언어 자료로 세상 보기
— 산업 분야의 언어 처리와 세종 말뭉치 운용

전채남
더아이엠씨

1. 빅데이터의 시대, 쌓이는 언어 자료

빅데이터의 시대는 소셜 미디어의 일상화로부터 시작되었다. 몇 년 사이에 카카오스토리, 페이스북, 트위터, 인스타그램, 유튜브 등 다양한 소셜 네트워크 서비스(SNS)가 등장하고 이용자들이 급증하면서 엄청난 양의 데이터들이, 또 다양한 형태의 데이터들이 실시간으로 생산되고 있다. 소셜 네트워크 서비스는 우리가 살아가는 세상의 모든 흔적을 데이터로 남기고 있기 때문이다. 이와 함께 정보 통신 기술(ICT)의 발달로 우리가 이용할 수 있는 데이터들은 엄청나게 늘어나고 있다.

소셜 미디어 시대가 되면서 수십 억 페이지에 이를 만큼 늘어난 반면 띄어쓰기 오류, 구어체, 미등록어, 오용어 등으로 인해 문서의 질은 낮아졌다. 비공식적, 비격식적 문서의 양이 절대적으로 늘어나면서 불완전한 문장 또는 문법에 어긋나는 표현들도 함께 늘어나게 되었다. 이러다 보니 사용자들은 정보 과부하에 의한 피로감을 해소하기 위해 필요한 것만 요약하길 원한다.

실제로 최근 2년 사이에 세계는 이전 인류 역사의 전체 기간보다 더 많은 데이터를 생산하였다. 그리고 2020년경이 되면 1초당 약 1.7메가바이트의 새로운 데이터가 생산될 것으로 예상된다. 이런 데이터는 페이스북, 카카오

톡, 메일 등을 통해 보내지는 메시지와 메일들로부터 생산되고 있을 뿐만 아니라 디지털 사진들과 점차 증가하는 비디오 데이터로부터도 생산되고 있다.

2020년경에는 전 세계에서 60억 개 이상의 데이터를 온전히 수집하는 스마트폰이 사용될 것으로 예상된다. 전화기가 스마트해지고 있을 뿐만 아니라 스마트 텔레비전, 스마트 워치, 스마트 미터, 스마트 홈, 스마트 테니스 라켓, 스마트 전구 등 우리 주변의 대부분 기기가 스마트해지고 있다. 2020년경에는 500억 개 이상의 인터넷 연결 기기(IoT)가 운영될 것이다. 이것은 엄청난 양의 다양한 데이터(텍스트와 비디오 데이터로부터 센스 데이터까지)가 상상할 수 없는 수준까지 증가할 것이라는 의미한다.

세상이 스마트해지는 만큼 빅데이터가 갈수록 중요해지고 있다. 각종 언론에서 연일 빅데이터와 관련된 보도를 하고 사람들은 빅데이터를 대명사처럼 사용하고 있다. 많은 정부 기관과 기업은 업무를 효율적으로 수행하기 위해서, 성과를 개선하기 위해서, 가치를 창출하기 위해서 빅데이터를 점점 더 활용하는 추세이다.

빅데이터는 몇 년 전까지는 불가능하였지만 현재는 가능해진 기술, 즉 데이터를 수집하고 분석할 수 있는 기술의 발달과 관련이 있다. 새로운 기술로 인해 향상된 능력을 가질 수 있게 되어 더 많은 데이터를 수집하고 저장, 분석할 수 있어 빅데이터 이용이 가능하게 되었다. 위키피디아(Wikipedia)는 “빅데이터를 기존 데이터베이스 관리 도구로 데이터를 수집, 저장, 관리, 분석할 수 있는 역량을 넘어서는 대량의 정형 또는 비정형 데이터 집합”이라고 정의하고 있다. 그리고 여기에는 이런 데이터로부터 가치를 추출하고 결과를 분석하는 기술까지 포함하고 있다. 쇤버거와 쿠키어(2013)는 “큰 규모를 활용해 더 작은 규모에서는 불가능했던 새로운 통찰이나 새로운 형태의 가치를 추출해 내는 일”로 빅데이터를 정의하고 있다. 그들은 이유를 아는 것보다 결과를 아는 것이 중요하기에 인과성(causality)에서 상관성

(correlation)으로 분석의 초점을 옮겨야 한다고 주장한다. 아이디시(IDC, 2011)는 대규모의 다양한 데이터로부터 수집, 검색, 분석을 신속하게 처리하여 경제적인 가치를 발굴하도록 설계된 차세대 기술 및 아키텍처로 4브이(Volume, Variety, Velocity, Value)를 특성으로 제시하고 있다. 비정형 데이터의 활용에 주목하여 '생각을 만드는 기술'이라고 빅데이터를 정의하기도 하는데, 이는 사람들의 자연스러운 디지털 대화를 수집하여 그 속에 내포된 인식, 이해, 의견, 반응 등을 읽어 내는 기술이라는 의미를 가지고 있다(김정선, 2015).

