안녕하십니까 비바프로 팀의 수행계획서를 발표하게 된 2015722057 조우진이라고 합니다.

저희가 수행하게 된 기업은 한화시스템이며, 주제는 인공지능 자연어 처리 기반 개체명 인식기 고도화 기술 개발 입니다.

목차는 개발동기 및 필요성, 과제 최종목표 및 개발내용, 과제해결 방법, 추진체계, 기대효과, 기타, 참고문헌 순으로 발표하겠습니다.

1 . 먼저 개발동기 및 필요성에서 개발동기를 살펴보겠습니다.

개체명 인식이 세계적으로 비약적인 발전을 하고 있지만, 한국어 개체명 인식은 그에 비해 많이 부진한 편입니다. 또한 개체명은 정보 검색에선 주요 검색 대상이며, 질의응답에선 주요 질의응답 대상으로 인식됩니다.
이런 과정에서 생기는 몇 가지 문제점들이 있는데, 번역이나 챗봇 응대 시스템 등에서 부정확한 경우가 많습니다. 또한 공식 문서 검색 시 키워드와 다르게 내용이 다르거나 필요하지 않은 결과가 출력되는 것을 볼 수 있습니다.
개체명의 인식 기술이 고도화 된다면 앞서 말씀드린 문제점들을 해결할 수 있고, 정보의 가독성 또한 증가할 것이며, 문서를 정리하기에 용이해지는 등 인간 생활의 편의성이 증대될 것이라 생각하여 개발하고자 생각했습니다.

이와 관련하여 국내의 지적재산권 현황을 조사한 결과 가장 유사한 “딥러닝을 이용한 한국어 개체명 인식 장치 및 방법” 이라는 것을 찾을 수 있었습니다.

전문을 읽어본 결과, bi-LSTM 을 사용하였고, 정확한 개체명 인식이 아닌 단어와 품사를 태그하는 방식으로 확인 됐습니다.

2. 다음으로 과제 최종 목표 및 개발내용에 대해 살펴보겠습니다.

Viva Pro 팀이 생각하는 최종 목표는 문장 내 더 다양한 개체명을 인식할 수 있게 태그 다양화와// 고도화가 기본 전제이므로 전년도 팀보다 더 높은 정확성을 최종 목표로 설정하였습니다.

개발내용으로는 가장 먼저 엑소브레인 말뭉치 및 기타 말뭉치 등 데이터 확보를 합니다. 그 다음으로 한글 형태소 분석기를 통해 토크나이징을 진행합니다.

다음 단계는 BERT 사용 시와 미사용 시로 나뉘어지게 되는데, 최종적으로는 분류, 임베딩, 태깅 및 훈련 등의 과정을 거치게 되고 그 중 효율성이 가장 높은 모델을 채택하게 됩니다.

마지막으로 성능 평가 및 목적 지향 대화 시스템에 사용될 것으로 보여집니다.

3. 다음으로 과제 해결 방법에 대해 살펴보겠습니다.

앞서 살펴본 내용을 좀 더 자세하게 보면 개체명 인식을 위해 클라이언트로 입력을 받은 뒤, 서버로 전해져 미캡, 코모란, 카히 등을 통해 토크나이징 합니다.

1단계에서 토크나이징 된 결과는 모델링 단계로 넘어가며, 사전학습 모델인 버트, 코버트, openAI GPT, 엘모 등의 모델을 사용합니다.

다음으로 파인 튜닝을 진행하여 데이터 셋을 이용해 태그를 추가하게 됩니다.

파인튜닝 방법을 다양하게 이용하여 진행 후, 성능 평가를 합니다.

최종적으로 클라우드와 amp를 사용하여 웹 서비스로 제공할 예정입니다.
가장 많이 사용되는 버트는 언어모델 중 하나로

다음으로 유명한 엘모는 문맥에 따라 단어를 다른 벡터로 표현해주는 워드 임베딩이 특징입니다.

고도화에 관한 개선점으로는 기존 KWBERT의 5개 태그에 7개 태그를 추가할 예정이고, 현재 데이터 셋을 준비하는 과정에 있습니다.

또한 띄어쓰기가 포함된 개체명의 경우 인식이 잘 안되는 오류가 있었지만 이를 해결하기 위해 BIO 태그법을 사용할 예정입니다. BIO 태그법은 개체명의 시작이 되는 단어에 B를, 중간 또는 끝에는 I, 그 외에는 O 태그를 붙이는 방법입니다. 즉 20세기 스튜디오를 예로 들자면 20세기에 OG_B, 스튜디오에 OG_I 를 붙여 데이터 셋을 만들 수 있습니다.

또 다른 오류로는 동음이의어에 대한 오류가 있는데 이런 단어에 대해서는 문맥에 따라 단어의 가중치가 달라지는 임베딩 단계에서 처리하는 기술을 활용할 예정입니다.

4. 다음으로 살펴볼 내용은 추진체계입니다.

역할 분담은 말뭉치 자료 확보와 언어모델에 관한 부분, 웹, 토크나이징 및 태깅, 신경망구축 등으로 나누었습니다. 일손이 많이 필요한 말뭉치는 모든 인원이 다 작업중이고, 그 외의 작업은 최대한 분담하여 진행할 계획입니다.

간트차트를 보시면, 말뭉치 자료 확보가 11월까지 예정되어 있으며, 현재 진행중인 상태입니다. 또한 전처리 과정 및 정제, 기존 프로젝트 및 관련기술 학습 또한 진행중에 있습니다.

5. 마지막으로 살펴볼 내용은 기대효과입니다.

여태 발표한 개체명 인식의 고도화를 성공적으로 이뤄낸다면 어떠한 분야에서 활용할 수 있을지.

자연어 처리가 필요한 의사결정 시스템, 질의응답 시스템, 정보검색 및 여러 응용 분야에서 활용가능할 것으로 보입니다.

이상으로 Viva Pro. 팀의 수행계획서 발표를 마치겠습니다.