분류 우선순위 적용과 후보정 규칙을 이용한 효과적인 한국어 화행 분류

Efficient Korean Speech-Act Classification using a Classification Priority Application and a Post Correction Rule

저자
(Authors)

송남훈, 배경만, 고영중
Namhoon Song, Kyoungman Bae, Youngjoong Ko

출처
(Source)

한국정보과학회 학술발표논문집 , 2015.6, 675-677 (3 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE06394182

APA Style

송남훈, 배경만, 고영중 (2015). 분류 우선순위 적용과 후보정 규칙을 이용한 효과적인 한국어 화행 분류. 한국정보과학회 학술발표논문집, 675-677.

분류 우선순위 적용과 후보정 규칙을 이용한 효과적인 한국어 화행 분류

송남훈○, 배경만, 고영중
동아대학교 컴퓨터공학과
{nh.song.89, kmbae0722, youngjoong.ko}@gmail.com

Efficient Korean Speech-Act Classification using a Classification Priority Application and a Post Correction Rule

Namhoon Song○, Kyoungman Bae, Youngjoong Ko
Department of Computer Engineering, Dong-A University

요 약

화행이란 발화 속에 포함되어 있는 화자에 의해 의도된 언어적 행위이다. 음성언어 시스템에서 입력된 발화에 적합한 화행을 분류하는 것은 중요하다. 기존의 화행 분류에 관한 연구는 규칙 기반과 통계 기반의 방법을 많이 사용한다. 통계 기반 방법은 어휘 자질과 담화 자질을 기반으로 지지벡터기계(SVM) 기반의 다중 분류기를 이용함으로써, 효과적으로 화행을 분류한다. 하지만, 기존의 통계기반 화행 분류 방법은 화행별 발화의 비율을 고려하지 않기 때문에 상대적으로 발화의 수가 적은 화행에 대한 분류 성능은 낮다. 본 논문에서는 발화의 수가 적은 화행을 효과적으로 분류하기 위해 화행별 발화 수의 비율을 고려한 SVM 기반의 개선된 화행 분류 방법을 제안한다. 또한, SVM 기반의 분류된 화행을 보정할 수 있는 변환 기반 학습(TBL)을 이용한 보정 규칙을 적용함으로써 향상된 화행분류 성능을 얻었다. 본 논문에서는 화행별 발화 수의 비율을 고려한 분류 우선순위 변화와 후보정 규칙을 이용한 화행 분류 방법을 실험을 통해 평가하였으며, 이는 발화수의 비율이 낮은 화행의 우선순위를 고려하지 않은 기존의 SVM 보다 1.03%가 향상된 86.54%의 화행 분류 성능을 얻었다.

1. 서 론

화행이란 발화 속에 포함되어 있는 화자에 의해 의도된 언어적 행위로, 자연어 대화를 처리하는 많은 시스템에서 화자의 의도를 파악하고 응답을 생성하는 중요한 역할을 한다[1]. 화행은 발화가 수행해야할 대화 내에서 역할을 결정하기 때문에 대화에 관한 연구에서 화행을 효과적으로 분류하는 것은 매우 중요하다. 따라서, 이러한 화행분류는 규칙 기반 방법과 통계 기반 방법을 기반한 많은 연구가 진행 되어왔다[2].

통계 기반 방법은 화행을 분류하기 위해 지지벡터기계(SVM) 기반의 다중 분류기를 사용한다. SVM을 기반으로 한 통계방법은 화행별로 분류기가 존재한다. 발화가 입력이 되면, 각 분류기별로 분류 가중치를 구하고, 가중치가 가장 높은 분류기의 화행을 선택한다. 하지만, 대화 코퍼스에는 화행별로 발화가 존재하며, 각 화행별로 존재하는 발화 비율은 다르다. 기존의 SVM을 이용한 화행 분류 모듈은 이러한 화행별 비율의 차이를 고려하지 않기 때문에 대화 코퍼스에 존재하는 비율이 낮은 화행의 발화는 비율이 높은 화행을 학습한 분류기로 분류되는 문제가 발행할 수 있다[3]. 본 논문에서는 발화 비율을 기반으로 분류기의 분류 우선순위를 변경함으로써 발화 비율이 낮은 화행분류 성능을 개선하는 방법을 제안한다. 하지만, 분류 우선순위를 변경하는 것만으로는 낮은 발화 비율을 가지는 화행을 효과적으로 분류하는 데는 부족하다.

