독일언어문학 제73집(2016.9). 147-166

대용어해소를 통한 독일어 텍스트 논조분석 방안

홍문표 (성균관대)

Ⅰ. 들어가는 말

일반적으로 사람들은 상품의 구매를 위한 의사결정이나 어떠한 선택을 하기 전에 다른 사람들의 의견이나 조언을 구하는 경우가 많다. 이 경우 우리는 되도록 많은 다른 사람들의 의견을 들어 보고자 하므로 특정 주제에 대해 다수의 사람들이 의견을 교환하는 인터넷상의 가상공간을 방문하게 된다.

소비자들의 여론동향을 파악하여 기업의 마케팅 및 경영전략을 수립하거나 국민여론의 동향을 파악하여 정책 등을 수립하기 위해서는 인터넷상에서 생성되고 배포되는 정보의 수집 및 분석이 매우 중요하다. 일반적으로 이러한 정보들은 짧은 시간 내에 대규모로 생성되고 배포되므로 몇 명의 모니터링 요원을 통한 정보의 수집 및 분석은 거의 불가능하다. 따라서 이를 위한 자동화 기술의 개발이 필요로 되고 최근 자연언어처리 분야에서 활발히 연구되고 있는데 이 분야를 전산언어학 분야에서는 논조분석 Sentimentanalyse이라 부른다.

어떤 텍스트를 통해 전달되는 저자의 주관적인 감정이나 생각을 컴퓨터를 이용해 자동으로 분석하는 것은 자연언어처리 Natural Language Processing 분야에서 최고 난이도의 작업으로 볼 수 있다. 이를 위해서는 일반적으로 텍스트를 구성하는 입력문장의 형태소 분석 morphologische Analyse, 구문관계의 분석을 위한 구조분석 syntaktische Analyse, 문장의 명제적 의미 파악을 위한 의미분석 semantische Analyse 뿐만 아니라 문장과 문장 간의 관계에 대한 담화분석 Diskursanalyse도 수행되어야 한다.

텍스트를 논조에 따라 긍정 positiv과 부정 negativ이라는 두 개의 극성 Polarität로 분류하기 위한 논조분석 분야는 텍스트를 주제에 따라 분류하는 텍스트 마이닝 분야의 기법들을 적용하면서 그 연구가 시작되었다. 논조분석의 초창기 연구들은 Pang et al. (2002)의 연구에서처럼 긍정 극성과 부정 극성 자질 Merkmal의 출현여

부 Präsenz 또는 출현빈도 Frequenz를 활용한 기계학습 maschinelles Lernen 기반의 방법론들이었다. 여기서 말하는 기계학습을 위한 자질이란 기계학습의 대상이 되는 학습데이터에 등장하는 단어를 포함한 모든 토큰 Token들이다.

학습데이터 Lerndaten에 등장하는 토큰을 알아내기 위해서는 형태소 분석만으로도 충분하다. 그러나 자질의 출현 및 빈도만으로는 텍스트의 논조를 정확하게 파악하는 것이 매우 어렵다. 왜냐하면 긍정이나 부정을 나타내는 텍스트의 토큰들, 예를 들어 단어들은 부정어를 통한 수식관계 (‘nicht gut’) 나 비현실 사태 (‘wäre gut’)를 나타내는 화법조동사 등의 사용 등을 통해 논조가 바뀔 수도 있기 때문이다. 즉, 이러한 관계를 파악하기 위해서는 형태소 분석 뿐만 아니라 입력문장의 구조정보를 반영하기 위한 구조 분석까지 수행되어야 한다.

Joshi & Pennstein-Rose (2009)와 홍문표 (2013)의 연구에서는 각각 영어와 독일어 텍스트의 논조분석을 위해 입력문장의 의존구조 Dependenzstruktur를 기계학습에 반영하는 방법론을 제안하였다. 이 연구에서는 기계학습을 위한 학습자질로서 학습코퍼스의 토큰 뿐만 아니라 문장의 의존구조도 활용하여 문장구조에 따른 극성의 변화현상 등을 처리하였다.

논조분석이 기술적으로 매우 어려운 이유 중의 하나는 문장의 명제적인 의미 뿐만 아니라 문장 속에 명시적으로 드러나지 않는 언표내적 illokutionär 의미까지도 알아내야 하기 때문이다. 화자들은 대상에 대한 직접적인 비판 등으로 인한 불편함 등을 모면하기 위해서나 청자의 체면 등을 살려주기 위해 반어적 표현 ironische Ausdrücke를 사용하는 경우가 많다(Vgl. Schwarz-Friesel 2009). 어떠한 문장이 반어적 의미를 전달하고 있는지를 파악하기 위해서는 의미분석이 수행되어야 한다. 홍문표 (2016)의 연구에서는 언표적 의미를 그대로 전달하는 일반의미 문장과 언표내적 의미를 전달하는 반어의미 문장의 분류를 위한 의미분석 방법론을 제안하였다.

