Hyunjoong Kim

Research

주제	설명	상태
형태소 분석기	단어 추출 능력이 추가된 한국어 형태소 분석기	• Lattice based tagger trainer 구현중
Word vector Inference	D_0 word2vec 에 D_1 단어 벡터 추가	• 실험 부분 제외 초고 완료 • Yelp data 는 domain specific 표현과 다의어가 많아서 embedding 이 잘 되지 않음 (Wikipedia 로 교체중) • IMDb data 는 수집된 데이터에 토크나이저 문제가 존재하여 재수집함 (재수집 완료)
Word embedding for NER	Embedding word or subwords with CRF potential function as features	• Dissertation 의 챕터로 넣지 않기로 함으로써 중단

Dissertation

Chapter	Contents
1. Introduction	• Natural Language Processing 과정에서의 어려운 점 정리 (out of vocabulary, out of domain data, noise, ambiguity) • 이 문제들의 발생 원인과 그 영향력에 대한 내용
2. Noise canceling	• CRF 보다 안전한 띄어쓰기 오류 교정기 (RNN 계열 모델 실험 추가할 예정)
3. Enhancing Korean morphological analysis with word extraction	• 명사와 어미 추출기가 내제된 lattice based tagger 로, 새로운 도메인의 out of vocabulary, data 문제를 완화. • 현재 tagger 개발 중 • Word embedding inference 로 tagging features 를 확장할 예정

Dissertation

Chapter	Contents
4. Keyword extraction	• 데이터셋의 주제가 일관될 경우와 다양할 경우 접근법이 다름 • Homogeneous domain data with KR-WordRank • Heterogeneous domain data with document clustering and labeling
5. Topic detection & tracking	• 한 주제 (query) 에 대하여 문서가 발생된 각 시점 (document id) 별로 document representation 을 학습하면 시계열 토픽 변화가 학습 (예: 특정 정치인 이름이 포함된 뉴스 기사) • Topic segmentation & labeling with keyword extraction • 현재 실험용 데이터 (정치인 뉴스) 수집 중