뉴스 기사의 자연어처리: 〈뉴스소스 베타〉를 중심으로*

박대민** 한국언론진흥재단 선임연구위원

뉴스 기사가 빅데이터화함에 따라 뉴스 분석에서 컴퓨터 보조 질적 자료분석 소프트웨어의 사용이나 컴퓨터 이용 내용분석, 의미연결망분석 등의 활용이 늘어나면서 그 과정에서 자연어처리를 이용하는 경우도 증가하고 있다. 하지만 일반적으로 언론학에서 자연어처리는 하나의 블랙박스로 간주되어 방법론적 절차에 대한 엄밀한 검토가 부족하다. 또한 다양한 주제에 대한 높은 수준의 논증을 담은 뉴스 담론분석을 위해서는 단어 중심의 구문분석에 초점을 둔 형식 언어학적 접근이 아니라, 개체명과 문장 수준에서 관계를 부여하고 가중치를 계산하는 데 필요한 자연어처리가 필요하다.

이에 따라 이 연구에서는 뉴스에 대한 컴퓨터 이용 내용분석을 위하여 개체명, 특히 정보원과 문장 수준의 분석에 초점을 둔 뉴스 빅데이터 분석시스템인 〈뉴스소스 베타〉를 소개한다. 〈뉴스소스 베타〉는 정보원 중심의 개체명 인식과 이에 따른 문장 다중분류, 저널리즘 관행에 따른 최소한의 부분 구문분석을 바탕으로 하는 의미중의성 해소와 대용어 해소, 군집화를 통한 중복기사와 중복 문장의 제거, 저널리즘 가치에 따라 정의된 뉴스 정보원 연결망분석 알고리즘을 통한 가중치 부여를 특징으로 한다. 이 연구는 〈뉴스소스 베

* 이 연구는 2013년도 미래창조부와 한국정보화진흥원의 〈빅데이터 활용 스마트서비스 시범사업〉의 지원을 통하여 연구되었으며 2015년 한국언론학회 봄철 정기학술대회에서 발표된 논문 '뉴스 자연어처리: 〈뉴스소스 베타〉를 중심으로'를 대폭 수정한 것입니다. 이 연구는 2016년 서울대학교 언론정보연구소 연구기금의 지원을 받았습니다.

** heathe1@snu.ac.kr; heathe0@gmail.com

타〉의 자연어처리 알고리즘을 설명하고, 분석사례를 소개한 뒤, 뉴스 자연어처리 성능을 개선하기 위해 필요한 점들에 대해 제언한다.

핵심어: 뉴스 자연어처리, 〈뉴스소스 베타〉, 컴퓨터 이용 내용분석, 담론분석, 뉴스 빅데이터 분석, 개체명 인식

1. 문제제기

언론학에서 컴퓨터 이용 내용분석(computerized content analysis), 컴퓨터 보조 질적 자료분석 소프트웨어(Computer Assisted Qualitative Data Analysis Software, CAQDAS), 의미연결망분석(semantic network analysis) 등의 활용이 늘어나면서 자연어처리(Natural Language Processing, NLP)를 하는 경우도 증가하고 있다. 하지만 언론학에서 수행하는 NLP는 하나의 블랙박스로 취급되어 방법론적 절차에 대한 엄밀한 검토가 부족하다. 즉 NLP 프로그램이 텍스트의 요소를 얼마나 정확하게 재현하는가 하는 성능 문제, 언론학 관점의 뉴스 분석을 위해 적절한 분석단위가 무엇인가 하는 분석수준 문제, 그리고 이러한 분석수준에 맞는 NLP를 위해 각 단계가 어떻게 제한적, 또는 추가로 수행되어야 하는가 하는 최적화 문제 등이 간과되는 경향이 있다.

이 연구에서는 특히 대규모 뉴스를 분석하기 위한 뉴스 빅데이터 분석의 방법으로서 NLP를 소개한다. 이를 위해 먼저 NLP의 개념과 절차를 뉴스 분석의 관점에서 살펴본다. 이어 언론학에서 수행해 온 내용분석(content analysis)과 담론분석(discourse analysis)의 연구주제를 간략히 검토하여 언론학의 뉴스 분석은 일반적인 언어학의 담론분석보다 복잡한 일종의 고도의 텍스트 추론(text inference)을 요구한다는 점을 지적할 것이다(최성필 · 송사광 · 정한민 · 황미녕, 2012).

언론학적 관점에서 뉴스에 담긴 복합논증을 분석하기 위해서는 형태소나 단어 중심의 완전 구문분석(complete parsing)에 초점을 둔 NLP로는 한계가 있다. 이 논문은 뉴스 분석을 위해 우선 모든 단어의 인식 성능이 아니라 인명(person), 장소(location), 조직(organization) 등 저널리즘 관행에 중요한 특정 유형의 개체명 인식(Named Entity Recognition, NER) 성능에 집중해야 한다고 본다. 또 문법적으로 치밀한 완전 구문분석보다는 최소한의 규칙(rule)에 따른 부분 구문분석(partial parsing)이 적합하다고 주장한다. 이러한 규칙은 저널리즘 관행에 대한 영역지식

(domain knowledge)에 의해 부여된다.

한편 이러한 관점에서 뉴스 NLP를 구현한 사례로 컴퓨터 보조 질적 자료분석 소프트웨어인 〈뉴스소스 베타〉의 NLP를 소개한다. 〈뉴스소스 베타〉의 NLP는 형태소분석(morphological analysis)과 부분 구문분석, 문장경계 인식(sentence boundary disambiguation), 개체명 인식, 문장 및 지면 분류(classification), 의미 중의성 해결(Word Sense Disambiguation, WSD)이나 대용어 해소(coreference resolution) 등의 의미분석(semantic analysis), 유사도(similarity), 군집화(clustering), 의미연결망분석과 가중치 부여(weighting), 순위화(ranking) 등을 포함한다. 이를 통해 〈뉴스소스 베타〉는 다량의 기사에 대한 담론분석을 정보원과 인용문을 중심으로 최소한의 인력과 비용으로 단기간에 수행할 수 있도록 도와준다.

끝으로 이 연구에서는 본격적인 분석은 아니지만, 〈뉴스소스 베타〉의 활용 사례를 간략하게 제시한다. 구체적으로는 2013년 6월 1일부터 1주일간 49개 매체의 '대통령' 관련 기사 1,383개에서 전체 정보원 729명에 대한 시각화와 개인실명정보원 575명, 2,482개 인용문에 대한 추출사례를 기술해 볼 것이다.

2. 뉴스 자연어처리의 개념과 절차

1) 자연어처리의 개념과 이론적 전개

NLP란 컴퓨터를 이용하여 사람 언어의 이해, 생성 및 분석을 다루는 인공지능 기술을 뜻한다(〈한국정보통신기술협회 IT용어사전〉, 2015). 여기서 자연어(natural language)란 인간 사회가 형성되면서 자연발생적으로 생겨나 진화한, 의사소통을 행하기 위한 수단으로서 사용되는 언어를 뜻한다. 자연어는 컴퓨터 프로그래밍을 위하여 특별히 개발된 인공어(artificial language) 또는 프로그래밍 언어(programming language)와는

대비되는 개념이다(김영택 · 권혁철 · 옥철영 · 서영훈 · 이호석 · 이근배 · 윤덕호 · 문유진 · 강승식 · 이하규 · 심광섭 · 장병탁 · 윤성희 · 양재형 · 서병락 · 양승현 · 이재원 · 김성동 · 김유섭 · 박성배 · 이종우 · 장정호 · 오장민 · 황규백 · 김선 · 신형주, 2001, 17쪽).

NLP는 일찍이 컴퓨터가 등장한 1940년대부터 시작됐다. 초기에는 실용적인 관점에서 이중어 사전(bilingual dictionary)을 이용한 기계번역 시도가 많았다. 그러나 NLP 이론은 언어학과 긴밀한 관계를 맺으며 체계적으로 발전해 왔다.

1950년대는 구조주의의 시대였다. 유럽에서는 프라하학파(Praque school)가 언어를 분석(analysis)하여 음의 변별적 자질(distinct feature)을 규명하고 음의 교체(commutation)에 의한 의미 변화를 파악하고자 했다(Vachek, 1966). 이는 언어를 작은 의미소 단위로 분석하여 설명할 수 있음을 시사하는 것이었다.

미국에서는 해리스가 실증주의에 입각하여, 언어자료를 수집하고, 분석과 교체를 통해 형태소 등을 추출한 뒤, 문장의 직접구성요소(immediate constituent)나 문법을 나타내는 구 구조규칙(phrase structure rules)을 재구성하는 구조주의 방법론을 주창했다(Harris, 1952). 이는 1990년대 컴퓨터를 이용한 대규모 말뭉치 분석(corpus analysis)과도 유사하지만, 당시에는 컴퓨터 성능의 한계로 언어자료가 불완전하다는 문제가 있었다.

1960년대에 촘스키가 변형생성문법(transformational generative grammar)을 제안했다(Chomsky, 1957, 1964). 그는 언어를 심층구조와 표층구조로 나누었다. 인간은 심층구조의 언어를 사용할 수 있는 생래적인 언어능력(competence)을 갖고 있으며, 언어수행(performance)을 통해 심층구조의 문장을 변형(transformation)규칙에 따라 표층구조의 문장으로 표현한다. 즉 발화자가 내용을 입력하면, 구 구조규칙을 저장한 구문부와 어휘를 저장한 사전에서 통사규칙과 어휘를 선택하고, 통사변형을 거친 뒤 발화체로 출력된다. 이러한 과정은 컴퓨터의 입출력 과정과 유

사하다는 점에서 전산언어학에 응용되기 쉽다. 그러나 이 모형에서는 자연어가 인공어로 과도하게 환원되는 한계가 있다. 인간은 언어능력 외에 지식(knowledge), 추론(inference), 그리고 지능(intelligence)을 바탕으로 언어활동을 하기 때문에 기계가 인간처럼 언어를 이해하고 생성할 수 없으리란 비판이 제기됐다(ALPAC, 1967).

1960년대에는 SAD-SAM, BASEBALL, STUDENT, ELIZA 등 다양한 NLP 시스템이 개발되기 시작했다. 이어 1970년대에는 복잡한 구문분석과 의미분석, 잘 정의된 어휘 지식과 실세계 지식을 활용하여 특정 분야(domain)에 특화된 NLP 연구가 활발하게 진행됐다(김영택 등, 2001, 32-33쪽).

1970년대 이후 1980년대까지는 점차 변형을 최소화한 통합기반문법(unification-based grammar)이 부상했다. 대표적으로 변형과 심층구조 없이 일반원리와 형식적 제약 위주로 분석하는 일반 구 구조문법(Generalized Phrase Structure Grammar, GPSG)을 비롯하여, GPSG를 모태로 중심어 역할을 중시하여 복잡한 문법규칙을 간소화한 중심어 주도 구 구조문법(Head-driven Phrase Structure Grammar, HPSG), 문장성분과 기능 등 어휘의 정보를 중시하는 LFG(Lexical Functional Grammar) 등의 문법이론이 등장했다(김영택 등, 2001; Gazdar, 1985; Kaplan & Bresnan, 1982; Pollard & Sag, 1994). 통합기반문법에서는 '자질-값' 쌍에 기초한 자질구조와, 자질구조의 정보결합 연산으로서 통합을 이용한다. 기본적으로 문맥자유문법(context-free grammar), 즉 문맥에 영향을 받지 않는 문법규칙을 이용하되, 어휘와 의미를 강조하여 문맥의존적인 요소를 자질구조를 이용해 해결한다.

1990년대에는 컴퓨터 성능이 크게 개선됨에 따라 구조주의에서 시도됐던 말뭉치를 다양하게 대규모로 구축하여 통계적으로 처리하는 NLP 방식이 발전했다. 특히 태깅된 말뭉치(tagged corpus), 이중언어 말뭉치(bilingual corpus) 구축이 활발하게 이루어졌다. 말뭉치는 어휘 간 관계를 통계적으로 분석해 품사나 의미를 파악하는 데 활용했다. 그러나

충분한 양의 말뭉치를 모으려면 여전히 막대한 비용이 들고, 개인별, 지역별, 영역별로 말뭉치가 크게 달라지며, 언어가 지속적으로 변화하기 때문에 말뭉치를 끊임없이 수정해 주어야 한다는 문제가 있었다. 또한 NLP의 모든 문제를 알고리즘을 통해 통계적으로 처리할 수 없는 데다 알고리즘의 성능 자체도 한계가 있다.

2000년대 이후에는 기계학습(machine learning)이 빠르게 발전하고 있다.[1] 최근에는 인공지능 연구가 활발해지면서 휴리스틱에 기초한 규칙, 수작업으로 구축된 온톨로지나 사전, 말뭉치 등에 대한 의존도를 최소화하고, 컴퓨터가 지속적으로 추가되는 문서를 바탕으로 스스로 학습하여 숨은 의미패턴을 자동으로 찾고 개선하는 딥 러닝(deep learning)이 본격적으로 연구되고 있다.

2) 뉴스 자연어처리의 일반적 절차

NLP 대상에는 음성도 포함될 수 있지만 이 연구에서는 텍스트의 NLP를 다룬다. 텍스트의 수준(level)에 따라 문자 또는 문자열(string), 형태소(morpheme), 단어(word), 구(phrase), 절(clause), 단문과 복문 등의 문장(sentence), 문단(paragraph), 문서(document), 복수 문서 등의 다양한 수준에서 NLP가 수행될 수 있다(Ingersoll, Morton, & Farris, 2013/2015, 41-42쪽).

---

1) NLP의 기계학습은 크게 기호귀납적 학습(symbolic inductive learning), 비기호연결주의 방법(subsymbolic and connectionist approach), 확률적 학습(probabilistic learning)으로 나눌 수 있다. 기호귀납적 학습으로는 사례 기반 학습(instance-based learning), 결정트리(decision tree), 귀납논리(inductive logic)가, 비기호연결주의 방법으로는 신경망(neural network), 유전자알고리즘(genetic algorithm)이, 확률적 학습으로는 베이지안망 학습(Bayesian network learning), 은닉 마코프 모델, 확률문법이 있다. 이 밖에 능동학습(active learning), 부스팅(boosting), 강화학습(reinforcement learning), 건설적 귀납(constructive induction)도 활용된다(김영택 등, 2001, 51-59쪽).

