한국어 언어자원에서의 자연어 처리 기술 현황 조사[1]

한만휘O 박성찬 이한빛 연종흠 이상구
서울대학교 컴퓨터공학부
{hanmanhui, baksalchan, hanbit2222, jonghm, sglee}@europa.snu.ac.kr

Natural language processing on Korean language: A survey

Manhui Han, Sungchan Park, Hanbit Lee, Jongheum Yeon, Sang-goo Lee
Department of Computer Science and Engineering, Seuol National University

요 약

다양한 언어 자원이 생성된 이후 한국어 처리 기술에 대한 많은 연구가 진행되고 있다. 이에 한국어에 대한 자연어처리 기술 각각의 성능을 조사해 보고 현재 가장 연구가 많이 진행된 영어와 비교해 본다. 전체적으로 대부분의 기술의 경우 영어와 비슷한 수준에 도달하거나 더 좋은 성능을 보였지만 몇몇 분야에서는 일관된 데이터 셋이나 지표의 부재로 인해 정확한 성능 비교가 불가능하였다.

1. 서 론

자연어 처리를 위해 한국어를 대상으로 하는 언어자원이 많이 존재하고 있다. 1998년부터 2007년까지 정부가 국가 차원의 한국어 디지털 언어자원 구축을 목표로 진행한 "21세기 세종 계획"을 통해 현재 한국어 언어자원 중 가장 큰 규모인 세종 말뭉치가 구축되었다. [2] 세종 말뭉치 외에도 부산대학교의 KorLex나 울산대의 U-WIN, 한국과학기술원(KAIST)의 CoreNet과 같은 어휘 의미망(WordNet)이 구축되어 있다. 이와 같은 한국어 언어자원들을 활용한 자연어 처리 기술이 활발히 연구되고 있다. 이러한 자연어 처리 기술들의 한국어에 대한 기술 현황을 가장 연구가 많이 진행된 영어에서의 자연어 처리 기술 수준과 비교해 본다. 본 논문의 나머지 부분은 다름과 같이 구성된다. 2절에서는 자연어 처리 기술 중 기본적 기술로 평가되는 품사 부탁과 구문 분석 기술 수준을 조사하고 3절에서는 응용 기술들인 교정, 중의성 해소, 상호 참조, 개체명 인식, 요약 기술들에 대해 조사해 본다. 4장에서는 결론으로 한국어 처리 기술 수준에 대해 전체적으로 요약해 보고 현재 문제점들에 대해 기술한다.

2. 기본 자연어 처리 기술

1) 품사 부착 (POS tagging)

품사 부착은 자연어 처리 기술 중 가장 기본이 되는 기술이다. 교착어에 해당되는 한국어는 굴절어인 영어와는 달리 품사 부착을 형태소 분석 후에 수행하는 것으로 일반화 되어 있다. 하지만 최근에 형태소 분석을 하지 않고 품사 부착을 수행하는 방안에 대한 연구가 진행되고 있다. 현재 조사된 방법들에 대한 비교는 표 1과 같다.

표 1. 한국어 품사 부착 성능 비교

방법	발표연도	정확도(AIR)
Lattice 기반[1]	2014	*91.27%
**단계별 전이모델[2]	2012	*96.44%
음절 단위 N-gram[3]	2013	98.9%
음절 단위 확률 모델[4]	2014	98.4%

(*) 서로 비교 방법이 다를 수도 있음
(**) 동형이의어 까지도 태깅함

영어의 경우 현재 97.32% 품사 부착 정확도를 보이고 있고 알려지지 않은 단어에 대한 품사 부착의 경우에도 90.79%의 성능을 보이고 있다.[5] 이를 통해 한국어 품사 부착의 경우, 영어와 비슷한 기술수준에 도달하였음을 알 수 있다. 다만 비교를 위한 데이터 셋이나 논문들 간에 공통으로 쓰일 수 있는 공신력 있는 지표가 없다는 점은 아쉬운 점이다.

