건설 산업 내 비정형 텍스트 데이터를 활용한 자연어 처리 (Natural Language Processing, NLP) 기반의 글로벌 연구 동향

이지희 Texas A&M University 박사후연구원, jhlee04@tamu.edu

KICEM

1. 서론

과거에 컴퓨터 기반의 자동번역 소프트웨어를 사용해본 독자라면 오늘날 구글 번역(Google Translation)이나 파파고에서 제공하는 번역 서비스의 품질이 얼마나 향상되었는지 짐작할 수 있을 것이다.구글 번역과 같은 인공지능 기반의 자동번역기에는 컴퓨터가 인간의 언어를 이해하도록 하는 자연어 처리 (Natural Language Processing, NLP) 기술이 핵심이 된다. 자연어가 사람들이 사용하는 언어로서 한국어, 영어, 중국어 등을 일컫는다면 자연어 처리(NLP)는 인간의 언어를 분석, 이해, 생성할 수 있는 딥러닝 기반의 기술을 의미한다. 최근 자연어 처리(NLP) 기술의 발전이 주목을 받고 있는 이유에는 인공지능에서 사용자의 명령을 인식하고 수행하는 시스템의 핵심 기술 중 하나가 자연어 처리이기 때문이다. 특히 음성인식 기술과 자연어 처리 기술이 결합하면 인간의 언어를 이해하고 반응할 수 있는 애플의 시리(Siri), 아마존의 알렉사(Alexa) 등과 같은 인공지능 서비스를 개발할 수 있다.

건설 산업에 자연어 처리(NLP) 기술이 활용된 사례는 아직까지 많이 보고되지 않고 있다. 건설 분야에 자연어 처리 기술 기반의 인공지능 서비스를 접목시킬 여지가 아직은 많지 않기 때문이라고 이해할 수도 있겠다. 그러나 건설 활동에서 발생하는 대다수의 정보들이 비정형의 텍스트 데이터 형태를 띄고 있다는 점에서 텍스트 문서 분석을 통한 건설 분야의 자연어 처리(NLP) 기술의 적용 가능성은 무궁무진하다고 할 수 있다. 실제로 건설 프로젝트는 방대한 양의 문서 작업의 결과라고 말할수도 있기 때문이다. 또한 건설 산업 내에서 꾸준히 이슈가 되고 있는 빅 데이터(Big Data)의 시각에서는 데이터의 양(Volume) 뿐만 아니라 데이터 유형의 다양성(Variety) 측면을 강조하기 때문에 자연어 처리(NLP)에 기반한 비정형 텍스트 데이터 분석은 건설 분야 빅 데이터 연구에 있어서도 중요한 과제라고 할 수 있다. 이에 본 고에서는 텍스트와 같은 비정형 형태의 데이터를 컴퓨터가 이해할 수 있도록 지원하는 자연어 처리(NLP) 기술의 개념과 관련 기술에 대해 소개하고, 건설 산업의 자연어 처리(NLP) 관련 연구 동향에 대해 살펴보고자 한다.

