CRF 를 이용한 영어작문 구성요소 자동분류기법

이한남*, 곽동민*, 박세원*, 엄진희*
*딥큐먼 AI 리서치 그룹
e-mail : jrhee17@deepcumen.com

Classification of Essay Discourse Elements Using Conditional Random Fields

John Rhee*, Dong-Min Kwak*, Sewon Park*, Jin-Hee, Um*
*DeepCumen AI Research Group

요 약

본 연구에서는 글의 구성요소를 추측하는 가장 높은 성능을 나타내는 알고리즘을 제시한다. 실험 방법은 글의 각 문장에 대한 자질을 추출, 자질 선택, 그리고 데이터에 대해 여러 기계학습 알고리즘을 학습시킨 후 성능을 비교하여 진행하였다. 또한 이 중 가장 높은 성능을 보이는 CRF 를 기존에 연구되어 있는 성능과도 비교하였다. 마지막으로 CRF 가 구성요소를 추측하는 데 있어서 가장 높은 성능을 보이는 이유에 대해 분석하였다.

국내의 유명 어학원 및 토플 웹사이트를 통해 1969 개의 토플 에세이를 수집했으며 2 명의 전문 평가자를 통해 각 문장을 8 개의 분류로 나누었다. 이를 CRF 를 적용한 결과 87.2%의 F score 가 나왔으며 기존 연구결과, 그리고 다른 알고리즘보다 높은 성능을 보였다.

1. 서론

토플은 ETS 가 주관하는 시험으로 매년 국내에서는 10 만명 정도의 학생이 본다. 시험 구성은 Reading, Listening, Speaking, Writing 으로 이루어져 있으며 현재 ETS 에서는 e-rater 라는 자동채점시스템을 채택하여 영 모의시험 채점에 사용하고 있다.

ETS e-rater 의 에세이 자동채점은 총 8 종류의 자질을 이용하여 계산하며 점수를 예측하는데 있어서 글의 구성(organization)이 가장 높은 영향(22%)을 주는 것으로 밝혀졌다.[1] 에세이의 구성을 분석하려면 에세이의 구성요소(discourse element)들을 찾아야 한다. ETS 에서는 이를 introductory material, thesis, main idea, supporting idea, 그리고 conclusion 으로 정의하며 위의 구성요소를 모두 갖춘 에세이가 점수를 잘 받는다고 한다.[1,2] 이러한 구성요소들을 이용하여 에세이의 일관성, 명료성 등의 세부적인 특징을 파악할 수 있다.[3]

기존 연구에서는 구성요소를 추측하기 위해 글이 문장 단위로 순차적인 관계를 가진다고 가정하고 조건부 확률의 독립성을 가정한 generative 모델을 통해서 에세이를 모델링했다.[4]

본 연구에서도 글이 하나의 주장을 하기 위해서는 그 안의 구성요소들이 순차적인 관계를 가질 것이라는 것을 가정하고 time-series 모델을 이용하여 이를 검증하였다. 여기에 더불어서 데이터를 모델링할 필요가 없는 discriminative 모델인 CRF 의 성능이 가장 잘 나올 것이라는 가정을 하고 여러 머신러닝 알고리즘을 이용하여 구성요소를 분류하였다.

이를 증명하기 위해 2 장에서는 관련연구 소개, 3 장에서는 데이터 소개, 4 장에서는 자질과 알고리즘 소개, 5 장에서는 실험 결과를 제시하며 6 장에서는 결론과 향후 연구 과제를 제시한다.

