사전학습된 대규모 언어 모델 사용한 한국어 문서 내 민감정보 자동 마스킹 처리 방법

인공지능융합학과
MUSAT
구선모, 김병건, 김예은, 김주성, 조시헌

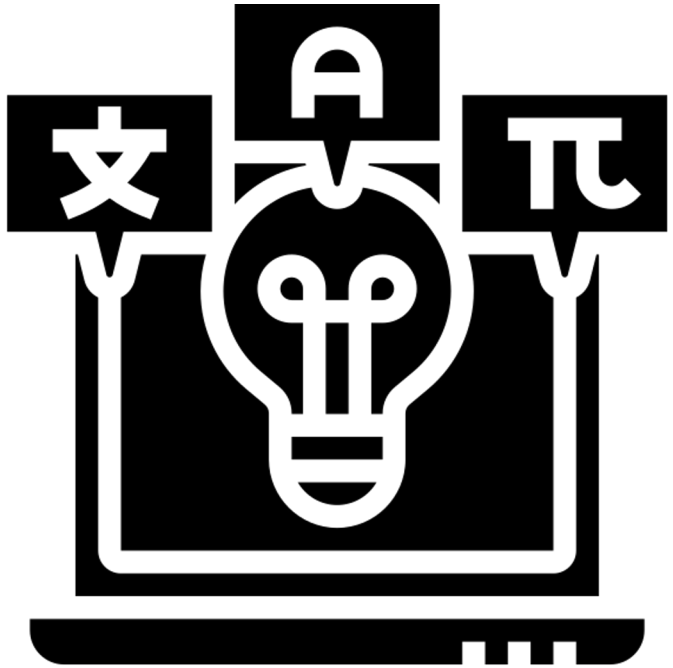
개요

문서 내 개인 민감 정보 식별이 필요한 이유

- 정보 검색과 요약
- 질문 답변
- 지식 베이스 구축
- 기계 번역 (Machine Translation, 이하 MT)의 품질 향상
- 사용자에게 맞춤형 번역 제공

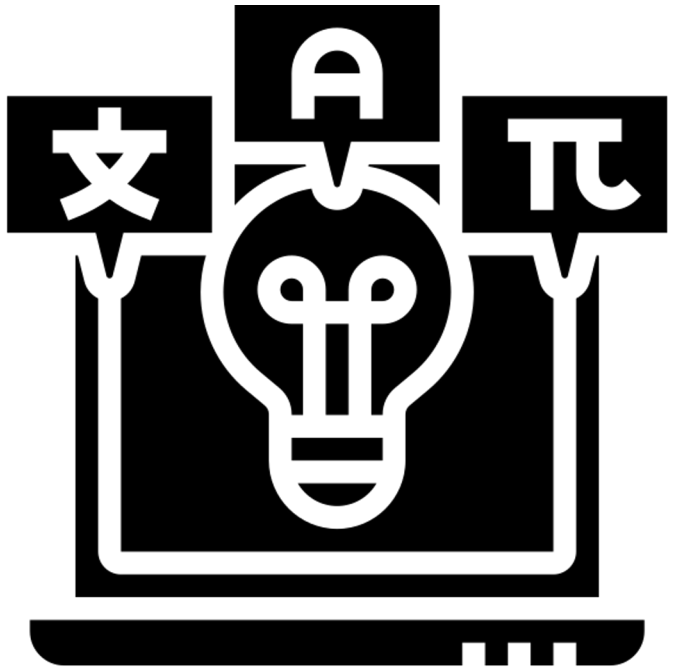
딥러닝 기반의 개체명 인식 모델

- **Rule-based Approaches**
- **LSTM (Bi-directional LSTM)**
- **LSTM (Bi-directional LSTM) - CRF**
- **BERT**
- **BERT - CRF**
- **RoBERTa**
- **RoBERTa - CRF**

