능동 학습 기법을 활용한 한국어 금융 도메인 개체명 인식 데이터 구축

정동호[a,O], 허민강[a], 김형철[b], 박상원[a]
딥네츄럴[a], KB국민은행[b]
dongho@deepnatural.ai, minkang@deepnatural.ai, yhdosu@kbfg.com, anson@deepnatural.ai

Constructing Korean Named Recognition Dataset for Financial Domain Using Active Learning

Dong-Ho Jeong[a,O], Min-Kang Heo[a], Hyung-Chul Kim[b], Sang-Won Park[a]
DeepNatural Inc[a], Kookmin Bank[b]

요 약

딥러닝 모델의 성능은 데이터의 품질과 양에 의해 향상된다. 그러나 데이터 구축은 많은 비용과 시간을 요구한다. 특히 전문 도메인의 데이터를 구축할 경우 도메인 지식을 갖춘 작업자를 활용할 비용과 시간이 더욱 제약적이다. 능동 학습 기법은 최소한의 데이터 구축으로 모델의 성능을 효율적으로 상승시키기 위한 방법이다. 다양한 데이터셋이 능동 학습 기법으로 구축된 바 있으나, 아직 전문 도메인의 한국어 데이터를 구축하는 연구는 활발히 수행되지 못한 것이 현실이다. 본 논문에서는 능동학습기법을 통해 금융 도메인의 개체명 인식 코퍼스를 구축하였고, 이를 통해 다음의 기여가 있다: (1) 금융 도메인 개체명 인식 코퍼스 구축에 능동 학습 기법이 효과적임을 확인하였고, (2) 이를 통해 금융 도메인 개체명 인식기를 개발하였다. 본 논문이 제안하는 방법을 통해 8,043문장 데이터를 구축하였고, 개체명 인식기의 성능은 80.84%로 달성되었다. 또한 본 논문이 제안하는 방법을 통해 약 12~25%의 예산 절감 효과가 있음을 실험으로 보였다.

주제어: 능동 학습, 개체명 인식, 금융 도메인

1. 서론

최근 인공지능 연구의 빠른 성장을 통해, 다양한 분야의 산업에서 이를 적용하고자 하고 있다. 자연언어처리 분야에서는 벤치마크 테스트인 GLUE[1]나 SQuAD[2] 등에서 딥러닝 모델이 인간보다 높은 성능을 보이기도 하였다. 이러한 딥러닝 모델은 전문가에 의한 모델 설계 이외에도, 모델 학습을 위한 컴퓨팅 파워의 비용, 시간, 그리고 데이터 구축 등의 다양한 비용이 발생하여 산업계에 쉽게 적용하기 어려운 요소가 된다. 특히 학습데이터 구축은 인간의 노력이 들어간다는 점에서 하드웨어와 달리 비용 절감이 쉽지 않으며, 또한 금융도메인과 같은 특정 분야의 산업에서는 전문 도메인 지식을 갖춘 작업자가 요구되어 더욱 비용과 시간이 많이 요구된다. 이러한 전문 도메인의 데이터 구축을 절감하기 위해 비지도 학습 기법[3]이나 크라우드소싱[4] 등의 다양한 연구가 수행되었다. 이러한 방법에 의한 구축된 데이터는 전문가에 의한 주석에 비해 양적으로는 더욱 많지만, 데이터의 품질이 상대적으로 낮다는 한계가 있다.