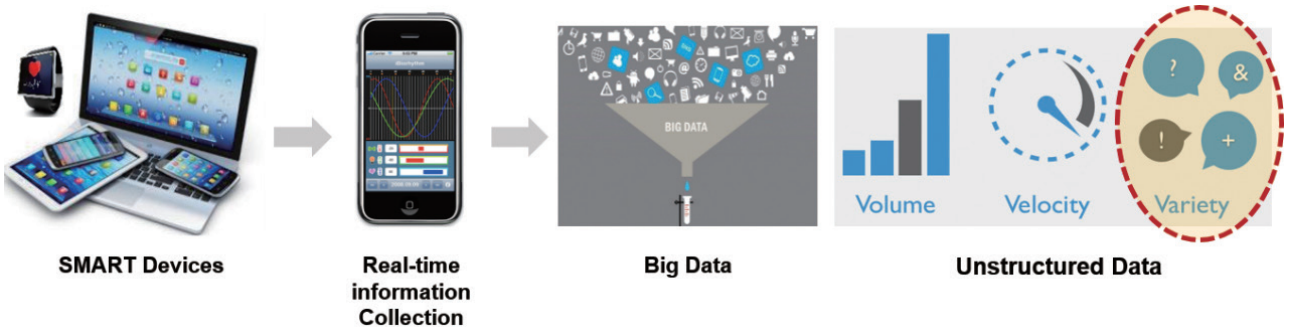

그림 1. 스마트 디바이스의 등장과 빅데이터 기술의 활용

2. 건설 분야의 자연어 처리(NLP) 관련 연구 동향

2.1 자연어 처리(NLP) 개념 및 적용 기술

기업에서 생산되는 데이터의 80% 이상은 비정형 데이터로 이루어져 있으며 (그림 2), 그 중에서도 텍스트 데이터가 차지하는 비중은 매우 높다. 이는 건설 프로젝트의 경우도 마찬가지인데, 건설공사를 수행하면서 활용하는 대다수의 정보는 계약서, 설계변경 보고서, RFI (Request for Information) 등과 같이 텍스트 기반의 비정형 데이터 형태로 이루어져 있다.

텍스트와 같은 비정형 데이터를 분석하기 위해서는 자연어 처리(NLP)를 통해 자연어 형태의 텍스트 데이터를 분석 가능한 구조화된 형태로 변환하게 된다. 자연어 처리(NLP)는 자연언어와 컴퓨터 간의 원활한 상호작용을 위해 기계학습(machine learning)을 통해 컴퓨터가 인간의 언어를 이해할 수 있도록 하는 작업이다(Chopra et al. 2016). 효과적인 자연어 처리(NLP)를 위해서는 단어의 의미 및 문법적 속성, 문법 규칙과 같은 리소스와 어휘집(lexicon), 온톨로지(ontology) 등과 같은 다양한 지식 표현기술을 사용하게 된다 (Kao and Poteet 2007).

자연어 처리(NLP) 기술은 기본적으로 두 가지 목적에 따라 발전해왔는데, 기계에 의한 언어 처리 자동화를 위한 목적과 인간과의 의사소통을 향상시키기 위한 목적이 그것이다(Tiwary and Siddiqui 2008). 즉, 오늘날 발전을 거듭하고 있는 기계 번역(machine translation)의 정확성과 포털 사이트 검색 기술의 발전에는 자연어 처리(NLP) 기술의 도움이 절대적이라고 할 수 있다 [1].

자연어 처리(NLP) 기술은 정보 검색(IR, Information Retrieval), 정보 추출(IE, Information Extraction) 분야에서 활발히 활용되고 있는데, 정보 검색(IR)이 거대한 규모의 정보 저장소에서 필요한 정보(주로 비정형 텍스트 정보)를 찾아내는 작업(Manning et al. 2008)이라면 정보 추출(IE)은 정보 검색(IR)과는 달리 추출하고자 하는 정보를 사전에 정의하여 해당 정보만을 추출하는 과정을 의미한다. 정보의 홍수 시대에 살고 있는 오늘날 정보 검색(IR)은 방대한 양의 정보에서 필요한 정보만을 획득하여 지식화 하는데 매우 유용한 기술이라고 하겠다. 결국, 구글(Google)과 같은 검색 포털 사이트의 발전은 자연어 처리(NLP)기술의 발전과 궤를 같이 한다고 볼 수 있다.

2.2 건설 분야 자연어 비정형 텍스트 데이터 및 자연어 처리 (NLP) 관련 연구 동향

건설 분야에서 자연어 처리(NLP)를 적용한 연구는 크게 통계 분석의 도구로서 텍스트 데이터 분석 방법을 적용한 연구와 활용 시스템으로서 자연어 처리(NLP) 기술을 적용한 연구로 구분할 수 있으며, 활용 시스템으로서 자연어 처리(NLP) 기술을 적용한 연구는 다시 1) 문서 분류 시스템에 자연어 처리(NLP)를 적용한 경우 (Caldas et al. 2002; Caldas and Soibelman 2003; Salama and El-Gohary 2016), 2) 정보 검색 시스템에 자연어 처리(NLP)를 적용한 경우(Fan and Li 2013; Gao et al. 2015; Tixier et al. 2016; Zou et al. 2017), 3) 텍스트 정보 자동 정보 추출을 위해 자연어 처리(NLP)를 적용한 경우(Mohemad et al. 2011; Zhang and El-Gohary 2013; Zhang and El-Gohary 2015; Tixier et al. 2016; Liu and El-Gohary 2017)로 분류할 수 있겠다.

