언어 자료의 보고, 빅데이터

이기황
다음소프트

1. 들어가는 말

이른바 빅데이터의 시대가 도래하였다. 거대 정보 기술(IT) 기업 가운데 하나인 아이비엠(IBM)이 최근 조사한 바에 따르면, 매일 2.5엑사바이트(1엑사바이트=1,000,000테라바이트)의 데이터가 생성되고 있다. 더욱 놀라운 사실은 전 세계에서 폭발적으로 생성되는 데이터의 90%가 최근 2년 이내에 생성되었다는 것이다(IBM, 2015). 이렇듯 방대한 양의 데이터는 미세먼지와 오존의 농도를 측정하는 센서, 카카오톡과 같은 메신저 서비스, 주식 거래소 등 다양한 원천에서 쉴 새 없이 생성되고 있다.

최근 빅데이터가 특별한 주목을 받는 이유는 그 규모 때문만은 아니다. 고도로 산업화된 오늘날 우리가 삶 속에서 겪는 여러 가지 문제를 해결하는 데 성공적으로 사용되고 있기 때문이다. 정부 주도의 빅데이터 활용 촉진 기관인 'K-ICT 빅데이터 센터(https://kbig.kr)'의 ≪빅데이터 글로벌 사례집≫(한국정보화진흥원, 2015, 2016)에서는 고객 관리, e-비즈니스, 의료, 제조, 재난·공공 등의 빅데이터 활용 분야를 소개하고 있는데 이는 수많은 빅데이터의 성공적인 적용 사례 중 극히 일부에 불과하다. 또한 최근 많은 화제를 몰고 온 인공 지능 바둑 에이전트 '알파고(AlphaGo)'는 대규모 데이

터의 유용성을 극명히 드러내었다.

주목할 것은 빅데이터의 80% 이상이 텍스트, 음성, 영상 등 구성 요소의 구조적 속성을 명시적으로 규정하기 어려운 반정형, 혹은 비정형 데이터로 구성되어 있으리라고 추정된다는 점이다(Economist, 2015). 여기서 텍스트라 함은 컴퓨터로 처리될 수 있는 형태로 저장된 글, 곧 언어 자료를 뜻한다. 실제로 앞서 소개한 빅데이터의 성공적인 적용 사례 가운데 상당수는 텍스트 자료의 분석을 통해 이루어진 것이다.

이와 같은 상황에서 우리는 빅데이터, 특히 텍스트로 이루어진 빅데이터를 언어의 탐구에 활용할 수 있는 가능성에 대하여 고려하게 된다. 언어 연구에 있어서 대규모 언어 자료인 말뭉치를 이용하는 것은 더 이상 낯선 일이 아니다. 그러므로 빅데이터를 언어 연구에 활용할 수 있는 방안에 대하여 고민하는 것은 매우 당연한 일이다.[1)]

이 글에서는 빅데이터의 개념과 특성을 언어 연구와 연관 지어 살펴보고 빅데이터를 언어 연구에 활용하기 위한 절차를 기술적 요건과 함께 소개하고자 한다. 그러나 자세한 기술적인 사항을 깊이 소개하는 것은 이 글의 범위를 벗어나는 일로 판단되어 개략적인 설명에 그쳤다.[2)] 또한 빅데이터를 언어 연구에 활용하는 일은 아직 걸음마 단계에 있으므로 명확한 방향을 제시하기 어려운 부분도 존재한다.

---

1) 언어 자료가 언어 연구에 유효한가에 대해서는 논쟁이 계속되고 있다. 촘스키는 최근 진행된 면담에서 제기된 빅데이터의 유효성에 관한 질문에 답변하면서 잘 설계된 실험을 통해 축적된 데이터의 사용에 대해서는 긍정적으로 평가하였으나 빅데이터의 유효성은 여전히 매우 부정적으로 평가하였다(뉴스센터, 2016).

2) 빅데이터를 언어 연구에 활용함에 있어서 적절한 기술의 도입과 활용은 필수적이다. 최근 말뭉치 언어학, 전산 언어학 등의 연구가 비교적 활발히 이루어지며 기술의 도입과 활용이 예전에 비해 활발해진 것은 사실이지만 빅데이터를 사용하기 위해서는 한 번의 도약이 더 필요하다.

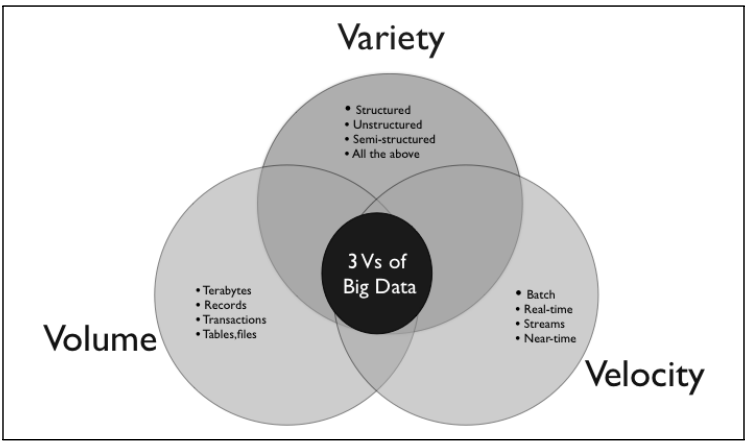
2. 빅데이터란 무엇인가?

2.1. 빅데이터의 정의

'빅데이터'라는 말은 이제 결코 생소한 용어가 아님에 틀림이 없지만, 어떠한 데이터가 빅데이터인지에 대해서는 명확히 규정하기가 쉽지 않은 것이 현실이다.[3] 그럼에도 불구하고 다음에 보이는 가트너(Gartner)의 정의는 가장 포괄적이면서도 고전적인 빅데이터의 정의로 널리 인용된다.

> 빅데이터의 정의(Gartner)
>
> Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

그림 1 빅데이터의 기본 속성: 3V



(https://commons.wikimedia.org/wiki/File:BigDataVs.png).

