요약

이번 연구는 개체인식의 (named entity recognition, NER) 일종인 중첩 개체 인식이라는 (nested NER) 다소 어려운 말뭉치 구축을 한 사례입니다. 구축시 공수가 많이 들어가고, 난이도도 높다보니 실제로 학계에서 쓰이는 말뭉치 손가락 안에 꼽을 정도로 적고, 언어 대응 또한 상당히 낮은 상황에 한국어에 먼저 지원을 할 수 있게 된 점에 대해서는 제안하신 세 분에게 학계와 산업계가 감사해야할 일로 생각됩니다.

종전 분류 방식의 제약사항을 극복하는 만큼 신규성이 높은 연구라고 생각됩니다. 세부적인 분류가 많아지면서, 이러한 다양한 태그 정보가 있어 기존 방식 대비 활용 가능성도 높아질 것으로 예상됩니다. 특히나 개체명에 대한 정확한 검출이 필요한 챗봇이나 knowledge base 같은 곳에 활용이 될 수 있는 방법론들이 많이 나올 수 있지 않을지 기대해봅니다.

마지막으로, 말뭉치 파일 형식을 표준 자료 포맷인 JSON, 그리고 표준은 아니지만 흔히 쓰이는 JSONL 형식으로 구축을 하는 새로운 경향에 대해서는 좋은 움직임으로 보입니다. 특히나 개인적으로 한국어 말뭉치의 해외 연구자 접근 제약, 그리고 말뭉치의 라이센스 문제는 항상 골치거리였는데 이번 연구에서는 이 또한 명쾌하게 공개를 해주신 점에서 대단히 높게 사고, 다른 말뭉치 과제에서도 본받을 필요가 있습니다.

질문

1. 해당 태그 분류는 이번 연구에서 새로이 제안하시는것으로 이해하면 될까요? 기존 GENIA/NNE/ACE 같은 말뭉치와의 호환성, 아울러 이 호환성을 통한 개선된 분류가 반대로 이 연구를 통해서 국제적으로 제안이 되어야 할지에 대해서 의견이 궁금합니다. 제안한 분류 체계는 종전 방식 대비 분류가 더 다양하게 가능한 만큼, 반대로 한국어 외에 다른 곳에도 활용이 되었으면 좋겠다는 생각이 듭니다.
2. 8번 슬라이드를 보면, 여러개의 개체명이 있고 중첩이 될 경우 넓은 범주에서 특정 개체명이 대표로 올라오는 현상을 볼 수 있습니다. 예를 들면 “노벨 경제학상 수상자”와 “노벨 경제학상 수상자 조지 에커로프”의 경우 각각 CV_POSITION과 PS_NAME이 가장 스팬이 넓어졌을때 특정 개체명으로 수렴하는 경향을 볼 수 있는데, 혹시 보였던 규칙 같은게 있었는지가 궁금합니다.
3. 일반적인 seq2seq 방식으로는 모델링이 아무래도 까다로운 태스크로 보입니다. 이에 대해서 베이스라인 모델을 혹시 준비할 계획이 있으신지가 궁금합니다.
4. 앞의 질문 두개가 합쳐진 것인데, 다른 각도로 seq2seq를 이용하여 모델링한 NER 모델에서 나온 단일 계층 개체명 태그를 이용해서 넓은 범주 태그를 출력하는 모델링 방식에 대해서 선행 연구가 있었는지가 궁금합니다.
5. 크게 의미가 있는 질문은 아닙니다만, 말뭉치 크기가 작아 JSONL까지 지원을 해야하는지에 대해서는 의문이 조금 듭니다. 한가지만 제공해도 충분하지 않을까요?
6.