통계 분석의 도구로서 자연어 처리(NLP) 기술을 활용한 연구들의 대부분은 얕은 자연어 처리(shallow NLP)를 통해 문서 내 주요

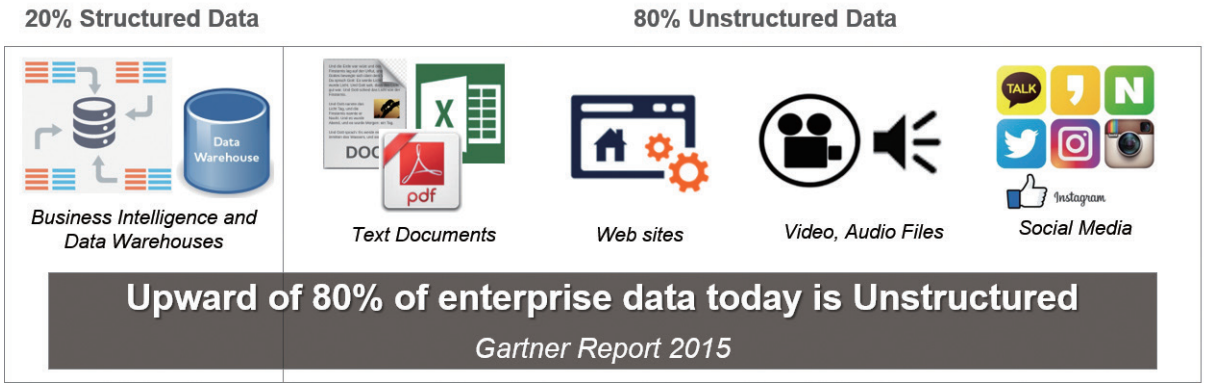

그림 2. 기업 생산 데이터 중 비정형 데이터의 비중

1) IBM에서 개발한 Watson은 2011년 2월 미국의 TV 퀴즈프로그램인 'Jeopardy! 퀴즈쇼'에서 인간 챔피언 2명과 대결해 우승하면서 크게 주목을 받았다. Watson은 고성능 하드웨어(90대의 IBM Power 750 서버)를 통해 대용량의 콘텐츠(약 200억 페이지 분량)를 대상으로 고정밀 자연어 처리 및 분석을 실시하여 구조화된 지식을 구축함으로써 질문에 대한 정답을 제시하는 질의응답시스템(question answering system)이다.

키워드를 중심으로 통계 분석을 실시한 연구가 주를 이룬다. 이와 같은 연구에서는 문장 단위(sentence-level)보다는 단어 단위(word-level)의 자연어 처리(NLP)가 이루어지기 때문에 대부분 형태소 분석에 의한 키워드 도출에 중점을 두고 있다. 따라서 키워드를 중심으로 산업 동향 및 트렌드를 분석하거나 온라인 정보들을 대상으로 한 연구들이 많이 수행되었는데, 데이터 마이닝 기반의 모델 예측 과정에 키워드 중심의 텍스트 정보가 활용되기도 하였다. Williams and Gong(2014)의 연구는 건설공사 입찰문서에 기록된 프로젝트 특성에 대한 정성적 정보(텍스트 정보)와 입찰 결과(입찰자 수, 입찰가격 등)에 대한 정량적 정보를 활용하여 건설 프로젝트의 공사비 증가를 예측하는 모델을 개발하였다. 이 연구는 입찰문서의 텍스트 정보와 숫자 정보를 통합하여 공사비 증가를 예측하는 리스크 모델을 만들었다는 점에서 의의가 있다. 그러나 프로젝트의 특성과 잠재적 위험요인이 내재된 문서들을 종합적으로 분석하지 못한 채 프로젝트 특성만을 짧게 기술한 요약 정보만을 텍스트 분석 대상으로 선정하였다는 점과, 단순 키워드 중심의 텍스트 분석만을 실시하였다는 점에서 분석된 결과의 정확도가 낮을 수밖에 없는 한계를 갖고 있다.