2. 관련연구

구성요소를 추측하려는 시도는 2000 년대 초반부터 2006 년까지 주로 ETS 에서 진행하였다. 초기에는 주로 일반적인 분류기를 이용해 추측하려는 시도가 있었다. Burstein 은 Naive Bayes 를 이용해서 Thesis 문장을 찾는 데 있어서 0.43 의 F-score 가 나왔다.[5] 또한 에세이의 Thesis 와 Conclusion 을 C5.0 트리를 이용해서 분류하려는 시도는 각각 0.54, 0.8 의 F-score 를 보였다.[6]

구성요소를 time-series 모델을 이용해 평가한 사례로는 TOEFL 의 e-rater 가 있으며 투표 시스템을 이용하여 구성요소를 추측하였다. 이 시스템에서는 time-series 모델과 일반 분류기를 혼합하여 사용하였다. Time-series 모델은 구성요소의 n-gram 을 통해 p(x|y)을 모델링 한 generative 모델을 사용하였으며 일반 분류 모델로는 C5.0 트리를 사용하였다. 이 두 모델의의 투표로 분류를 진행하였으며 이를 이용해 intro, thesis, main idea, supporting idea, conclusion 을 분류하는 데 있어서 0.85 의 F-score 를 이루어냈다. [4] 현재 Burstein 은 구성요소를 추측하는 것보다는 구성요소를 이용해서 글의 일관성, 등의 특징을 파악하는 방향으로 연구를 진행하고 있다.[3]

CRF 역시 2000 년대 초반에 관련연구가 활발했으며 알고리즘을 소개하는 논문이 잘 나와 있다.[7] CRF 는 time-series 모델 중에서도 discriminative 한 성격을 띄고 있어 데이터에 대해 독립성의 가정을 할 필요가 없다. 이러한 이유로 각 자질의 조건부 확률의 독립성을 가정할 수 없는 데이터에 대해서는 HMM 보다 성능이 잘 나온다는 것이 밝혀졌다.[8]

특히 NLP 분야의 특성상 글의 각 자질들에 대해 독립성을 가정하기 어렵기 때문에 CRF 를 이용해 순서적인 데이터를 분석한 사례가 많다. 대표적으로 sentence parsing 문제에 CRF 를 적용한 경우 기존의 generative 모델보다 성능이 잘 나왔다는 것이 밝혀졌다.[9] 문장 단위의 sentiment 를 평가하는 데 있어서 CRF 가 Naïve Bayes 보다 성능이 좋다는 것도 보여졌다. [10] 마지막으로 Noun Phrase Chunking 을 하는데 있어서 CRF 가 다른 알고리즘에 비해 성능이 우월하다는 것을 보였다.[11]

본 연구에서는 CRF 가 time-series 모델이면서 discriminative 모델이라는 이점이 구성요소 추측에 적합할 것이라고 보고 이를 기존의 연구결과, 그리고 다른 알고리즘들의 성능과 비교하였다.

3. 데이터

본 논문은 국내 유명 어학원 및 토플 웹사이트를 통해 수집된 1969 개의 토플 에세이를 이용했다. 각 에세이에 대해서는 해외 영문학 석/박사 출신의 평가자들이 각 문장을 thesis(ts), main idea(mi), 그리고 supporting idea(si)로 분류를 했다. Thesis 는 에세이가 전체적으로 주장하는 요지, main idea 는 thesis 를 뒷받침하는 근거, 그리고 supporting idea 는 main idea 를 뒷받침하는 근거로 정의하였으며 어떠한 분류에도 해당하지 않는 경우 unknown(unk)으로 처리하였다.

```
unk ts ts

mi1 mi1 si1 si1 si1 si1
si1 si1

mi2 si2 si2 si2 si2 si2
si2

unk unk
```

<그림 1> 에세이 문장의 분류 예시

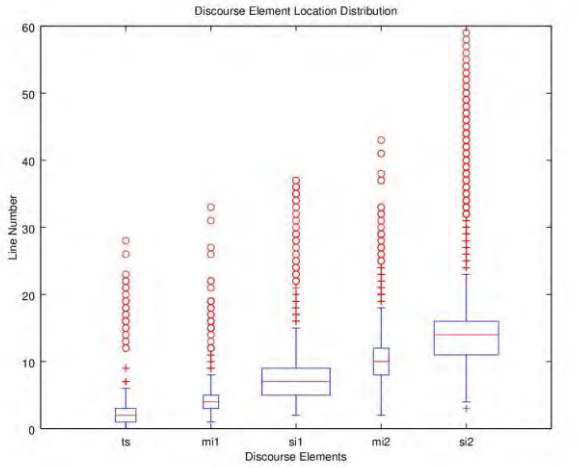
그림 1 에서 하나의 에세이의 구성요소에 대한 예시를 볼 수 있다.