3) 빅데이터의 소개 자료는 이루 나열하기 힘들 정도로 많다. 다소 비즈니스 편향적이기는 하지만 IDG(2012)는 빅데이터에 대한 간략한 이해에 많은 도움이 된다. 한국소프트웨어기술인협회(2016)는 교과서로 활용될 수 있도록 단권으로 구성된 개론서이다.

위 정의의 핵심은 이른바 3V로 알려진 빅데이터의 기본 속성에 있다. [그림 1]은 빅데이터의 기본 속성 3V를 요약적으로 보여 준다. 이들 속성에 대하여 차례로 살펴보자.

첫 번째 V는 '규모' 혹은 '용량'을 뜻하는 볼륨(Volume)이다. 빅데이터의 '빅'은 규모가 크다는 빅데이터의 속성을 글자 그대로 드러낸다. 데이터 규모가 데이터의 가치를 결정짓는 유일한 속성은 아니지만 어느 정도의 질이 보장된다면 규모가 큰 데이터에서 추출된 정보의 신뢰성이 상대적으로 높다는 것은 일반적으로 알려져 있다. 그러므로 데이터에 기반을 둔 연구에서는 가능한 한 규모가 큰 데이터를 확보하기 위한 노력을 기울인다. 다만 데이터 수집, 저장, 처리 등의 공정에서 맞닥뜨리게 되는 현실적 한계는 무시할 수 없는 문제이다. 그런데 최근 컴퓨터 하드웨어와 소프트웨어의 급격한 발달로 이 한계가 급속히 무너지고 있다. 예를 들어, 2003년 인간 게놈 프로젝트를 통해 30억 개의 염기쌍 해독을 하는 데에 13년간 총 30억 달러의 비용이 들었는데 현재는 약 3.2기가바이트 용량의 인간 게놈 서열을 2시간 내에 1,000달러의 비용으로 해독할 수 있다고 한다 (IDG, 2012).

그렇다면 얼마나 규모가 큰 데이터가 빅데이터인가? 아이비엠이 2012년에 1,000명이 넘는 관련 분야 전문가들을 대상으로 실시한 설문 결과에 따르면 절반이 넘는 응답자가 적어도 1테라바이트는 넘어야 빅데이터라고 부를 수 있다고 하였다(슈렉 외, 2012). 1테라바이트는 디브이디(DVD) 220장의 저장 용량과 맞먹는 규모이다. 그런데 서두에 언급한 대로 매일매일 생성되는 데이터가 엑사바이트급인 오늘날 테라바이트급이 아니라 페타바이트급 데이터도 그리 희귀하지는 않다.

앞서 언급한 대로 데이터의 규모는 기본적으로 상대적인 개념일 수밖에 없다. 점점 더 많은 데이터가 생성될 것이고 이를 저장할 수 있는 저장 매체의 용량도 점점 더 커질 것이다. 또 한 가지 데이터 규모의 상대성에

영향을 미치는 요소는 데이터의 종류이다. 예를 들어, 같은 용량의 데이터라고 해도 그 데이터가 데이터베이스에 저장된 정형 데이터인지 동영상 데이터인지에 따라 전혀 다른 데이터 처리 방법이 요구되므로, 빅데이터의 정의는 특정한 유형의 데이터가 사용되는 산업과 응용 분야에 따라 달라진다.

결국 데이터의 규모는 중요한 빅데이터의 정의 요소 가운데 하나임에 틀림없지만 어느 정도 규모가 빅데이터에 해당한다고 규정하는 것은 의미가 없다. 따라서 앞서 보인 가트너의 정의와 같이 '혁신적인 형태의 자료 처리 방법이 필요할 정도의 규모'라고 정의하는 것이 합리적이라고 결론지을 수 있다. 이러한 생각을 적극적으로 확장하면, 규모로는 스몰데이터이더라도 그 데이터를 바라보는 새로운 관점과 시각이 동반된 새로운 방식의 자료 처리와 해석이 더해진다면 빅데이터로 볼 수 있을 것이라는 주장도 가능할 수 있다.

두 번째 V는 '속도'를 뜻하는 벨로시티(Velocity)이다. 빅데이터는 그 규모가 클 뿐만 아니라 매우 빠른 속도로 생성되는 데이터를 말한다. 앞서 보인 바와 같이 빅데이터는 다양한 원천으로부터 생성되는데 이들 원천의 공통된 특징은 데이터 생성 속도가 매우 빠르다는 것이다. 기상 정보 측정 장치는 시시각각으로 변하는 기상 정보를 측정하여 기록하는데 그 데이터 생성 속도는 오로지 미리 정한 데이터 측정 사이클에 의해 결정된다. 미국의 대형 마트 체인인 '월마트'에서는 시간당 100만 건 이상의 거래 정보를 처리한다고 한다(쿠키어, 2010).

소셜 네트워크 서비스(SNS)에서 생성되는 메시지 역시 매우 빠른 속도로 생성된다. 우리는 소셜 네트워크 서비스를 통하여 엄청난 속도로 소식이 퍼져 나가는 것을 여러 번 목격한 바 있다. 소식의 확산은 바로 데이터의 생성에 의해 이루어지는 것이다. 오늘날 스마트폰으로 대표되는 휴대 가능한 기기의 확산으로 개인화된 데이터가 엄청난 속도와 양으로 생성되고 있음은 널리 알려진 사실이다.

이렇듯 데이터의 엄청난 대응 속도에 반응하여 데이터의 분석 또한 실시간으로 이루어져야 하는 요구가 발생하였다. 실시간 교통 안내 시스템은 그러한 예의 하나다. 지속적으로 수집되는 교통량과 통행 정보를 바탕으로 한, 실시간으로 최적화된 교통 안내를 할 수 없는 시스템은 아무런 쓸모가 없다. 유용한 정보를 제공하는 빅데이터의 실시간 분석을 위해서는 새로운 기술의 개발이 필수적으로 요구된다. 그런 기술의 목표는 고속으로 생성되어 사라져 갈 수밖에 없는 데이터로부터 실시간으로 유의미한 정보와 지식을 산출해 내는 것이다.

