2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

# 미관측문맥 모델링을 위한 다중단어카테고리 결정

한명수, 정민화
서강대학교 컴퓨터학과

# Determining Multiple Word Category Membership for Modeling Unseen Context

**Myungsoo Han, Minhwa Chung**
**Department of Computer Science, Sogang University**
**E-mail: uuhan@nlpeng.sogang.ac.kr, mchung@ccs.sogang.ac.kr**

## 요약

본 논문에서는 연속음성인식에 사용되는 언어모델이 학습 코퍼스에서 나타나지 않는 문맥에 대하여 신뢰할만한 확률을 생성할 수 있도록 하는 방안으로 다중 단어 카테고리 결정방법을 제안하였다. 제안된 다중 단어 카테고리 결정 방법은 기존의 카테고리 기반 언어모델에서의 미관측 문맥에 대한 모델링 능력을 유지하면서 동형이의어에 대한 확률의 과도한 일반화를 방지한다. 제안된 방법을 이용한 언어모델의 성능을 측정하기위해 미관측 문맥이 31% 포함된 인식문장에 대한 N-Best rescoring을 수행한 결과 word accuracy는 1-Best 문장에 대해서 3.2%의 향상을 얻었고 기존의 카테고리기반 언어모델을 적용한 결과에 비하여 0.8%의 향상을 얻을 수 있었다.

## 1. 서론

연속음성인식 시스템은 음향모델부과 언어학적 처리부로 구성된다. 언어학적 처리부에서 가장 기본이 되는 부분은 언어모델(Language Model)로서 가능한 단어열에 대한 정보를 줌으로써 정확한 인식을 위한 제약조건을 제공하는 역할을 한다. 대부분의 연속음성 인식기에서 사용되고있는 통계적 언어모델은 코퍼스로부터 규칙을 취득하기 때문에 자동으로 규칙을 획득할 수 있지만 많은 양의 데이터를 필요로 하기 때문에 학습을 위한 코퍼스가 부족하게 되는 문제가 있다.

학습데이터 부족현상은 발생 가능한 문맥에 대해 신뢰할 만한 확률을 얻을 수 있을 만큼 모든 단어가 최소한의 발생 빈도를 보이지 못하는데 원인이 있다. 이러한 문제의 해결을 위해 유사한 문맥을 가지는 단어들을 동일 카테고리로 결정하고 카테고리간의 확률을 사용하여 언어모델을 구성하는 카테고리 기반 언어모델이 제안되어 사용되고 있다. 단어를 카테고리의 멤버로 결정하는 방식에는 전문가가 수작업으로 각 단어를 결정하는 방식과 형태소 태그를 사용하는 방법[3] 그리고 코퍼스로부터 자동 생성하는 방식[1][2]이 있는데 우선 수작업으로 결정하는 방법은 규칙기반으로 언어모델을 구성하는 경우와 마찬가지로 단어규모가 커질 경우 문제가 되고 형태소 태그를 이용하는 경우는 형태소 태그가 부착된 대량의 코퍼스가 필요하며 클래스의 수의 최적화가 힘들다. 코퍼스로부터 자동으로 카테고리를 생성하는 방법은 카테고리의 수를 임의로 결정 가능하다는 장점이 있지만 단어의 수와 카테고리의 수가 커질 경우 클러스터링 시간이 과도하게 커지는 단점이 있다.

본 논문에서는 자동 카테고리 결정 알고리즘을 사용하지만 한 단어가 여러 카테고리에 속하는 것이 가능하도록 확장하였다. 다대일 대응으로 단어를 카테고리에 속하게 할 경우에는 단어가 여러 가지 기능으로 사용될 경우 그에 따른 다양한 문맥을 반영하는 하나의 카테고리는 존재하지 않으므로 각각의 문맥에 맞는 카테고리에 가중치를 가지고 속할 수 있도록 하였다. 이 방법을 통해 단어별 특성에 알맞은 카테고리의 집합을 구성함으로써 확률의 왜곡을 막고 학습 데이터에 나타나지 않은 문맥에 대한 모델링의 범위를 넓힌다.

본 논문의 구성은 2장에서는 다중 카테고리 결정 방법과 이를 이용한 언어모델 구성에 대해서 설명하고 3장에서 음성인식에 적용한 결과를 보이고 이에 대해 분석한다.

