연구보고서 2016-006

지능정보기술 동향 조사

Survey on the trends of Intelligence-information technology

- 자연어 처리를 중심으로
Focused on Natural language processing

안성원/김태호/송지환/추형석 외

2017.04.

SPRi 소프트웨어정책연구소

연구기관 및 참여연구원 : 소프트웨어정책연구소
안성원
김태호
송지환
추형석

창원대학교
차정원
박은실
박태호
박다솔
신창욱
박성재
신영태

이 보고서는 2016년도 미래창조과학부 정보통신진흥기금을 지원받아 수행한 연구결과로 보고서 내용은 연구자의 견해이며, 미래창조과학부의 공식입장과 다를 수 있습니다.

목 차

제1장 서 론 ····· 1

제2장 자연어처리 기술 ····· 3

제1절 자연어처리 개요 ····· 3

제2절 형태소 분석 및 형태소 품사 부착 ····· 3

제3절 함의(Entailment) ····· 5

제4절 질의 응답(Question Answering System) ····· 7

제5절 정보 추출(Information Extraction) ····· 8

제6절 구문 분석 ····· 9

제7절 상호 참조 ····· 10

제8절 개체명 ····· 11

제9절 의미역 ····· 12

제10절 감성 분석 ····· 14

제3장 기술 동향 ····· 15

제1절 학술 대회 ····· 15

1. ICON ····· 15

2. ACL ····· 16

3. 한글 및 한국어 정보처리 학술대회 ····· 19

제2절 경진 대회 ····· 24
1. NLP tools contest ····· 24
2. EVALITA ····· 26
3. TAC ····· 40
4. 국어 정보 처리 시스템 경진대회 ····· 43

제4장 언어지능 기술이 적용된 실제 사례 조사 ····· 45
제1절 Virtual Assistance ····· 45
제2절 Social Network analytics tools ····· 46
1. Keyhole ····· 47
2. Signal Media ····· 49
3. Brandwatch ····· 50

제5장 언어처리 기술 수준 ····· 52
제1절 공개된 SW의 종류와 성능 ····· 52
1. CoreNLP ····· 52
2. Natural Language Toolkit ····· 53
3. KoNLPy ····· 57
4. Google SyntaxNet ····· 58
5. Apache OpenNLP ····· 59
6. Apache Lucene ····· 61

7. UIMA ····· 61
8. MALLET ····· 62
9. ESPRESSO ····· 63
10. 꼬꼬마 ····· 64
11. 코모란 ····· 65
12. 동아대학교 띄어쓰기 모델 ····· 65
제2절 연구소, 대학, 기업의 연구 수준 ····· 66
1. Stanford ····· 66
2. Berkeley group ····· 67
3. MIT ····· 68
4. Sheffield ····· 69
5. Univ. of Washington ····· 69
6. MS Human language technologies ····· 70
7. IBM NLP group ····· 71
8. 다이퀘스트 ····· 72
9. ETRI(한국전자통신연구원) ····· 73

제6장 지식데이터베이스 현황 조사 ····· 74
제1절 한국어 지식 데이터베이스 ····· 74
1. 서울대학교 ····· 74
2. 언어정보연구원 ····· 75

3. 국어 국립원 ····· 76
4. ETRI ····· 77
5. 한국과학기술정보연구원과 충남대학교 ····· 78
6. 울산대학교 ····· 79
7. 고려대학교 ····· 79
8. 펜실베니아 대학교 ····· 80
9. LDC ····· 80
10. kAIST ····· 81
11. 언어자원은행 ····· 81
제2절 해외 지식 데이터 베이스 ····· 83
1. 영국 코퍼스 ····· 83
2. 미국 코퍼스 ····· 84
3. 체코 코퍼스 ····· 84
4. 헝가리 코퍼스 ····· 85
5. 러시아 코퍼스 ····· 85
6. 이탈리아 코퍼스 ····· 86
7. 그리스 코퍼스 ····· 86
8. 독일 코퍼스 ····· 87
9. 중국 코퍼스 ····· 87
제3절 모니터 코퍼스 ····· 88
1. The Bank of English(BoE) 코퍼스 ····· 88

2. Global English Monitor 코퍼스 ····· 88
제4절 Brown 계열 코퍼스 ····· 88
제5절 공시적 코퍼스 ····· 89
1. The International Corpus of English(ICE) ····· 90
2. The Longman/Lancaster Corpus ····· 90
3. The Longman Written American Corpus ····· 91
4. The CREA corpus ····· 91
5. The LIVAC corpus ····· 92
제6절 통시적 코퍼스 ····· 92
1. The Helsinki corpus ····· 92
2. The ARCHER corpus ····· 93
3. The Corpus of Early English Correspondence ····· 93
4. The Zurich English Newspaper Corpus(ZEN) ····· 94
5. 영어대화 코퍼스 ····· 94
제7절 음성 코퍼스 ····· 94
1. The London-Lund Corpus ····· 94
2. 10대 언어의 Bergen corpus ····· 95
3. The Spoken corpus of the Survey of English Dialect ····· 95
4. The Intonational Variation in English Corpus ····· 95
5. The HongKong Corpus of Conversational English(HKCCE) ····· 96
제8절 학업 및 전문영어 코퍼스 ····· 97

1. The Michigan Corpus of Academic Spoken English ····· 97
2. The British Academic Spoken English Corpus(BASEC) ····· 98
3. The Academic Corpus ····· 98
4. The Corpus of Professional English(CPE) ····· 99
제9절 구문 분석 코퍼스 ····· 99
1. The Lancaster-Leeds Treebank ····· 100
2. The Lancaster Parsed Corpus ····· 100
3. The SUSANNE corpus ····· 100
4. The Penn Treebank ····· 100
제10절 Developmental and learner corpus ····· 101
1. The Child Language Data Exchange System ····· 101
2. The Louvain Corpus of Native English Essays ····· 101
3. The Polytechnic of Wales corpus ····· 102
4. The Cambridge Learner Corpus ····· 102
5. The HKUST Corpus of Learner English ····· 102
6. The Japanese EFL Learner corpus ····· 102
7. The Standard Speaking Test Corpus ····· 103
제11절 다국어 코퍼스 ····· 103
1. The English-Norwegian parallel Corpus ····· 103
2. The English-Swedish Parallel Corpus ····· 103
3. The Oslo Multilingual Corpus ····· 104

4. The IJS-ELAN Slovene-English Parallel Corpus ····· 104
5. The CLUVI parallel corpus ····· 104
6. The PAROLE corpus ····· 105
7. The EMILLE corpus ····· 105
8. Hong Kong parallel Text ····· 106
제12절 영어 이외의 단일 언어 코퍼스 ····· 106
1. The COSMAS corpus ····· 106
2. The CETEMPúblico Corpus ····· 107
3. The INL corpus ····· 107
4. The CEG corpus ····· 107
5. The Scottish Corpus of Texts and Speech ····· 108
6. Academia Sinica Balanced Corpus ····· 108
7. Spoken Chinese Corpus of Situated Discourse ····· 109

제7장 결론 및 시사점 ····· 110

참고문헌 ····· 111

CONTENTS

Chapter 1. Introduction ····· 1

Chapter 2. Natural language processing technology ····· 3

Part 1. Natural language processing overview ····· 3

Part 2. Morphological analysis and morphological parts ····· 3

Part 3. Entailment ····· 5

Part 4. Question Answering System ····· 7

Part 5. Information Extraction ····· 8

Part 6. Syntatic Analysis ····· 9

Part 7. Co-reference ····· 10

Part 8. Named Entity ····· 11

Part 9. Semantic Role ····· 12

Part 10. Emotional Analysis ····· 14

Chapter 3. Technology trends ····· 15

Part 1. Conference ····· 15

Part 2. Contest ····· 24

Chapter 4. Practical case study with language intelligence technology ····· 45

Part 1. Virtual Assistance ····· 45

Part 2. Social Network analytics tools ····· 46

Chapter 5. Language processing technology level ····· 52

Part 1. Types and performance of open SW ····· 52

Part 2. Research level of research institutes, universities and companies ··· 66

Chapter 6. Knowledge database status survey ····· 74

Part 1. Korean knowledge database ····· 74

Part 2. Global knowledge database ····· 83

Part 3. Monitor corpus ····· 88

Part 4. Brown Corpus ····· 88

Part 5. Disclosed Corpus ····· 89

Part 6. Transitive Corpus ····· 92

Part 7. Voice Corpus ····· 94

Part 8. Academic and professional English corpus ····· 97

Part 9. Parsing Corpus ····· 99

Part 10. Developmental and learner corpus ····· 101

Part 11. Multilingual Corpus ····· 103

Part 12. Non-English single language corpus ····· 106

Chapter 7. Conclusion and implications ····· 110

Reference ····· 111

표 목 차

<표 2-1> 프레임 예 ····· 8

<표 3-1> 28회 한글 및 한국어 정보처리 학술대회 우수논문 ····· 24

<표 3-2> POS Tagset ····· 26

<표 3-3> 속성값을 부착하지 않는 경우 예시 ····· 31

<표 3-4> FactA 주석 부착 예시 ····· 31

<표 3-5> NEEL-IT의 분류체계 ····· 34

<표 3-6> PoSTWITA의 데이터 예 ····· 35

<표 3-7> PoSTWITA 태그 셋 ····· 36

<표 3-8> SENTIPOLC 데이터 정보 ····· 39

<표 4-1> Keyhole에서 확인할 수 있는 정보 ····· 48

<표 5-1> CoreNLP 개요 ····· 52

<표 5-2> CoreNLP의 지원 언어 내역 ····· 52

<표 5-3> CoreNLP의 모듈별 성능과 레퍼런스 ····· 53

<표 5-4> NLTK 개요 ····· 53

<표 5-5> NLTK의 Dependency Parser 모듈 ····· 54

<표 5-6> NLTK의 Sentiment Analysis 모듈 ····· 55

<표 5-7> KoNLPy 개요 ····· 57

<표 5-8> KoNLPy의 모듈 ····· 57

<표 5-9> Google SyntaxNet ····· 58

<표 5-10> Google SyntaxNet의 성능 ····· 58

<표 5-11> Apache OpenNLP 개요 ····· 59

<표 5-12> Apache OpenNLP의 모듈 ····· 60
<표 5-13> Apache Lucene ····· 61
<표 5-14> UIMA 개요 ····· 61
<표 5-15> MALLET 개요 ····· 63
<표 5-16> MALLET의 모듈 ····· 63
<표 5-17> ESPRESSO 개요 ····· 64
<표 5-18> ESPRESSO의 모듈과 알고리즘 ····· 64
<표 5-19> 꼬꼬마 개요 ····· 64
<표 5-20> 꼬꼬마 형태소 분석기의 성능 ····· 65
<표 5-21> 코모란 개요 ····· 65
<표 5-22> 동아대학교 띄어쓰기 모델 ····· 66
<표 5-23> MS에서 제공하는 API들 ····· 71
<표 5-24> IBM에서 제공하는 API들 ····· 72
<표 5-25> ETRI에서 제공하는 기술이전 일부 ····· 73
<표 6-1> 연세 코퍼스 구성 ····· 75
<표 6-2> 전자사전 구성 ····· 76
<표 6-3> 문어 코퍼스 양 ····· 77
<표 6-4> 구어 코퍼스 양 ····· 77
<표 6-5> 배포 중인 데이터베이스 목록 ····· 78
<표 6-6> 음성 데이터 정보 ····· 81
<표 6-7> 전자사전 정보 ····· 82
<표 6-8> 코퍼스 정보 ····· 82
<표 6-9> 코어넷 정보 ····· 82

〈표 6-10〉 영국 텍스트 코퍼스 구성 ····· 83
〈표 6-11〉 영국 음성 코퍼스 구성 ····· 83
〈표 6-12〉 미국 코퍼스의 텍스트 카테고리 ····· 84
〈표 6-13〉 체코 코퍼스 공시적 언어의 구성 ····· 84
〈표 6-14〉 체코 코퍼스 통시적 언어의 구성 ····· 85
〈표 6-15〉 헝가리 코퍼스 카테고리 ····· 85
〈표 6-16〉 러시아 코퍼스 카테고리 ····· 86
〈표 6-17〉 CORIS 코퍼스의 구성 ····· 86
〈표 6-18〉 독일 코퍼스의 구성 ····· 87
〈표 6-19〉 중국어 MCLC 코퍼스의 구성 ····· 87
〈표 6-20〉 Brown 계열 코퍼스의 구성 ····· 89
〈표 6-21〉 ICE 음성 데이터의 구성 ····· 90
〈표 6-22〉 Helsinki Diachronic Corpus의 기간별 정보 ····· 92
〈표 6-23〉 ZEN 데이터의 구성 ····· 94
〈표 6-24〉 Bergen Corpus의 나이 그룹 ····· 95
〈표 6-25〉 HKCCE의 설계 기준 ····· 96
〈표 6-26〉 MICASE corpus의 정보 ····· 98
〈표 6-27〉 BASEC corpus의 정보 ····· 98
〈표 6-28〉 The Academic corpus의 세부 구성 ····· 99
〈표 6-29〉 Penn Treebank 1의 구성 요소 ····· 101
〈표 6-30〉 Oslo Multilingual 코퍼스의 정보 ····· 104
〈표 6-31〉 CLUVI 코퍼스의 정보 ····· 105
〈표 6-32〉 Academia Sinica Balanced 코퍼스의 정보 ····· 108

Contents of Table

<Table 2-1> Frame example ····· 8

<Table 3-1> 28th Excellent conference on korean Information Processing ····· 24

<Table 3-2> POS Tagset ····· 26

<Table 3-3> Example of not attaching attribute values ····· 31

<Table 3-4> FatcA annotation example ····· 31

<Table 3-5> NEEL-IT's classification system ····· 34

<Table 3-6> PoSTWITA data example ····· 35

<Table 3-7> PoSTWITA tag set ····· 36

<Table 3-8> SENTIPOLC data information ····· 39

<Table 4-1> Information that can find on Keyhole ····· 48

<Table 5-1> CoreNLP overview ····· 52

<Table 5-2> Supported languages of CoreNLP ····· 52

<Table 5-3> CoreNLP module performance and reference ····· 53

<Table 5-4> NLTK overview ····· 53

<Table 5-5> Dependency Parser module of NLTK ····· 54

<Table 5-6> Sentiment Analysis module of NLTK ····· 55

<Table 5-7> KoNLPy overview ····· 57

<Table 5-8> KoNLPy module ····· 57

<Table 5-9> Google SyntaxNet ····· 58

<Table 5-10> Performance of Google SyntaxNet ····· 58

<Table 5-11> Apache OpenNLP overview ····· 59

<Table 5-12> Apache OpenNLP module ····· 60
<Table 5-13> Apache Lucene ····· 61
<Table 5-14> UIMA overview ····· 61
<Table 5-15> MALLET overview ····· 63
<Table 5-16> MALLET module ····· 63
<Table 5-17> ESPRESSO overview ····· 64
<Table 5-18> ESPRESSO module and algorithm ····· 64
<Table 5-19> GGoGGoMa overview ····· 64
<Table 5-20> Performance of GGoGGoMa morpheme ····· 65
<Table 5-21> KOMORAN Overview ····· 65
<Table 5-22> Dong-A university spacing model ····· 66
<Table 5-23> APIs provided by MS ····· 71
<Table 5-24> APIs provided by IBM ····· 72
<Table 5-25> Some technology transfer from ETR ····· 73
<Table 6-1> Organization of Yonsei Corpus ····· 75
<Table 6-2> Electronic dictionary configuration ····· 76
<Table 6-3> Amount of literary corpus ····· 77
<Table 6-4> Amount of colloquial corpus ····· 77
<Table 6-5> List of databases being deployed ····· 78
<Table 6-6> Voice data information ····· 81
<Table 6-7> Electronic dictionary information ····· 82
<Table 6-8> Corpus Information ····· 82
<Table 6-9> CoreNet information ····· 82

<Table 6-10> Composition of English text corpus ····· 83

<Table 6-11> British speech corpus composition ····· 83

<Table 6-12> US Corpus Text Category ····· 84

<Table 6-13> Composition of the Czech corpus disclosure language ····· 84

<Table 6-14> Composition of the Czech corpus language ····· 85

<Table 6-15> Hungarian corpus category ····· 85

<Table 6-16> Russian corpus category ····· 86

<Table 6-17> Composition of CORIS corpus ····· 86

<Table 6-18> Composition of the German corpus ····· 87

<Table 6-19> Composition of Chinese MCLC corpus ····· 87

<Table 6-20> Composition of brown corpus ····· 89

<Table 6-21> Composition of ICE voice data ····· 90

<Table 6-22> Periodic information from Helsinki Diachronic Corpus ····· 92

<Table 6-23> Configuration of ZEN data ····· 94

<Table 6-24> Bergen Corpus Age Group ····· 95

<Table 6-25> Design criteria of HKCCE ····· 96

<Table 6-26> Information about MICASE corpus ····· 98

<Table 6-27> Information on BASEC corpus ····· 98

<Table 6-28> Details of The Academic corpus ····· 99

<Table 6-29> Components of Penn Treebank 1 ····· 101

<Table 6-30> Oslo Multilingual Corpus Information ····· 104

<Table 6-31> CLUVI Corpus Information ····· 105

<Table 6-32> Information of Academia Sinica Balanced Corpus ····· 108

그 림 목 차

[그림 2-1] 영어권 구문 분석의 예 ····· 10

[그림 2-2] 상호 참조 예 ····· 11

[그림 2-3] 한국어 의미격 인식 결과 ····· 13

[그림 3-1] FactA의 이벤트 인식 및 주석 방법 예시 ····· 28

[그림 4-1] Keyhole의 분석 예 ····· 48

[그림 4-2] Brandwatch의 분석 예 ····· 51

[그림 5-1] NLTK parse tree display의 예 ····· 56

[그림 5-2] UIMA의 구조 ····· 62

[그림 6-1] IDG - “글에서 감정을 읽다” 표지 ····· 74

Contents of Figure

[Figure 2-1] Examples of English-language parsing ····· 10

[Figure 2-2] Cross-reference example ····· 11

[Figure 2-3] Korean semantic recognition result ····· 13

[Figure 3-1] FatcA's example of event recognition and annotation ····· 28

[Figure 4-1] Example of keyhole analysis ····· 48

[Figure 4-2] Example of Brandwatch analysis ····· 51

[Figure 5-1] Example of NLTK parse tree display ····· 56

[Figure 5-2] Structure of UIMA ····· 62

[Figure 6-1] IDG - "Read emotion in the article" Cover ····· 74

요 약 문

<table>
<tr><td>과제명</td><td colspan="2">지능정보기술 동향 조사 – 자연어 처리를 중심으로</td></tr>
<tr><td>연구 목적</td><td colspan="2">최근 자연어처리 분야 연구 동향 및 SW/DB 조사·분석</td></tr>
<tr><td>연구 필요성</td><td colspan="2">최근 여러 연구기관에서 다양한 자연어처리 세부분야의 연구를 진행</td></tr>
<tr><td rowspan="5">지능정보기술 동향 조사 요약</td><td>1. 자연어처리</td><td>- 자연어처리의 기본 개념과 하위 분류 조사</td></tr>
<tr><td>2. 학술대회 및 경진대회</td><td>- 학술대회, 경진대회별 역사와 범위, 특징 조사</td></tr>
<tr><td>3. 실제 사례</td><td>- 언어처리 기술이 적용된 실제 사례 조사
- 사례별 역사와 특징 나열</td></tr>
<tr><td>4. 언어처리 기술 수준</td><td>- 공개된 SW의 성능과 지원 모듈 조사</td></tr>
<tr><td>5. 지식 데이터 베이스 현황조사</td><td>- 공개된 국내외 지식 데이터베이스 현황 조사
- 각 지식데이터베이스의 문서량과 종류 조사</td></tr>
<tr><td>연구 성과</td><td colspan="2">- 학술대회와 경진대회의 특징과 역사 조사 성과물
- 언어처리 기술이 적용된 실제사례 조사 성과물
- 각 기관별 언어처리 기술 수준 조사 성과물
- 공개된 데이터베이스 조사 및 정리 성과물</td></tr>
<tr><td>기대효과</td><td colspan="2">국내·외 학회 또는 경진대회의 특징 및 분야를 참고하면 앞으로의 동향을 읽는 데에 도움이 될 것이다. 또한 언어처리 기술이 적용된 실제 사례를 참고해 공개된 데이터베이스를 활용하면 새로운 연구를 진행하는 데 좋은 기반이 될 수 있을 것으로 예상된다.</td></tr>
</table>

SUMMARY

Title	Survey on the trends of Intelligence-information technology - Focused on Natural language processing	
Purpose	Recent Research Trends in Natural Language Processing and Research of SW/DB	
Need	In recent years, various research institutes have been conducting research on various fields of natural language processing.	
Summary of Intelligent Information Technology Trend	1. NLP (Natural Language Processing)	Basic concept and sub-classification of NLP
	2. Related academic conferences and contests	History, scope and characteristics of domestic and international academic conferences and contests
	3. Real case study	survey on NLP Real Case case History and List features
	4. NLP technology level	Survey of NLP open source's performance and spec
	5. Survey on Knowledge Database for NLP	Survey on domestic and international open Knowledge Database for NLP Survey on characteristic of each Knowledge Database
Research Results	- Survey on domestic and international academic conferences and contests - Real case study on NLP technology - Research on NLP technology level for each application - Categorization on open knowledge databases for NLP	
Expectations	his report will help understand trends for NLP with its details for academic conferences and content. Additionally, it will contribute to the future works by using the real case study and the free knowledge database for NLP.	

제1장 서 론

지능정보기술은 제4차산업혁명 시대의 도래와 함께 핵심기술로 부상하였다. 지능정보기술은 지능을 뜻하는 인공지능과 정보를 뜻하는 사물인터넷(IoT), 클라우드(Cloud), 빅데이터(BigData), 모바일(Mobile)과 같은 기술들이 융합된 형태를 의미한다.

이런 지능정보기술의 발전된 모습 중 하나인 언어지능은 인간의 언어를 컴퓨터가 이해하여 다양한 서비스를 제공하기 위한 기반기술이자, 인간과 컴퓨터 간의 상호작용(HCI: Human-computer Interaction), 즉, 보다 편리한 인터페이스를 제공하기 위한 역할을 한다. 여기에는 음성인식이나 문자 및 구문인식과 같은 기술 등이 있고, 결국 인간이 사용하는 자연어(또는 자연언어 : Natural Language)를 얼마나 빠르고 정확하게 분석하는가가 중요한 이슈이다.

지구상에는 약 6900여 가지의 언어가 존재 한다[1]. 물론 대부분의 언어는 극소수만이 사용되는 언어이고, 전 세계 인구의 99%에 이르는 수가 사용하는 언어는 1284개 이다. 이렇듯 다양한 언어를 컴퓨터가 이해하고 활용하기 위해서는 각 언어별로, 정해진 문법과 어휘, 어절, 문맥상의 의미, 그리고 이들 간의 다양한 조합에 따라 생성되는 또 다른 의미들에 대한 데이터베이스를 보유하고, 학습을 통한 최적화된 자연어처리 알고리즘이 필요하다.

최근에는 실시간 통역 애플리케이션(Application)이나 완성도가 높아진 번역기 프로그램들이 등장 하면서 자연어처리를 통한 언어지능은 더욱 주목을 받고 있다. 자연어처리 기술이 발전하면 앞으로는 타 국가의 언어를 공부하는데 시간을 들이거나, 이해하기 위해 애쓰지 않아도 될 것이다. 또한, 컴퓨터가 인간과 완벽한 대화가 가능해질 날이 머지않을 것이다.

현재 세계는 언어지능을 비롯한 지능정보기술에 대한 개발 경쟁과 원천기술 선점을 위한 각축전을 벌이고 있다. 이러한 시점에서 국·내외 자연어처리 기술에 대한 동향을 살펴보고 적절한 대응책을 모색해 나가는 것이 필요하다.

이 보고서는 지능정보기술의 자연어 처리 분야를 중심으로 관련 기술과 기술수준, 사례 등의 동향을 정리하여 제공하는 것을 목적으로 한다. 이를 위해

1) 참조 : 언어다양성 보존활용 센터, Ethnologue language of the world
http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

자연어 처리 분야의 기본 개념을 정리하고 언어 처리 분야에서 전 세계의 분야별 동향을 조사하였다. 이 보고서는 향후 지능정보기술 발전을 위한 정책방향을 도출하는데 참고자료로 활용할 수 있다.

보고서의 구성은 다음과 같다. 2장에서는 자연어처리 기술의 정의와 그 하위분류에 있는 다양한 자연어 처리 기술의 내용을 기술한다. 3장에서는 자연어처리 분야에서 진행되고 있는 국내·외 학술 대회와 자연어처리 기술경진을 통해 기술 수준의 발전을 도모하는 각국의 경진 대회들을 소개한다. 4장에서는 현재 서비스되고 있는 다양한 자연어처리 기술이 접목된 실 사례에 대해 살펴본다. 5장에서는 언어처리 기술 수준에서는 지금까지 공개된 Software들과 각각의 성능을 조사하였고, 연구소나 대학, 기업의 연구 수준에 대해 함께 조사를 진행하였다. 6장에서는 국내뿐만 아니라 해외 각국의 자연어처리를 위한 다양한 지식데이터베이스 수준을 조사하였다.

제2장 자연어처리 기술

제1절 자연어처리의 개요

자연어처리(NLP : Natural Language Processing)는 인간의 언어를 컴퓨터가 이해할 수 있도록 다양한 분석 방법을 통해 기계적인 형태로 변환하는 기술을 의미한다. 또한 이를 다시 인간이 해석할 수 있는 형태소 만드는 기술도 포함한다. 인공지능의 하위 분야로 1960년대의 인공지능을 만들려던 시도가 실패한 후에 인간의 언어를 분석하고 이해하는 기술이 세분화되면서 파생된 학문 분야이다. 언어공학, 인공지능, 전산언어학의 연구 분야이다.

자연어(Natural Language)는 프로그래밍 언어처럼 사람이 인공적으로 만든 언어가 아닌 과거에 오랜 시간을 거쳐 자연스럽게 발생한 의사소통을 위해 사용해 온 한국어나 영어 같은 언어를 의미한다. 일반적으로 컴퓨터 공학에서 언어라고 하면 C나 JAVA와 같은 프로그래밍 언어를 떠오르기 때문에 자연어를 구분지어 부르고 있다.

2016년 현재 우리가 상상하는 인공지능의 모습처럼 컴퓨터가 인간의 언어를 완벽하게 이해하지는 못하고 있다. 하지만 많은 연구를 통해 활률과 통계 기법을 사용하여 수준 높은 자연어 처리 기술들이 실생활 깊숙이에 활용되고 있다. 구글의 검색 엔진이나 Apple의 Siri같은 시스템이 그 예이다. 자연어처리 기술은 데이터 소스에서 다양한 정보 추출과 의미 분석을 통해 인공 지능을 위한 꾸준한 연구가 계속되고 있다.

이러한 자연어처리에는 많은 하위 연구 분류가 있다. 문서를 분석하는 기술에는 형태소 분석 및 품사 부착, 구문 분석, 개체명 인식 및 분류, 의미역 결정, 상호 참조, 관계 정보 추출 등이 있다. 이러한 문서 분석 정보를 활용하여 대화 모델이나 정보 검색, 질의응답 등의 기술이 서비스되고 있다.[1][2)]

제2절 형태소 분석 및 형태소 품사 부착

형태소는 POS(Part-Of-Speech)라고 하며, 의미가 있는 최소의 단위를 말하며

2) 참고문헌 [1] 참조, 이하 내용상의 인용에 대해 참고문헌 번호[ ]로 표기

문법적, 관계적인 뜻을 나타내는 단어 또는 단어의 부분이라고 정의된다. 즉, 언어를 분석하기 위한 기본 단위로 의미를 가지는 요소로서 더 이상 분석할 수 없는 가장 작은 문법 단위이다.

자연어 처리에서의 형태소 분석은 형태소 분석을 하는 대상 어절에서 분석 가능한 모든 형태소를 분리하여 분석하는 것을 의미한다. 영어의 경우 한 어절이 하나의 단어로 이루어지기 때문에 한 단어가 하나의 형태소가 된다. 하지만 한국어의 특성상 하나의 어절에는 하나 이상의 형태소가 존재한다. 따라서 한 어절에서 나타나는 모든 형태소를 분리하여 분석하는 과정이 필요하다. 이러한 과정은 명사나 대명사 같은 '의미형태소'와 조사나 어미 등의 '기능형태소'를 분리하는 작업이나 복합명사를 분리하는 작업에 필요하다.

다음은 형태소를 분리하는 작업의 예이다.

```
철수는 바나나우유를 마셨다.
철수 + 는  + 바나나우유 + 를 + 마시 + 었 + 다 + .
```

위 예시 문장에서 '철수는'이라는 어절은 고유명사인 '철수'와 보조사인 '는'으로 분리되었다. '바나나우유를'이라는 어절은 '바나나우유'와 '를'로 분리되었는데, '바나나우유'의 경우는 복합 명사를 분리하여 형태소 분석을 할 경우, '바나나'와 '우유'로 분리되어 분석된다.

POS tagging(Part-Of-Speech Tagging)은 형태소 분석이다. 형태소 분석은 자연어 처리의 첫 단계로 단어 또는 어절 단위로 분리된 입력 문자열로 부터 각각의 형태소를 분리하고 용언의 불규칙 활용이나 굴절 현상이 일어난 단어에 대해 원형을 복원하는 과정으로 기술된다.

형태소 분석을 통해 나온 결과 중 가장 적합한 형태의 품사를 부착하는 것을 말한다. 보통 태거라고 하는 모듈이 이 기능을 수행한다. 이는 형태소 분석기가 출력한 다양한 분석 결과 중에서 문맥에 적합한 하나의 분석 결과를 선택하는 모듈이라 할 수 있다. 분석 시 문맥 좌우에 위치한 중의성 해소의 힌트가 되는 정보를 이용해서 적합한 분석 결과를 선택한다. 보통 태거는 대규모의 품사부착 코퍼스를 이용해서 구현하는데 은닉 마르코프 모델(HMM)[3]이 널리 사용되고

3) 은닉 마르코프 모델(HMM)은 통계적 마르코프 모델 중 하나로 은닉된 상태와 관찰되는 상태 정보를 통해 실제 상태를 예측하는 시스템이다.

있다.

형태소 분석의 일반적 과정은 아래와 같다.

① 전처리 단계 → ② 분석 후보 생성 → ③ 결합 제약 검사 → ④ 옳은 후보 선택 → ⑤ 후처리 단계

전처리 단계에서는 문장 부호 분리 및 단어 추출하고 특수 문자열이나 숫자를 처리하는 단계이다. 분석 후보 생성 단계는 형태소 분리 및 불규칙 원형 복원을 하고, 모음조화와 형태소 결합 제약(음운 형상에 따른 제약 등)을 검사하는 결합 제약 검사 단계를 거친다. 사전 탐색 및 단어 형성 규칙을 통해 옳은 후보를 선택하고 마지막으로 복합명사 추정, 사전 미등록어 · 준말 처리를 하는 후처리 단계를 거친다.

제3절 함의(Entailment)

함의(Entailment)는 문장과 문장 사이의 의미 관계를 나타내는 용어이다. 어떤 문장 S1에서 다른 문장 S2를 사실적 지식에 의존하지 않고 오직 논리형식이나 언어 의미에 의해 논리적으로 추론할 수 있을 때 즉, S1이 참일 때 반드시 S2가 참이 되는 경우 ‘S1은 S2를 함의 한다’ 라고 한다.

함의는 의미론적 함의와 사실적 함의로 구분되고, 의미론적 함의는 다시 어휘 의미에서의 함의와 통사구조에서의 함의로 나눌 수 있다. 어휘 의미에서의 함의는 상의어와 하의어의 관계에 놓이는 단어 자체에서 함의 관계를 형성한다. 이 외에도 조사나 어미 등에 의해서도 함의는 일어날 수 있다.

1. 어휘 의미에서의 함의

먼저 어휘 의미에서의 함의의 예를 살펴보자.

* T : 창수는 총각이다. ⇒ H : 창수는 인간이다. * T : 창수는 총각이다. ⇒ H : 창수는 남성이다. * T : 창수는 총각이다. ⇒ H : 창수는 성인이다. * T : 창수는 총각이다. ⇒ H : 창수는 미혼이다.