세 번째 V는 '다양성'을 뜻하는 버라이어티(Variety)이다. 역시 앞서 언급한 대로 빅데이터는 다양한 원천으로부터 생성되기에 다양한 형태를 지닌다. 과거 컴퓨터를 이용한 데이터 처리에 있어서 처리 대상은 각종 측정치, 계산값 등을 기록한 수치 데이터가 주종을 이루었으며, 텍스트가 포함되었다고 해도 가로와 세로가 잘 구성된 목록형 데이터, 다른 말로 정형 데이터가 대부분이었다. 그런데 컴퓨터의 활용 분야가 확대되면서 전자우편, 소셜 미디어 포스팅과 같은 비정형 텍스트, 그리고 방대한 음성과 영상 데이터가 축적되고 있다. 글머리에서 언급한 바와 같이 이러한 반정형, 혹은 비정형 데이터가 오늘날 생성되고 있는 데이터의 80% 이상을 차지할 것이라 추정되고 있다.

다양한 형태의 데이터에 대한 관심은 빅데이터가 유행하기 전에도 존재하였다. 그런데 다양성이 빅데이터의 정의 요소로까지 중요해진 것은 최근의 기술적 진보와 무관하지 않다. 예를 들어, 최근 이미지 처리 기술이 급격히 발전하여 안면 인식을 통해 고객의 성별과 나이 등을 파악하고 이를 마케팅에 활용하는 일이 현실화되기 시작했다(간도미와 하이더, 2015). 즉 종전에는 축적되기는 하여도 제대로 활용할 수 없었던, 특수한 처리 기술이 요구되는 다양한 형태의 데이터가 풍부하고도 유용한 정보를 제공할 수 있는 소중한 자원의 지위를 갖게 되었다.

이상으로 가트너의 '빅데이터의 정의'에 나타난 3V, 즉 '볼륨(Volume), 벨로시티(Velocity), 그리고 버라이어티(Variety)'에 대하여 알아보았다. 이들 속성은 빅데이터에만 있는 고유 속성이라기보다는 빅데이터라는 '현상'을 이해하기 위한 상대적인 속성으로 이해해야 한다. 이 상대적 속성을 드러내는 핵심 요소는 혁신적인 데이터 처리와 해석 방법에 있다.

한편 위의 3V에 더하여 몇몇 기업들이 다음과 같은 V를 추가로 제시하였다(간도미와 하이더, 2015).

- Veracity: 아이비엠(IBM)이 추가한 네 번째 V로 '진실성'을 뜻한다. 이는 데이터 원천의 특성상 어느 정도 존재할 수밖에 없는 데이터의 불확실성, 비신뢰성 등을 지적한 것이다. 특히 소셜 미디어 등에 나타난 소비자의 의견 등은 어느 정도의 불확실성을 가질 수밖에 없다. 그러나 이러한 데이터의 유용성 자체를 부정할 수는 없다.
- Variability: 새스(SAS)는 빅데이터의 추가 속성으로 variability와 complexity, 즉 '가변성'과 '복잡성'을 제시하였다. 가변성은 데이터 생성 속도가 변할 수 있음을 지적한 것이고 복잡성은 데이터가 단일 원천으로부터가 아니라 여러 원천이 복잡하게 뒤엉켜 있는 상태에서 생성될 수 있음을 지적한 것이다.
- Value: 오라클(Oracle)이 추가한 것으로 '가치'를 뜻한다. 오라클에 의하면 빅데이터에 포함되는 원본 데이터들은 상대적으로 규모에 비해 가치가 적다. 그런데 이와 같이 '저가치 밀도'의 데이터를 대량으로 분석했을 때에 큰 가치가 창출될 수 있다는 것이다.

2.2. 소셜 빅데이터

앞서 빅데이터의 80% 이상이 비정형 데이터로 추정되며, 이 가운데 특히 텍스트 데이터가 매우 중요한 위치를 차지하고 있음을 언급하였다. 빅데이터를 구성하는 텍스트 데이터의 주요 원천은 소셜 미디어이다.[4)]

소셜 미디어는 인간의 사고와 행위를 인간 스스로 기록하여 생성하는 장이라는 점에서 특수한 가치를 지니고 있다. 사용자들은 소셜 미디어를 통해 연결되어 온라인 환경에서 새로운 공동체를 형성하고 서로의 생각과 일상을 공유하며 그 기록을 텍스트로 남긴다. 이와 같은 현상은 온라인 공동체의 형태를 지닌 온라인 카페나 게시판에서도 관찰된다. 나아가 포털 서비스의 포스팅, 뉴스 기사 등에 대한 댓글 또한 온라인 공간에서의 미디어 소비와 여론 생성 현장을 고스란히 기록하고 있다.

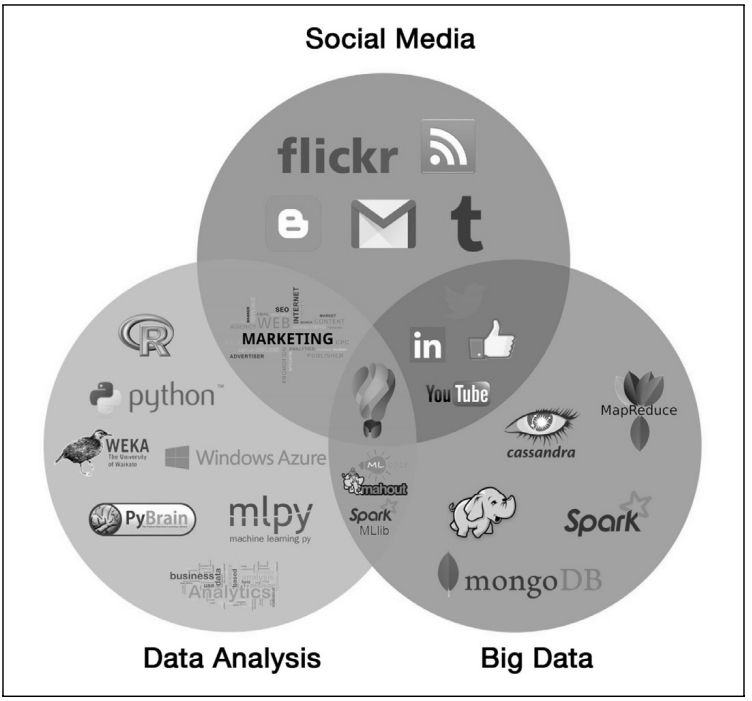
위와 같은 배경에서 미디어, 혹은 플랫폼으로서의 소셜 미디어와 빅데이터 현상이 결합된 소셜 빅데이터라는 개념이 등장하였다(송길영, 2012, 2015; 벨로-오르가즈 외, 2016). 언어 자료로서의 빅데이터를 이야기할 때 소셜 빅데이터의 개념을 다루지 않을 수 없다. 그러므로 이 글에서 빅데이터라 함은 곧 소셜 빅데이터를 가리킨다. [그림 2]는 벨로-오르가즈 외(2016)에서 보인 소셜 빅데이터의 개념을 나타내는 그림이다.