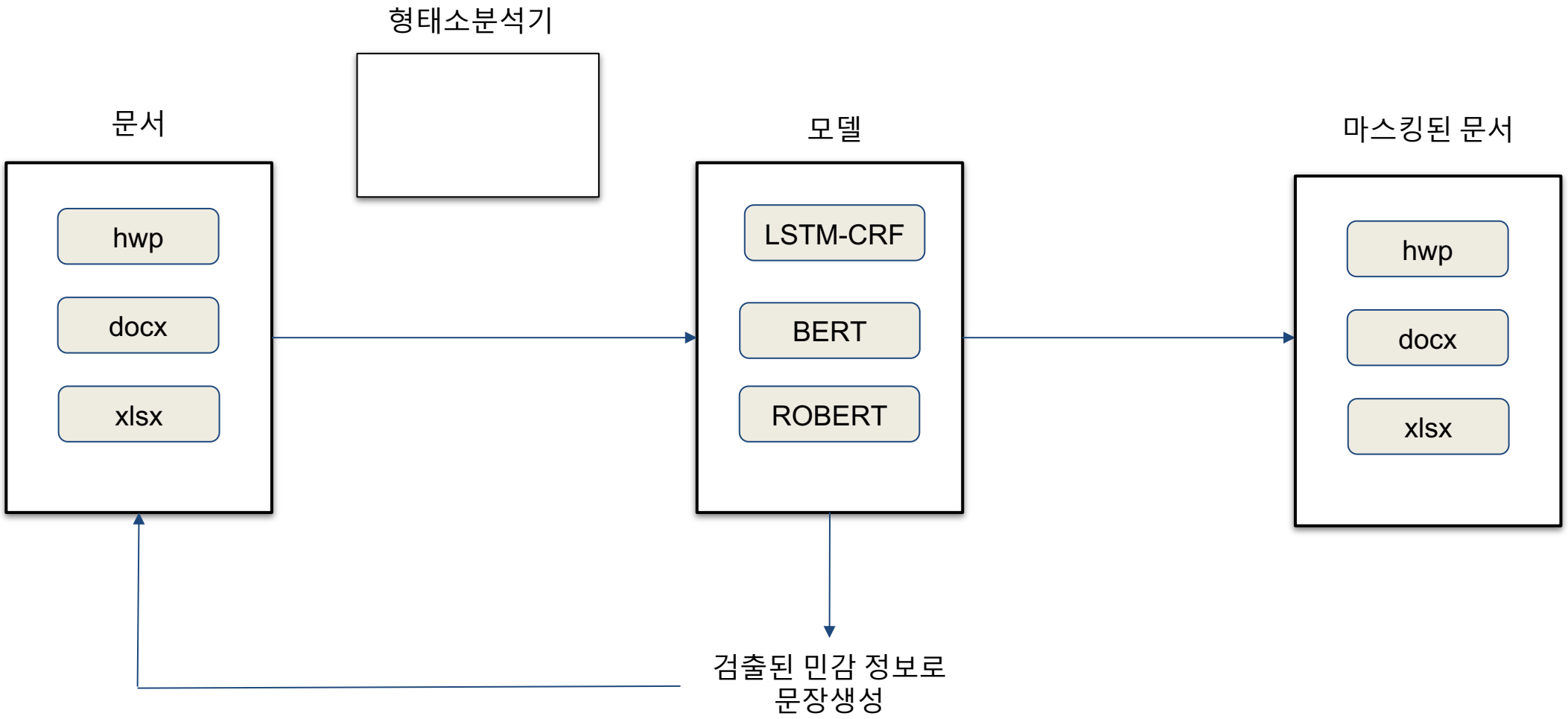


한국어 문서 내 자동 민감 정보 인식

- **규칙기반 개체명 인식**
  - 고유명사 사전이나 접사 사전, 결합명사 사전과 같은 사전 이용
  - 문장에 자주 발생하는 문맥을 활용하는 방법
  - 접사 사전과 결합 규칙을 이용한 방법
  - 규칙과 문맥을 다단계로 적용한 방법
- **통계 기반 개체명 인식**
  - 결정 트리 기반
  - 최대 엔트로피 모델에 기반한 방법