이를 고려해 NLP를 분류하면, 형태소분석, 구문분석(syntax analysis, 또는 parsing), 의미분석, 담론분석으로 나눌 수 있다.[2] 형태소분석이란 문서, 문장 등 입력된 문자열을 분석하여 형태소를 기본 단위로 분류하는 것이다. 구문분석은 형태소들이 결합하여 구문이나 문장을 만드는 규칙인 통사규칙에 따라 문장 내에서 형태소의 역할이나 상호관계를 분석하는 것이다. 의미분석은 구문분석 결과를 해석하여 문장 내 단어의 의미를 보다 명확하게 구분하고 문장성분 간 의미관계를 파악하는 작업이다. 담론분석은 문맥 속에서 단어나 문장 등에 어떤 의미가 있는지 분석하는 것이다. 형태소분석, 구문분석, 의미분석, 담론분석은 각각 형식 언어학(formal linguistics)의 형태론(morphology), 구문론 또는 통사론(syntax), 의미론(semantics), 담론분석 또는 화용론(pragmatics)에 대응한다(김영택 등, 2001, 24-26쪽).

방법론적으로는 편의상 크게 알고리즘(algorithm)에 의한 통계적 접근(statistical approach)과 영역지식에 기초한 휴리스틱(heuristic)에 의존하는 규칙 기반 접근(rule-based approach), 그리고 둘을 혼합한 복합적 접근(hybrid approach)으로 나눠 볼 수 있다(김영택 등, 2001, 99-106쪽). 대부분 실용적인 서비스에서는 복합적 접근을 통해 성능을 극대화한다.

---

2) 이 같은 구분은 특히 보편연산문법(universal applicative grammar)에서 자연언어를 분석하는 형태-구문 구조 층위와 논리-문법적 표상 층위에 집중한 것으로도 볼 수 있다. 전자는 형태소분석과 구문분석에, 후자는 의미분석과 담론분석에 해당한다. 자연어처리 알고리즘이 시사하는 것처럼, 보편연산문법은 자연어를 의미소로 나눈 뒤 이를 연산 가능한 형태로 표현할 수 있다. 보편연산문법은 촘스키 이래로 다양한 자연언어를 묶는 하나의 보편언어가 존재한다고 보는 보편주의 관점에 속한다. 그러나 보편주의는 보편언어를 인도유럽어의 문법 요인으로 환원하는 로고스중심주의(logocentrism)의 한계를 갖는다. 이에 대해 데끌레는 인지연산문법(cognitive and applicative grammar)를 제안한다. 그는 앞선 두 층위 외에 인지-의미 표상 층위를 추가한다. 이는 보편언어를 문법적 수준이 아닌 인지적 수준에서 정의하려는 시도다. 의미소는 인지불변소(invariant cognitif)인 원시소(primitive sémantiques)로 대체되며, 원시소의 관계는 인지의미도식(schéma sémantico-cognif)으로 표현된다(서종석, 2011; 손호건, 2012; 윤석만, 2006).

규칙은 비록 조작적으로 정의된 범위에 한정된다고 할지라도, 최소한의 절차, 즉 소수의 단순한 'if-then' 제어 논리만으로 높은 성능을 구현할 수 있어야 가치가 있다.

(1) 자료수집

뉴스 NLP를 위해서는 우선 기사를 수집해야 한다. 이 단계에서 먼저 고려해야 할 점은 사회적 문제이다. 일반적으로 자료는 개인정보 문제와 저작권 문제가 발생한다. 뉴스의 경우 개인정보 문제보다는 저작권 문제가 해결하기 어렵다. 저작권법상의 문제[3]를 검토해야 할 뿐만 아니라 합법적이더라도 저작권자인 언론사가 뉴스 이용 자체에 반대할 수도 있다. 따라서 매우 방대한 자료를 분석하고자 할 경우, 저작권자의 동의 없이 임의로 뉴스를 기계적으로 수집하기보다는, 먼저 저작권자인 언론사나 한국언론진흥재단 같은 저작권 대행사 등과 자료 이용에 대한 계약을 체결한 후 다운로드 받는 편이 낫다. 이때 비용이 소요될 수도 있지만, 연구목적으로 사용범위를 한정하면 부담이 크지 않다.

웹사이트나 소셜미디어의 페이지나 메타데이터 등의 필요한 정보를 자동으로 수집하는 크롤러(crawler)를 사용하여 언론사 홈페이지나 뉴스 포탈에서 기사를 자동으로 수집할 수도 있다. 그러나 크롤링에는 난관이 적지 않다. 우선 해당 언론사나 포탈이 제공하는 API(Application Programming Interface)를 이용할 수도 있지만, 그런 사례는 드물다. 저작권 문제를 고려하면 원문을 제외한 메타데이터만 수집하는 것이 적절하다. 문제는 제목과 날짜 등을 제외한 대부분의 메타데이터가 없는 경우가 많다는 점이다. 일반적으로 언론사는 NewsML(News Markup Language)이라는 데이터 구조 표준을 일부 변형하여 메타데이터를 축적한다(김명기 · 최진순, 2007). 그런데 NewsML 데이터 입력을 대부분 수작업에 의존하기 때문에 비용이 많이 들어 언론사가 NewsML 데이터를

3) 국가법령정보센터 저작권법(http://www.law.go.kr/lsInfoP.do?lsiSeq=148848&efYd=20140701#0000) 참조.

입력하지 않는 경우가 많다. 따라서 부득이하게 연구용에 한해 본문을 수집하여 데이터를 추출할 경우도 적지 않다. 이때 크롤링을 하면서 한꺼번에 대규모 요청을 보내면 언론사 서비스를 방해할 위험이 있다. 일부 해외 주요 언론사는 자사 사이트에 크롤링(crawling) 방지 기술을 도입하기도 하고, 데이터 구조도 언론사마다 조금씩 차이가 나기 때문에 이에 맞게 크롤러를 운용할 수 있는 개발자를 고용해야 한다. 이 경우 추가 비용이 든다. 따라서 필요한 기사 자료가 많지 않다면, 언론사 사이트나 뉴스 포털에서 수작업으로 수집하는 편이 나을 수 있다.

기술적 문제도 있다. 우선 고(古)신문 자료처럼 아직 디지털화되지 않은 자료는 디지털화해야 한다. 디지털화됐더라도 PDF(Portable Document Format) 같은 이미지 파일 형태라면 OCR(Optical Character Reader/Recognition) 등을 이용한 텍스트 변환 작업이 필요하다. 자료의 출처나 형식이 다양하다면, 사전에 데이터 형식을 표준화해야 한다.

요컨대 기사 수집에는 적지 않은 노력과 비용이 들기 때문에, 적정한 수준에서 수집 범위를 정해야 한다. 특정 연구만을 위해서라면 수작업이나 크롤링을 통해 최소한의 기사만 수집하는 것이 낫다. 그러나 향후 다양한 연구를 위해서는 저작권자에게 구입하는 식으로 대규모로 자료를 수집하는 것이 나을 수 있다.

(2) 형태소분석

기사의 형태소분석을 위해서는 우선 기사를 글자나 단어, 어절, 문장처럼 작은 의미단위인 토큰(token)으로 나누어야 하는데 이를 토큰 분리(tokenization)라고 한다. 토큰 분리 후에는 형태소 사전과 품사 사전을 활용하여 어간추출(stemming)과 품사부착(Part Of Speech tagging, POS-tagging) 과정을 거친다.

국내에서는 흔히 21세기 세종계획에 따라 구축한 세종말뭉치(Sejong corpus)를 바탕으로 보완하여 형태소분석에 활용하는 경우가 많다. 참고로 고려대학교 민족문화연구원의 〈물결21 말뭉치〉는 〈조선일보〉,

표 1. 형태소분석사례

러시아는	러시아/NNP+는/JX	NNP : 고유명사, JX: 보조사
이제	이제/MAG	MAG : 일반부사
헌법규정에	헌법/NNG+규정/NNG+에/JKB	NNG : 일반명사, JKB: 부사격 조사
따라	따르/VV+아/EM	VV : 동사, EM: 어말어미
3개월	3/SN+개월/NNB	SN : 숫자, NNB: 의존명사
이내에	이내/NNG+에/JKB	
대통령	대통령/NNG	
선거를	선거/NNG+를/JKO	JKO : 목적격 조사
치르게	치르/VV+게/EM	
된다.	되/VV+ㄴ다/EM+./SF	SF : 마침표, 물음표, 느낌표

〈중앙일보〉, 〈동아일보〉, 〈한겨레〉 등 4개 신문사의 14년 치(2000~2013년) 기사 6억 어절에 달하는 말뭉치를 구축하고 있는데 이에 따른 형태소분석의 사례는 〈표 1〉과 같다(김일환 · 이도길 · 강범모, 2010).

(3) 구문분석, 개체명 인식

뉴스의 구문분석은 개체명 인식이나 의미분석의 성능 개선을 도와준다. 구문분석은 크게 완전 구문분석과 부분 구문분석으로 나뉜다. 일반적으로 영역에 맞는 부분 구문분석이 완전 구문분석보다 더 유용하다(강승식 · 우종우 · 윤보현 · 박상규, 2001, 29쪽). 뉴스 NLP에서 중요한 것이 개체명 인식이다. 개체명 인식에는 완전 구문분석이 필요 없다. 경우에 따라 부분 구문분석을 활용하면 된다. 예컨대 정보원의 이름, 기관, 직함을 찾기 위해, 큰따옴표가 있는 인용문에서 앞쪽의 큰따옴표 앞에 조사 '은/는/이/가'가 붙는 단어 중 명사들에서 사람 이름을 찾는 식으로 규칙을 줄 수 있다. 이때 개체명 인식을 위해 모든 문장에 대한 구문분석은 필요 없고, 인용문, 그것도 큰따옴표 앞부분에 대해서만 부분 구문분석을 수행하면 된다.

개체명 인식은 형태소분석에서 고유명사로 분류된 것을 의미에 따라 재분류하는 작업으로 볼 수 있다. 개체명은 뉴스에서 중시되는 5W1H

(who, where, when, what, why, how)와 관련성이 높다. 뉴스에서 중요한 개체명으로는 인명(person), 장소(location), 기관(organization), 직급명, 직업명, 상품명, 작품명, 학술용어, 개념어 등의 용어(terminology)가 있다. 이 밖에 시간, 통화량, 비율 등 각종 도량형의 값이 되는 수치(number)와 그 단위도 고유명사는 아니지만 중시된다. 외국어 개체명의 한국어 표기 인식은 난제이지만 꼭 필요하다.

개체명 인식에는 개체명 사전이 활용된다. 개체명 사전은 수작업으로 구축할 수도 있지만, 이미 디지털화된 표준 사전을 이식하는 것이 편하다. 전화번호부, 우편번호부, 백과사전, 인명사전, 각종 전문용어사전을 비롯해, 영어권의 경우 위키피디아(Wikipedia), 프리베이스(Freebase), 지오네임(Geoname), 위키데이터(Wikidata) 등을 활용할 수 있다.[4] 일반적으로 공신력 있는 표준 사전을 활용하는 것이 수작업보다 바람직하다. 우선 이런 사전들은 범용 분류체계를 제공하기 때문에 개발팀에 상관없이 호환이 가능하다. 개체명을 수작업으로 입력할 경우 다른 개발팀이 구축한 개체명 사전과 호환성이 없다. 또 우편번호 변경처럼 추후 분류체계가 전부 바뀔 경우에도 사전을 일괄적으로 업데이트하기 쉽다. 다만 기존 사전의 분류체계가 너무 복잡해서 서비스 목적에 맞게 축소된 범주를 활용해야 할 수도 있다. 또한 표준 사전은 1년 단위로 일괄 갱신되는 경우가 많다. 뉴스는 날짜가 부여되므로 각 연도의 개체명 사전을 해당 연도의 뉴스와 연계해 적용함으로써 개체명 인식 성능을 더욱 높일 수도 있다. 표준 사전이 없는 경우 해당 분야의 전문가와 협력하여 사전을 구축해야 한다.

한편 개체명 인식을 통해 문장을 내용에 따라 분류할 수도 있다. 예컨대 이름과 기관이 모두 있는 정보원의 인용문은 개인실명인용문, 기관만 있는 정보원의 인용문은 기관인용문, 이름, 기관 모두 없는 정보

4) 참고로 영문 문서의 의미분석을 위해서는 지식그래프(Knowledge Graph), 워드넷(WordNet), 프레임넷(Framenet), 바벨넷(BabelNet), 놈뱅크(NomBank), 시맨틱 미디어위키(Semantic MediaWiki) 등을 활용할 수 있다.

원은 익명인용문으로 분류할 수 있다. 이를 위해서는 우선 문장경계 인식 내지 문장 분리(sentence separation)가 필요하다. 문장경계 인식은 한마디로 문장이 어디서 시작되고 끝나는지를 파악하는 것으로 규칙과 기계학습 등을 활용하여 높은 성능을 기대할 수 있다(김주희 · 서정연, 2010; 박수혁, 2008).

(4) 의미분석

의미분석의 문제는 다양하다. 우선 단어와 문장 수준에서 동의, 유의, 중의, 다의, 반의, 상위개념 또는 하위개념, 관련어, 옛말 등을 파악하는 문제가 있다. 또한 동형이의어나 이형동의어의 파악, 비유나 대용어의 파악, 연상, 대체어 등의 파악, 패러프레이징 등이 문제가 될 수 있다. 예컨대 '아름답다'는 '곱다', '깨끗하다', '찬란하다', '찬연하다', '청아하다', '훌륭하다'와 같은 단어와 유의어이며, '못생기다', '추하다', '밉다'의 반대말이고, 그 옛말은 '아롬답다'가 되는데, 이러한 관계를 파악하는 것이 의미분석의 한 과제다.