능동 학습(Active Learning)이란 전문가에 의해 데이터를 구축하되 보다 적은 데이터로 최적의 성능을 달성하는 것을 목적으로 하는 기법이다. 능동 학습 기법은 학습 데이터가 제한된 상황에서 어떤 데이터를 우선 학습할 것인지를 사용자와의 상호작용을 통해 결정한다. 이러한 상호작용의 목적은 학습 모델의 성능을 효율적으로 향상시킬 수 있는 데이터를 파악하는 것이다. 능동 학습 기법은 학습 초기 소수의 학습 데이터로 학습된 모델을 사용하여, 모델의 성능 향상에 효율적인 데이터만을 추가 학습데이터로 작업자에게 요청(쿼리, Query)하여 최소한의 데이터를 구축하는 비용 절감을 위한 기법이다. 이러한 능동 학습 기법은 정보 추출, 개체명 인식, 텍스트 분류 등의 다양한 자연언어처리 분야에서 효과적인 것으로 알려졌다[5]. 해외의 연구에서는 일반 도메인 및 전문 도메인에 대한 다양한 능동 학습 연구가 수행되었지만, 국내의 연구에서는 아직 전문 도메인 데이터 구축을 위한 능동 학습 연구가 활발히 수행되지는 못하였다.

본 논문은 금융 도메인의 개체명 인식 시스템을 위한 능동 학습 기반의 데이터 구축을 연구하였다. 이를 통해 본 논문은 다음의 기여가 있다:

(1) 능동 학습 기법이 금융 도메인 개체명 인식 코퍼스 구축에도 효과적임을 실험을 통해 확인하였다. 특히, 3000~3500 문장 데이터를 구축하는 과정에서 12.5%~25%의 예산절감 효과가 있음을 확인하였다.

(2) 이를 통해 80.84%의 성능을 갖는 금융 도메인 개체명 인식기를 개발하였다.

본 논문의 구성은 다음과 같다. 2장에서는 능동 학습 기법과 관련된 국내외 논문과을 살펴보며 능동 학습 기법의 연구 방향을 살펴보고, 3장에서는 금융 도메인 개체명 인식 코퍼스를 구축하기 위한 시스템 구성도를 기술한다. 4장에서는 이를 평가하고, 그 결과를 5장에서 논의하였다.

2. 관련 연구

2.1 능동 학습 기법

능동 학습 기법과 구분되는 개념으로는 수동 학습 기법(Passive learning)이 있다. 수동 학습 기법은 능동 학습 기법처럼 쿼리를 통해 학습 데이터를 선택하는 과정 없이 데이터의 순서대로 학습을 진행하거나 임의로 학습 데이터를 선택하여 학습하는 일반적인 기계학습을 의미한다. 능동 학습 기법은 수동 학습 기법과 비교하여 다양한 학습 데이터 선택 방식이 존재한다[5]. 어떤 학습 데이터를 선택하여 우선 학습할지에 대한 가장 간단하고 방식은 불확실성 샘플링(Uncertainty sampling) 방식이다. 특정 데이터에 대해 모델 예측이 불확실한 경우, 해당 데이터에 대한 정보가 부족하다 가정하여 우선 학습 대상 데이터로 선택하는 직관적인 방식이다. 그리고 불확실함을 측정하는 기준에 따라 Least Confidence(LC), Margin Sampling, Entropy Sampling 등으로 샘플링 방식이 나뉘는데, 본 논문에서는 가장 직관적인 LC 샘플링 방식을 사용하였다. LC 샘플링 방식은 학습된 모델의 예측에 대해 Softmax 등의 확률값이 가장 낮게 측정된 데이터들을 우선 사용하는 방법이다. 다시 말해, 현재의 모델이 잘 풀지 못하는 데이터를 먼저 학습데이터로 선택하는 방법이다.

2.2 국내외 연구 동향

국외에서는 개체명 인식[6]을 포함해 다양한 자연언어처리 분야에서 능동 학습 기법 적용 연구가 이루어졌고[5], 또한 금융 도메인을 포함해[8] 다양한 도메인 적용에 대한 연구[7]도 이루어 졌다. 국내의 경우, 문장 분류[9], 개체명 인식[10] 등의 분야에서 능동 학습 기법 적용이 이루어진 바 있으나, 금융 도메인과 같은 전문 지식이 요구되는 도메인에 대한 적용은 아직 활발히 이루어진 바 없다. 본 논문은 PLO(사람, 장소, 조직)으로 대표되는 일반 도메인의 개체명 인식이 아닌, 40개의 금융 관련 도메인의 개체명(금융기관, 방송국, 경제 관련 기관, 펀드 및 파생상품 등의 금융상품 등)을 인식하는 금융 도메인 개체명 인식 코퍼스 구축에 대한 능동 학습 기법 적용의 연구를 수행하였다.