이를 위해, 본 논문에서는 SVM 분류기의 결과를 보정하는 규칙을 추가함으로써 낮은 비율의 화행을 효과적으로 분류하는 방법을 제안한다. 우선, 발화의 비율이 낮은 순서로 분류기들을 오름차순으로 정렬하고, 발화가 들어오면, 비율이 가장 낮은 분류기부터 차례로 분류를 진행한다. 상위60%(비율이 낮은 순서로 9개의 분류기)의 발화에 대해서 분류를 진행하고, 가장 처음 분류기의 가중치가 양수가 되는 화행을 선택한다. 만약, 상위 60%의 분류기에서 양수의 가중치를 가지는 분류기가 존재하지 않으면, 일반적인 SVM 방식으로 가장 가중치가 큰 분류기의 화행으로 분류한다. 하지만, SVM을 이용한 분류기에서 비율이 낮은 화행과 연관된 발화의 경우 SVM 분류기의 모든 가중치가 음수로 나오며 상대적으로 다른 화행에 비해 많이 존재한다. SVM 분류기의 가중치가 모두 음수가 나오는 경우 분류결과가 잘못될 확률이 높다.

본 논문에서는 이를 해결하기기 위해, SVM 분류 결과가 모두 음수인 발화에 대해 검증 데이터(validation data)를 기반으로 실험을 통해 얻은 임계값 이하의 발화에 대해서 변환 기반 학습(TBL) 기법을 기반으로 만들어진 보정 규칙을 적용한다. 이러한 후 보정 규칙을 생성하기 위해 이전화행, 동사, 명사를 이용한 5-gram 기반의 규칙 틀을 이용한다.

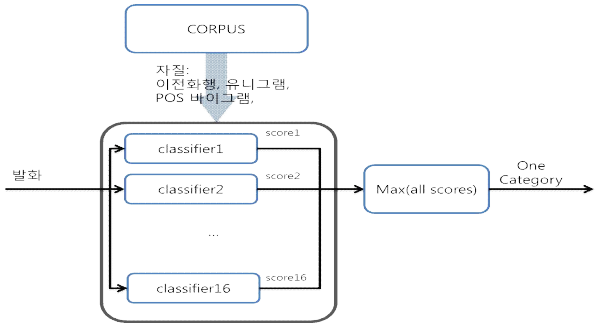
실험을 통해 제안한 방법을 평가하였으며, 유니그램 자질(어휘, 품사), 바이그램 자질(품사 바이그램), 이전화행을 기반으로 학습된 분류기를 사용하였다. 제안한 화행별 발화의 비율을 고려한 우선순위 적용과 후 보정 규칙을 적용함으로써 기존의 SVM보다 1.03% 향상된 86.5%의 F1 값을 얻었다. 본 논문의 구성은 다음과 같다. 2절에서는 SVM 기반의 화행 분류기와 TBL 기법에 대해 살펴보고, 3절에서 본 논문에서 제안하는 방법을 살펴본다. 4절에서는 제안한 방법을 실험을 통해 평가하고, 마지막으로 5절에서는 결론을 낸다.

2. 기존 연구

2.1 통계 기반 모델

통계 기반 모델은 화행 분류를 위해 대량의 대화 코퍼스를 사용하여, 기계 학습을 진행 하고, 학습된 분류기를 기반으로 각 발화의 화행을 분류한다[4]. 기존의 한국어 화행 분류 방법은 유니그램 자질(어휘, 품사), 바이그램 자질(품사 바이그램), 이전화행을 사용하여 분류기를 학습했으며, 발화가 입력되면 각 분류기를 통하여 나오는 분류 가중치 중 가장 높은 가중치의 화행으로 분류된다. 아래의 그림 1은 일반적인 SVM 기반의

화행 분류기를 나타낸다. 본 논문에서는 이를 베이스라인 시스템으로 사용한다.



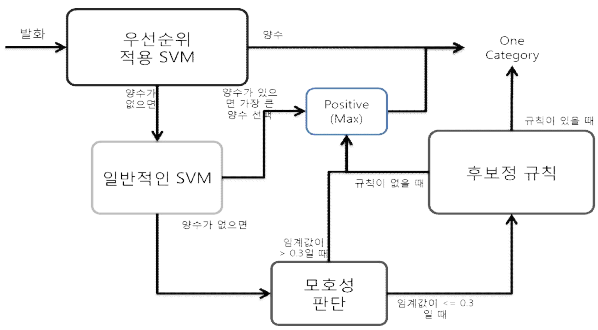
[그림 1] 일반적인 SVM 기반의 다중 분류기

2.2 규칙 기반 모델

규칙 기반 모델이란 시스템 설계자가 화행을 결정하기 위한 규칙들을 작성하는 것으로, 해당 화행의 지식을 포함하는 언어 정보 규칙과 문맥 규칙을 사용하여 화행을 결정하는 방법이다 [5].