지금까지 간략하게 살펴 본 것처럼 논조분석을 위해서는 형태소 분석, 구조 분석, 의미 분석의 단계가 필요하다. 본 연구에서는 분석의 더 깊은 단계에까지 들어가 담화레벨 Diskursebene의 분석을 통해 논조분석의 정확률을 높일 수 있는 방법을 다루고자 한다. 담화분석 Diskursanalyse이라는 분야에는 또 다시 많은 세부 분야들이 존재하지만 본 연구에서는 문장에 등장하는 대명사가 지시하는 선행사를

찾아내는 대용어해소[1] Anaphernresolution 문제에 집중하고자 한다.

어떠한 제품에 대한 전문가 리뷰 review나 사용자들의 사용후기 등에서는 하나의 텍스트 안에서도 빈번히 주제 혹은 토픽이 바뀌게 된다. 예를 들어, 갤럭시 노트7에 대한 리뷰 안에서도 단락의 주제가 갤럭시 노트7이 아니라 갤럭시 S7 엣지 또는 아이폰 6 Plus가 될 수도 있다. 왜냐하면 저자들은 (1)의 예문에서처럼 갤럭시 노트7의 장단점을 이야기하기 위해 관련제품인 갤럭시 S7 엣지나 경쟁제품인 아이폰 6 Plus를 자주 언급하기 때문이다.[2]

(1) Im Grunde verhält sich der Übergang vom Note 5 zum Galaxy Note 7 wie der Sprung von der S6-Reihe auf die S7-Reihe. Es gibt keine große Revolution beim Note 7, aber die kleinen Detailverbesserungen und ¬optimierungen sind in Summe aus meiner Sicht besser als die große technische und optische Revolution. Trotz nur einer begrenzten Zeit mit dem Note 7 sieht es danach aus, dass ich schon bald einen neuen Artikel mit dem Titel „Für dieses Smartphone würde ich mein iPhone 6 Plus aufgeben“ schreiben müsste, nur das nun statt eines Galaxy S7 Edge ein Note 7 den Platz meines iPhone 6 Plus einnimmt. Ob das auch nach dem ausführlichen Test hier auf AndroidPIT.de Bestand hat, muss sich noch zeigen…

(1)의 텍스트에서 밑줄 친 명사구 ‘dieses Smartphone’이 가리키는 선행사가 무엇인지에 대한 고려가 없이는 저자의 긍정적인 생각이 무엇에 대한 것인지 정확하게 파악할 수 없다. 대명사를 포함한 대용어의 선행사를 찾기 위해서는 텍스트 전체 구조에 대한 이해가 필요하다. 그러나 현재로서는 텍스트에 대한 컴퓨터 처리시 텍스트의 전체 구조를 분석하고 파악한다는 것은 거의 불가능한 일이다. 따라서 본 연구에서는 기계학습을 이용한 대명사의 선행사 복원 방법론을 제안하며, 이를 통해 논조분석의 정확률이 향상될 수 있음을 보일 것이다.

---

1) ‘대용어해소’라는 용어는 영미권 문헌에 사용되는 ‘anaphora resolution’을 그대로 번역한 용어로서, 한국어 처리분야의 학술담론에서도 그대로 사용되고 있다. 개인적인 의견으로는 ‘resolution’을 ‘해소’라고 번역하는 것은 적절하지 못한 것으로 생각되나, 그럼에도 불구하고 본 연구에서는 일반적으로 통용되는 ‘해소’ 또는 ‘복원’이라는 용어를 사용하기로 한다.

2) (1)의 예문은 https://www.androidpit.de/samsung-galaxy-note-7-test에서 발췌하였음.

이를 위해 본 논문은 다음과 같이 구성되어 있다. 2장에서는 기존의 논조분석에 대한 연구동향을 정리하며, 대용어의 선행사 복원이 논조분석에서 갖는 중요성에 대해 다룰 것이다. 3장에서는 대용어 해소를 위한 기존의 방법론들을 소개한다. 4장에서는 본 연구에서 제안하는 독일어 대용어 해소 방법론을 소개하고, 이를 기존의 논조분석 방법론에 적용하는 방안에 대해서도 논의한다. 또한 이 방법론이 논조분석의 정확률 향상에 미치는 영향을 계량적으로 측정하는 실험에 대해서도 논의한다. 마지막으로 5장에서는 본 연구의 내용을 정리하며 향후 연구방향을 제시한다.