뉴스 NLP에서도 의미분석 문제는 다양하게 나타날 수 있지만, 이 연구에서는 크게 대용어 해소와 의미 중의성 해결을 중심으로 살펴본다. 뉴스 NLP에서는 특히 개체명과 관련된 의미분석이 중요하고, 대용어 해소가 의미 중의성 해결보다 더 빈번한 문제이다. 뉴스에서는 구체성과 간결성을 위해서 개체명과 그에 대한 대용어가 널리 사용된다. 예컨대 '홍길동 OO그룹 회장'을 '홍 회장' 등 성과 직함 또는 '그'와 같은 인칭대명사로 표기하거나, '한국은행'을 '한은'과 같은 두문자어(acronym) 내지 약어(abbreviation)로 쓰는 경우, '전년 대비'와 같이 수치를 대신하는 경우 등이다.

한편 뉴스에는 수많은 개체명이 등장하기 때문에 분석대상이 되는 기사가 많아지면 의미 중의성 해결 문제가 나타난다. '사과'가 먹는 사과인지 사과하는 행위인지 구분하는 것이 의미 중의성 해결의 대표적 사례다. 뉴스 개체명 인식에서 '수지'가 가수인지 장소명인지 아니면 보통

명사인지, '박 대통령'이 '박근혜 대통령'인지 '박정희 대통령'인지를 구분하는 문제, 또는 '이재용'이 삼성전자 부회장인지 아나운서인지 파악하는 동명이인 문제나 '박근혜 대표'와 '박근혜 대통령'이 동일인임을 파악하는 이명동인 문제 등 인명과 관련된 의미 중의성 해결의 문제가 중요하다.

(5) 담론분석

담론분석은 고난도의 NLP에 속한다. 뉴스에서는 흔히 군집화, 중복 기사의 제거, 지면 분류, 순위화, 기사 요약(summarization), 태그 추천(tagging), 이슈 트래킹(tracking), 평판분석(opinion mining)이나 감성분석(sentimental analysis) 등이 시도되어 왔다.

한국어 NLP의 성능은 만족스럽지 않은 편이다. 군집화의 경우 관련성을 저널리즘 관행에 맞게 정의할지, 얼마나 관련 있어야 하나의 군집으로 묶을지 등의 문제가 발생한다. 지면 분류 역시 기계학습의 기본적인 성능 문제 외에도 언론사별로 지면 분류기준도 모두 다른 상황에서 통일된 분류기준을 어떻게 만들지, 한 기사가 복수의 지면으로 분류될 수 있을 때 이를 어떻게 처리할지 등의 문제가 있다.

가중치 부여나 기사 요약에서는 기사와 문장의 다양한 중요도를 지향하는 저널리즘 가치와 어떻게 연계할 것인가, 중요도를 판정하는 기간 등은 어떻게 설정할 것인가 등의 문제가 발생한다. 예컨대 속보성 기준이면 시간을, 인기도 기준이면 클릭 수를 기준으로 삼을 수 있지만, 사실성이나 다양성, 비판성 등 저널리즘 가치를 반영하여 가중치를 어떻게 부여할지는 난제이다. 태그 추천이나 이슈 트래킹은 기사 내의 중요한 단어를 빈도나 각종 토픽 모델링(topic modeling) 기법을 통해 추출하고 시계열 변화를 살펴보는 방식이다. 문제는 개념어 사전이 구축되어 있지 않은 경우, 기계적으로 추천한 태그가 너무 일반적인 의미를 담고 있는 경우가 많아서, 수작업으로 다시 선별하고 태깅하는 경우가 적지 않다는 점이다.

평판분석과 감성분석의 경우 외적 타당도가 낮다. 무엇보다 뉴스는 형식적 중립성을 지향하여 찬반 양쪽의 의견을 다 담고 있기 때문에 의견이 중립으로 판정되는 경우가 많다. 게다가 의견은 단순히 찬반으로 양분되지 않는다. 즉 찬성도 그 이유가 다 다르며 찬성 간에도 내용이 양립할 수 없는 경우도 적지 않다. 또한 같은 문장도 맥락에 따라 의견이 다르게 해석되는 등 어려움이 있다.

이러한 난점에도 불구하고 뉴스의 담론분석은 언론학이나 저널리즘 측면에서 가장 널리 관심을 받았던 분야이다. 따라서 일단은 조작적 정의를 통해 기계적으로 가능한 범위 내에서 담론분석을 시작하고, 이를 단계적으로 확장하여 수작업을 줄여 가는 과정이 필요하다.

구체적으로는 먼저 중복기사를 제거하거나 유사한 내용 또는 관련된 내용을 담은 기사와 문장을 묶어서 검토에 드는 수고를 덜 수 있다. 또 기사의 영역을 나타내는 지면을 자동으로 다중분류하고 기사의 주제를 자동으로 부착하여 영역별로 주제를 다양하게 검토할 수 있다. 문장을 분할하는 등 개체명 인식 결과를 활용해 다중분류하면 필요한 문장만 빠르게 검토할 수 있다. 특히 인명, 기관명, 직함에 대한 개체명 인식과 이에 따른 인용문 추출이 가능하므로 정보원 분석과 인용문 분석을 대규모로 반복 가능한(replicable) 방식으로 수행할 수 있다. 문장 단위로 주제를 추출할 문장을 주제별로 손쉽게 검토할 수도 있다. 무엇보다 의미연결망분석 등과 결합되어 개체명과 주제, 문장, 기사, 매체에 따라 순위를 매겨, 중요한 정보원과 기관, 주제의 시계열적인 변화나 정보원 간, 기관 간, 매체 간의 관계 및 그 강도를 전수에 대해 정확히 파악할 수도 있다. 더 나아가 어떤 정보원이나 기관이 어떤 주제를 어떻게 다루었으며, 어떤 주제가 어떤 하위주제 또는 상위주제와 함께 다루어졌는지도 쉽게 파악할 수 있다(박대민, 2013, 2014a, 2015; 김선호·박대민, 2015).

3. 단어 중심 뉴스 자연어처리의 한계와 개체명 및 문장 중심 분석의 필요성

최근 언론학에서 컴퓨터 이용 내용분석에 대한 관심이 확대되면서 뉴스 분석에서 NLP를 활용하는 사례가 많아지고 있다. 특히 의미연결망 분석을 위한 전처리 작업을 위해 NLP가 활용되는 경우가 늘고 있다. 기존에는 영미권 기사나 국내 신문의 영문판 뉴스를 대상으로 CatPac, WordLink, LIWC(Linguistic Inquiry and Word Count) 등 영어 NLP 프로그램을 활용한 연구가 진행됐다(백영민 · 최문호 · 장지연, 2014; 장하용, 1995, 2000; 최수진, 2014). 이후 LIWC를 국문에 맞게 변형한 K-LIWC가 개발되어 활용되기도 했다(이창환 · 심정미 · 윤애선, 2005; 박종민 · 이창환, 2011). 이 밖에 강남준 등은 한글 형태소분석기인 KLT(Korean Language Technology)[5]를 사용하여 신문 표절을 판정했다(강남준 · 이종영 · 오지연, 2008). 표절이나 저자 판별, 문체 등을 식별하기 위해 기사의 유사도 등의 NLP를 시도한 연구도 등장했다(강남준 등, 2008; 강남준 · 이종영 · 최운호, 2010; 윤상길 · 김영희 · 최운호, 2011).

이 가운데 국내 뉴스 NLP에서 널리 쓰이는 프로그램은 영어 텍스트 분석 프로그램인 FullText를 변형한 KrKwic이다(박한우 · Leydesdorff, 2004)[6]. KrKwic는 의미연결망분석을 위해 텍스트에서 단어를 추출하고 빈도를 간결하게 제시해 준다. 또한 공기(co-occurrence) 여부에 따른 단어의 1원 행렬(1 mode matrix)을 제공하여 Ucinet, Pajek 등 연결망분석 프로그램과 연동이 쉽다. KrKwic는 비교적 짧은 길이의 텍스트를 분석하는 데 적합하므로 적은 양의 기사를 분석하기 위해 주로 활용된다(김만재 · 전방욱, 2012; 박한우 · 남인용, 2007; 이완수 · 최명일, 2014; 최윤정 · 권상희, 2014). 문장 단위의 공기 빈도를 제공하는 등 KrKwic를 개선하고 온라인으로 구현한 TextoM[7]도 나와 있다.

---

5) http://nlp.kookmin.ac.kr/HAM/kor/index.html

6) http://www.leydesdorff.net/krkwic/

그러나 KrKwic을 비롯한 기존 프로그램에는 몇 가지 한계가 있다. 우선 이들에는 형태소분석 기능이 없거나, 있더라도 일반적으로 형태소 분석과 품사 부착, 조사나 부사 및 의존명사 등 불용어 제거 등만 자동으로 수행하고, 개체명 인식이나 의미분석 기능이 없거나 성능이 낮다. 이 경우 컴퓨터가 기사에 개체명이 실제로 있는데 없는 것으로 판별하는 2종 오류를 범하거나(낮은 재현율), 개체명이 아닌데 개체명으로 판별하는 1종 오류를 범할 수 있다(낮은 정확도). 특히 정확도는 수작업으로 보정할 수 있지만, 재현율은 보정이 어렵다. 개체명 인식의 재현율이 낮거나 의미분석의 성능이 떨어질 경우, 기사에서 실제로는 더 많이 나타났음에도 불구하고 빈도를 과소평가하거나, 아예 특정 개체명이 결측되어 의미연결망분석의 각종 중앙성 계산 등에서 빠지는 등의 오류가 나타날 수 있다. 즉, 개체명의 가중치를 과소평가할 가능성이 높다. 요컨대 NLP 프로그램에서 개체명 인식이나 의미분석 등의 기능이 있는지, 있다면 어느 정도의 성능을 구현하는지, 기능이 없거나 성능이 낮다면 그러한 문제를 어떻게 해석에 반영해야 할지 고려해야 한다.

예컨대 〈그림 1〉은 〈한국일보〉, 〈서울신문〉, 〈동아일보〉, 〈문화일보〉, 〈한겨레〉, 〈국민일보〉, 〈세계일보〉, 〈경향신문〉에서 2004년 7월 1일부터 2005년 6월 30일까지 나온 기사 중 '청년', '대학생', '청소년', '청춘'이라는 단어가 포함된 기사 797개의 인용문 872개를 분석해 단어 클라우드를 그린 것이다. 왼쪽은 인용문에서 3개씩 수작업으로 추출한 개념어를 형태소분석한 채로, 오른쪽은 개념어를 개체명 상태로 그대로 둔 채로 단어를 등장 빈도에 따라 가중치를 부여해 시각화했다. 형태소 클라우드에서 주요어는 '스포츠'(138), '청년'(123), '실업'(103) 순이었으며, 개념 클라우드에서 주요어는 '청년실업'(49회), '신용불량자'(18회), '일자리창출'(8) 순이었음을 알 수 있다. 주요어들의 순위도 달랐지만 무엇보다 오른쪽 그림에서 가장 중시된 '청년실업'이라는 단어는

7) http://www.textom.co.kr/

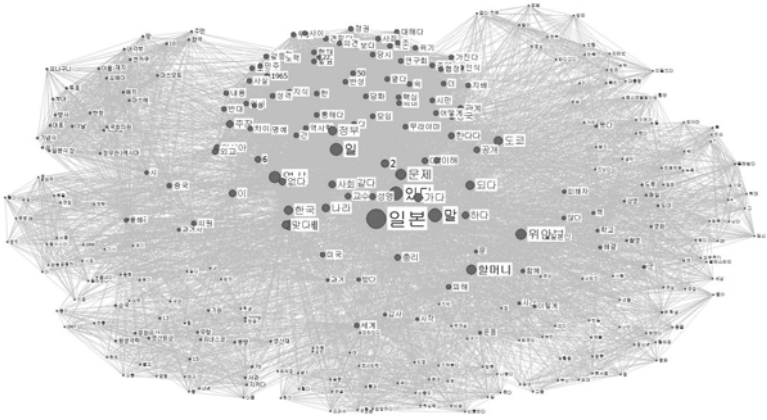
그림 1. 2005년 청년담론의 형태소 단위 단어클라우드와
개념 단위 단어클라우드

형태소 단위의 단어클라우드　　개념 단위의 단어클라우드

왼쪽 그림에서는 아예 나타나지 않았다. 이는 '청년실업'이란 단어가 형태소분석 결과 '청년'과 '실업'으로 모두 쪼개지기 때문이다.

게다가 단어 단위의 분석은 기사 단위의 분석보다도 분석을 더 어렵게 만드는 경향이 있다. 예컨대 〈그림 2〉의 상단 그림은 2015년 1월 1일부터 2015년 6월 30일까지 '일본 망언'으로 검색된 〈한겨레〉와 〈동아일보〉의 기사 14개를 단어를 결점으로 해서 기사 공동출현 기준으로 의미연결망으로 그린 것이다. 하단 그림은 같은 기간, 같은 검색어로 〈경향신문〉, 〈국민일보〉, 〈동아일보〉, 〈문화일보〉, 〈서울신문〉, 〈세계일보〉, 〈한겨레〉, 〈한국일보〉 등 8개 매체 48개 기사의 정보원 69명을 개체명, 즉 '인명+기관명+직함'을 결점으로 하여 기사 공동출현 의미연결망을 도출한 것이다. 직관적으로 보아도 상단 그림의 단어연결망은 적은 기사를 시각화했음에도 불구하고 너무 많은 결점과 연결이 출현해서 무엇과 무엇이 유의미하게 연결되는지를 명쾌하게 파악하기 어렵다. 하나의 기사에도 수백 개의 단어가 있는 데다가, 기사 공동출현 기준으로 연결을 정의하면 이 단어가 모두 연결된 완전연결망으로 구성되기 때문이다. 설사 결점을 단어가 아닌 명사로 한정한다고 해도 마찬가지

그림 2. 2015년 상반기 일본 망언 담론의 단어연결망과 정보원연결망



단어연결망

정보원연결망

이다. 반면 정보원연결망은 더 많은 기사를 더 간결하게 표현하며, 어떤 정보원이 함께 기사에 출현했는지도 보다 쉽고 명확하게 파악할 수 있다. NLP와 의미연결망분석과 같은 컴퓨터 보조 분석이 대규모 데이터를 다양하게 축약하여 간결하게 처리하는 데 유용성이 있다는 점을 감안할 때 단어 수준의 분석은 이러한 장점을 살리지 못하는 셈이다.