3. 제안하는 방법

본 논문에서는 발화의 비율이 낮은 화행 분류기의 우선순위를 조정하고, SVM 결과를 후보정하는 규칙을 적용함으로써 발화의 비율이 낮은 화행의 분류 성능을 개선하는 방법을 제안한다. 그림 2는 제안하는 방법을 도식화 한 것이다.

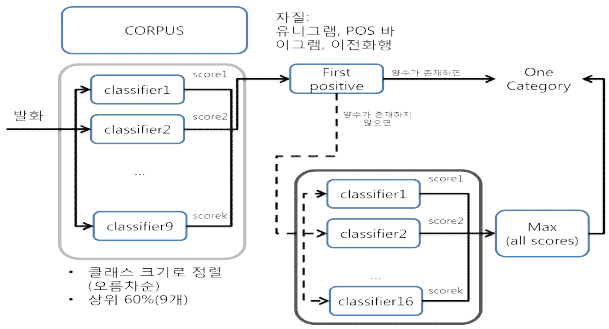


[그림 2] 본 논문에서 제안하는 최종 시스템

3.1 분류 우선순위 적용 SVM

본 논문에서는 화행별 발화의 비율을 기반으로 화행 분류의 우선순위를 변경함으로써 비율이 낮은 발화(Offer, Suggest, Accept 등)의 화행분류 성능을 개선하는 방법을 제안한다. 그림 3은 제안하는 분류 우선순위 적용 SVM를 도식화 한 것이다. 일반적인 SVM 기반의 분류에서는 각 화행별 분류기에서 추정된 분류 가중치 중 가장 큰 가중치를 가지는 분류기의 화행을 선택한다. 이러한 방식은 발화의 비율이 높은 분류기의 경우 분류기를 학습에 상대적으로 많은 발화를 기반으로 학습이 진행된다. 적은 수의 발화로 학습을 진행한 화행의 분류기는 많은 수의 발화로 학습한 분류기에 비해 상대적으로 낮은 분류 가중치를 추정할 확률이 높다. 일반적으로 SVM 분류기에서 양수의 분류 가중치를 가지면, 그 분류기의 화행일 확률이 높다. 하지만, 발화 비율이 낮은 화행을 가지는 발화가 입력되었을 때 비율이 낮은 화행의 분류기에서 양수의 분류 가중치를 가짐에도 불구하고, 양수 분류 가중치를 가지는 비율이 높은 화행의 분류기로 잘못 분류되는 경우가 발생한다.

본 논문에서는 이를 해결하기 위해 비율이 낮은 화행의 분류기에 대해 분류 우선순위를 줌으로써 비율이 낮은 화행을 가지는 발화에 대한 분류 성능을 개선시켰다. 그림 3에서와 같이 화행들을 발화의 비율이 낮은 순서부터 오름차순 정렬을 하여 상위 60%(9개의 화행)에 해당하는 분류기에 대해서 순차적으로 분류를 진행한다. 분류 가중치가 양수를 가지는 분류기가 있으면, 그 분류기의 화행으로 분류하고, 만약, 상위 60%의 화행 분류기에서 양수 가중치를 가지는 분류를 찾기 못한다면, 기존의 방법과 동일하게 모든 화행 분류기 중에서 가장 큰 가중치를 가지는 분류기의 화행을 선택한다.