Ⅱ. 논조분석과 대용어해소

Ⅱ.1. 논조분석의 레벨

논조분석은 자연언어 텍스트와 같은 비정형 데이터 unstrukturierte Daten로부터 정보를 추출하는 작업이다. 텍스트의 논조를 성공적으로 분석하기 위해서는 형태소 분석과 구조 분석 이외에도 I장에서 언급한 바와 같이 의미분석, 담화분석 등의 과정이 필요하다. 이 중에서도 특히 자연언어처리 분야에서 해결되지 못한 많은 문제를 갖고 있는 공지시관계해소 Koreferenzresolution, 부정어처리 Behandlung der Negation, 대용어해소 Anaphernresolution, 의미모호성해소 Disambiguierung des Wortsinnes의 처리가 매우 중요하다(Vgl. Cambria et al. 2013).

이 중 텍스트의 담화구조를 반영하기 위한 시도는 논조분석 연구의 초창기부터 있었다. 그러나 그 시도는 매우 단순하고 비직관적이기 때문에 진정한 의미의 담화구조의 반영이라고 말하기는 어렵다. Pang et al. (2002)의 연구에서는 리뷰나 블로그와 같은 텍스트 형태들의 경우 저자의 의견을 강하게 나타내는 문장이 텍스트의 가장 앞부분이나 가장 마지막 부분에 나타난다는 경향성에 주목하여, 텍스트 내에서의 문장의 위치를 기계학습을 위한 자질로 하여 가중치를 부여하였다. 이를 통해 텍스트의 담화구조가 간접적으로 반영될 수는 있으나 모든 텍스트가 그러한 경향성을 보이는 것은 아니며, 이 방법론은 대용어해소 등의 문제를 해결하지 못

하는 단점을 가지고 있다.

논조분석에 대한 연구는 문서레벨 Dokumentebene의 논조분석에서 문장레벨 Satzebene의 논조분석을 거쳐 구레벨 Phrasenebene 및 양상레벨[3] Aspektebene의 논조분석에까지 이르고 있다. 어떠한 문서가 특정 주제에 대해 긍정적인 의견을 전달하느냐 부정적인 의견을 전달하느냐를 분류하는 문서레벨의 논조분석은 Pang et al. (2002)의 연구에서 처음으로 제안되었으며, 기존 텍스트마이닝 Textmining 분야의 주제분류 Topikklassifizierung 기법을 그대로 적용하였다.

이 방법론에서는 문서단위에서 '긍정'과 '부정'의 레이블 label이 부착된 학습데이터를 활용해 문서의 논조를 분석하는 분류기 Klassifizierer를 개발하는 것을 목표로 한다. 학습데이터에 포함된 각 문서로부터 학습을 위한 자질들이 추출되고, 이 자질 (주로 토큰)의 출현여부 또는 출현빈도를 값으로 하여 지지기반벡터머신 Support Vector Machine과 같은 기계학습 알고리즘을 적용하여 분류기가 만들어진다.

그러나 논조분석의 주요 목적이 사용자의 의견 및 여론 분석 등을 통한 마케팅/경영 전략 수립 등이라는 점에서 문서레벨의 논조분석은 그 활용도가 다소 떨어진다. 왜냐하면 하나의 문서에서도 특정 주제에 대해 다양한 논조가 드러날 수 있기 때문이다. 예를 들어, 갤럭시 노트7에 대한 문서 내에 나타나는 모든 문장이 항상 긍정 혹은 부정 하나만의 극성을 갖지는 않는다. 활용의 측면에서는 어떤 문서 내에서 모든 문장들을 객관적 의미의 문장과 주관적 의미의 문장으로 분류한 후, 주관적 의미의 문장을 긍정논조의 문장 및 부정논조의 문장으로 분류하는 것이 훨씬 더 활용도가 높을 수 있다.

Liu (2012)의 연구는 논조분석의 단위를 문서레벨에서 문장레벨로 전환한 대표적인 연구이다. 문장레벨의 논조분석은 기계학습데이터의 구성을 문장에 대한 토큰벡터 Token vector로 구성한다는 차이가 있을 뿐 방법론상으로는 문서레벨 논조분석과 큰 차이가 없다. 그러나 이와 같이 문장레벨에서의 논조를 분석하면서 문장의 구조를 고려하지 않는 단어뭉치 bag of words 방식의 접근법은 부정어구의 수

3) 여기서의 '양상'은 'Aktionsarten'과 유사한 개념으로서의 '양상'이 아니라 대상의 여러 특징 또는 자질을 의미한다. 예를 들어 스마트폰의 논조분석을 위한 양상으로는 '통화음질', '배터리', '카메라', '디자인', '편의성' 등을 들 수 있다.

식, 비현실 사태 등으로 인해 정확한 분석에 한계가 있다.

이러한 문제를 해결하기 위해 등장한 방법론들이 바이그램 bigram을 활용하거나 문장의 구조정보를 기계학습에서 직접 자질로 활용하는 방법들이다. 홍문표(2014)에서는 바이그램을 기계학습의 자질로 활용할 경우 문장구조를 반영하는 효과를 볼 수 있음을 실험을 통해 입증하였다. Joshi & Penstein-Rose (2009)와 홍문표(2013)의 연구에서는 학습데이터를 구성할 때 구조분석을 수행한 후 파악된 의존관계를 기계학습을 위한 자질로 활용하여 문장의 구조에 따른 논조의 변화를 포착하였다.

