의학용 영어 품사 태거 구현

이현구○, 안혁주, 김학수
강원대학교 컴퓨터정보통신공학과
nlphglee@kangwon.ac.kr, zingiskan12@kangwon.ac.kr, nlpdrkim@kangwon.ac.kr

Implementation of an English POS Tagger for Medical

Hyeon-Gu Lee○, HyeokJu Ahn, HarkSoo Kim
Kangwon National University Computer and Communication Engineering

요 약

자연어처리의 여러 분야에서 기본요소로 사용되는 영어 품사 태거를 UMLS의 의학용어 어휘정보와 OANC(Open American National Corpus) 말뭉치를 이용해 의학용 문서도 분석 가능한 의학용 영어 품사 태거를 제안한다. TRIE구조를 이용한 단어 묶음 모델로 여러 어절의 의학용어를 하나로 묶고 HMM(Hiden Markov Model)을 이용한 품사 태거로 해당하는 품사를 부착한다.

주제어: 영어 품사 태거, UMLS, TRIE구조, OANC

1. 서론

영어 품사 태거는 문장을 분석하는 가장 기본요소로 자연어처리의 관계 추출, 정보 검색, 질의응답 시스템 등 여러 분야에서 사용된다. 최근에는 단순한 문장이 아닌 의학용 문서를 분석하는 연구[1]가 진행되어 의학용 품사 태거가 필요하게 됐다. 그러나 기존 품사 태거는 품사를 부착하는 단위가 한 어절이지만 의학용어는 어절이 하나 이상인 경우가 많아 여러 어절의 의학용어를 하나로 묶는 작업이 필요하다. 본 논문에서는 TRIE 구조를 이용해 단어를 묶고 품사를 부착하는 의학용 영어 품사 태거를 제안한다. 먼저 2장에서 관련 연구에 대해 알아보고, 3장에서 TRIE 구조를 이용한 단어 묶음 모델과 그 결과를 이용한 영어 품사 태거를 제안하고 4장에서 결론을 맺는다.

2. 관련 연구

기존에 의학용 품사 태거로는 Genia 말뭉치[2]를 학습한 Genia Tagger[3]가 있다. 하지만 Genia Tagger는 여러 분야의 의학용어가 아닌 유전자 분야에 치중된 Genia 말뭉치를 학습하여 의학용 영어 품사 태거로는 부족함이 있다. 또한 여러 어절로 이루어진 의학용어가 각 어절별로 분석되어 오류가 발생 할 수 있다. 본 논문에서는 이러한 문제를 해결하기 위해 여러 분야의 의학용어로 이루어진 UMLS[4]의 어휘와 단어 묶음 모델을 통해 여러 어절의 의학용어를 묶어 오류를 해결한다.

3. 의학용 영어 품사 태거

제안 시스템은 [그림 1]과 같다. 의학정보 온톨로지 UMLS(Unified Medical Language System)에서 의학용어를 추출해 단어 묶음 사전을 구축한다. 구축된 사전을 이용해 단어 묶음 모델을 생성하고 문장이 입력되면 여러 어절의 의학용어를 하나로 묶은 후 영어 품사 태거를 통해 품사를 부착한다.

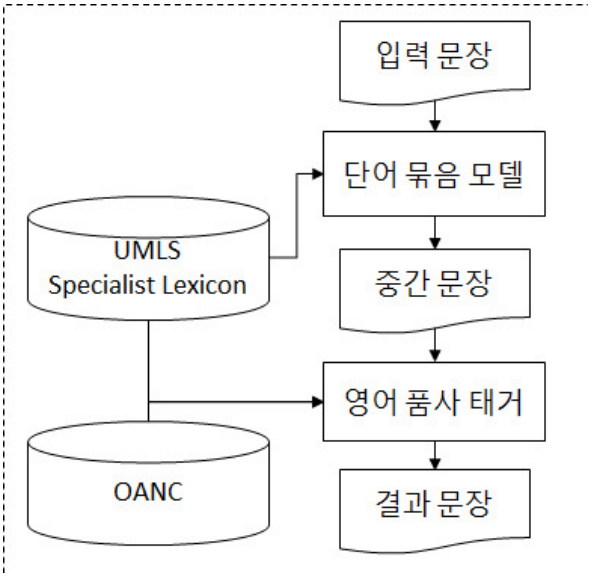

그림 1. 시스템 구성도

3.1 단어 묶음 모델

본 시스템에서는 TRIE 구조를 이용한 단어 묶음 모델을 제안한다. TRIE 구조는 한번의 검색으로 최장일치 단어를 찾아내 불필요한 사전 검색을 억제하고 사전의 메모리 크기를 축소 할 수 있는 자료구조이다.

제안 시스템에서는 TRIE 구조에 의학용어들을 입력하고 그 정보들을 이용해 입력된 문장에서 최장일치의 의학용어를 찾은 후 여러 어절의 단어인 경우 하나로 묶어준다. 여기서 사용되는 의학용어는 UMLS의 Specialist Lexicon에 있는 체언류 어휘들이다. [그림 2]는 단어 묶음 모델의 예시다.