이 문장들에서 ‘총각’이라는 단어의 의미는 (+인간)(+남성)(+성인)(+미혼) 총 4가지 의미를 모두 포함하고 있기 때문에 각각의 T 문장은 H 문장을 함의한다. 비슷한 예로,

* T : 영아는 사과를 먹었다. ⇒ H : 영아는 과일을 먹었다.

‘과일’은 ‘사과’의 상의어이기 때문에 T 문장은 H 문장을 함의한다.

2. 통사구조에서의 함의

한편, 통사구조에서의 함의는 태의 변화, 역의 반의어, 연접 구성에 의한 함의를 말한다. 아래의 예를 살펴보자.

* T : 성호가 수미의 남편이다. ⇒ H : 수미는 성호의 아내이다.

역의 반의어는 서로 대체될 수 있는 두 단어 사이의 관계를 말하고, ‘남편’과 ‘아내’는 역의 반의어 관계이기 때문에 T 문장은 H 문장을 함의한다. 또 다른 예를 들어보면,

* T : 바둑이가 온 집안을 난장판으로 만들었다. ⇒ H : 온 집안이 바둑이에 의해 난장판이 되었다.

두 문장이 능동과 피동의 관계 즉, 태의 변화이기 때문에 T 문장은 H 문장을 함의한다.

3. 사실적 함의

사실적 함의는 언어 내부의 논리적 관계가 아닌 세상지식에 의하여 생성되는 관계를 말한다. 아래의 예에 해당된다.

* T : 음주 단속 결과 혈중 알코올 농도가 0.1%를 넘었다. ⇒ H : 운전면허가 취소된다.

이 문장에서의 세상의 지식 이란 음주 운전에 대한 법규를 의미한다.

제4절 질의응답(Question Answering System)

질의응답 시스템(Question Answering System, QA)은 자연어로 된 사용자의 질의에 대한 답을 찾아서 제공하는 서비스이다. 사용자가 입력한 키워드와 연관되는 정보를 랭킹 순서에 따라 리스트 형식으로 제공해주는 일반적인 검색 서비스와 다르게, QA 서비스는 사용자에게 단답형으로 대응한다.

사용자는 자연어로 질문을 전달하고 QA 서비스는 사용자의 자연어 질문을 분석한 후 여러 단계의 프로세스를 통하여 답을 도출한다. 찾아진 답은 다시 자연어 형태로 변환되어 사용자에게 제공된다.

QA 시스템은 3가지 유형의 데이터를 사용하여 사용자에게 적절한 답변을 제공한다.

1. 정형 데이터를 사용하는 QA 시스템

사용자의 질의를 받은 후 미리 정의된 규칙에 일치하는 서비스를 사용자에게 제공한다. 또한 질문에 대한 사용자의 요청에 대해서 미리 구축해 놓은 정형 데이터에 대한 쿼리를 통해 답변을 제공한다. 이를 수행하지 못하는 경우 일반 검색 엔진을 이용하여 얻어진 검색 결과를 제공한다.

2. 비정형 데이터를 사용하는 QA 시스템

사용자의 질의를 분석한 후 검색을 통해 관련된 문서들 중에서 답변을 추출하여 제공하는 방법이 있다. 검색을 통해 결과로 얻어진 문서에서 몇 가지 기준을 통해 스코어를 계산할 수 있는 알고리즘이 적용되고 이를 통하여 최종 답변이 도출된다.

3. 정형 데이터와 비정형 데이터를 모두 이용하는 QA 시스템

비정형 데이터로부터 정형화된 데이터를 추출하여 구축하는 시스템이다. 비정형 데이터에 대해 여러 검색 엔진을 이용하여 인덱스를 설정한 후 사용자의 질의와 관련이 높은 데이터를 찾아내는 데 사용한다.

제5절 정보 추출(Information Extraction)

정보 추출(Information Extraction)은 미리 사용되지 않은 텍스트를 입력으로 받아서 구조화된 형태의 출력으로 결과를 내는 기술을 말한다. 사용자가 원하는 정보를 검색하거나 추출하는 연구 분야로서 대용량의 정보 자료들로부터 미리 정의된 주제나 관심 분야의 정보만을 인식하여 요약된 형태로 가공하는 것이다. 즉, 정보 추출 시스템은 자연어 처리 기술을 바탕으로 문서의 내용을 분석하여 주어진 주제 영역에서 유용한 정보들을 구조화된 형식으로 저장한다.

> **예시)** 삼성전자가 2017년 1월 5일부터 8일까지(현지시간) 미국 라스베이거스에서 열리는 가전전시회 'CES 2017'에서 이전 제품과 비교해 한층 얇은 두께의 2017년형 '파워봇' 신제품을 선보인다. <IT 동아, 2016년 12월 28일 기사>

예시와 같은 문장의 키워드는 아래 <표 2-1>과 같은 프레임 형식으로 표현할 수 있다.

<표 2-1> 프레임 예

슬롯	슬롯 값
제작사	삼성전자
연월일	2017년 1월 5일부터 8일까지
지역	미국 라스베이거스
제품	파워봇

정보 추출의 방법에는 여러 가지이다. 물론 문장을 구문·의미 해석하여 그 결과를 바탕으로 프레임의 슬롯을 메울 수도 있지만, 대부분의 경우는 패턴 매칭을 이용한다. 위의 슬롯과 같이 공통으로 슬롯을 지정해놓은 형식을 알고 있다면 패턴 매칭만으로도 간단히 정보 추출을 할 수 있다. “<제작사>는 <제품>의 신제품을 선보인다.“라는 문장의 패턴으로 제작사와 제품의 슬롯 값을 찾을 수 있다는 것이다.

제6절 구문 분석

하나의 문장은 보통 여러 개의 단어로 구성되어 있고, 각각의 단어들 사이에는 그 단어가 문장 내에서 맡은 역할에 따른 연결 관계가 존재한다. 따라서 자연어 처리를 위해서 문장을 구성하는 구성성분과 그 성분들 간의 관계를 분석할 필요성이 있다. 이 과정을 "구조분석"이라고 한다. 문장의 의미를 가지는 최소 단위로 분리하고(형태소 분석), 이들 간의 연관 관계 및 문장 내에서의 역할을 파악한다(구문 분석).

구문 분석(Syntatic Analysis)는 하나의 문장 즉, 일련의 문자열을 형태소 단위로 분리하고 해당 구문론적 관계를 도식화하여 명확히 하는 분석 절차이다. 그리고 형태소 분석의 결과를 기반으로 문장을 이루고 있는 명사구, 동사구, 부사구 등의 구문을 묶어주는 것 뿐만 아니라 주어, 술어, 목적어 등과 같은 주요한 문장 구성성분을 밝혀내고 그들 사이의 구문관계를 분석하여 문장의 문법적 구조를 결정하는 기술이다.

구문 분석은 다음과 같은 과정을 거쳐서 분석된다.

Tagging Result ➜ Apply Grammar ➜ Parsing ➜ Parsing Tree ➜ Choose Best Tree

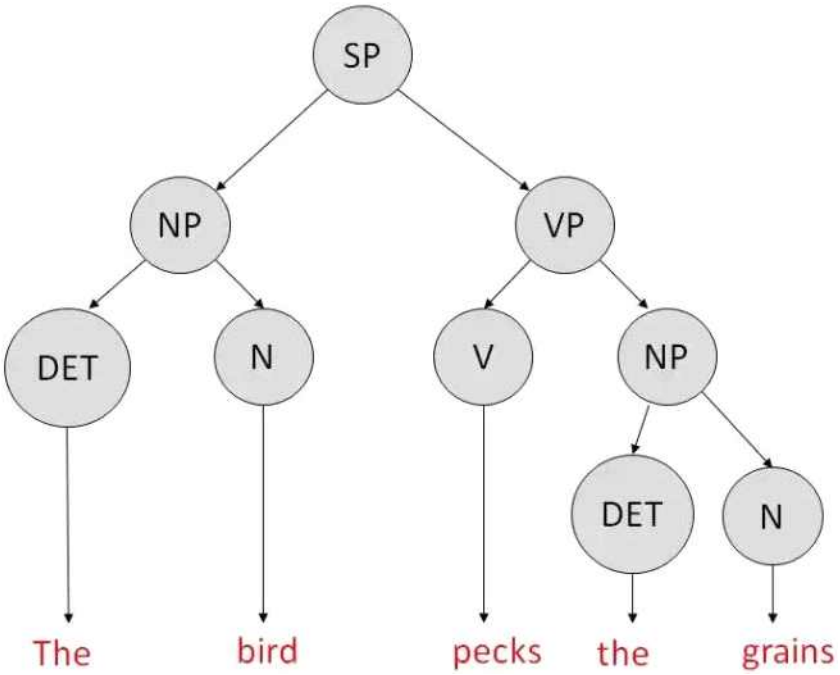
일반적으로 문맥 자유 문법(Context Free Grammar)를 사용하여 구문 분석 후 여러 개의 Parsing Tree를 반환한다. 문맥 자유 문법은 다음과 같은 요소로 이루어져있다.

G = <NT, T, P, S>
- Start Sysmol(S) : Sentence
- Non-Terminal(NT) : Syntactic constituents
- Terminals(T) : Lexical entities/words
- Productions(P) ⊆ NT× (NT∪T)+ ≡ grammar rules

구문 분석 방법은 일반적으로 2가지이다. 하향식(Top Down) 구문 분석과 상향식(Bottom Up) 구문분석으로 나눌 수 있는데, 하향식 구문 분석은 문장을 좌측에서 우측으로 한 문자씩 읽으며 [그림 2-1]과 같이 뿌리 노드로부터 시작하여 단말 노드를 생성하는 방식이다. 문법 규칙을 좌변에서 우변으로 바꿔 쓰기 규칙으로 간주하며 개시기호(S) 로부터 시작해서 입력문(Bottom)을 만들어

가는 알고리즘이다.

반대로 상향식 구문 분석은 문법규칙을 우변으로부터 좌변으로의 바꿔 쓰기 규칙이라고 간주하며, 입력문에서 시작하여 개시기호를 만들어 가는 방법이다. 즉, 트리를 단말노드부터 뿌리 노드를 향하여 위쪽 방향으로 파스트리를 생성해나가는 방식이다.



※이미지출처 : http://www.w3ii.com/ko/artificial_intelligence/artificial_intelligence_natural_language_processing.html

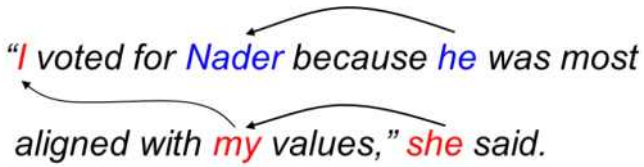
[그림 2-1] 영어권 구문 분석의 예

제7절 상호 참조

상호참조(Coreference)는 문서 내에서 선행하는 명사구와 현재 등장한 명사구 간에 같은 개체를 의미하는 지를 결정하는 문제이다. 상호참조(Coreference)는 같은 개체(entity)를 나타내는 명사구들이 종종 다른 형태와 어휘로 나타나기도 한다. 앞서 등장한 명사구가 이후에 재표현될 때 대명사의 형태로 나타나거나 축약되기도 한다. 문서 내에 같은 개체를 의미하지만 단어가 변형되거나 혹은 대명사 형태와 같이 축약된 명사구들로 나타나는 것을 참조라고 한다.

임의의 개체(entity)에 대하여 다른 표현으로 사용되는 단어들을 찾아, 서로

같은 개체로 연결해주는 자연어처리 문제이다. 하나의 개체를 다른 단어로 표현하는 경우는 별명, 약어, 대명사, 한정사구 등이 있으며, 이들 간의 참조 관계를 올바르게 찾아낼 수 있으면 담화나 문서 내에서 언급하는 대상에 대한 정보를 일관성 있게 유지할 수 있고, 정확하게 전달할 수 있다.



[그림 2-2] 상호 참조 예[3]

[그림 2-2]에서 볼 수 있듯이 'I'와 'my', 'she'는 모두 같은 대상을 가리키고, 'Nader'와 'he'는 같은 대상을 가리킨다. 이처럼 문장 내에서 서로 다른 형태로 나타났으나 같은 대상을 가리키는 단어들의 관계를 생성하는 것이 상호 참조이다.

제8절 개체명

개체명(Named Entity)이란 문서에서 특정한 의미를 가지고 있는 단어 또는 어구를 말한다. 정보 검색에서 개체명은 주요 검색 대상이 된다. 이러한 개체명을 추출하기 위해 자연어 처리 분야에서 개체명 인식 및 분류 (Named Entity Recognition and Classification) 연구가 활발하게 진행되었다.

한국어 텍스트에서 개체명은 대부분 인명, 지명, 기관명 등과 같은 고유명사로 구성된다. 개체명 인식은 정보검색, 정보추출, 질의응답 등에 있어서 중요한 기술이고, 특히 질의응답 시스템의 성능에 큰 역할을 한다. 개체명 인식 기술은 단순히 사전 기반으로 구현될 수도 있지만 도메인 변경 시에 추가적인 사전 구축비용과 구축시간이 필요하고, 신조어에 대해서도 성능이 떨어지는 문제가 있다. 그러므로 개체명 태깅 코퍼스로부터 분류 모델을 학습하는 기계학습 기반 개체명 인식 기술이 최근에는 주로 사용되고 있다. 기계학습 기반 개체명 인식기를 개발하기 위해서는 수작업으로 정답 개체명이 태깅된 태깅 코퍼스가 반드시 필요하다. 다음은 개체명 인식의 한 예이다.

<박찬호:PS> 선수는 <한화 이글스:ORG>와 계약을 연장했다.

위의 예문에서 ‘박찬호’는 인명의 개채명을 띄고, ‘한화 이글스’는 단체의 개체명을 띈다. 이처럼 문장에서 찾을 수 있는 개체명을 인식하고, 해당 개체명에 알맞은 태그를 부착하는 것이 개체명 인식 및 분류이다. [4]

제9절 의미역

의미역(Semantic Role)은 문장 내의 서술어와 관계있는 명사구의 의미 역할을 말하고, 의미역이 부여된 명사구를 논항(argument)이라고 한다. 의미역 결정(Semantic Role Labeling)은 문장 내의 서술어를 인식하여 서술어의 의미와 명사구들의 역할을 부여하는 자연어처리의 한 분야이다. 명사구들의 역할은 서술어와 의미 관계에 따라 “행위주”, “피동작주”, “착점” 등의 역할을 부여한다. 의미역 결정은 구문 분석과 유사하지만 문장 구조가 바뀌어도 의미가 변하지 않는 특성 때문에 어려움이 따른다.

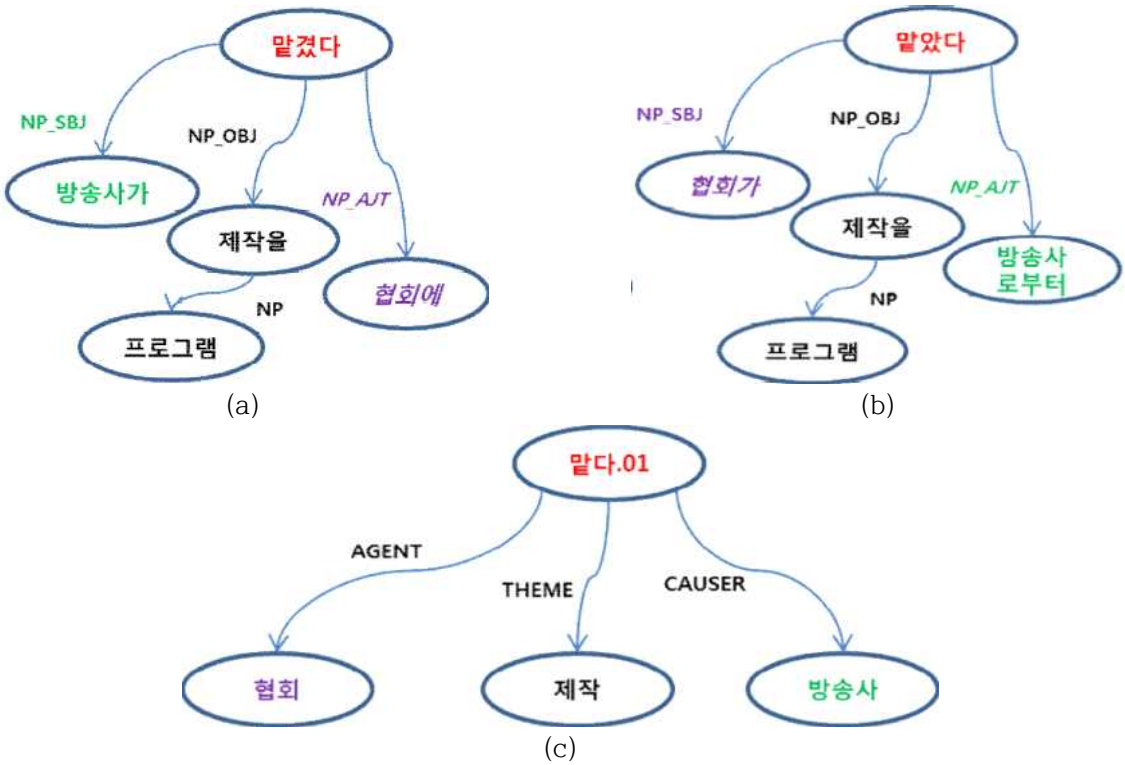
(가) 해커는 서버를 공격했다. (나) 서버는 해커에게 공격받았다.

(가)의 문장에서 해커는 주어이며 ‘공격하다’라는 서술어의 ‘행위주’이다. (나)의 문장에서 서버는 주어이며, ‘공격받다’라는 서술어의 ‘피동작주’이고, ‘행위주’는 서버로 동일하다. 이처럼 문장 구조가 바뀜에 따라 구문 정보는 변하지만 의미역은 변하지 않고, 문장 구조처럼 겉으로 들어나는 정보가 아닌 문장이 지닌 의미를 이해해야하는 작업이기 때문에 어려움이 따른다.

한국어 의미역 결정을 위해서 격 정보를 활용하는 방법과 기계학습을 활용하는 방법이 있다. 격 정보는 서술어에 따라 사용되는 격 정보를 파악하여 해당 격에 위치하는 명사구에 적합한 역할을 부여하는 방식이다. 이 방법은 오류가 적게 발생하나 모든 서술어에 대해 격이 존재하지 않기 때문에 의미역을 부착하지 못하는 경우가 발생한다. 기계학습을 활용한 방법은 CRFs(Conditional Random Fields)나 SVM(Support Vector Machine) 등의 알고리즘을 이용하여 의미역 결정 모델을 생성하여 자동으로 의미역을 결정하는 방법이다.

기계학습을 활용할 시, 격 정보가 없이도 학습에 사용한 자질에 따라 의미역을 결정할 수 있는 장점이 있으나 오류가 다수 발생할 수 있다는 단점이 있다.

대용량의 텍스트를 최소한의 비용으로 처리하기 위해서는 일반적으로 기계학습 방식을 사용한다. 현재 기계학습을 이용한 방법의 성능은 영어권에서 88%의 성능을 보이고 한국어에서는 75% 정도의 성능을 보이고 있다. [5] 다음 [그림 2-3]은 구문 분서과 의미역 결정의 차이를 보이는 그림이다.



※이미지출처 : 임수종, 김현기, “순차적 레이블링을 이용한 한국어 의미격 인식”, 한국컴퓨터종합학술대회 논문집, 2014.

[그림 2-3] 한국어 의미격 인식 결과

구문 분석은 문장에서 단어의 순서가 바뀌면 각 단어가 지니는 구문 역할이 바뀐다. 예를 들어 [그림 2-3]에서 (a)의 문장은 ‘방송사가 프로그램 제작을 협회에 맡겼다.’이고 (b)의 문장은 ‘협회가 방송사로부터 프로그램 제작을 맡았다.’이다. (a)에서 주어(NP_SBJ)는 ‘방송사가’이다. 하지만 (b)의 주어는 ‘협회가’이다. 하지만 사실상 두 문장의 의미는 동일하다. 따라서 두 문장에서 의미역은 구문의 기능들처럼 바뀌지 않는다. [그림 2-3]의 (c)는 (a)와 (b)에서 나타난 두 문장의 용언인 ‘맡다’의 행위주와 대상이 동일함을 알 수 있다.

제10절 감성 분석

감성 분석은 문장에서 화자의 감성을 다양한 방법과 분류로 인식하는 자연어 처리 기술이다. 감성 분석은 일반적으로 문장에서 언급하는 감성 대상에 대해 긍정 또는 부정의 감성을 판단하는 작업을 수행한다. 이때 감성을 나타나는 단어를 '감성 단어(Opinion Word)'라고 하고, 감성 대상이 되는 단어를 '감성 대상(Opinion Target)'이라고 한다.

즉, 감성 분석은 문장에서 감성 단어와 감성 대상을 찾는 작업이다. 감성 단어는 일반적으로 사전을 사용하여 검색 가능하다. 하지만 중의적인 표현이나 반어법 표현 등의 문장 표현 방식으로 인해 사전만으로는 해결하기 어려운 점이 있다. 따라서 이를 해결하기 위해 문장의 문맥을 파악하여 사전을 사용하는 방법 등이 제시되고 있다.

또한 감성 단어를 검색 후에 해당 감성 단어가 지칭하는 감성 대상을 찾아야한다. 감성 대상의 후보는 문장 내에서 다수 발견 가능하나 현재 화자의 감성의 대상이 되는 명확한 단어를 찾는 것이 가장 중요한 점이다.

나는 호텔의 조식은 좋았지만, 룸의 전망은 별로였다.

위의 예문에서 '좋았다' 라는 긍정 감성과 '별로다' 라는 부정 감성이 함께 나타났다. 이때 감성 대상이 될 수 있는 단어들은 '호텔' 과 '조식', '룸', '전망' 이 있다. '호텔' 이 좋다고 할 때, 절대적으로 틀렸다고는 할 수 없다. '좋았다' 라는 감성의 정확한 대상은 '조식' 이기 때문에 이를 이해하고 '조식' 과 '좋았다' 라는 단어에 관계 쌍을 생성해 주는 것이 중요하다.

감성 분석은 트렌트 트래킹이나 기업의 제품에 대한 소비자의 관점 등을 분석하는데 사용된다. 기업의 입장에서는 고객의 소리를 즉각적으로 수집하여 마케팅에 활용할 수 있기 때문에 최근 트위터나 블로그 등의 SNS에서 수집되는 문서에서 감성 분석 결과를 수집하는 것에 관심이 높아지고 있다.

제3장 기술 동향

이 장에서는 학술대회 및 경진대회를 통한 자연어처리 기술동향을 살펴본다.

제1절 학술 대회

1. ICON

ICON[7]은 “International Conference on Natural Language Processing” 의 약자로 자연어처리 국제 학술대회 중 하나이다. ICON은 자연어 처리 분야의 연구자들 간의 상호 작용을 촉진하기 위한 포럼이다. ACL 산하의 전산언어학 전자 논문집인 ACL Anthology의 주관으로 시행되고 있으며, 주요 국제 학회를 위한 전산 언어학 연구 논문의 디지털 아카이브이다. 2016년도에 13회 차수에 달하였다. ICON에서는 형태분석부터 구문, 의미 등의 다양한 자연어처리 분야의 대한 논문이 다뤄지고 있다. 논문은 자연어 처리의 모든 측면에서 실용적이고 독창적인 연구에 관련되어 있으며, 남아시아 언어를 중심으로 세계 언어 모두 연구하고 있다.

학술 대회에 접수하기 위해서는 영어로 된 10페이지 이하의 논문을 온라인 포털을 통해 접수해야 한다. 제출되는 논문은 약 100~200 단어의 초록을 포함해야 한다. 논문 양식은 다운로드 페이지에서 다운받을 수 있다. 논문 제출시 논문의 글자 및 스타일 등이 양식을 준수하지 않을 경우 논문 접수에 제한이 있을 수 있다. ICON에 제출하는 논문은 두 개의 트랙 중 하나를 선택하여 제출할 수 있다. 하나는 메인 컨퍼런스를 위한 논문이고, 다른 하나는 학생 논문 경연이다. 자연어처리 도구 콘테스트를 위해 작성된 논문의 경우 경진 대회에서 따로 접수한다.

2016년 ICON은 바라나시의 Banaras Hindu University(BHU)에서 개최되었다. 12월 17일부터 20일까지 나흘 동안 진행되었으며 메인 학회는 19, 20일 이틀에 걸쳐 진행되었다.

2. ACL

전산 언어학 협회(Association for Computational Linguistics)[8]는 전산 언어학 또는 자연어 처리(Natural Language Processing)라고 불리는 인간 언어와 관련된 전산 문제에 종사하는 사람들을 위한 최고의 국제 과학 및 전문 협회이다. ACL은 1962년에 설립되었으며 원래 기계 번역 및 전산 언어학 협회(Association for Machine Translation and Computational Linguistics)로 명명되었으며, 1968년 ACL이 되었다. ACL의 활동에는 매년 여름마다 개최되는 연례 회의 개최와 Computational Linguistics 저널 등의 활동을 하고 있다.

ACL에서 발행하는 저널은 크게 두 가지로 Computational Linguistics 저널과 Transactions of the Association for Computational Linguistics 저널이 있다. Computational Linguistics은 다시 MIT Press[9]와 MacQuarie University[10]로 분류된다.

Computational Linguistics - MIT Press 저널은 언어의 계산 및 수학적 특성과 자연어 처리 시스템의 설계 및 분석에만 전념하는 최장수 출판물이다. 수학자, 컴퓨터 언어학자, 인공 지능 및 기계 학습 연구자, 인지 과학자, 언어 전문가 및 철학자는 언어 연구의 모든 면에서 계산 가능한 측면에 대한 정보를 수집한다. 전산 언어학 협회(the Association for Computational Linguistics)가 후원하며, 1988년 이래로 MIT Press의 ACL에 게시되었으며 2009년 초부터 Open Access 되었다.

TACL(Transactions of the Association for Computational Linguistics)은 전산 언어학 및 자연어 처리의 모든 영역에서 논문을 출판하는 ACL 후원 저널이다. TACL은 저널 스타일 검정 프로세스를 가지고 있다. 또한 TACL에 게재되는 논문은 특정 ACL이 후원하는 컨퍼런스에서 발표할 수 있다. 따라서 저널의 장점과 주요 컨퍼런스에서 작품을 발표할 수 있는 이점이 있다. TACL은 일년 내내 논문을 접수하고 있다.

가. Annual meeting of the ACL

2016년 ACL 연례회의[11]가 처음으로 독일에서 개최되었다. ACL 2016은 2016년 8월 7일부터 12일까지 베를린의 훔볼트 대학에서 개회되었다. 지난 수년간과 마찬가지로 이번 학회 프로그램에서는 주요 발표 외에도 포스터 세션,

튜토리얼, 워크샵 및 시연이 포함되어 있다. ACL은 전산 언어학 분야의 최고 회의로써 자연 언어에 대한 전산 접근법과 관련된 다양한 연구 분야를 총 망라한다. ACL 2016에서는 다양한 튜토리얼과 워크샵이 진행되었다. 다음은 그 일부이다.

(1) Multimodal Learning and Reasoning

자연어 처리는 점점 더 많은 언어 이해와 추론 작업을 다루기 위해 영역을 넓혀왔다. 핵심 NLP 작업은 우리가 사람이 지각적으로 풍부한 환경에서 의사 소통을 하면서 언어를 습득하고 사용한다는 사실에도 불구하고 언어 입력에 초점을 맞추어 연구되어 왔다. 인간 수준의 인공 지능으로 이동하기 위해서는 언어 외에도 여러 가지 양식을 통합하고 모델링해야할 필요가 있다. 따라서 멀티 모달 신호를 처리할 수 있는 NLP 분야의 기술이 필요하다. NLP 작업에서 멀티 모달 신호를 사용하는 것은 다양한 데이터 양식을 보완하는 하나의 기술이다. 예를 들어, 텍스트 바코드에서 노란색 바나나나 또는 나무 의자에 대한 설명은 없지만 이에 대한 시각적 특성은 이미지에서 쉽게 추출이 가능하다. 이런 시각, 청각 또는 후각적 데이터와 같은 멀티 모달 신호는 단어의 유사성 및 관련성, 자동 이미지 및 비디오 설명, 관련 단어를 예측하는 모델에서 유용하게 사용 가능하다.

(2) Semantic Representations of Word Senses and Concepts

해석 가능한 형태로 언어 항목의 의미를 표현하는 것은 자연어 처리의 가장 중요한 목표였다. 다양한 언어 항목의 범위 중에서 단어는 가장 많은 연구 관심을 끌고 있다. 그러나 단어 표현에는 몇 가지 제한이 있다. 단어의 다른 의미를 단일 벡터로 결합하는 방식은 단어 감각의 표현이 가지는 고유한 한계를 극복할 가능성이 있다. 실제로 개별 단어 감각과 개념의 표현은 단어 수준에서 깊은 감각과 개념 수준으로 이동할 때 여러 가지 NLP 응용 프로그램에서 상당한 성능 향상을 보이고 있다. 개념 및 단어 감각 표현에 관한 또 다른 흥미로운 점은 이러한 모델이 단어, 구, 문장 등과 같은 다른 언어 항목에 원활하게 적용될 수 있다는 점이다.

(3) Game Theory and Natural Language: Origin, Evolution and Processing

게임 이론적 틀은 언어 기원과 진화를 연구하기 위해 여러 가지 방식으로

사용되었다. 또한 소위 게임은 철학자와 언어학자들이 언어가 어떻게 진화되었고, 어떻게 작동하는지 설명하기 위해서 사용되었다. 예를 들어 루드비히 비트겐슈타인(Ludwig Wittgenstein)은 언어의 전통적 성격을 설명하기 위해 언어 게임의 개념을 도입한 것으로 유명하다. 인구 내의 연사들 사이의 상호 작용으로부터 서서히 나타나는 공통 언어의 자발적 형성에 대한 아이디어를 제시한 바 있다.

이러한 개념은 언어적 해석과 언어적 단위, 그 상호 작용으로 구성된 복잡한 적응 시스템으로서 길을 열어 구조적 특성의 출현을 가져온다. 그것은 고전 게임 이론과 진화 게임 이론에 기초한 언어의 많은 계산 모델의 핵심 부분이 된다. 같은 맥락에서 언어의 다른 특성이 비슷한 역동성을 따르는 방법을 설명하기 위해 많은 시도가 있었다.

상호 작용 상황을 모델링하고 여러 소스의 정보를 통합할 수 있는 능력 덕분에 게임 이론, 특히 진화 이론은 언어의 생성과 언어의 감각 모호성 제거와 같은 자연어 처리 및 정보 검색의 특정 문제를 해결하는데 사용되었다.

나. Association for Computational Linguistics - European Chapter

EACL(European Chapter of the ACL)의 유럽 지부 구성은 1982년 토론토에서 열린 ACL 연례 회의에서 발표되었다. 이 학회의 주요 목표는 유럽 회원권에 대한 실질적인 장애물을 없애는 것이다. 회비 납부에 대한 높은 은행 수수료와 우편물 발송에 대한 높은 우편 요금 및 북미에서만 독점적으로 개최되는 회의에 대한 대서양 횡단 비용 등 이러한 요구를 해결하기 위해 EACL은 유럽에서 격년제로 진행되어 왔다. 주력 행사는 1983년 9월 피사에서 열린 Giacomo Ferrari와 함께 진행된 EACL 회의이다.

1983년을 시작으로 2014년까지 총 14회의 학회가 진행되었다. 지난 EACL 2014는 2014년 4월 26일부터 30일까지 진행되었다. 현지에서 준비는 언어 기술 센터(Centre for Language Technology)에서 주관하여 진행되었으며, CLT는 4개의 참여 부서에서 언어 기술의 연구 영역을 조정하였다. 4개의 연구 영역은 응용 정보 기술과 컴퓨터 공학, 언어학 및 과학 이론, 스웨덴어이다.

2014년 EACL에서는 다음과 같은 Computational modelling of metaphor 튜토리얼이 이루어졌다.

은유(Metaphor) 처리는 NLP에서 최근 급속하게 성장하는 영역이다. 은유의

편재성은 여러 가지 코퍼스 연구에서 확립되어 왔고, 인간의 추론에서 역할이 심리적 실험에서 확인되었다. 이는 은유를 컴퓨터 언어학 및 인지공학에서 중요한 연구 영역으로 전환되는 계기가 되었고, 의미론적 NLP 응용에 필수적인 식별 및 해석을 요구한다.