지금까지 연구자들의 빅데이터 정의가 다소 차이는 있지만 전반적으로 데이터 수집, 저장, 정제, 분석 등과 관련된 기술뿐만 아니라 이를 활용하는 해석 능력, 이를 통해 가치를 창출 할 수 있는 통찰력을 포함하고 있다.

빅데이터에 대한 장점을 쇤버거와 쿠키어(2013)는 다음의 세 가지로 정리하였다. 첫째, 빅데이터로 인해 훨씬 더 많은 데이터를 분석할 수 있고 어떤 때는 특정 현상과 관련된 모든 데이터를 분석할 수 있다. 둘째, 빅데이터와 같은 방대한 데이터를 들여다볼 때는 정밀성에 대한 욕구가 다소 느슨해져서 샘플링 오류가 줄어들어 측정 오류에 대해서는 좀 더 관대해질 수 있다. 셋째, 거대 규모의 데이터를 취함으로써 인과관계 추구라는 오래된 습관에서 멀어지는 대신 패턴이나 상관성을 찾아내어 새로운 이해와 귀중한 통찰을 얻을 수 있다.

2. 텍스트 마이닝(Text Mining), 언어 자료 처리

빅데이터가 있다고 해도 우리가 잘 활용하여 데이터를 통한 통찰력을 발휘할 수 없다면 빅데이터는 거의 가치가 없다. 데이터를 잘 활용하기 위해서는, 즉 통찰력으로 유용한 결과를 도출하기 위해서는 데이터를 분석할

수 있도록 데이터의 정제와 처리가 필요하다. 그런데 일반적으로 텍스트 데이터는 비정형 데이터로서 복잡한 구조를 갖기 때문에, 이를 정제하고 처리하는 일은 쉽지 않다.

데이터 정제의 방법에는 수동적인 방법과 자동적인 방법이 있다. 수동적인 방법은 데이터를 수집한 이후에 연구자가 워드나 엑셀 프로그램을 사용하여 일일이 특수문자, 조사, 띄어쓰기 등을 확인한 후 분석 가능한 형태로 수정하고 정리하는 것이다. 때때로 텍스트 마이닝의 과정에서 수집된 단어를 정제하는 단계에서 컴퓨터 프로그래밍을 통해 자동으로 실시하기도 한다.

텍스트 데이터를 자동으로 정제하는 방법은 언어 정보 처리의 한 분야이다. 언어 정보 처리는 컴퓨터와 인간 사이의 언어 소통을 강화하는 일로, 컴퓨터를 비롯한 기기가 인간의 언어를 잘 알아듣도록 하는 일과 인터넷이나 문서로 축적되어 있는 언어 자료를 인간이 잘 이용할 수 있도록 하는 일이다. 정보 통신 기술의 발달로 사람들의 스마트 기기 사용이 늘어나면서 점점 언어 정보 처리가 중요해지고 있다.

언어 정보 처리는 컴퓨터가 사람의 일상 언어를 이해하고 생성할 수 있도록 함으로써, 컴퓨터를 인간의 지적 활동의 보조자 및 지원 도구로 활용하도록 한다. 다시 말해 언어 정보 처리의 한 목적은 컴퓨터가 인간의 언어를 자동 번역, 요약, 인식하도록 하는 것이다.

우리나라에서는 국어 정보 자료를 구축하여 정보를 쉽고 편리하게 활용할 수 있는 국어 정보화 기반을 조성하기 위해 1998년부터 '21세기 세종계획: 국어 정보화 추진 중장기 사업'을 실시하였다. 이에 따라 본격적으로 한국어 말뭉치(Corpus), 즉 '세종 말뭉치'가 구축되었다. 일정 규모 이상의 크기를 갖추고 그 시대의 언어 현실을 골고루 반영한 기계 가독형 국어 자료의 집합체인 말뭉치를 만들기 위해 현대 국어, 역사 자료, 구어 자료 등 다양한 분야의 자료를 망라하였다. 구축된 자료는 사전 편찬을 위한 용례 추출, 어휘의 빈도 조사, 언어 교육, 철자 교정기 및 번역 프로그램을 만들기 위한

분석 대상 자료 등으로 활용되고 있다.

언어 정보 처리 기술은 기술적인 특성에 따라 언어 처리 기반 기술과 응용 기술로 구분할 수 있다. 언어 처리 기반 기술은 입력된 텍스트의 형태, 구문, 의미, 구조 등을 자동 추출하거나 문장을 생성하는 기술을 말하며 언어 정보 처리 응용 기술은 정보 검색, 자동 번역, 텍스트 마이닝 기술 등을 포함한다.