2) 구문 분석 (parsing)

한국어는 어순이 자유롭고, 주어를 포함한 필수 논항의 생략이 빈번히 일어나는 특징으로 인해 구문 분석의 난이도가 높다. 현재까지 진행된 국내 구문 분석 연구 현황은 표 2와 같다.

표 2. 한국어 구문 분석 성능 비교

방법	발표 연도	F1-Measure	문장 정확도
스탠포드 구문분석기[6]	2012	74.65%	N/A
버클리 구문분석기[6]	2012	78.74%	N/A
키어절[7]	2013	N/A	45.91%
PNU-KLParser 2.0[8]	2014	N/A	*100%

(*) 복수개의 답안을 출력하므로 다른 방법과 비교 불가

[1] 본 연구는 미래창조과학부 (MSIP) 의 출연과 한국연구재단 (NRF) 의 지원으로 수행되었음. (No. 20110030812)
[2] http://www.sejong.or.kr

대부분의 연구가 세종 구문 분석 말뭉치를 사용하였지만 논문 상에서 다른 연구들과의 비교 실험이 없거나 서로 다른 지표를 사용하여 정확한 성능 비교가 불가능하였다.

영어의 경우 PARSEVAL이라는 지표로 정확도를 나타내며 현재 92.1%를 기록하고 있다[9]. 이와 비교해 볼 때 아직 한국어 구문 분석은 아직 연구가 더 필요한 부분으로 볼 수 있다.

3. 응용 자연어 처리 기술

1) 교 정 (spelling)

한국어 교정 기술의 경우 크게 띄어쓰기 교정과 철자 교정 기술로 나누어진다. 한국어 띄어쓰기 연구의 경우 주로 말뭉치 기반의 통계기반 방식으로 진행되었다. 각 연구의 비교는 표3과 같다. 음절 단위의 띄어쓰기는 거의 완벽하지만 복합 어절의 존재로 인해 어절 단위 정확도는 최대 98.39%에 머물러 있음을 볼 수 있다.

표 3. 한국어 띄어쓰기 교정 성능 비교

사용 모델	발표 연도	음절 단위 정확도	어절 단위 정확도
CRF[10]	2011	*98.84%	*95.99%
Structural-SVM[11]	2013	99.01%	95.47%
Modified Viterbi Search[12]	2014	99.64%	98.39%

(*) 평가 데이터 셋이 다름

한국어 철자 교정 연구의 경우 한국어가 가지는 교착어의 특성으로 인해 N-gram 방식의 접근이 어렵다. 따라서 한국어의 철자 교정은 rule-based 방식으로 교정어휘 쌍을 사용하여 접근되고 있다. 현재 한국어 철자 교정은 부산대에서 개발한 Koran Spelling and Grammar Checker(KSGC)[3] 라는 툴이 많이 쓰이고 해당 툴에 기반한 연구도 진행되고 있다. 해당 툴은 각각 95.19%와 37.56%의 정확도와 재현율을 기록하고 있다[13]. 규칙 기반 방식이기 때문에 정확도가 높은 것을 볼 수 있고 현재 추가로 재현율을 더 상승시키기 위한 연구들이 진행되었다[14, 15].

영어 철자 교정의 경우 'Web 1T 3-gram' 데이터 셋이 발표되면서 3-gram 방식의 연구가 진행되었고 각각 76.3%와 38.1%의 재현율과 정확도를 기록하고 있다[16]. 규칙기반 접근의 경우 한국과 마찬가지로 교정어휘 쌍을 사용하는 방법이 사용되고 있으며 현재 95% 내외의 정확도를 보이고 있다[17].

2) 중의성 해소 (Word Sense Disambiguation, WSD)

한국어 중의성 해소 연구는 주로 소규모의 의미 태그 부착 말뭉치나 사전 정보 등을 이용하여 엔트로피 정보, 조건부 확률, 상호 정보 등을 사용하여 다양하게 진행되었다. 한국어 중의성 해소 기술 수준은 표 4와 같다.