벨로-오르가즈 외(2016)가 [그림 2]를 통해 특히 강조하고자 하는 것은 소셜 빅데이터 분석이 근본적으로 학제적이라는 것이다. 이 논문에서는 관련된 분야로 데이터 마이닝, 기계 학습, 통계학, 그래프 마이닝, 정보 검색, 언어학, 자연 언어 처리, 시맨틱 웹, 온톨로지, 빅데이터 컴퓨팅을 나열하고 있다. 즉 수치 데이터와 정형 데이터 중심의 일반적인 빅데이터에 비해 비정

4) 소셜 미디어는 "개방, 참여, 공유의 가치로 요약되는 웹 2.0 시대의 도래에 따라 소셜 네트워크의 기반 위에서 개인의 생각이나 의견, 경험, 정보 등을 서로 공유하고 타인과의 관계를 생성 또는 확장할 수 있는 개방화된 온라인 플랫폼"으로 정의된다(위키피디아). 소셜 미디어에는 블로그, 소셜 네트워크 서비스, 위키, 마이크로 블로그 등이 포함된다.

형의 텍스트 중심이 되는 소셜 빅데이터의 분석에서는 언어학, 자연 언어 처리 분야 등의 참여가 필수적으로 요구되는 것이다. 바꾸어 말하면 소셜 빅데이터야말로 언어학의 연구 대상이라고 할 수 있다.

그림 2 소셜 빅데이터의 개념(벨로-오르가즈 외, 2016)



3. 언어 자료로서의 빅데이터

3.1. 균형 말뭉치와 모니터 말뭉치

언어 연구를 위한 대표적 언어 자료인 말뭉치는 여러 가지 기준에 따라 다양한 유형으로 분류할 수 있다. 그 가운데 하나는 원자료의 수집 방식에

따라 말뭉치를 균형, 혹은 표본 말뭉치와 모니터 말뭉치로 구분하는 것이다(맥에너리 외, 2011).

균형 말뭉치란 탐구 대상이 되는 언어 전체, 즉 모집단의 표본 자료로서의 역할에 충실하기 위해, 즉 대표성을 극대화하기 위해 말뭉치에 포함되는 원자료들을 다양한 매체, 장르에 걸쳐 선정하고 이들의 포함 비율을 적절히 조절하여 구성한 말뭉치를 말한다(맥에너리 외, 2006). 그러므로 균형 말뭉치는 조심스럽게 준비된 원칙에 따라 한 번 구성되면 그 내용과 규모가 고정된다.

이에 비해 모니터 말뭉치는 시간이 흐름에 따라 규모가 점점 커지는 말뭉치를 말한다. 원자료의 추가는 연간, 월간, 그리고 일간으로 이루어질 수도 있다. 말뭉치의 구성을 미리 설계할 수 없으므로 포함된 원자료의 균형성을 보장하는 것은 불가능하다. 모니터 말뭉치의 개념은 싱클레어(1991)에서 처음 주창된 것으로 균형 말뭉치의 폐쇄성이 살아 있는 언어 현상의 발굴, 그리고 매우 드물게 발생하지만 유의미한 언어 현상의 관찰에 적합하지 않다는 점을 지적하며 고안된 것이다.

맥에너리 외(2006)에서는 모니터 말뭉치의 강점으로 언어 변화의 관찰이 가능하다는 점을 들면서 신어의 등장과 사멸 등에 대한 연구를 예로 들었다. 또한 모니터 말뭉치가 매우 오랜 기간 축적되면 문법의 변화 등도 추적이 가능할 것이라고 보았다. 그러나 모니터 말뭉치는 균형성을 보장할 수 없으므로 신뢰도가 높은 통계 정보의 추출이 불가능하고 내용과 규모가 고정되지 않으므로 연구 결과의 비교가 불가능하다는 문제점을 지적하였다.

3.2. 말뭉치로서의 웹

모니터 말뭉치와 유사한 개념으로 말뭉치로서의 웹이 있다(킬가리프와 그레펜슈테트, 2003). 현재 주어진 가장 방대한 언어 자료임에 틀림없는

웹을 언어 연구에 활용하려는 시도이다. 웹에 존재하는 모든 언어 자료를 오프라인 사용을 위해 저장하는 것은 불가능한 일이므로 이 접근에서는 구글 등의 상용 검색 엔진, 그리고 언어 연구를 위해 특별히 고안된 인터페이스인 '웹콥(WebCorp)'(르누프, 2003)을 사용한다.

말뭉치로서 웹을 사용하는 가장 큰 장점은 역시 방대한 규모로부터 온다. 항상 그러한 것은 아니지만 보통의 균형 말뭉치에서 그 용례를 찾기 어려운 비교적 희귀한 언어 현상의 경우에도 웹에서는 상당히 많은 건수의 용례를 찾을 가능성이 있다. 또한 모니터 말뭉치의 경우와 마찬가지로 새로이 생성되는 데이터의 반영이 매우 빠르므로 새로이 등장한 단어나 표현의 추적에도 매우 유용하다.

그러나 말뭉치로서 웹을 사용할 때에는 검색 엔진 등의 제한적인 방법을 사용할 수밖에 없다는 근본적인 한계에서 벗어나기 어려워 그 용도가 제한될 수밖에 없다. 또한 검색 엔진의 검색 결과 수의 표시가 어떤 과정을 통해 생성되는지 알 수 없기 때문에 안정적인 통계 데이터를 얻을 수 있다는 보장을 하기 어렵다.