각 에세이는 두 명의 평가자들에 의해 평가되었으며 각 카테고리의 개수는 Table 1 과 같았다.

<표 1> 구성요소 분포

분류	개수	분류	갯수
ts	2825	si2	8856
mi1	2122	mi3	341
si1	9355	si3	931
mi2	2118	unk	11957

또한 각 구성요소는 그림 2 와 같이 특정 위치에 따른 분포가 생긴다는 것을 확인할 수 있었다.



<그림 2> 구성요소 위치 분포

그림 2 를 보면 에세이 내에서 각 구성요소가 비교적 작은 표준편차를 가지고 서로 다른 위치에 분포한다는 것을 알 수 있다. 이를 바탕으로 하나의 구성요소의 위치가 다음에 올 구성요소에 영향을 줄 수 있다는 것을 추측했으며 이를 증명하기 위해 본 논문에서는 CRF 를 비롯한 여러 time-series 모델을 사용하였다.

4. 실험방법

4.1 자질추출

각 문장에서 추출한 자질은 그림 3 에 나와 있다.

1. 주제의 성격을 나타내는 자질(7 개의 자질)
   ex: 선호, 비교, 표현...
2. 문장의 성격을 나타내는 자질(7 개의 자질)
   ex: 인과, 가정, 설명...
3. 문장의 긍정.부정 정도(1 개의 자질)
   ex: $-1 < x < 1$, x 가 1 에 근접할수록 긍정
4. 구두점 갯수(4 개의 자질)
   ex: 물음표의 개수, 느낌표의 개수 ...
5. 문단위치, 문장위치에 대한 자질(6 개의 자질)
   ex: 문단번호, 문장번호, 문단 내의 문장번호 ...
6. 문장의 문법적 요소 존재 여부(45 개의 자질)
   ex: 과거형 동사, 대명사, 부사...
7. 문장의 unigram 단어 존재 여부 (48 개의 자질)
   ex: third, finally...

<그림 3>사용된 자질

총 118 개의 자질이 사용되었으나 HMM, GMM, NB 의 경우 자질의 개수가 너무 많아 성능이 굉장히 낮

았다. 이를 보완하기 위해 정보이득(Information gain)을 통해 위 알고리즘들의 성능을 가장 높이는 자질의 개수(10 개)를 선정하고 사용하였다.

4.2 알고리즘

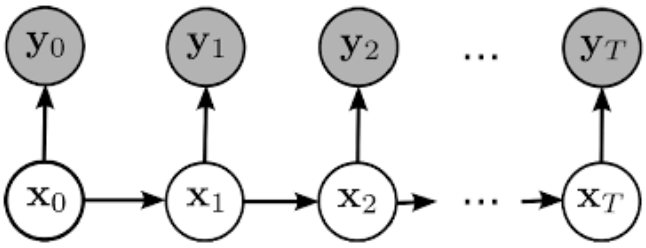
본 연구에서는 TOEFL 에세이의 구성요소 사이의 순서적인 흐름을 모델링하기 위해 CRF 를 이용하였다. CRF 는 그림 4 의 공식으로 모델링되며 feature function (f_{Ak}) 정의방법에 따라 모델의 형태가 달라진다.

$$p(x,y) = \frac{1}{Z}\prod_{A}\psi_A(x_A, y_A)$$

$$\psi_A(x_A, y_A) = exp\left(\sum_k \theta_{Ak} f_{Ak}(x_A, y_A)\right)$$

<그림 4> CRF 모델 공식

HMM 과 같은 형태의 모델을 적용하기 위해 그림 5 의 형태로 되어 있는 CRF 를 이용하여 분석했다. 최적화하기 위해서는 L-BFGS 를 이용하였으며 inference 를 위해 forward-backward 알고리즘을 사용하였다.