## 2. 다중 단어 카테고리 결정

### 2.1 다중 단어 카테고리

카테고리 기반의 언어모델은 코퍼스에서 관측되지 않은 문맥(unseen context)에 의해 발생하는 문제들에 대한 대처가 가능하게 한다. 그러나 카테고리 기반의 언어모델을 구성하기 위해서 자동 카테고리 결정 알고리즘을 사용할 경우 한 단어는 한 카테고리에만 속하게 되므로 단어가 여러 용도를 가지고 있으면 결정된 카테고리는 그 단어의 특성을 충분히 대변하지 못하게 된다. 본 논문에서는 이에 대한 해결책으로서 자동 단어 카테고리 결정 알고리즘에서 한 단어가 복수의 카테고리에 속하도록 결정함으로써 각 단어의 다중적 특성을 카테고리 결정에 반영하는 자동 다중 단어 카테고리 결정 알고리즘을 제안한다.

```
Collect unigram and bigram statistics from corpus

Clustering
    Start with random class distribution
    Iterate until convergence criterion is met{
        For all w in vocabulary{
            For all g' in G{
                Calculate the difference in L when
                is moved from g to g'
            }
            Move w to g' that results in bigges
            improvement in L
        }
    }
End Clustering

Decision Multiple Category membership
    For all w in vocabulary{
        For all g' in G{
                Calculate the difference in L when
                is moved from g to g'
        }
        Record M g's that result in smallist degration in L
    }
End Decision
```

[그림 1] 다중 카테고리 결정 알고리즘

제안된 다중 단어 카테고리 결정 알고리즘은 [그림 1]에서 보이는 바와 같이 단일 카테고리 결정과 다중 카테고리 결정의 두 단계로 구성하였다. 단일 카테고리 결정 단계는 코퍼스로부터 공기정보를 추출하여 likelihood를 최대화하는 k-means 유형의 단일 카테고리 결정 알고리즘을 사용하였으며 두 번째 단계는 각 단어가 최우선 카테고리 이외의 카테고리에 속했을 경우의 likelihood를 측정하여 순위 별로 다중 카테고리를 구성한다.

클러스터링 준비 단계로서 코퍼스로부터 unigram과 bigram 정보를 추출하여 테이블로 저장한다. 이 단계에서는 전체 문장에서 bigram 문맥과 발생횟수, 단어 당 발생횟수 등을 기록하여 클러스터링에서 사용할 데이터 테이블을 만든다.

첫번째 클러스터링 단계에서는 클러스터링 대상 어휘에 대해서 무작위로 카테고리를 할당한 후 모든 단어가 더 이상 카테고리의 이동이 없을 경우까지 반복하여 계산을 수행한다. 이 알고리즘에서 카테고리의 변화는 카테고리 bigram의 likelihood를 최대로 하는 방향으로 진행되고 초기 분포에 의존하는 local optimum solution을 얻는다

다중 카테고리 결정 단계에서는 단어를 카테고리로부터 제거한 후 다른 가능한 카테고리에 포함시킨 후 likelihood를 최대로 하는 그 단어에 대한 상위 M 개의 후보 카테고리를 결정한다.

### 2.2 다중 단어 카테고리 기반 언어모델

다중 단어 카테고리를 기반으로 언어모델을 구성할 경우에는 단어가 각 카테고리에 속하는 정도를 표현하는 가중치가 필요하게 된다. 카테고리 결정 연산자 $V(\cdot)$를 단일 카테고리에 대해서 정의할 경우에는 확률을 다음과 같이 나타낼 수 있다.

$$P(w(i) \mid V(w(i))) = \frac{N(w(i))}{N(V(w(i)))} \qquad (1)$$

즉 단어의 발생횟수를 카테고리의 발생횟수로 나눈 값으로서 단어가 카테고리에 기여하는 정도를 표현한다. 그러나 $V(w(i))$가 단어를 다중 카테고리에 대응시키는 경우에는 단어의 발생횟수를 이용하여 단어가 카테고리에 속할 조건부확률을 구할 수 없게 된다. 그 이유는 각 단어는 여러 가지 종류의 문맥에서 발생하는데 각각의 카테고리에 기여하는 문맥의 비율에 따라 가중치가 결정돼야 하기 때문이다.

본 논문에서는 다중 카테고리에 대한 가중치를 표현하기 위해서 단어가 각각의 카테고리에 속할 경우의 log likelihood의 차이를 이용하였다. 단어가 속하는 카테고리에 대한 log

likelihood는 최선의 카테고리에 속했을 경우 가장 큰 값을 가지게 되고 차선의 카테고리들에 속했을 경우에는 단어를 예측할 확률의 차이만큼 작은 값을 가지게 된다. 단어가 카테고리 v에 속할 경우의 log likelihood를 $LL_v$ 라고 하고, 카테고리 문맥의 일치로 인한 likelihood 증가치를 $\Delta LL_v$ 라고 하면 단어1회 발생에 의한 log likelihood 증가치는 $\Delta LL_v / N(w(i))$ 가 된다.