3.3. 말뭉치로서의 빅데이터

위의 원자료 수집 방식에 따른 말뭉치의 유형 분류로 보면 빅데이터는 모니터 말뭉치의 부류에 속한다. 한편 웹에서 생성된 데이터로 구성되었다는 점에서 말뭉치로서의 웹의 성격도 어느 정도 지니고 있다. 다만 빅데이터는 분석을 위해 모든 데이터를 데이터 원천으로부터 수집, 저장하여 사용하기 때문에 말뭉치로서의 웹이 갖는 한계는 없다.

말뭉치로서 빅데이터가 갖는 첫 번째 가치는 그 규모이다. 글쓴이의 일터에서 경험한 바에 따르면 대표적인 마이크로 블로깅 서비스인 트위터에서 생성되는 한국어 작성 트윗은 하루에 최소 500만 건에 이르며 대표적인

블로그 서비스인 네이버 블로그에서 생성되는 블로그 포스트는 하루 최소 50만 건에 이른다. 이 규모는 물론 말뭉치로서의 웹의 규모에는 훨씬 미치지 못하지만 그 어느 한국어 말뭉치보다도 규모가 크다.[5]

물론 아무리 규모가 크다고 해도 모든 한국어 사용자가 트위터, 혹은 블로그 서비스를 이용하는 것이 아니며 여기에 한국어의 양상이 모두 반영되어 있다고 할 수는 없다. 그럼에도 불구하고 한국어의 언어적 특성에 대한 탐구에 있어서 기존 말뭉치가 주지 못하는 풍부한 용례를 제공할 수 있다. 또한 맥에너리 외(2006)가 모니터 말뭉치에 대하여 지적한 대로 빅데이터는 균형성과 대표성에 있어서 문제가 있다고 볼 수 있다. 균형성과 대표성은 통계적 유의성에 기반을 둔 일반적인 통계적 연구 방법, 즉 모집단으로부터 추출한 비교적 작은 규모의 표본에서 통계적 유의성을 바탕으로 결론을 도출하고 이를 모집단으로 일반화하는 연구 방법에서 제기되는 문제이다. 그런데 모집단은 아닐지라도 모집단의 상당 부분을 포함하는 빅데이터에 있어서는 통계적 유의성이 그렇게 큰 의미를 갖지 못한다(간도미와 하이더, 2015). 그러므로 빅데이터로부터 통계적 정보를 얻기 위해서는 기존의 통계적 방법이 아닌 새로운 방법의 개발이 요구된다.[6]

대규모 말뭉치로서 빅데이터가 지니는 진정한 가치는 데이터의 원천인 소셜 미디어의 특성에서 찾아야 할 것이다. 예를 들어, 빅데이터는 언어 사용의 맥락과 언어 공동체에 대한 새로운 시각을 제공할 수 있을 것이다. 빅데이터에 포함된 언어 자료를 생성한 사람들은 넓게 보면 한국어라는 특정한 언어를 사용하며 이 시대를 살아가는 언어 공동체의 일원이다. 그러나 자료 생성자들은 지역, 직업, 연령, 관심사 등에 따라 각자 다른 맥락에서 한국어를 사용한다.

---

5) 한국어 트윗과 네이버 블로그 포스트의 일일 생성량은 글쓴이의 일터에서 측정한 것으로 실제 생성량과는 차이가 있을 것이다.

6) 특히 최근에는 베이지안 통계 기법의 활용이 여러 분야에서 시도되고 있다(알렌비 외, 2014; 스콧 외, 2016).

빅데이터는 언어 사용의 현장에서 동떨어지고 고립되어 존재하는 언어의 조각들이 아닌 무한히 확장될 수 있는 맥락 속의 언어를 들여다볼 수 있게 해 준다. 이를 통해 진정으로 동적인 언어 공동체의 생성과 발전의 양상을 살펴볼 수 있을 것이다. 이는 빅데이터가 단발적인 언어 사용을 담는 것에서 그치는 것이 아니라 다양한 환경에 처한 매우 많은 언어 사용자들의 언어 사용 양상을 비교적 장시간 지속적으로 담을 수 있기에 가능한 일이다.

말뭉치로서 빅데이터가 갖는 또 하나의 가치는 앞서 논의한 데이터 생성 속도와 관련이 있다. 소셜 미디어 서비스, 특히 마이크로 블로깅 서비스인 트위터에서는 초 단위로 새로운 트윗이 생성된다. 이를 통해 언어 연구자들은 언어 현장의 시간성을 정확히 파악할 수 있다. 특정한 발화가 이루어진 계절, 날짜, 시간은 물론이고 그 발화에 영향을 미쳤을 수도 있는 언어 외적 요소들에 대한 추적도 어느 정도 가능하다. 예를 들어, 우리 사회에 큰 영향을 미친 사건이 사람들의 언어 사용에 끼친 영향들을 관찰할 수 있을 것이다. 또한 특정한 언어 사용 양상이 사람들 사이에서 어떻게 펴져 나가는지, 즉 언어 사용 양상의 확산에 관한 연구도 가능할 것이다.

앞서 언급한 대로 빅데이터의 주요 속성 가운데 하나는 그 형식의 다양성이다. 이제까지의 언어 연구는 어쩔 수 없는 기술적, 또는 자료 수집의 제약으로 글말 중심으로 이루어져 왔다. 그런데 최근 기술의 발전 양상을 볼 때에 동영상을 언어 연구에 적극적으로 활용하게 될 날이 그리 멀지 않아 보인다. 먼저는 동영상에 포함된 음성의 인식이 가능하게 될 것이다. 이어서 동영상의 배경과 참여자를 인식하여 수많은 동영상을 자동으로 분류하고 이를 맥락화하는 일이 가능해질 것이다. 이는 언어 연구의 방법론과 대상에 있어서 작지 않은 변혁을 불러올 것으로 기대된다.[7]

---

7) 앞서 기술한 대로 빅데이터는 언어 연구의 대상과 방법에 상당한 변화를 가져올 것으로 보인다. 글쓴이는 한걸음 더 나아가 빅데이터가 기존 언어학의 확장이 아닌 전혀 새로운 시각의 언어학 출현, 즉 패러다임의 변화를 불러일으키지 않을까 조심스럽게 점쳐 본다.