<그림 5> Time-series 모델에 사용된 그래프

실험을 진행하기 위해 HMM, CRF, 그리고 Decision Tree(DT)는 Mallet 라이브러리를 이용해 적용하였으며 Naive Bayes(NB), Logistic Regression(LR), Gaussian Mixture Model(GMM)는 weka 라이브러리를 사용하였다. 모든 실험은 10-fold cross validation 을 이용하여 성능을 평가하였다.

5. 결과

우선 CRF 를 적용한 결과를 기존의 사례들 중 가장 성능이 높은 시스템과 비교하였다. 이 투표시스템은 여러 알고리즘들을 통합한 시스템이다.[4]

<표 2> 기존연구와 CRF 의 F-score 비교

	CRF	Voting System
Ts	0.712	0.73
Mi	0.893	0.76
Si	0.910	0.91
Overall	0.892	0.85

전체적인 성능은 CRF 가 4%정도 앞선다는 것을 볼 수 있다. 또한 글의 ts 를 찾는 성능은 다소 떨어지지만 mi 를 찾는 성능은 월등히 뛰어나다는 것도 볼 수 있다.

기존의 나와 있는 시스템에 비해 CRF 가 성능이 좋다는 것을 확인한 후 다른 알고리즘들과 CRF 를 비교해보았다.

<표 3> CRF 와 다른 알고리즘의 F-score 비교

	GMM	NB	HMM	LR	DT	CRF
Ts	0.556	0.608	0.673	0.740	0.727	0.745
mi1	0.547	0.324	0.779	0.925	0.925	0.937
si1	0.463	0.585	0.415	0.898	0.898	0.922
mi2	0.04	0.631	0.183	0.914	0.914	0.920
si2	0.32	0.0	0.607	0.891	0.891	0.906
mi3	0.009	0.0	0.796	0.847	0.847	0.851
si3	0.046	0.0	0.032	0.816	0.816	0.836
Unk	0.448	0.731	0.806	0.806	0.806	0.822
total	0.396	0.554	0.609	0.852	0.856	0.872

표 3 을 보면 CRF 를 통한 구성요소 예측이 가장 높다는 것을 알 수 있다. 또한 모든 알고리즘이 전체적으로 mi3, si3 에 대한 성능이 낮게 나온다는 점을 확인할 수 있으며 이는 에세이에 따라 3 개의 주장을 하지 않는 에세이도 있어 성능이 낮게 나오는 것이라고 추측한다. 하지만 CRF 를 비롯한 discriminative 한 모델들은 전반적으로 모든 성분에 대해 고르게 성능이 월등하다는 것을 확인할 수 있다.

마지막으로 흥미롭게 눈 여겨 볼 사항은 generative 모델 중에서는 HMM 이, discriminative 모델 중에서는 CRF 가 가장 성능이 높다는 점이다. 즉, generative model, discriminative model 내에서는 time-series 모델의 성능이 일반 분류 알고리즘보다 성능이 더 높게 나왔다는 점을 확인할 수 있다.

6. 결과

6.1 Time-Series vs. Non Time-Series

글의 구성요소가 순서적인 영향을 받는지 확인하기 위해 각 에세이 내의 문장의 순서를 랜덤으로 뒤바꾼 후 다시 CRF 와 HMM 을 적용한 결과 표 4 와 같은 결과가 나왔다.

<표 4> 문장의 순서를 뒤바꾼 것과의 성능비교

	HMM	CRF
Normal	0.7650	0.872
Random	0.3750	0.7163

문장의 순서를 뒤바꾼 경우 HMM 은 절반 이상의 성능이, CRF 는 경우에는 16% 이상의 차이가 나왔으며 이를 통해 토플 에세이의 구성요소가 어느 정도 순서적인 영향을 받는다는 것을 알 수 있다.