3. 능동 학습을 사용한 금융 도메인 개체명 코퍼스 구축

본 장에서는 본 연구에 적용된 능동 학습 기법을 기술한다.

본 연구에서는, 금융 도메인에 대해 능동 학습을 적용해 개체명 인식 데이터를 구축하였다. 개체명 인식이란 주어진 문장에서 사람, 장소, 조직 등을 나타내는 개체를 발견하고 분류하는 문제이다. 금융 도메인에서는 그 이외에도 금융기관과 일반 기관을 구분하고, 금융상품(펀드 등) 및 금융 이론과 현상 등에 대한 개체명 태깅이 필요한, 전문 지식이 요구되는 주석 작업이다. 이러한 전문 도메인에 대한 개체명 코퍼스 구축은 전문 지식이 필요하여 데이터 구축의 비용은 물론, 주석 작업을 수행할 작업자도 현실적인 제약이 있다. 따라서 이러한 문제를 다루기 위해 능동 학습 기법을 적용해 최소한의 주석 작업만으로 모델의 성능 향상을 보일 수 있는지를 연구하였다. 그림 1은 능동 학습을 사용한 금융 도메인 개체명 코퍼스 구축 작업 전반의 개념도이다.

작업 순서는 다음과 같다. (1) 먼저, 금융 도메인 원시 코퍼스로부터 임의의 주석 대상 데이터를 선정한다(예: 500문장). (2) 해당 데이터에 대해 주석 플랫폼을 통해 개체명 주석 작업을 수행한다 .(3) 이렇게 만들어진 데이터를 사용, 모델을 학습한다. (4) 해당 모델을 사용하여 원시 코퍼스에 대해 예측을 수행한다. 이 과정에서 특정 조건에 속하는 주석 대상 데이터를 선정한다. 본 연구에서는 2.1장에서 논의된 LC 방법을 적용하여 주석 대상 데이터를 선정하였다. (1) ~ (4) 과정을 반복하면서 모델의 성능이 수렴하는 지점에서 데이터 구축을 종료한다. 이러한 작업은 모델이 예측하지 못하는 문장들을 먼저 학습하는 과정을 통해 모델의 성능에 보다 많이 기여할 것으로 기대되는 데이터들을 선별한다. 다시 말해, 특정 규모 이상의 데이터는 성능 향상에 점차 적게 기여하는 경향이 있기 때문에 최소한의 데이터 만으로 모델의 성능이 효율적으로 최적화 되는 것을 기대할 수 있다. 결과적으로, 임의 선정(random sampling) 방법에 비해 보다 적은 데이터를 구축하여도 보다 높은 성능을 기대할 수 있다. 이러한 과정을 통해 데이터 구축 비용의 절감 효과를 기대할 수 있다.