4. 빅데이터 활용의 절차의 기술적 요건

앞에서 언급하였듯이 빅데이터를 언어 자료로 활용하기 위해서는 일정한 절차를 거쳐야 하며 각 절차에는 적절한 기술적 요건이 따른다. 이 글에서는 라브리니디스 외(2012)에서 도식화한 빅데이터 분석의 과정을 언어 연구의 관점에 맞추어 설명한다.

4.1. 데이터 수집

빅데이터를 언어 연구에 활용하기 위한 가장 첫 단계는 데이터 수집 단계이다. 소셜 미디어 서비스로부터의 데이터 수집에는 크게 세 가지 방법을 이용할 수 있다.

(1) 데이터 제공 서비스 이용

소셜 미디어 서비스로부터 데이터 제공 업무를 대행하는 업체의 서비스를 이용하는 방법으로, 가장 안정적으로 데이터를 수집할 수 있다. 대표적인 서비스 업체로는 트위터 데이터를 공급하는 '지닙(GNIP, www.gnip.com)'이 있다. 이 업체의 서비스를 이용하면 실시간으로 생성되는 모든 트윗, 혹은 표본 데이터를 수집할 수 있다. 이 업체에서 제공하는 가장 특징적인 서비스는 과거에 작성된 트윗에 접근할 수 있도록 해 주는 서비스이다. 과거

현존하는 가장 영향력 있는 과학 철학자 중 한 사람인 이언 해킹은 그의 저서 ≪우연을 길들이다≫에서, 19세기 초까지 모든 과학을 지배하던 결정론적인 믿음을 뚫고 다른 어느 법칙이나 원리로 환원될 수 없는 '우연'이라는 개념이 받아들여지는 과정을 보였다. 해킹은 우연과 확률은 과학에 있어서 거대한 사고의 전환을 가져 왔으며, 오늘날 가장 엄정한 과학으로 인정받는 양자론의 근간을 불확정성의 원리가 이루게 되었음을 논증하였다.
빅데이터는 자연 과학이 경험한 패러다임의 변화를 언어학도 마찬가지로 경험하게 될 것이라 믿는다. 이세돌과 알파고의 바둑 대국을 보면서 과연 알파고가 바둑을 이해하고 있는지에 대한 논쟁이 벌어졌던 것처럼 인간과 대화를 나누고 소설을 쓰는 컴퓨터가 과연 인간의 언어를 정확히 이해하고 있는지에 대한 논쟁이 벌어지는 날이 올 것이고, 그때 우리는 언어, 그리고 언어 연구에 대한 생각을 많이 바꾸어야 할지도 모른다.

에 생성된 트윗을 수집할 수 있는 유일한 방법은 이 서비스를 이용하는 것이다. 이와 같은 장점을 지닌 이 서비스를 사용하는 데에 있어서 가장 큰 난관은 사용료이다.

(2) 오픈 에이피아이(Open API) 사용

두 번째 방법은 소셜 미디어 서비스 업체에서 제공하는 오픈 에이피아이(Open API)를 이용하여 데이터를 수집하는 방법이다. 소셜 미디어 서비스는 다른 서비스와의 연동이 매우 중요하므로 서비스 업체에서는 다양한 형태로 데이터를 생성하거나 데이터에 접근할 수 있는 오픈 에이피아이를 제공한다. 오픈 에이피아이를 사용하기 위해서는 이를 주어진 규격에 따라 사용하는 컴퓨터 프로그램을 작성해야 한다.[8]

트위터의 경우 트윗의 수집에 이용할 수 있는 샘플 에이피아이, 검색 에이피아이, 스트리밍 에이피아이, 그리고 레스트 에이피아이를 제공한다. 이 가운데 스트리밍 에이피아이는 검색어를 지정하여 실시간으로 생성되는 트윗들을 수집할 수 있도록 해 준다. 한 번 실행할 때에 지정할 수 있는 검색어의 수에 제한이 있고 에이피아이 호출 간격에도 시간제한이 있기 때문에 대량의 트윗 수집을 위해서는 여러 컴퓨터에서 수집 프로그램을 구동해야 한다.

(3) 웹 접근 수집

마지막 방법은 오픈 에이피아이가 제공되지 않는 자료원으로부터 데이터를 수집할 때에 사용하는 방법으로, 인간이 웹브라우저를 통해 해당 서비스를 이용하는 것을 흉내 내는 프로그램을 작성하여 데이터를 수집하는 것이다.

---

8) 트위터 오픈 에이피아이(Open API)를 쉽게 사용할 수 있도록 도와주는 라이브러리들이 프로그래밍 언어별로 존재한다.

데이터 접근 스케줄링을 비롯한 많은 고려 사항이 따르는 방법이나 에이피아이가 제공되지 않는 서비스에 대한 유일한 데이터 수집 방법이다.

4.2. 데이터 정제와 정보 추출

많은 경우에 수집된 자료는 바로 사용할 수가 없고 일정한 정제 과정을 거쳐야 한다.

(1) 필터링

필터링이란 연구 목적에 부합하지 않거나, 나아가 연구 목적 성취에 방해가 되는 데이터를 걸러내는 과정이다. 트위터의 경우 자동으로 트윗을 생성하는 '봇'의 트윗을 제거한다든지, 이벤트성 트윗을 제거한다든지 등의 처리를 할 수 있다. 블로그의 경우 상당수를 차지하는 광고성 포스트를 제거할 수 있다. 물론 이 과정은 연구 목적에 따라 다른 접근을 하게 될 수도 있다.

(2) 중복 제거

소셜 미디어에서 생성된 데이터는 다양한 형태의 데이터 중복이 존재한다. 트위터의 경우에는 '리트윗'이라는 형태의 적극적인 데이터 전파 기능이 있어서 데이터 중복이 발생한다. 블로그의 경우에도 소위 '퍼나르기'에 의한 데이터 중복이 발생한다. 이러한 데이터 중복을 어떻게 처리할 것인가도 연구 목적에 따라 결정된다.