NLP와 AI의 은유에 대한 전산 작업은 1970년대에 불을 붙였으며, 1980년대에 모멘텀을 얻었고, 현상의 형태나 구조, 메커니즘에 대한 풍부한 아이디어를 제공하였다. 지난 수십년 동안 자료 기반의 통계 방법으로 자연 언어 계산 방법의 기술적 도약이 이루어졌다. 그러나 통계적인 은유 처리 방법에는 제한된 영역이 존재하기 때문에 개념의 현상이나 부분 집합에 관점을 두고 연구가 진행되어 왔다. 의미론적 과제에 대한 기계 학습을 위해 광범위한 NLP 방법이 적용되었고, 은유 해결을 위해서 다양한 연구가 진행되었다.

3. 한글 및 한국어 정보처리 학술대회

언어공학연구회는 회원 상호간에 학술활동과 친목을 목적으로 1991년 창립되었으며, 한국어를 중심으로 자연언어의 언어 공학적 연구에 관심 있는 모든 사람들이 회원이 될 수 있다. 특히 한글 부호체계, 한국어의 음성 인식 및 합성, 한국어의 기계에 의한 이해 및 생성, 한글 문자 및 문서 인식, 한국어의 언어학적 연구 등 우리말과 글의 정보처리에 관련된 기반 분야는 물론, 기계번역 및 정보검색 등과 같은 응용 분야에도 중점을 두고 있다.

연구회에서 추진하고 있는 학술활동으로는 매년 “한글 및 한국어정보처리 학술대회[12]”를 한국인지과학회와 공동으로 개최하고 있으며, 수시로 자연언어처리 튜토리얼, 워크샵, 소식지 발간 등 다양한 활동을 하고 있다.

2016년 28회를 맞이한 한글 및 한국어 정보처리 학술대회는 10월 7일부터 8일까지 이틀에 걸쳐 진행되었다. 논문 모집 분야는 자연언어처리, 한국어 정보처리, 음성인식, 음성합성, 감성인식, 문자인식, 기계번역, 텍스트 마이닝, 정보검색, 질의응답, 정보추출, 문서요약, 문서분류, 감정분류, 온톨로지, 시맨틱웹, 언어학, 인지과학, 언어지식 획득 및 구축, 언어습득 및 발달, 언어코퍼스[4] 분석, EEG 기반 언어분석, Eyetracking 기반 언어분석, 기타 한글 및 한국어 정보처리 관련 분야 모두 접수 하였다.

4) 코퍼스(Corpus) : 말뭉치를 뜻하며, 자연언어 연구를 위해 특정 목적을 가지고 언어의 표본을 추출한 집합.

이번 학술대회에서는 다음과 같은 초청 강연이 진행되었다.

가. 한국어 문서로부터의 시간정보 추출 (정영섭 박사, Naver Labs)

사람은 문서로부터 정보를 습득한다. 문서의 양이 기하급수적으로 증가함에 따라, 문서를 자동으로 분석하여 필요한 정보를 추출하는 기술을 개발해왔으며, 이 정보를 바탕으로 질의응답 시스템, 추천 시스템 등을 개발하는 것이 가능해졌다. 이 시스템들은 정보를 바탕으로 개발되므로, 정보의 종류와 질에 의해 서비스 종류와 질이 달라지게 된다. 따라서, 보다 다양한 정보를 정확하게 추출하는 기술이 필요하다. 특히, 대부분의 문서에는 시간 정보가 포함되어있으며, 시간정보는 시스템 성능에 큰 영향을 끼치기도 한다. 예를 들어, 질의응답 시스템에 “10 년 전 대통령이 누구야?” 라고 질문을 할 경우, ‘10 년 전’ 에 해당하는 시간 정보를 추출하지 못한다면 정확한 답변을 제공할 수 없게 된다. 시간 정보를 정확하게 추출하는 기술 개발을 위해 약 20여년 동안 활발히 연구되어왔지만, 한국어 문서로부터의 시간정보 추출 연구는 초기 수준에 머물러있다. 한국어 문서로부터 시간 정보를 추출할 때 발생하는 각종 이슈들과 현재 상황에 대해 공유한다.

나. 커뮤니티 기반 질문-응답 서비스를 위한 질문 분류 및 검색 기법 (배경만 박사, ETRI)

최근 지식iN, Yahoo! Answer, 바이두 쯔다오와 같이 사용자가 자신이 원하는 정보에 대한 질문을 등록하면, 다른 사용자가 질문에 대한 응답을 해주는 커뮤니티 기반의 질문-응답 서비스(community-based Question Answering Service, 이하 cQA)의 중요성이 늘어나고 있다. 효과적인 cQA 서비스를 구축하기 위해서는 질문을 분류하는 방법과 유사한 질문을 효과적으로 검색하는 방법이 중요하다.

질문 분류는 입력된 질문에 대해 적절한 범주를 자동으로 결정하는 방법으로 입력된 질문이 적은 수의 단어로 이루어져 있기 때문에 질문 분류에 필요한 정보가 부족하다는 문제를 가지고 있기 때문에 짧은 질문 길이 문제(short length problem)를 해결하는 것이 첫 번째 도전과제이다. 질문 검색은 입력된 질문과 의미적으로 유사한 질문을 찾아 주는 것으로써 질문 사이의 연관성을 효과적으로 계산할 수 있는 검색 모델을 설계하는 것이 중요하다. 추가적으로 의미적으로 같지만 다른 형태의 단어들로 구성된 질문들이 가지는 단어 불일치

문제(lexical gap problem)를 해결하는 것이 필요하고, 이를 해결하기 위한 방법들을 두 번째 도전과제로 지정한다. 그리고 의미적으로 유사한 질문은 비슷하거나 동일한 주제를 가질 것이라는 가정 하에 질문 검색의 성능 향상을 위해 질문 분류 결과를 활용하는 방법에 대한 연구를 진행하였으며, 이를 세 번째 도전과제로 정의한다.

첫 번째 도전과제를 해결하기 위해서 범주(category) 정보를 이용한 단어 가중치 계산 방법과 효과적인 질문 확장 방법에 대한 연구를 진행하였다. 먼저 단어 가중치 계산을 위해 현재 범주에서 분포와 다른 범주에서 단어들이 가지는 분포의 평균과의 비율을 고려한 단어 가중치 계산 방법을 제안한다. 다음으로 제안한 단어 가중치 계산 방법과 의사 적합성 피드백 방법을 이용하여 질문과 연관된 양질의 단어들을 효과적으로 확장하는 질문 확장 방법을 제안한다. 추가적으로 제안하는 가중치 계산 방법은 범주정보가 필요하며 테스트 질문의 잠재적인 범주를 결정하는 문제가 존재한다. 이를 해결하기 위해 엔트로피 기반의 결정 방법을 이용한 테스트 질문의 효과적인 표현 방법을 제안한다.

두 번째 도전과제를 위해서 의존 파서 결과를 기반으로 바이그램 자질 추출하고, 문장에서 중요한 성분인 주어, 서술어, 목적어, 보어로 이루어진 격틀 자질을 추출하여 새로운 질문 검색 모델을 제안한다. 추가적으로 단어 간의 연관성을 효과적으로 계산할 수 있는 워드 임베딩을 활용함으로써 유사한 질문을 효과적으로 검색할 수 있는 새로운 질문 검색 모델들을 제안한다. 마지막으로 제안한 질문 분류 방법과 단어 가중치 계산 방법을 기반으로 주제에서의 단어 분포를 효과적으로 반영할 수 있는 질문 검색 방법에 대해 제안한다.

제안한 방법에 대해 실험을 통해 평가를 진행하였다. 실험 결과를 통해 제안한 질문 분류 방법과 질문 검색 방법을 효과적으로 조합했을 때 좋은 성능을 얻을 수 있었다.

다. 다국어 공동 공간상의 개체 군집화를 통한 개체 요약 향상 기법 연구 (김은경 박사, KAIST 정보전자연구소 시맨틱웹연구센터)

시맨틱 웹과 링크드 데이터 기술의 발전으로 최근 몇 년간 상호 연결된 웹 데이터의 양이 방대하게 증가하고 있다. 특히 개체 중심의 데이터가 빠르게 증가하는 환경에서 <개체-속성-목적어>로 이루어진 지식베이스 규모가

방대해짐에따라, 서로 다른 지식 정보원에서 발행된 데이터 간의 중복적인 사실과 의미 없는 연결 정보가 비 정렬적으로 혼재되어 있어, 개체의 주요 본질을 신속하게 식별하기 어렵다. 본 강연에서는 링크드 데이터를 실현하기 위한 허브 역할을 하는 디비피디아로부터 다국어 연결 데이터를 확보하고 이를 활용하여 지식베이스에서 개체 단위의 주요 트리플을 추출하는 개체 요약 방법을 소개하고자한다. 특히 개체를 다국어 공동 공간상으로 투영한 자질로 군집화하고, 군집별로 중요한 속성 관계를 가려내 소형 요약본에 포함시키는 방식을 제안함으로써, 개체 본질을 기술하기 위한 필수 항목을 포함시키는 전문가의 요약 방식과 최대한 가깝게 재현한 기법을 소개한다. 제시된 기법은 개체 요약의 최신 기법과의 비교 품질 평가 결과를 통하여 기존의 독립된 언어 중심 정보 추출에서 간과할 수 있는 다양성 및 연관성을 효율적으로 확인할 수 있는지 분석한다.

라. 언어적 특징을 이용한 의생명 분야 초록의 구조화 (남세진 박사, 충남대)

의생명 분야에서는 1980년대 후반부터 구조화된 초록을 사용하고 있다. 구조화된 초록은 Introduction, Methods, Results, 그리고 Discussion(IMRAD)과 같은 섹션명으로 구분되어, 연구자들이 논문의 세부 내용을 쉽게 파악할 수 있도록 하여 논문 검색과 선택이 효율적이도록 도와준다. 그러나, 많은 의생명분야의 저널에서 구조화된 초록을 사용하고는 있지만, 여전히 MEDLINE의 초록 중 75%가 구조화되지 않은 상태로 있다. 해당 연구는 이렇게 비구조화된 의생명 분야의 초록을 구조화된 초록으로 변환하는 것을 목표로 초록의 문장을 IMRAD중 하나로 분류(Classification)하는 내용을 담고 있다. 문장 분류에서 가장 일반적으로 사용되는 특징(Feature)은 Bag-of-words이지만, 해당 연구에서는 의생명 분야의 초록에 나타난 언어적 특징을 문장 분류에 사용했다는 점에서 차별성을 가지고 있다. 강연을 통해 대규모 의생명 초록 코퍼스에서 추출한 초록의 섹션별 언어적 특징과 이러한 특징을 이용하여 문장을 분류한 결과를 공유한다.

마. 훈민정음의 과학성과 공학화 (변정용 교수, 동국대학교)

국보70호이면서 1997년 유네스코 기록유산으로 등재된 훈민정음은 우리의 문자생활에 큰 기여를 하여 왔다. 1933년 한글맞춤법이 제정되면서 표기체계가 정립되는 성과를 이루었다. 정보화시대를 맞이하여 한글 정보화는 한글부호

제정이 중요하고 그 중심에 보호화 대상을 음절로 할 것인가 음소로 할 것인가의 논란이 있었다. 결론은 음절로 하였고 그것이 곧 한글완성형부호이다. 그리고 유니코드에 한글자모, 한글호환자모, 한글음절 등 세가지 종류의 부호가 반영되고 최근에 한글자모가 추가되면 과학적인 문자에 누가 되고 있다. 지금 훈민정음 과학성을 논하는 것은 여러 가지 새로운 표현요구에 문제가 있고, 유니코드에 반영된 3종류 부호가 합당한지에 대하여 근본적 해결을 모색하고 방향을 제시하고자 함이다.

바. 언어장벽 해소를 위한 자동통번역 기술 동향 (김영길 실장, ETRI)

최근 글로벌화의 가속화로 언어장벽 해소를 위한 자동통번역 기술의 중요성이 커지고 있다. Google, MS 등 선진글로벌 기업을 중심으로 다양한 다국어 자동통번역 서비스를 경쟁적으로 출시하고 있으며, 국내에서도 시스트란인터내셔널, 한컴인터프리, 네이버 등에서 자동통번역 서비스를 내놓으면서 한국어를 중심으로 영어, 중국어, 일어에 대한 자동통번역 기술에 대한 경쟁력을 확보하고 있으며, 불어, 스페인어, 독어, 러시아어, 동남아권 언어에 대한 확장을 시도하고 있다.

그리고, 자동번역 방법론으로는 규칙기반(Rule-based Machine Translation)과 통계적 자동번역(SMT; Statistical Machine Translation)에 이어 신경망 자동번역(Neural Machine Translation)까지 범위가 확대되고 있으며, 이들 방법론의 장단점을 활용한 하이브리드 자동번역 기술도 활발히 연구되고 있다. 자동번역의 요소기술인 형태소 분석, 구문 분석, 의미 분석 등도 딥러닝 기술이 다양하게 적용되면서 성능 향상이 이루어지고 있다.

학술대회를 통해 발표된 논문 중에서 우수논문으로 선정된 논문은 논문지에 추천을 받아 게재할 수 있게 된다. 28회 학술대회에서는 총 10편의 논문이 우수논문으로 선정되었으며, 정보과학회와 인지과학회에 추천을 받았다. <표 3-1>은 28회 한글 및 한국어 정보처리 학술대회에서 우수 논문으로 선정된 10편의 논문이다.

<표 3-1> 제 28회 한글 및 한국어 정보처리 학술대회 우수논문

번호	제목	저자	추천논문지	제출자 성명
1	Stack LSTM 기반 한국어 의존 파싱을 위한 음절과 형태소의 결합 단어 표상 방법	나승훈 (전북대), 신종훈 (ETRI), 김강일 (건국대)	정보과학회	나승훈
2	Copy Mechanism과 Input Feeding을 이용한 End-to-End 한국어 문서요약	최경호, 이창기 (강원대)	정보과학회	최경호
3	2-Phase CNN을 이용한 SNS 글의 논쟁 유발성 판별	허상민, 이연수, 이호엽 (엔씨소프트)	정보과학회	허상민
4	효율적인 자동 주석을 위한 단어 임베딩 인공 신경 정리 증명계 구축	양원석, 박한철, 박종철 (KAIST)	정보과학회	양원석
5	딥러닝 모형 기반 한국어 개체명 연결	손대능, 이동주, 이용훈, 정유진, 강인호 (네이버 검색연구)	정보과학회	손대능
6	한국어 질의응답 시스템을 위한 프레임 시멘틱스 기반 질의 의미 분석	함영균, 남상하, 최기선 (KAIST)	인지과학회	함영균
7	품사 임베딩과 음절 단위 개체명 분포 기반의 Bidirectional LSTM CRFs를 이용한 개체명 인식	유홍연, 고영중 (동아대)	정보과학회	유홍연
8	포인터 네트워크를 이용한 멘션탐지	박천음, 이창기 (강원대)	정보과학회	박천음
9	두벌식 한글 입력 방식의 정형적인 기술과 분석	김용묵 (연세대), 김국 (서경대)	인지과학회	김용묵
10	한국어 수사구조 분류체계 수립 및 주석 코퍼스 구축	노은정, 이연수 (엔씨소프트), 김연우, 이도길 (고려대)	인지과학회	노은정

제2절 경진 대회

1. NLP tools contest

NLP tools contests at ICON는 ICON에서 진행하는 프로그램 중 하나로 자연어처리 도구 콘테스트로 해마다 새로운 주제를 통해 다양한 자연어처리 도구 기술의 발전을 도모하고 있다. 2015년도에는 통계기반 기계 번역인 SMT 분야를 주제로 하여 진행되었으며, 2016년에는 인도 소셜 미디어 텍스트를 형태소 분석 및 품사 부착하는 것이 주제로 선정되었다. 최근 소셜 네트워크 서비스를 통해 수집 가능한 텍스트가 급증하는 현상에 따라 소셜 미디어를 통해 생산되는 텍스트를 분석하는데 관심이 집중되고 있다. 형태소 분석 및 품사 부착은 자연어처리의 가장 기본이 되며 오랫동안 연구되어온 주제이다. 하지만 소셜 미디어 텍스트는 오타나 신조어, 약어 등이 빈번히 사용되어 분석이

쉽지가 않다. 따라서 이번 대회를 통해 연구되는 형태소 분석 및 품사 부착 시스템은 향후 소셜 미디어 텍스트 분석의 기반이 될 것으로 예측된다.

2016 ICON에서 진행한 NLP Tools contests는 "TOOL CONTEST ON POS TAGGING FOR CODE-MIXED INDIAN SOCIAL MEDIA (FACEBOOK, TWITTER, AND WHATSAPP) TEXT"라는 주제로 진행되었다. 대회에서 대상으로 하고 있는 소셜 미디어 텍스트는 인도어와 영어이다. 인도의 경우에는 수 백 가지의 언어를 사용하고 있으며, 소셜 미디어에서 인도어와 영어를 혼용하여 사용하고 있다. 따라서 이를 분리하여 분석 가능한 시스템 구축이 필요하다. 다음의 예문은 영어와 인도어가 혼용되어 사용되는 예이다.

> ICON 2016 Varanasi me hold hoga! Great chance to see the pracheen nagari!
>
> (※해석 : ICON 2016 바라나시에서 개최됩니다. 고대 도시를 볼 수 있는 좋은 기회입니다.)

참가자들은 POS 태깅 시스템을 개발 및 평가하기 위해 학습, 개발, 평가 데이터를 제공받게 된다. 제공되는 데이터는 영어-힌디어, 영어-벵골어, 영어-텔루구어 총 3가지의 혼합어 트위터 텍스트 데이터이다. 각각의 데이터는 언어에 따라 추가적인 정보가 포함되어 있다. 시스템 성능 측정은 정밀도와 재현율, 이를 조화 평균한 F-1 Score를 사용한다. 예선 심사를 거쳐 후보에 오른 참가자들은 본선에서 특별 세션을 통해 발표를 진행한다.

제공되는 학습 데이터의 상세 내용으로는 영어-힌디어, 영어-벵골어, 영어-텔루구어의 3가지 언어 세트가 제공된다. 각 언어 세트는 모두 트위터 1천 문장, 페이스북 1천 문장, 왓츠앱 1천 문장이 포함되어 있다. 제공되는 텍스트 데이터에는 영어와 인도어의 혼합 또는 인도어 단일의 텍스트가 포함되어 있으며, 영어 단일의 텍스트는 전부 제외하였다. 또한 텍스트에 포함된 영어와 인도어의 비율이 최대한 비슷하도록 균형을 유지하여 데이터를 구축하였다. 제공되는 텍스트 데이터는 대부분 2중의 언어 혼합 데이터이지만 부분적으로 3중 또는 4중의 언어가 혼합된 데이터도 포함되어 있다. 예를 들어 영어-힌디어 텍스트에서 벵골어가 나타나는 경우이다.

혼합어 텍스트 데이터 생성시 언어의 균형을 위해서 코드 혼합 지수인 CMI(Code-Mixing Index)를 도입하였다. CMI는 텍스트에서 가장 빈도 높은 언어를 찾은 다음에 그 외에 존재하는 다른 언어를 검사하여 단어의 빈도를 측정한다. CMI를 적용하여 제공되는 데이터는 페이스북, 트위터, 왓츠앱 각

3개와 영어-힌디어, 영어-벵골어, 영어-텔루구어 3개 그리고 데이터의 세분화에 따라 2개로 총 18가지의 텍스트 데이터를 제공한다. <표 3-2>는 제공되는 데이터의 POS 태그셋 정보이다.

<표 3-2> POS Tagset

Category	Type	Description
Noun (G_N)	N_NN	Common Noun
	N_NNV	Verbal Noun
	N_NST	Spatio-temporal
	N_NNP	Proper Noun
Pronoun (G_PRP)	PR_PRP	Personal
	PR_PRL	Relative
	PR_PRF	Reflexive
	PR_PRC	Reciprocal
	PR_PRQ	Wh-Word
Verb (G_V)	V_VM	Main
	V_VAUX	Auxiliary
Adjective (G_J)	JJ	Adjective
Adverb (G_R)	RB_ALC	Locative Adverb
	RB_AMN	Adverb of Manner
Demonstrative (G_PRP)	DM_DMD	Absolute
	DM_DMI	Indefinite
	DM_DMQ	Wh-word
	DM_DMR	Relative

Category	Type	Description
Quantifier (G_SYM)	QT_QTF	General
	QT_QTC	Cardinal
	QT_QTO	Ordinal
Particles (G_PRT)	RP_RPD	Default
	RP_NEG	Negation
	RP_INTF	Intensifier
	RP_INJ	Interjection
Residual (G_X)	RD_RDF	Foreign Word
	RD_SYM	Symbol
	RD_PUNC	Punctuation
	RD_UNK	Unknown
	RD_ECH	Echo Word
Conjunction, Pre- & Postposition	CC	Conjunction
	PSP	Pre-/Postposition
Numeral	&	Numeral
Determiner	DT	Determiner
Twitter-Specific (Gimpel *et al.* 2011) (G_X)	@	At-mention
	~	Re-Tweet/discourse
	E	Emoticon
	U	URL or email
	#	Hashtag

※출처 : http://amitavadas.com/Code-Mixing.html

2. EVALITA

EVALITA(Evaluation of NLP Tools for Italian)[14]는 이탈리아 전산언어학회(Italian Association for Computational Linguistics) 주관으로 2007년 처음으로 시작되었으며, 매 진행마다 당시 이슈에 맞는 새로운 주제를 선정하여 다양한 자연어처리와 음성 처리 기술의 발전을 도모하였다. 여기서는 자연어처리 및 이탈리아어용 음성 처리 도구에 대한 주기적인 평가 캠패인을 진행한다. EVALITA는 2016년 5회차를 맞았으며 다음의 6개 기술에 대한 평가 캠패인을 진행하였다.

가. ArtiPhone (Articulatory Phone Recognition)

ArtiPhone은 자동 음성 인식 기술 도구에 대하여 평가가 진행되며 제공되는

학습데이터는 일정한 속도의 음성 데이터이지만 실제 평가 시에는 대화 속도가 불규칙적인 데이터가 사용된다.

자동 음성 인식 기술은 지난 수 년 동안 많은 기술의 발전을 이루어왔다. 하지만 수천 시간 이상의 충분한 리소스에도 불구하고, 음성 데이터의 상이한 음향 환경이나 데이터 사용 환경에 따라 여전히 열악한 성능을 내고 있다. 이것은 대개 불일치 문제라고 부른다. 불일치 문제는 이미 학습된 단어임에도 불구하고 억양이나 말하는 속도 등의 문제로 오류를 발생시킨다. 따라서 이 도전 과제는 일치하지 않는 말하기 속도의 음성 데이터를 사용하여 전화 인식 시스템을 구축해야 한다. 학습에 사용되는 데이터는 평균적인 속도로 진행되는 음성 데이터이지만 평가에 사용되는 데이터는 학습 데이터보다 느리거나 빠른 음성 데이터를 사용한다.

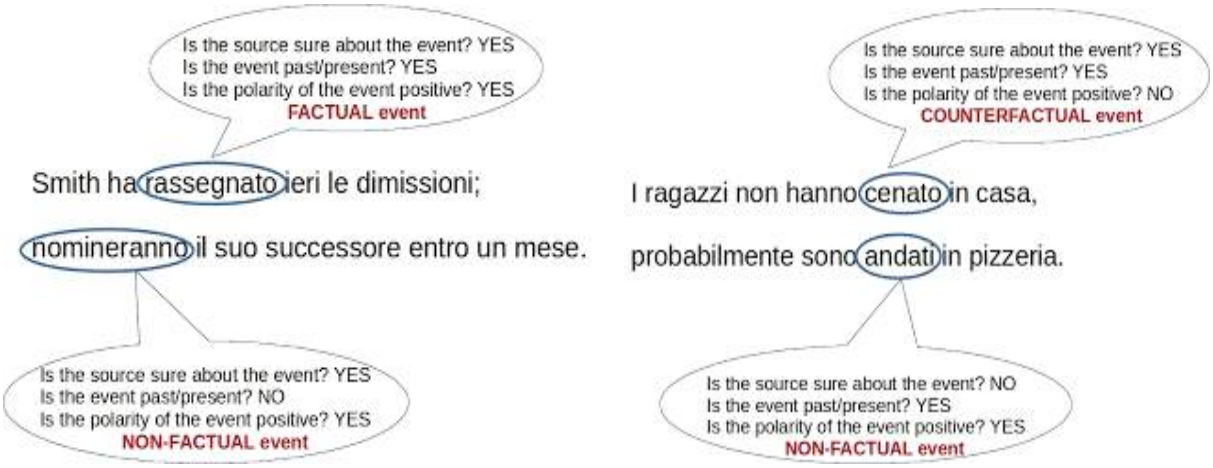
전화 음성 인식 시스템은 음향 전용 데이터에 대한 학습이나 조음 데이터에 대한 학습을 받을 수 있다. 조음 데이터는 x, y, z의 3차원 위치에서 7개의 육체점(위턱, 아래턱, 혀끝, 입술 등)으로 구성되어 있다. 제공되는 음성 데이터인 MSPKA는 다양한 말하기 스타일과 음성 및 음성의 궤적을 동시에 녹음한 이탈리아 음성 데이터이다. 총 2가지 음성 데이터가 제공되는데 하나는 약 500문장의 텍스트를 남성 1명, 여성 2명이 말하는 약 2시간 정도의 음성 데이터이다. 다른 하나는 첫 번째의 음성 데이터에서 과장된 데이터를 교정한 약 3시간 정도의 음성 데이터이다.

음성 인식 시스템의 평가는 오류율(Phone recognition Error Rate)로 측정한다. 2개의 기준 시스템 및 PER 시스템이 제공된다. 2개의 기준 시스템은 음향 정보만을 사용한 시스템과 음향 및 조음 정보를 모두 사용한 시스템이다. 두 기준 시스템 모두 심층 신경망 모델이며, 은닉 마르코프 모델(HMM)이 결합된 DNN-HMM 시스템으로 이를 이용하여 참가자들은 자신의 시스템을 구축한다. 참가자들은 기준 시스템의 코드를 자유롭게 변경 가능하다.

나. FactA (Event Factuality Annotation)

FactA는 문서 내에서 발상한 사건과 사건의 분류, 사실성 여부 판단 등의 일을 수행하는 도구에 대한 평가가 진행된다. 평가 참가자들은 사건의 확실성과 시간, 극성의 세 요소에 대해 판단하도록 시스템을 구성하여야 한다. [그림 3-1]은 문장에서 이벤트를 추출하는 예시이다. 이벤트를 인식할 때는 몇 가지의

확인사항이 있다. 이벤트의 출처와 과거 또는 미래의 사건 그리고 사건의 긍/부정적 극성이 있다. 이와 같은 사항에 따라 FACTUAL, NON-FACTUAL 그리고 COUNTER FACTUAL event인지 판단한다.



※이미지출처 : http://facta-evalita20

[그림 3-1] FactA의 이벤트 인식 및 주석 방법 예시

FactA의 중점은 뉴스 와이어 도메인에 맞추어져 있다. 학습 및 평가 데이터는 모두 뉴스 기사로 구성되어 있다. 이 외에도 소셜 미디어 텍스트 중 사실성 인식 및 추출에 대한 파일럿 작업을 제안한다.

FactA의 시스템에서 사건과 관련된 세 가지 속성의 조합을 통해 이벤트의 사실성을 확인한다.

- 확실성 - 확실성은 자료에 언급된 사건에 대한 확실한 방법과 관련된다. (certain, non_certain, and underspecified.)
- 시간 - 시간은 사건이 발생했거나 발생할 것으로 보고 된 시간을 말한다. (past/present, future, and underspecified.)
- 극성 - 극성은 사건이 확인되거나 무효화되면 결과에 따라 양수 또는 음수가 될 수 있다. 사건의 극성을 판단하기에 정보가 충분하지 않을 때는 명시하지 않는다.

평가 데이터에는 주석이 달린 정답 사건이 포함되어 있다. 참가자는 주어진 사건에 대해 3가지 사실 관계 속성에 대해 올바른 값을 할당하여야 한다.

시스템 산출물의 평가는 사실 관계 속성값에 대하여 진행한다. 시스템 성능은 정밀도, 재현율, 정밀도와 재현율의 조화 평균인 F-1 점수를 사용하여 평가한다. 공식 순위는 F-1 점수를 사용한다.

제공되는 데이터는 CAT(Content Annotation Tool) 레이블 형식의 XML 기반의 데이터이다. 독립 실행형 형식으로 서로 다른 주석 영역이 별도의 문서의 섹션에 저장되며 서로 관련된 포인터 정보를 활용할 수 있다. 각 문서는 한 줄에 하나의 토큰인 토큰 목록을 포함하며 사건의 경우 <EVENT>라는 태그가 부착되어 있다.

참가자는 주석 지침에 따라 세 가지의 속성값을 부착한 결과 데이터를 제출해야 한다. 이 결과를 FV(Factuality Value)를 통해 점수를 측정한다. 시스템의 순위는 FV의 F1 점수를 기반으로 한다. 아래는 제공되는 토큰 데이터의 일부이다.

```
<EVENT m_id="1" certainty="CERTAIN" time="PAST/PRESENT" polarity="POS">
<EVENT m_id="2" certainty="CERTAIN" time="PAST/PRESENT" polarity="POS">
<EVENT m_id="3" certainty="UNCERTAIN" time="PAST/PRESENT" polarity="POS">
```

FactA에서는 정확한 평가를 위해서 다음과 같은 주석 부착 가이드 라인을 제시하고 있다. 사건의 사실성에 대한 주석을 위해 2014년에 작성된 가이드 라인으로 확실성, 시간, 극성 3가지 속성을 표현하는 방법을 제시한다.

(1) 확실성

확실성 속성은 CERTAIN, NON_CERTAIN, UNDERSPECIFIED로 분류되며 각각의 내용은 다음과 같다. CERTAIN은 사건이 일어났는지 일어나지 않았는지 혹은 일어나거나 일어나지 않을 것인지 확실한 경우에 부착한다. NON_CERTAIN은 사건이 일어났거나 일어나지 않았거나 혹은 일어나거나 일어나지 않을 것으로 확신하지 못할 경우에 부착한다. UNDERSPECIFIED는 문서에서 사건의 확실성에 대하여 판단할 수 있는 정보가 없을 경우 부착한다.

확실성을 판단할 수 있는 정보는 확률성을 나타내는 등의 가능성에 대한 정보이다. 따라서 희망이나 기대와 같은 불확실한 근거는 정보로써 활용하지 않는다. 예를 들어 “나는 내일 우승할 것 같다.”와 같은 문장은 확실성을 판단할 근거가 없는 문장이다. 따라서 ‘UNDERSPECIFIED’ 값을 부착한다. 확실성을 판단할 수 있는 문장은 다음과 같다.

① "내일 우리 팀이 승리할 확률은 희박하다."
② "잘 기억나지 않지만 아마 2시쯤 집에 도착한 것 같다."

①은 확률에 근거하여 사건을 예측하고 있기 때문에 CERTAIN 값을 부착한다. ②의 경우에는 사건에 대해 정보를 제시하고 있으나 불분명하게 나타나고 있다. 따라서 NON_CERTAIN값을 부착한다.

(2) 시간

시간 속성은 사건의 시간적 해석 또는 의미론적 시간 값을 지정하고 사건이 과거에 나타났는지 혹은 미래에 나타날 예정인지를 판단한다. 하지만 시간 속성은 시제 정보에 기반하지 않는다. 사건을 해석할 때 전체적인 시간적 흐름과 문서의 발화 사이의 복합적인 관계를 포함하여 속성값을 부착해야 한다.

시간 속성은 PAST/PRESENT, FUTURE, UNDERSPECIFIE 3가지로 분류된다. PAST/PRESENT는 현재 및 과거의 사건을 나타내고, FUTURE는 미래에 발생할 사건을 나타낸다. UNDERSPECIFIED는 시간적 해석을 할 수 있는 정보가 제공되지 않는 경우에 부착한다. 몇몇의 발화의 경우 항상 발화 시간과 관계있거나 항상 발화 시간과 무관한 경우도 있다. 따라서 이런 예외를 주의하여 주석을 부착하여야 한다. 또한 일반적인 명령이나 사건의 특징을 설명하는 경우에는 시간 속성이 UNDERSPECIFIED로 부착된다.

(3) 극성

극성 속성은 문장의 긍정과 부정의 차이를 판단하며 POS, NEG, UNDERSPECIFIED 3가지의 속성값을 사용한다. POS는 긍정적인 의미가 있는 사건에 부착한다. NEG는 부정적인 의미의 사건에 부착되며, UNDER-SPECIFIED는 사건의 긍정과 부정을 판단할 수 없는 경우에 부착한다.