활용 시스템으로서 자연어 처리(NLP) 기술을 적용한 연구는 적용 목적에 따라 문서 분류, 정보 검색, 정보 추출 관련 연구로 구분할 수 있다. 문서 자동 분류를 위해 자연어 처리(NLP) 기술을 활용한 연구 중에서는 Caldas et al.(2002) 및 Caldas and Soibelman(2003)의 연구가 대표적이라고 할 수 있는데, 이 연구에서는 PMIS와 같은 건설 프로젝트 정보 시스템에 저장된 수많은 건설문서들을 자동으로 분류하기 위해 텍스트 마이닝 기반의 기계학습(machine learning)을 실시하여 건설 문서 자동분류 시스템의 프로로타입을 개발하였다. 이 연구에서는 각기 다른 기관에서 만든 다양한 형태의 텍스트 정보들을 어떻게 검색(retrieve)하고, 분류(classify)하고, 통합(integrate)할 것인지에 대한 고민으로부터 시작하여 건설문서의 자동분류 시스템을 개발하였다.

Zou et al.(2017)의 연구에서는 건설사고 관련 데이터베이스에서 유사 사고 사례를 검색하는 CBR(Case-based Reasoning) 기반의 정보검색 모델을 제시하였다〈그림 3〉.이 연구에서는 기존의 정보검색 모델들이 키워드 검색 기반으로 이루어져 있어 정확도가 낮다는 문제를 해결하기 위해 의미론적 쿼리 확장(semantic query expansion) 방법을 적용하여 쿼리와 검색 문서들간의 유사도를 평가하였다. 이 연구에서는 워드넷(WordNet)이라고 불리는 어휘사전 및 도메인 지식이 포함된 어휘집(lexicon)을 기반으로 의미론적 분석(semantic analysis)을 가능케 하였으며, 그 결과 단순 키워드 중심의 검색 방법에 비해 높은 정확도를 얻을 수 있었다.

Zhang and El-Gohary(2013)의 연구에서는 정보 추출(IE) 방법을 토대로 복잡한 건설 법규문서로부터 패턴 매칭을 통한 의미론적 정보 추출을 가능케 하는 연구를 수행하였는데, 문서에서 일부 정보만을 선택적으로 추출하던 과거의 방식에서 한 단계 더 나아가 심층 자연어 처리(NLP)를 통한 문장 의미요소 전체를 분석할 수 있는 성과를 도출하였다〈그림 4〉. 이는 건설 분야에서 비정형 텍스트 정보에 대한 심층 의미 분석이 가능함을 보여준 의미 있는 성과로 볼 수 있다.