(3) 가공

데이터 가공은 오픈 에이피아이가 아닌 웹 접근 수집에 모아진 데이터일 경우 주로 이루어져야 하는 일이다. 즉 렌더링을 위해 부가된 에이치티엠엘(HTML) 태그 등을 제거하고 순수 텍스트만 추출하는 과정을 거쳐야 한다.

단순히 제거할 뿐만 아니라 최소한의 구조적 정보인 포스트의 제목, 본문을 구분하고 작성자, 작성 날짜와 시간, 태그 등을 분절해야 한다.

(4) 언어 처리

정보 추출 단계에서 이루어져야 할 일은 언어 처리이다. 언어 처리라 함은 자동화된 언어의 형식적 분석을 말하는데 현실적으로 한국어 데이터에 대하여 할 수 있는 언어 처리는 형태소 분석이다. 형태소 분석이 이루어지지 않은 데이터를 언어 연구에 이용하는 일은 불가능하지는 않다. 그러나 많은 경우에 형태소 분석은 효과적인 언어 연구를 위한 최소한의 언어 처리 단계일 것이다.

과거에는 일반 연구자들이 자동화된 데이터의 처리에 사용할 수 있는 형태소 분석기가 거의 없었지만 최근에는 무료로 사용할 수 있는 공개 형태소 분석기들이 등장하여 많은 연구자에게 큰 도움이 되고 있다. 그러나 형태소 분석기를 연구 목적에 맞게 조절하여 사용하는 일은 결코 쉽지 않은 일이다.

4.3. 데이터의 구조화와 통합

데이터의 구조화는 연구자들이 언어 처리가 적용된 데이터에 쉽고도 효과적으로 접근할 수 있도록 해 주는 일이다. 즉 자소, 음절, 형태소, 어절, 연어, 구 등의 언어 단위별로 다양한 질의 조건을 부가하여 데이터에 접근할 수 있어야 한다.

또한 데이터 통합에 의해 다양한 원천으로부터 수집된 데이터를 하나로 통합하여 접근할 수 있어야 하며 각종 메타 데이터에도 접근이 가능해야 한다.

매우 방대한 양의 데이터를 효율적으로 저장해야 하기 때문에 여러 대의

컴퓨터로 이루어진 분산 파일 시스템이나 분산 데이터베이스를 사용해야 하는 경우가 있다.[9]

4.4. 데이터 모델링과 분석

이 단계에서는 구조화된 데이터로부터 데이터를 효율적으로 질의하여 데이터에 대한 분석이 이루어져야 한다. 예를 들어, 특정 단어의 의미 변화에 대한 연구를 수행한다면 그 단어의 의미를 파악할 수 있는 실마리 문맥을 분류하고 그 변화를 추적할 수 있어야 한다.

빅데이터를 활용할 때에는 매우 많은 양의 데이터를 사용하게 되므로 자동화된 데이터 마이닝 기법의 도움을 받지 않을 수 없다. 다양한 데이터 마이닝 기법이 언어 연구에 어떻게 접목될 수 있는지에 대해서는 다양한 실험과 검증을 통해 밝혀져야 할 것이다.

이 단계에서는 통계적 분석도 수행하게 된다. 앞서 언급한 바와 같이 통계적 유의성에 기반을 둔 전통적 통계 분석 방법이 빅데이터에서는 큰 의미가 없다는 지적이 있다. 그러나 그 대안은 아직 마련되지 않았다.

한편 웹 규모의 빅데이터를 이용한 언어 처리의 경험을 간략히 요약한 해일러비 외(2009)는 빅데이터를 이용한 언어 연구에서도 참고할 만하다. 이 논문에서는 다음과 같은 '교훈'을 역설한다.

- "존재하지 않는 주석된 데이터를 기대하지 말고 존재하는 대규모의 데이터를 이용하라." 데이터를 이용한 연구에서는 탐구 대상 데이터를 해석하고 이용하기에 편리한 주석을 중요하게 여긴다. 나아가 주석된

9) 빅데이터의 분산 저장과 처리에 관련하여 많은 기술적 진보가 있었고 지금도 진행 중이다. 특히 아파치 하둡(http://hadoop.apache.org)과 아파치 스파크(http://spark.apache.org)는 오늘날 빅데이터 처리의 핵심 기반 기술이다.

데이터의 부재가 연구의 발전을 가로막는 장애임을 지적하기도 한다. 언어 연구에 있어서 주석된 데이터라 함은 형태소, 단어, 구, 문장 등의 언어 단위의 분절과 최소한의 해석이 이루어진 데이터를 말할 것이다. 이러한 언어 주석 데이터가 언어 처리와 언어 연구에 큰 도움이 됨은 틀림이 없다. 그러나 이러한 주석 데이터를 구축하는 데에는 엄청난 비용과 시간이 소요되며, 일반적인 규모를 훨씬 뛰어넘는 빅데이터에 주석을 부가하는 일은 비현실적이다. 그러므로 주어질 가능성이 거의 없는 주석 데이터에 의존하지 않고 대규모로 주어지는 원시 데이터를 어떻게 이용할 수 있는지에 대하여 깊이 고민해 보아야 한다.

- "정교하고 일반화된 규칙보다는 개별 사실에 집중하라." 이 논문의 저자들은 최근의 기계 번역에서 기억된(memorized) 개별 번역 사례의 중요성을 예로 들면서 일반화된 규칙보다 개별 사실을 최대한 이용할 것을 권장한다. 이 교훈은 언어 현상을 간명히 설명할 수 있는 일반화된 규칙의 작성에 관심을 두는 언어학 연구에서는 받아들이기 힘들 수도 있다. 다만 소규모 데이터에서 도출된 규칙은 언제든지 그 적용 범위에 한계가 올 수 있다는 점을 알아야 한다는 점에는 동의할 수 있을 것이다. 나아가 지식의 표현이 일차술어논리 형식의 간결한 규칙으로 되어야만 한다는 것 또한 편견일 수 있다는 사실을 인정해야 한다. 수많은 개별 사실과 개별 사실들의 조합으로부터 도출된 확률적 표현 또한 훌륭한 지식 표현의 방법 가운데 하나이다.