속성값을 부착할 때 단어를 인식하였으나 속성값을 부착하지 않는 경우도 있다. 이는 사건을 설명하기 위해 있는 문장이지만 실제 사건을 판단하기 위해 사용되지 않는 문장일 경우이다. 예로 다음과 같은 문장을 분석한 주석 부착은 <표 3-3>과 같다.

La Volkswagen non ha detto quanto costerà la costruzione dell'XL1

(※해석 : 폭스바겐은 XL1의 건설비용을 언급하지 않았습니다.)

<표 3-3> 속성값을 부착하지 않는 경우 예시

predicate (서술어)	certainty (확실성)	polarity (극성)	time (시간)
detto	certain (확실)	neg (부정)	past/present (과거 또는 현재)
costerà	-	-	-
costruzione	-	-	-

<표 3-4>는 FactA에서 사용되는 다음 예문의 주석 부착 예시이다.

Giovanni non sa se Maria verrà

(※해석 : 요한은 마리아가 올 것인지 알지 못했습니다.)

<표 3-4> FactA 주석 부착 예시

predicate (서술어)	certainty (확실성)	polarity (극성)	time (시간)
sa	certain	neg	past/present
verrà	non_certain	underspecified	past/present

다. NEEL-IT (Named Entity Recognition and Linking in Italian Tweets)

NEEL-IT는 Named Entity rEcongition and Linking(NEEL)을 목적으로 진행되며, 트윗 텍스트를 대상으로 동작하는 시스템을 평가한다. 트윗 텍스트는 최근 트렌드와 사용자의 행동을 실시간으로 반영하는 데이터로 유용한 정보가 포함되어 있다. 하지만 트윗 텍스트에는 다수의 노이즈가 포함되어 있어 분석이 어렵다. 따라서 이에 대응 가능한 NEEL이 수행 가능하도록 시스템을 개발하는 것이 중요하다. EVALITA는 이 분야의 기술을 발전시켜 DBpedia와 같은 지식 기반의 데이터에 연결(Linking)을 진행하고자 한다. NEEL-IT는 NER 및 NEL

시스템을 위한 자연어 처리 도구 개발을 장려하고 이탈리아 커뮤니티를 위한 평가 체계를 확립할 수 있는 기회를 제공한다. NEEL-IT는 이탈리아의 트위터 문서에서 참조 및 평가 프레임 워크로 자리매김할 수 있는 비전을 가지고 있다.

NEEL-IT는 트위터의 English Micropost에 대한 NEEL challenge와 비슷한 형식을 따른다. 이 작업은 지식 기반(DBpedia)에 링크를 연결하여 텍스트에 있는 명명된 엔티티의 설명과 같은 주석 정보로 구성된다. 주석 부착 프로세스에서 명명된 엔티티는 분류 체계에서 지정된 범주 중에 속하거나 DBpedia의 개념에 연결할 수 있는 적절한 명사를 나타내는 트위터 문자열이다. 따라서 분류 체계에 따라 엔티티를 가질 수 있으나 DBpedia와 연결 정보가 없을 경우 “NIL DBpedia” 라는 참조 정보를 부착한다. NEEL-IT에서는 사물, 사건, 캐릭터, 장소, 조직, 사람, 제품의 총 7가지의 분류 체계를 가진다. 주석을 부착하는 엔티티는 관사나 접두어를 포함하여 부착한다. 참가자는 주석을 부착할 때 주석마다 한 줄씩 주석 정보를 생성해야 한다.

참가자는 자신의 시스템 결과를 TSV 파일로 최대 3회까지 제출할 수 있다. 제출 형식의 예제는 개발 세트과 함께 배포된다. 참가자의 결과물을 평가하기 위해서 TAC KBP Scorer[16] 시스템을 사용한다. 평가시에는 다음의 세 가지 정보에 대해 중점을 두고 측정한다. 첫 번째로 tagging. tagging은 엔티티의 단어 경계와 분류 정보를 확인한다. 두 번째는 linking으로 linking은 DBpedia와 연결된 링크 정보를 확인한다. 마지막 세 번째는 clustering으로 NIL정보에 대해 정확하게 분류되었는지를 확인한다. 세 가지 모두 F-1 점수를 평가에 사용한다.

NEEL-IT는 트위터의 고유 명사나 약어 등에 대하여 엔티티를 분류하고 DBpedia와 링크를 연결하는 작업이다. 엔티티의 범위는 이름을 나타내는 전체 문자열이며 단어 앞의 관사나 접두어 등의 인용 부호를 모두 포함한다. 복합 명사는 단어를 분리하여 엔티티를 부착한다. NEEL-IT는 크게 두 가지 작업을 진행한다. 첫 번째는 분류 체계에 따른 엔티티 인식 및 분류이고, 두 번째는 트위터의 문맥을 고려하여 DBpedia 항목에 연결 또는 NIL 정보를 부착하는 것이다.

엔티티를 인식할 때 몇 가지 유의사항이 있다. 먼저 대명사는 엔티티에서 제외한다. 맞춤법 오류 중 트위터에서 흔히 사용하는 스펠링의 변형이나 약어 같은 경우에는 고유 명사에 사상될 수 있다면 엔티티로 판단한다. 또한 완전히 엔티티로 분류된 문자열의 부분 문자열은 엔티티로 판단하지 않는다. 예를 들어

“Galaxy Note 2”의 경우 “Galaxy”와 “Note 2”로 분리하여 엔티티를 생성할 수 있지만 “Galaxy Note 2”로 엔티티 하나만 생성한다. 명사구의 경우는 명사구 전체를 엔티티로 인식하지 않고, 명사구 내에 엔티티를 분류할 수 있는 문자열이 있을 때 해당 문자열만 엔티티로 분류한다.

트위터 텍스트의 경우 트위터만의 특수한 엔티티 정보를 포함하고 있다. 따라서 다음의 몇 가지 사항에 대해 가이드 라인을 제시한다. 먼저 트위터 텍스트의 경우 해쉬 태그나 @ 기호를 사용하여 엔티티를 생성할 수 있다. 해쉬 태그나 @ 기호는 트위터에서 참조의 목적으로 사용되는데 이 중 엔티티가 될 수 있는 명사들이 다수 포함된다.

하지만 이는 해쉬 태그와 @ 기호에 따른 참조 문자열이 엔티티가 될 수 있다는 가능성을 나타내는 것이지 모든 해쉬 태그가 엔티티가 된다는 의미는 아니다. 따라서 다음의 경우에 엔티티가 될 수 있는지를 고려해야 한다. 먼저 기본적인 엔티티 생성에서 언급한 명사구에 대한 해쉬 태그이다. 명사구는 엔티티로 인정하지 않기 때문에 해쉬 태그가 부착된 명사구 역시 엔티티로 분류하지 않는다. 또한 이미 엔티티로 분류된 문자열의 부분 문자열을 해쉬 태그로 사용한 경우도 엔티티가 되지 않는다.

트위터 텍스트의 경우 닉네임을 사용하는 경우가 흔하기 때문에 트위터에서 사용되는 닉네임에도 엔티티를 부착한다. 예를 들어 “Harry Styles”에 “Hazza”라는 닉네임을 붙여 사용하였다면 “Hazza”에 “Harry Styles”의 엔티티를 참조하여 동일하게 엔티티를 부착한다.

DBpedia 연결 정보는 DBpedia2015-10버전을 기준으로 엔티티와 연결한다. DBpedia는 널리 사용되는 Linked Data이며 RDF (Resource Description Framework) 리소스로 구성되어 있다. 각 리소스는 URI (Uniform Resource Identifier)로 고유하게 식별되어 있다. 식별된 엔티티는 DBpedia의 리소스(트리플)과 연결되어야 한다. 만약 DBpedia와 연결할 수 없다면 NIL 정보를 부착한다.

제공되는 데이터는 트위터 텍스트 파일과 주석 파일로 두 개가 제공된다. 트위터 텍스트는 트위터 id정보와 함께 나열된 트위터 목록이고, 주석 파일은 각 트위터 id에 대한 엔티티와 링크 정보를 포함하고 있다.

NEEL-IT에서 사용하는 분류 체계 총 7가지로 다음과 같다.

<표 3-5> NEEL-IT의 분류체계

분류 체계	항목	예시
사물(Thing)	언어(languages), 인종(ethnic groups), 국적(nationalities), 종교(religions), 질병(diseases), 스포츠(sports), 천체(astronomical objects)	영어, 한국어, 유럽인, 기독교, 홍역, 달, 지구, 축구
사건(Event)	기념일(holidays), 스포츠 행사(sport events), 정치 사건(political events), 사회적 사건(social events)	추수감사절, 브라질 월드컵, 워터게이트 사건
캐릭터(Character)	허구 인물(fictional character), 만화 캐릭터(comics character), 표제 인물(title character)	셜록 홈즈, 스파이더맨
장소(Location)	공공 시설물(public places), 지역(regions), 상업 장소(commercial places), 빌딩(buildings)	오페라 하우스, 뉴욕시, 63빌딩
조직(Organization)	회사(companies), 기업의 세분류(subdivisions of companies), 브랜드명(brands), 정당(political parties), 정부 기관(government bodies), 언론사(press names), 공공 기관(public organizations), 협회 및 단체(collection of people)	삼성, 삼성 전자, 청와대, 민주당, 한국일보, 소프트웨어산업협회
사람(Person)	사람의 이름(people's names)	오바마, 트럼프
제품(Product)	영화(movies), TV 프로그램(tv series), 음악 앨범(music albums), 출판물(press products), 장비(devices), 운영 시스템(operating systems), 프로그래밍 언어(programming languages)	바람과 함께 사라지다, CSI, IOS, Windows

라. PoSTWITA - POS tagging for Italian Social Media Texts

POS(Part-of-Speech) 태그 부착 작업은 주로 표준화된 텍스트에 오랜 기간 동안 집중되어 연구되었다. 그러나 최근 소셜 미디어 텍스트의 자동 평가에 대한 관심이 상당히 증가되고 있다. 특히 트위터나 마이크로 블로그의 문서들이 크게 증가하고 있는데, 이와 같은 소셜 미디어 텍스트에서 추출할 수 있는 정보는 이미 다양한 분야에서 추세 및 향후 사건 인식 등을 식별하기 위해 활용하고 있다.

소셜 미디어 텍스트의 본질은 사용되는 어휘나 배포 속성 등과 관련하여

기존의 표준 텍스트와는 분명하게 다른 점이 있다. 따라서 기존의 자연어 처리 방법으로 해결하기에는 문제점이 존재하기 때문에 이에 알맞은 시스템 연구가 필요하다. 따라서 EVALITA에서는 트위터 텍스트에 POS 태그를 부착하는 작업을 제안한다.

2016년도 EVALITA의 PoSTWITA에서는 2014년에 사용된 SENTIPLOC 코퍼스의 일부를 재사용하였다. 개발 및 평가 세트는 각각 4,041문장과 1,749문장이고 수동으로 태그를 부착하였다. 참가자는 시스템 성능 향상을 위해 배포되는 리소스 외에 다른 리소스를 사용할 수 있다. PoSTWITA에서는 몇 가지 태그 부착 주의 사항이 있다. 먼저 개체명의 경우 개체명을 하나의 고유한 토큰으로 간주하여 개체명 전체에 'PROPN' 태그를 부착한다.

텍스트 분할(토큰화) 문제는 POS 태그 지정 및 평가에서 핵심적인 문제이다. 텍스트 분할의 기준이 명확하지 않으면 서로 다른 분할 기준으로 인해 상이한 결과를 내기 때문이다. 따라서 PoSTWITA는 토큰화된 형식의 개발 및 평가 데이터를 제공한다. 한 줄에 하나의 토큰과 해당 토큰의 태그가 부착되어 있다. <표 3-6>은 제공되는 데이터의 예이며, <표 3-7>는 PoSTWITA의 태그셋이다.

<표 3-6> PoSTWITA의 데이터 예

개 체	태 그	개 체	태 그
샘플데이터 ____162545185920778240____		샘플데이터 ____192902763032743936____	
Governo	PROPN	#Ferrara	HASHTAG
Monti	PROPN	critica	VERB
:	PUNCT	#Grillo	HASHTAG
decreto	NOUN	perché	SCONJ
in	ADP	dice	VERB
cdm	PROPN	cose	NOUN
per	ADP	che	PRON
approvazione	NOUN	dicevano	VERB
!	PUNCT	Berlusconi	PROPN
http://t.co/Z76KLLGP	URL	e	CONJ
		Bossi	PROPN
		.	PUNCT
		E	CONJ
		che	PRON
		non	ADV
		hanno	AUX
		fatto	VERB
		.	PUNCT

※출처 : http://corpora.ficlit.unibo.it/PoSTWITA/index.php?slab=data

<표 3-7> PoSTWITA 태그 셋

TAGSET		CATEGORY	EXAMPLES
UD	EVALITA16		
ADJ	ADJ	Adjective	-
ADP	ADP	Adposition (simple prep.)	di, a, da, in, con, su, per…
	ADP_A	Adposition (prep.+Article)	dalla, nella, sulla, dell'…
ADV	ADV	Adverb	-
AUX	AUX	Auxiliary Verb	-
CONJ	CONJ	Coordinating Conjunction	-
DET	DET	Determiner	-
INTJ	INTJ	Interjection	-
NOUN	NOUN	Noun	-
NUM	NUM	Numeral	-
PART	PART	Particle	-
PRON	PRON	Pronoun	-
PROPN	PROPN	Proper Noun	-
PUNCT	PUNCT	punctuation	-
SCONJ	SCONJ	Subordinating Conjunction	-
SYM	SYM	Symbol	-
	EMO	Emoticon/Emoji	:-) ^_^ ♥ :P □
	URL	Web Address	http://www.somewhere.it
	EMAIL	Email Address	someone@somewhere.com
	HASHTAG	Hashtag	#staisereno
	MENTION	Mention	@someone
VERB	VERB	Verb	-
	VERB_CLIT	Verb + Clitic pronoun cluster	mangiarlo, donarglielo…
X	X	Other or RT/rt	-

※출처 : http://corpora.ficlit.unibo.it/PoSTWITA/index.php?slab=guidelines

마. QA- 4FAQ - Question Answering for Frequently Asked Questions

웹 사이트의 FAQ(Frequently Asked Questions) 페이지에서 검색하는 것은 많은 사람들이 이용하는 서비스이고, 중요한 작업이다. 고객은 찾고자 하는 정보와 관련 없는 수많은 질문 중에서 적합한 FAQ를 찾는데 어려움을 겪고 있다. QA- 4FAQ는 사용자가 생성한 쿼리를 분석하여 FAQ 목록에서 적합한 답변을 검색하는 작업으로 구성되어 있다. AQP(Acquedotto Pugliese)는 QA(Question Answering) 기술을 기반으로 AQP Risponde라는 FAQ를 위한 의미론적 검색 엔진을 개발하였다. 이 시스템을 통해 고객은 자신의 질문과 관련된 FAQ를 쉽고 정확하게 검색할 수 있다. 또한 고객은 시스템이 제시한 FAQ 중 적합한 답변을 선택하고 해당 답변의 정확도에 대한 피드백을 제공할 수 있다. AQP Risponde는 QA 구조에서 이탈리아어 사용법과 고객의 요구 사항을 공식화하기 위하여 다양한 표현 방법에 대해 연구를 진행한다. QA- 4FAQ는 궁극적으로 커뮤니티에서 작성한 토론 포럼에서 새로운 질문에 대해 적합한 답변을 찾는 과정을 자동화하는데 도움이 된다. 더욱이 QA- 4FAQ는 최근 몇 년 동안 주목받고 있는 ‘Textual Similarity’ 작업과 공통점이 있다.

AQP는 약 25,000개의 질문을 받았으며 약 2,500명의 사용자 피드백을 수집하였다. QA- 4FAQ는 약 470개의 FAQ에 대한 지식 기반의 평가 데이터를 구축하였다. 이 데이터는 각 FAQ에 대한 답변과 범주 및 태그 정보가 포함되어 있다. 또한 참가자를 평가하기 위해서 수동으로 노이즈를 포함하였다.

QA- 4FAQ는 학습 데이터를 제공하지 않는다. AQP는 AQP Risponde가 사용자 피드백 없이 좋은 성능을 제공할 수 있어야 한다고 제시하고 있다. 따라서 비지도 학습 시스템을 구축해야 한다.

FAQ 데이터는 CSV 형식으로 제공된다. 각 FAQ는 id, question, answer, tag가 포함되어 있다.

바. SENTIPOLC - SENTIment POLarity Classification

트위터와 같은 온라인 소셜 네트워크 및 마이크로 블로그 플랫폼에서 스트리밍 되는 엄청난 양의 정보가 점점 많은 종류의 연구원과 실무자들의 관심을 끌고 있다. 특히 소셜 미디어의 언어학적 분석은 여러 연구 주제로 자리 잡았다.

Sentiment Analysis (SA)는 비즈니스 서비스를 제공하는 소셜 미디어 모니터링 플랫폼에서 중요한 정보로 인식되고 있다. 실제로 소셜 미디어의 사용자 제작 컨텐츠는 기업에서 고객의 요구와 선호도를 직접 활용할 수 있는 중요한 정보이다. 예를 들어 트위터는 특정 주제나 제품에 대해 사람들의 잡단 의견을 파악할 수 있는 귀중한 리스소 창고이다.

전반적으로 트위터에서 표현되는 정서를 추출하는 것은 정치적 정보나 비상 사태에 대한 정서, 대중의 관심과 기분 등을 분석하기 위해 진행되어 왔다.

SENTIPOLC에서는 트위터 글 중 이탈리아 텍스트에 중점을 두고 있다. 주요 과제는 트위터 텍스트의 정서의 극성 분류이다. 트위터로 표현된 감성은 일반적으로 긍정, 부정 또는 중립으로 분류된다. 하지만 일부 긍정과 부정이 혼합되어 사용될 수도 있다.

트위터 텍스트의 감성 분석에서 일부 문제점은 트위터를 이용한 대화에서는 반어적인 표현이 다수 사용된다는 점이다. 반어적인 표현은 텍스트의 극성을 반전시키기 때문에 시스템의 정확성을 낮출 수 있다. 이러한 문제는 비유적인 언어 표현에서 함께 연구되었다. SENTIPOLC는 이 문제에 대한 해결을 위해 반어적인 표현이 포함된 데이터를 사용하며 참가자들은 반어적 표현이 포함된 텍스트에 대응할 수 있는 시스템을 구축해야 한다.

SENTIPOLC는 이탈리아어 트위터의 감성 극성을 분류할 때 세 가지의 하위 작업으로 분류하여 진행된다. 참가자는 그 중 하나의 하위 작업에 참여할 수 있다. 첫 번째는 주관 분류로 트위터 메시지가 객관적인지 혹은 주관적인지를 판단한다. 두 번째는 극성 분류로 주어진 트위터 메시지가 긍정인지 부정인지 중립 또는 복합된 감성인지 결정한다. 마지막으로는 반어적 표현 탐지로 주어진 메시지에 표현된 반어적 표현을 판단한다.

SENTIPOLC는 NEEL-IT와 공동 작업을 수행하고 있다. 일반적으로 트위터에서 엔티티에 대해 관심을 표하고 이때 정서 표현이 함께 나타나는 경우가 많다. 전자 상거래가 일상 생활의 일부가 되었고, 소셜 미디어 플랫폼이 마케팅 및 잠재 고객의 신뢰를 키우는 새로운 장소로 간주됨에 따라 트위터의 의견 추출은 큰 관심을 받고 있다. 따라서 엔티티와 관련있는 감성 표현 정보를 추출하는 것은 중요한 문제로 인식되고 있다. 이에 SENTIPOLC는 엔티티 기반의 감성 분석 시스템을 구축할 수 있는 장기 목표의 첫 번째 단계이다.

테스트 데이터는 트위터에서 추출한 짧은 문서가 포함되어 있다. Entity

Linking 작업에서 사용되는 데이터를 공통으로 사용하며 감성 분석을 위한 주석을 추가적으로 부착하여 공통 데이터 세트를 만든다.

제공되는 데이터에는 정치적 또는 비정치적 태그가 포함되어 있다. 참가자가 이 정보를 사용하는 것은 자유지만 실제 평가에는 사용되지 않는 정보이다. 단, 이 정보를 수정해서는 안 된다. SENTIPOLC의 데이터 정보는 <표 3-8>과 같다.

<표 3-8> SENTIPOLC 데이터 정보

idtwitter	트위터 상태 ID API에 의한 실제 트위터 ID 정보
subj	주관적인 트위터는 subj = 1. 객관적인 트위터는 subj = 0.
opos	긍정 전체 극성(문장 전체에 대한 극성) 긍정 극성이 나타나는 트위터는 opos = 1; 긍정 극성이 없는 트위터는 opos = 0.
oneg	부정 전체 극성(문장 전체에 대한 극성) 부정 극성이 나타나는 트위터는 neg = 1; 부정 극성이 없는 트위터는 neg = 0.
iro	반어적인 트위터는 iro = 1, 그 외는 iro = 0.
lpos	긍정 리터럴 극성(문장 전체와는 다를 수 있는 부분적 극성) 긍정 극성이 나타나는 트위터는 pos = 1; 긍정 극성이 없는 트위터는 pos = 0.
lneg	부정 리터럴 극성(문장 전체와는 다를 수 있는 부분적 극성) 부정 극성이 나타나는 트위터는 neg = 1; 부정 극성이 없는 트위터는 neg = 0.
top	정치적 주제 정치적 주제가 명시된 트위터는 top = 1, 사회적 정치 주제가 명시된 트위터는 top = 2, 그 외는 top = 0.
text	트위터 텍스트

그리고 주석을 부착할 때는 다음과 같은 유의사항이 따른다.

- 객관적인 트위터는 극성이나 반어적 표현이 없으므로 subj =0, opos =0, oneg =0, iro =0, lpos =0, lneg =0 이다.

- 주관적이고 비반어적 표현 트위터는 긍정과 부정이 함께 나타날 수 있기 때문에 opos =1 과 oneg =1이 함께 나올 수 있다.
- 반어적 표현 트위터는 항상 주관적이며 하나의 정의된 극성을 가져야하므로 iro =1은 동일한 값을 갖는 opos와 oneg의 조합은 나올 수 없다.
- 반어적 표현이 없는 주관적인 트위터는 iro =0이고, 전체 극성(opos와 oneg)과 리터럴 극성(lpos와 lneg)은 항상 일관되게 주석이 붙는다.(즉, opos = lpos, oneg = lneg)

3. TAC

정보 기술 연구에 대한 활동이 활발해짐에 따라 텍스트 분석에 대한 관심이 커지고 있다. 텍스트 분석 회의 (TAC : Text Analysis Conference) [19] 는 자연어처리 기술의 대규모 평가를 위한 인프라를 제공하는 일련의 워크샵이다.

TAC의 주된 업무는 NLP 방법론의 대규모 평가에 필요한 인프라를 제공함으로써 자연어 처리 커뮤니티 내의 연구를 지원하는 것이다. TAC의 주요 목적은 평가 결과를 통해 최첨단 기술을 발전시키는 데 있다. 특히 TAC 워크샵은 다음과 같은 목적을 지닌다.

- 대형 공통 테스트 컬렉션을 기반으로 하는 NLP 연구를 촉진시킨다.
- NLP에 대한 평가 방법론 및 측정 방법을 개선한다.
- 연구 아이디어 교환을 위한 포럼을 만들어 산학간의 의사 소통을 증진시킨다.
- 실제 문제에 대한 NLP 방법론의 실질적인 향상을 입증함으로써 상용 제품 기술 이전을 가속화한다.

TAC는 국립 표준 기술 연구소(NIST; National Institute of Standards and Technology)의 정보 액세스 부서(IAD; Information Access Division)에서 그룹을 구성한다. 2008년에 시작된 TAC는 텍스트 요약을 위한 NIST의 문서 이해 회의(DUC; Document Understanding Conference)와 텍스트 검색 회의(TREC; Text Retrieval Conference)의 질의 응답 트랙에서 발전하였다.

TAC에서 진행되는 트랙은 여러 가지 용도로 사용된다. 트랙은 새로운 연구 분야를 실험하기 위한 인큐베이터 역할을 수행한다. 트랙의 기본은 실제로 겪고 있는 문제가 무엇인지 정의하고, 해당 문제를 해결할 수 있는 연구를 지원하는데 필요한 인프라를 만든다. 또한 더 많은 그룹의 연구 관심사와 일치하는 작업을 제공하여 TAC를 더 넓은 커뮤니티에 이목을 끌도록 한다. TAC 자문위원회는 트랙 제안서를 토대로 해당 연도의 TAC에서 실행될 트랙을 결정한다.

각 TAC 주기마다 NIST는 각 트랙에 대한 테스트 데이터를 배포한다. 참가자는 데이터에서 자체 NLP 시스템을 실행하고 결과를 NIST에 제출한다. NIST는 개별 심사를 정리하고 정확성을 판단하여 결과를 전달한다. 데이터 및 평가 결과는 TAC 주기가 끝날 때까지 비참가자에게 제공되지 않는다. TAC 주기는 참가자들이 경험을 공유하고 미래의 과제/평가를 계획하는 워크숍이다. 워크숍은 일반적으로 미국 추수감사절 연휴 마지막 1주일 동안 NIST에서 열린다.

2016년에 진행된 TAC는 “Knowledge Base Population(KBP)”라는 주제로 진행되었다. TAC KBP의 목표는 구조화되지 않은 텍스트에서 지식 기반(Knowledge Base) 정보를 채우는 기술을 개발하고 평가하는 것이다. 현재 지식 기반 개요는 특정 개인(PER), 조식(ORG), 지정학적 엔티티(GPE), 위치(LOC), 시설(FAC)로 구성된다. 트랙에 접수된 Cold Start KBs는 비워진 지식 기반 정보를 채우는 시스템의 정확성을 평가한다. 이러한 지식 기반 정보 추출은 크게 두 가지 작업으로 분류된다. 첫째는 문서에서 PER, ORG, GPE, LOC, FAC 엔티티를 이름과 분류 정보를 추출하고 각 엔티티를 지식 기반 노드에 연결하는 것이다. 둘째는 Cold Start 문서 목록을 검색하여 특정 항목에 대한 슬롯의 값을 채우는 것이다.

KBP 2016의 트랙은 정보가 구조화된 지식 기반에 입력으로 적합하도록 구조화되지 않은 텍스트에서 정보를 추출하는 것을 목표로 한다. 2016년의 모든 트랙은 다국어(영어, 중국어, 스페인어)로 진행된다. Cold Start EDL, SF, KB 구축 작업은 여러 언어로 진행된다.

모든 KBP 2016 작업의 원본 데이터는 최대 90,000개의 영어, 중국어, 스페인어 뉴스와 토론 포럼의 문서로 구성되어 있다. 이 중 약 500개의 핵심 문서에는 지침에 따라 엔티티, 관계 및 이벤트 정보가 주석으로 추가되어 있으며, 이벤트

및 감성 트랙에 대한 시스템 응답을 평가하는데 사용된다.

제출 후 평가 절차는 Cold Start SF/KB 구축 작업에 대한 시스템 응답을 평가하고 이벤트 인수 추출 및 링크 작업에서 이벤트 프레임을 문서화 한다.

Cold Start KBP는 KB를 구성할 엔티티 및 관계에 대한 사전 정의된 스키마와 주어진 문서 모음을 사용하여 지식 기반을 구축하는 작업이다. Cold Start KB에서 사용하는 엔티티 유형은 5 가지로 PER, ORG, GPE, LOC, FAC이다. 또한 KBP 슬롯에 해당하는 2진 관계 유형으로 구성된다.

주어진 문서 모음을 감안할 때, Cold Start KB 구축 시스템은 컬렉션의 특정 PER, ORG, GPE, LOC, FAC 엔티티의 모든 이름을 찾아 각 엔티티에 대해 지식 기반 노드를 만들고 각 이름들을 연결해야 한다.

문서 수집에서 입증된 모든 Cold Start 관계는 주제, 술어 및 대상에 대한 오프셋으로 구성되어 신뢰도 및 출처와 함께 KB에 기록되어야 한다. 객체가 Cold Start의 PER, ORG 또는 GPE 엔티티와 관련된 관계일 경우 시스템은 올바른 엔티티를 나타내는 KB 노드에 연결해야 한다.

KB에 동일한 주체-술어-개체 트리플의 인스턴스가 포함되어 있는 경우 가장 신뢰도가 높은 값을 갖는 트리플이 평가되고 낮은 신뢰도를 갖는 트리플은 평가되지 않는다.

Cold Start KB 구축 작업에서 시스템은 슬록 채우기 및 엔티티 추출/링킹 두 가지 측면에서 평가된다. 엔티티 검색 및 링크 연결은 90,000개의 문서 중 약 500개의 핵심 문서가 선택되면 엔티티와 엔티티 노드는 KBP 2016 엔티티 검색 및 링크 트랙의 클러스터링 매트릭을 사용하여 평가된다. 슬롯 채우기의 평가 쿼리에는 문서 컬렉션의 명명된 엔티티와 하나 이상의 관계 유형 시퀀스가 포함된다. 슬롯 채우기는 전통적인 영어 슬롯 채우기와 동일한 방법으로 평가된다.

2016 Cold Start SF/KB 트랙은 다음과 같은 사항을 따른다.

- 영어 이외에도 중국어, 스페인어가 문서에 포함되어 있다. 교차 언어 작업의 경우 3 언어의 모든 조합에서 엔티티를 추출해야 한다. 또한 평가를 위해서 단일 언어 데이터를 함께 제공한다.
- PER, ORG 그리고 GPE 엔티티 유형 외에도 KB 구축 시스템은 LOC와

FAC 엔티티에 대한 이름을 반환해야 한다.

- 이름을 추출하는 것 외에도 KB 구축 시스템은 특정 PER, ORG, GPE, LOC, FAC 엔티티의 명목상 멘션을 추출하고 링크를 생성해야 한다.

참가자의 시스템의 주석 부착은 엔티티 멘션을 나타내는 정확한 텍스트 문자열(문자, 숫자, 구두점, 기호 등을 포함할 수 있음)을 선택하도록 요구한다. 정확성은 엔티티의 유형과 올바른 클러스터 내의 그룹 정보를 기준으로 측정한다. 참가자는 주석 부착을 위해 각 문서 내의 컨텍스트를 참조하는 것이 좋다.

4. 국어 정보 처리 시스템 경진 대회

국어 정보 처리 시스템 경진 대회[20]는 문화체육관광부와 국립국어원이 주최하고, 한국정보과학회 언어공학연구회가 주관하여 진행되는 경진대회이다. 국어 정보 처리 시스템 경진대회는 2009년도부터 시작되어 2016년 8회 차를 맞이하였다. 2016 국어 정보 처리 시스템 경진 대회는 2016년 10월 7일에 동아대학교 부민 캠퍼스에서 진행되었다. 국어 정보 처리 시스템 경진 대회는 지정 분야와 일반 분야로 나누어 신청 가능하다. 지정 분야는 2년마다 주제를 바꾸어 진행되며 심사 시에 지정 분야와 일반 분야를 구분하여 진행된다.

2016년도 지정 분야는 개체명 인식 시스템 개발 및 적용으로 사용자 인터페이스는 제외하여 심사가 진행되었다. 일반 분야는 지정 분야를 제외한 모든 분야의 한국어 처리 도구와 한글 활용 및 학습에 관련된 소프트웨어가 대상이 되었다. 참가를 위해서는 실행 동영상과 실행 파일, 사용 설명서를 함께 제출해야된다. 참가 자격은 개인 또는 팀으로 참가 가능하며, 대학, 고등학교, 연구소 등 모든 단체에서 참가 가능하다.

2016 국어 정보 처리 시스템 경진 대회 심사 내역에 따른 수상은 대상(1팀), 금상(1팀), 은상(1팀), 동상(2팀), 장려상(2팀)이다. 대상 팀에게는 문화체육관광부 장관상 및 상금 300만원이 수여된다.

2016 국어 정보 처리 시스템 경진 대회에 지정 분야는 총 6팀이 최종 본선에 진출하였고, 지정 분야 역시 6팀이 본선에 진출하였다. 지정 분야는 개체명 인식 및 분류기의 성능을 기준으로 평가가 진행되었다. 일반 분야는 시스템의 성능과

아이디어, 활용 방안 등 다양한 관점에서 평가가 진행되었다. 2016 국어 정보 처리 시스템 경진 대회 결과 대상은 서울대학교의 KoNER가 수상하였다.