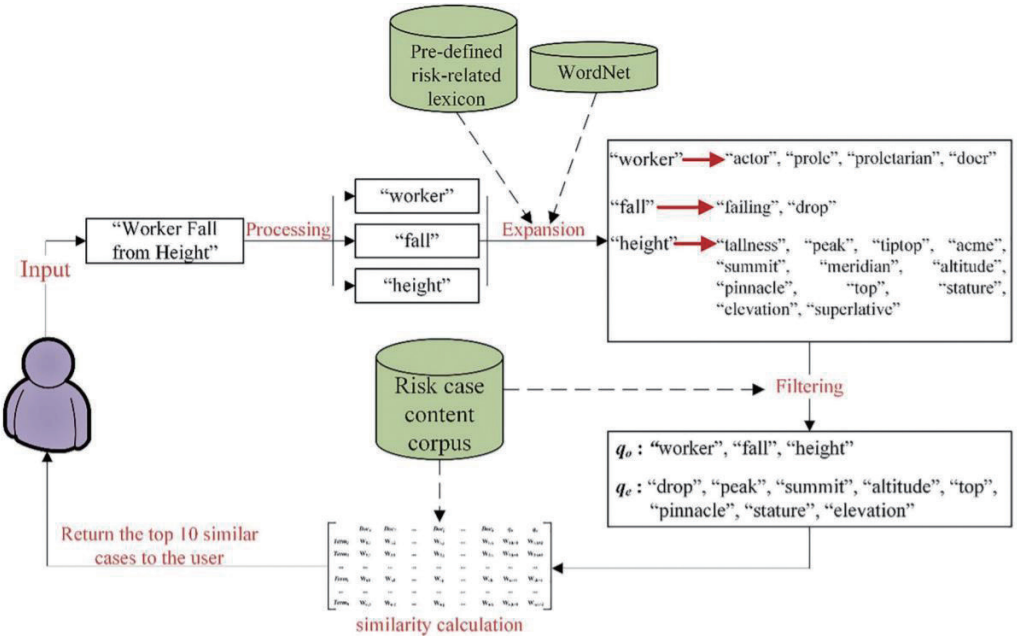

그림 3. 유사 사례 검색을 위해 자연어 처리(NLP)를 적용한 연구 사례(Zou et al. 2017)

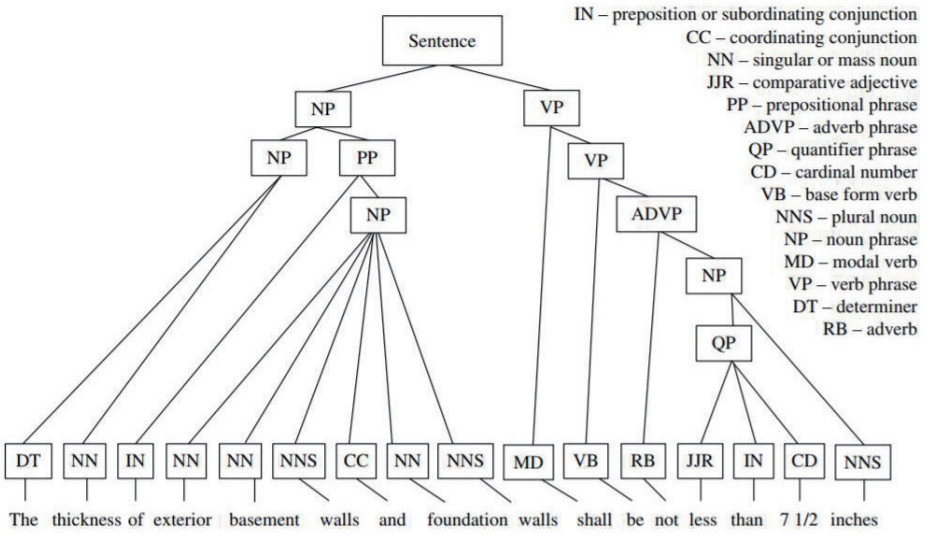

그림 4. 문장 내 단어, 구(phrase) 구조와 문법적 규칙에 의한 텍스트 분석 사례(Zhang and El-Gohary 2013)

이상에서 살펴본 자연어 처리(NLP)의 활용 분야과 건설 산업에서의 관련 연구 동향 고찰을 통해 다음과 같은 몇 가지 시사점을 얻을 수 있었다. 첫째, 건설 분야에서 비정형 텍스트 데이터 분석에 대한 연구가 본격적으로 시작된 것은 일부 연구를 제외하고 대부분 4-5년 이내에 이루어졌다는 것이다. 이는 건설 프로젝트를 통해 생산되는 정보의 대부분이 텍스트 형태로 이루어져있음을 생각할 때 아직까지 활용되지 못하고 사장(死藏)되고 있는 텍스트 정보들이 여전히 많음을 의미함과 동시에, 향후 건설 도메인의 많은 텍스트 데이터를 활용한 연구 분야의 발전 가능성이 무궁무진함을 뜻하는 것이기도 하겠다. 둘째, 건설 텍스트 분석 연구에서 활용되는 데이터의 대부분은 웹 정보, SNS 데이터 등과 같이 온라인에서 실시간으로 생성되고 있는 방대한 양의 빅 데이터(Big Data) 라기보다는 과거에 축적된 경험 데이터, 프로젝트 수행 데이터 등과 같이 상대적으로 적은 규모의 데이터이다. 따라서 방대한 양의 데이터를 통해서 얻을 수 있는 데이터 학습 효과가 상대적으로 낮을 수밖에 없으며, 이는 적은 양의 데이터를 통해 자연어 처리(NLP)를 할 수 있는 적절한 방법의 선택이 무엇보다 중요한 이유이기도 하다.