기계학습을 활용한 논조분석 방법은 기계학습의 대상이 되는 학습코퍼스의 성격에 따라 성능이 크게 좌우된다. 예를 들어 스마트폰 분야의 코퍼스로 학습된 논조분석시스템은 정치사회 분야 텍스트의 논조를 분석할 경우 그 정확률이 하락할 가능성이 높다. 마찬가지로 스마트폰 분야의 코퍼스로 학습된 논조분석시스템을 여행분야의 텍스트에 적용할 경우에도 정확률이 하락하게 된다. 이러한 문제를 해결하기 위해서는 다양한 분야별로 학습코퍼스를 구축해야 하는데, 일반적으로 학습코퍼스의 구축을 위해서는 수작업의 단계가 필요하다는 점에서 이는 매우 큰 시간과 비용을 필요로 하는 비현실적인 해결방법이다.

이를 위해 다양한 분야의 문장레벨 논조분석을 위한 선극성사전 prior-polarity dictionary이 고안되었다.[4] 선극성사전은 문맥의 영향이 없는 중립적인 상황에서 어떤 단어에 대해 일반 화자들이 느끼는 공시적인 의미 Konnotation가 기술된 사전이다. 예를 들어 독일어의 경우, ‘Unfall’, ‘blöd’, ‘kaputt’ 등은 부정적인 공시의 단어들이며, ‘schön’, ‘Meisterstück’, ‘sauber’ 등은 긍정적인 공시의 단어들이다. 이러한 단어들은 사용되는 분야, 예를 들어 정보통신, 경제, 스포츠, 정치 등에 상관없이 항상 동일한 공시의를 가지고 있으므로, 이 단어들을 기계학습을 위한 자질로 사용하게 되면 주제분야에 상관없이 비교적 고른 분석성능을 보이는 것으로 알려져 있다(Vgl. 홍문표 2015).

문장레벨의 논조분석도 어떠한 대상 및 그것이 가지고 있는 각종 양상 Aspekt에 대한 저자의 논조를 분류하는데 어려움이 있다. 예를 들어 스마트폰이라는 대상은

4) 대표적인 독일어 선극성사전으로는 Waltinger (2010)의 ‘German Polarity Clues’, Remus et al. (2010)의 ‘SentiWS’ 등이 있다.

‘가격’, ‘디자인’, ‘통화품질’, ‘배터리’, ‘카메라’, ‘무게’ 등과 같은 다양한 양상을 가지고 있으며, 사람들은 일반적으로 어떠한 대상에 대해 좋고 싫음을 얘기할 때 막연하게 호감/비호감을 얘기하는 것이 아니라 특정한 양상에 대한 호감/비호감을 얘기하는 경우가 더 많다. 즉, 예를 들어 ‘갤럭시노트 7은 디자인은 좋은데 가격이 좀 비싸다’라는 문장은 두 개의 양상 (‘디자인’, ‘가격’)에 대해 언급하고 있으며 각 양상에 대한 극성값 또한 다르다 (‘호감’, ‘비호감’).

(2) Der Akku des iPhone 6 scheint etwas länger durchzuhalten als es beim iPhone 5S der Fall war.
(3) Das iPhone 6 Plus bietet hier dank seinem riesigen Bildschirm noch jede Menge Potential.

위의 독일어 예문 (2)와 (3)은 모두 긍정적인 논조를 전달한다. 그러나 위 글의 저자는 주제에 대한 일반적인 긍정의 의견을 표출하는 것이 아니라 특정한 양상에 대한 자신의 긍정적인 의견을 피력하고 있다. (2)의 예문에서는 배터리에 대한 긍정의견이 전달되고 있으며, (3)의 예문에서는 화면크기에 대한 긍정의견이 나타난다.

이렇듯 사용자들은 대상이 가지고 있는 여러 가지 양상 Aspekt에 대해 의견을 나타낼 수 있는데, 이것을 정확하게 분석하기 위해서는 문장의 레벨보다 더 세부적으로 구레벨 Phrasenebene 또는 양상레벨 Aspektebene의 분석이 수행되어야 한다. Wilson et al. (2005)은 논조분석의 대상을 문장레벨에서 구레벨로 전환한 대표적인 연구이다. 이 연구에서는 단어와 극성에 관한 자질 10개를 사용한 기계학습 기반의 구레벨 논조분석시스템을 소개하였다.