이는 근본적으로 텍스트분석에서의 분석수준(level of analysis) 문제, 즉 연구목적에 따라 형태소, 단어, 품사, 개체명, 문장, 기사, 매체의 기사 전체 등 어떤 분석수준을 결정하는 문제와 연관된다. 기존의 뉴스 NLP 활용 연구는 형태소분석을 주로 활용하여 단어, 또는 명사를 분석단위로 삼았다. 그리고 이를 제목이나 기사에서 추출했다. 물론 기존의 내용분석 연구에서도 이러한 표면적 또는 언어적 측면을 중심으로 한 분석이 없지는 않다. 우선 흔히 기사의 양(몇 번이나 보도됐나), 단어의 등장횟수 등 단순한 빈도가 분석된다. 제목의 비문 요소, 특수어 또는 인용부호 사용 여부와 효과, 제목 간의 일치도, 술어의 특성 등을 따져본 연구도 있다(김관규, 2013; 김춘식·이강형, 2008; 김형·양혜승, 2013). 개체명인 정보원이나 보도대상, 정보원이 소속된 기관과 직함, 그리고 드물지만 장소 측면에서 기사 내용을 분석한 경우도 있다(이완수, 2006; 이준웅 등, 2007; 임영호, 2012; 임영호·이현주, 2001). 그러나 언론학에서 수행된 전통적인 내용분석이나 담론분석에서는 대부분 형태소나 구문적 특성으로 환원되기 어려운 거시 수준의 의미를 파악하는 데 더 많은 공을 들이고 있다. 구체적으로는 기사의 내용이나 주제, 기사의 형식(역 피라미드 형식, 내러티브 형식 등), 기사 유형(스트레이트, 피처 등), 기사 생산주체(언론인, 비언론인), 대상독자, 태도, 틀 짓기, 의제설정, 뉴스가치(사실성, 정파성, 선정성, 공정성, 중립성, 다양성, 심층성 등), 논증 품질, 논조 변화, 관점, 이념, 이데올로기, 권력관계, 집합기억 등 고차원적인 의미를 파악하는 데 초점을 두고 있다(김남일·백선기, 2008; 김동윤·김성해·유용민, 2013; 김성해, 2006; 김세은, 2009; 김춘식, 2002; 남재일, 2010; 방은주·김성태, 2012; 설진아, 2013; 손승

혜 · 황하성 · 장윤재, 2011; 이완수 · 박재영 · 노성종 · 이수미 · 강충구, 2009; 이정교 · 서영남 · 최수진, 2009; 임봉수 · 이완수 · 이민규, 2014; 최민재 · 김위근, 2006; 최원석 · 반현, 2006; 홍원식 · 김은정, 2013).

문제는 단어들만 보면 담론의 맥락은 물론, 간단한 의미조차 즉시 파악할 수 없는 경우가 많다는 것이다. 실제로는 단어의 의미를 파악하기 위해 단어를 선별적으로 이리저리 연결해야 하는데, 이때 해석자가 이미 가진 배경지식에 의존하거나, 배경지식이 없다면 결국 원문기사를 읽어야 한다. 예컨대 2004년 8월 3일 〈조선일보〉의 '기술혁신형 중소기업 육성'이라는 기사에서 가장 많이 등장한 단어는 '기업', '지역', '이노', '육성', '비즈', '부산', '중기청', '울산', '600'인데 이것만으로는 기사가 무슨 내용을 담고 있는지 알기 어렵다. 반면 인용문과 같은 문장으로 추출하면 다음과 같이 내용을 직관적으로 알 수 있다(박한우 · Leydesdorff, 2004).

> 부산 · 울산지방중소기업청은 2일 "최근 중국경제 부상 이후 중소기업이 원자재난과 인력난, 사회적 인식저하 등으로 침체위기를 겪고 있는 가운데 이를 극복하고 지역경제에 활력을 불어넣기 위해 부산 · 울산지역의 이노비즈 기업 600여 곳을 발굴해 집중 육성할 것"이라고 밝혔다.

사실 텍스트학에서 지적하는 것처럼 내용적인 측면에서 텍스트의 최소 의미단위는 문장이다(Vater, 2001/2006, 98쪽). 뉴스도 사실과 의견을 담은 명제인 문장들로 구성된다. 따라서 뉴스를 해석할 때 분석단위를 단어가 아니라 문장으로 삼는 것이 더 낫다. 즉, 컴퓨터 이용 내용분석을 위해, NLP를 통해 제시된 분석대상 자료는 기사의 내용을 빠짐없이 체계적으로 축약하면서도 개체명 단위나 문장 단위로 제시될 필요가 있다.

이 연구에서 소개하는 〈뉴스소스 베타〉 외에도 개체명과 연계된 문장 중심으로 뉴스를 NLP하여 서비스하는 모델은 최근 들어 해외에서도

그림 3. 뉴스 익스플로어

베타 서비스 수준에서 시도되고 있다. 비정형 데이터인 기사 텍스트에서 기본적인 메타데이터는 물론, 개체명에 해당하는 정보를 수작업으로, 또는 기계적으로 추출하여 정형데이터 형태로 관리하여 다양한 콘텐츠를 재생산하고자 하는 스트럭처 저널리즘(structured journalism)이 대표적이다. 예컨대 인공지능 플랫폼으로 유명한 IBM 왓슨은 뉴스에서 인명(person), 회사명(company), 기관명(organiztion), 장소(location), 주제(topics), 시간(time) 등을 추출하여 이를 의미연결망과 지도, 단어 클라우드, 타임라인 등으로 시각화해 보여 주는 '뉴스 익스플로어'(news explorer)[8]를 2015년 7월 공개했다. BBC도 2013년 뉴스 스토리라인 온톨로지(news storyline ontology) 모형을 제안한 바 있다. 뉴스 스토리라인 온톨로지란 뉴스를 구성하는 단어들과 이들 간의 관계를 분석하여 데이터 모델링하여 구축한 일종의 사전으로 이를 활용해 '스토리라인'을 만들 수 있다. BBC의 뉴스 스토리라인 온톨로지는 실용화되지는 않았다. 그러나 BBC는 2015년 7월 '스트럭처 저널리즘 선언'(a manifesto for structured journalism)을 발표하는 등 보다 나은 스트럭처 저널리즘 방법

8) http://news-explorer.mybluemix.net/

론을 꾸준히 모색하고 있는 것으로 보인다(오세욱 · 김선호 · 박대민, 2014).

4. 뉴스 자연어처리의 사례: 〈뉴스소스 베타〉를 중심으로

1) 〈뉴스소스 베타〉의 개요

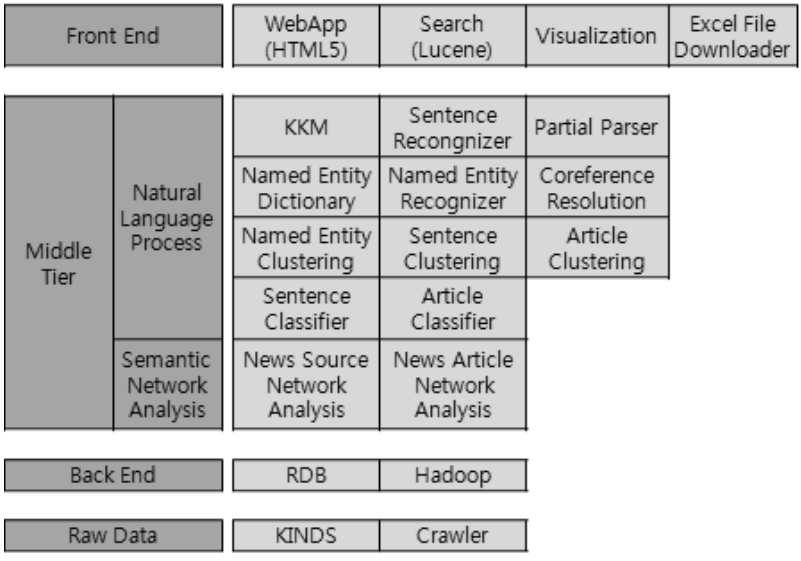
〈뉴스소스 베타〉는 뉴스 정보원 연결망분석(news source network analysis) 프로그램이다. 한국언론진흥재단 뉴스 아카이브인 카인즈[9]에 저장된 반정형 자료(semi-structured data) 형태의 뉴스 기사를 NLP하여 정형자료(structured data)로 변환해 공동 인용 정보원의 연결망을 분석한다. 2012년 개발된 프로토타입(prototype)을 바탕으로 한다.[10]

〈그림 4〉는 〈뉴스소스 베타〉의 개략적인 시스템 구성을 보여 준다. 간략히 설명하면 프론트 엔드(front end)는 이용자에게 보이는 서비스 측면을 말하는 것으로 PC와 모바일 등 기기에 따라 화면이 최적화되는 웹앱, 공개소프트웨어인 루씬(Lucene)을 기초로 만든 검색엔진, 연결망 시각화, 엑셀 파일 다운로드 등으로 구성된다. 미들티어는 NLP와 의미연결망분석 부분으로, 형태소분석기인 서울대학교 IDS(Intelligent Data System) 연구실에서 개발한 오픈소스인 꼬꼬마 형태소분석기(KKM), 문장경계 인식(sentence recognizer), 부분 구문분석기(partial parser), 이름, 소속, 직함에 관한 개체명 사전(named entity dictionary), 개체명 인식기(named entity recognizer), 대용어 해소 도구, 개체명, 문장, 기사의

9) www.kinds.or.kr

10) 프로토타입은 사전 없이 규칙 기반 부분 구문분석만으로 인명+기관명+직함, 인용문 등을 추출해 매칭하고 연결정도중앙성을 구해 시각화하는 기능을 포함한다. 이를 활용해 기사 205,391건의 정보원 33,611명을 분석하여 뉴스 정보원 연결망의 분포가 두터운 꼬리 형태의 멱함수임을 밝힌 바 있다(박대민, 2014b).

그림 4. 〈뉴스소스 베타〉의 시스템 구성



군집화 도구, 인용문, 수치문 등 문장 분류기(sentence classifier), 지면 분류기(article classifier), 뉴스 정보원 연결망분석(news source network analysis) 도구와 뉴스 기사 연결망분석(news article network analysis) 도구로 이루어진다.

백 엔드(Back End)는 자료를 저장하고 처리, 관리하는 데이터베이스(Data Base, DB) 부분이다. 리눅스 센토스(Linux Centos) 운영체제 위에서 구동하며, 받아 온 기사 자료를 저장하는 관계형 데이터베이스(relational Database)와 NLP 및 연결망분석을 수행하여 저장하는 분산형 데이터베이스인 하둡으로 구성된다. 이 밖에 자료를 수집해 놓은 뉴스 아카이브인 카인즈와 카인즈에서 기사를 받아 오는 크롤러로 구성된다. 개발 언어는 백 엔드와 미들 티어(Middle Tier) 구현에는 자바(JAVA)를, 프론트 엔드 구현에는 HTML5를 사용했다.

서비스 측면에서 〈뉴스소스 베타〉는 크게 일반인 버전(오늘의 뉴스, 검색)[11]과 전문가 버전(뉴스 정보원 연결망 시각화, NLP 데이터 다운로드)

12)으로 구성된다. 다운로드 파일은 엑셀 형태로 제공되며 내용을 담은 파일(reference file)과 가중치를 담은 파일(degree file)로 나눠진다. 다운로드 파일은 매체명, 날짜, 기사 제목, 기사 본문, 정보원 이름, 인용문, 검색어, 검색 기간, 검색 매체, 정보원 가중치 등을 포함한다.

2) 자료수집, 형태소분석

〈뉴스소스 베타〉에서는 1990년 1월 1일부터 2013년 12월 30일까지 약 3천만 건의 뉴스 데이터를 처리하여 베타서비스를 재단 홈페이지에 제공하고 운영하는 조건으로 한국언론진흥재단과 양해각서를 체결하여 데이터를 확보했다. 상술하면, 카인즈에 축적된 1990년 1월 1일부터 2013년 9월 23일까지의 기사는 재단을 직접 방문해 일괄 다운로드했다. 이후 1일 단위로 추가로 쌓이는 기사는 정기적으로 ftp를 통해 크롤러로 수집했다. 즉, 재단 카인즈 DB 서버와 〈뉴스소스 베타〉의 전처리 서버에 크롤러를 탑재하고, 이를 통신량이 적은 새벽 2시에 활성화하여, 당일 데이터와 혹시 있을 2주 이내의 미수집 데이터를 수집했다. 크롤러는 2014년 2월 14일까지 작동됐다. 〈뉴스소스 베타〉의 크롤러 구성도는 〈그림 5〉와 같다(차세대융합과학기술원, 2013).

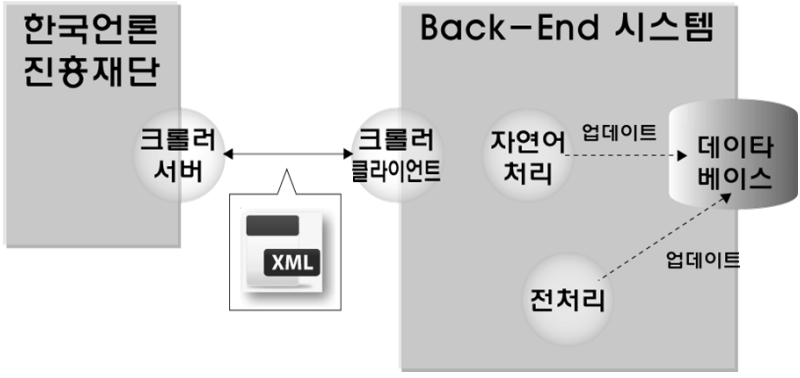
파일은 대부분 NewsML 형태로 txt 파일이나 jpg 파일 등도 일부 있다. 수집된 데이터의 항목은 원문기사 ID, 기사 제목, 언론사, 발행일, 지면 분류이다. 참고로 분석 결과 추가되는 데이터 항목은 문장 ID, 문장 유형, 문장 내용, 정보원 ID, 정보원 이름, 정보원 소속, 정보원 직함, 정보원 유형 등이다. 형태소분석에 사용된 꼬꼬마 형태소분석기는 비상업적 사용이 가능한 부분적인 공개소프트웨어로 뉴스의 완전 형태소분석에 대해 81%의 정확도를 보여 주는 무난한 성능을 구현한다(이동주 · 연종흠 · 황인범 · 이상구, 2010).

