|  |
| --- |
| **이종 사이트 메타데이터 특징을 고려한 확장 가능한 동명이인 처리 기법** |
| Jun-Hyeok Jang\*, Do-Jin Choi\*, Hyeon-byeong Lee\*, Jong-Tae Lim\*, Kyoung-Soo Bok\*\*, Jae-Soo Yoo\*†  \*Chungbuk National University, \*\*Wonkwang University  E-mail : jbdata@cbungbuk.ac.kr, mycdj91@cbnu.ac.kr, [lhb@cbnu.ac.kr](mailto:lhb@cbnu.ac.kr), jtlim@chungbuk.ac.kr, ksbok@wku.ac.kr, yjs@cbnu.ac.kr |
|  |

# Introduction

실생활에서 여러 사람이 같은 이름을 갖는 것은 흔한 일이다. 논문에는 저자의 이름과 소속, 공동저자 등의 여러 정보가 존재한다. 이런 상황에서 내가 원하는 논문을 찾기 위해 저자의 이름으로 논문을 검색할 때 동명이인이 존재하는 경우 내가 원하는 논문의 저자인지 알 수 없다. 이렇듯 같은 이름에 대한 문제를 해결하기 위해서 다양한 연구들이 진행되었으며, 대표적으로 [1]에서는 rule-based algorithm을 이용한 휴리스틱 규칙과 분류기를 사용하여 나온 유사성으로 HAC에 적용하여 동명이인을 구별하는 방법을 제시하였다. 또한, [2]에서는 논문과 저자를 동일한 공간에 임베딩하고, 그 공간에서 구조적 정보 및 속성을 이용해 deep learning algorithm으로 밀접하게 연결된 개체의 근접성을 찾아 HAC에 적용하여 동명이인을 구별하는 방법을 제시하였다. 하지만, 논문을 제공하는 학술 데이터베이스마다 메타데이터를 고려하는 특징이 다르다. 그렇기 때문에 앞에서 소개한 동명이인 구별 알고리즘들은 학술 데이터베이스에 필요로 하는 메타데이터 속성이 존재하지 않으면 알고리즘 실행 자체가 불가능하기 때문에 본 논문에서 제안하는 기법이 필요하다.

본 논문에서는 메타데이터 특징을 고려한 확장 가능한 동명이인 처리 알고리즘을 제안한다. 제안하는 기법은 동명이인을 구별하는 알고리즘을 실행하기 위해 학술DB마다 다르게 존재하는 메타데이터를 처리하여 데이터의 특징이 다양하게 고려될 수 있는 다중분류기를 활용한다. 또한, 이 다중분류기에서는 사용자 입력 margin 기반의 다중 분류를 수행하고, 상황에 맞는 알고리즘을 채택하여 동명이인을 구별할 수 있게 한다.

# The proposed scheme

2.1 전체 처리 과정

본 논문에서는 Multi-Classifier를 이용하여 데이터 속성에 알맞은 알고리즘을 채택한다. Figure 1은 제안하는 기법의 전체적인 시스템을 나타낸다.

인풋 데이터인 ambiguous data는 내가 검색한 저자 그리고 저자와 같은 이름을 가진 모든 동명이인들의 데이터이며, 아웃풋 데이터인 Unambiguous data는 알고리즘 과정을 거쳐서 나온 내가 원하는 저자 한 명의 데이터를 의미한다. 다음은 이 시스템이 진행되는 순서이다.

첫째, ambiguous data가 입력되면 동명이인 처리 알고리즘 모두에 적용하여 클러스터링 결괏값을 얻어내고 loss 값을 이용하여 Multi-Classifier를 학습시킨다. 둘째, 학습된 알고리즘에 새로운 ambiguous data가 들어오면 어떤 알고리즘을 이용할지 학습된 Multi-Classifier가 결정을 하여 그 알고리즘에 데이터를 할당하고 처리한다. 이렇게, 기존의 방식에서는 메타데이터의 형식에 맞추어 제한된 선택을 하며 알고리즘을 실행하였지만, 본 논문의 시스템을 이용하여 메타데이터의 형태에 따라 다중 알고리즘을 선택하여 사용할 수 있게 되었다.