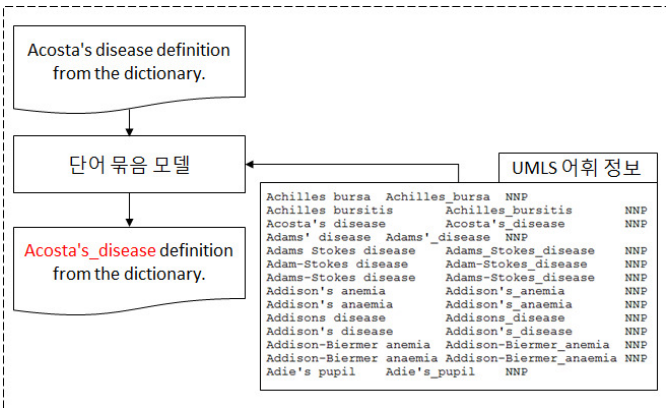

그림 2. 단어 묶음 모델의 예

[그림 2]의 예문 "Acosta's disease definition from the dictionary."을 단어 묶음 모델에 입력하면 TRIE 구조를 통해 "Acosta's disease"가 매칭되고 단어가 묶인 중간 문장이 결과로 나온다.

3.2 품사 태거

앞서 설명한 단어 묶음 모델을 바탕으로 입력 문장에서 의학용어가 발견될 경우를 고려하는 품사 태깅 모델이 필요하다. 따라서 본 논문에서는 기존에 많이 사용되는 HMM(Hidden Markov Model)을 이용한 품사 태거[5]를 이용한다. OANC(Open American national Corpus)[6]에 UMLS의 Specialist Lexicon단어를 추가하여 사전을 생성하고 단어 묶음 모델에서 묶인 의학용어를 선별해 분리한다. 그리고 HMM 모델에 필요한 관측(observation)확률과 전이(transition)확률을 계산한다. [그림 3]은 [그림 2]의 결과인 중간 문장에 품사를 부착 시 예시다.

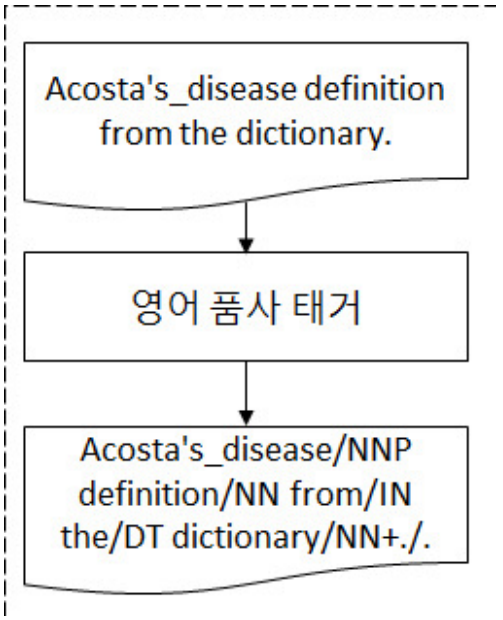

그림 3. 영어 품사 태거의 예

3.3 개선 방안

본 논문의 품사 태거는 단어 묶음 모델을 통해 묶인 의학용어들의 관측 확률이 최솟값으로 설정되는 문제가 있어 품사부착이 잘못되는 경우가 발생한다. 최솟값으로 설정되는 이유는 일반단어들과는 다르게 묶인 의학용어들은 말뭉치가 아닌 사전형태의 Specialist Lexicon단어를 추가한 것이라 빈도수를 셀 수가 없다. 따라서 [수식 1]을 이용하여 관측확률을 재설정한다.

$$P(Word) = P(Word_{head}) \qquad (1)$$

[수식 1]은 단어 묶음 모델을 통해 묶인 의학용어의 head단어를 구하고 head단어의 관측확률로 재설정한다. 예를들어 "Muenke syndrome"의 경우 "syndorme"이 head단어로 뽑히며 그에 해당하는 관측확률로 "Muenke syndrome"의 관측확률을 재설정한다.

4. 결론

본 논문에서는 UMLS의 Specialist Lexicon과 TRIE 구조를 이용한 단어 묶음 모델과 그 결과를 이용한 의학용 영어 품사 태거를 제안하였다. 제안 모델을 이용하여 의학용 문서를 효과적으로 분석할 수 있고 비슷하게 단어 묶음을 해야 하는 분야에서 효과적으로 사용 할 수 있을 것이다.

감사의 글

본 연구는 LG전자 산학연구용역 과제의 지원을 받아 수행되었음. 또한 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2013R1A1A4A01005074)

참고문헌

[1] Devarakonda, Murthy, and Ching-Huei Tsou. "Automated Problem List Generation from Electronic Medical Records in IBM Watson." Twenty-Seventh IAAI Conference, 2015.

[2] KIM, J.-D., et al. "GENIA corpus-a semantically annotated corpus for bio-textmining." Bioinformatics, 19.suppl 1: i180-i182, 2003.

[3] Tateisi, Yuka, and Jun'ichi Tsujii. "Part-of-Speech Annotation of Biology Research Abstracts." LREC, 2004.

[4] Unified Medical Language System, "http://www.nlm.nih.gov/research/umls/"

[5] 안혁주, 최맹식, 김학수, "말뭉치 기반 영어 품사 태거 구현", 2013 제7회 한국정보과학회, 한국 정보처리학회 공동학술심포지움, pp.73-75, 2013.

[6] Open American National Corpus, "http://www.anc.org/" .