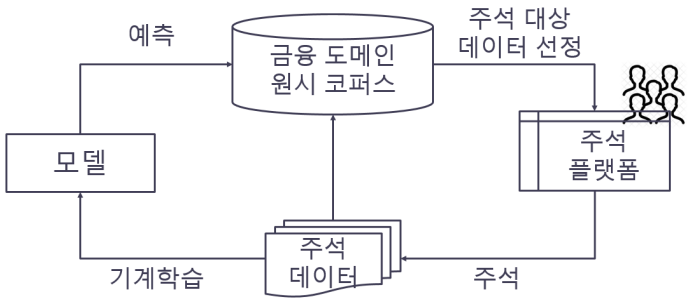

그림 1. 능동 학습을 사용한 금융 도메인 개체명 코퍼스 구축 개념도

4. 실험 진행 및 결과 논의

4.1 데이터

금융 도메인에서의 능동 학습의 효과를 확인하기 위해 웹 상의 금융 도메인 텍스트들을 수집하였다.[1] 해당 텍스트는 금융 관련 키워드를 포함한 문장들을 크롤링하는 방법을 통해 수집되었다. 결과적으로 총 9,043 문장을 수집하였다. 본 연구에서는 먼저 1,043개 문장에 대해 개체명 태깅을 수행하여 이를 평가데이터로 사용하고, 나머지 8,000문장을 그림 1에서의 금융 도메인 원시 코퍼스로 간주하였다. 개체명 태깅 작업은 그림 2과 같은 형식의 주석 플랫폼[2]을 사용하여 총 40개의 개체명 태그셋을 사용하여 이루어졌다.

1) 본 데이터셋은 KB의 협력을 통해 수집되었습니다.
2) http://app.deepnatural.ai

4.2 실험 방법

실험은 금융 도메인 원시 코퍼스로부터 초기 주석 대상 데이터를 임의로 선택하며 시작된다. 선택된 주석 대상 데이터는 그림 2의 주석 플랫폼을 통해 주석 대상 문장에 대해 개체명 인식 태그를 주석하였다. 주석된 문장은 KB에서 제공한 형태소 분석기를 활용해 형태소 단위로 분리되어 학습 데이터로 사용하였다. 본 논문에서는 개체명 인식기를 개발하기 위해 Multilingual-BERT모델을 원 논문의 방법으로 미세조정(fine-tuning) 하는 방식으로 개체명 인식기를 구현하였다[10].

실험 방법은 다음과 같다. Least Confident 기법을 활용하여 대상 데이터를 선택한(LC Sampling)하는 능동 학습 기법으로 모델을 점차 개선하는 방법과, 임의의 문장들을 선택하는 방법(Random Sampling)으로 모델을 각각 학습하였다. 각각의 기법으로 학습된 모델 간 성능 향상 추이를 비교하기 위해서 모델의 학습과 평가는 학습 문장 수를 500씩 증가시키며 진행되었다. 500문장부터 4000문장까지 총 8번의 스텝을 거쳐 모델을 생성하였다. 또한, 실험의 객관성 확보를 위해 각각의 스텝을 10번 반복하였다. 실험에 사용한 하이퍼 파라미터는 표 1과 같다.