텍스트 마이닝은 텍스트 데이터에서 형태소 분석과 자연 언어 처리(Natural Language Processing) 기술을 활용하여 일정한 의미 단위로 구획한 뒤 유용한 정보를 추출하는 기법이다. 텍스트 마이닝 기술을 통해 방대한 텍스트 뭉치에서 의미 있는 정보를 추출해 내고, 다른 정보와의 연계성을 파악하며, 텍스트의 빈도를 계산하거나 카테고리를 찾아내는 등 단순한 정보 검색 그 이상의 결과를 얻어낼 수 있다. 자연 언어 처리(NLP)는 인간이 발화하는 언어 현상을 기계적으로 분석해서 컴퓨터가 이해할 수 있는 형태로 만드는 자연 언어 이해, 혹은 컴퓨터의 언어를 다시 인간이 이해할 수 있는 언어로 표현하는 제반 기술을 의미한다. 컴퓨터가 인간이 사용하는 언어, 즉 자연 언어를 분석하고 그 안에 숨겨진 정보를 발굴해 내기 위해 대용량 언어 자원과 통계적, 규칙적 알고리즘을 활용한 형태소 분석이 사용되고 있다.

형태소 분석은 자연 언어 처리의 가장 기본적인 단계로, 어절을 의미를 갖는 가장 작은 단위인 형태소로 분리하고 품사를 찾아내는 것이다(곽수정 외, 2013). 빠른 속도의 형태소 분석을 위해 세종 형태 분석 말뭉치와 같은 기분석 말뭉치를 활용하기도 한다. 세종 형태 분석 말뭉치는 장기간에 걸친 형태 분석 및 검토 과정을 거쳐 형태 분석의 신뢰도가 상대적으로 높기 때문에, 규칙이나 통계를 기반으로 한 형태소 분석기와 결합하여 사용된다.

규칙에 의한 접근은 정해진 규칙에 따른 형태소 분석으로 특수한 부분에 대한 조건까지 제시해 줄 수 있으므로 중의성을 해결하는 성능이 우수한 반면 구축하고 유지 관리하는 데 부담이 많다. 시간과 비용이 많이 들어 많은 규칙을 만들기가 어렵고 규칙 관리자의 능력이 중요하므로 전문가가

필요하다. 띄어쓰기의 규칙으로 휴리스틱(heuristic), 바이어블 프리픽스(viable prefix)를 이용한 최장 일치 기법, 접두 명사 및 접두사와 이웃하는 명사 간의 조합 규칙 이용법, 형태소 분석 결과 이용법 등이 있다.

통계에 의한 접근은 단순하고 계산적으로도 부담이 적으며, 언어의 생산성에 잘 대처할 수 있고 영역 지식의 쉬운 활용이 가능하다. 그러나 이 접근은 통계를 낼 수 있는 정도의 자료(대용량 말뭉치 필요)를 확보하여야 하고 통계를 추출한 해당 말뭉치 분야나 유사 분야에 대해 우수한 성능을 보이지만 다른 분야에는 적용이 힘들다.

텍스트 마이닝은 텍스트로부터 고품질의 정보를 도출하는 과정과 관련이 있다. 고품질의 정보는 일반적으로 통계적 패턴 학습과 같은 수단을 가지고 패턴과 트렌드의 장치(devising)를 통해 도출된다. 텍스트 마이닝은 보통 입력 텍스트의 구조화－보통 파싱(parsing), 몇 가지 파생된 언어적 특징의 추가와 제거, 그리고 데이터베이스에 후속적 추가－, 구조화된 데이터 안에서의 패턴 도출, 그리고 마지막으로 출력의 평가와 이해 등의 과정을 포함한다.

텍스트 마이닝에 있어 고품질은 통상 관련성, 참신함, 그리고 관심도 등의 몇 가지 조합을 의미한다. 전형적인 텍스트 마이닝 기술은 텍스트 범주화(categorization), 텍스트 군집(clustering), 콘셉트/개체 도출, 알갱이(granular) 분류의 생산, 감성 분석, 문장 요약, 개체 관계 모형화(예: 개체 인식 사이의 관계 학습) 등을 포함한다.

텍스트 분석은 정보 검색, 단어 빈도 분포를 연구하는 어휘 분석(lexical analysis), 패턴 재인(pattern recognition), 태깅/주석(tagging/annotation), 정보 추출, 링크와 연상 분석을 포함하는 데이터 마이닝 기법, 시각화(visualization), 그리고 예측 분석(predictive analytics)을 포함한다. 핵심적으로 무엇보다도 중요한 목표는 분석을 위해 자연 언어 처리(NLP)와 분석 방법을 활용하여 텍스트를 데이터로 바꾸는 것이다.