---

11) http://147.47.125.161/NSNA/

12) http://147.47.123.2/expert/

그림 5. 〈뉴스소스 베타〉의 크롤러 구성도



3) 구문분석, 개체명 인식, 분류

(1) 문장경계 인식 및 문장 대분류

〈뉴스소스 베타〉는 최소한의 부분 구문분석만 실시했으며 주로 표면적 자질과 규칙을 활용했다. 우선 구문분석을 위해 문장경계 인식과 간단한 문장 분류를 수행했다. 문장경계 인식 방법을 소개하면 우선 문장 전후에 찍힌 구두점을 기준으로 한 문장을 식별한다. 물론 인용문 안에 구두점이 하나 이상인 경우가 있거나 카인즈 데이터 특성상 기사가 광고나 관련 기사 제목 등 문장이 아닌 구문이 있을 수 있다. 따라서 약간의 추가적인 구문분석을 함께 수행했다.

다소 논란은 있지만 많은 연구는 한국 언론이 대체로 가치규범적으로나 취재관행 측면에서나 사실성 중시, 전문직주의, 역피라미드 기사 양식 등 다양한 형태의 객관주의 저널리즘을 지향한다고 지적한다(김경모·신의경, 2013; 김수영·박승관, 2010; 남재일, 2008; 박재영·이완수, 2008; 이준웅, 2010; 정동우·황용석, 2012). 객관주의 저널리즘 관행에서는 인용과 수치를 사실성을 나타내는 전형적인 방법으로 활용한다(Van Dijk, 1988). 이에 기초하여 〈뉴스소스 베타〉는 문장을 크게 인용문, 수치문, 기타문으로 나눴다. 인용문은 직접인용문만 인용문으로

간주했다. 이는 보다 강한 사실성 효과를 위해 직접인용을 간접인용보다 선호하는 저널리즘 관행을 반영한 것이다. 직접인용문은 분할된 문장 중 큰따옴표가 포함된 문장으로 찾을 수 있다. 수치문은 숫자가 포함된 문장으로 제한했는데, 이는 기사에서 정보가 되는 수치는 한글이 아닌 숫자로 표기하는 관행을 반영한 것이다. 기타문은 인용문과 수치문이 아닌 모든 문장으로 한다.

뉴스는 다른 텍스트에 비해 문법적 오류가 적다. 덕분에 문장경계 인식과 문장 분류에서 공백이나 구두점, 큰따옴표, 수치 등 표면적 자질과 최소한의 구문적 지식만을 이용하는 부분 구문분석만으로도 100%에 가까운 재현율과 정확도가 나온다.[13]

(2) 개체명 인식

우선 개체명 인식을 위해 개체명 사전을 정보원, 즉 인명, 기관, 직함에 대해서 구축했다. 구체적으로는 2012년 한국전화번호부, 2012년 한경기업총람, 위키피디아에서 인명, 기관명, 직함을 추출했다. 최종적으로 개체명은 한국인명 116,688명, 외국인명 7,051명, 기관명 145,748개, 직함명 959개 등 총 270,446개의 개체명 사전이 구축됐다. 충분한 수는 아니지만, 단기간에 저비용으로 일정 수준의 사전을 구축했다는 점에서 의의가 있다.

기관명은 표준산업분류상의 업종분류코드를 활용하여 다중분류했다. 즉 다섯 자리 업종분류코드를 출입처 관행을 반영하여 편의상 세 자리 분류코드(1의 자리 대분류, 10의 자리 중분류, 100의 자리 소분류, 미분류 코드 999)로 재분류했다. 예컨대 중앙부처는 231로 분류되는데, 여기서 1은 대분류인 정치를, 10의 자리의 3은 중분류인 행정부를, 100의 자리인 2는 가장 구체적인 소분류를 뜻한다.

13) 다만 카인즈 아카이브에는 링크 기사 제목이나 광고 문구처럼 기사가 아닌 부분이 기사에 포함된 경우, 복수의 기사가 하나의 기사로 저장되는 경우가 있는데 이 경우 NLP가 적절히 수행되지 않을 수 있다.

〈뉴스소스 베타〉에서는 정치, 경제, 사회, 문화, 국제 등 5개의 대분류와, 법원, 농림수산, 학계 등 47개 항목의 중소분류를 활용했다. 중소분류는 실제 1개 유수 신문사의 출입처 관행을 참조하여 정치, 경제, 사회는 소분류 위주로 자세하게, 문화, 국제는 중분류 중심으로 느슨하게 분류했다. 이렇게 분류코드를 세분화한 것은 검색 순위에서 규칙 기반 가중치를 부여하기 위해서이다.

개체명 인식은 인용된 정보원을 찾는 데 활용했다. 즉, 인용문에 한정하여 인용문에서 인명, 기관, 직함을 찾는다. 이는 원칙적으로 다음과 같은 간단한 부분 구문분석 절차를 통해 수행한다.

① 인용문의 앞 큰따옴표를 기준으로 인용문 내에 주격조사(은/는/이/가)가 있는지 식별한다.
② 주격조사 앞에 5개 어절 이내에 개체명 사전의 개체명과 일치하는 인명, 기관, 직함이 있는지 식별한다.
③ 개체명과 일치할 경우 형태소분석을 통해 명사에 해당하면 인식된 개체명으로 간주한다.

정보원은 크게 셋으로 나눴다. 첫째, 인명이 있는 경우 개인실명정보원이다. 일반적으로 기사에서 개인실명정보원은 인명, 기관, 직함이 함께 나온다. 둘째, 집단정보원은 인명 없이 기관이 나온 정보원이다. 셋째, 인명, 기관, 직함이 모두 없는 정보원이다. 사실 익명의 경우 '관계자는'과 같이 익명 표지가 나오는 경우가 대부분이지만, 〈뉴스소스 베타〉에서는 이러한 익명 표지를 활용하지 않고 개체명 인식이 안 된 경우를 익명으로 간주했다.

개체명 인식의 성능은 사전에 크게 영향을 받는다. 인명, 기관, 직함 각각의 개체명 인식의 오류에 따라 정보원의 분류도 잘못될 가능성은 있다. 정보원 분류가 잘못될 가능성을 확인하기 위해서는 각 개체명 인식의 성능을 따져봐야 한다. 개체명 인식의 성능은 〈표 2〉에 기술한 것처럼 정확도의 경우 인명, 기관, 직함이 각각 90.3%, 92.7%, 97.8%,

표 2. 〈뉴스소스〉 베타의 개체명 인식 성능

	인명	기관	직함	정보원 분류	분류상 오류 가능성
재현율	87.4%	50.9%	75.0%		
정확도	90.3%	92.7%	97.8%		
개체명 인식 결과	O*	O	O	개인실명	인명의 재현율과 정확도가 높아 개인실명정보원으로 분류한 정보원이 잘못 분류될 가능성은 낮음
	O	X*	O	개인실명 (대용어 해소)	
	O	O	X	개인실명	
	O	X	X	개인실명	
	X	O	O	집단	기관의 재현율이 낮아 집단정보원이 인식되지 않을 수는 있지만, 일단 집단정보원으로 인식된 정보원은 제대로 분류됐을 가능성이 높음
	X	O	X	집단	
	X	X	O	익명	기관의 재현율이 낮아 집단정보원일 수 있지만 개인실명정보원일 가능성은 낮으며, 집단정보원이 아니라면 인용문의 인식의 정확도와 재현율이 높기 때문에 개체명 인식이 안 된 경우 익명정보원으로 제대로 분류됐을 가능성이 큼
	X	X	X	익명	

* O는 측정 값 있음, X는 측정 값 없음.

재현율은 87.4%, 50.9%, 75%가 나왔다.

인명의 재현율과 정확도는 상용 프로그램과 비교해도 높은 편이다. 예컨대 NLP 업체 솔트룩스가 2016년 구축한 한국언론진흥재단 빅카인즈의 경우 인명 인식의 재현율은 79.9%, 정확도는 85.1%이다.[14] 일반적으로 개체명 인식에서는 재현율이 높으면 정확도가 떨어지고, 정확도가 높으면 재현율이 낮아지는 트레이드오프(trade-off)가 있다. 정확도가 높다는 것은 인명, 기관, 직함으로 인식된 개체명은 실제로 인명, 기관, 직함일 가능성이 높다는 것을 의미한다. 재현율의 경우 인명은

14) 참고로 빅카인즈의 기관명 인식의 재현율은 80.4% 정확도는 85.8%이며, 직함의 경우 재현율 73.5%, 정확도는 84.0%로 인명 및 직함의 개체명 인식 성능은 〈뉴스소스 베타〉가 약간 더 뛰어나다.

높지만 기관이 현저히 낮은데, 이는 기사 내에서 인명은 잘 찾아내지만, 기관은 찾지 못하는 경우가 많으며, 직함도 찾지 못하는 경우가 꽤 있다는 것을 의미한다. 기관명 인식 성능이 비교적 낮은 이유는 다양하다. 우선 기관명은 보통명사의 다양한 조합으로 이뤄진 경우가 많기 때문에 개체명 인식이 오히려 어렵다. 또 기관명은 약자표기가 많아서 성능 개선을 위해서는 기관명에 대한 대용어 해결이 필요한데 〈뉴스소스 베타〉는 인명에 대한 대용어 해소 기능밖에 없다. 또 기관명은 인명에 비해 매년 자주 바뀌는데 〈뉴스소스 베타〉는 2012년에 현존하는 기관이 수록된 사전을 중심으로 활용했다. 외국어 한글표기 기관명, 인터넷 카페나 소규모 시민단체, 사적 모임 등은 사전에 누락된 경우가 많다. 이를 개선하기 위해서는 공기어분석 등을 통한 기관명 사전 자동 구축 기능 추가, 기관에 대한 대용어 해소 기능 추가, 기관명을 포함한 외국어 개체명의 한글표기 사전 구축, 연도별 기관명 사전 활용, 기타 다양한 사전 추가 등이 필요하다.

개인실명정보원을 중시하는 객관주의 저널리즘 관행과 이를 반영한 〈뉴스소스 베타〉의 기획 취지에 비추어 보면, 이러한 개체명 인식 성능의 한계는 관리가 가능한 수준이다. 즉 〈뉴스소스 베타〉는 우선 개인실명정보원은 제대로 분류한다. 또 집단정보원의 경우, 결측된 집단정보원이 있기는 하지만, 일단 인식된 정보원은 개인실명정보원이나 익명정보원이 아닌 집단정보원이 맞을 가능성이 크다. 재현율이 높기 때문에 인명이 있다면 인식됐을 것이기 때문이다. 또 익명정보원의 경우, 집단정보원일 가능성이 있기는 하지만 최소한 개인실명정보원은 아니다.

인명, 기관, 직함 모두가 재현되고 정확한가를 따진다면, 재현율은 33.4%, 정확도는 81.9%로 재현율이 너무 낮아서 문제인 것처럼 보인다. 하지만 실제로는 인명, 기관, 직함이 모두 식별되지 않을 확률은 15.5%, 식별된 개체명이 모두 맞지 않을 확률은 15.6%로, 바꿔 말하면 개체명 세 종류 중 어느 하나라도 식별될 확률은 84.5%이고 이중 하나라도 실제로 제대로 인식됐을 확률은 85.4%이다.

요컨대 개인실명정보원과 그들의 인용문을 중심으로 분석할 경우, 〈뉴스소스 베타〉는 활용에 큰 문제가 없다는 점을 뜻한다. 게다가 인용문의 재현율이나 정확도는 100%에 가깝고, 인용문 인식 후에 개체명 인식을 수행하므로, 인명, 기관, 직함 모두가 제대로 인식되지 않더라도 인용문은 추출할 수 있고, 따라서 개체명은 수작업으로 수정 입력도 가능하다. 만일 기관정보원 및 익명정보원과 그들의 인용문을 분석하고자 한다면, 인명이 식별되지 않은 정보원을 추려서 그들의 인용문을 중심으로 검토하면 된다. 기관정보원과 익명정보원을 구별해서 분석하고자 한다면, 인용문에 근거해 기관명이 식별되지 않은 익명정보원 위주로 정제작업을 수행한 뒤 분석하면 된다. 기관명 인식은 재현율은 떨어지지만 정확도는 높기 때문에, 인명이 식별되지 않고 기관명이 식별됐다면 일단 해당 정보원은 실제로 그 기관에 속한 집단정보원일 가능성이 높다.

(3) 문장 분류 심화 및 지면 분류

인용문은 개체명 인식 결과를 반영하여 규칙기반방식으로 개인실명정보원의 인용문은 개인실명인용문으로, 집단정보원의 인용문은 집단인용문으로, 익명정보원의 인용문은 익명인용문으로 세분화했다. 수치문이나 기타문의 다중분류는 개발 범위에 포함하지 않았지만, 추후 개체명 인식기능과 부분 구문분석을 추가하면 더 세분화할 수 있다.

카인즈의 기사는 2015년 4월 8일 현재 전체 32,069,983건의 기사 중 6,767,344건의 기사가 15개의 지면[15]으로 수작업으로 지면 분류되어 있다. 〈뉴스소스 베타〉에서는 이를 일단 기관의 대분류와 마찬가지로 5개 지면 분류[16]로 나누고, 학습집합을 생성한 뒤 단순 베이지안 분류

15) 종합, 정보통신·과학, 사회, 매체, 경제, 오피니언·인물, 지역, 스포츠, 특집, 문화, 국제·외신, 생활·여성, 북한, 방송·연예, 정치·해설 등.