Pontiki et al. (2014)의 연구는 양상레벨의 논조분석에 대한 대표적인 연구 중의 하나이다. 이 연구에서는 단순히 문장의 논조를 긍정, 부정으로 나누는 것이 아니라 우선 문장을 통해 기술되는 양상을 파악한 후, 해당 양상에 대한 극성값을 계산하는 방법론을 소개한다. 이를 위해 양상을 나타내는 단어들을 탐지한 후, 각 단어들이 어떠한 양상범주에 속하는지 계산하고 나서 각 양상별 논조를 계산한다. 이 결과 사용자들은 그림 1에서 보는 것과 같이 어떤 제품의 다양한 양상별로 보다 정확한 정보를 얻을 수 있게 된다.

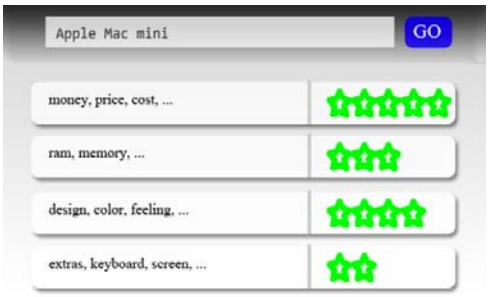

그림 1 : 양상레벨 논조분석의 예 (Pontiki et al. (2014)에서 발췌)

지금까지 우리는 이 장에서 논조분석의 대상이 되는 레벨들에 대해 살펴보았다. 논조분석이 실질적으로 산업계에서 보다 더 유용하게 활용되기 위해서는 최소한 문장레벨의 논조분석 또는 더 나아가 양상레벨의 논조분석이 필요함을 보았다. 다음 장에서는 대용어해소가 문장레벨의 논조분석 및 양상레벨의 논조분석에 있어서 매우 중요함을 보일 것이다.

II.2. 대용어해소를 고려한 논조분석

(3) Die Kamera$_i$ des iPhone 6 hat mich ebenfalls überrascht. Sie$_i$ ist unglaublich gut.

(4) Das iPhone 6$_i$ ist für mich das perfekte Smartphone. Es$_i$ leistet sich nirgends Schwächen und ist eigentlich überall ganz stark unterwegs.

(5) Durch den eingebauten optischen Bildstabilisator verfügt das iPhone 6 Plus hier noch über mehr Reserven als das kleinere iPhone 6$_i$. Aber auch dessen$_i$ Kamera macht klasse Fotos.

문장레벨과 양상레벨에서의 논조분석에서 대명사가 지시하는 선행사를 찾아내

는 것은 (3)~(5)의 예문에서 보는 것과 같이 매우 중요하다. (3)의 경우, 첫 번째 문장만으로는 컴퓨터가 아니라 사람조차도 아이폰6의 카메라 기능에 대한 글쓴이의 논조를 정확하게 파악하기 어렵다. 왜냐하면 'überraschen'은 긍정적인 의미로도 사용될 수 있지만 부정적인 의미로도 사용될 수 있기 때문이다. 그러나 이어지는 문장에서 사용된 대명사 'sie'가 지시하는 선행사가 'die Kamera'임을 파악하게 되면, 두 문장 모두 아이폰의 사진기능에 대해 긍정적인 의견을 전달하는 것임을 알 수 있게 된다.

이는 (5)의 예문에서처럼 주격 대명사가 아닌 소유격 대명사('dessen')이 사용된 경우에도 마찬가지이다. 이 문장에서 소유격 대명사('dessen')이 지시하는 선행사는 아이폰6로서, 글쓴이는 비교대상인 아이폰6 플러스와 함께 아이폰6의 카메라 기능에도 좋은 점수를 주고 있음을 알 수 있다.

다음 예문 (6)은 논조분석에서 대명사의 선행사 복원만이 중요한 것이 아님을 보여준다. Nikolov et al. (2008)의 연구에서 발췌한 예문 (6)을 보면, 첫 번째 문장만으로는 이 짧은 텍스트의 토픽인 'Zune'에 대한 논조가 긍정인지 부정인지 알기가 어렵다. 만약 이 문장에 대해 전통적 방식인 단어뭉치 기반의 기계학습 알고리즘을 적용한다면 부정논조의 단어인 'flaw' 때문에 이 문장은 부정적 논조의 문장으로 분류될 가능성이 높다. 그러나 두 번째 문장의 한정명사구 definite NP인 'the upgraded player'가 'Zune'과 공지시 Koreferenz 관계를 갖기 때문에 첫 번째 문장을 포함한 두 문장 모두 긍정논조의 문장임을 알 수 있게 된다.

(6) Microsoft retools Zune$_i$ to target Apple's flaws. The upgraded player$_i$ and a new strategy helps Redmond gain ground in its battel to beat the iPod

Nikolov et al. (2008)에 따르면 대용어해소를 포함한 공지시관계의 복원을 통해 논조분석의 정확률이 약 10% 정도 향상될 수 있는 것으로 알려져 있으며, Hurst & Nigam (2003)도 비슷한 연구결과를 보이고 있다.