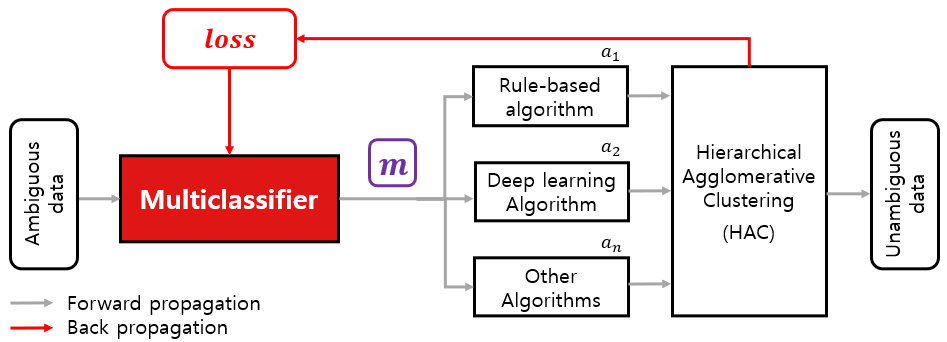


Figure 1. An overview the proposed disambiguation system

2.2 Multi-classifier

다음은 Multi-classifier을 학습하는 과정은 다음과 같다.

ambiguous data의 속성에 맞는 알고리즘을 채택하기 위해서 다중 분류를 처리할 수 있는 활성화 함수 Softmax를 사용하여 Multi-classifier를 학습시킨다. Softmax 함수는 결괏값들의 총합이 1이 되도록 압축하고, 그중 가장 높은 값을 정답으로 선택할 수 있게 한다. Figure 2는 multi-classifier를 학습하는 과정을 나타낸다. 다음은 Multi-classifier을 학습하는 과정이다.

1. 학습되지 않은 Multi-classifier에 새롭게 입력된 ambiguous data 벡터 값에 임의로 지정된 값을 이용해 가중치 곱을 진행하여 알고리즘의 입력 확률값을 구한다.

2. 소프트맥스를 이용하여 총합이 1이 되는 0~1 사이의 예측값을 얻는다.

3. 예측값으로 해당 알고리즘에 메타데이터를 입력하여 loss 값을 구한다.

4. loss 값을 이용해 Back propagation을 진행하여 가중치를 업데이트한다.

다음은 학습된 Multi-classifier을 이용하여 알고리즘을 채택하는 과정은 다음과 같다.

1. 업데이트된 예측값으로 이 예측값을 기준으로 다중분류를 수행하여 알고리즘을 채택한다.

2. 다중 분류 결괏값에 대하여 margin 임계값을 정한다. 채택된 알고리즘과 정해진 margin 임계값 안에 들어오는 알고리즘도 같이 채택하여 사용자에게 제시한다.

3. margin 임계값 안에 들어오는 결괏값으로 다시 loss 값을 구해서 가중치를 업데이트한다. Figure 2에서 은 마진을 나타낸다.

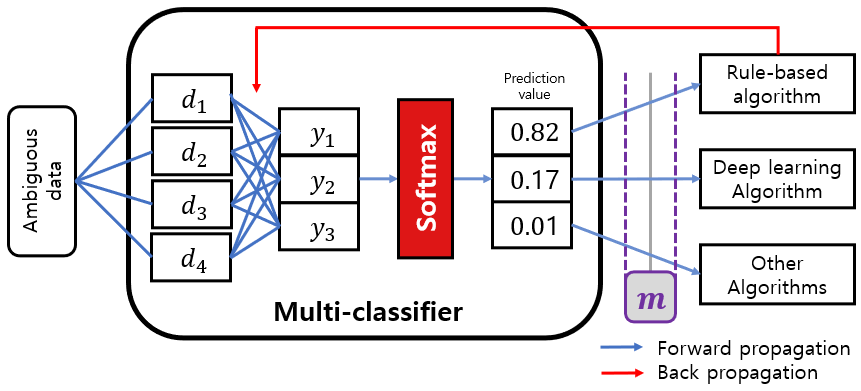


Figure 2. The training of Multi-classifier

1. Conclusion

본 논문에서는 메타데이터 특징을 고려한 확장 가능한 동명이인 처리 알고리즘을 제안하였다. 제안하는 기법은 학술 DB에 각기 다르게 존재하는 메타데이터를 이용하여 동명이인을 구별하는 알고리즘을 실행하기 위해서 데이터의 다양한 특징을 고려할 수 있는 다중 분류기를 제안하였다. 또한 다중 분류기에서 사용자 입력 margin 기반으로 다중 분류를 수행하여 가장 알맞은 알고리즘을 채택하는 기법을 제안하였다.

제안하는 기법은 학술 데이터베이스 특징에 맞는 알고리즘을 선택할 수 있다. 또한 새로운 동명이인 처리 알고리즘을 추가할 수 있는 확장 가능한 시스템을 설계하였다. 향후 연구로는 실제 구현을 통하여 이종 사이트 메타데이터 특징을 고려한 확장 가능한 동명이인 처리 기법의 타당성을 검증할 예정이다.

References

[1] Protasiewicz, J., & Dadas, S. “A hybrid knowledge-based framework for author name disambiguation.” IEEE International Conference on Systems, Man, and Cybernetics. 2016.

[2] Chen, Y., Yuan, H., Liu, T., & Ding, N. “Name Disambiguation Based on Graph Convolutional Network.” Scientific Programming. 2021.