단어 w의 1회 발생에 의한 log likelihood 증가에 exponential을 취하면 단어 w를 카테고리 v 에 속하도록 한 경우의 모델링 확률에 비례한다. 즉 여기서 카테고리 가중치 계산을 위해

$$K_j = e^{\frac{\Delta LL_{v_j}}{N(w(i))}} \qquad (2)$$

라고 정의하고 최선의 카테고리를 $V_1$, 차선의 카테고리를 $V_2$ 라고 하면 최선의 카테고리에 의한 확률 $P_1$은 (3)과 같다.

$$\begin{aligned} P_1 &= P(w(i)|V_1)\cdot P(V_1|V(w_{i-1})) \\ &\approx \frac{K_1}{\sum_{j=1}^{N_v} K_j}\cdot\frac{N(w(i))}{N(V_1)}\cdot P(V_1|V(w_{i-1})) \end{aligned} \qquad (3)$$

또한 같은 방법으로 차선의 카테고리에 속했을 경우의 확률 $P_2, P_3, \ldots, P_{N_v}$ 를 추정할 수 있다.

이때 이전 단어에 해당하는 카테고리 V(w(i-1))는 다중 카테고리가 아닌 단일 카테고리를 사용하였는데 이유는 다중 카테고리 수에 따라 계산량이 늘어나게 되기 때문이다. 또한 식(3)의 다중 카테고리 계산은 카테고리의 개수 $N_v$ 만큼 가능하지만 모든 카테고리에 대해 다중으로 가중치를 구하면 파라미터가 과도하게 많아지므로 $K_0 >> K_M$ 의 양상을 보이는 상위 M개의 카테고리만을 다중 카테고리로 사용하여 근사화 하면

$$P(w(i)|w(i-1)) \cong \sum_{j=1}^{M} P_j \qquad (4)$$

를 얻을 수 있다. 본 논문에서는 M의 값을 4로 고정시켜서 실험하였는데 이는 각 단어의 특성에 따라 4가지 카테고리로 맵핑시킨다는 것을 의미한다.

## 3. 실험 결과 및 분석

### 3.1 음향모델과 언어모델 학습

본 논문의 실험을 위하여 사용한 음성 데이터는 삼성 종기원의 낭독체 문장 음성 데이터중 일부분인 3154문장이다. 문장들은 주로 신문, 서적등에서 수집된 광범위한 영역의 문장으로 이루어져 있다. 화자 당 100문장씩 발화하였으며 음성은 16KHz로 샘플링 되어있다. 각 문장은 형태소 단위로 태깅하였고, 각 문장은 평균 20형태소로 이루어져 있다. 형태소 태깅된 5000문장의 음성 코퍼스 중에서 30명의 화자(남성 16명, 여성 14명)가 발화한 3000문장은 음향모델과 언어모델의 학습을 위해서 사용되었고 별도의 문장에서 미등록 어휘가 없는 154문장(20명화자)을 추출하여 인식 실험용으로 사용하였다 word bigram은 absolute discounting 방법을 사용하였으며 카테고리 기반 bigram은 클러스터링을 사용한 카테고리 자동결정 알고리즘과 다중 단어 카테고리 언어모델을 위한 다중 단어 카테고리 결정 알고리즘을 구현하여 사용하였다. Word bigram의 인식 문장에 대한 perplexity는 224.0이다.

### 3.2 자동 단어 카테고리 결정 결과

단어 카테고리 결정 알고리즘을 적용하여 학습문장에 대하여 50, 100, 200, 300 개의 카테고리를 결정하였다. 카테고리 결정에는 단어의 cut-off없이 모두 클러스터링 하였으며 총 4개의 다중 카테고리에 속하도록 하였다.

| | Word bigram | 카테고리 기반 언어모델 | | | |
|---|---|---|---|---|---|
| | | 300 | 200 | 100 | 50 |
| Bigram | 35475 | 10653 | 8276 | 4021 | 1525 |
| Trigram | 51103 | 33791 | 30598 | 21278 | 10460 |

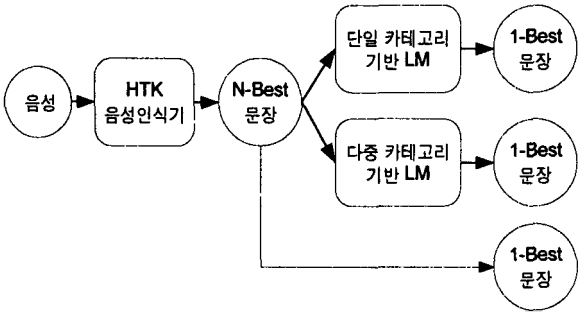
[표 1] 카테고리 개수의 변화에 따른 문맥 수 변화

[표 1]에서 bigram문맥의 수는 카테고리를 사용하여 계산하였을 때 word bigram을 사용할 경우보다 급격하게 감소하는 것을 볼 수 있다. 이는 인식대상 문장에 대해 모델링 가능한 범위가 넓어진다는 것을 나타낸다.