텍스트 마이닝의 전형적인 활용은 자연 언어로 쓰인 일련의 문서를 스캔하

거나, 예측 분류 목적을 위해 문서군(document set)을 모형화하고, 혹은 추출된 정보를 가지고 데이터베이스 또는 검색 지수에 덧붙이는 것이다.

3. 빅데이터 분석, 언어 자료의 활용

빅데이터를 잘 활용하기 위해서는 우선 데이터를 잘 분석할 필요가 있다. 빅데이터 시대에 맞추어 데이터 분석의 방법도 놀랄 만한 진전이 이루어지고 있어 웹과 소셜 네트워크 서비스상의 다양한 데이터를 수집하고 분석할 수 있게 되었다. 웹과 소셜 네트워크 서비스에는 이용자들의 자발적 참여에 의한 비정형 데이터들이 축적되어 있어 특정 현상에 대한 색다른 분석을 해 볼 수 있다.

한편 디프 러닝(Deep Learning)과 같은 기계 학습(Machine Learning)의 알고리즘이 빅데이터 분석에 활용되기 시작하였다. 알고리즘은 현재 발성된 단어들을 이해할 수 있고 음성 단어를 텍스트로 바꾸어 기술할 수 있다. 그리고 내용, 의미, 감성을 알기 위해 이런 텍스트를 분석할 수 있다. 예를 들면 우리가 어떤 사람이나 사물에 대하여 좋게 이야기하는지 아닌지를 텍스트의 감성 분석을 통해 알 수 있다. 사람들이 세상을 이해하고 미래를 예측할 수 있도록 매일 점점 더 향상된 알고리즘들이 나타나고 있다. 이런 알고리즘은 기계 학습과 인공지능(Artificial Intelligence)－독자적으로 학습하고 의사결정하는 알고리즘의 능력－이 짝을 이루어 괄목할 만한 성과를 내고 있다. 이세돌 9단과 세기의 바둑 대결을 벌여 널리 알려진 알파고(AlphaGo)가 대표적인 사례이다.

빅데이터 분석은 통계학과 전산학의 분석 방법을 주로 사용한다. 정형 빅데이터의 분석에는 데이터 마이닝과 기계 학습의 알고리즘을 대규모 데이터 처리에 맞도록 개선하여 활용하고 있다. 인터넷과 소셜 미디어의 생활화로

비정형 빅데이터가 폭발적으로 증가하면서 비정형 빅데이터의 분석에는 주로 시맨틱 네트워크 분석, 감성 분석, 군집 분석 등을 많이 활용하고 있다.

시맨틱 네트워크 분석(Semantic Network Analysis)은 소셜 네트워크 분석의 한 종류이다. 소셜 네트워크 분석(Social Network Analytics)은 수학의 그래프 이론(Graph Theory)에 뿌리를 두고 있는데 사람이나 사물의 관계를 노드와 링크의 구조로 파악하는 기법이다. 소셜 네트워크의 연결 구조, 연결 중심, 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 측정하여, 소셜 네트워크상에서 입소문의 중심이나 허브 역할을 하는 노드를 찾는데 주로 활용된다.

시맨틱 네트워크 분석은 소셜 네트워크 분석을 텍스트 데이터에 응용하여 특정 현상의 인식이나 개념의 해석에 있어서 의미의 관계를 중심으로 분석하는 방법이다. 소셜 네트워크 분석이 사람들 사이의 특정 네트워크 특성으로 네트워크에 포함된 사람들의 사회적 행위를 설명하는 시도라면, 시맨틱 네트워크 분석은 소셜 네트워크를 기반으로 개념을 노드로 나타내고 개념 간의 관계를 연결로 나타낸 그래프이다. 개념은 단어나 구로 표현되는 정보 단위이며 의미는 다른 개념들과의 관계 속에 내재되어 있는 것이고 관계는 개념들 간의 연결을 나타내는 개념의 특정 범주를 의미한다. 시맨틱 네트워크는 다양한 개념들을 연결하고 의미가 주요하다는 관점에서 개념(concept)은 관련된 단어들의 합성체로서 사회 네트워크에서의 노드와 같고 개념 간 연결은 서술(statement)이며 네트워크 분석의 선이다.