제4장 언어지능 기술이 적용된 실제 사례 조사

제1절 Virtual Assistance

Virtual Assistance[21] 는 언어지능 기술이 적용된 대표적인 사례 중의 하나로 음성인식을 이용한 가상비서 서비스를 들 수 있다. Apple의 Siri가 2011년 발표된 후로, Google의 Google Now 그리고 Microsoft의 Cortana는 현재 모바일 디바이스에서 사용가능한 대표적인 3개의 가상비서 서비스이다.

그 중 Siri는 SRI International에서 DARPA의 지원으로 시작된 프로젝트였다. DARPA의 지원이 끝난 후 Siri International로 독립하였고, 2010년 AppStore에 무료 앱으로 등록한 후 2500만 달러에 달하는 수익을 거두었다. 직후 Apple은 Siri를 약 2억 달러에 인수하였다. 그리고 Apple은 Siri를 iOS에 포함시켰다. 무료 앱이었던 Siri는 2011년 9월 앱스토어에서 내려왔고, 그해 10월, Apple이 iPhone 4S와 함께 Siri를 공개하였다.

이 이전의 사용자들은 음성 인터페이스에 대해서 부정적인 견해를 갖는 경우가 많았다. 그러나 Siri를 비롯한 요즈음의 가상비서들은 그 편견을 깨고 있다. 그 뒤에는 자연어처리 모듈의 성능 향상이 자리매김하고 있다.

일반적인 음성 비서의 프로세스는 먼저 디바이스에서 음성을 서버로 보내는 것으로부터 시작한다. 서버에서는 높은 성능의 STT(Speech To Text) 프로그램을 통해 이를 텍스트로 변환하고, 텍스트를 분석한 후 수행할 작업을 결정한다. 텍스트 분석 단계는 또 다시 높은 성능의 NLU(Natural Language Understanding) 프로세스가 동작할 것이다. 이렇게 함으로써 기업은 많은 데이터를 데이터 센터에 축적할 수 있고, 모든 사용자에게 질높은 사용자 경험을 선사할 수 있다.

프로세스에 대해 조금 더 이야기하자면, 일반적으로 특정 입력에 대해 어떻게 행동해야 한다는 정해진 규칙이 없다. 상황에 따라 달라질 수 있고, 사용자의 특성에 대해서도 마찬가지이다. 대신에 가상비서 서비스들은 대화를 이해하고 사용자가 원하는 사항을 알아낸다. 이해의 단계도 비단 단어에 머무는 것이 아니라, 문장 자체를 분석하고 행동한다. 무엇보다도 가장 중요한 것은 Siri가 Siri의 데이터 센터의 리소스를 활용해 대화를 관리한다는 것이다.

Siri는 또한 사용자로부터 배우기도 한다. Siri는 사용자로 하여금 사용자의 악

센트를 더 잘 이해할 수 있도록 Siri를 개선하는 방안을 제공한다. 이 또한 서버에서 수백만 건 이상의 음성을 검토하고, 개선할 부분을 찾는다. 그 예로, Siri가 처음 출시되었을 때 'Texas'의 발음은 'Tee X'와 같이 들렸다. 그러나 약 일주일 후에는 그것이 개선되어 훌륭하게 'Texas'를 발음할 수 있게 되었다.

제2절 Social Network analytics tools

언어지능 기술을 이용한 대표적인 사례는 SNS 데이터 분석이 있다. SNS 데이터는 문법이 잘 지켜지지 않는 경우도 많고, 많은 양의 분석을 빠르게 수행하여야 의미가 있기 때문에 분석속도와 정확도가 상당히 중요하게 작용하는 응용분야라고 말할 수 있다.

소셜 애널리틱(Social Analytics)은 사람, 화제, 아이디어 간 상호 작용과 연계 결과를 측정, 분석하는 과정을 기술하는 용어를 말한다. 소셜 필터링, 소셜 네트워크 분석, 정서분석, 소셜미디어 분석기법과 같은 전문적인 분석 테크닉 다수를 포괄하는 상위용어이다.

소셜 네트워크 분석은 조직 내 커뮤니케이션 패턴과 정보흐름을 파악할 수 있는 유용한 분석이고 주요 관심사를 키워드로 등록하여 블로그나 트위터상 올라오는 텍스트 데이터를 수집하여 분류하고 분석함으로써 데이터 속에 숨겨진 패턴 및 의미를 찾아내는 데이터 마이닝이 가능한 솔루션이다.

소셜 네트워크 분석은 개체 간 관계 데이터와 개체 속성 데이터의 '분석대상'을 이용하여 관계 구조 속의 Business Insight 발굴 및 비지니스 의사결정 지원을 하는 '분석결과'를 나타내는 것을 말한다.

SNS 데이터 분석은 먼저 목표를 설정하는 것에서부터 시작한다. 그리고 설정한 목표를 얼마나 달성했는지 SNS 데이터 분석을 이용해 확인한다. 예를 들면 “like 또는 share의 횟수”, “댓글의 수”, “링크의 클릭 수(가장 중요한 것)” 와 같은 지표를 사용할 수 있다.

이번 절에서는 SNS 데이터 분석 서비스를 소개하고 각각의 개요와 특징을 기술한다.

1. Keyhole

Keyhole[22]은 트위터 트윗과 인스타그램 포스트(이하 포스트)를 다방면으로 분석한 후에 그 결과를 제공해주는 서비스이며, 웹페이지로 서비스하고 있다. Keyhole에서 제공하는 서비스들은 다음과 같다.

가. Hashtag & Keyword 분석

Keyhole에서 제공하는 Hashtag & Keyword 분석 서비스는 Hashtag 혹은 Keyword가 포함된 총 포스트의 수, 포스트의 총 도달 수, 노출 수 분석을 제공한다. 분석은 기본적으로 웹페이지로 제공되지만, PDF 또는 Excel로도 다운로드할 수 있다.

영향력 있는 유저에 대한 분석을 제공한다. 지역별 / 키워드별 영향 요인 분석을 제공하여 마케팅에 도움을 준다.

연관 해쉬태그 : 대쉬보드에서는 해당 Hashtag 또는 Keyword와 함께 사용되는 Hashtag를 확인할 수 있다.

나. 사용자 계정 분석

Optimize Posting : Keyhole의 사용자 계정 분석은 계정의 포스트에 대한 일괄적인 분석을 제공한다. 언제 포스트를 작성하고 언제 그 포스트를 공유해야할 지에 대한 통찰을 제공한다. 모든 정보는 차트와 그래프로 정리되어 있어 쉽게 파악할 수 있다.

팔로워 추이 : Keyhole은 일일, 월별, 연간 팔로워 활동에 대한 데이터를 제공하여 잠재고객의 수를 관리할 수 있도록 도와준다. 포스트의 수와 참여 통계를 이용해 사용자의 관심이 어떻게 변해가고 있는지에 대한 추이를 제공한다.

자동화된 KPI 보고서 : Keyhole은 당신의 키워드에 대한 실시간 성장, 참여, 최적화 보고서를 제공한다. 이러한 보고서로부터 제품에 대한 마케팅 전략을 수립할 수 있다.

※이미지출처 : http://keyhole.co/

[그림 4-1] Keyhole의 분석 예

다. 확인할 수 있는 정보

Keyhole에서 확인할 수 있는 정보는 다음과 같다.

<표 4-1> Keyhole에서 확인할 수 있는 정보

구 분	내 용
Timeline	시간별 작성된 연관 포스트의 수
Top Posts	리트윗 횟수와 Like 횟수를 합하여 점수를 매긴 후, 높은 점수를 받은 상위 포스트
Related Topics	연관 포스트에서 분석한 연관 단어 워드 클라우드
Most Influential	연관 포스트로부터 분석한 가장 영향력있는 계정
Recent Users	연관 포스트를 최근 작성한 계정
Top Sites	연관 포스트 내에 포함된 URL/도메인 순위와 갯수
Location	연관 포스트가 작성된 나라
Share of Posts	원 포스트와 리트윗/답변의 비율
Sentiment	포스트의 감성분석 결과
Top Sources	데스크탑 웹, 모바일 웹, 모바일 OS별 작성 비율
Demographics	포스트의 성별 비율

2. Signal Media

런던에 본사를 둔 Smart Media[23]는 75,000개 이상의 온라인 뉴스 소스, 350만개의 블로그, 100개의 소셜 네트워크 연구 출판물 및 기타 비정형 데이터를 크롤링하여 실시간 정보 수집 및 필터링을 자동화하는 서비스를 구축했다.

Signal은 미디어 모니터링, 정보 검색 및 데이터 분석 소프트웨어 플랫폼이다. 매일 300만 개가 넘는 언론 매체, 블로그 및 소셜 미디어 플랫폼을 분석하여 세계 뉴스를 추적할 수 있는 가장 진보되고 저렴한 솔루션을 제공한다.

대상 시장과 사용자는 홍보 및 통신 회사나 금융, 투자 자금 및 기타 법인체, 다양한 분야와 전문 서비스를 가지는 미디어 및 콘텐츠 마케터이다.

거의 3년 전에 설립된 영국의 신생기업은 NLP와 기계학습을 사용하여 수백만 개의 온라인 뉴스 소스를 처리하고 분석하는 핵심 텍스트 분석 플랫폼을 구축하는데 주력하여 6개월 전 상업 서비스를 시작했다.

Feedly, Google Alerts 및 Flipboard와 같은 소비자가 이용할 수 있는 무료 툴과 Precise, Gorkana 및 Cision와 같은 계약을 맺고 기업들이 이용할 수 있는 툴 사이에서의 차이점은 전자는 기능이 제한되어 있으며, 후자는 다루기 힘들고 변함없이 매우 비싸다.

일부 기존의 미디어 모니터링 기술과는 달리 Signal은 키워드 검색 만 하는 것이 아니다. 데이터를 필터링하여 “가치를 추출“하는 것을 목표로 하는 “일련의 텍스트 분석 구성 요소“에 중점을 둔다.

Signal은 정교한 자연 언어 처리, 엔티티 인식, 중복 제거 및 클러스터링이 기능을 사용하여 맞춤식으로 매우 정확한 결과를 실시간으로 보장한다. 인터페이스가 사용하기 쉽기 때문에 교육 및 외부 분석가에 대한 고객의 의존도가 줄어들어 사용자 비용이 절감된다.

Signal의 플랫폼은 사용자가 관심 있는 주제를 정의하고 해당하는 주제에 있어야 할 것과 하지 말아야할 것의 알고리즘을 학습할 수 있는 기계 학습 기반 주제 분류 시스템을 제공한다. 이 플랫폼의 또 다른 기술을 ’엔티디 인식(Entity recognition)’이라고 하며 예를 들어 ’apple’로 했을 경우 이 단어가 회사 ’apple’과 과일 종류의 ’apple’를 지능적으로 구별하기 위해 기사 내의 언어 컨텍스트를 분석한다.

c-level 임원에서 일선 직원까지 누구나 사용할 수 있는 직관적인 도구로 설계했다. 우리의 플랫폼은 매우 사용하기 쉽고 커스터마이징이 쉽기 때문에 파악하기 쉽다.

Signal의 플랫폼에는 요약 기능이 있어서 텍스트에서 중요한 사실을 추출하여 사용자에게 다이제스트를 제공할 수 있다.

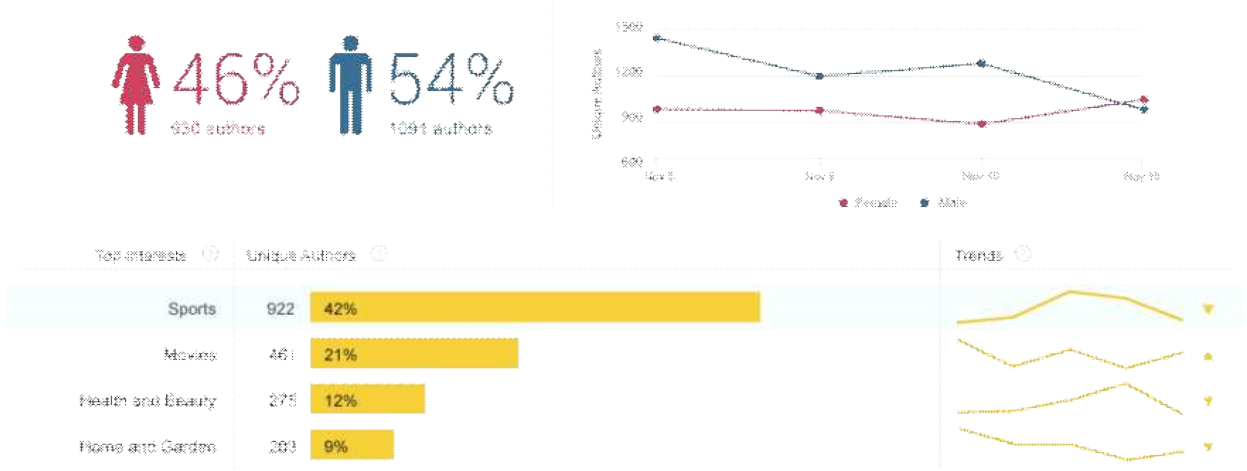
Signal은 “사용자에게 앞서 나가기 위해서 실시간으로 세상을 추적하십시오.” 라는 문구로 시선을 끌며 실시간(Real-time), 제한 없는 검색 및 콘텐츠, 하이퍼 관련, 단순함(Simple)의 장점을 내세운다.

실시간성은 실시간으로 뉴스가 받은 편지함으로 전송되고 가능한 빠른 통계를 제공한다. 제한 없는 검색 및 콘텐츠는 사용자가 원하는 뉴스와 간행물을 각각 고정된 비용을 지불하고 추적할 수 있다. 하이퍼 관련은 인공지능이 뉴스 검색작업을 수행하게 되는데, 사과(apple)과 기업 Apple을 선별할 수 있는 수준으로 보인다. 단순함은 사용자가 조직내의 시스템과 통합한 형태의 직관적인 플랫폼 내에서 보고서를 작성하고 전송할 수 있다.

Signal의 분석은 기본적으로 정보를 시각화하는 데에 중점을 두고 있다. 세계의 방송 및 라디오 채널에서부터 신문 인쇄, 온라인 미디어, 법률 및 규제 문서 등의 컨텐츠를 Signal 시스템에 전송하면 인공지능이 분석하고 분류하는 작업에 착수한다. 자연어 처리는 언급된 회사의 이름에서 논의 되는 주제 이르기 까지 모든 관련 정보를 자동으로 식별하고, 컨텍스트 정보를 사용하여 관련성 있는 스토리를 함께 클러스터링한다. 이후, 각 기사의 모든 정보에 태그를 추가하고 실시간 메일 알람 또는 Signal의 검색 기능을 통해 사용자에게 전달한다.

3. Brandwatch

Brandwatch[24]는 소비자의 의견에 대해 깊은 통찰력을 얻을 수 있도록 도움을 주는 서비스이다. 예를 들면, 무인자동차에 대한 최신 의견과 세계에서 가장 좋아하는 아이스크림 맛, 사람들이 이번 휴가 시즌에 어떤 계획을 세우고 있는지 파악하는 데 도움을 줄 수 있다. Brandwatch를 사용하면 고객 피드백을 수집하고 브랜드, 제품 및 마켓 플레이스에 영향을 주는 추세를 쉽게 발견 할 수 있다.



※이미지출처 : https://www.brandwatch.com/

[그림 4-2] Brandwatch의 분석 예

Brandwatch에서 제공하는 Brandwatch Analytics 서비스는 브랜드가 비즈니스에서 변화하는 양상을 소셜 대화를 통해서 추출한다. 이 기능은 소셜 데이터에서 더 많은 답변 찾기 및 매일 온라인에서 발생하는 수십억 가지 대화에 컨텍스트를 추가한다. 모든 공개 웹의 대화를 즉시 액세스 할 수 있는데, 액세스 범위는 약 8,000만 개까지 가능하다. 서비스되는 브랜드는 전 세계의 약 1,000개 이다.

Brandwatch에서 제공하는 서비스는 구글의 데이터를 이용한다. 구글은 모든 자체 데이터를 크롤링하고 저장하며 시장에서 간단하지만 정확한 query builder를 보유하고 있어서 스팸 및 중복이 없는 고품질의 유용한 데이터를 제공한다. 고급 필터링, 태그 지정 및 분류, 유연한 차트 작성, 주제 및 정서 분석, 많은 데이터 관리 기능 등 분석을 위한 기능들도 제공한다. 또한, 워크플로우 기능을 통한 사용자 제약을 해소했다. 비디오나 웹을 통해 사용자 매뉴얼도 제공한다.

제5장 언어처리 기술 수준

제1절 공개된 SW의 종류와 성능

1. CoreNLP

Stanford CoreNLP[25]는 영어 등 6개 언어를 지원하는 자연어처리 분석도구 모음이다. 입력된 문장으로부터 단어 원형 추출, Parts of speech, 개체명, 구문 분석, Coreference resolution, Sentiment analysis, Relation extraction을 수행할 수 있다. 옵션으로 간편하게 어떤 모듈을 사용할 지 선택할 수 있으며, stanford 모델이 아닌 다른 모듈과도 함께 사용할 수 있도록 지원하고 있다.

<표 5-1> CoreNLP 개요

구 분	내 용
License	GNU General Public License v2 (Dependency Parser의 경우 상용라이센스 협의 가능)
Requirement	Java 1.8+
Reference	The Stanford CoreNLP Natural Language Processing Toolkit (http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf)

<표 5-2> CoreNLP의 지원 언어 내역

ANNOTATOR	AR	ZH	EN	FR	DE	ES
Tokenize / Segment	O	O	O	O	X	O
Sentence Split	O	O	O	O	O	O
Part of Speech	O	O	O	O	O	O
Lemma	X	X	O	X	X	X
Named Entities	X	O	O	X	O	O
Constituency Parsing	O	O	O	O	O	O
Dependency Parsing	X	O	O	O	O	X
Sentiment Analysis	X	X	O	X	X	X
Mention Detection	X	O	O	X	X	X
Coreference	X	O	O	X	X	X
Open IE	X	X	O	X	X	X

<표 5-3> CoreNLP의 모듈별 성능과 레퍼런스

모듈	성능	Reference
POS	Overall 96.86%	Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger (http://nlp.stanford.edu/~manning/papers/emnlp2000.pdf)
NER	CMU 91.85%	Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. (http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf)
Parser	UAS 92.0% LAS 90.7%	A Fast and Accurate Dependency Parser using Neural Networks (http://cs.stanford.edu/people/danqi/papers/emnlp2014.pdf)
Coreference Resolution	B3-F1 64.78%	The Life and Death of Discourse Entities: Identifying Singleton Mentions (http://nlp.stanford.edu/pubs/discourse-referent-lifespans.pdf)
Sentiment Analysis	Negated Positive 71.40% Negated Negative 81.80%	Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank (http://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf)
Bootstrapped pattern learning	-	Improved Pattern Learning for Bootstrapped Entity Extraction (http://nlp.stanford.edu/pubs/gupta14evalpatterns.pdf)
Open IE	F1 28.30%	Leveraging Linguistic Structure For Open Domain Information Extraction (http://nlp.stanford.edu/pubs/2015angeli-openie.pdf)

2. Natural Language Toolkit(NLTK)

NLTK[26]는 자연어처리를 위한 파이썬 라이브러리이다. NLTK 프로젝트는 200년에 시작되어 수십 명의 자발적인 개발자들의 참여로 진행되었다. <표 5-4>는 NLTK의 개요를 나타낸다.

<표 5-4> NLTK 개요

구 분	내 용
License	Apache 2.0
Requirement	Python 2.7 or 3.2+
Reference	Bird, Steven, Edward Loper and Ewan Klein (2009). Natural Language Processing with Python.

NLTK는 2015년엔 초기 지원 범위로부터 ‘의존성 분석’, ‘기계 번역’, ‘감정 분석’, ‘트위터 처리’의 기술을 추가하였다. 2016년 12월 기준으로는, 코퍼스와 모델을 포함한 총 105개 모듈을 제공하고 있으며, 구문분석과 개체명 인식, 감성 분석, 토크나이저 등의 모델을 제공한다.

가. Dependency Parsing

NLTK에는 총 3개의 종속성 구문분석기(Dependency Parsing)[27]가 구현되어 있고, 코퍼스를 읽기 위한 도구, 구문 트리를 시각화하기 위한 도구 등이 포함되어 있다. 구문분석은 스웨덴의 Peter Ljunglöf팀[28]이 개발하였다.

<표 5-5> NLTK의 Dependency Parser 모듈

분류	모듈명	레퍼런스
Existing functionality	projective dependency parser	https://github.com/nltk/nltk/blob/develop/nltk/parse/projectivedependencyparser.py, Eisner 1996 Model C
	Non-projective dependency parser	https://github.com/nltk/nltk/blob/develop/nltk/parse/nonprojectivedependencyparser.py
	MaltParser	https://github.com/nltk/nltk/blob/develop/nltk/parse/malt.py
Tools	Corpus reader for the CoNLL 2007 shared task 10% sample of the Penn Treebank Corpus	https://github.com/nltk/nltk/blob/develop/nltk/corpus/reader/dependency.py
	Nivre's arc-eager and arc-standard algorithms	https://github.com/nltk/nltk/blob/develop/nltk/parse/transitionparser.py
	Visualization of dependency trees	https://github.com/nltk/nltk/blob/develop/nltk/parse/dependencygraph.py#L140
Planned	Add interface to Universal Dependencies Dataset	https://github.com/nltk/nltk/issues/875
	Adapt Honnibal's shift-reduce dependency parser	https://github.com/nltk/nltk/issues/694#issuecomment-69795719

나. Sentiment Analysis

Sentiment Analysis[29]를 위한 NLTK의 모듈들은 다음과 같다. 현재는 일부 사전과 데이터만이 수집되어 있는 상태이다. 의미 분석은 북미의 Dan Garrette 팀[30]이 개발하였다.

〈표 5-6〉 NLTK의 Sentiment Analysis 모듈

분류	사전/코퍼스명	레퍼런스
사전	Opinion Lexicon by Bing Liu	Mining and summarizing customer reviews http://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf
	MPQA Subjectivity Lexicon	Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis http://people.cs.pitt.edu/~wiebe/pubs/papers/emnlp05polarity.pdf
	Harvard General Inquirer	The General Inquirer: A Computer Approach to Content Analysis http://psycnet.apa.org/psycinfo/1967-04539-000
코퍼스	Customer Review Dataset	Mining and summarizing customer reviews (http://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf)
	Pros and Cons Dataset	Mining Opinions in Comparative Sentences (http://www.cs.uic.edu/~liub/FBS/Coling-2008-camera-ready.pdf)
	Comparative Sentences	Identifying Comparative Sentences in Text Documents http://www.cs.uic.edu/~liub/publications/sigir06-comp.pdf Mining Opinion Features in Customer Reviews http://www.cs.uic.edu/~liub/publications/aaai04-featureExtract.pdf
	STS-Gold	Evaluation datasets for twitter sentiment analysis http://ceur-ws.org/Vol-1096/paper1.pdf

다음은 NLTK의 모듈 사용 예제이다. python을 실행하여 nltk를 import하고 'pos_tag' 함수에 단어 리스트를 전달하면 tuple형태의 형태소 품사 태그 부착 결과를 얻을 수 있다. 이외에도 'ne_chunk' 함수로 개체명을 추출할 수 있고, 'parsed_sents' 함수로 구문 분석 결과를 얻을 수 있다.

(1) Tokenizer

```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged[0:6]
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'), ('on', 'IN'),
('Thursday', 'NNP'), ('morning', 'NN')]
```

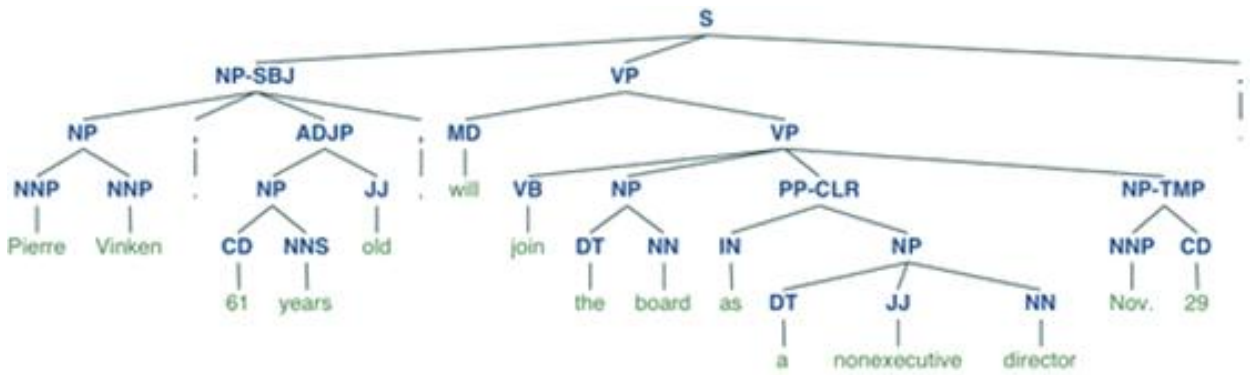
(2) Identify Named Entity

```
>>> entities = nltk.chunk.ne_chunk(tagged)
>>> entities
Tree('S', [('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'),
          ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN'),
      Tree('PERSON', [('Arthur', 'NNP')]),
          ('did', 'VBD'), ("n't", 'RB'), ('feel', 'VB'),
          ('very', 'RB'), ('good', 'JJ'), ('.', '.')])
```

(3) Display a parse tree

```
>>> from nltk.corpus import treebank
>>> t = treebank.parsed_sents('wsj_0001.mrg')[0]
>>> t.draw()
```

다음 [그림 5-1]은 NLTK의 구문 분석기를 이용한 결과를 parse tree로 표현한 예이다. ‘NP-SBJ’는 주어를 뜻하며 ‘VP’는 용언을 뜻한다.



※ 출처 : Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O’Reilly Media Inc.

[그림 5-1] NLTK parse tree display의 예

3. KoNLPy

KoNLPy[32, 33]는 서울대학교의 박은정, 조성준의 연구로 진행되었으며, 공개된 한국어 자연어처리 프로그램을 파이썬에서 쉽게 호출할 수 있도록 구현한 라이브러리이다. 현재 형태소분석기와 데이터 모듈이 작성되어 있으며 계속해서 발전중이다. 형태소분석기 라이브러리로는 한나눔, 꼬꼬마, 코모란, Mecab, Twitter 형태소분석기를 사용할 수 있도록 지원하고 있다. <표 5-7>은 KoNLPy의 개요를, <표 5-8>은 KoNLPy의 모듈과 관련 레퍼런스를 나타낸다.

<표 5-7> KoNLPy 개요

구 분	내 용
License	GNU General Public License v3 or later
Requirement	Java, Python
Reference	KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지 (http://dmlab.snu.ac.kr/~lucypark/docs/2014-10-10-hclt.pdf)

<표 5-8> KoNLPy의 모듈

구분	모듈명	URL
POS Tagger	Hannanum	http://konlpy.org/ko/latest/api/konlpy.tag/#module-konlpy.tag._hannanum
	Kkma	http://konlpy.org/ko/latest/api/konlpy.tag/#module-konlpy.tag._kkma
	Komoran	http://konlpy.org/ko/latest/api/konlpy.tag/#module-konlpy.tag._komoran
	Mecab	http://konlpy.org/ko/latest/api/konlpy.tag/#module-konlpy.tag._mecab
	Twitter	http://konlpy.org/ko/latest/api/konlpy.tag/#module-konlpy.tag._twitter

예제) POS tagging example using Mecab

```
>>> from konlpy.tag import Mecab
>>> mecab = Mecab()
>>> print(mecab.morphs(u'영등포구청역에 있는 맛집 좀 알려주세요.'))
['영등포구', '청역', '에', '있', '는', '맛집', '좀', '알려', '주', '세요', '.']
>>> print(mecab.nouns(u'우리나라에는 무릎 치료를 잘하는 정형외과가 없는가!'))
['우리', '나라', '무릎', '치료', '정형외과']
>>> print(mecab.pos(u'자연주의 쇼핑몰은 어떤 곳인가?'))
[('자연', 'NNG'), ('주', 'NNG'), ('의', 'JKG'), ('쇼핑몰', 'NNG'), ('은', 'JX'), ('어떤', 'MM'),
('곳', 'NNG'), ('인가', 'VCP+EF'), ('?', 'SF')]
```

4. Google SyntaxNet

SyntaxNet은 Google이 2016년 5월 12일 블로그를 통해 자연어 처리 시스템인 SyntaxNet을 오픈소스로 공개하였다. SyntaxNet은 2016년 12월 기준 40여 개의 언어에 대한 형태소 분석, 구문분석 모델을 제공하는 tensorflow의 프로젝트이다. 학습에는 Universal Dependencies Dataset이 사용되었고, 재학습할 수 있는 소스코드를 제공한다.

<표 5-9> Google SyntaxNet 개요

구 분	내 용
License	Apache 2.0
Requirement	Python2.7, pip, bazel, swig, protocol buffers, tensorflow, asciitree, numpy, mock
Reference	Globally Normalized Transition-Based Neural Networks (http://arxiv.org/abs/1603.06042) github (https://github.com/tensorflow/models/tree/master/syntaxnet)

<표 5-10> Google SyntaxNet의 성능[28]

Language	No. tokens	POS	fPOS	Morph	UAS	LAS
Spanish-Ancora	53594	98.28%	98.28%	97.82%	89.26%	86.50%
Spanish	7953	95.27%	-	95.74%	85.06%	81.53%
English-LinES	8481	95.34%	93.11%	-	81.50%	77.37%
English	25096	90.48%	89.71%	91.30%	84.79%	80.38%
Hindi	35430	96.45%	95.77%	90.98%	93.04%	89.32%
Arabic	28268	95.65%	91.03%	91.23%	81.49%	75.82%
Portuguese-BR	29438	97.07%	97.07%	99.91%	87.91%	85.44%
Portuguese	6262	96.81%	90.67%	94.22%	85.12%	81.28%
Russian-SynTagRus	107737	98.27%	-	94.91%	91.68%	87.44
Russian	9573	95.27%	95.02%	87.75%	81.75%	77.71%
Average	-	94.27%	92.63%	90.38%	81.12%	75.85%

5. Apache OpenNLP

Apache OpenNLP[34]는 자연어처리를 위한 도구이다. 토크나이저, 문장 분리, POS 태깅, 개체명 인식, 청킹, 구문 분석, Coreference resolution 모듈을 제공한다. 또한 Maximum entropy와 perceptron을 학습할 수 있도록 구현하여 제공하고 있다. Apache OpenNLP를 사용하기 위해서는 아래 (1)에서처럼 사용하고자 하는 model을 호출하여 사용한다. (2)는 model의 함수를 호출하여 파라미터로 전달한 문장을 자연어처리를 거친 문자열로 변환해 준다. (3)은 tokenizer 호출을 옵션과 파라미터를 이용하여 문서 입/출력 방식으로 동작하는 예시이다.

<표 5-11> Apache OpenNLP 개요

구 분	내 용
License	Apache 2.0
Requirement	JDK 5, Maven 3.0

(1) Example to open model

```
InputStream modelIn = new FileInputStream("lang-model-name.bin");

try {
  SomeModel model = new SomeModel(modelIn);
}
catch (IOException e) {
  //handle the exception
}
finally {
  if (null != modelIn) {
    try {
      modelIn.close();
    }
    catch (IOException e) {
    }
  }
}
Tool tool = new Tool(model)
```

(2) Example to process a task

```
String output[] = tool.executeTask("This is a sample text.");
```

(3) CLI Example

```
$ opennlp TokenizerTrainer ##will output help message
Usage : opennlp TokenizerTrainer[.namefinder|.conllx|.pos] [-abbDict path] ...
#Processing
$ opennlp Tokenizer lang-model-name.bin < input.txt > output.txt
```

<표 5-12> Apache OpenNLP의 모듈[35]

구분	설명	URL
Sentence Detector	문서에서 문장을 탐색	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#tools.sentdetect
Tokenizer	문장을 지정한 기준으로 토큰화	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#tools.tokenizer
Name Finder	개체명 인식 및 분류	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#tools.namefind
Document Categorizer	문서 카테고리 분류	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#tools.doccat
POS Tagger	형태소 품사 태거 부착	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#tools.postagger
Chunker	명사 또는 동사 그룹 청킹	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#tools.chunker
Parser	구문 분석	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#tools.parser
Coreference Resolution	상호 참조	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#tools.coref
Corpora	CONLL 코퍼스	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#tools.corpora
Machine Learning Algorithm	기계 학습 알고리즘	http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html#opennlp.ml

6. Apache Lucene

Apache Lucene[36]은 고성능 다목적을 지향하는 텍스트 검색 엔진이며, Java로 작성되었다. 고성능이라는 소개에 걸맞게 2016년 12월 기준 시간당 500GB이상의 텍스트 데이터를 처리할 수 있다.