### 3.3 인식 실험 결과 및 분석

기본 인식실험에서는 HTK를 사용하여 word bigram을 언어모델로 N-Best 인식결과를 얻었다. 본 논문에서 제안한 알고

리즘의 성능을 평가하기위해서 N-Best결과를 기존의 단일 카테고리 기반 언어모델로 rescoring한 결과와 제안한 다중 카테고리 기반 언어모델을 사용하여 rescoring하는 방법을 사용하였다. Rescoring에서는 word bigram과 카테고리 기반




모델을 interpolation하여 사용하였다.

[그림 1] 언어모델 평가방법

기본 인식 결과는 [표 2], rescoring 결과는 [표 3]과 같다.

| 언어모델 / 인식평가(단어) | 200 Best 문장 | 1-best 문장 |
|---|---|---|
| | Word bigram | |
| Correctness | 75.11 | **63.82** |
| Accuracy | 72.24 | **58.38** |

[표 2] Word bigram 사용한 인식결과(%)

| 언어모델 / 인식평가(단어) | | 단일 카테고리 | | 다중 카테고리 | |
|---|---|---|---|---|---|
| | | bigram | trigram | bigram | trigram |
| 300 카테고리 | Correctness | **65.76** | 64.97 | **66.81** | 65.76 |
| | Accuracy | **60.78** | 60.59 | **61.61** | 60.92 |
| 200 카테고리 | Correctness | 65.43 | 65.13 | 65.92 | 65.72 |
| | Accuracy | 60.65 | 60.75 | 61.18 | 60.82 |
| 100 카테고리 | Correctness | 65.29 | 65.10 | 65.62 | 65.82 |
| | Accuracy | 61.08 | 60.65 | 61.18 | 61.28 |
| 50 카테고리 | Correctness | 64.90 | **65.35** | 65.43 | 65.66 |
| | Accuracy | 60.68 | **60.94** | 61.21 | 61.54 |

[표 3] 카테고리 기반 언어모델 사용 rescoring 결과(%)

이상의 결과에서 300 다중 카테고리 bigram 을 사용하여 rescoring한 것이 가장 높은 인식률을 보인다. 실험에 사용한 인식 문장을 학습 코퍼스가 충분히 설명하지 못하는 상황에서 word trigram의 경우 word bigram보다 오히려 혼잡도가 커지는 현상을 보인다. 카테고리의 개수가 많아질수록 카테고리 기반 언어모델은 word n-gram의 성격에 가까워지는 것이 300카테고리 trigram에서의 성능 향상이 크지 않은 이유로 보인다. 또 다른 이유는 단어 카테고리 결정이 bigram에 최적화하는 클러스터링 알고리즘으로 구성되어있다는 점이다. 카테고리 bigram의 수는 word-bigram문맥의 수에 비해서 월등히 작은 크기로 최적화 되지만 카테고리 trigram 문맥의 수는 word trigram 문맥의 수에 비해서 큰 차이가 나지 않는다. 따라서 trigram의 수가 가장 작은 50카테고리의 경우에만 카테고리 trigram의 성능이 bigram의 성능을 능가하는 현상을 보였다.

## 4. 결론

본 논문에서는 음성인식에서 사용되는 카테고리 기반 언어모델의 성능향상을 위해서 다중 단어 카테고리의 자동 결정을 가능하게 하는 알고리즘을 제안하였다. 자동 다중 단어 카테고리 결정 알고리즘은 각 단어가 가진 다중적 특성을 카테고리 결정에 반영하여 확률의 신뢰도를 높이면서 카테고리 기반 언어모델의 장점인 학습 코퍼스 부족현상에 대처하는 능력을 강화한다.

본 논문의 실험에 사용된 음성 코퍼스는 삼성 종합기술원의 HCI 팀이 본 연구실과 공동으로 제작한 낭독체 음성DB의 일부입니다. 사용을 허락해 주신데 대해 감사드립니다.

## 5. 참고문헌

[1] Peter F. Brown, "Class-based n-gram models of natural language", Computational Linguistics vol.8, 1992

[2] Reinhard Kneser , Hermann Ney ,"Improved clustering techniques for class-based statistical language modelling", Proceedings of Eurospeech, 1993

[3] T.R. Niesler, "Category-based statistical language models", PhD. thesis, Dept. Engineering, University of Cambridge, 1997

[4] 김우성, 구명완 , "통계적 언어 모델의 clustering 알고리즘과 음성인식에의 적용", 한글 및 한국어정보처리 학술발표 논문집, 1996