[그림 3] 화행별 발화 비율을 고려한 분류 우선순위 적용 SVM

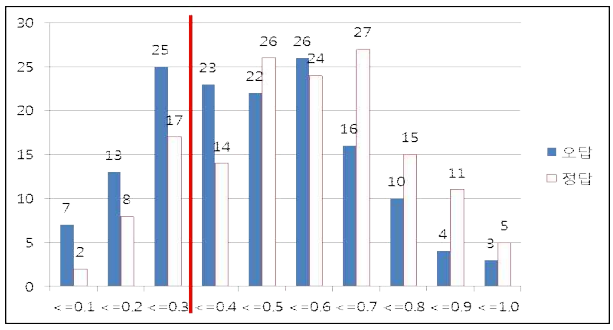
3.2 SVM 후보정 규칙

앞서 언급한 것과 같이 일반적인 SVM을 이용한 분류 방식은 분류기의 분류 가중치가 가장 큰 분류기의 화행을 선택한다. 입력된 발화에 대한 모든 분류기의 분류 가중치가 음수인 경우가 존재한다. 이 경우 역시 가중 큰 분류 가중치를 가지는 분류기로 분류가 되는데 모든 분류 가중치가 음수인 경우 양수 분류 가중치를 가지는 경우보다 분류가 잘못될 확률이 높다. 특히, 발화 비율이 낮은 화행을 가지는 발화는 상대적으로 분류가 잘못될 확률이 높다. 본 논문에서는 이러한 문제를 해결하기 위해 모든 분류 가중치가 음수 값을 가지는 결과들의 모호성을 계산하여, 실험을 통해 얻은 임계값 보다 낮은 발화에 대해 SVM 분류 결과를 보정할 수 있는 후보정 규칙을 제안한다.

3.2.1 모호성 판단 방법

모든 분류기의 분류 가중치가 음수인 경우도 가중 큰 가중치를 선택하면, 분류가 잘 이루어지는 발화가 존재한다. 모든 분류기의 분류 가중치가 음수인 모든 발화에 대해 보정을 진행했을 때 오히려 성능이 떨어질 수 있다. 본 논문에서는 분류 결과가 틀리는 발화의 경우 분류 가중치가 큰 상위 3개의 가중치의 값의 차이가 적을 것이라는 가정 하에 아래 수식을 이용해 모호성을 계산한다. ①, ②, ③은 분류기의 분류 가중치가 큰 기준으로 상위 1, 2, 3 순위의 분류 가중치를 나타낸다.

$$(① - ②) + (① - ③) \quad (1)$$



[그림 3] 수식(1)을 통하여 일반화 시킨 분포

검증 데이터(validation set)을 기반으로 수식 (1)을 이용해 모든 분류 가중치가 음수를 가지는 발화들에 대해서 음수임에

도 분류 결과가 맞은 발화의 수와 틀린 발화의 수는 그림 3과 같다. X축의 값은 수식(1)에서 계산된 모호성 값을 정규화한 값이다. 그림 3과 같이 정답보다 오답의 비율이 높은 0.3을 임계값을 결정하였다.

3.2.2 후보정 규칙

후보정 규칙은 학습 데이터와 검증 데이터로 나눈 후 학습데이터만을 이용해서 학습된 분류기를 기반으로 검증데이터에 대해 화행 분류를 진행하였다. 분류 후 3.2.1에서 제안한 방법을 이용해서 모호성을 판단하고, 모호성이 존재하는 발화를 변환 기반 학습(TBL)의 학습 데이터로 사용하였다. 규칙 생성을 위한 템플릿의 자질로는 이전화행(PS), 명사태그(NN, NC), 동사태그(PV, PA)를 기반으로 5-gram을 사용하였다.

아래의 표 1은 규칙 템플릿 중 바이그램에 대한 예이다. 모든 템플릿에 이전화행이 들어가고, 각 4개의 태그별로 템플릿을 생성하였다. 규칙 템플릿 56개를 통해 4267개의 후보정 규칙을 생성하였다.

[표 1] 규칙 템플릿의 예

바이그램	이전화행 + 명사태그 + 단어
	이전화행 + 단어 + 명사태그
	이전화행 + 동사태그 + 단어
	이전화행 + 단어 + 동사태그

4. 실험 및 결과

본 논문에서 선행된 한국어 화행 분석 시스템들에서 사용된 대화 코퍼스를 이용하였다[7-10]. 이 코퍼스는 호텔, 비행기, 여행 예약에서 실제로 사용되는 대화를 녹음 후 정제한 코퍼스이다. 이렇게 정제된 코퍼스는 10,285개의 발화와 17개의 화행으로 구성되어있다. 이중 중 발화의 빈도가 너무 낮은 하나의 화행을 제거한 10,280개의 발화와 16개의 화행을 사용한다. 학습 데이터로는 8,348개의 발화를 그리고 테스트 데이터로는 1,932개의 발화를 사용하였다. 아래의 표 2는 테스트 데이터의 세부 정보이다. 평가는 F1 값을 이용하여 진행한다.