<표 5-13> Apache Lucene

구 분	내 용
License	Apache
Requirement	Java 1.8+

Lucene는 다음과 같은 특징을 가지고 있다. 먼저 검색 결과의 순위를 제공한다. 검색 쿼리와 연관성에 따라 검색 결과의 순위를 사용자에게 제공한다. 또한 검색시 다양한 쿼리 타입을 사용할 수 있다. 검색 시 사용할 수 있는 쿼리 타입으로는 'Phrase query', 'Wildcard query', 'Proximity query', 'Range query'가 있다. 검색 결과의 필드를 순차 정렬을 통해 원하는 항목에 따라 우선도를 결정할 수 있다. 그리고 다중 인덱스 검색으로 검색을 폭 넓은 검색 범위를 지니고 있으며, 순수하게 JAVA 언어로 동작한다.

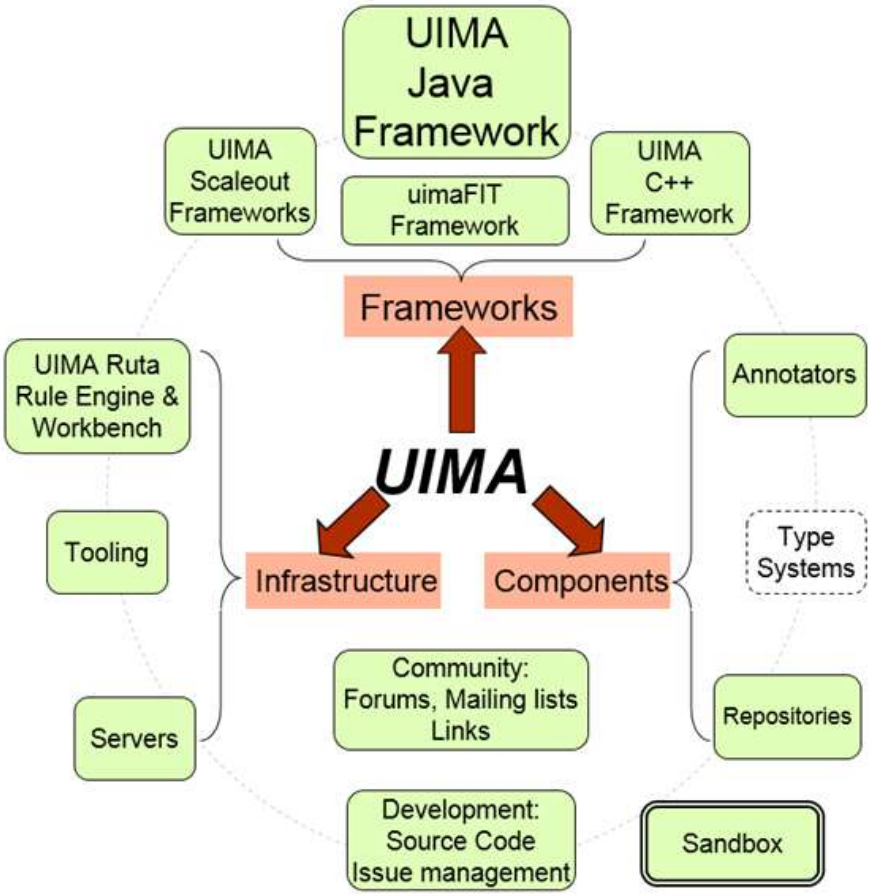
7. UIMA

Unstructured Information Management Applications[37] 줄여서 UIMA는 많은 양의 텍스트 정보로부터 유저와 관련된 지식을 추출하는 데에 목적을 둔 정보 검색 소프트웨어이다. UIMA는 어플리케이션을 컴포넌트로 나누어 관리할 수 있게 해준다. 예를 들어, '언어 인식', '문장 분리', '개체명 인식'과 같은 컴포넌트이다.

<표 5-14> UIMA 개요

구 분	내 용
License	Apache 2.0
Requirement	Python2.4+, Java 7 or 8

[그림 5-2]는 UIMA의 구조도를 나타낸다. UIMA는 다양한 언어의 플랫폼을 바탕으로 동작 가능한 프레임워크를 제공한다. UIMA는 규칙 기반 엔진과 서버를 지닌 구조이며, 주석 부착기와 데이터 저장소를 포함한다.



※출처 : https://uima.apache.org/

[그림 5-2] UIMA의 구조[37]

8. MALLET

MALLET[38]은 문서 분류, 개체명 인식, 토픽 모델링, 정보 추출 등의 언어분석 모듈을 제공하는 언어 분석 도구이다. 문서 분류로는 Naive Bayes, Maximum Entropy, Decision Tree 등의 알고리즘이 구현되어 있고, 분석 성능을 측정할 수 있는 코드도 작성되어 있다.

<표 5-15> MALLET 개요

구 분	내 용
License	CPL 1.0
Requirement	Java, Ant

MALLET에서 지원하는 모듈은 다음과 같다.

<표 5-16> MALLET의 모듈

분류	알고리즘	URL
Document Classification	Naive Bayes, Maximum Entropy, C45, DicisionTree	http://mallet.cs.umass.edu/classification.php
Sequence Tagging	HMM, CRFs	http://mallet.cs.umass.edu/sequences.php
Topic Modeling	LDA, Pachinko Alloction, Hierarchical LDA	http://mallet.cs.umass.edu/topics.php

9. ESPRESSO

Espresso[39]는 창원대학교 적응지능연구실에서 개발한 한국어 분석 도구이다. Espresso는 C++로 개발되었으며, 형태소분석, 구문분석, 개체명 인식 및 분류, 의미역분석, 상호참조까지를 하나의 커맨드로 분석할 수 있다.

형태소분석은 직접 구현한 HMM으로 동작한다. 형태소 분석에는 확률 사전과 패턴 사전을 사용하여 해당 사전의 확률 값을 이용하여 HMM에서 계산할 때 사용한다. 상호참조는 직접 구현한 Centering theory로 동작한다.

구문분석부터 의미역분석까지는 CRFs를 이용한 분석을 수행하고 있다. CRFs에 사용하는 자질은 형태소 분석 결과와 세종전자사전 의미 분류 정보, Word2Vector로 분류한 군집 정보 등을 사용하고 있다.

2016년 12월 기준으로 이 모든 자연어처리를 한 번에 수행하면서 초당 150문장을 처리 가능하며 형태소 분석 성능은 97%, 구문 분석은 81%, 의미역 결정은 68%이다.

<표 5-17> ESPRESSO 개요

구 분	내 용
Requirement	C++, Boost C++, crfpp
데모	http://air.changwon.ac.kr/~airdemo/Espresso/

<표 5-18> ESPRESSO의 모듈과 알고리즘

Module	Algorithm
POS	HMM
Parsing	CRFs
Named Entity Recognition	CRFs
Semantic Role Labeling	CRFs
Coreference Resolution	Centering theory

10. 꼬꼬마

꼬꼬마[40, 41] 프로젝트는 서울대학교 IDS 연구실에서 진행하는 프로젝트로, 크게 '형태소 분석기 등 자연어 분석 모듈 개발'과 '세종 코퍼스 활용 시스템'으로 구분하고 있다. '꼬꼬마' 팀에서는 세종 코퍼스를 다양한 용도로 활용할 수 있도록 코퍼스를 구조화하여 데이터베이스에 저장하고, 저장된 코퍼스로부터 다양한 통계 데이터를 생성하고, 저장된 코퍼스 및 통계 정보를 다양한 방법으로 조회 가능하다.

<표 5-19> 꼬꼬마 개요

구 분	내 용
License	GPL 2.0
Requirement	C++, Boost C++, crfpp
Reference	인접 조건 검사에 의한 초고속 한글 형태소 분석기 (https://www.dbpia.co.kr/Journal/ArticleDetail/NODE00617598) 어절 내의 형태소 범주 패턴에 기반한 통계적 자동 띄어쓰기 시스템 (http://www.dbpia.co.kr/Journal/ArticleDetail/NODE00788421) MADE: 형태소 분석기 개발 환경 (http://cs.sungshin.ac.kr/~shim/download/07_08.pdf)

<표 5-20> 꼬꼬마 형태소 분석기의 성능

유형	글 수	문장 수	평균 정확도
뉴스	10	1057	0.81
블로그	10	2199	0.75
상품평	10	2414	0.70

11. 코모란

코모란(Komoran: Korean Morphological Analyzer)[42]은 Shineware에서 작성해 공개한 Java 기반 형태소분석기이다. 순수하게 자바로만 개발되었기 때문에 자바가 설치된 환경이라면 어디서든지 사용 가능하며, 자제 제작한 라이브러리들만을 사용하기 때문에 외부 라이브러리에 대한 의존성 문제가 발생하지 않는다. 또한 50MB 정도의 적은 메모리 상에서도 동작 가능하다.

<표 5-21> 코모란 개요

구 분	내 용
License	Apache 2.0
Requirement	Java, gradle
Reference	@misc{komoran, author = {Junsoo Shin}, title = {komoran}, publisher = {GitHub}, journal = {GitHub repository}, howpublished = {₩url{https://github.com/shin285/KOMORAN}}
분석속도	초당 1.0MB
성능	어절 92.12% / 형태소 95.36%

12. 동아대학교 띄어쓰기 모델

동아대학교에서 개발한 띄어쓰기 모델[43]은 아주 작은 사전을 이용하면서도 94%의 성능을 내는 경량 띄어쓰기 모델이다. 경량화에 초점을 맞추어 확률 정보만을 기반으로 동작하여, 1MB 정도의 용량으로도 동작 가능하다. 음절 단위 패턴과 전이확률을 이용하여 구현하였다.

<표 5-22> 동아대학교 띄어쓰기 모델

구 분	내 용
Reference	음절 자질을 이용한 경량화된 자동 띄어쓰기 시스템
메모리	1MB 가량
성능	94%
데모	http://islab.donga.ac.kr/demo2.htm

제2절 연구소, 대학, 기업의 연구 수준

이번 절에서는 대학 및 연구소, 기업에서 수행한 연구들을 간략하게 소개한다. 자연어처리 기술의 발전과 실용 소프트웨어 개발을 위해 많은 대학과 연구소 그리고 기업에서 지속적인 연구&개발을 진행하고 있다. 자연어처리 기술 관련 학회와 경진대회를 통해 꾸준히 많은 대학, 연구소, 기업들이 참여하며 기술 발전과 정보 공유를 통한 협력 관계 형성을 위해 힘쓰고 있다. 소개 대상은 Stanford, Berkeley Group, MIT, Sheffield, Univ. of Washington, MS, IBM, 다이퀘스트, ETRI 이다.

1. Stanford

Stanford 대학의 자연어처리 그룹[44]은 컴퓨터가 인간의 언어를 이해하고 처리할 수 있게 해주는 알고리즘에 대해 함께 연구하는 교수진과 박사후 과정, 프로그래머 및 학생들로 이루어진 팀이다. Stanford 자연어처리 그룹은 언어학의 기초 연구부터 문장 이해, 자동 질문 응답, 기계 번역, 구문 분석 및 감정 분석 등 다양한 자연어 처리 시스템을 연구해왔다. 다음은 2016년도 진행한 연구 중 일부를 소개한다.

- Data Recombination for Neural Semantic Parsing
  - 코퍼스로부터 CFG 생성 -> CFG로 문장 증축 -> seq2seq 학습
  - 성능 GEO(89.3%), ATIS(83.3%), OVERNIGHT(77.5%)
- How Much is 131 Million Dollars? Putting Numbers in Perspective with

Compositional Descriptions

- 금액을 다른 친숙한 표현으로 생성해주는 연구
- 입력 예) Cristiano Ronaldo, the player who Madrid acquired for […] $131 million.
- 출력 예) About the cost to employ everyone in Texas over a lunch period.
- 성능 BLEU(Train 81.50%, Test 69.79%)

▪ A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task

- 문단과 질의를 입력하면, 문단의 정보를 이용하여 질의에 응답해주는 시스템
- 성능 CNN(Dev 72.4% Test 72.4%), Daily Mail(Dev 76.9%, Test 75.8%)

▪ Unanimous Prediction for 100% Precision with Application to Learning Semantic Mappings

- 인공신경망 기반 모델들의 confidence에 관한 연구
- 성능 검증 : GeoQuery dataset

2. Berkeley group

Berkeley group[45]은 언어를 이해하는데 컴퓨터를 사용하여 동작할 수 있도록 다양한 실험과 연구를 진행하고 있다. 인간의 언어의 풍부함과 미묘함을 대처 가능한 시스템을 만드는 것을 목표로 하고 있다. 연구하는 분야는 ‘언어 분석’과 ‘기계 번역’, ‘전산 언어학’, ‘의미 분석’, ‘비지도 학습’ 등을 연구하고 있다.

▪ Learning-Based Single-Document Summarization with Compression and Anaphoricity Constraints

- 단일 문서 요약 연구
- 성능 : ROUGE-1 42.20% ROUGE-2 25.90%

▪ Capturing Semantic Similarity for Entity Linking with Convolutional

Neural Networks

- Entity linking
- 성능 ACE 89.9% CoNLL 85.5% WP 90.7% Wiki 82.2%

▪ Learning to Compose Neural Networks for Question Answering

- 이미지와 구조화 데이터를 함께 이용한 QA
- 성능 59.4%

3. MIT

MIT NLP Group[46]은 기본 언어 현상을 모델링하는 것부터 실용적인 텍스트 처리 시스템을 설계하는 것과 새로운 기계 학습 방법을 개발하는 것을 포함하는 NLP 의 모든 측면을 다양한 방식으로 연구하였다.

MIT의 연구 핵심 주제는 번거로운 수동 주석 부착을 요구하지 않고 훈련할 수 있는 NLP 시스템을 구축하는 것이다. 최근에는 NLP의 경계를 새로운 과제로 확장하는데 관심을 가지고 있다. 예를 들어 텍스트 매뉴얼을 통해 자동으로 게임을 하는 플레이어를 개발하거나 텍스트 및 이미지 등의 기타 정보를 통해 중요한 사회 문제를 해결하는 기술을 유도하는 독창적인 알고리즘 개발에 힘쓰고 있다.

▪ Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings

- 단어 번역 쌍을 이용한 정확한 multilingual POS tagger.
- 6개 언어에 대해 성능을 측정한 결과 67.5%

▪ Rationalizing Neural Predictions

- 추론의 근거를 입력에서 추출할 수 있는 연구

▪ Randomized Greedy Inference for Joint Segmentation, POS Tagging and Dependency Parsing

- POS Tagging과 Parsing을 한 번에 수행하는 구조 제안

4. Sheffield

Sheffield NLP Group[47]은 1993년에 설립되어 영국에서 현재 가장 연구가 활발한 그룹중 하나이다.

Sheffield Group은 Information Access, Language Resources and Architectures for NLP, Machine Translation, Human-Computer Dialogue System, Detection of Reuse and Anomaly, Foundational Topics 등의 NLP 분야에 대한 연구를 진행하고 있다. 또한, EU, European Research Council, the UK Research Councils(EPSRC, BBSRC, MRC, AHRC), 정부, GlaxoSmithKline, IBM 등으로부터의 지원을 받고 있다.

Sheffield group은 자연어 처리의 다른 영역으로 연구를 확장시킬 수 있도록, Speech and Hearing, 정보 검색 그룹과 많은 유대를 맺고 있다.

- Representation and Learning of Temporal Relations
  - 사건과 시간의 상관 순서를 파악하기
  - 성능 : Before 97.72%, After 95.43%, Overlap 55.23%, During 89.02%
- Broad Twitter Corpus: A Diverse Named Entity Recognition Resource
  - 개체명 인식 트위터 코퍼스 공개 및 분석
- Understanding the Lexical Simplification Needs of non-Native Speakers of English
  - 복잡한 단어를 비슷하지만 쉬운 대체어로 변환하는 연구. Non-native를 위함.

5. Univ. of Washington (UW)

UW CSE NLP 그룹[48]은 요즘 각광받고 있는 도전 과제(이를테면 SNS 데이터 분석, 이미지 분석과 자연어 처리 기술의 융합)와 함께 핵심 NLP 기술(이를테면 구문분석, 정보 추출, 기계 번역 등)을 연구하고 있다. 또한, 워싱턴 대학 내의 다른 분야의 전문가들과 업계의 파트너들(Microsoft Research, Google, Facebook, Intel)과도 지속적인 협력관계를 유지하고 있다.

- LSTM CCG Parsing

- LSTM CCG를 이용한 state-of-the-art Parser

- Human-in-the-Loop Parsing
  - Crowdsourcing을 이용한 Active learning CCG dependency parsing
- Are Elephants Bigger than Butterflies? Reasoning about Sizes of Objects
  - 이미지와 텍스트를 이용한 사물 간 크기 비교 연구
- Global Neural CCG Parsing with Optimality Guarantees
  - CCG와 Recursive Network를 이용한 Graph-based dependency parsing algorithm 제안

6. MS Human language technologies

MS의 human language technologies 팀[49]은 자연어처리, speech recognition, 대화 시스템, 구어체 이해, 인간 대화의 뉘앙스 등에 대해서 연구하고 있다. HLT는 웹 검색 엔진 기술과 음성 인식, 음성 합성, 광학 문자 인식, 기계 번역, 전자 사전, 맞춤법 검사, 문법 검사, 워드 프로세싱, 컴퓨터 지원 언어 교육, TV 자동 자막 등의 인간 언어와 정보 기술이 만나는 모든 분야의 기술에 관심을 가지고 연구하고 있다.

- Focus areas
  - Deep learning, Dialog Systems, Speech Recognition, Machine Reading
- Research news
  - 전문적인 transcriptionist보다 낮은 에러를 기록한 speech recognition system 개발
  - 음성/대화 기술을 위한 툴킷 Computational Network Toolkit(CNTK)를 출시 tensorflow가 지원하지 않는 다중 서버*다중 GPU를 지원
- APIs
  - <표 5-23> 참조.

<표 5-23> MS에서 제공하는 API들

분류	URL
안면인식+감성 분류	https://www.microsoft.com/cognitive-services/
Spell Checker	https://www.microsoft.com/cognitive-services/en-us/bing-spell-check-api
POS +Constituency parsing +Tokenize	https://www.microsoft.com/cognitive-services/en-us/linguistic-analysis-api
Sentiment + Key phrase extraction + Topic detection + Language detection	https://www.microsoft.com/cognitive-services/en-us/text-analytics-api
Translation	https://www.microsoft.com/cognitive-services/en-us/translator-api
Word breaking	https://www.microsoft.com/cognitive-services/en-us/web-language-model-api

7. IBM NLP group

IBM Research[50]는 NLP에서 오래되고 유명한 역사를 지니고 있다. IBM은 기계 번역에서 정보 추출, 질의 응답에 이르는 광범위한 주제를 다루고 있다. IBM은 NLP Professional Interest Community(PIC)를 통해서 학계와 상호 작용 및 광범위한 연구 커뮤니티를 유지하며 NLP에 대한 많은 연구 자료 정보를 제공하고 있다.

- Generating High Quality Proposition Banks for Multilingual Semantic Role Labeling
  - 반자동 SRL 코퍼스 생성 방법 제안
- An In-depth Analysis of the Effect of Text Normalization in Social Media
  - SNS 데이터를 normalize하는 방안에 대한 연구 소개
- VINERy: A Visual IDE for Information Extraction
  - Unstructured / semi-structured data를 분석하는 데 유용하게 사용될 수 있는 visual IDE 소개.

<표 5-24> IBM에서 제공하는 API들

분류	URL
통합 언어 모델	https://www.ibm.com/watson/developercloud/alchemy-language.html
Chatbot	https://www.ibm.com/watson/developercloud/conversation.html
Dialog	https://www.ibm.com/watson/developercloud/dialog.html | | Translation
Translation	https://www.ibm.com/watson/developercloud/language-translator.html
STT	https://www.ibm.com/watson/developercloud/speech-to-text.html
TTS	https://www.ibm.com/watson/developercloud/text-to-speech.html
Natural Language Classifier	https://www.ibm.com/watson/developercloud/nl-classifier.html
Tone Analyzer	https://www.ibm.com/watson/developercloud/tone-analyzer.html

8. 다이퀘스트

다이퀘스트[51]는 2000년부터 자연어처리 연구를 수행하고 있는 국내 기업이다. 100여 개가 넘는 구축사례를 가지고 있으며, 스무 곳 이상의 기업과 파트너쉽을 맺고 있다.

- DISA : 비정형 데이터의 의미정보 추출 시스템
  - 다양한 형태의 비정형 데이터에서 의미 개체, 관계 정보, 감성 정보 등의 의미 정보를 추출하여 정형화된 데이터로 정제.
- DQ-Cat : 문서 자동 분류기
  - 문서 범주의 대표 문서를 직접 할당함으로써 자동 문서 분류가 가능한 문서 분류기
- DQ-Clu : 문서 자동 군집기
  - 문서간 유사도를 기준으로 군집화한 데이터를 제공함으로써 그룹별 주제어휘에 따른 검색결과를 손쉽게 파악, 분류된 정보로 인한 정보 접근성 상승
- DRAMA : 연관도 분석기
  - 형태소 분석으로 검색 키워드와 연관 검색어, 관계도, 유사 문서 분석 등의 결과 제공

- SEMON : 시맨틱 서비스 플랫폼
  - 사용자 의도에 적합한 결과를 다양한 분석 방법을 통해 생성하는 시맨틱 프레임워크

9. ETRI

ETRI(한국전자통신연구원)[52]는 신뢰받는 연구개발 성과를 창출하고, 우리나라 SW·콘텐츠 기술 경쟁력을 도약시키기 위한 연구개발을 수행하고 있으며, 창의적 SW·콘텐츠미래기술 선점을 위한 차세대콘텐츠, 빅데이터 인텔리전스, 자동통역, 고성능컴퓨팅, 임베디드SW, 지능형 인지, 스마트 모빌리티, SW원천연구분야에서 핵심기술을 연구개발하고 있다.

- 무인자율차 전용 프로세서 첫 개발
  - 1W 내외의 저전력으로 동작하는 GHz급 자동차 전용 프로세서 개발
- T-SDN 표준기술 개발 국내생태계 활성화
  - 광전송네트워크 SDN 표준 기술을 제정하여 국제 표준을 선도하고자 함

<표 5-25> IBM에서 제공하는 API들

기술명	개요
타액기반 발병원인물질 정량화 프로세스 기술	치매 특히 알츠하이머 환자에게 나타나는 발병원인물질인 베타아밀로이드 40과 베타아밀로이드 42를 타액 내 어느 정도의 양으로 존재하는지 알아내는 방법에 대한 절차 기술. 알츠하이머 조기 진단에 활용 가능
스마트 광분배함 및 관리 플랫폼 기술	광가입자망에서 광섬유와 광분배 노드로 구성된 광분배망을 효과적으로 실시간 관리할 수 있는 체계 구축에 관한 기술, 실시간 광계층 모니터링을 위한 스마트 광분배함 및 국제표준 기반 스마트 광분배망 통합관리 플랫폼을 포함함. 광 인프라 자원 이용률 100% 달성 및 회선 제공 시간을 기존 대비 1/3로 줄일 수 있음
광시야각 영상 재생 스케쥴러 기술	본 기술은 넓은 광시야각(8K X 2K 이상)을 가지는 고품질 대용량 영상들을 스케쥴링하고 재생하는 소프트웨어. 기존에 개발된 광시야각 영상 렌더링 시스템과 연동하여 광시야각 영상과 패키징된 콘텐츠(오디오, 자막 등)의 원하는 구간을 선택하여 스케쥴링 및 재생하는 기능을 지원함

제6장 지식데이터베이스 현황 조사

제1절 한국어 지식 데이터베이스

한국어 자연어 처리 연구를 위한 지식 데이터베이스에는 문자, 음성, 이미지 등 다양한 종류가 있다. 각 대학교, 기관이 구축한 지식 데이터베이스 현황에 대한 현황 조사이다.

1. 서울대학교

서울대학교의 코퍼스[53]는 한국어 감성 코퍼스이다. 영어의 경우, 약 10,000개의 문장 안에 나타나는 감성 표현들로 구축된 사전인 MPQA(Multi-perspective Question Answering)와 같은 역할을 하는 한국어 감성 코퍼스, KOSACKorean Sentiment Analysis Corpus)을 구축했다.

세종 구문 코퍼스로부터 선별한 322개 신문 기사, 7744개 문장을 주석 대상으로 삼아 총 17,582개의 감성 표현을 주석한 코퍼스이다.

※이미지출처 : IDG Tech Report - "글에서 감정을 읽다" 감성 분석의 이해

[그림 6-1] IDG - "글에서 감정을 읽다" 표지

2. 언어정보연구원

연세대학교 언어정보연구원[54]의 코퍼스는 띄어쓰기를 기준으로 약 8,879만 어절 규모이고 1980년대 후반부터 구축하기 시작하였으며, 구축 대상을 현대 한국어 문어만으로 한정 했던 것에서 벗어나 다양한 언어자료를 전산화하기 시작했다. <표 6-1>은 '연세 코퍼스'의 구성이다.

<표 6-1> 연세 코퍼스 구성

코퍼스 이름	코퍼스 소개	어절 수
연세 코퍼스1 (균형 코퍼스)	현대 한국인의 독서 실태'에 관한 설문 조사에 의해 구축	290만
연세 코퍼스2 (균형 코퍼스)	도서 대출 빈도에 따른 균형 잡힌 코퍼스를 구성할 목적으로 구축 한국어 문헌을 크게 열 가지 대 범주(철학, 종교, 사회, 과학, 언어학, 순수과학 등의 분야)로 나누어 수집	110만
연세 코퍼스3 (1980년대 코퍼스)	1980년대 우수 출판물 목록에서 표본을 선정	598만
연세 코퍼스4 (1990년대 입 코퍼스)	실제 사용된 말을 녹음한 구어와 희곡, 방송대본 등을 중심으로 한 준구어로 구성 발화자 참여자 나의 ,성별, 직업 등 정보 포함	77만
연세 코퍼스5 (1970년대)	1970년대 문헌을 대상 신문(10%), 소설·수필(50%), 일반서적(35%) 등의 문헌 자료를 수집함으로써 구축	860만
연세 코퍼스6 (1960년대)	해방 이후의 우리말의 모습을 제대로 반영한 코퍼스 목적으로 구축 1960년대 문헌을 대상	723만
연세 코퍼스7 (1990년대)	1990년대 중반까지의 자료로 구성 주로 소설과 수필로 구성	1,367만
연세 코퍼스8 (교과서)	초등학교 전 과목 및 중·고등학교의 국어, 사회를 중심으로 구성	87만
연세 코퍼스9 (아동도서)	순 우리말의 사용빈도가 높고 문장도 비교적 우리말 구조에 맞게 사용 되는 아동 교육용 도서로 구성됨	150만
연세 코퍼스10 (1945~1965년 단행본)	<연세 현대한국어사전> 편찬을 위해 보완된 1945년~1965년 코퍼스 중 단행본으로 구성	78만
연세 코퍼스11 (1945~1965년 교과서)	<연세 현대한국어사전> 편찬을 위해 보완된 1945년~1965년 코퍼스 중 교과서로 구성	73만

3. 국립국어원

국립국어원[55]은 1984년에 설립한 ‘국어연구소’의 전통을 이어 1991년에 ‘국립국어원’으로 승격되었다. 국립국어원은 한국어가 처한 환경을 이해하고 이를 해결할 수 있는 문제에 대해 다양한 정책을 펼쳐 다양한 한국어 자원을 성과물로 내놓았다. 국립국어원에는 ‘21세기 세종계획’의 성과물로서, 전자사전과 세종 코퍼스 등이 있다.

가. 21세기 세종계획

(1) 전자사전

전자사전은 해당 어휘가 가질 수 있는 언어적 정보를 모두 ‘코드화’하여 기술하고 있는 사전이다. 해당 어휘의 의미와, 의미 분류, 예문이 포함됨과 동시에 결합 정보, 격틀 등등 갈래뜻(sense) 단위로 되어있다. 전자 사전의 항목은 약 50,000 항목으로 이루어져 있으며, <표 6-2>는 전자 사전의 구성이다.

<표 6-2> 전자사전 구성

품 사	항목 수
체 언	25,458
동 사	15,180
형용사	4,398
부 사	4,320

(2) 세종 코퍼스

세종 코퍼스는 코퍼스 종류(유형)에 따라 다시 여러 가지로 구분된다. 코퍼스의 종류는 원시 코퍼스, 형태 분석 코퍼스, 의미 분석 코퍼스, 구문 분석 코퍼스가 있다. 원시 코퍼스는 원문을 입력해 놓은 코퍼스이고, 형태 분석 코퍼스는 ‘원시 코퍼스’에 형태(어휘) 단위의 정보를 부가한 코퍼스이고, 의미 분석 코퍼스는 ‘형태 분석 코퍼스’에 동형어와 갈래뜻(sense) 구분 표지를 부가한 코퍼스이며 구문 분석 코퍼스는 ‘의미 분석 코퍼스’에 통사 구조/기능 표지를 부가한 코퍼스이다.<표 6-3>은 문어 코퍼스의 양이고, <표 6-4>는 구어 코퍼스의 양이다.

<표 6-3> 문어 코퍼스 양

세부 코퍼스	어절 수
원시 코퍼스	약 3,700만 (36,879,143어절)
형태 분석 코퍼스	약 1,000만 (10,066,722어절)
의미 분석 코퍼스	약 1,000만 (9,071,054어절)
구문 분석 코퍼스	약 1,000만 (9,071,054어절)

<표 6-4> 구어 코퍼스 양

세부 코퍼스	어절 수
원시 코퍼스	약 80만 (805,646어절)
형태 분석 코퍼스	약 80만 (805,646어절)

나. 서울말 낭독체 발화 코퍼스

2005년에 개발된 서울말 낭독체 발화 코퍼스는 2대 이상 서울•경기 지역에 거주해 온 서울말 화자 120명의 낭독체 말소리를 담아 만든 것이다. 여기에는 소설, 수필, 논설 등 다양한 형태의 글을 화자 각자가 편안하게 읽은 음성 파일이 약 88,800여개가 수록되었다.

다. 한국어 자연발화 음성 코퍼스(서울 코퍼스)

연령별 (10대, 20대, 30대, 40대) 서울 지역의 화자 각 10명씩 총 40명으로 연령대에는 남녀 각 3명씩 구성되었다. 화자 별로 약 1시간가량 녹음한 인터뷰이며 주제는 일상적인 얘기에서 정치, 경제, 사회, 문화 등 다양한 내용을 다룬다. 인터뷰는 질문에 대답하는 방식 그리고 자발적으로 자연스럽게 자신의 의견을 표현하고 있다.

4. ETRI

ETRI[56]는 ICT 발전을 위해 다양한 연구&개발을 진행하고 있다. 자연어처리 분야 뿐만 아니라 다양한 기술 분야에서 많은 연구가 진행되고 있다. 그중 자연어 처리 기술 발전을 위해 다양한 대학 연구소 및 기업과 연계하여 기초 코퍼스부터 자연어처리 분석 도구 개발 등의 많은 성과를 보유하고 있다.

ETRI의 지식 데이터베이스는 텍스트 뿐 만 아니라 음성 데이터도 보유하고 있으며 지난 2004년부터 2015년까지 배포중인 DB목록이다.