시맨틱 네트워크 분석은 행위자 간 연결성을 중시하는 소셜 네트워크 분석과 달리 단어들의 공유된 의미를 토대로 체계적 구조를 분석하는 데 주안점을 두고 있다. 시맨틱 네트워크 분석은 핵심 단어 사이의 의미론적 연관이 중요한 요소이고, 핵심 단어의 동시 발생 빈도는 소셜 네트워크 관점의 중요한 요소이다. 시맨틱 네트워크 분석의 장점은 표준화되지 않은 텍스트 자료로부터 구조화된 형태의 정보를 추출함으로써 커뮤니케이션 과정의

양상을 시각화할 수 있다는 점이다. 검색된 결과를 추출하여 핵심 단어의 빈도와 매트릭스 자료를 만들어 핵심 단어 간 관계를 알아봄으로써 전체 데이터에 대한 구조화된 자료를 시각적으로 나타낼 수 있다. 시맨틱 네트워크 분석 방법은 핵심 어휘 및 단어 간의 의미론적 관련성을 규명하는 데 있어서 객관성을 확보하기 위해 유시넷(UCINET), 노드엑셀(NodeXL), 게피(Gephi), 파옉(Pajek) 등 자동화 도구인 소프트웨어가 개발되어 사용되고 있다.

시맨틱 네트워크 분석은 자동화된 도구인 소프트웨어를 활용하여 수행하는데, 가장 먼저 분석의 대상이 되는 데이터의 수집으로부터 시작된다. 데이터 수집의 원천으로부터 획득된 개별 객체들은 메시지 내 주요 단어의 빈도 분석을 통해서 주요 키워드를 도출하는 단계로 이어지고 이를 기반으로 네트워크 다이어그램의 설계를 통한 의미론적 분석이 진행된다.

네트워크 분석을 이용하여 연결 구조의 특성을 파악하는 것은 여러 지표를 통해 이루어진다. 분석은 네트워크를 구성하는 단위들을 노드로 단위들의 관계를 링크로 정의하여 이루어지는데 링크의 연결 정도(degree), 밀도(density) 등을 통해 네트워크가 어떻게 얼마나 결속되어 있는지 그 형태를 알아볼 수 있다. 시맨틱 네트워크 분석에서 활용되는 여러 지표 중에서 가장 중요한 개념이자 많이 쓰이는 측정 방법 중 하나는 중심성(centrality)이다. 중심성은 노드가 전체 네트워크에서 중심에 위치하는 정도를 표현하는 지표를 의미하는데, 이는 연결 중심성(degree centrality), 근접 중심성(closeness centrality), 매개 중심성(betweenness centrality) 등으로 세분화할 수 있다(freeman, 1978).

시맨틱 네트워크 분석을 실시한 후에 특정 대상에 대한 에고 네트워크 분석(Ego Network Analysis)을 실시하여 감성 분석(Sentiment Analysis)을 실시할 수 있다. 감성 분석은 소셜 미디어와 인터넷 데이터, 문서 등의 시맨틱 텍스트를 긍정, 부정, 중립으로 판별하여 선호도를 측정하는 기법이

다. 감성 분석은 특정 서비스 및 상품에 대한 시장 규모 예측, 소비자의 반응, 입소문 분석 등에 활용되고 있다. 정확한 감성 분석을 위해서는 전문가와 데이터에 의한 선호도를 나타내는 표현이나 단어 자원의 축적이 필요하다.

또 시맨틱 네트워크 분석의 한 분야로 군집 분석(Cluster Analysis)을 실시할 수도 있다. 군집 분석은 비슷한 특성을 가진 개체를 합쳐 가면서 최종적으로 유사 특성의 집단을 발굴하는 데 사용된다. 예를 들어 페이스북상에 주로 여행에 대해 이야기하는 사용자군이 있고, 자동차에 관심 있는 사용자군이 있을 때 이러한 관심사나 취미에 따른 사용자군을 군집 분석을 통해 분류할 수 있다. 군집 분석은 시맨틱 네트워크 분석에 활용하여 비슷한 의미를 표현하기 위해 사용하는 단어들을 묶어 가면서 담론(discourse)을 발굴하는 데 사용한다.

인터넷의 방대한 언어 자료를 활용하기 위해 거의 자동에 가까운 텍스트 마이닝을 실시하는 다양한 분석 솔루션들이 개발되고 있다. 다음소프트의 '소셜 매트릭스'는 온라인과 소셜 네트워크 서비스에서 공개된 데이터를 실시간으로 수집하고 자연어 처리와 텍스트 마이닝을 통해 언급 횟수, 연관 긍·부정어, 인플루언서(영향력자) 도출 등의 결과를 제공한다. 아르스프락시아(Ars Praxia)의 '심플'은 수집된 텍스트 데이터의 언어 정보 처리를 통해 온라인 위기 관리 모니터링을 온타임에 가까운 실시간성으로 단어의 양뿐만 아니라 양태를 함께 보여 주면서 단어의 파급력과 파급 효과의 확산 속도, 예측 상황 등을 고유 지표를 통해 다각도로 보여 준다. 유저스토리랩(Userstorylab)의 '트렌드믹스'는 주요 포털 사이트의 데이터를 수집한 후 내부의 언어 정보 처리 로직에 따라 분석하여 하나의 주요 단어에 대해 제일 파급력이 높은 사용자가 누구인지, 중요 이슈는 무엇인지 등의 결과를 제공한다. 더아이엠씨(The IMC)의 '텍스톰'은 텍스트 데이터 일관 처리 솔루션으로 웹과 소셜 네트워크 서비스에 수집된 데이터뿐만 아니라 연구자의 내부 데이터도

원시 데이터(Raw data)로 활용하여 전처리, 형태소 분석 등 텍스트 마이닝을 자동적으로 실시하여 단어 빈도, 트렌드 등의 결과를 제공하고 추가 분석을 위한 매트릭스 데이터도 제공한다.