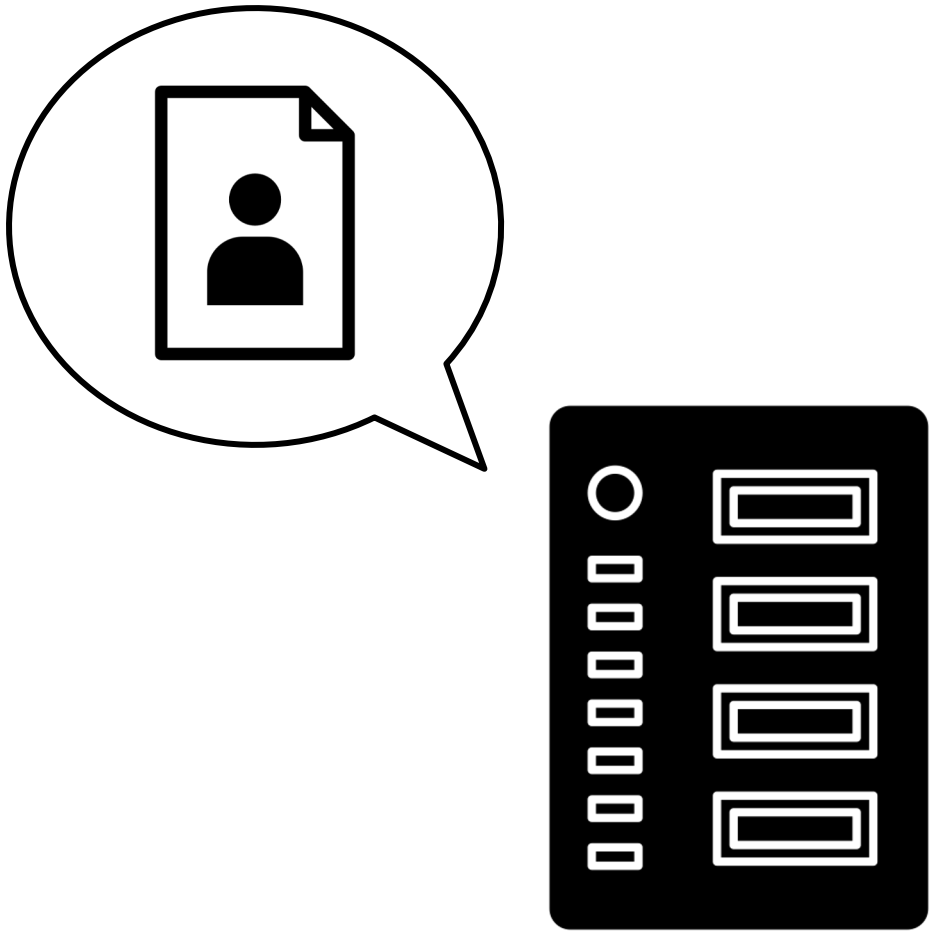
한국어 민감 정보 자동 마스킹을 위한 파이프라인



데이터 셋과 딥러닝 모델

데이터 셋 종류

- exo (Google)
- Naver
- v3
- etc.



딥러닝 모델 종류

- LSTM
- CNN
- GRU
- ElMo
- BERT
- RoBERTa

TABLE 3
Summary of recent works on neural NER. LSTM: long short-term memory, CNN: convolutional neural network, GRU: gated recurrent unit, LM: language model, ID-CNN: iterated dilated convolutional neural network, BRNN: bidirectional recursive neural network, MLP: multi-layer perceptron, CRF: conditional random field, Semi-CRF: Semi-markov conditional random field, FOFE: fixed-size ordinally forgetting encoding.

