답에 해당하는 문서마다 precision을 계산 한 뒤, 이를 0%, 10%, 20%, .., 100% 의 11개

recall point에 대해 보간(interpolation) 한 값이다. 예를 들어, d_1, d_2, ..., d_{10} 이 검색 결과로 주어진 문서들이고, 이 중 d_2와 d_8만이 정답이라 하면, d_2의 recall과 precision은 각각 *1/2=0.5* (총 2개의 정답 중 1개가 검색됨), *1/2=0.5* (d1, d2중 d2만이 정답임)에 해당한다. 또 d_8의 recall과 precision은 *2/2=1*(d_2, d_8이 모두 검색되었음), *2/8=0.25*(d_1, ..., d_8중 d_2와 d_8만이 정답임)에 해당한다. 11-point는 이들 값을 11개 recall에 대해 보간한 값이며, MAP는 11-point precision들의 평균으로 구했다.

4.4. 실험 결과

표 1은 스크립트 검색 결과의 상위 5, 10, 20개 문서를 정답으로 간주하고, $EM(i)$, $EM_{BI}(i)$를 적용한 경우와 기존의 방법을 비교한 결과이다.

<표 1> 제안한 방법과 기존 방법의 정확도/속도 비교
(음영은 기존 방법에 비해 정확도가 저하된 경우를 의미함)

Recall	$EM(i)$			$EM_{BI}(i)$			기존 방법		
	5	10	15	5	10	15	5	10	15
0.0	0.464	0.545	0.668	0.464	0.545	0.669	0.464	0.545	0.669
0.1	0.464	0.545	0.551	0.464	0.545	0.551	0.464	0.545	0.551
0.2	0.464	0.436	0.402	0.464	0.436	0.403	0.464	0.436	0.403
0.3	0.279	0.228	0.318	0.282	0.226	0.322	0.282	0.226	0.324
0.4	0.279	0.176	0.249	0.282	0.178	0.250	0.282	0.178	0.252
0.5	0.095	0.147	0.171	0.094	0.148	0.171	0.094	0.148	0.174
0.6	0.095	0.126	0.143	0.094	0.125	0.144	0.094	0.128	0.144
0.7	0.037	0.080	0.119	0.036	0.080	0.121	0.037	0.081	0.122
0.8	0.037	0.035	0.086	0.036	0.036	0.085	0.037	0.036	0.085
0.9	0.012	0.024	0.034	0.013	0.025	0.036	0.013	0.025	0.036
1.0	0.012	0.016	0.018	0.013	0.019	0.024	0.013	0.019	0.024
MAP	0.204	0.214	0.251	0.204	0.215	0.252	0.204	0.215	0.253
검색 시간 (초)	62			195			859		

실험 결과, 기존의 방법에 비해 $EM(i)$는 13.9배 빠르면서 0.2%의 MAP 저하만을, $EM_{BI}(i)$는 4.4배 빠르면서 0.1%의 MAP 저하만을 보였다.

V. 결 론

본 논문에서는 음소 n-gram을 사용한 음성 문서 검색 시 문서 확장 과정에서 나타나는 속도 저하를 해결하기 위해 *Expected Matching Score* 기반의 유사 검색을 제안하였다. 이를 통해 기존의 기법에 비해 검색 속도를 13.9~4.4배 향상 시켰으며, 반면 검색 정확도의 저하는 0.2%~0.1%에 그쳤다. 향후 연구에서는 *Expected Matching Score* 계산을 좀 더 정교하게 수행함으로써 불필요한 음소 n-gram간의 비교로 인해 생겨나는 검색 정확도 저하를 개선하기 위한 연구를 수행할 계획이다.

참고 문헌

[1] E. Chang, F. Seide, H. M. Meng, Z. Chen, Y. Shi, and Y.-C. Li, "A system for spoken query information retrieval on mobile devices," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 8, pp. 531-541, 2002.

[2] N. Moreau, H. Kim,, and T. Sikora, "Phonetic confusion based document expansion for spoken document retrieval," *in ICSLP*, 2004.

[3] K. Ng, "Towards robust methods for spoken document retrieval," *in ICSLP*, 1998.

[4] K. Ng, Subword-based Approaches for Spoken Document Retrieval, Ph.D. thesis, *MIT*, 2000.

[5] P. Yu, K. Chen, C. Ma, and F. Seide, "Vocabulary independent indexing of spontaneous speech," *Speech and Audio Processing, IEEE Transactions on,* vol. 13, no. 5, pp. 635-643, 2005.

[6] F. Crestani, ""Combination of similarity measures for effective spoken document retrieval,"" *Journal of Information Science*, vol. 29, no. 2, pp. 87-96, 2003.

[7] D. Abberley, D. Kirby, S. Renals, and T. Robinson, "The thisl broadcast news retrieval system," *in ESCA ETRW Workshop on Accessing Information in Spoken Audio*, 1999.

[8] J. Allan, J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu, "Inquery does battle with TREC-6," *in Text REtrieval Conference*, pp. 169-206, 1997.

[9] A. Singhal and F. Pereira, "Document expansion for speech retrieval," *in Research and Development in Information Retrieval*, pp. 34-41, 1999.

[10] K. F. Lee, H. W. Hon, and R. Reddy, "An overview of the sphinx speech recognition system," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 38, no. 1, pp. 35-45, 1990.

[11] W. Walker an P. Lamere and P. Kwok, "Freetts - a performance case study," *Tech. Rep.*, Sun Microsystems, Inc., 2002.

[12] R. B.aeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, ADDISON WESLEY, 1999.

[13] TREC, "Common Evaluation Measures," *in TREC*, pp. A-14, Nov., 2001.