4. 언어 자료로 세상 보기: '알파고' 열풍 분석

소셜 미디어의 일상화로 생활 속의 다양한 장면에 설명과 대상에 대한 인지 및 특정 현상에 대한 의견과 태도를 표현하는 비정형 빅데이터가 실시간으로 인터넷에 축적되고 있다. 소셜 미디어 대부분의 데이터는 언어 자료이고 비개입적 상황에서 자연스럽게 생산되었기에 이를 수집하여 발달된 알고리즘으로 분석할 수 있다면 정확하게 세상을 볼 수 있다. 즉 사회 여론, 기업과 브랜드에 대한 소비자 인식, 브랜드 태도 등을 세종 말뭉치와 언어 정보 처리를 기반으로 하는 텍스트 마이닝을 통한 언어 자료로써 다양하게 조사할 수 있다. 특히 응답자들의 소극적 협조에 의해 조사의 정확도가 떨어지고 점점 예측도 빗나가고 있는 여론조사를 대체해 나가고 있다. 텍스트 마이닝을 통해 정제된 데이터는 연구자가 주요 단어들을 선택한 후 매트릭스 데이터를 만들어 추가적으로 시맨틱 네트워크 분석(semantic network analysis)을 할 수 있다. 주요 의미 단어를 찾을 수 있고 인플루언서(영향력자)와 특정 의미 단어에 대한 인지, 연상(association)을 파악할 수 있다.

지난 3월 우리나라에 바둑과 인공지능(AI) 열풍을 몰고 온 '알파고'를 통해 산업 분야의 언어 처리와 세종 말뭉치 운영의 실제를 살펴보았다. 알파고 빅데이터 분석은 이세돌 9단과 인공지능 알파고의 바둑 대국이 온 나라의 뜨거운 화제가 되자, 알파고의 주요 의미 단어를 통해 연상을 살펴보고 시민들의 인식과 의견을 알아보려는 목적으로 실시하였다.

데이터의 수집은 빅데이터 일관 처리 솔루션인 '텍스톰(TextoM)'을 사용

하여 2016년 3월 3일에서 21일까지 우리나라 대표 포털인 N과 D 사이트의 뉴스, 블로그, 카페 등에서 '알파고'를 수집 단어로 하여 실시하였다. 수집한 데이터의 양은 약 17만 2천여 개였다.

표 1 채널별 상위 빈도 의미 단어

전체	뉴스	블로그	카페
이세돌	이세돌	이세돌	이세돌
인공지능	인공지능	인공지능	인공지능
바둑	바둑	바둑	바둑
구글	구글	구글	인간
인간	인간	인간	구글
승리	딥마인드	승리	사람
딥마인드	승리	딥마인드	컴퓨터
프로그램	프로그램	컴퓨터	프로그램
컴퓨터	개발	프로그램	관련주
개발	쇼크	충격	실수

수집한 알파고 빅데이터의 전처리와 형태소 분석은, 세종 말뭉치를 기반으로 통계적 형태소 분석을 하는 텍스톰을 사용하여 실시하였고, 그 결과는 [표 1]과 같다.

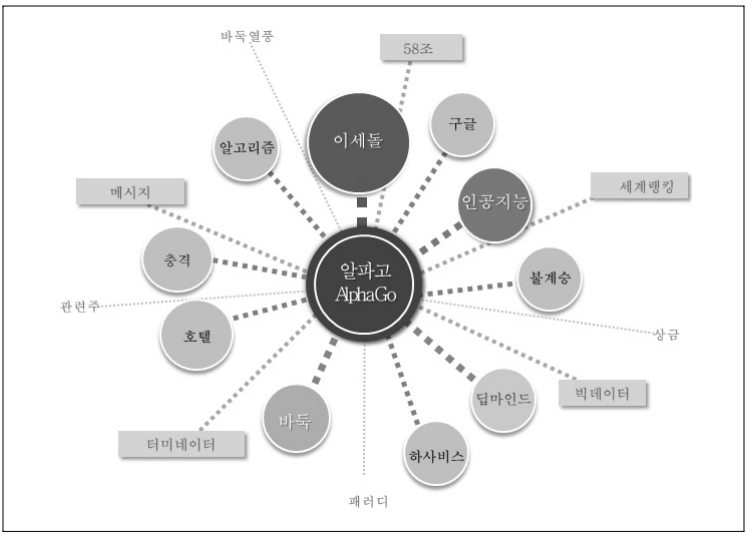
알파고 빅데이터 전체의 상위 빈도 의미 단어를 살펴보면 '이세돌'이 첫 번째에 위치하고 있었다. 알파고와 세기의 바둑 대국을 펼친 이세돌이 알파고 빅데이터에서 가장 많이 등장하고 있어, 알파고가 이세돌과 밀접한 관련이 있음을 보여 주었다. 다음으로 '인공지능'과 '바둑'이 높은 빈도수를 보여 알파고는 인공지능 바둑 프로그램이라는 정체성을 보여 주고 있었다. 그 밖에 '구글', '인간', '승리', '딥마인드', '프로그램', '컴퓨터', '개발' 등의 알파고 개발 주체와 바둑 대국과 관련된 의미 단어들이 나타나고 있었다.