Work	Input representation			Context encoder	Tag decoder	Performance (F-score)
	Character	Word	Hybrid			
[94]	-	Trained on PubMed	POS	CNN	CRF	GENIA: 71.01%
[89]	-	Trained on Gigaword	-	GRU	GRU	ACE 2005: 80.00%
[95]	-	Random	-	LSTM	Pointer Network	ATIS: 96.86%
[90]	-	Trained on NYT	-	LSTM	LSTM	NYT: 49.50%
[91]	-	SENNA	Word shape	ID-CNN	CRF	CoNLL03: 90.65%; OntoNotes5.0: 86.84%
[96]	-	Google word2vec	-	LSTM	LSTM	CoNLL04: 75.0%
[100]	LSTM	-	-	LSTM	CRF	CoNLL03: 84.52%
[97]	CNN	GloVe	-	LSTM	CRF	CoNLL03: 91.21%
[105]	LSTM	Google word2vec	-	LSTM	CRF	CoNLL03: 84.09%
[19]	LSTM	SENNA	-	LSTM	CRF	CoNLL03: 90.94%
[106]	GRU	SENNA	-	GRU	CRF	CoNLL03: 90.94%
[98]	CNN	GloVe	POS	BRNN	Softmax	OntoNotes5.0: 87.21%
[107]	LSTM-LM	-	-	LSTM	CRF	CoNLL03: 93.09%; OntoNotes5.0: 89.71%
[103]	CNN-LSTM-LM	-	-	LSTM	CRF	CoNLL03: 92.22%
[17]	-	Random	POS	CNN	CRF	CoNLL03: 89.86%
[18]	-	SENNA	Spelling, n-gram, gazetteer	LSTM	CRF	CoNLL03: 90.10%
[20]	CNN	SENNA	capitalization, lexicons	LSTM	CRF	CoNLL03: 91.62%; OntoNotes5.0: 86.34%
[116]	-	-	FOFE	MLP	CRF	CoNLL03: 91.17%
[101]	LSTM	GloVe	-	LSTM	CRF	CoNLL03: 91.07%
[113]	LSTM	GloVe	Syntactic	LSTM	CRF	W-NUT17: 40.42%
[102]	CNN	SENNA	-	LSTM	Reranker	CoNLL03: 91.62%
[114]	CNN	Twitter Word2vec	POS	LSTM	CRF	W-NUT17: 41.86%
[115]	LSTM	GloVe	POS, topics	LSTM	CRF	W-NUT17: 41.81%
[118]	LSTM	GloVe	Images	LSTM	CRF	SnapCaptions: 52.4%
[109]	LSTM	SSKIP	Lexical	LSTM	CRF	CoNLL03: 91.73%; OntoNotes5.0: 87.95%
[119]	-	WordPiece	Segment, position	Transformer	Softmax	CoNLL03: 92.8%
[121]	LSTM	SENNA	-	LSTM	Softmax	CoNLL03: 91.48%
[124]	LSTM	Google Word2vec	-	LSTM	CRF	CoNLL03: 86.26%
[21]	GRU	SENNA	LM	GRU	CRF	CoNLL03: 91.93%
[126]	LSTM	GloVe	-	LSTM	CRF	CoNLL03: 91.71%
[142]	-	SENNA	POS, gazetteers	CNN	Semi-CRF	CoNLL03: 90.87%
[143]	LSTM	GloVe	-	LSTM	Semi-CRF	CoNLL03: 91.38%
[88]	CNN	Trained on Gigaword	-	LSTM	LSTM	CoNLL03: 90.69%; OntoNotes5.0: 86.15%
[110]	-	GloVe	ELMo, dependency	LSTM	CRF	CoNLL03: 92.4%; OntoNotes5.0: 89.88%
[108]	CNN	GloVe	ELMo, gazetteers	LSTM	Semi-CRF	CoNLL03: 92.75%; OntoNotes5.0: 89.94%
[133]	LSTM	GloVe	ELMo, POS	LSTM	Softmax	CoNLL03: 92.28%
[137]	-	-	BERT	-	Softmax	CoNLL03: 93.04%; OntoNotes5.0: 91.11%
[138]	-	-	BERT	-	Softmax +Dice Loss	CoNLL03: 93.33%; **OntoNotes5.0: 92.07%**
[134]	LSTM	GloVe	BERT, document-level embeddings	LSTM	CRF	CoNLL03: 93.37%; OntoNotes5.0: 90.3%
[135]	CNN	GloVe	BERT, global embeddings	GRU	GRU	CoNLL03: 93.47%
[132]	CNN	-	Cloze-style LM embeddings	LSTM	CRF	**CoNLL03: 93.5%**
[136]	-	GloVe	Plooled contextual embeddings	RNN	CRF	CoNLL03: 93.47%

실험 및 결과

데이터 셋 분석 (태깅 종류별 개수 및 비율)

Category	Count	Frequency
O	1487289	84.67%
PER_B	12659	0.72%
PER_I	1976	0.11%
COM_B	34077	1.94%
COM_I	77445	4.41%
LOC_B	11594	0.66%
LOC_I	49290	2.81%
POS_B	8446	0.48%
POS_I	5026	0.29%
EDU_B	5818	0.33%
EDU_I	13297	0.76%
AFF_B	16499	0.94%
AFF_I	33059	1.88%