16) 정치(북한, 정치/해설), 경제(정보통신/과학, 경제), 종합/사회(지역, 사회, 생활, 여성, 종합, 특집, 인물/오피니언), 문화(방송연예, 스포츠, 매체, 문화), 국

기(Naive Bayesian classifier)를 이용한 기계학습을 통해 미분류된 기사를 분류했다. 분류기 성능은 다양한 테스트 집합에 대해 평균 75%의 정확도로 나타났다.

추후에 지면 분류기의 성능을 더 높이기 위해서는 간단한 규칙 기반 방식을 추가하는 것도 고려할 만하다. 즉 출입처 관행을 반영하여, 우선 기사 내에 인식된 기관의 대분류를 기준으로, 전부 일치할 경우 해당 분류로 기사의 지면을 분류한다. 전부 일치하지는 않을 경우, 기사 내 등장한 각 분류별 기관 수를 전체 기관 수로 나눈 가중치로 기사의 지면을 복수로 지정한다. 복수 분류를 허용하면, 추후 서비스 필요에 따라 중복 노출시키거나, 복수로 지정된 기사를 종합으로 분류할 수 있다.

4) 의미분석

(1) 대용어 해소

〈뉴스소스 베타〉에서 대용어 해소 문제는 어떤 인용문을 누가 말했나, 즉 인용문의 정보원 찾기 문제로 한정된다. 예컨대 "박 대표가 '……'라고 말했다"라고 했을 때 성과 직함만 나온 경우나, "그가 '……'라고 말했다"라고 했을 때 대명사가 지칭하는 인명을 찾는 문제가 된다. 이러한 대용어 해소를 위해서는 아래와 같은 간단한 부분 구문분석을 활용한다.

① 개체명 인식을 수행한다. 인용문 내에서 '인명+기관+직함', 또는 '기관'의 개체명이 인식된 경우, 해당 인용문의 발화자는 인식된 개인실명정보원 또는 집단정보원이다.

② '성+직함'만 나온 경우 인용문으로부터 앞으로 거슬러 올라가서 발견된 '인명+기관+직함' 중 '성+직함'이 일치하는 경우 해당 인용문의 발화자인 개인실명정보원이다. 이때 인용문의 '성+직함'과 인용문 앞에서 발

제(국제, 외신) 등.

견된 '성+이름+기관+직함'의 문자열에 대해 SVM(Support Vector Machine)를 수행해 일치 여부를 확인한다.

③ 형태소분석 결과 인용문의 주어가 인칭대명사로 나타난 경우, 인용문으로부터 앞으로 거슬러 올라가서 처음으로 나타난 '인명+기관+직함'이 해당 인용문의 발화자인 개인실명정보원이다.

대용어 해소 성능은 무작위로 선정된 300쌍의 개체 쌍에 대해 95%의 높은 정확도를 기록했다.

(2) 의미 중의성 해결

〈뉴스소스 베타〉에서 의미 중의성 해결 문제는 정보원과 관련된 동명이인과 이명동인 문제로 한정한다. 〈뉴스소스 베타〉에서는 인명, 기관, 직함이 모두 같아야 동일인물이고 어느 하나라도 다르면 다른 인물이다. 이 경우 이명동인 문제는 흔히 나타나지만, 동명이인 문제는 장기적으로 희소하게 나타난다.

동명이인 문제는 하나 또는 그 이상의 기사에 등장하는 같은 인명의 정보원이 동일인물인지 아니면 동명이인인지 판정하는 문제이다. 인명, 기관, 직함이 같은 경우, 기간을 막론하고 다른 매체의 다른 기사에 등장했더라도 동일인물일 가능성이 매우 높다.

인명, 소속, 직함이 모두 일치할 경우만을 동일인으로 판단할 경우 이명동인 문제가 발생한다. 개체명 인식이 제대로 이루어졌다고 가정할 경우, 같은 날짜에 이명동인, 즉 이름, 소속, 직함이 같은 서로 다른 두 사람은 존재할 수 없다. 한편, 다른 날짜의 뉴스에서 인명이 다른 경우 이름을 바꾸는 드문 경우가 아닌 한 동일인이 아니기 때문에 이름이 다르면 근사적으로 다른 인물이다. 문제는 인명은 같지만 기관과 직함이 다르면서도 실제로 동일인일 경우이다. 〈뉴스소스 베타〉에서는 이런 정보원을 다른 사람으로 조작적으로 정의한다. 이는 취재나 출입처 배정 시 정보원의 인격보다는 소속기관과 역할을 중시하는 저널리즘 관행을 반영한 것이다. 예컨대 '홍길동 OO그룹 회장'과 'OO당 국회의

원 홍길동'은 설사 동일인물일지라도 다른 소속이기 때문에 다른 출입 기자가 담당을 맡아 서로 다른 주제로 인용하는 기능적으로 다른 정보원이다.

5) 군집화, 순위화, 담론분석

(1) 유사도에 따른 문장과 기사의 군집화

〈뉴스소스 베타〉에서는 유사도를 VSM(Vector Space Model), 코사인 유사도(cosine similarity), 역색인 구조(inverted index structure), TF-IDF(Term Frequency-Inverse Document Frequency) 등 정보검색에서 널리 사용되는 알고리즘을 활용해 계산했다. 각 알고리즘을 간단하게 설명하면 다음과 같다(Ingersoll et al., 2013/2015, 87-100쪽).

① 역색인: 비교 대상이 되는 전체 기사 각각에 어떤 주요 단어, 즉 ID가 부여된 색인어가 있는지를 나타내는 색인 작업과 함께, 각 색인어가 어느 기사에 있는지, 즉 역색인 작업도 함께 수행한다. 이를 통해 유사도 계산의 속도를 개선할 수 있다.

② VSM: 같은 날짜의 기사들에 대해 전체 기사에 출현하는 n개의 단어로 이루어진 n차원 벡터공간을 가정한다. 이를 통해 단어의 유무에 따라 1, 0등의 값을 부여할 수 있다. 이때 〈뉴스소스 베타〉는 유사도 계산의 효율성을 위해 단어를 명사, 수치, 동사로 제한했다. 각 기사를 해당 단어가 출현하는지 여부에 따라 값을 부여한 단어벡터로 나타낸다.

③ TF-IDF: 각 단어의 벡터 값을 출현 시 1, 출현하지 않으면 0과 같은 단순한 출현 여부에 따라 부여하지 않고, 각 단어가 각 기사를 얼마나 대표하는지에 따라 부여할 수도 있다. TF는 단어 빈도로 한 기사에 특정 단어가 등장한 횟수이다. DF는 문서 빈도로 전체 기사에 특정 단어가 등장한 횟수이다. IDF는 DF의 역수이며 TF-IDF는 TF와 IDF를 곱한 값이다. 즉 전체 기사에서는 적게 등장하는 데 비해 특정 기사에 많이 등장하는 단어가 해당 기사를 대표하는 단어가 된다. TF만으

로 가중치를 부여하면 경제 기사에서 '경제'나 '주식'과 같이 모든 분석 대상 기사에서 자주 등장하는 상투어가 중요하게 취급될 수 있다.

④ 코사인 유사도: 비교 대상인 두 기사의 단어벡터 간 각도를 θ라고 할 때, 이 각도에 대한 코사인 값을 계산하여 유사도를 계산한다. 완전히 일치할 경우 θ가 0이므로 코사인 값은 1이 나온다. 임계값(threshold)을 정하여 이 값을 넘으면 유사한 것으로, 그렇지 않으면 유사하지 않은 것으로 판정한다. 유사한 것끼리 묶으면 기사와 문장의 군집을 만들 수 있다(clustering). 이러한 군집 안에서 TD-IDF 알고리즘에 의한 순위를 활용하여 대표문장과 대표기사를 제시할 수 있다.

〈뉴스소스 베타〉는 기사의 유사도를 먼저 판정하고 문장의 유사도를 판정했다. 기사 유사도와 문장 유사도는 위와 같은 방식으로 동일하게 수행했다. 기사나 문장의 유사도를 계산하면 중복을 제거하고 대표 기사나 문장만 볼 수도 있다. 그런데 코사인 유사도 값이 1인 완전 중복이 아닌 경우, 중복기사라고 해도 서로 다른 내용을 담고 있을 수 있다. 따라서 중복기사 중에 대표기사를 제시하고, 중복기사 중에 중요한 부분을 대표기사와 함께 보여 준다면, 여러 문서를 요약하는 효과를 기대할 수 있다.

〈뉴스소스 베타〉의 경우, 대표기사와 함께 중복기사라고 해도 중복되지 않은 인용문과 수치문을 보여 준다. 인용문과 수치문 역시 유사도를 계산했기 때문에, 대표문장을 제시할 수 있다. 다만 〈뉴스소스 베타〉의 경우 코사인 유사도 임계값이 0.45로 비교적 높게 정해져 있어서, 중복성이 상당히 크지 않으면 기사나 문장이 독립적으로 제시된다.[17]

17) 기사 간 유사도 임계값의 최적치는 정답 군집과 유사도에 의해 자동화된 군집을 비교하여 재현율과 정확도가 모두 충분히 높은 최적값을 결정했다. 참고로 유사도 기준을 낮추면 중복기사가 줄어들 수 있지만 대표기사만 검토할 때 너무 많은 기사를 건너뛰게 되는 트레이드오프가 발생하므로 실제 서비스나 분석에서는 목표에 맞춰 조정해야 한다.

(2) 의미연결망분석 및 규칙 기반 순위화

〈뉴스소스 베타〉는 다양한 순위화 알고리즘과 규칙을 도입 또는 개발해 적용하고 있지만, 개체명, 특히 정보원과 기사의 순위화를 위해 의미연결망분석의 일종인 뉴스 정보원 연결망분석을 가장 중요하게 활용한다. 뉴스 정보원 연결망은 같은 기사에 두 정보원이 직접인용문으로 함께 인용됐을 경우 이 정보원들 간에 서로 의미론적인 관계가 있는 것으로 보고 간접적으로 만드는 준연결망(quasi network)을 뜻한다. 정보원의 가중치는 뉴스 정보원 연결망분석을 통해 각종 중앙성(centrality)을 활용하여 다양하게 부여할 수 있다. 가장 단순한 연결정도 중앙성(degree centrality)의 경우 한 정보원의 중요도는 공동 인용된 정보원의 수에 따라 결정된다. 대체로 사실적이고 논쟁적인 여러 기사에서 다양한 정보원과 함께 다양한 주제에 대해 인용된 개인실명정보원의 가중치를 높게 계산한다. 즉 기사를 한 주제를 다루는 일종의 토론장으로, 개인실명정보원을 논객으로 간주하고, 논쟁적인 논객일수록 높게 평가한다(박대민, 2013).

연결정도 중앙성은 지역중앙성(local centrality)으로 각 정보원의 가중치는 단지 인접한 정보원의 수만 세어도 파악할 수 있어 효율적이다. 그렇지만 장기간의 분석에 대한 신속한 전처리 및 결과저장을 위해서는 하둡 등 분산처리 및 분산저장이 가능한 빅데이터 DBMS(Database Management System)를 활용해야 한다.

한편 〈뉴스소스 베타〉에서는 뉴스 기사 연결망분석도 수행했다. 뉴스 기사 연결망은 기사 유사도가 높거나 같은 정보원을 인용할 경우 관계가 있는 것으로 정의했다. 정보원 가중치와 마찬가지로 각 기사의 가중치 역시 뉴스 기사 연결망분석을 통해 얻은 연결정도 중앙성 값으로 부여했다. 하지만 이러한 뉴스 기사 연결망분석은 외적 타당도가 떨어져 추후 다른 알고리즘으로 대체해야 한다.

의미연결망분석 외에도 〈뉴스소스 베타〉는 저널리즘 관점에서 가중치를 부여하는 인위적인 규칙을 몇 가지 도입했다. 첫째, 정보원의 가

중치는 개인실명정보원, 집단정보원, 익명정보원 순이다. 개인실명정보원 사이에서는 뉴스 정보원 연결망분석에 따른 가중치를, 집단정보원과 익명정보원 각각은 무작위 가중치를 부여했다. 문장의 가중치는 이를 반영하여 개인실명정보원의 인용문, 집단정보원의 인용문, 수치문, 익명정보원의 인용문 순으로 노출된다. 둘째, 다양성을 제고하기 위해, 의미연결망분석에 따른 가중치가 낮더라도 소속이 다른 정보원과 그 인용문을 먼저 보여 준다. 즉 의미연결망분석에 따른 정보원 가중치가 '소속 1-정보원 A', '소속 1-정보원 B', '소속 2-정보원 C' 순일 경우, 비록 정보원 B가 정보원 C보다 더 중요하더라도 정보원 B는 정보원 A와 같은 소속이므로 검색결과는 정보원 A, 정보원 C, 정보원 B 순으로 제시한다.

5. 분석사례

이 연구는 뉴스 기사에 대한 NLP 방법 소개에 초점을 두고 있다. 따라서 본격적인 내용분석은 수행하지 않고 간단한 분석사례만 제시한다. 〈그림 6〉은 〈뉴스소스 베타〉의 전문가 버전 화면 예시로 '대통령'을 검색어로 입력했을 때 출력되는 2013년 6월 1일부터 1주간의 뉴스 정보원 연결망을 보여 준 것이다. 49개 매체 1,383개 기사에서 전체 정보원 수 729명 가운데, 고립자를 제외한 458명의 개인실명정보원이 시각화되어 있다. 박근혜가 결점의 크기가 가장 크고 중심에 위치하고 있어 가장 중요한 정보원임을 직관적으로 알 수 있으며, 최경환, 황우여, 김한길 등이 눈에 띈다.