3. 결언

본 고에서는 건설 산업에서 적용되고 있는 비정형 텍스트 데이터 분석 기반의 자연어 처리(NLP) 관련 연구들에 대해 살펴보았다. 자연어 처리(NLP)와 관련한 연구는 의료, 법률, 문헌 정보 등과 같은 분야에서는 이미 상당한 성과가 이루어진 분야이기도 하다. 반면, 건설 분야의 관련 연구는 아직 상대적으로 미진한 상태라고 할 수 있다. 그러나 한편으로는 그만큼 아직 많은 발전 가능성이 있음을 시사하는 것으로 볼 수도 있겠다. 건설 프로젝트에 존재하는 다양한 유형의 텍스트 정보, 문서들을 토대로 과거의 경험적 정보를 지식화하고, 이러한 정보들이 이미지(image), 오디오(audio), 비디오(video) 등과 같은 또다른 유형의 비정형 데이터들과 통합적으로 분석이 될 수 있다면 건설 프로젝트에 존재하는 다양한 유형의 정보 활용 측면에서 보다 확장적 연구가 가능해질 것으로 기대한다.

참고문헌

1. Caldas, C. H., Soibelman, L., and Han, J. (2002). Automated classification of construction project documents. Journal of Computing in Civil Engineering, 16(4), 234 – 243.
2. Caldas, C. H., and Soibelman, L. (2003). Automating hierarchical document classification for construction management information systems. Automation in Construction, 12(4), 395 – 406.
3. Chopra, D., Joshi, N., and Mathur, I. (2016). Mastering Natural Language Processing with Python. Packt Publishing Ltd., Birmingham, UK.
4. Fan, H., and Li, H. (2013). Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques. Automation in construction, 34, 85–91.
5. Gao, G., Liu, Y. S., Wang, M., Gu, M., and Yong, J. H. (2015). A query expansion method for retrieving online BIM resources based on Industry Foundation Classes. Automation in construction, 56, 14–25.
6. Kao, A., and Poteet, S. R. (Eds.). (2007). Natural language processing and text mining. Springer Science & Business Media.

7. Liu, K., and El-Gohary, N. (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. Automation in Construction. 81, 313-327.
8. Manning, C. D., Raghavan P. and Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press, New York, USA.
9. Mohemad, R., Hamdan, A. R., Othman, Z. A., and Noor, N. M. M. (2011). Ontological-based information extraction of construction tender documents. Proceedings of Advances in Intelligent Web Mastering - 3 Springer, Berlin, Heidelberg, 153-162.
10. Salama, D. M., and El-Gohary, N. M. (2016). Semantic Text Classification for Supporting Automated Compliance Checking in Construction. Journal of Computing in Civil Engineering, 30(1). doi:10.1061/(ASCE)CP.1943-5487.0000301.
11. Tiwary, U. S., and Siddiqui, T. (2008). Natural language processing and information retrieval. Oxford University Press, Inc.. New Delhi, India.
12. Williams, T. P., and Gong, J. (2014). Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. Automation in Construction, 43, 23-29.
13. Zhang, J., and El-Gohary, N. M. (2013). Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. Journal of Computing in Civil Engineering, 30(2), 04015014.
14. Zhang, J., and El-Gohary, N. M. (2015). Automated information transformation for automated regulatory compliance checking in construction. Journal of Computing in Civil Engineering, 29(4), B4015001.
15. Zou, Y., Kiviniemi, A., and Jones, S. W. (2017). Retrieving similar cases for construction project risk management using natural language processing techniques. Automation in Construction, 80, 66-76.