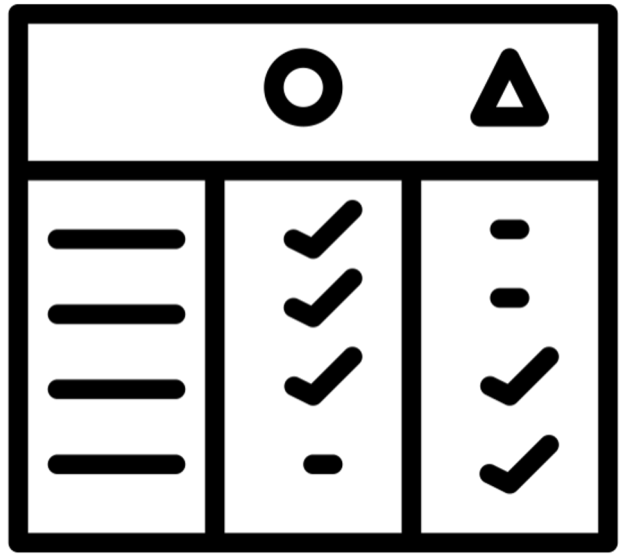
Table 1. train v3 dataset tag distribution

Category	Count	Frequency
O	152461	92.70%
DT_B	1563	0.95%
DT_I	2046	1.24%
LC_B	1956	1.19%
LC_I	81	0.05%
OG_B	1529	0.93%
OG_I	291	0.18%
PS_B	3686	2.24%
PS_I	389	0.24%
TI_B	297	0.18%
TI_I	174	0.11%

Table 2. train exo dataset tag distribution.

모델에 따른 NER 정확도 결과

- **LSTM-CRF**
- **BERT-base**
- **RoBERTa-small**
- **RoBERTa-base**
- **RoBERTa-large (didn’t use, computation capacity)**



모델에 따른 NER 정확도 결과

Weighted F1 score	v3	exo	Naver	Total
LSTM-CRF	69.78	72.44	50.47	81.39
BERT-base	**92.50**	**93.05**	**85.90**	**90.91**
RoBERTa-small	88.61	89.32	77.51	86.40
RoBERTa-base	85.75	89.17	78.01	87.81

Unweighted F1 score	v3	exo	Naver	Total
LSTM-CRF	74.55	54.86	48.14	81.55
BERT-base	**92.22**	**92.42**	**74.33**	**90.94**
RoBERTa-small	86.27	89.70	67.15	86.41
RoBERTa-base	87.01	89.15	66.58	87.68

모델에 따른 예측 결과 분석

sentence 1	병역제도 출항 후에는 부산을 이유로 한 이상네트웍스가 창단했다 . (18970) total
True tagging	O O O LOC_B O O COM_B O O
LSTM-CRF	O O O O O O O O O
BERT-base	O O O LOC_B O NUM_B COM_B O O
RoBERTa-small	O O O LOC_B O O COM_B O O

sentence 2	기술 보안 서약서 소속 : (주) 평화 엔지니어링 총무 팀 직위 : 선임 성명 : 김영우 (4905) v3
True tagging	O O O O O COM_B COM_I COM_I COM_I COM_I AFF_B AFF_I O O POS_B O O PER_B
LSTM-CRF	O O O O O COM_I COM_I COM_I COM_I COM_I AFF_B AFF_I O O POS_B O O PER_B
BERT-base	O O O O O COM_B COM_I COM_I COM_I COM_I AFF_B AFF_I O O POS_B O O PER_B
RoBERTa-small	O O O O O COM_B COM_I COM_I COM_I COM_I AFF_B AFF_I O O POS_B O O PER_B

모델에 따른 예측 결과 분석 (cont'd)

sentence 3	-저는 5월 MBC드라마에서 하프파이프로 데뷔하는 동방신기의 비숙련자 유노윤호 씨를 만나고 왔습니다 . (total 19415)
True tagging	O NUM_B COM_B O O PER_B O O O O O O
LSTM-CRF	O NUM_B O O O PER_B O PER_B O O O O
BERT-base	O NUM_B COM_B O O PER_B O PER_B O O O O
RoBERTa-small	O NUM_B O O O O O PER_B O O O O

sentence 4	5월에는 추모전도 둘러보시고 김약국의 딸들에서 집터까지 박경리 선생의 책도 한차례 읽어보시는 건 어떨까요 . (total18261)
True tagging	NUM_B O O LOC_B O O PER_B O O NUM_B O O O O
LSTM-CRF	NUM_B O O O O O PER_B O O NUM_B O O O O
BERT-base	NUM_I O O PER_B O O PER_B O O NUM_B O O O O
RoBERTa-small	NUM_B O O PER_B O O PER_B O O NUM_B O O O O