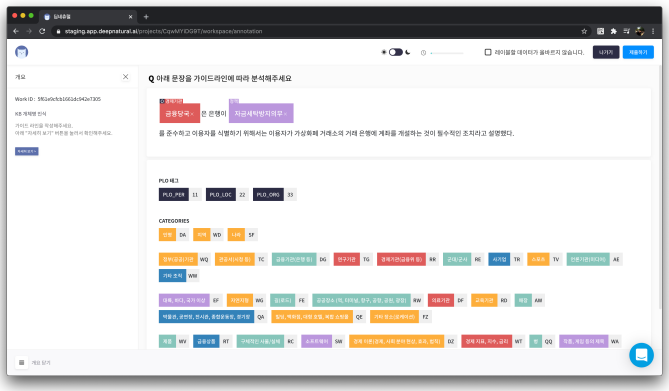

그림 2. 주석 플랫폼의 가공 도구

표 1. 하이퍼 파라미터 정보

하이퍼 파라미터	값
최대 문장 길이	256
배치 사이즈	6
학습률	3e-5
Optimizer	Adam
Epoch	10

4.3 실험 결과

능동 학습 기법(LC와 임의로 대상 데이터를 선택한 방법(Random Sampling)을 통해 학습된 모델에 대한 F1 평가는 그림 3의 그래프로 나타났다. 먼저, 학습 문장 수가 적은 상황(500~1,500 문장)인 경우 모델 성능에 유의미한 차이가 없었다. 이는 다시 말해 모델 성능에 필요한 의미 있는 정보는 최소한 1,500개의 문장이 필요하다고 보여진다. 이후, 학습 문장 수가 2000개가 넘어가면 LC 모델 학습법이 Random Sampling에 비해 좋은 성능을 가진 모델을 생성해냄을 확인할 수 있다. 그리고 학습데이터가 4,000문장을 넘어서는 경우에는 LC 모델과 Random Sampling 모델이 점차 성능이 비슷해지는 경향을 확인할 수 있었다. 특히, Random sampling의 F1은 79%로서, LC 모델이 3000~3500 문장을 학습했을 때의 성능과 유사하다. 이를 통해 능동 학습 기법을 활용하면 12.5 ~ 25%의 예산을 절감할 수 있다는 가능성을 보였다.

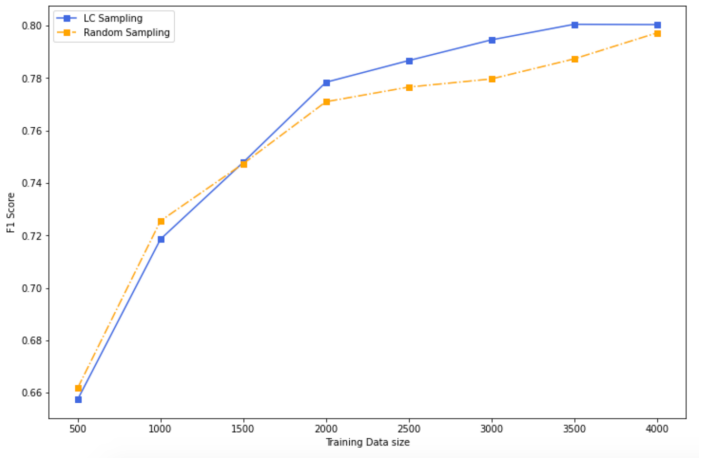

그림 3. 각각의 기법을 활용한 모델 성능 평가 추이

주석 대상 데이터를 생성하는 총 8번의 단계 중, 각각의 단계에서 선택된 500 문장의 태그셋 비율에 대한 정보는 표 2와 같다. 표 2에서는 지면의 한계 상 4번째 단계까지만 명시하였다.

표 2. 단계에 따른 상위 태그셋 5개 비율(%)

STEP	LC Sampling 상위 5개 태그 비율(%)	Random sampling 상위 5개 태그 비율(%)
#1	AMOUNT(25)	AMOUNT(20)
	PRICE(20)	DATETIME(17)
	DATETIME(15)	PRICE(15)
	FINANCIAL_INSTITUTION(10)	FINANCIAL_INSTITUTION(15)
	PERSON(4)	PERSON(5)
#2	FINANCIAL_INSTITUTION(15)	AMOUNT(24)
	COMPANY_AND_BRAND(8)	PRICE(17)
	GOVERNMENT(7)	DATETIME(17)
	FINANCIAL_PRODUCT(7)	FINANCIAL_INSTITUTION(10)
	VALUE(6)	VALUE(6)
#3	FINANCIAL_INSTITUTION(15)	AMOUNT(23)
	DATETIME(12)	PRICE(16)
	AMOUNT(11)	DATETIME(14)
	AREA(6)	FINANCIAL_INSTITUTION(13)
	GOVERNMENT(5)	PERSON(5)
#4	PRICE(21)	AMOUNT(28)
	FINANCIAL_INSTITUTION(14)	PRICE(19)
	DATETIME(11)	DATETIME(13)
	AMOUNT(8)	FINANCIAL_INSTITUTION(9)
	FINANCIAL_PRODUCT(6)	VALUE(5)

Random Sampling을 통해 선택된 문장들의 경우, ‘AMOUNT,’ ‘PRICE’, ‘DATETIME’, ‘FINANCIAL_INSTITUION’등의 태그셋들이 주로 선택되었음을 알 수 있다. 이는 실제 수집된 전체 원시 코퍼스에서 빈번하게 등장하는 태그셋을 의미한다. 주석 처리된 전체 문장에 대한 태그셋 분포 역시 ‘DATETIME’, ‘AMOUNT’, ‘FINANCIAL_INSTITUION’, ‘PRICE’ 등의 태그가 많이 등장하는 것으로 파악되었다. LC Sampling의 경우, 전체 데이터의 태그 셋 비율과 무관하게 다양한 태그 셋들이 상위 태그 셋으로 등장하였다.