4.5. 결과의 해석

가장 어려운 단계이다. 연구 가설이 주어진 연구였다면 빅데이터에 의해 가설이 지지되는지 그러지 않는지를 검증하여야 하며, 연구 가설이 주어지지 않은 탐색적 연구였다면 연구 결과가 다른 연구로 이어질 수 있도록 정리해야

한다.

결과의 해석을 효과적으로 전달하기 위하여 적절한 시각화 기법의 활용을 적극적으로 고려해 볼 필요가 있다. 방대한 데이터로부터 도출된 복잡한 결론을 글로만 표현하는 데에는 한계가 있을 때가 많기 때문이다.

5. 맺는말

이 글에서는 빅데이터의 특성을 먼저 살펴보고, 빅데이터, 특히 소셜 빅데이터가 언어 연구에 새로운 전기를 마련해 줄 수 있는 언어 자원으로서의 가치가 있음을 논하였다. 이어서 빅데이터를 언어 연구에 활용하기 위한 절차를 기술적 요건과 함께 간략히 설명하였다.

앞서 언급한 대로 빅데이터를 언어 연구에 활용하는 일은 아직 걸음마 단계에 있다. 그리고 해결해야 할 문제도 다수 존재한다. 특히 개인 정보 보호의 문제는 연구 윤리에 있어서 매우 중요한 문제이다. 또한 비즈니스의 목적으로 서비스되고 있는 데이터를 이용하기 때문에 데이터의 공유 등에 있어서 자유롭지 못한 부분이 많은 것도 문제이다.

이미 우리는 빅데이터의 시대에 살고 있고 어떠한 형태로든 빅데이터와 연관이 되어 있다. 이러한 시대에 빅데이터를 언어 연구에 활용하는 것은 필연적인 일일 수도 있다. 활발한 토론과 다양한 시도가 이루어지기를 기대해 본다.

참고 문헌

송길영(2012), ≪여기에 당신의 욕망이 보인다≫, 쌤앤파커스.

_____(2015), ≪상상하지 말라≫. 북스톤.

이안 해킹 저·정혜경 역(2012), ≪우연을 길들이다≫, 바다출판사. / Hacking, I.(1990), *The Taming of Chance*, Cambridge University Press.

한국소프트웨어기술인협회 빅데이터전략연구소(2016), ≪빅데이터 개론≫, 광문각.

한국 IDG(2012), 빅 데이터의 이해, ≪IDG Tech Report≫. http://kbig.kr/index.php?sv=title&q=knowledge/pds_&tgt=view&idx=15326/(검색일: 2016. 5. 29.).

한국정보화진흥원 미래전략센터(2015), ≪2015년 빅데이터 글로벌 사례집≫. http://kbig.kr/index.php?sv=title&q=knowledge/pds_&tgt=view&idx=15614&sv=title/(검색일: 2016. 5. 29.).

한국정보화진흥원 ICT융합본부(2016), ≪2016 글로벌 빅데이터 융합 사례집≫. http://kbig.kr/index.php?sv=title&q=knowledge/pds_&tgt=view&idx=16137/(검색일: 2016. 5. 29.).

Allenby, G. M., Bradlow, E. T., George, E. I., Liechty, J. and McCulloch, R. E.(2014), Perspectives on Bayesian Methods and Big Data, *Customer Needs and Solutions*, 1(3): 169~175.

Bello-Orgaz, G., Jung, J. J. and Camacho, D.(2016), Social big data: Recent achievements and new challenges. *Information Fusion*, 28: 45~59.

Cukier, K.(2010). The Economist, Data, data everywhere: A special report on managing information. http://www.economist.com/node/15557443/(검색일: 2016. 5. 29.).

Gandomi, A. and Heider, M.(2014), Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35: 137~144.

Gartner, IT Glossary: Big Data. http://blogs.gartner.com/it-glossary/big-data/(검색일: 2016. 5. 29.).

Halevy, A., Norvig, P. and Pereira, F.(2009), The Unresonable Effectiveness of Data, *IEEE Intelligent Systems*, 8~12.

Economist(2015), The data deluge: Five years on. http://www.veritas.com/content/dam/Veritas/docs/reports/EIU-veritas-data-deluge.pdf/(검색일: 2016. 5. 29.).

IBM(2015), Big Data and Analytics. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html/(검색일: 2016. 5. 29.).

Kilgarriff, A. and Grefenstette, G.(2003), Introduction to the special issue on the Web as Corpus, *Computational Linguistics*, 29(3): 333~347.

Labrinidis, A. and Jagadish, H. V.(2012), Challenges and opportunities with big data, *Proceedings of the VLDB Endowment*, 5(12): 2032~2033.

McEnery, T. and Hardie, A.(2011), *Corpus Linguistics: Method, Theory and Practice*, Cambridge University Press.

McEnery, T., Xiao, R. and Tono, Y.(2006), *Corpus-based Language Studies*, Rutledge.

Newscenter, Conversations on linguistics and politics with Noam Chomsky, 2016년 4월 18일 자. http://www.rochester.edu/newscenter/conversations-on-linguistics-and-politics-with-noam-chomsky-152592/(검색일: 2016. 5. 29.).

Renouf, A.(2003), WebCorp: providing a renewable data source for corpus linguists, S. Granger and S. Petch-Tyson(eds.) *Extending the Scope of Corpus-based Research: New Applications, New Challenges*, Rodopi, 39~58.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D. and Tufano, P.(2012), Analytics: The real-world use of big data. How innovative enterprises extract value from uncertain data, *IBM Institute for Business Value*. http://www-03.ibm.com/systems/hu/resources/therealworduseofbigdata.pdf/(검색일: 2016. 5. 29.).

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E.(2016), Bayes and big data: The consensus Monte Carlo algorithm, *International Journal of Management Science and Engineering Management*, 11(2): 78~88.

Sinclair, J.(1991), *Corpus, Concordance, Collocation*, Oxford University Press.