<표 6-5> 배포 중인 데이터베이스 목록

원시 자료명	원시자료 주요내용	자료 형태	수량
개체명 사전 DB	정보 검색을 위한 개체명 사전	텍스트	250만 엔트리
질문대화 코퍼스	질의응답용 질문대화 코퍼스	텍스트	100만 엔트리 5,000 대화셋
구문구조부착 코퍼스	품사 및 구문정보를 위한 코퍼스	텍스트	10만 문장
언어대역 패턴 DB	한영 자동번역용 연어 대역 패턴	텍스트	280만 패턴
한국어 대화체 텍스트 DB	한국어 대화체 텍스트 DB	텍스트	100만 문장
대화 음성 및 태깅 코퍼스	대화체 음성 및 대화 정보 태깅 코퍼스 DB	음성 /텍스트	2만 발화 2천 대화
구어체 한국어 코퍼스	일반인 작성 구어체 한국어 대화 코퍼스	텍스트	A4 3,500면
원거리/다채널 음성인식 단어/ 문장 DB	지능형 로봇용 음성인터페이스 개발용 원거리/다채널 음성인식 단어/문장 DB	음성	200시간
핵심어 검출용 음성인식 문장 DB	지능형 로봇용 음성인터페이스 개발용 핵심어 검출용 음성인식 문장 DB	음성	120시간
한국어 감정표현 음성합성용 대화체 문장 음성 DB	한국어 감정표현 음성합성용 대화체문장 음성 DB	음성	60시간
한국어 대화체 음성 DB	표준어 사용자 대상 한국어 음성 DB	음성	5,810분

5. 한국과학기술정보연구원과 충남대학교

한국 과학기술 정보연구원과 충남대가 공동으로 개발한 지식 데이터베이스[57]는 1998년부터 2003년까지 개발한 정보검색 시스템 평가를 위한 한글 테스트 컬렉션인 HANTEC이다. HANTEC은 테스트 문서와 TREC형식의 질의와 적합문서집합과 시스템 평가 프로그램으로 구성되어 있다. 테스트 문서는 12만 건으로, 한국경제신문, 여성개발원 보고서, 회의록을 포함하는 사회과학 분야와

과학기술부 보고서, 과학기술 동향, 논문 서지를 포함하는 과학기술 분야 그리고 한국일보, 인터넷 웹문서를 포함하는 일반종합 분야로 되어있다.

TREC 형식의 질의는 50개로 전문영역 분류와 사용자 분류로 구분되어 있다. 전문영역은 사회과학, 과학기술, 일반종합 질의로 분류되고 사용자는 일반인, 영역 전문가, 중·고등학생으로 분류되고 적합문서집합은 적합성 정도에 의한 8종의 적합문서 집합이 있으며 SMART 정보검색시스템 평가프로그램으로 구성되어 있다.

6. 울산대학교

울산대학교는 표준국어대사전(2002년 CD버전)을 기반으로 명사의미망(상•하위어), 용언 논항의 의미제약정보, 부사와 의미적으로 결합하는 용언, 부사, 명사 등을 연결한 어휘의미망(Lexical Semantic Network)인 UWordMap[58]이 있으며, 어휘지도는 DB자체로는 제공하지 않으며, UTagger용 파일 DB 및 API로 연구용으로 제공되고 있다.

한국어 형태소 분석기와 동형이의어 분별을 동시에 수행하는 UTagger이 있다. 품사태그는 세종태그셋을 사용하고, 동형이의어 번호 체계는 세종도 기준으로 하며 대체로 국립국어원의 표준국어대사전과 일치한다. 기본적으로 세종 코퍼스를 학습하여 작동하며, '사용자 코퍼스'을 이용하여 신조어, 용언의 활용형, 인접 두 어절 간의 새로운 문맥을 실시간으로 학습하는 기술을 제공하고 있다.

울산대 코퍼스인 UCorpus에는 3가지가 포함되어 있다. 형태의미주석된 세종 코퍼스, 신문, 사전 등 약 6,300만어절을 가지는 형태의미주석코퍼스인 UCorpus-HG이다. 세종구구조코퍼스를 의존관계로 변환 후 의미역으로 부착한 의존관계/의미역주석코퍼스인 UCorpus-DP/SR이고 90,010개의 다의어를 가지는 용언을 대상으로 논항의 의미역을 부착시킨 용언격틀/의미역 사전인 UPropBank가 있다.

7. 고려대학교

1995년에 1,000만 어절 규모의 현대 한국어 말모둠인 '고려대학교 한국어 말모둠 1' (KOREA-1 Corpus)[59]을 구축했다. 텍스트의 매체 및 내용에 따른

분류를 기반으로 다양하게 구성되었다. 텍스트 분류는 4개의 숫자로 표현하며 1) 문• 구어 구분, 2) 전달 매체에 따른 구분(신문, 잡지, 방송 등), 3) 텍스트 내용에 따른 구분(인문, 사회, 자연 등) 4) 문어의 경우 더욱 자세한 내용의 구분/구어의 경우 담화상황(전화, 대담 등)을 표현한다.

2000년도 이후에 발간된 국내 4대 일간지인 동아일보, 조선일보, 중앙일보, 한겨레신문으로 부터 제공받은 신문기사를 바탕으로 수집하고 가공한 대규모 신문자원 코퍼스인 물결21 코퍼스이다. 매년 전년도의 기사를 추가하고 있으며 2014년도 기준으로 2000년부터 2013년(14년치)에 이르는 모든 기사를 포함하여 총 6억 어절 규모로 구축되었다.

8. 펜실베니아 대학교

미국의 펜실베니아 대학에서 영어와 중국어와 함께 한국어를 대상으로 하여 만든 구문분석 코퍼스인 Penn Korean Treebank(PKT)[60]가 있다. 군사 교본과 뉴스기사로 대상이 한정되었다 있으며 약 30만 어절 규모이다. Penn Korean Treebank는 최근에도 많은 논문에서 학습 코퍼스로 사용되고 있는 중요한 언어 자원으로 여겨지고 있다.

9. LDC

LDC(Linguistic Data Consortium)[61]는 군대 도메인의 Virginia 코퍼스(약 54,000 단어)와 경제 신문 도메인의 Newswire 코퍼스(약 130,000 단어)로 구성되었다 있는 Korean PropBank이다. Virginia 코퍼스와 Newswire 코퍼스에 등장한 모든 서술어에 대해서 서술어의 의미 별로 논항의 의미역을 정해 놓은 격틀 파일(frame file)을 포함하고 있다.

LDC 코퍼스는 의미역 결정 모델의 학습 코퍼스로 사용되고 있다. 문장에 나타난 서술어에 따라 각 서술어가 가질 수 있는 의미 논항을 표시하고 있으며, 각 서술어의 격틀 정보를 따로 정리한 격틀 사전 정보를 포함하고 있다. 격틀 사전에는 서술어가 지닐 수 있는 의미 논항의 격 정보가 포함되어 있으며 서술어의 의미 분류 번호에 따른 활용 정보가 포함되어 있다.

10. kAIST

KAIST(한국과학기술원)는 SWRC(Semantic Web Research Center)에서 Off-line Korean letters Database[62]를 제공한다. 이는 오프라인 한글 글씨 인식에 관한 연구를 수행하는 개발자들에게 한글 필기체 영상 데이터를 제공한다. 이를 위해 광범위한 직업, 연령, 성별층에 대한 정보를 포함하고 있는 필기체를 선정하여 KSC 완성형 한글 1,200자에 대한 글씨 데이터 1,000벌을 수집하였고 이를 데이터베이스화 하였다. 1997년에 구축을 하여 국어정보처리기술 개발 프로젝트를 수행하였다.

11. 언어자원은행

언어자원은행(BORA : Bank of Resource for Language and Annotation)[63]은 과학재단의 지원을 받아 연구소재를 연구자가 효율적으로 확보할 수 있도록 하자는 취지에 의해 만들어 졌다. 언어자원은행은 지식기반 소프트웨어 개발을 위한 언어자원을 체계적이고 종합적으로 개발 및 관리하는 역할을 한다. 언어자원은행은 한국어, 영어, 중국어, 일어 등의 원시 코퍼스 및 분석 코퍼스를 비롯하여 품사별 고유명사, 복합어, 격 등에 대한 전자사전(Electronic Dictionaries)을 보유하고 있다. 또한 이를 기반으로한 다양한 자연어처리 소프트웨어를 보유하고 있다. <표 6-6>은 언어자원 은행의 한국어 음성데이터(Speech Data)에 대한 정보를 나타낸다.

<표 6-6> 음성 데이터 정보

구분	자원 정보	개수
한국어 음성DB1	화자 70명, 아나운서 2명	450어절, 2,000어절
한국어 음성DB2	36개 지명, 남성/여성	36어휘
한국어 음성DB3	아나운서 2명, 화자70명	1문단
한국어 음성DB4	화자 70명(4회), 아나운서 2명(2회)	기수 32개, 한정사 9개
한국어 음성DB5	화자 70명(4회), 아나운서 2명(2회)	단순숫자 4개, 기수 35개
한국어 음성DB6	화자 20명	539문장, 일상문장 50세트
한국어 음성DB7		단순기수 32개, 단순숫자 1,620종

<표 6-7>,<표 6-8>,<표 6-9>는 각각 전자사전 정보, 코퍼스 정보, 코어넷에 대한 정보이다.

<표 6-7> 전자사전 정보

구 분	개 수
기본어휘 사전	54,796어휘
고유명사 사전	37,000어휘
복합명사 사전	26,000어휘
개념체계 사전	2,954개념(53,628개 의미)
동사격틀 사전	4,560어휘
복합명사 논항구조 사전	4,297개
형용사 구문정보 사전	5,930개

<표 6-8> 코퍼스 정보

구 분	개 수
원시 코퍼스	70,000,000어절
정련된 형태소분석 코퍼스	1,000,000어절
자동형태소 분석 코퍼스	40,000,000어절
정련된 구문분석 코퍼스	3,000문장
수동 구문분석 코퍼스	30,000문장
신문 코퍼스 (한겨레 신문)	620파일
신문 코퍼스 (동아-한영일중)	1791파일

<표 6-9> 코어넷 정보

구 분	개 수
코어넷 한국어편	2,937개의 개념망에 연결된 40,644어휘
코어넷 한국어 명사편	21,401어휘 (51,607 의미)
코어넷 한국어 동사편	1,758어휘 (5,290 의미)
코어넷 한국어 형용사편	813어휘 (2,801 의미)
코어넷 한국어 동사격틀	406어휘 (957의미)
코어넷 한국어 형용사격틀	759어휘 (1,109 의미)

제2절 해외 지식 데이터베이스

지능정보기술 발전의 필수 요소인 지식 데이터베이스가 존재하듯이 다른 주요 국가에도 코퍼스들이 존재한다. 그 코퍼스들에 대한 조사를 진행하였다.

1. 영국 코퍼스

영국 코퍼스는 현대 영국영어의 넓은 범위를 표현하기 위해 설계 되었으며 대략 100만 단어가 포함되었다. 영국 코퍼스의 구성은 텍스트의 콘텐츠 유형인 '도메인', 텍스트 생산 기간인 '시간', 책 또는 정기간행물과 같은 텍스트 출판물의 유형인 '매체'로 나눌 수 있다. <표 6-10>은 텍스트 코퍼스의 구성을 나타낸다. 또한, 음성 데이터는 지역, 유형, 주제로 나눌 수 있다. <표 6-11>는 음성 코퍼스 구성을 나타낸다.

<표 6-10> 영국 텍스트 코퍼스 구성

도메인(%)	시간(%)	매체(%)
창작물 (21.91%) 예술 (8.08%) 종교 (3.40%) 경제/금융 (7.93%) 레저 (11.13%) 자연/순수과학 (4.18%) 응용과학 (8.21%) 사회과학 (14.80%) 세계정세 (18.39%) 분류되지 않음 (1.93%)	1960년~1974년 (2.26%) 1975년~1993년 (89.23%) 분류되지 않음 (8.49%)	책 (58.58%) 정기 간행물 (31.08%) 기타 출판 (4.38%) 연설문 (1.52%) 분류되지 않음 (0.40%)

<표 6-11> 영국 음성 코퍼스 구성

지역(%)	유형(%)	주제(%)
남부 지방 (45.61%) 중부 지방 (23.33%) 북부 지방 (25.43%) 분류되지 않음 (5.61%)	독백 (18.64%) 대화 (74.87%) 분류되지 않음 (6.48%)	교육/정보 (20.56%) 비즈니스 (21.47%) 기관 (21.86%) 레저 (23.71%) 분류되지 않음 (12.38%)

2. 미국 코퍼스

미국 코퍼스는 텍스트 83,000단어와 음성 32,000단어 총 115,000단어로 이루어져있다. 현재는 Linguistic Data Consortium(LDC)에서 이용 가능하며 <표 6-12>는 미국 코퍼스의 텍스트 카테고리이다.

<표 6-12> 미국 코퍼스의 텍스트 카테고리

구분	텍스트 카테고리	%
텍스트	책	55
	뉴스, 잡지, 신문	20
	메일, 웹페이지 등	10
	여러 종류(출판 그리고 비출판)	5
음성	대화, 연설, 미팅	10

3. 체코 코퍼스

체코 코퍼스는 두 가지 섹션으로 구성되어 있다. 공시적 언어와 통시적 언어이며, 공시적 언어는 역사적인 시대의 흐름이 아닌 그 시대 상황만을 고려하고, 역사적 사실을 동시대의 관점으로 비교, 해석하는 것이다. 공시적 언어와 반대되는 개념의 통시적 언어는 역사적으로 고찰하는, 시간의 흐름을 고려하고 시간적, 문화적, 공간적 차이를 고려하여 해석하는 것이다. 공시적 언어는 100만 단어로 구성되어 있으며 <표 6-13>은 공시적 언어의 구성이다.

<표 6-13> 체코 코퍼스 공시적 언어의 구성

주요 카테코리	장르	%
창작물	소설	11.02
	시	0.81
	드라마	0.21
	다른 문학작품	0.36
	짧은 텍스트 유형	2.6
정보제공물	저널	60
	기술/특수 텍스트	25

통시적 언어는 80만 단어로 구성되어 있으며 <표 6-14>는 통시적 언어의 샘플링 프레임 구성이다.

<표 6-14> 체코 코퍼스 통시적 언어의 구성

표준	유형	%
화자 성별	남성	50
	여성	50
화자 나이	21~35세	50
	35세 이상	50
교육 레벨	중학교	50
	대학교	50
발화 유형	공식	50
	비공식	50

4. 헝가리 코퍼스

헝가리 국가 코퍼스에는 153만 7천단어가 포함되어 있으며 <표 6-15>는 헝가리 코퍼스의 포함 단어의 구성이다.

<표 6-15> 헝가리 코퍼스 카테고리

카테고리	%
미디어	52.7
문학	9.43
과학문서	13.34
공식 문서	12.95
비공식 문서 (예 : 전자포럼 토론)	11.58

5. 러시아 코퍼스

100만 단어로 구성되어 있으며 POS를 위한 주석이 포함되어 있고, TEI-compliant SGML로 인코딩되어있다. <표 6-16>은 러시아 코퍼스의 샘플링 프레임이다.

<표 6-16> 러시아 코퍼스 카테고리

텍스트 카테고리	%
음성	5
전기	30
자연과학	5
응용과학	10
사회과학	12
정치	15
상업/거래	5
예술	5
지역과 철학	3
레저	10

6. 이탈리아 코퍼스

100만 단어로 구성되어 있으며 대부분의 국가 코퍼스와는 달리 2년 마다 업데이트 된다는 점이 특징이다. <표 6-17>은 CORIS 코퍼스의 구성이다.

<표 6-17> CORIS 코퍼스의 구성

카테고리	하위 카테고리	%
언론	뉴스, 정기간행물, 부록	38
허구	소설, 짧은 이야기	25
교육 산문	인간과학, 자연과학, 물리학, 실험과학	12
법률 및 행정 산문	법률, 정치, 행정문서	10
여러 종류	종교, 여행, 요리, 취미 등 도서	10
일회성 인쇄물	편지, 전단지, 지시	5

7. 그리스 코퍼스

32만 단어로 구성되어 있으며 기사, 에세이, 문학 작품, 보고서, 전기 등 다양한 장르 및 도메인을 포함하는 여러 출판 매체(경제, 의학, 레저, 예술 등)으로 부터 샘플데이터를 이용하여 구축한 데이터이다. 책 (15.75%), 뉴스 (69.01%), 정기간행물(6.97%), 여러 종류(통신, 전자 텍스트, 수기 작성/입력)(8.27%)로 4가지 유형으로 구성되어 있다.

8. 독일 코퍼스

100만 단어로 구성되어 있으며, 데이터에 대한 원형화, 단어 수준의 POS와 의미 주석, 전치사구/명사구 인식으로 구성되어 있고 <표 6-18>은 독일 코퍼스의 구성이다. '특수' 카테고리에는 광고, 매뉴얼 등이 포함되고, '음성' 카테고리에는 일상 언어, 방송 토론, 방언 등이 포함되어 있다.

<표 6-18> 독일 코퍼스의 구성

텍스트 카테고리	%
문학	25
언론 산문	25
과학	20
특수	20
음성	10

9. 중국 코퍼스

현대 중국어 코퍼스로 The Modern Chinese Language Corpus(MCLC)이다. 1919년부터 1992년 까지 14억 문자 텍스트에서 체계적으로 샘플링한 7억 문자 데이터로 구축 되었고, 새로운 데이터는 연례로 350만개의 비율로 필터링되었다. 현재는 850만을 포함하고 있다. 40개 이상의 하위 카테고리를 포함하는 4개의 카테고리로 분류하고 있다. <표 6-19>는 MCLC의 구성이다.

<표 6-19> 중국어 MCLC 코퍼스의 구성

카테고리	하위 카테고리	%
인문 사회 과학	정치와 법률, 역사, 사회, 경제, 예술, 문학, 군사학 및 체육학, 생활 (8개의 하위 카테고리)	59.6
자연 과학	수학 및 물리학, 의료 및 건강 생물학, 화학, 천문과 지리, 해양 및 기상, 농업 및 임업 (6개의 하위 카테고리)	17.24
여러 종류	공식 문서, 규정, 법적 문서, 비지니스 문서, 연설문, 일회성 인쇄물 (6개의 하위 카테고리)	9.36
신문		13.79

제3절 모니터 코퍼스

언어가 가지는 가변성으로 인하여 시대와 상황, 지역과 사람에 따라 다양하게 변한다. 사람들은 더 명쾌하고 참신한 표현을 원하고 그것이 잘 맞으면 오랜 기간 사용하기도 하기도 하고 오래 가지 못하고 사라지는 언어가 있다. 이러한 언어들이 시간이 갈수록 가속화되며 그 양도 많아지고 있다. 당연히 양이 많아질수록 오래가지 못하고 사라지는 언어들도 많다. 이러한 점 때문에 '모니터 코퍼스'가 있다. '모니터 코퍼스'는 빠른 언어의 변화를 추적하기 위해 지속적으로 업데이트 되는 코퍼스를 이야기 한다.

1. The Bank of English(BoE) 코퍼스

가장 잘 알려진 The Bank of English(BoE) 코퍼스이다. 영국 영어(70%), 미국 영어(20%), 다른 종류의 언어(10%)로 나타내고 524만 단어를 포함하고 있다. 뉴스, 잡지, 소설, 실화, 보고서, 웹사이트로 이루어진 텍스트는 75%를 차지하고, tv, 라디오, 미팅, 인터뷰, 토론, 회의에 대한 사본으로 구성되어 있는 음성은 25%를 차지한다.

2. Global English Monitor 코퍼스

Global English Monitor Corpus는 1961년 미국에서 출판된 영문 텍스트이며, 샘플링 프레임은 Brown University Library와 Providence Athenaeum에서 수집한 서적과 정기 간행물의 목록이고 대상 모집단은 15개의 텍스트 카테고리로 그룹화한다. 2,000개의 단어 중 500개의 샘플을 각 텍스트 카테고리에 추출하는 방법을 적용함으로써 약 100만 단어로 구성되어 있다.

제4절 Brown 계열 코퍼스

대상 모집단은 먼저 15개의 텍스트 카테고리로 그룹화하고 그 중 2,000개의 단어 샘플이 각 텍스트 카테고리에 비례하여 약 100만 단어가 되었다. Brown corpus는 영문 또는 다른 언어로 자료를 추가로 준비하고 그것을 표현하기 위한

표준을 설정하기 위해 비교 연구를 염두에 두고 구축했다. 완성 후 통용적 연구뿐 아니라 언어 간 대조를 위한 많은 자료에 사용되었다. <표 6-20>은 Brown 계열의 코퍼스 비교표이다.

<표 6-20> Brown 계열 코퍼스의 구성

코퍼스	언어	기간	샘플	단어 수
Brown	미국 영어	1961년	500	100만
Frown	미국 영어	1991년~1992년	500	100만
LOB	영국 영어	1961년	500	100만
Pre-LOB	영국 영어	1928년~1934년	500	100만
FLOB	영국 영어	1991년~1992년	500	100만
Kolhapur	인도 영어	1978년	500	100만
ACE	호주 영어	1986년	500	100만
WWC	뉴질랜드 영어	1986년~1990년	500	100만
LCMC	맨다린 중국어	1988년~1994년	500	100만

샘플링 기간은 동시 비교를 위해 비슷하거나, 통시적 비교(시대의 변화에 따라 달라지는 어느 시점부터 다른 시점까지의 시간을 비교)를 위해 약 30년 정도 기간차를 둔다. 코퍼스의 비교의 예를 들자면 Brown과 LOB를 사용하여 1960년대 초반의 미국 영어와 영국 영어를 비교할 수 있다. Frown과 FLOB을 사용하여 1990년대 초반의 미국 영어와 영국 영어를 비교할 수 있다. 자료를 동기적으로 비교할 때는 샘플링 기간이 유사하도록 주의를 기울여야 하고, Brown과 FLOB를 비교하는 것은 언어의 다양성 뿐 만 아니라 언어의 변화도 포함한다.

제5절 공시적 코퍼스

공시적 코퍼스는 현재 상황만 고려한 코퍼스로 역사적 관점과 시간의 흐름을 이해하여 생성한 통시적 코퍼스와 대비된다. 공시적 코퍼스는 현재 미디어나 도서 등의 다양한 플랫폼에서 획득할 수 있는 모든 언어 자원을 의미한다.

1. The International Corpus of English(ICE)

영어의 국제 코퍼스이다. The International Corpus of English(ICE)는 세계 영어의 동시성 연구를 위해 특별히 구축되었고 1990년~1994년 동안 작성된 서면 및 구어체 영어로 100만 단어의 20개 모음집으로 되어 있다. 음성 데이터가 60%를 차지하고, 서면 데이터가 40%를 차지한다. <표 6-21>은 ICE의 음성 데이터에 대한 상세한 구성 정보이다. 아래의 두 표에 명시된 괄호 안의 수는 텍스트 수를 의미한다.

<표 6-21> ICE 음성 데이터의 구성

음성 (300)	대화(180)	개인(100)	대화(90) 전화 통화(10)
		공공(80)	수업(20) 방송 토론(20) 방송 인터뷰(10) 의회 토론(10) 교차 검증(10) 상거래(10)
	독백(120)	즉흥(70)	주석(20) 즉흥적 연설(30) 데모(10) 법률 프리젠테이션(10)
		스크립(50)	방송 뉴스(20)방송 회담(20)비방송(10)
서면 (200)	비인쇄물(50)	학생 작문(20)	학생 에세이(10) 시험 스크립트(10)
		편지(30)	공식 편지(15) 비지니스 편지(15)
	인쇄물(150)	학술(40)	인문(10)사회 과학(10) 자연 과학(10)기술(10)
		대중(40)	인문(10) 사회 과학(10) 자연 과학(10)기술(10)
		보도(20)	언론 보도(20)
		교육(20)	관리(10)기술/취미(10)
		주장(10)	사설(10)
		창작(20)	소설(20)

2. The Longman/Lancaster Corpus

The Longman/Lancaster Corpus는 약 30만 단어로 영국 데이터 50%, 영어

데이터 40%, 호주, 아프리카 및 아일랜드 영어와 같은 다른 종류 데이터 10%로 구성되어 있으며, 샘플 중 절반을 무작위로 선택하고, 나머지 절반은 전문가 패널에 의해 선택되었다. 코퍼스는 창작물 텍스트와 정보제공 텍스트로 범주를 나눈다. 창작물 텍스트는 유명한 문학 작품에서 나왔으며 인쇄된 책에서 무작위로 추출한 작품이다. 주로 책 형태의 소설 작품을 의미한다. 정보제공 텍스트는 자연 및 사회 과학, 세계 정세, 상업 및 금융, 예술 여가 등에서 비롯되었고 책, 신문, 저널 등으로 구성되어 있다. 텍스트 선택에 있어서 지역, 기간, 매체, 기준 총 4가지의 외부 기준이 적용되었다. 언어의 다양성으로 인한 지역, 1900년부터 1980년대까지의 기간, 서적(80%), 정기 간행물(13.3%), 극단(6.7%)의 매체, 인기의 유무와 텍스트의 카테고리 유형에 따라 기준을 나누어 적용했다. 이 코퍼스는 the Longman Corpus Network의 일환으로 일반인이 이용할 수 없다.

3. The Longman Written American Corpus

The Longman Written American Corpus는 1억 단어가 넘는 코퍼스이며 현재 신문, 저널, 잡지, 베스트셀러 소설, 기술 및 과학 서적 등을 포함하고 있으며 새로운 자료가 계속 추가되면서 증가하고 있다. The Longman Written American Corpus은 일반적인 디자인과 영국 국립 코퍼스의 구성 요소를 기반으로 구축되었다. Longman Corpus Network의 다른 코퍼스와 마찬가지로 단어를 조합하여 새로운 단어 목록을 만들고 통계 기능을 분석하여 영국식과 미국식 영어 사용법을 비교해 볼 수 있다.

4. The CREA corpus

스페인어의 The CREA corpus이다. 5년 기간으로 나누어 1975년부터 1999년 사이의 기간의 1억 3,300만 단어를 포함하고 있다. 모든 스페인어 사용국가에서 작성된 다양한 서면 데이터와 음성 데이터로 구성되어 있으며 서면 데이터는 90%를 차지하고, 음성 데이터는 10%를 차지한다. 텍스트는 과학과 기술, 사회 과학, 종교, 기술, 정치와 경제, 미술, 건강 및 소설 등으로 되어 있으며 지속적으로 업데이트 되는 모니터 코퍼스로 설계되었다.

5. The LIVAC corpus

중국어의 the LIVAC corpus이다. 마다린 중국어를 공부하기 위한 동기식 코퍼스를 구축을 목적으로 하며 홍콩, 대만, 베이징, 상하이, 마카오 및 싱가포르의 6개 중국어 커뮤니티의 전자 매체와 마다린 중국 대표 신문을 통해 4일에 한번 정기적으로 데이터를 수집한다. 수집된 데이터에는 사설 및 첫 페이지의 모든 기사, 국제 및 지역 뉴스 페이지 그리고 리뷰가 포함되어 있다. 2003년 말을 기준으로 하여 코퍼스는 14억개의 한자를 포함하고 있으며 사전에는 64만개 단어가 포함되어 있다.

제6절 통시적 코퍼스

통시적 코퍼스는 언어의 진화의 변화를 추적하는데 사용되며, 시대의 변화에 달라지는 어느 시점부터 다른 시점의 기간의 언어에 대한 코퍼스이다. 따라서 시간별 코퍼스가 구분되어 구성된다.

1. The Helsinki corpus

8세기부터 18세기까지의 400개 텍스트 샘플링으로부터 약 150만 영어 단어로 구성되어 있으며 3가지 기간과 11개의 하위 기관으로 구분된다. <표 6-22>는 Helsinki Diachronic Corpus의 기간 구분에 대한 정보이다.

<표 6-22> Helsinki Diachronic Corpus의 기간별 정보

기간	세부 기간	단어	%	합계
초기 영어	~850년	2,190	0.5	413,250
	850년~950년	92,050	22.3	
	950~1050년	251,630	60.9	
	1050년~1150년	67,380	16.3	
	합계	413,250	100	26.27%

중기 영어	1150년~1250년	113,010	18.6	608,570
	1250년~1350년	97,480	16.0	
	1350년~1420년	184,230	30.3	
	1420년~1500년	213,850	35.1	
	합계	608,570	100	38.70%
현대 영어	1500년~1570년	190,160	34.5	551,000
	1570년~1640년	189,800	34.5	
	1640년~1710년	171,040	31.0	
	합계	551,000	100	100%

2. The ARCHER corpus

1650년부터 1990년을 다루는 역사시대에 샘플링된 1,037개의 텍스트로 되어 있으며 170만 단어를 포함하고 있다. 데이터는 영국 데이터와 미국 데이터로 구분되어 있으며, 영국이 데이터의 2/3을 차지하고 미국이 1/3을 차지한다. 영국 데이터는 저널, 일기, 편지, 소설, 뉴스 및 과학 등 7가지의 서면 데이터와 대화, 드라마, 설교 코퍼스인 3가지 언어 기반으로 되어있다. 미국 데이터는 1750년~1799년, 1850~1899년, 1950년~1990년 기간에만 사용할 수 있는 데이터로 한정되어 있다. 각 50년 하위 코퍼스는 약 2,000~3,000단어로 이루어진 10개의 텍스트로서 총 20,000~30,000 단어를 포함한다. 저작권 문제로 인해 공개적으로 제공되지 않고 있다.

3. The Corpus of Early English Correspondence

96개의 컬렉션으로 1417년~1681년 사이에 778명의 사람들이 작성한 6,000개의 개인 편지로 구성되어 있으며 270만 단어로 포함하고 있다. 발신자 데이터베이스가 수반되어 있어 발신자의 연령, 성별, 출생지, 직업, 사회적 계급, 소재지 및 수신자와의 관계를 비롯하여 다양한 정보를 엑세스 할 수 있다. 저작권으로 인해 전체의 자료 공개가 차단되어 있으며 일부만 공개되어 있다.

4. The Zurich English Newspaper Corpus(ZEN)

The Zurich English Newspaper Corpus(ZEN)은 영국 영어 신문사의 1671년~1791년(120년)의 기사를 다루고 있으며 120만 단어를 포함하고 있다. 다양한 종류의 신문과 신문 당 최대 10개의 이슈가 10년 간격으로 선정되었다. <표 6-23>는 ZEN 데이터의 구성이다.

<표 6-23> ZEN 데이터의 구성

섹션	기간	단어 수	문장 수
A	1670년~1709년	242,758	7,642
B	1710년~1739년	347,825	12,163
C	1740년~1769년	339,362	14,112
D	1770년~1799년	298,249	11,843
합계		1,228,194	45,760

5. 영어대화 코퍼스

1560년~1760년의 200년 동안 현대 영어 대화 텍스트로 구성되어 있으며 130만 단어를 포함하고 있다. 연극, 대화 형식의 핸드북, 언어 교과서 등 6가지 카테고리로 되어 있으며 Lancaster 대학교와 Uppsala 대학교에서 구축을 진행하고 있다.

제7절 음성 코퍼스

많은 연구와 개발을 통해 음성 데이터를 텍스트로 변환하거나(STT) 텍스트를 음성으로 변환하는(TTS) 기술이 꾸준히 발전해왔다. 따라서 음성 데이터도 자연어 처리에서 중요한 코퍼스 자원으로 사용되고 있다. 다음은 음성 코퍼스를 보유하고 있는 기관과 보유한 코퍼스에 대한 소개이다.

1. The London-Lund Corpus

The London-Lund Corpus이다. 1953년~1987년에 녹음된 영어 음성 코퍼스이며

5,000개 단어로 구성된 100개의 텍스트로 50만개의 단어가 구성되어 있다. 직접 대면 대화, 전화 대화, 공개토론 등 '대화'와 '독백'으로 구분되어 있다.

2. 10대 언어의 Bergen corpus

런던의 10대 언어의 Bergen corpus이다. 31명의 자발적인 지원자의 10대 이야기로 구성되어 있으며 약 55시간 분량인 50억 단어가 녹음되어 있다. 나이 그룹은 <표 6-24>와 같이 6개로 나뉘어져 있다.

<표 6-24> Bergen Corpus의 나이 그룹

기준	나이
사춘기 이전	0세 ~ 9세
이른 청소년기	10세 ~ 13세
중간 청소년기	14세 ~ 16세
늦은 청소년기	17세 ~ 18세
청년기	20세 ~ 29세
어른	30세 이상

코퍼스의 명칭이 10대 언어 코퍼스인 만큼 전체의 청소년이 94%를 차지하고 있다. 이른 청소년기(24%), 중간 청소년기(61%), 늦은 청소년기(9%), 성인그룹(6%)를 차지한다.

3. The Spoken corpus of the Survey of English Dialect

영어 방언을 연구하기 위해 제작된 음성 코퍼스이며 영국의 Servey of English Dialects 지역 289개에서 314개의 녹음 본으로 구성되어 있고 약 80만개 단어를 포함하고 있다.

4. The Intonational Variation in English Corpus

36시간의 음성데이터이고 5가지 다른 말하기 스타일을 포함하고 있다.

음성학적으로 제어되는 문장(구문 마커가 없는 질문, 육하원칙에 관한 질문), 읽기 텍스트, 같은 텍스트의 다시 말하는 버전, 예를 들어, “작은 마을을 돌아서 가세요.”와 같은 지도나 표지판의 문구 텍스트, 자유로운 대화이다. 116명이 참여를 하였고 이 코퍼스는 비상업적 용도로만 무료로 사용이 가능하다.