표 4. 한국어 중의성 해소 성능 비교

방법	발표연도	정확도
어휘 의미망 사용[18]	2011	*86.2%
단어공간모델[19]	2012	94.02%

(*) 평가를 위해 SENSEVAL-2 한국어 학습 데이터 셋 이용

중의성 해소 연구의 경우 다른 논문들과 비교할 수 있는 SENSEVAL-2 한국어 학습 데이터 셋이 존재하였다. SENSEVAL은 ACL SIGLEX와 EURALEX의 후원 하에 개최된 중의성 해소 기술 평가 대회로 2번째 대회에 한국어가 포함되어 SENSEVAL-2 한국어 학습 데이터에는 눈, 손, 말 등 중의성이 있는 어휘에 대해 각각 수십 개의 데이터가 포함되어 있다. 영어의 경우 Semeval-2007 데이터 셋에 대해 의미 추출 수준에 따라 coarse-grain에 대해서는 82.50%, fine-grain에 대해서는 59.1%의 성능을 보였다[20].

3) 상호 참조 (Coreference Resolution)

한국어 상호 참조 문제에 대해서는 최근 전체 명사에 대해 Stanford의 multi-pass sieve 시스템을 한국어에 적용해 문제를 해결한 연구가 발표되었다[21]. Mean(CoNLL) 지표를 사용하여 60.65%를 기록하였고, 영어의 경우 2011년 공식 기록이 59.3%인 것으로 보아 한국어 상호 참조 문제는 영어와 비슷한 수준에 도달한 것으로 볼 수 있다.

4) 개체명 인식 (Named Entity Recognition, NER)

한국어 개체명 인식의 경우, 다양한 도메인에서 다양한 모델을 사용하여 연구가 진행되었다[22, 23]. 문제 자체가 도메인 별로 서로 다른 성능을 내기 때문에 통합하여 비교하기는 어렵지만 대체로 80% 중후반 대의 성능(F1-Measure)을 보이고 있다. 영어의 경우 93.39%의 성능을 보인다.

5) 요 약 (Summarization)

한국어 요약과 관련된 연구는 주로 뉴스 기사의 주요 문장을 추출하는 방향으로 진행되었다. 진행된 연구로는 TextRank 알고리즘을 사용한 연구[24]가 있으며 ROUGE 평가 프로그램을 이용해 영어와 비교했을 때 더 좋은 성능을 보였다.

4. 결 론

전체적으로 위에서 서술한 대부분의 자연어 처리 기술들이 영어와 비슷한 수준이거나 더 좋은 성능을 보이는 것을 볼 수 있었다. 다만 구문 분석의 경우에는

[3] KSGC: http://speller.cs.pusan.ac.kr

자연어 처리의 기본적인 기술임에도 불구하고 영어와 비교해 떨어지는 성능을 보였고, 개체명 인식의 경우에도 도메인을 아우를 수 있는 데이터 셋과 지표의 부재로 인해 정확한 성능 비교 평가가 불가능하였다.

참 고 문 헌

[1] 나승훈, 김창현, 김영길. “래티스상의 구조적 분류에 기반한 한국어 형태소 분석 및 품사 태깅,” 정보과학회논문지 : 소프트웨어 및 응용 제 41 권 제 6 호, pp. 523-532, 2014

[2] 신준철, 옥철영. "한국어 품사 및 동형이의어 태깅을 위한 단계별 전이모델," 정보과학회논문지 : 소프트웨어 및 응용 제 39 권 제 11 호, pp. 889-901, 2012

[3] 심광섭. "음절 단위의 한국어 품사 태깅에서의 원형 복원," 정보과학회논문지 : 소프트웨어 및 응용 제 40 권 제 3 호, pp. 182-189, 2013

[4] 심광섭. "한국어 형태소 분석을 위한 음절 단위 확률 모델," 정보과학회논문지 제 41 권 제 9 호, pp. 642-651, 2014

[5] Manning, Christopher D. "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?," Proceedings of the 12th international conference on Computational linguistics and intelligent text processing, Part I, pp. 171-189, 2011