문서에 따른 민감 정보 마스킹 결과
(BERT-base trained with total dataset) – example 1

자 기 소 개 서 **Before**

성장과정

2018년 학부생 연구원으로 ANSYS(구조해석 툴)를 이용하여 데모엔지니어링에서 새로 개발 중인 천공기의 CAD파일을 받아 전처리 과정을 거친 후 ANSYS를 이용하여 구조해석을 하는 게 주요 업무입니다.

매주 교수님과 세미나를 갖는데, 처음 CAD파일을 받고 막막했을 때, 교수님께서 바로 구조해석을 시작하는 것도 좋지만 정역학과 재료역학에서 배운 내용을 토대로 직접 손으로 구해보는 것도 좋을 것 같다고 조언을 해주셨습니다.

손으로 자유물체도를 그려서 받는 하중과 무게중심을 고려하여 가장 하중이 많이 받으며 구조적으로 취약한 부분을 찾고 계산하는데 2주 넘게 걸렸습니다. 그러다보니 처음 2주간 교수님과 세미나에서 다른 연구원들은 전처리 과정을 보여준 반면 저는 보여드릴게 없어서 많이 아쉬웠지만 꿋꿋이 직접 계산하고 결과를 찾았습니다.

찾고나니 다른 연구원과는 조금 다른 부분이 취약하다고 나왔습니다. 그리고 저도 해석프로그램을 돌렸을 때 친구들과 같은 부분으로 나와서, 프로그램상 실수를 찾기위해 엔시스 책을 선배에게 빌렸습니다. 유압을 넣는 부분을 잘못하고 있었기 때문에 저와 친구들의 해석 결과와 제가 직접 구해본 결과와 다르다는 것을 알게됐고, 이 부분을 수정하였더니 손으로 구한 결과와 같은 것을 알게되었고, 그 다음번 세미나때 다른 해석결과 때문에 교수님께서 대략적으로 자유물체도를 그리셔서 구해보니 제 해석결과가 맞다고 하셨습니다.

자신을 믿고 해낼 수 있는 힘이 저에게 있다는 것을 깨달았습니다.

자 기 소 개 서 **After**

성장과정

xxx 학부생 xxx ANSYS(구조해석 툴)를 이용하여 xxx 새로 개발 중인 천공기의 CAD파일을 받아 전처리 과정을 거친 후 ANSYS를 이용하여 구조해석을 하는 게 주요 업무입니다.

xxx 교수님과 세미나를 갖는데, 처음 CAD파일을 받고 막막했을 때, 교수님께서 바로 구조해석을 시작하는 것도 좋지만 정역학과 재료역학에서 배운 내용을 토대로 직접 손으로 구해보는 것도 좋을 것 같다고 조언을 해주셨습니다.

손으로 자유물체도를 그려서 받는 하중과 무게중심을 고려하여 가장 하중이 많이 받으며 구조적으로 취약한 부분을 찾고 계산하는데 xxx 넘게 걸렸습니다. 그러다보니 처음 xxx 교수님과 세미나에서 다른 연구원들은 전처리 과정을 보여준 반면 저는 보여드릴게 없어서 많이 아쉬웠지만 꿋꿋이 직접 계산하고 결과를 찾았습니다.

찾고나니 다른 연구원과는 조금 다른 부분이 취약하다고 나왔습니다. 그리고 저도 해석프로그램을 돌렸을 때 친구들과 같은 부분으로 나와서, 프로그램상 실수를 찾기위해 xxx 책을 선배에게 빌렸습니다. 유압을 넣는 부분을 잘못하고 있었기 때문에 저와 친구들의 해석 결과와 제가 직접 구해본 결과와 다르다는 것을 알게됐고, 이 부분을 수정하였더니 손으로 구한 결과와 같은 것을 알게되었고, 그 다음번 세미나때 다른 해석결과 때문에 교수님께서 대략적으로 자유물체도를 그리셔서 구해보니 제 해석결과가 맞다고 하셨습니다.