5. The HongKong Corpus of Conversational English(HKCCE)

총 341명이 참여 하였으며 130번의 개별 대화로 약 50시간 분량의 녹음으로 구성되어 있다. 대화 길이는 평균적으로 약 23분이고 코퍼스는 대화, 학술담화, 비지니스 담화, 공식적 담화와 같이 4개의 하위 카테고리로 나뉜다. 약 50만 단어로 되어 있으며 1990년대 중반 홍콩 중국인과 주로 영어를 모국어로 사용하는 사람 간의 대화를 통해 이루어진다. <표 6-25>는 HKCCE의 설계 기준이다.

<표 6-25> HKCCE의 설계 기준

기준	유형	%
성별	남성(모국어가 영어인 남성)	34
	여성(모국어가 영어인 여성)	18
	남성(모국어가 영어가 아닌 남성)	24
	여성(모국어가 영어가 아닌 여성)	24
나이	18세~29세	40
	30세~39세	21
	40세~49세	27
	50세~59세	10
	60세 이상	2
교육	중학교	8
	고등학교	5
	대학교	82
	기타	5

도메인	교육	35
	비지니스	23
	관리	11
	공학	8
	서비스 부문	7
	예술	5
	법률	3
	미디어	2
	항공 산업	2
	의료	2
	사용하지 않음	2
화자 수	2명	58
	3명	23
	4명	10
	4명 이상	9

제8절 학업 및 전문영어 코퍼스

코퍼스는 코퍼스를 형성하고 있는 주제나 작성 배경에 따라 특정 도메인을 지닌다. 이러한 도메인은 목적에 따라 한정되어 사용할 수 있다. 그 중 학업 및 전문영어 코퍼스는 일반 코퍼스보다 전문적인 영어 코퍼스를 제공한다.

1. The Michigan Corpus of Academic Spoken English

The Michigan Corpus of Academic Spoken English(Michigan 대학 학술 영어 코퍼스)는 1,571명의 발표자가 약 200시간 녹음한 152개의 사본으로 구성되어 있으며 약 180만개 단어가 포함되어 있다. 발표자의 연령은 17세~23세, 24세~30세, 31세~50세, 50세 이상과 같이 4가지로 나뉜다. 학력 또한 중학교, 고등학교, 대학교, 대학원생 및 연구원과 같이 4가지 그룹으로 나뉜다. 언어

상태는 원어민(북미 영어), 다른 원어민(비 미국 영어), 모국어가 아닌 사람 총 3가지 그룹으로 나뉜다. <표 6-26>은 The MICASE corpus에 대한 정보이다.

<표 6-26> MICASE corpus의 정보

기준	분포
화자 성별	남성(46%), 여성(54%)
교육	학부(49%), 학생(44%)
언어 지역	원어민(88%), 비원어민(12%)
학술 부문	인문학 및 예술(26%), 사회 과학 및 교육(25%), 생명 및 건강 과학(19%), 물리 과학 및 공학(21%), 기타(9%)
기초 담화 모드	독백(33%), 패널(8%), 대화형(42%), 혼합(17%)
음성 유형	상담(3.5%), 대화집(8.9%), 토론 섹션(4.4%), 논문 발표(3.4%), 면접(0.8%),대형 강의(15.2%), 작은 강의(18.9%), 회의(4.1%), 세미나(8.9%), 스터디 그룹(7.7%), 학생 발표(8.5%)

2. The British Academic Spoken English Corpus(BASEC)

Reading에서 63회, Warwick에서 97회 등 127회 녹음한 160개의 강의 및 39회 세미나 데이터로 구성되어 있다. 강의 및 세미나는 4가지 주제 영역으로 구분되며 <표 6-27>과 같다.

<표 6-27> BASEC corpus의 정보

주제 영역	강의	세미나
예술과 인문학	42	10
사회 및 과학	40	11
물리 과학	40	8
생명 및 의료과학	38	10
합계	160	39

3. The Academic Corpus

The Academic Corpus는 4개의 섹션과 28개의 주제 분야를 다루며 약 350만

단어를 포함하고 있다. 각 섹션은 각 12만 5천 단어로 7개 주제로 나뉘며 총 87만 5천단어로 구성되어 있다. 저널 기사, 서적, 설문조사, 통합문서, 실험실 매뉴얼 등 400명이 넘는 414개의 학술 문서로 되어 있으며 출처는 뉴질랜드(64%), 영국(20%), 미국(13%), 캐나다(2%), 호주(1%)이다. <표 6-28>는 The Academic corpus의 세부 구성이다.

<표 6-28> The Academic corpus의 세부 구성

분류	예술	경제	법률	과학	합계
텍스트	122	107	72	113	414
단어	883,214	879,547	874,723	875,846	351,333
주제영역	교육 역사 언어학 철학 정치 심리학 사회학	회계 경제학 금융 노사관계 관리 마케팅 공공 정책	헌법 형법 가정 법률 및 법의학 국제법 순수상법 준상법 권리 및 구제	생물학 화학 컴퓨터과학 지리학 지질학 수학 물리학	

4. The Corpus of Professional English(CPE)

100만 단어 코퍼스를 만드는 것을 목표로 과학, 공학, 법률 및 의학, 금융 등 실무 전문가와 훈련 전문가의 음성 및 서면 담화가 모두 포함될 것으로 예상되고, 1995년~2001년 사이의 출판된 미국 영어와 영국 영어로 1,000여 건의 연구 저널, 무역 잡지 및 교과서 등으로 구성되어 있다.

제9절 구문 분석 코퍼스

코퍼스는 원시 코퍼스로 존재하거나 혹은 특정 자연어 처리 기술이 거쳐 주석이 부착된 형태로 존재한다. 다음은 구문 분석 결과가 부착된 구문 분석 코퍼스이다.

1. The Lancaster-Leeds Treebank

Brown 계열의 코퍼스인 LOB 코퍼스의 모든 텍스트 카테고리에서 가져와 문구 14개, 문장 클래스 28개, 문장 부호 5개 총 47개의 레이블로 구성되어 있다.

2. The Lancaster Parsed Corpus

LOB 코퍼스의 13.29% 차지하는 145개 파일로 이루어져 있으며 구문 분석된 LOB 코퍼스에서 가져온 약 14만 4천단어로 구성되어 있다. Parser가 20단어~25단어 이상의 문장을 처리 할 수 없기 때문에 긴 문장을 제외하였으며, 자동 파싱으로 인한 오류는 코퍼스에 오류가 없도록 수기로 수정되었다.

3. The SUSANNE corpus

미국 영어 Brown 코퍼스에서 파생되었고 13만 단어의 하위 샘플이 있다. 카테고리는 보도자료, 전기 및 회고록, 글쓰기, 서양 소설로 나누어진 4가지 텍스트 카테고리에서 각각 16개를 추출한 총 64개의 텍스트 샘플로 구성되어 있다.

4. The Penn Treebank

최초 버전(Penn Treebank 1)은 450만 단어 이상의 미국 영어 데이터가 포함되어 있다. 전체 자료 POS 태그이며 데이터의 2/3는 구문분석이 되어있다. 이 모든 자료는 자동 처리 후 수작업으로 수정되었다. <표 6-29>은 Penn Treebank 1의 구성 요소이다.

Penn Treebank 1은 Lancaster 구문 분석 체계를 기반으로 확장되고 수정된 구문 분석 체계를 적용했다. 두 주석 방법 모두 명사, 동사, 형용사, 부사 및 전치사 구를 포함하는 구문 구조를 사용하지만 Lancaster 체계는 부사절, 비교절, 명사구 절 및 상대적 절과 같은 다른 유형도 구별하지만 Penn Treebank는 부정사 또는 필수의 이해대상을 나타내는 다양한 null 요소와 제로(Zero) 변형이 포함된다.

<표 6-29> Penn Treebank 1의 구성 요소

구성	태그된 단어	구문 분석된 단어
Dow Jones 뉴스	3,065,776	1,061,166
Brown 코퍼스	1,172,041	1,172,041
에너지 파트 요약	231,404	231,404
MUC-3 메세지	111,828	111,828
미국 도서관 텍스트	105,652	105,652
IBM 매뉴얼	89,121	89,121
농무부 게시판	78,555	78,555
ATIS 문장	19,832	19,832
WBUR 라디오 대본	11,589	11,589
합계	4,855,798	2,881,188

제10절 Developmental and learner corpus

는 아동이 모국어를 습득하여 구성된 자료를 의미하고, Learner corpus는 학습자가 제 2 언어를 습득하여 작성한 글 또는 연설로 구성된 자료이다. 따라서 국적, 연령, 언어 수준 등을 고려하여 코퍼스가 작성되었다.

1. The Child Language Data Exchange System

The Child Language Data Exchange System이다. 1차 및 2차 언어 습득 연구를 위해 구성된 국제 데이터베이스이며 제 1 언어와 제 2 언어를 모두 배우는 어린이와 성인으로부터 수집한 데이터이다. 약 2억 단어이며 25개의 언어를 지원하고 있다.

2. The Louvain Corpus of Native English Essays

The Louvain Corpus of Native English Essays이다. 네이티브 영국계 미국인

대학생들이 작성한 다양한 주제에 대한 논쟁적인 에세이 모음이다. 영어 학생 A 레벨 에세이 114개 (60,209단어), 영국 대학생 에세이 90개(95,695단어), 미국 대학생 에세이 232개(168,400단어)로 총 324,304단어로 구성되어 있다.

3. The Polytechnic of Wales corpus

The Polytechnic of Wales corpus이다. 사우스 웨일즈 주에서 1978년~1984년 사이에 수집되었으며 6세~12세 어린이 약 12만명에 대한 비공식 대화로 구성되어 있으며 65,000 단어가 포함되어 있다. 웨일스 말이나 다른 제 2 언어의 영향을 최소화하기 위해 선정되었으며 6세, 8세, 10세, 12세의 어린이를 각 30명씩 4개의 그룹으로 나누었고, 성별과 사회 경제적 계층으로도 세분화되어 있다. 코퍼스는 184개의 파일에 11,396의 구문 분석 트리를 포함한다.

4. The Cambridge Learner Corpus

전 세계 영어 글쓰기 예제 모음이며 2,000만 단어가 넘으며 계속 확장되고 있다. 전 세계 Cambridge ESOL 영어 시험을 치른 학생들이 작성한 익명의 시험 스크립트가 출처이며 이 코퍼스는 현재 150개국에서 50,000개의 스크립트를 포함하고 있다. 각 스크립트는 학생의 제 1 언어, 국적, 영어 수준, 나이 등의 정보로 코딩되어 있다. Cambridge 대학 출판사에서 개발한 학습자 오류 코딩 시스템을 이용하여 8억 단어가 오류 코드로 작성되어 있다.

5. The HKUST Corpus of Learner English

2,500만 단어의 에세이와 홍콩의 중등, 고등 중국어의 시험 스크립트로 구성되어 있으며 에세이의 평균길이는 1,000단어이다.

6. The Japanese EFL Learner corpus

3가지 구성요소로 이루어진 100만 단어로 구성되어 있다. 중학교 7학년~12학년에서 작성된 일본어 EFL 학습자에 의해 쓰여진 서면 데이터(40만 단어)와 음성 데이터(5만 단어) 그리고 일본에서 사용되는 15만 단어의 영어

교재를 포함하는 언어 코퍼스와 일반적인 일본어 작문에 사용되는 10만 단어로 구성되어 있다.

7. The Standard Speaking Test Corpus

The Standard Speaking Test Corpus는 일본어 학습자가 제작한 100만 단어의 말하기 오류 음성 데이터이며 '표준말하기 시험'이라는 영어 구술 면접 시험의 오디오 녹음을 근거로 하고 있다. 15분 구두 면접으로 1,200개의 샘플(총 300시간 녹음)으로 구성되어 있다. 담화 태그 지정과 오류 태깅이라는 두 가지 유형의 태깅을 사용하고 있다. 오류 태깅은 47개의 태그로 지정되어 있고 각 태그는 품사, 문법적/어휘 규칙 및 수정된 형식 3가지 유형의 정보를 표시하고 있다.

제11절 다국어 코퍼스

다국어 코퍼스는 기계 번역 등에 사용하기 위해 동일한 문서를 2개 이상의 언어로 작성한 코퍼스이다. 일반적으로 서로 다른 두 언어의 코퍼스 쌍을 병렬 코퍼스라고 한다.

1. The English-Norwegian parallel Corpus

The English-Norwegian parallel Corpus이다. 영어로 된 10,000 단어~15,000단어 사이의 100개 기존 텍스트와 그것을 번역한 텍스트로 구성되어 있고 260만 단어를 포함하고 있다. 텍스트는 소설과 비소설 텍스트로 샘플링을 하며 소설에는 어린이 소설, 탐정 소설, 일반적 소설 등이 있고 비소설에는 지역, 사회 과학, 법률 자연 과학, 의학 및 예술 등이 있다.

2. The English-Swedish Parallel Corpus

64개의 영어 텍스트 샘플을 스웨덴 언어로 번역을 하고, 72개의 스웨덴어 텍스트 샘플을 영어로 번역한 것으로 구성되어 있으며 280만 단어가 포함되어

있다. 샘플은 소설과 비소설 두 가지 텍스트 카테고리에서 가져왔다. 어린이 소설, 범죄/미스테리 소설, 일반적인 소설이 소설에 포함되며 생물학, 지리학, 자연 과학, 사회 과학, 법률 문서, 연설문 등이 비소설에 포함된다.

3. The Oslo Multilingual Corpus

The Oslo Multilingual Corpus는 현재 구축중인 코퍼스이며, 여러 언어와의 병렬 구성이다. <표 6-30>은 구성 정보이다.

<표 6-30> Oslo Multilingual 코퍼스의 정보

구분	단어 수
영어 - 독일어	130만
프랑스어 - 노르웨이어	50만
독일어 - 노르웨이어	150만
영어 - 네델란드어	30만
노르웨이어 - 프랑스어 - 포르투갈어	150만
영어 - 필란드어	30만

4. The IJS-ELAN Slovene-English Parallel Corpus

15개 언어로 된 텍스트로 구성되어 있으며 100만 단어를 포함한다. 전체 코퍼스 중 11개 슬로베니아어 텍스트와 그것을 영어 번역 텍스트로 되어 있고, 나머지는 영어의 텍스트와 그것을 슬로베니아어로 번역한 텍스트이다.

5. The CLUVI parallel corpus

소설, 저널리즘, 법률 및 행정 분야에서 가저온 전문 단어 공개 텍스트 코퍼스이며 800만 단어를 포함하고 있다. 5개의 주요 섹션으로 구분되어 있으며 <표 6-31>에 구성 정보이다.

코퍼스는 갈라시아어 - 스페인어 경제 텍스트, 영어 - 포루투갈어 문학 텍스트, 영어 - 스페인어 문학 텍스트, 영어 - 갈라시아어 영화 자막 4개의 섹션으로 확장된다.

<표 6-31> CLUVI 코퍼스의 정보

구성 정보	단어 수
갈라시아어 - 스페인 법률 병렬 텍스트	190만
영어 - 갈라시아어 - 프랑스 - 스페인 과학기술 병렬 텍스트	180만
영어 - 갈라시아어 문학 병렬 텍스트	130만
프랑스 - 갈라시아어 문학 병렬 텍스트	130만
바스트 - 스페인 법률 병렬 텍스트	240만

6. The PAROLE corpus

카탈로니아어, 덴마크어, 네덜란드어, 영어, 프랑스어, 핀란드어, 독일어, 그리스어, 아일랜드어, 이탈리아어, 노르웨이어, 포르투갈어, 스웨덴어를 포함한 14개 언어가 포함되어 있으며 1996년~1998년까지 동일한 구성을 이용하여 각 언어에 대해 2,00만 단어와 20,000개 항목이 들어있는 사전을 작성했다. 이 코퍼스에는 도서 카테고리(20%), 신문(65%), 정기 간행물(5%), 기타(10%)의 텍스트가 특정 비율로 포함되어 있다.

7. The EMILLE corpus

The EMILLE corpus는 남아시아 언어의 언어자원을 개발하는 Enableing Moinority Language Engineering 프로젝트의 결과물이며 비상업용, 연구용으로 무료로 배포되는 EMILLE/CIIL corpus와 상업용으로만 배포되는 EMILLE/Lanscater corpus가 있다.

가. EMILLE/CIIL corpus

EMILLE/CIIL corpus는 3가지 요소로 구성되어 있다. 단일 언어, 병렬 및 주석 처리된 코퍼스, 14개 단일 언어를 포함한 14개의 단일 언어 자료가 있다. 약 900백만 단어가 포함되어 있다. 병렬 자료는 영어와 함께 번역된 힌디어, 벵갈어 등 20만 단어의 텍스트로 구성된다.

나. EMILLE/Lancaster corpus

EMILLE/Lancaster corpus도 3가지 요소로 구성되어 있다. 단일 언어, 병렬

코퍼스, 주석 처리된 코퍼스이다. 7개의 남아시아 언어를 포함하는 단일 언어로 구성된 EMILLE/CIIL corpus와 다른 5,800만 단어가 포함되어 있으며 병렬 코퍼스와 주석 처리된 코퍼스는 EMILLE/CIIL corpus와 동일하다.

8. Hong Kong parallel Text

2004년 LDC가 발표한 대규모 병렬자료 모음이다. 이 자료집에는 약 5,900만개의 영어 단어와 4,900만개의 중국어 단어가 포함되어 있다. 2000년에 발행된 3가지 병렬 자료인 홍콩 한사드(Hong Kong Hansards), 홍콩 법률(Hong Kong Law), 홍콩 뉴스(Hong Kong News)의 업데이트로 구성된다. 홍콩 한사드는 1985년 10월~2003년 4월까지의 HKSAR 입법회의의 공식 기록에서 발췌한 내용이 포함되어 있으며 3,600만개의 영어 단어와 5,600만개의 한자가 포함되어 있다. 홍콩 법률은 HKSAR 법무부에서 2000년 까지 발행한 영어와 중국어의 홍콩 법령, 헌법 문서, 국가 법률 및 기타 관련 문서가 포함되어 있으며 8,300만개의 영어 단어와 14,800만개의 한자가 포함되어 있다. 홍콩 뉴스에는 1997년 7월~2003년 10월까지 HKSAR의 정보 서비스 부서에서 14,700개의 영어 단어와 26,600개의 한자에 대한 보도 자료가 포함되어 있다.

제12절 영어 이외의 단일 언어 코퍼스

일반적으로 영어로 생성된 언어 자원이 다른 언어에 비해 많이 존재한다. 특히 자연어처리 연구가 뒤 늦게 시작된 나라일수록 모국어 언어 자원이 부족한 상황이다. 따라서 자연어처리 기술 발전을 위해 일부 특별하게 관리되는 영어 외 단일 언어 코퍼스가 존재한다.

1. The COSMAS corpus

Corpus Search, Management and Analysis System(COSMAS)는 독일 텍스트 코퍼스의 대규모 모음이다. 약 20억 단어를 포함하고 있으며 언어 연구를 위한 계속 규모가 커지는 컬렉션이다. 다양한 출처를 다루고 있는데 고전 문학 텍스트, 국가 및 지역 신문, 구문론적 주석이 달린 텍스트 그리고 독특한

코퍼스를 포함하고 있다. COSMAS 컬렉션은 COSMAS 온라인 툴을 이용해 무료로 공개적으로 사용할 수 있다. 비상업적 용도로만 사용할 수 있으며 익명의 사용자에게는 60분으로 제한을 둔다.

2. The CETEMPúblico Corpus

The CETEMPúblico Corpus(Corpus de Extractos de Textos Electrónicos MCT/Público)는 1991년~1999년 사이에 작성된 포르투갈어 일간지 Público의 약 2,600여판의 텍스트를 포함하며 18,000만 단어를 포함하고 있다. 연구 및 기술 개발에 사용할 수 있지만 상업적 용도는 허용되지 않는다.

3. The INL corpus

네덜란드어 어휘 학회(INL)은 웹을 통해 3개의 코퍼스를 제공했다. 500만 단어 코퍼스는 다양한 구성을 가지고 있다. 저널리즘, 정치, 환경, 언어학, 레저 및 비지니스/고용과 같은 주제를 다루는 책, 잡지, 신문 및 TV 방송을 포함하여 1989년~1994년에 작성된 17개의 텍스트 출처에서 나온 현재 네덜란드 텍스트로 구성된다. 2,700만 단어의 네덜란드 신문 코퍼스는 1994년~1995년에 주요 전국 신문인 NRC에 의해 발행된 신문 텍스트로 구성되어 있다. 3,800만 단어 코퍼스는 1996년에 3가지 주요 구성 요소를 가지고 있다. 1970년~1995년까지의 다양한 구성(책, 잡지, 신문 텍스트, TV 방송, 국회 보고서)으로 인한 1,270만 단어와 1992년~1995년의 신문의 1,240만 단어, 네덜란드 법조계 조약인 1,290만 단어의 초기 문서로 되어 있다. 이 코퍼스는 비상업적인 연구 목적으로만 이용이 가능하다.

4. The CEG corpus

CEG corpus(Cronfa Electroneg o Gymraeg)는 100만 단어의 웨일스어 산문이 포함되어 있다. 코퍼스는 웨일즈어의 산문으로 쓰였으며 1970년 이후의 현대적인 글을 보여주기 위해 대표적인 텍스트에서 선택된 500개의 2,000 단어 샘플로 구성된 Brown과 LOB코퍼스와 동일하게 웨일스어로 설계되었다. 그러나 텍스트 범주와 비율은 Brown 및 LOB와 다르다. 사실적인 산문과 소설의 두

가지 범주로 분류된다. 소설에는 소설이나 단편소설과 같은 7가지 유형의 소설이 포함되어 있으며 사실적인 산문은 다양한 종류의 언론 자료, 행정 문서, 학술 텍스트 와 같이 22가지 카테고리로 나뉜다.

5. The Scottish Corpus of Texts and Speech

The Scottish Corpus of Texts and Speech(SCOTS)는 스코틀랜드의 언어에 대한 글과 음성 모두의 대형 전자 코퍼스를 만드는 것을 목표로 하고, 현재 진행 중인 프로젝트이다. 프로젝트의 첫 번째 결과물로 113명의 사람들이 참여한 385개의 텍스트(약 50만 단어)로 구성되고 대화, 인터뷰, 시, 소설 및 산문으로 구성되며 특히 시 또는 문학 산문이 많이 사용되었다. 두 번째 결과물로 800개의 텍스트, 400만 단어이며 이 중 20%는 음성으로 되어 있다.

6. Academia Sinica Balanced Corpus

Academia Sinica Balanced Corpus(ASBC)는 현대 중국어의 첫 번째로 주석이 달린 코퍼스이다. 코퍼스는 대만에서 사용되는 만다린 중국어의 대표적인 샘플이다. 장르, 스타일, 모드, 주제, 출처로 5가지 기준에 따라 500만 단어의 텍스트가 서로 다른 영역에서 샘플링된다. <표 6-32>는 텍스트 범주의 비율을 보여준다.

<표 6-32> Academia Sinica Balanced 코퍼스의 정보

기준	종류
유형	보도(56.25%) / 보도 리뷰(10.01%) / 광고(0.59%) / 편지(1.29%) / 소설(10.12%) / 에세이(8.48%) / 전기 및 일기(0.50%) / 시(0.29%) / 설명서 (2.03%) / 재생 스트립트(0.05%) / 공공 연설(8.19%) / 대화(1.34%)
스타일	서술 텍스트(70.66%) / 논쟁 텍스트(12.24%) / 해설 텍스트(14.72%) / 설명 텍스트(2.83%)
방법	글(90.14%) / 읽기 위한 글(1.38%) / 말하기 위한 글(0.82%) / 음성(7.29%) / 읽기 위한 음성 (0.35%)
이야기	철학(8.68%) / 자연 과학(12.97%) / 사회 과학(34.99%) / 예술(9.28%) / 일반레저(17.89%) / 문학(16.20%)
출처	신문(31.28%) / 일반 잡지(29.18%) / 학술저널(0.70%) / 교재(4.08%) / 참고 도서(0.13%) / 논문(1.36%) / 일반 책(8.45%) / 오디오 및 비디오 매체(22.83%) / 대화 및 인터뷰(1.63%) / 공공 연설(0.25%)

7. Spoken Chinese Corpus of Situated Discourse

Spoken Chinese Corpus of Situated Discourse(SCCSD)는 중국 사회 과학 아카데미의 후원 하에 진행되는 프로젝트로, 중국에서 사용되는 만다린어를 1,000시간동안 녹음한 코퍼스를 목표로 하고 있다. 코퍼스는 3가지 하위 코퍼스를 가지며 워크샵 담화를 위한 코퍼스, 중국의 주요 방언을 위한 코퍼스, 연설을 위한 코퍼스로 구성되어 있다. 현재 650시간의 오디오와 150시간의 비디오 녹음이 수집되었다.

제7장 결론 및 시사점

4차산업혁명 시대의 시작과 함께 현재 세계 각국은 자국의 언어지능 개발을 위해 언어자원을 조직적인 활동을 통해 확보하고 있다. 자연어 처리의 경우에는 현재 영어에 대한 자원이 가장 많지만, 각국은 다국어 자원 구축을 일반적으로 추진하는 경향을 보이고 있다. 또한 이에 따라 영어권 국가에서도 한국어 자원을 구축하는 움직임도 포착되고 있어 우리의 대응 또한 필요하다.

자연어는 비정형 데이터이자 자연어를 이루는 어휘, 어절, 품사, 격 등은 그 양이 매우 방대하여 빅데이터의 특성을 고스란히 갖는다. 이 데이터를 다루고 분석하기 위한 방편으로는 지능정보기술인 인공지능, 클라우드, 빅데이터 등의 기술들의 발전과 지원이 필수적으로 동반되어야 한다.

최근 하드웨어의 발전과 딥러닝(Deep-Learning) 기술의 출현으로 대규모 학습에 대한 여건은 과거보다 더 잘 갖추어 졌다. 현재의 딥러닝은 지난 알파고에서처럼 그 성능의 완성도를 점차 입증해가고 있다. 앞으로는 언어 자원을 구축하고 언어에 대한 학습과 정확한 이해, 이를 기반으로 한 최적화된 서비스의 실현이 더욱 가속화 될 것이다.

또한, 최근 앞 다투어 공개되고 있는 인공지능 알고리즘의 활용으로 상대적인 기술 진입장벽이 낮아진 이때에, 데이터에 대한 확보와 무결성에 대한 중요성은 더욱 높아지고 있다. 앞으로는 얼마나 더 다양하고 정확한 데이터의 확보 및 관리가 자연어처리 기술 발전에 더 큰 영향을 미칠 것이다.

한국어의 경우, 지난 1998년부터 2007년까지 수행한 ‘21세기 세종 계획’의 결과물이 공개되어 있으나 다수의 오류로 인해 외면 받고 있는 실정이다. 한국어 표준 말뭉치에 대한 제대로 된 표준화 작업과 꼼꼼하게 계획한 투자가 있어야 할 것으로 사료된다.

최근 각국의 다국어 분석 방법의 발달로 인하여 한국어 분석기만의 효용성은 많이 줄어들었다. 다국어 분석 방법을 골자로 하되, 효용성 차원의 국내 서비스와 한국어 심층분석에서 필요한 기능에 집중할 필요가 있다. 또한 인공지능 기술 발전을 위해서 연구 개발팀 간의 shared task 형태의 workshop 등과 같은 적극적인 교류가 필요하다.

참 고 문 헌(Reference)

[1] https://ko.wikipedia.org/wiki/자연_언어_처리

[2] http://www.w3ii.com/ko/artificial_intelligence/artificial_intelligence_natural_language_processing.html

[3] http://nlp.stanford.edu/projects/coref.shtml

[4] “Word Embeddings 자질을 이용한 한국어 개체명 인식 및 분류”, 최윤수, 차정원.

[5] “CRFs 기반의 한국어 의미역 부착 성능 향상을 위한 자질 선택”, 박태호, 차정원.

[6] “순차적 레이블링을 이용한 한국어 의미격 인식”, 임수종, 김현기.

[7] http://ltrc.iiit.ac.in/icon2016/

[8] https://www.aclweb.org/

[9] http://www.mitpressjournals.org/

[10] http://cljournal.org/

[11] http://acl2016.org/index.php?article_id=46

[12] https://sites.google.com/site/2016hclt/home

[13] http://amitavadas.com/Code-Mixing.html

[14] http://www.evalita.it/

[15] http://facta-evalita20

[16] https://github.com/wikilinks/neleval/wiki/Evaluation

[17] http://corpora.ficlit.unibo.it/PoSTWITA/index.php?slab=guidelines

[18] http://corpora.ficlit.unibo.it/PoSTWITA/index.php?slab=data

[19] https://tac.nist.gov//

[20] https://ithub.korean.go.kr/user/contest/contestIntroView.do

[21] http://www.venturewerks.com/blog-siriprimer.html

[22] http://keyhole.co/

[23] http://eng.signalmedia.ru/

[24] https://www.brandwatch.com/

[25] http://stanfordnlp.github.io/CoreNLP/

[26] http://www.nltk.org/

[27] https://github.com/nltk/nltk/wiki/Dependency-Parsing

[28] http://www.cse.chalmers.se/~peb/

[29] https://github.com/nltk/nltk/wiki/Sentiment-Analysis

[30] http://www.dhgarrette.com/

[31] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O’Reilly Media Inc.

[32] http://konlpy.org/ko/latest/

[33] “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 박은정, 조성준, 2014

[34] https://github.com/tensorflow/models/blob/master/syntaxnet/universal.md

[35] http://opennlp.apache.org/

[36] http://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html

[37] https://uima.apache.org/

[38] http://mallet.cs.umass.edu/

[39] https://air.changwon.ac.kr/software/

[40] http://kkma.snu.ac.kr/

[41] 이동주, 연종흠, 황인범, 이상구, 꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구, 2010, 정보과학회논문지: 컴퓨팅의 실제 및 레터 (Journal of KIISE: Computing Practices and Letters), Volume 16, No.11, Page 1046-1050

[42] http://shineware.tistory.com/category/Project/Library

[43] http://islab.donga.ac.kr/demo2.htm

[44] http://nlp.stanford.edu/

[45] http://nlp.cs.berkeley.edu

[46] http://nlp.csail.mit.edu/

[47] http://nlp.shef.ac.uk/

[48] https://www.cs.washington.edu/research/nlp

[49] https://www.microsoft.com/en-us/research/

[50] http://researcher.watson.ibm.com/researcher/view_group.php?id=147

[51] http://www.diquest.com/

[52] https://www.etri.re.kr/korcon/sub7/sub7_02.etri

[53] IDG Tech Report - “글에서 감정을 읽다” 감성 분석의 이해

[54] https://ilis.yonsei.ac.kr/subbbs/sub.php?grid=g7&top_id=g7_b2&skin=basic1&id=2

[55] http://www.korean.go.kr/nkview/nklife/2016_2/26_0204.html

[56] http://m.blog.naver.com/htiger31/220669389279

[57] http://www.kristalinfo.com/download/

[58] http://nlplab.ulsan.ac.kr/doku.php?id=start

[59] 고려대학교 한국어 말모둠1 설계 및 구성(강범모, 김홍규) 1996.

[60] 한국어 용언 위계구조 자동구축, 송상헌, 최재웅, 2012

[61] https://www.ldc.upenn.edu/

[62] http://semanticweb.kaist.ac.kr/home/index.php/Imagedata

[63] http://semanticweb.kaist.ac.kr/org/bora/04_01.php

[64] https://www.ethnologue.com/statistics/size

연구보고서 2016-006

지능정보기술 동향 조사 - 자연어 처리를 중심으로

2017년 05월 인쇄
2017년 04월 발행

발행처 정보통신산업진흥원 부설 소프트웨어정책연구소
경기도 성남시 분당구 대왕판교로712번길22 A동 4층
Homepage: www.spri.kr

ISBN : 978-89-6108-367-6

주 의

1. 이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
2. 이 보고서의 내용을 발표할 때에는 반드시 소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.

ISBN : 978-89-6108-367-6

[소프트웨어정책연구소]에 의해 작성된 **[SPRI 보고서]**는 공공저작물 자유이용허락 표시기준 제 4유형**(출처표시-상업적이용금지-변경금지)**에 따라 이용할 수 있습니다.
(출처를 밝히면 자유로운 이용이 가능하지만, 영리목적으로 이용할 수 없고, 변경 없이 그대로 이용해야 합니다.)