[6] 최동현, 박정열, 임경태, 함영균, 최기선. "구문 분석기 성능 향상을 위한 세종 트리뱅크 변환 방법," 한국정보과학회 2012 한국컴퓨터종합학술대회 논문집 제 39 권 제 1 호(B), pp. 342-344, 2012

[7] 오진영, 차정원. "키어절을 이용한 새로운 한국어 구문분석," 정보과학회논문지 : 소프트웨어 및 응용 제 40 권 제 10 호, pp. 600-608, 2013

[8] 남웅, 윤애선, 권혁철. "무제한 한국어 의존 구문분석기 'PNU-KLParser 2.0'의 개발," 정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 20 권 제 6 호, pp. 354-358, 2014

[9] McClosky, D., Charnik, E., Johnson, M. "Effective self-training for parsing," Proceedings of HLT/NAACL, pp. 152-159, 2006

[10] 심광섭. "CRF 를 이용한 한국어 자동 띄어쓰기," 인지과학 제 22 권 제 2 호, pp. 217-233, 2011

[11] Changki Lee, Hyunki Kim. "Automatic Korean word spacing using Pegasos algorithm," Information Processing & Management, vol. 49, issue 1, pp. 370-379, 2013

[12] 이창기. "사용자가 입력한 띄어쓰기 정보를 이용한 Structural SVM 기반 한국어 띄어쓰기," 정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 20 권 제 5 호, 2014

[13] 최현수, 윤애선, 권혁철. "통합적 방식을 이용한 한국어 문맥의존 철자오류 교정규칙의 재현율 향상," 한국정보과학회 2014 한국컴퓨터종합학술대회 논문집, pp. 577-579, 2014

[14] 서한영, 최성기, 권혁철. "한국어 어휘 의미망을 이용한 통계적 문맥 의존 철자오류 교정 성능 향상," 한국정보과학회 2014 한국컴퓨터종합학술대회 논문집, pp. 607-609, 2014

[15] 최현수, 윤애선, 권혁철. "조사제약 조건의 완화에 의한 문맥의존 철자오류 교정의 재현율 향상 방식," 정보과학회논문지 : 소프트웨어 및 응용 제 41 권 제 3 호, pp. 249-256, 2014

[16] Islam, Aminul, Diana Inkpen. "Real-word spelling correction using Google Web 1T 3-grams", Proceeding of International Conference on Natural Language Processing and Knowledge Engineering, vol. 3, pp. 1241-1249, 2009

[17] Golding, Andrew R., Dan Roth, J. Mooney, Claire Cardie, "A window-based approach to context-sensitive spelling correction," Machine Learning, vol. 34, pp. 107-130, 1999

[18] 김민호, 권혁철, "한국어 어휘의미망의 의미 관계를 이용한 어의 중의성 해소," 정보과학회논문지 : 소프트웨어 및 응용 제 38 권 제 10 호, pp. 554-564, 2011

[19] 박용민, 이재성, "한국어 단어 공간 모델을 이용한 단어 의미 중의성 해소." 한국콘텐츠학회논문지 제 12 권 제 6 호, pp. 41-47, 2012

[20] Roberto Navigli. "Word sense disambiguation: A survey", ACM Computing Surveys (CSUR), vol. 41 issue 2, 2009

[21] 박천음, 최경호, 이창기, "Multi-pass Sieve 를 이용한 한국어 상호참조해결," 정보과학회논문지 제 41 권 제 11 호, pp. 992-1005, 2014

[22] 이창기, 장명길, "Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식," 인지과학 제 21 권 제 4 호, pp. 655-667, 2010

[23] 김성원, 나동렬, "2 단계 최대 엔트로피 모델을 이용한 한국어 개체명 인식," 2008 한국정보과학회 학술 심포지움 논문집 제 2 권 제 1 호, pp. 81-86, 2008

[24] 홍진표, 차정원, "TextRank 알고리즘을 이용한 한국어 중요 문장 추출," 한국정보과학회 2009 한국컴퓨터종합학술대회 논문집 제 36 권 제 1 호(C), pp. 311-314, 2009