채널별로 보면, 뉴스의 경우 인공지능의 성능에 대한 충격과 이와 연관된

개발의 필요성에 대해 좀 더 의미를 가지는 단어가 상위에 올라와 있었으며 블로그의 경우는 알파고 4승과 이세돌 1승의 바둑 대결의 승패에 대한 의미 단어가 상위에 있었다. 카페의 경우는 알파고와 관련한 주식에 대한 관심도를 보여 주는 의미 단어의 순위가 높았고 알파고와 이세돌의 바둑 대국에서 승패를 좌우했던 바둑 수에 대한 평가와 관련된 의미 단어들이 상위에 있었다. 이와 같은 텍스트 마이닝을 통해 알파고의 화제성과 유명세를 알게 해 주는 의미 단어들이 많이 출현하고 있음을 알 수 있었다.

주요 의미 단어의 영향력과 연상을 파악하기 위해 알파고에 대한 시맨틱 네트워크 분석을 실시한 결과는 [그림 1]과 같다. 알파고 전체 데이터에서 유의미한 단어 100개를 선정하여 네트워크를 그린 결과, 알파고 의미망에서도 바둑 대국을 펼친 이세돌과 가장 강한 연결을 보이고 있었다. 인간과 인공지능의 대국으로 관심이 높아지면서 이세돌과 알파고의 연결성이 매우 높게 나타났다.

그림 1 알파고 의미망



다음으로 '구글', '딥마인드', '하사비스'와 같이 알파고의 개발자에 대한 관심도가 매우 높게 나타났다. 알파고를 개발한 하사비스(Demis Hassabis)는 컴퓨터공학과를 졸업하여 게임 개발자로 활동하면서 인공지능 알고리즘에 대해 관심을 갖고 인간의 뇌구조에 대해 공부했으며, 마침내 인공지능 알파고를 탄생시켰다. 이렇게 알파고를 만든 개발자의 이력이나 경력, 구글과의 관계 등 알파고 프로그램 개발자와 구글 회사에 대한 관심도가 매우 높았다는 것을 의미망을 통해 알 수 있었다.

또한 '인공지능', '구글', '딥마인드', '빅데이터', '알고리즘'과 같은 단어들과도 큰 연결을 보이고 있었다. 알파고가 어떻게 바둑의 수를 스스로 계산하여 둘 수 있느냐에 대한 궁금증으로 이와 연관된 텍스트가 많았으며 이와 관련된 다양한 의견을 보여 주었다. 인공지능 바둑 프로그램인 알파고는 엄청난 경우의 수를 가진 바둑에서 통계적으로 이길 수 있는 확률에 의해 수를 둔다. 사람들은 이러한 알파고의 바둑 두는 방식에 대해 많은 관심을 가지고 있었으며, 이세돌이라는 한 사람과 1,200여 명의 바둑 기사들이 동시에 바둑 대결을 벌이는 상황에 비유하며 대국 자체가 불공정 대국이라는 의견도 있었다.

또한 '불계승', '호텔', '58조', '상금', '바둑 열풍', '바둑 랭킹' 등의 단어는 이번 바둑 대결로 인해 생겨난 유행 현상과 홍보 효과을 보여 주고 있다. 구글의 자산 가치 상승과 대국이 치러진 호텔, 바둑 용어와 바둑에 대한 관심 등 알파고 전체 네트워크에서 관련 기업의 홍보와 마케팅 측면에서의 효과를 보여 주는 단어들과 바둑 열풍과 연관된 단어와의 연결이 높았다.