5. 결론

본 논문에서는 능동 학습 기법을 통해 금융 도메인의 개체명 인식 코퍼스를 구축하였다. 실험을□통해 능동 학습 기법을 적용할 경우, 보다 빠르게 모델의 성능을 효율적으로 향상시킬 수 있음을 보였다. 구축된 데이터셋은 모델이 잘 해석하지 못하는 태그셋을 중심으로 구축되어, 기계학습 관점에서 좀 더 유의미한 품질의 데이터가 구축되었음을 확인할 수 있었다. 또한 비용의 측면에서도 능동 학습 기법이 12.5%~25%의 예산 절감 효과가 있음을 실험으로 확인되었다. 본 논문을 통해 총 8,043 문장 데이터가 구축되었고, 성능 80.84%의 금융 도메인 개체명 인식기가 개발되었다.

감사의 글

본 연구는 중소벤처기업부에서 지원하는 2019년도 창업성장기술개발사업(팁스프로그램, No. S2816383)의 연구수행으로 인한 결과물임을 밝힙니다. KB국민은행과 함께 진행한 과제의 결과물입니다.

참고문헌

[1] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.

[2] Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD." arXiv preprint arXiv:1806.03822, 2018.

[3] Taghipour, Kaveh, and Hwee Tou Ng. "Semi-supervised word sense disambiguation using word embeddings in general and specific domains." Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies. 2015.

[4] OOMEN, Johan; AROYO, Lora. Crowdsourcing in the cultural heritage domain: opportunities and challenges. In: Proceedings of the 5th International Conference on Communities and Technologies. p. 138-149. 2011.

[5] Settles, B. "Active learning literature survey: Computer sciences technical report 1648", University of Wisconsin-Madison. 2009.

[6] Shen, Y., Yun, H., Lipton, Z. C., Kronrod, Y., & Anandkumar, A. Deep active learning for named entity recognition. arXiv preprint arXiv:1707.05928. 2017.

[7] Lianghao Li, Xiaoming Jin, Sinno Jialin Pan, and JianTao Sun. Multi-domain active learning for text classification. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1086-1094). 2012.

[8] Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. Stream-based active learning for sentiment analysis in the financial domain. Information sciences, 285, 181-203. 2014.

[9] 김제욱, 김한준, & 이상구. 베이지언 문서분류시스템을 위한 능동적 학습 기반의 학습문서집합 구성방법. 정보과학회논문지: 소프트웨어 및 응용, 29(11 · 12), 966-978. 2012.

[10] 윤보현, 오효정. 능동 학습 기법을 활용한 개체명 사전 반자동 구축 도구 개발. 컴퓨터교육학회 논문지, 18(6), 81-88. 2015.

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.

부록 1. 모델 성능 평가 결과

#Sentence	LC Sampling			Random Sampling		
	P	R	F	P	R	F
500	61.93	70.53	65.75	62.46	69.57	66.19
1000	69.55	73.65	71.86	70.58	75.44	72.55
1500	72.67	78.14	74.79	71.88	75.16	74.74
2000	76.32	79.39	77.84	75.20	77.95	77.10
2500	77.17	79.97	78.67	75.84	78.05	77.66
3000	78.09	81.04	79.46	75.87	79.27	77.97
3500	78.22	81.84	80.05	77.23	81.21	78.73
4000	78.26	82.09	80.04	78.69	80.59	79.72

P = Precision, R = Recall, F = F1-score