다운로드한 엑셀 파일의 예는 〈그림 7〉과 같다. 〈그림 7〉은 인용문 파일로 첫 테이블(reference)의 첫 열(INFOSRC_NAME)은 개인실명정보원의 이름, 두 번째 열(STN_CONTENT)은 인용문을, 세 번째 열(ART_ID)은 기사 ID를 나타낸다. 기사 ID에서 앞의 8자리는 매체 ID,

그림 6. 〈뉴스소스 베타〉 전문가 버전 화면

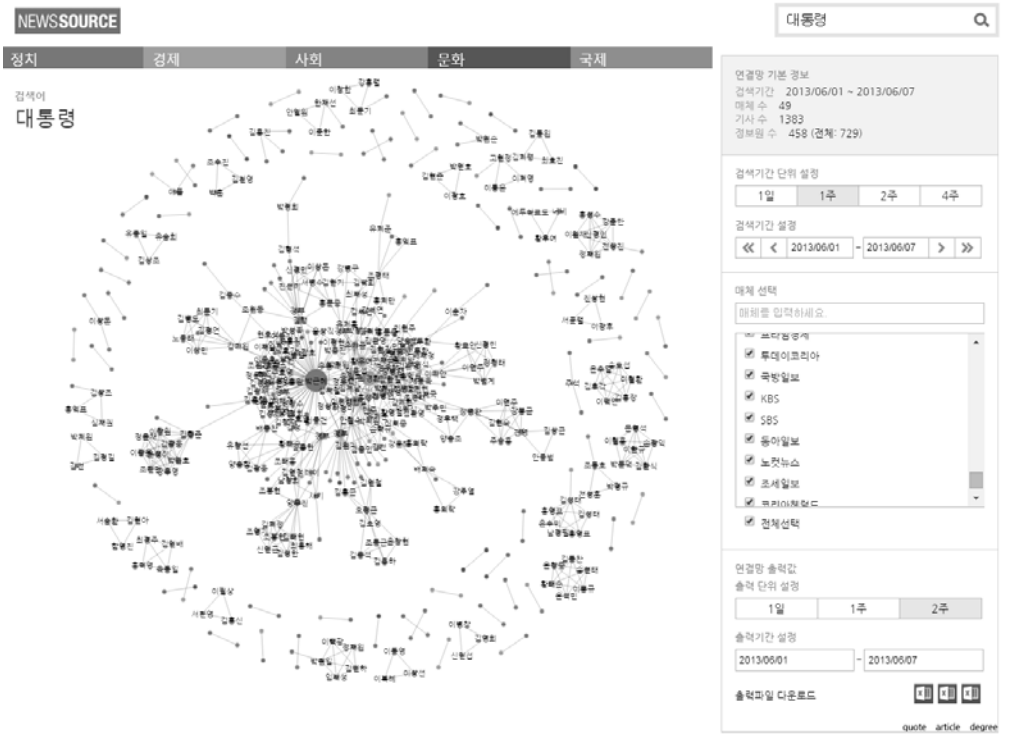


그림 7. 〈뉴스소스 베타〉의 정보원, 인용문 다운로드 기능

1	INFOSRC_I	STN_CONTENT	ART_ID
2	박근혜	박근혜(대통령) : "오늘의 대한민국은 그 분들의 숭고	"08100101.20130606100000200
3	박근혜	박근혜(대통령) : "국가를 위해 헌신한 분들을 위해 여	"08100101.20130606100000200
4	박근혜	박근혜(대통령) : "나라를 위해서 몸 받치신 분들이 흩	"08100101.20130606100000200
5	박근혜	朴 대통령 "대구국가産團 적극 지원"	"01500801.20130606111275115
6	박근혜	박근혜 대통령은 5일 "대구국가산업단지가 대구·경	"01500801.20130606111275115
7	박근혜	박 대통령은 앞서 기공식 축사를 통해 "이제 우리 산	"01500801.20130606111275115
8	박근혜	박 대통령은 이어 "대구를 비롯한 전국의 국가산업단	"01500801.20130606111275115
9	박근혜	박 대통령은 또 창조경제와 관련, "정보기술(IT)이나	"01500801.20130606111275115
10	박근혜	박 대통령은 이날 대구수목원에서 열린 환경의 날 기	"01500801.20130606111275115
11	박근혜	박근혜 대통령이 취임 100일을 하루 앞두고 "국민과	"02100101.20130603100000400

다음 8자리는 연월일이다. 보이지는 않지만 두 번째 테이블은 검색어와 검색기간, 지정매체가 기록된다. 식별된 인용문 수는 2,482개였다. reference 파일을 다운로드 받아 보면, 예컨대 '정부'처럼 인명이 아니지만 인명으로 식별된 오류가 있다. 앞서 〈뉴스소스 베타〉의 인명 식별과 성능의 정확도가 90%라는 것은 이러한 오류가 인명 10개 중 1개 정도라는 것을 의미한다. reference 파일을 통해 재현율은 확인하기 어렵지만 정확도는 따져 볼 수 있다. 식별된 개인실명정보원 수가 575명인데 이 경우 58명 정도 오류가 난다고 예상할 수 있다. 그러나 실제로는 오류가 17명 정도로 그 수가 훨씬 적다.

이 밖에 카인즈에서 제공되는 기사 원문과 제목, 날짜, 매체명 등의 메타데이터도 다운로드 받을 수 있다. 실제 분석에서는 정보원 분석과 인용문 분석을 할 수 있다. 즉 논란의 중심에 있는 정보원이 누구인지, 기사에 인용된 개인실명정보원들 전수의 명단과 중요도는 어떻게 되는지, 그들이 언제 어떤 매체에서 무슨 발언을 했는지, 소속별 정보원의 비중은 어떻게 되는지를 파악할 수 있다. 또 주요 정보원을 중심으로 인용문의 내용을 살펴봄으로써 담론의 흐름을 축약하여 파악할 수 있으며, 기사 원문도 살펴볼 수 있다. 엑셀의 다양한 필터 기능과 그래프 기능을 활용할 수 있으며, UCINET과 같은 연결망분석 프로그램에서 엑셀 파일을 불러들여서 다양한 분석과 시각화를 시도할 수도 있다.

6. 나가며

이 연구는 기존 언론학 연구에서 수행된 NLP의 한계와 NLP가 뉴스 영역에서 수행될 때 고려사항을 살펴보고, 담론분석을 위한 문장 중심의 뉴스 NLP의 필요성을 역설했다. 이어 뉴스 NLP의 사례로서 〈뉴스소스 베타〉의 NLP를 설명했다. 이 연구에서 제안한 1) NLP 의미분석을 추가하고 2) 최적화된 부분 구문분석을 지향하며 3) NLP 성능을 명기하고 4) 문장 수준의 분석을 통하여 담론분석을 추구하며, 5) 단어 중에서도 개체명을 선별하여 분석하여 보다 간결한 분석이 가능하다는 차이가 있다. 이를 기존 NLP와 대조하면 〈표 3〉과 같다.

이 연구는 새로운 NLP 알고리즘을 제안하는 식의 성능 개선은 아니지만, 저널리즘 관련 영역지식에 기초한 간단한 구문규칙을 추가해 뉴스에 관한 한 개체명 인식 및 의미분석의 성능을 개선했다. 또 언론학 분야에서 처음으로 의미분석 등 본격적인 NLP 기법을 활용한 컴퓨터 보조 질적 자료분석 소프트웨어를 개발하여 이를 3천만 건이 넘는 국내 뉴스 데이터에 적용했다는 점에서 의의를 갖는다. 특히 컴퓨터공학 분야에서도 아직 방향성도 제대로 잡히지 않은 난제인 문장 수준의 분석을 개체명 중심으로 유형화하고 가중치를 부여하여 분석하는 시도는 전례

표 3. 〈뉴스소스 베타〉 NLP와 기존의 NLP 주요 차이

	〈뉴스소스 베타〉	기존 방식
기능 및 성능	재현율 및 정확도 등 성능 명시, 개체명 인식, 의미 중의성 해결, 대용어 해소 기능 포함	어절분석, 수작업에 의한 수정
주요 분석수준	문장 중심	어절, 단어 중심
단어분석 핵심방법	개체명 인식	형태소분석
문장분석 핵심방법	의미 중심의 부분 구문분석	문법 중심의 완전 구문분석

가 없다. 향후 개체명 인식의 범위가 확대되고 다중분류가 심화되면, 담론분석에 보다 최적화된 컴퓨터 보조 질적 자료분석 소프트웨어가 될 수 있을 것으로 기대된다.

최근 언론학에서 컴퓨터 보조 질적 자료분석 소프트웨어를 활용하는 경우가 많아지면서, 언론학자가 직접 프로그램 개발 능력을 갖추면 바람직하다고 생각하는 이들도 늘고 있다. 그러나 현실적으로 언론학자가 컴퓨터공학 전공자나 숙련된 현업 개발자만큼 개발 능력을 갖추는 것은 어렵다. 그보다는 언론학자가 약간의 컴퓨터공학 용어나 알고리즘을 이해하고, 연구목적에 맞는 요구사항을 개발자에게 보다 정확히 설명하는 것이 현실적인 대안이다. 다른 한편으로는 저널리즘이나 언론학에 유용한 프로그램은 컴퓨터공학이나 언어학, 문헌정보학의 범용 프로그램을 활용하거나, 언론에 대한 이해가 부족한 개발자들만으로는 만들 수 없고 기자나 언론학자가 기획자로서 참여하여 기능을 제시하고 규칙을 제공해야 가능하다.

이 연구는 언론학자가 이해하면 좋을 개념이나 알고리즘을 컴퓨터공학에 대한 지식 없이도 이해할 수 있도록 수식을 배제하고 인문사회과학적 용어로 최대한 풀어서 설명하고자 했다. 하지만 개념이 상당히 많고 생소하기 때문에 충분히 쉽게 설명되지 못했을 가능성은 있다. 다만 이 연구가 언론학에서 NLP 기술을 도입할 때 고려할 요소들을 이해하는 출발점이 되기를 기대한다.

〈뉴스소스 베타〉는 물론 뉴스 NLP는 아직 넘어야 할 산이 많다. 우선 개체명 인식 성능의 개선이 요청된다. 구체적으로는 우선 조직의 개체명 인식 성능 개선을 위한 개체명 사전 확장이 필요하다. 수치 개체명 인식도 시간, 통화 등 다양한 단위와 연계하여 다중분류하면 더 많은 정보를 구체적으로 추출할 수 있을 것이다. 또 '인명+조직+직함' 외에 정보원이 '직업+인명'으로 나타나는 경우가 적지 않으므로 이에 대한 개체명 인식을 위해 '직업' 관련 개체명 사전을 추가로 구축해야 한다. 이 외에도 장소, 외국어 개체명의 한국어 표기, 상품명, 전문용어에 대

한 개체명 인식기능을 구현하면 좋을 것이다.

낮은 지면 분류 성능은 조직 개체명 인식을 개선하면 나아질 수 있다. 개체명 인식기능을 확대하면, 문장 분류도 더욱 확대될 수 있다. 예컨대 각종 수치에 대한 문장이나 장소, 상품 등과 관련된 문장을 따로 추출할 수도 있다. 개체명 인식기능을 개선하는 것은 대용어 해소 측면에서도 도움이 될 수 있다. 또한 '전년 대비'와 같이 수치와 관련된 대용어나, 두문자어, 약어의 대용어 해소 기능도 추가될 필요가 있다. 의미 중의성 해결 측면에서는 경력정보가 포함된 인명사전을 활용하면 이음동인의 식별을 좀더 강화할 수 있을 것으로 기대된다.

의미연결망분석 측면에서는 정보원 이외의 개체명에 대한 의미연결망분석이 가능한지, 연결정도 외에 다른 가중치 부여 방식이 어떤 저널리즘 관행을 반영할 수 있는지 추가로 연구해 볼 필요가 있다. 또 문장과 기사에 대한 의미연결망분석과 함께, 개체명, 문장, 기사 간의 다수준 의미연결망분석 연구도 가치가 있다. 예컨대 뉴스 정보원 연결망분석을 통해 도출한 정보원의 가중치, 즉 개체명 수준의 가중치가 문장이나 기사 등 다른 수준의 가중치에 어떤 영향을 주는가와 같은 연구가 가능하다. 특히 인공지능 연구 중 최근 각광을 받고 있는 딥 러닝(deep learning) 같은 신경망(neuron network) 연구 등을 참조할 수 있다. 추가로 기계학습이나 토픽 모델링 등 다양한 통계적 기법을 활용하여 보다 깊이 있는 뉴스 NLP를 구현할 수 있을 것이다.

참고문헌

강남준 · 이종영 · 오지연 (2008). 신문기사의 표절 가능성 여부 판정에 관한 연구: 컴퓨터를 활용한 형태소 매칭 기법을 중심으로. 〈한국언론학보〉, 52권 1호, 437-466.