[표 2] 테스트 데이터

화행	개수	화행	개수
Offer	8	Request	84
Reject	22	Ask-if	101
Suggest	37	Expressive	113
Promise	40	Opening	125
Accept	50	Introducing-oneself	141
Acknowledge	69	Inform	250
Closing	69	Ask-Ref	257
Ask-confirm	82	Response	484

표 3에서 각 모델들의 성능 평가의 결과를 나타내고 있다. 우선순위 모델은 우선순위SVM을 적용하였다. 마지막으로 최종 모델은 우선순위SVM의 결과에 TBL을 적용한 결과를 나타낸다. 기존 모델 보다 1.03%로를 개선하였다.

[표 3]각 모델들의 성능 평가 결과 (%)

화행	F-Measure		
	베이스라인	우선순위조정	최종 모델
micro	85.51	85.71(+0.20)	**86.54(+1.03)**
macro	75.51	75.76(+0.25)	**78.09(+2.58)**

표 4는 빈도수가 낮은 화행 별로 성능을 평가한 내용이다. Offer 화행은 기존 모델 보다 20%를 개선하였고, Accept 화행은 약 5%의 성능을 개선시켰다. 나머지 화행들은 약 1~3%정도의 개선된 것을 볼 수 있다.

[표 4]빈도수가 낮은 상위 5개의 화행의 성능 평가 (%)

화행	F-Measure		
	베이스라인	우선순위 조정	최종 모델
Offer	0.00	0.00	**20.00**
Reject	73.17	73.17	**75.00**
Suggest	63.64	63.64	**64.62**
Promise	72.00	72.00	**72.73**
Accept	60.76	60.76	**65.82**

5. 결론

SVM를 이용하여 화행 결정에 있어서 효과적인 정보들을 제공해줄 수 있다. 하지만, 기존의 SVM을 이용한 화행 분류 모듈은 이러한 화행별 비율의 차이를 고려하지 않기 때문에 대화 코퍼스에 존재하는 비율이 낮은 화행의 발화는 비율이 높은 화행을 학습한 분류기로 분류되는 문제가 발행할 수 있다. 그렇기 때문에 비율이 낮은 화행의 발화에 먼저 우선순위를 부여하여 성능을 개선하였고 SVM이 해결하지 못하는 부분을 규칙기반인 변환 기반 학습 기법을 통해 생성된 보정규칙을 이용하여 보다 나은 성능을 내었다.

6. 참고 문헌

[1] 김경선, 서정연, "자질 선택 기법을 이용한 한국어 화행 결정", 정보과학회논문지, Vol.30, No.3, pp. 278-284, 2003.

[2] 김민정, 박재현, 김상범, 임해창, 이도길, "한국어 화행 분류를 위한 최적의 자질 인식 및 조합의 비교 연구", 정보과학회논문지, Vol.35, No.11, pp. 681-691, 2008.

[3] A. Omuya, V. Prabhakaran, O. Rambow, "Improving the Quality of Minority Class Identification in Dialog Act Tagging", Proc NAACL-HLT, pp.802-807, 2013.

[4] Kyoungman Bae and Youngjoong Ko, An Effective Category Classification Method Based on a Language Model for Question Category Recommendation on a cQA service, CIKM 2012, pp.2255-2258, 2012.

[5] 김세종, 이용훈, 이종혁, "이전 문장 자질과 다음 발화의 후보 화행을 이용한 한국어 화행 분석", 정보과학회논문지, Vol.35, No.6, pp. 374-385, 2008.

[6] 양재형, "규칙 기반 학습에 의한 한국어의 기반 명사구 인식", 정보과학회논문지, Vol.27, No.10, pp.1062-1071, 2000.

[7] Won Seug Choi, Jeong-Mi Cho, and Jungyun Seo. "Analysis System of Speech Acts and Discourse Structures Using Maximum Entropy Model", In Proceedings of the 37th Annual Meeting of the Association for computational Linguistics, pp. 230-237, 1999.

[8] Lee, Songwook, and Jungyun Seo, "Korean Speech Act Analysis Using Decision Tree", In Proceedings of the Conference on Hangul and Korean Language Information Processing, pp.377-381, 1999.

[9] Lee, Jae-won, Jungyun Seo, Gilchang Kim, "A dialogue analysis Model with statistical speech act processing for Dialogue Machine Translation", Workshop in conjunction with (E)ACL'97, pp.10-15, 1997.

[10] 이현정, 이재원, 서정연, "자동통역을 위한 한국어 대화 문장의 화행 분석 모델", 정보과학회논문지(B), Vol.25, No.10, pp.1443-1452, 1998.