자신을 믿고 해낼 수 있는 힘이 저에게 있다는 것을 깨달았습니다.

문서에 따른 민감 정보 마스킹 결과
(BERT-base trained with total dataset) – example 2

보 안 서 약 서

Before

본인 신진호 은 강원도 삼척시 원덕읍 삼척로 446 에 소재하고 있는 남선알미늄 에서 퇴직함에 있어 다음 사항을 숙지하고 이를 이행하지 않을 경우 관계 법령에 의거 처벌받을 것은 물론 남선알미늄 에 손해를 입힐 경우에는 그 손해액을 변상할 것을 엄숙히 서약합니다.

1. 남선알미늄 에 근무 중 지득한 국가보안 등에 관한 제반 비밀과 직무상 지득한 과학기술정보 관련 제반 비밀사항 및 주요 기술비밀을 관련 법령, 인사규정 제 10조, 취업규칙 제 2조의 규정에 따라 일체 누설하거나 도용하지 않는다.

2. 남선알미늄 에 근무 중의 모든 발명, 고안, 창작 및 발견 등에 대하여 남선알미늄 총무팀 에게 이를 공개, 양도할 것에 동의하고 그 절차에 적극 협력한다.

3. 남선알미늄 에 근무 중의 모든 연구자료 및 연구결과 보고서, 설계서, 청사진 등과 보조기억장치 등에 대하여는 누락없이 남선알미늄 총무팀에게 인계하고 이를 소지하거나 유출하지 않는다.

4. 퇴직 후 2년 간은 남선알미늄 의 사전 승인 없이 남선알미늄 의 연구자료, 연구결과 등과 직무 발명, 고안, 창작 및 발견사항 등의 지적재산권을 이용하여 자신 또는 제 3자를 위하여 창업하거나, 기업체에 전직, 동업 또는 자문하지 않는다.

5. 위 사항을 위반하는 경우에는 관련 법규(국가보안법, 형법, 부정경쟁방지 및 영업비밀보호에 관련 법률)에 따른 어떠한 처벌도 감수한다.

2019 년 12 월 22 일

서약인 주소 : 경상북도 경주시 한빛길28번길 26

서약인 주민등록번호 : 971014-1457745

서약인 성명 : 신진호

남선알미늄 귀하

보 안 서 약 서

After

본인 xxx 은 xxx xxx xxx xxx xxx 에 소재하고 있는 xxx 에서 퇴직함에 있어 다음 사항을 숙지하고 이를 이행하지 않을 경우 관계 법령에 의거 처벌받을 것은 물론 xxx 에 손해를 입힐 경우에는 그 손해액을 변상할 것을 엄숙히 서약합니다.

xxx. xxx 에 근무 중 지득한 국가보안 등에 관한 제반 비밀과 직무상 지득한 과학기술정보 관련 제반 비밀사항 및 주요 기술비밀을 관련 법령, 인사규정 제 xxx0조, 취업규칙 제 xxx조의 규정에 따라 일체 누설하거나 도용하지 않는다.

xxx. xxx 에 근무 중의 모든 발명, 고안, 창작 및 발견 등에 대하여 xxx 총무팀 에게 이를 공개, 양도할 것에 동의하고 그 절차에 적극 협력한다.

xxx. xxx 에 근무 중의 모든 연구자료 및 연구결과 보고서, 설계서, 청사진 등과 보조기억장치 등에 대하여는 누락없이 xxx 총무팀에게 인계하고 이를 소지하거나 유출하지 않는다.

xxx. 퇴직 후 xxx년 간은 xxx 의 사전 승인 없이 xxx 의 연구자료, 연구결과 등과 직무 발명, 고안, 창작 및 발견사항 등의 지적재산권을 이용하여 자신 또는 제 xxx자를 위하여 창업하거나, 기업체에 전직, 동업 또는 자문하지 않는다.

xxx. 위 사항을 위반하는 경우에는 관련 법규(국가보안법, 형법, 부정경쟁방지 및 영업비밀보호에 관련 법률)에 따른 어떠한 처벌도 감수한다.

xxx0xxx9 년 xxxxxx 월 xxxxxx 일

서약인 주소 : xxx xxx xxx xxx

서약인 주민등록번호 : 97xxx0xxxxxx-xxxxxxxxx77xxxxxx

서약인 성명 : xxx

xxx 귀하

문서에 따른 민감 정보 마스킹 결과
(BERT-base trained with total dataset) – example 3

기술보안 서약서

Before

소속 : 대성물류건설(주) 보안팀

직위 : 상무보

성명 : 서인혁

본인 서인혁 은 울산광역시 남구 돋질로145번길 33 소재지에 있는 대성물류건설(주) 의 영업비밀 관리규정을 충분히 숙지, 이해하였으며 다음의 사항을 준수할 것을 엄숙히 서약합니다.