강남준 · 이종영 · 최운호 (2010). 〈독립신문〉 논설의 형태 주석 말뭉치를 활용한 논설 저자 판별 연구. 〈한국사전학〉, 15호, 73-101.
강승식 · 우종우 · 윤보현 · 박상규 (2001). 정보 추출. 〈정보과학학회지〉, 19권 10호, 27-39.
김경모 · 신의경 (2013). 저널리즘의 환경 변화와 전문직주의 현실. 〈언론과학연구〉, 13권 2호, 41-84.
김관규 (2013). 인쇄신문과 인터넷신문의 동일 기사 제호 비교분석에 관한 연구. 〈언론과학연구〉, 13권 4호, 135-172.
김남일 · 백선기 (2008). TV뉴스의 특정지역 담론화와 사회문화적 함의: KBS-TV 서울 '강남권역' 보도의 담론 형성을 중심으로. 〈한국언론학보〉, 52권 2호, 125-150.
김동윤 · 김성해 · 유용민 (2013). 의견지면을 통해 본 한국 신문의 정파성 지형: 공정한 중재자인가, 편파적 대변자인가. 〈언론과학연구〉, 13권 3호, 75-122.
김만재 · 전방욱 (2012). 언어네트워크 분석 기법을 활용한 인간배아복제 신문보도 분석. 〈생명윤리〉, 26호, 19-34.
김선호 · 박대민 (2015). 청년실업 언론보도와 국민인식. 〈미디어이슈〉, 1권 14호. 서울: 한국언론진흥재단.
김성해 (2006). 언론과 (대외) 경제정책: 문화엘리트 모델의 시각에서 바라본 미국 언론의 정치성. 〈한국언론학보〉, 50권 5호, 30-54.
김세은 (2009). 한국 문화 저널리즘의 진단과 모색: 하나의 탐색적 논의. 〈미디어, 젠더&문화〉, 11호, 5-40.
김수영 · 박승관 (2010). 방송 경제위기 뉴스의 정치 의미화 과정에 관한 연구. 〈한국언론학보〉, 54권 5호, 301-326.
김영택 · 권혁철 · 옥철영 · 서영훈 · 이호석 · 이근배 · 윤덕호 · 문유진 · 강승식 · 이하규 · 심광섭 · 장병탁 · 윤성희 · 양재형 · 서병락 · 양승현 · 이재원 · 김성동 · 김유섭 · 박성배 · 이종우 · 장정호 · 오장민 · 황규백 · 김선 · 신형주 (2001). 〈자연 언어 처리〉. 서울: 생능출판사.
김일환 · 이도길 · 강범모 (2010). SJ-RIKS Corpus: 세종 형태의미 분석 말뭉치를 넘어서. 〈민족문화연구〉, 52호, 373-403.
김주희 · 서정연 (2010). 비형식적인 문서에 강건한 문장경계 인식. 〈한국정보과학회 학술발표논문집〉, 37권 1C호, 266-270.
김춘식 (2002). 한국 정치인의 부정적 캠페인에 관한 연구: 2002년 민주당 국

민경선제에 관한 언론보도 내용분석을 중심으로. 〈한국방송학보〉, 16권 3호, 207-231.
김춘식 · 이강형 (2008). 언론의 선거보도에 나타난 캠페인 관련 인용구: 2007년 대통령선거에 관한 신문보도 분석을 중심으로. 〈한국언론학보〉, 52권 4호, 377-400.
김형 · 양혜승 (2013). 저널리즘 원칙으로 본 온라인 뉴스제목의 형식적 및 내용적 문제점 분석. 〈언론학연구〉, 17권 3호, 87-114.
나이스신용평가정보 (2013). 〈한경기업총람〉. 서울: 한국경제신문사.
남인용 · 박한우 (2007). 대권 예비후보자 관련 신문기사의 네트워크 분석과 홍보전략. 〈한국정당학회보〉, 6권 1호, 79-107.
남재일 (2010). 한국 신문의 자살보도의 담론적 성격: 〈동아일보〉와 〈한겨레신문〉을 중심으로. 〈언론과학연구〉, 10권 3호, 191-224.
_____ (2008). 한국 객관주의 관행의 문화적 특수성. 〈언론과학연구〉, 8권 3호, 233-270.
박대민 (2013). 뉴스 기사의 빅데이터 분석 방법으로서 뉴스 정보원 연결망분석. 〈한국언론학보〉, 57권 6호, 233-261.
_____ (2014a). 〈담론의 금융화: 서민주택담론으로 본 한국 금융통치성의 대두〉. 서울대학교 언론정보대학원 박사학위논문.
_____ (2014b). 뉴스 정보원 인용에서의 폭발성과 언론의 편향성. 〈커뮤니케이션이론〉, 10권 1호, 295-324.
_____ (2015). 망언의 네트워크: 신문뉴스 빅데이터 분석으로 본 일본 망언 보도 10년사. 〈미디어이슈〉, 1권 12호. 서울: 한국언론진흥재단.
박수혁 (2008). 〈기계학습 기법을 이용한 문장경계 인식〉. 고려대학교 컴퓨터정보통신대학원 박사학위논문.
박재영 · 이완수 (2008). 역피라미드 구조의 한계에 대한 이론적 논의. 〈커뮤니케이션 이론〉, 4권 2호, 112-154.
박종민 · 이창환 (2011). 한국어 분석 프로그램(KLIWC)을 이용한 남북한 방송극의 언어문화 구조 차이 분석. 〈방송과 커뮤니케이션〉, 12권 3호, 5-30.
박한우 · Leydesdorff (2004). 한국어의 내용분석을 위한 KrKwic 프로그램의 이해와 적용: Daum.net에서 제공된 지역혁신에 관한 뉴스를 대상으로. 〈Journal of the Korean Data Analysis Society〉, 6권 5호, 1377-1387.

방은주 · 김성태 (2012). 국내 주요 신문의 소셜미디어 정보원 뉴스보도 분석. 〈사이버커뮤니케이션 학보〉, 29권 4호, 145-189.
백영민 · 최문호 · 장지연 (2014). 한미 정권교체에 따른 주한 미대사관 외교문서의 주제와 감정표현 변화: 위키리크스 공개 외교전문의 컴퓨터 언어처리 분석. 〈언론정보연구〉, 51권 1호, 133-179.
서울대학교 융합과학기술대학원 (2012). 〈글로벌 · 사회적 문제해결형 융합연구 사업모델 수립에 관한 연구〉. 한국연구재단.
서종석 (2011). 인지연산문법의 역동적 원시소. 〈언어와 언어학〉, 52권, 75-96.
설진아 (2013). 소셜 뉴스의 기사유형 및 뉴스특성에 관한 연구. 〈한국언론학보〉, 57권 6호, 149-175.
손동원 (2002). 〈사회 네트워크 분석〉. 서울: 경문사.
손승혜 · 황하성 · 장윤재(2011). 한국 언론의 교육보도 특성과 뉴스가치 분석. 〈미디어와 교육〉, 1권 1호, 115-145.
손호건 (2012). 인지연산문법틀 내에서의 동사 의미 기술과 형식화 작업. 〈언어와 언어학〉, 56권, 133-163.
오세욱 · 김선호 · 박대민 (2014). 스트럭처 저널리즘, 데이터 저널리즘을 넘어서. 〈2015 해외미디어동향〉(213-255쪽). 한국언론진흥재단.
윤상길 · 김영희 · 최운호 (2011). 대한매일신보 국문논설 번역문체 판별의 어휘적 준거에 대한 탐색적 연구: '형태주석 말뭉치'의 어휘를 중심으로. 〈언론정보연구〉, 48권 1호, 188-228.
윤석만 (2006). 자연언어 자동처리를 위한 언어이론모델 연구: J.-P. Desclés의 '적용인지문법'. 〈언어와 언어학〉, 38권, 35-57.
이동주 · 연종흠 · 황인범 · 이상구 (2010). 꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용도구. 〈정보과학회논문지: 컴퓨팅의 실제 및 레터〉, 16권 11호, 1046-1050.
이완수 (2006). 인물뉴스의 특성과 결정요인 연구: 사회자본(social capital) 이론을 중심으로. 〈한국언론정보학보〉, 32호, 295-332.
이완수 · 박재영 · 노성종 · 이수미 · 강충구 (2009). '메멘토 모리'(Memento Mori)의 정치학: 부음(訃音) 기사: 중앙일보 〈삶과 추억〉에 나타난 집합기억과 망각의 구성. 〈한국언론학보〉, 53권 5호, 221-243.
이완수 · 최명일 (2014). 한국 대통령 죽음에 대한 집단기억: 김대중 · 노무현 대통령 사후평가에 대한 미디어의 언어구성. 〈한국언론학보〉, 58권 2

호, 123-152.
이정교 · 서영남 · 최수진 (2009). 한국 신문에 나타난 강간보도의 통시적 분석: 강간통념과 양가적 성차별주의를 중심으로. 〈한국언론정보학보〉, 45호, 425-462.
이준웅 (2010). 한국 언론의 경향성과 이른바 사실과 의견의 분리 문제. 〈한국언론학보〉, 54권 2호, 187-209.
이준웅 · 양승목 · 김규찬 · 송현주 (2007). 기사 제목에 포함된 직접인용부호 사용의 문제점과 원인. 〈한국언론학보〉, 51권 3호, 64-90.
이창환 · 심정미 · 윤애선 (2005). 언어적 특성을 이용한 '심리학적 한국어 글 분석 프로그램'(KLIWC) 개발과정에 대한 고찰. 〈인지과학〉, 16권 2호, 93-121.
임봉수 · 이완수 · 이민규 (2014). 뉴스와 광고의 은밀한 동거: 광고주에 대한 언론의 뉴스구성. 〈한국언론정보학보〉, 66호, 133-158.
임영호 (2012). 신문 매체의 지리적 특성, 전국 · 지역 종합지와 무료신문의 배포와 기사, 광고 특성 비교. 〈언론학연구〉, 16권 1호, 287-314.
임영호 · 이현주 (2001). 신문기사에 나타난 정보원의 권력 분포: 1949-1999년 〈동아일보〉 기사의 내용분석. 〈언론과학연구〉, 1권 1호, 300-330.
장하용 (1995). 조직과 매스미디어 메시지의 상징 네트워크 분석에 관한 연구: 홍보, 광고, 경영진 상호교류의 네트워크를 중심으로. 〈한국언론학보〉, 36호, 111-120.
_____ (2000). 언론 메시지 분석의 새로운 접근: 경제위기 담론의 상징 네트워크 분석. 〈한국언론정보학보〉, 14호, 244-266.
전산용어사전편찬위원회 (2011). 〈컴퓨터 IT용어 대사전〉. 서울: 일진사.
정동우 · 황용석 (2012). 공정성 개념에 대한 신문기자들의 인식차이 연구. 〈언론과 사회〉, 20권 3호, 120-158.
차세대융합과학기술원 (2013). 〈빅데이터 기술을 활용하여 스마트 뉴스를 제공하는 모바일 앱 개발〉. 한국정보화진흥원.
최민재 · 김위근 (2006). 포털 사이트 뉴스서비스의 의제설정 기능에 관한 연구: 제공된 뉴스와 선호된 뉴스의 특성 차이를 중심으로. 〈한국언론학보〉, 50권 4호, 437-463.
최성필 · 송사광 · 정한민 · 황미녕 (2012). 텍스트 추론(Textual Inference) 연구 동향 분석. 〈정보과학회지〉, 30권 11호, 68-77.

최수진 (2014). 한류에 대한 미·중 언론보도 프레임 및 정서적 톤 분석: 싸이의 '강남스타일' 이후를 중심으로. 〈한국언론학보〉, 58권 2호, 505-532.
최원석·반현 (2006). 공중 의견과 행동에 대한 의제설정 효과 모형의 검증: 부동산 이슈보도를 중심으로. 〈한국언론학보〉, 50권 1호, 406-435.
최윤정·권상희 (2014). '빅데이터' 관련 신문기사의 의미연결망분석. 〈사이버커뮤니케이션학보〉, 31권 1호. 241-286.
한국전화번호부 (2013). 〈2013 사업체 CD 번호부〉. ㈜ 한국전화번호부.
한국정보통신기술협회 (2006). 〈정보통신용어사전〉. 서울: 두산동아.
홍원식, 김은정 (2013). TV 미디어 비평의 어제와 오늘: 〈미디어비평(KBS)〉 10년, 내용분석. 〈한국언론정보학보〉, 64호, 59-84.

Automatic Language Processing Advisory Committee National Research Council. (1966). *Language and machines: Computers in translation and linguistics*. Washington, DC: National Academy of Sciences.
Chomsky, N. (1964). *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.
_____ (1957). *Syntactic structures*. Mouton & Co.
Gazdar, G. (1985). *Generalized phrase structure grammar*. Harvard University Press.
Harris, Z. S. (1951). *Methods in structural linguistics*. University of Chicago.
Ingersoll, G. S., Morton, T. S., & Farris, A. L. (2013). *Taming text: How to find, organize, and manipulate it*. 임혜연 (역) (2015). 〈자연어텍스트 처리를 통한 검색 시스템 구축: 아파치 솔라, 루씬, OpenNLP 등 오픈소스 활용〉. 의왕: 에이콘.
Kaplan, R. M., & Bresnan, J. (1982). Lexical-functional grammar: A formal system for grammatical representation. In J. Bresnan (Eds.). *The mental representation of grammatical relations* (pp. 173-281). Cambridge, MA: The MIT Press.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284.
Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*.

University of Chicago Press.
Vachek, J. (1966). *The linguistic school of prague: An introduction to its theory and practice*. Indiana University Press.
Van Dijk, T. A. (1988). *News as discourse*. NJ: Lawrence Erlbaum.
Vater, H. (2001). *Einfuhrung in die textlinguistik: Struktur und verstehen von texten*. 이성만 (역) (2006). 〈텍스트의 구조와 이해: 텍스트언어학의 새 지평〉. 서울: 배재대 출판부.

국가법령정보센터 www. law. go. kr
꼬꼬마 프로젝트 kkma. snu. ac. kr
뉴스소스 베타 일반인버전 http://147. 47. 123. 2/NSNA/
뉴스소스 베타 전문가버전 http://147. 47. 123. 2/expert/
미디어가온 www. mediagaon. or. kr
카인즈 www. kinds. or. kr
한국정보통신기술협회 IT용어사전 word. tta. or. kr/terms/terms. jsp
KLT nlp. kookmin. ac. kr/HAM/kor/index. html
KrKwic www. leydesdorff. net/krkwic/
News Explorer http://news-explorer. mybluemix. net/

투고일자 2016년 01월 06일
심사일자 2016년 02월 19일
게재확정일자 2016년 03월 02일

Abstract

Natural Language Processing of News Articles: A Case of 'NewsSource beta'

Daemin Park
Senior Researcher, Korea Press Foundation

The use of natural language processing (NLP) to analyze news articles has increased gradually for computerized content analysis, computer assisted qualitative data analysis software, and semantic network analysis. However, the methodology of NLP has been considered as a black box in communication studies and not closely verified yet. This study argues that the level of analysis to perform discourse analysis of news articles should be named entities or sentences, not words. 'NewsSource beta', a news big data analytics system, has functions of NLP including not only morphological analysis and partial parsing, but also sentence boundary disambiguation, named entity recognition, classification of news articles and sentences, and semantic analysis such as word sense disambiguation and coreference resolution. Clustering and ranking algorithm by journalistic values like criticism, is adopted as well. This study explains NLP algorithm of 'NewsSource beta' and shows pilot analysis, and discuss how to improve NLP performance of news articles.

Keywords: Natural language processing of news articles, 'NewsSource beta', Computerized content analysis, Discourse analysis, News big data analytics, Named entity recognition