Ⅲ. 기계학습기반 대용어해소

Ⅲ.1. 대용어해소를 위한 언어학적 제약

대용어의 선행사를 찾아내기 위해서는 많은 언어학적 지식이 필요하다. 여기서 말하는 언어학적 지식에는 형태론적 지식, 통사론적 지식, 의미론적 지식, 화용론적 지식, 세상지식 등이 포함된다. 형태론적 지식에는 단어의 성 Genus, 수 Numerus, 격 Kasus, 동사의 굴절 Flexion 등에 대한 지식이 포함된다. 명사의 문법적 성이 존재하지 않는 영어와는 달리 독일어의 대용어 해소에서는 형태론적 지식이 매우 중요한 요소로 작용한다.

(7) Das iPhone 5S$_i$ kommt einem sofort klein vor, gleichzeitig aber auch recht dick. Es$_i$ ist wie so oft bei Apple wenn diese neue iPhones herausbringt.

(7)의 예문에서 대명사 'es'의 선행사 후보 중 고유명사 'das iPhone 5S'는 대명사와 성과 수 정보가 일치하므로 대명사의 선행사로 복원될 수 있다.

통사론적 지식 중 대용어의 해소를 위해 가장 유용하게 사용될 수 있는 정보는 술어의 논항선택에 관한 선택제약 selektionelle Restriktion이다. 선택제약은 술어가 논항들을 하위범주 Subkategorisierung하는 경우 논항에게 요구하는 문법적 형태, 격, 의미 등에 대한 제약을 말한다.

(8) In dem Bus gab es nur den Fahrer$_i$. Er$_i$ begrüßte mich.

(8)의 예문에서 대명사 'er'의 선행사 후보로는 'Bus'와 'Fahrer'가 있다. 이 중 'Fahrer'만이 동사 'begrüßen'의 주어위치에 대한 의미제약을 만족시키므로 대명사의 선행사로 복원된다.

홍문표(2011)의 연구에서는 센터링이론 Centering Theory (Vgl. Grosz & Sidner 1986)를 활용한 대명사 복원 방법론을 제안하였다. 대명사 복원의 대상이 되는 선행 명사들은 모두 센터 center라고 보고 현가성이 가장 높은 센터를 선호되는 센터

preferred center라고 부른다. 일반적으로 현가성이 가장 높은 센터가 이어지는 담화 내에서 대명사로 나타나는 경우가 많다고 보았다.

이 연구에서는 담화구조를 반영하는 센터링이론에서 대명사의 선행사가 될 수 있는 선행사 후보들이 문법기능 Grammatische Funktionen에 따라 현가성 Saliency가 달라지는 현상에 주목하여, 선행사 후보가 여럿일 때 가장 현가성이 높은 주어, 목적어 순으로 선행사를 결정하는 방법론을 제안하였다.

이상에서 제시한 언어학적 지식에 기반한 대용어해소 방법론의 단점은 알고리즘이 결정론적 deterministisch이라는 점이다. 즉, 선행사 후보 중 어떠한 하나의 후보가 언어학적 지식에 기반한 제약들 constraint을 위반하지 않고 만족시킨다면 무조건 그 후보가 대용어의 선행사로 결정된다는 점이다. 이러한 접근법의 문제점은 다수개의 선행사 후보가 언어학적 제약들을 위반하지 않고 만족시키는 경우 그 중 어떠한 후보를 선행사로 취할 것인가에 대한 선택방법이 없다는 것이다.

Ⅲ.2. 기계학습과 대용어해소

박아름 & 홍문표 (2015:a)의 연구에서는 영형대명사 Null-Subjekt를 포함한 한국어 주어 대명사의 선행사 복원 문제를 다루었다. 이 연구에서는 그 동안 다양한 자연언어처리 응용 분야에서 분류 Klassifizierung 문제를 해결하는데 성공적으로 사용되어 온 기계학습 방법을 사용했다. 기계학습 방법은 비결정론적 nicht-deterministisch인 특징을 가지고 있다. 즉, 여러 가지의 제약을 만족하는 다수 개의 후보가 존재하더라도 통계적인 계산을 통해 상황에 가장 적합한 단일 후보를 선택할 수 있다는 장점이 있다.

이를 위해 이 연구에서는 기계 학습을 위한 총 12개의 자질을 제안하였다. 이 12개의 자질은 문장 내 등장하는 동사의 형태론적 정보를 반영하고, 주관 그리고 객관 형용사의 등장 유무에 관한 것이었다. 이 연구에서는 또한 센터링 이론을 반영하여 담화 내의 정보와 관련된 자질도 제안하였다. 위 방법론의 성능평가에서는 89.3%의 정확도가 측정되었는데, 이는 홍문표 (2011)에서 한국어 대화체에 등장하는 영형 주어를 처리하기 위해 결정론적 방식의 언어학적 제약만을 적용한 방법론이 73.29%의 정확도를 보였기 때문에, 이와 비교하여 약 16%가 향상된 결과였다.