한편 '충격', '터미네이터'와 같은 단어들과도 연결되어 많은 사람들이 이번 알파고를 영화 속 인공지능과 겹쳐 생각하여 무섭다거나, 섬뜩하다거나 하는 감정을 드러내기도 하는 것을 알 수 있었다. '패러디'는 텔레비전 프로그램에서 알파고의 모습을 패러디하거나 알파고와 관련한 유머러스한 내용들이 있었음을 보여 준다. 알파고가 자사고, 특목고와 같이 특수 고등학교를 일컫는 말이라는 유머를 통해 사회를 풍자하고 뿐만 아니라, 알파고의 천적은

두꺼비집, 바이러스로 컴퓨터 프로그램을 다운시키는 방법들로 바둑 대결을 이길 수 있다는 의미 단어도 있었다. 이처럼 여러 분야에서 알파고를 패러디한 의미 단어가 많았으며 그만큼 알파고가 큰 화제였다는 것을 보여 주고 있다.

5. 언어 자료의 미래

실제로 언어 자료를 통해 세상 보기가 가능한지를 알아보기 위해 '알파고' 빅데이터를 수집하여 분석해 본 결과 알파고의 대결자, 개발자, 인공지능으로 바둑을 두는 방식 등 바둑 대국과 알파고 프로그램 자체에 대한 관심도가 매우 높게 형성되어 있다는 것을 알 수 있었다. 알파고 전체 네트워크에서도 알파고의 주요 의미 단어, 연상, 의견, 태도 등을 통해 세상을 볼 수 있었고 알파고의 화제성을 확인할 수 있었다.

앞으로 정보 통신 기술이 점점 더 발달하면 인간과 디바이스의 커뮤니케이션은 더 일상화될 것이다. 그러면 데이터 생성 속도는 더 빨라지고 데이터양은 기하급수적으로 늘어나며 데이터의 형태도 다양해질 것이다. 이와 같은 정보 통신 환경의 변화에 적응하면서 데이터를 적극 활용하기 위해서는 향상된 언어 정보 처리 기술이 필요하다. 또 데이터 기반의 인공지능이 발달하면 할수록 기계가 데이터를 잘 읽고 말할 수 있도록 하는 언어 정보 처리도 점점 더 중요해질 것이다. 언어 정보 처리가 발달하면서 세종 말뭉치의 활용은 더 늘어나고 중요성이 높아질 것이다. 언어 정보 처리와 세종 말뭉치는 사회간접자본(Social Overhead Capital)과 같은 기반 기술이기에 연구개발을 더 강화해야 한다.

언어 자료가 빅데이터가 되고 분석 알고리즘이 발달하면 더 정확하게 세상을 보고 더 나아가서 미래를 예측할 수도 있을 것이다. 언어 자료의 가치를 높이기 위해서는 시대와 상황의 변화에 적응하는 언어 생활에 맞도록 세종 말뭉치를 지속적으로 구축하고 더 정교화할 필요가 있다.

참고 문헌

권혁철(2004), 인터넷 환경에서 언어 정보 처리 기술의 응용과 발전 방향, 한국어정보처리연구실, http://klpl.re.pusan.ac.kr/.../20040720104458_충북대-인터넷환경에서언어정보처리기술의응용과발전방향.ppt/.

곽수정·김보겸·이재성(2013), 한국어 형태소 분석을 위한 효율적 기분석 사전의 구성 방법, 정보처리학회논문지, ≪소프트웨어 및 데이터 공학≫ 제2권 제12호, 881~888.

김유경(2002), 전자정보통신 연구계를 움직이는 사람들: 언어정보처리, ≪전자신문≫, 2002년 8월 7일 자.

김윤정·이조은·이유리(2016), 네트워크 분석 기법을 통한 패션 상권의 특성 분석, ≪한국의류학회지≫, 제40권 제2호, 203~221.

김정선(2015), 혁신기술로서의 빅데이터 국내 기술수용 초기 특성 연구, 박사학위 논문, 이화여자대학교 대학원.

문화관광부(2003), ≪(21세기 세종계획) 말뭉치 활용 방안 연구≫, 문화관광부.

빅토르 마이어 쇤버거·케네스 쿠키어(2013), ≪빅데이터가 만드는 세상≫, 21세기북스.

서일원·전채남·이덕희(2013), 시맨틱 네트워크 분석을 이용한 원천기술 분야의 잠재적 기술 수요 발굴기법에 관한 연구, ≪기술혁신연구≫, 제21권 제1호, 279~301.

IDC(2011), The 2011 IDC Digital Universe Study.

Marr, B.(2016), *Bigdata in Practice,* Wiley, West Sussex.

Wikipedia. http://ko.wikipedia.org/wiki/자연_언어_처리/(검색일: 2016. 5. 31.).

_____. http://en.wikipedia.org/wiki/Big_data/(검색일: 2016. 5. 31.).

_____. http://en.wikipedia.org/wiki/Text_mining/(검색일: 2016. 5. 31.).