6.2 Generative vs. Discriminative

실험을 진행하면서 한가지 특이사항은 time-series 모델인 HMM 이 몇몇 분류 알고리즘보다 성능이 떨어진다는 점이다. 하지만 표 3 을 보면 generative 모델이 전반적으로 discriminative 모델보다 성능이 떨어진다는 것을 볼 수 있다.

Generative 모델은 결합확률분포 argmaxyp(x|y)p(y)에

기반한다. 따라서 p(x|y)를 구해야 하며 이는 데이터를 모델링해야 한다는 것을 의미한다. 데이터의 자질이 많아질수록 모델링하기 어려워지며 이를 간소화하기 위해 그림 6 과 같이 조건부 확률의 독립성을 가정한다. [8]

$$p(C_k, x_1, x_2, \ldots, x_n) = p(C_k)p(x_1, x_2, \ldots, x_n|C_k)$$
$$= p(C_k)p(x_1|C_k)p(x_2|x_1, C_k)\ldots$$
$$\approx p(C_k)p(x_1|C_k)p(x_2|C_k)\ldots p(x_n|C_k)$$

<그림 6> 조건부 확률의 독립성

하지만 구성요소가 주어졌을 때 각 자질의 독립성을 가정하기는 어렵다. 단적인 예로 문단 번호는 분명히 문장 번호와 독립적인 자질은 아니다.

이에 반해 discriminative 모델은 argmaxyp(y|x)에 기반하기 때문에 데이터에 대한 가정을 하지 않는다. 이러한 이유로 현재 데이터를 통해 추출된 자질들은 discriminative 모델을 이용해 분석하는 것이 더 적합하며 discriminative 모델들이 generative 모델들보다 성능이 더 잘 나온 것을 설명할 수 있다.

6.3 향후 연구과제

향후에는 ts, mi, si 뿐만 아니라 추가적으로 introductory material, conclusion 등의 추가적인 구성요소 분류를 예측하는 것이 연구과제가 될 것이다.
또한 RBM, SVM 등의 추가적인 알고리즘과의 성능비교 및 분석도 흥미로운 연구과제가 될 것이다.

참고문헌

[1] Yigal Attali. Jill Burstein. Automated Essay Scoring With e-rater V.2. The Journal of Technology, Learning, and Assessment Volume 4, Number 3. February 2006

[2] Jill Burstein. Martin Chodorow. Claudia Leacock. Automated Essay Evaluation: The Criterion Online Writing Service AI Magazine Volume 25 Number 3 2004

[3] Derrick Higgins. Jill Burstein. Daniel Marcu. Claudia Gentile. Evaluating Multiple Aspects of Coherence in Student Essays. Handbook of Automated Essay Evaluation: Current Applications and New Directions. 267-280 2013

[4] Jill Burstein. Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. IEEE Computer Society. 2003.

[5] Jill Burstein. Daniel Marcu. Slava Andreyev. Towards Automatic Classification of Discourse Elements in Essays. In Proceedings of ACL-01 2001

[6] Jill Burstein. Daniel Marcu. A Machine Learning Approach for Identification of Thesis and Conclusion Statements in Student Essays. Computer and the Humanities 37: 455-467. 2003

[7] Charles Sutton. Andrew McCallum. An Introduction to Conditional Random Fields for Relational Learning.

[8] John Lafferty. Andrew McCallum. Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th International Conference on Machine Learning 2001

[9] Jennry Rose Finkel. Alex Kleeman. Christopher D. Manning. Efficient, Feature-Based, Conditional Random Field Parsing. Proceedings of ACL-08: HLT, pages 959–967. 2008

[10] Yi Mao. Guy Lebanon. Sequential Models for Sentiment Prediction Proceedings of the ICML Workshop on Learning in Structured Output Spaces 2006

[11] Sha, Fei. Pereira, Fernando. Shallow Parsing with Conditional Random Fields. In Proceedings of the 2003 Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics. 2003