박아름 & 홍문표 (2015:b)의 연구에서는 앞서 소개한 연구와 비슷한 맥락에서 한국어 영형목적어 Null-Objekt를 포함한 목적격 대명사의 선행사를 복원하는 문제를 다루었다. 목적어 복원을 위해서는 총 8개의 기계학습 자질을 사용하였으며 73.37%의 정확도가 측정되었다. 주어 대명사의 경우보다는 16% 정도 정확도가 낮지만, 비교의 대상이 되는 순수언어학적 제약 기반의 방법론과 비교하여 무려 61%의 정확도 향상 효과가 있음을 보였다.

Park & Hong (2014)의 연구에서는 스페인어 영형 주어대명사의 해소문제를 다루었다. 여기서도 기계학습 기반의 방법론을 제안하였는데, 스페인어 동사의 굴절형태에 관한 정보 및 총 11개의 자질을 사용하였다. 이를 활용한 성능평가 결과 83.6%의 선행사 복원 정확도가 측정되었다. 또한 스페인어 영형 주어대명사 해소를 위해 가장 효과적인 자질을 알아내는 실험을 수행하였고, 그 결과 2개의 자질(선행사 후보의 성, 동사의 인칭)이 빈도수가 높은 주어 유형을 분류하기 위해 중요한 역할을 수행한다는 사실을 알아냈다. 이는 독일어 대명사의 선행사 복원에도 큰 영향을 미칠 것으로 예상된다.

Ⅳ. 제안하는 방법론

본 연구에서는 독일어의 대용어 중 명백한 선행사를 지니는 주격 nominativ 대명사의 선행사 복원문제에만 집중하도록 한다. (5)의 예문에서처럼 주격 대명사 이외에도 소유격 genitiv이나 목적격 akkusativ 대명사의 선행사 복원도 매우 중요한 문제이지만, 본 연구에서는 우선 주격 대명사의 선행사 복원만을 다루기로 한다. 여기에 포함된 대명사는 ‘er’, ‘sie’, ‘es’, ‘dieser’, ‘diese’, ‘dieses’이다.

또한 아래의 예문 (9), (10)과 같이 주격 대명사라 할지라도 허사 expletive ‘es’ 등의 용법으로 사용된 주격 대명사는 논의에서 제외한다.

(9) Der Blickwinkel von welchem man auf die Screens schauen kann ohne das sich die Farben oder Kontrast verändern wurde noch einmal vergrössert. Es ist so zum Beispiel mit dem iPhone 6 Plus ohne weiteres möglich zu zweit einen Film oder Bilder

anzuschauen.

(10) Dadurch gibt es keine Ränder unten und oben wie das beim Wechsel vom iPhone 4S auf das iPhone 5 am Anfang bei vielen Apps der Fall war.

앞 장에서 언급한 바와 같이 순수언어학적 제약기반의 결정론적 방법론은 다수의 선행사 후보가 제약들을 만족할 때 선택의 문제를 해결하기 어렵기 때문에 본 연구에서는 Park & Hong (2014) 등의 연구에서와 마찬가지로 기계학습 기반의 방법론을 제안한다.

다만 Park & Hong (2014)에서 스페인어 대명사의 해소를 위해 제안한 11개의 자질 중에는 스페인어의 고유 특성에 인한 것들도 존재하고, 실제 평가결과 많은 자질들이 대용어 해소를 위해 큰 역할을 못하는 것으로 밝혀졌기 때문에, 본 연구에서는 단 6 개만의 자질을 가지고 대명사의 선행사를 찾아내는 방법론을 고안하였다. 본 연구에서 사용한 6 개의 자질 및 자질의 값은 다음과 같다.

- 선행사 후보의 성 : {masc, neut, fem}
- 선행사 후보의 수 : {sg, pl}
- 선행사 후보의 의미부류 : {hum, thing, abs}
- 선행사 후보의 문법기능 : {subj, obj, obl}
- 선행사 후보와 대명사간의 거리 : {1, 2, 3, 4}
- 대명사의 선행사 여부 : {yes, no}

선행사 후보의 성과 수에 관한 자질은 형태론적 지식을 반영한 기계학습 자질이다. 우선적으로 선행사 후보의 성과 수가 대명사와 일치하는지의 여부는 선행사 후보 중 형태론적으로 적합하지 않은 후보들을 제외하는 역할을 할 수 있다. 선행사 후보의 의미부류에 관한 자질은 의미정보의 반영을 위함이다. 여기서는 선행사 후보의 의미를 간단하게 '사람 (hum)', '사물 (thing)', '추상 (abs)'으로 분류하였다.