1. 대성물류건설(주) 의 영업비밀 관리 규정과 이에 관련한 명령을 성실히 이행하겠습니다.

2. 대성물류건설(주) 의 영업비밀은 재직 중은 물론 퇴직 후에도 회사의 허가 없이 사용하거나 제 3자에게 무단 누설하거나 경쟁회사에 유출하지 않겠습니다.

3. 본인 서인혁 이 알고 있는 제 3자의 영업비밀은 여하한 일이 있어도 비밀 보유자의 승낙 없이 회사에 공개하거나 회사의 업무에 부정하게 사용하지 않겠습니다.

4. 대성물류건설(주) 재직 시 지득한 영업비밀과 관련하여 경쟁회사에서는 이와 동일한 보안팀 업무를 담당하지 않겠습니다.

5. 재직 시는 물론 퇴직 후에도 대성물류건설(주) 재직 시 지득한 영업비밀을 가지고 창업을 하거나 경쟁 회사에 전직 또는 동업을 하지 않겠습니다.

6. 만약 이 서약서에 위반할 경우에는 부정경쟁방지법의 관련 규정과 대성물류건설(주) 의 영업비밀관리 규정에 의한 어떠한 조치도 감수하겠습니다.

2018 년 12 월 11 일

서약인 : 서인혁 (인)

대성물류건설(주) 귀하

기술보안 서약서

After

소속 : xxx 보안팀

직위 : xxx

성명 : xxx

본인 xxx 은 xxx xxx 돋질로xxxxxxxxx번길 xxx 소재지에 있는 xxx 의 영업비밀 관리규정을 충분히 숙지, 이해하였으며 다음의 사항을 준수할 것을 엄숙히 서약합니다.

xxx. xxx 의 영업비밀 관리 규정과 이에 관련한 명령을 성실히 이행하겠습니다.

xxx. xxx 의 영업비밀은 재직 중은 물론 퇴직 후에도 회사의 허가 없이 사용하거나 xxx xxx자에게 무단 누설하거나 경쟁회사에 유출하지 않겠습니다.

xxx. 본인 xxx 이 알고 있는 xxx xxx자의 영업비밀은 여하한 일이 있어도 비밀 보유자의 승낙 없이 회사에 공개하거나 회사의 업무에 부정하게 사용하지 않겠습니다.

xxx. xxx 재직 시 지득한 영업비밀과 관련하여 경쟁회사에서는 이와 동일한 보안팀 업무를 담당하지 않겠습니다.

xxx. 재직 시는 물론 퇴직 후에도 xxx 재직 시 지득한 영업비밀을 가지고 창업을 하거나 경쟁 회사에 전직 또는 동업을 하지 않겠습니다.

xxx. 만약 이 서약서에 위반할 경우에는 부정경쟁방지법의 관련 규정과 xxx 의 영업비밀관리 규정에 의한 어떠한 조치도 감수하겠습니다.

xxx0xxx8 년 xxxxxx 월 xxxxxx 일

서약인 : xxx (인)

xxx 귀하

결론

한계점 및 정리

- Learning continuous representation of words has a long history in NLP(*Rumelhart* et al., 1988)
- These represenation are derived from large unlabeled corpora using co-occurrence statistics (*Deerwester* et al., 1990; *Schutze*, 1992; *Lund and Burgess*, 1996)
- The distributional **semantics** has sudied the properties of these methods (modelling) (*Turney* et al., 2010; *Baroni and Lenci*, 2010)
- In neural network, word embedding(semantic model) (*Collobert and Weston*, 2008)
- The simple log-bilinear model efficiently (*Mikolov* et al, 2013b)

  => It is also **linear neural network** based model

  (It can decrease the training complexity)

  => based idea of the paper