센터링 이론에 따르면 선행사 후보들의 문법기능에 따라 현가성이 결정되고 일반적으로 현가성이 가장 높은 후보가 대명사와 호응하는 경우가 많다. 이러한 현

상을 반영하기 위해 선행사 후보의 문법기능을 자질로 고려하였으며, 그 값으로는 주어 (‘subj’), 목적어 (‘obj’), 그 외의 문법기능 (‘obl’)이 올 수 있다.

선행사 후보와 대명사 간의 거리 또한 중요한 역할을 할 수 있다. 본 연구에서는 주격대명사가 사용된 경우 선행사를 탐색하는 범위를 선행 세 문장으로 제한하였다. 세 문장 내에 등장하는 모든 명사구에 대해 대명사에서 가장 가까운 명사구부터 가장 먼 명사구까지 순서대로 1에서 4까지의 값을 부여하였다. 즉, 예를 들어 2의 의미는 대명사로부터 2번째로 가까운 명사구라는 의미이며, 만약 명사구가 4개 이상 존재할 경우, 4번째 이후의 모든 명사구들의 값은 4로 고정하였다. 마지막으로 선행사 후보 명사구가 실제로 대명사의 선행사인 경우 ‘yes’값을, 그렇지 않은 경우 ‘no’값을 갖도록 하여 기계학습에 활용하였다.

기계학습을 위한 학습데이터는 독일어권 네티즌들의 블로그에서 발췌하였다. 주격 대명사가 등장하는 300개의 IT 분야 문장으로 학습데이터를 구성하였으며 각 자질에 대한 값은 저자가 수작업을 통해 부착하였다. 따라서 학습데이터를 만드는데 컴퓨터를 통한 자동처리가 들어가지 않았으므로 각 자질에 대한 값은 매우 정확할 것으로 볼 수 있다.

학습데이터에 대해서는 지지기반 벡터머신 Support Vector Machine 기계학습알고리즘을 적용하여 대용어해소 시스템을 구현하였다. 본 연구에서 제안하는 방법론의 성능을 비교 평가하기 위한 벤치마킹 시스템으로는 언어학적 제약을 활용한 방법론을 택하였다. 언어학적 제약기반 방법론은 선행사 후보의 형태론적 정보와 위치 정보만을 고려하여 대명사의 선행사를 복원하는 방법을 취한다. 즉, 대명사로부터 가장 가까이 위치하면서 대명사와 성과 수 정보가 일치하는 후보명사를 선행사로 보았다.

웨카 WEKA 3.6.9 프로그램을 활용하여 수행한 5 분할균등 상호교차평가 5-fold cross validation에서 기계학습 기반의 시스템은 90.38%의 정확도 Accuracy를 나타내었다. 벤치마킹 대상인 언어학적 제약기반 시스템은 42%의 정확도를 나타내어, 기계학습 기반의 시스템이 무려 2 배가 넘는 정확도의 향상을 가져옴을 볼 수 있었다.

기계학습을 위해 선택한 6개의 자질이 적절하게 선택되었는지를 판단하기 위해 웨카 프로그램의 자질 평가 기능을 활용하였다. ‘InfoGainAttributeEval’ 알고리즘을 사용하여 평가한 결과, 다음 자질들의 순서대로 대용어해소에 가장 큰 영향을

미치는 것으로 조사되었다.

1. 선행사 후보의 성
2. 선행사 후보의 수
3. 선행사 후보와 대명사간의 거리
4. 선행사 후보의 문법기능
5. 선행사 후보의 의미부류

실험 결과 역시 예상대로 선행사의 형태론적 정보가 대용어 해소에 가장 큰 역할을 하고, 선행사와 대명사간의 거리도 큰 역할을 한다는 것을 알 수 있었다. 또한 센터링 이론에서 밝힌 바와 같이 선행사 후보의 문법기능도 적지 않은 역할을 하면서, 마지막으로 선행사의 의미부류를 통한 선택제약 정보도 대용어 해소에 다소의 영향을 미친다는 것을 알 수 있었다.

이어서 논조분석에 대용어해소 방법론을 적용할 경우 어느 정도의 성능변화가 있는지를 보고자 하였다. 대용어해소를 위해 사용한 300개의 평가문장을 논조분석을 위해서도 사용하였다. 논조분석 방법론은 홍문표 (2014)에서 제안한 바이그램 기반의 논조분석 방법이었다. 즉, 이것은 학습데이터에 등장하는 유니그램 및 바이그램들을 기계학습의 자질로 삼아 분류기를 구현하는 방법이다.

대용어해소와 마찬가지로 웨카 3.6.9 프로그램을 이용하여 구현하였으며 5 분할 균등 상호교차평가의 방식으로 정확률을 측정하였다. 대용어해소 없이 문장레벨에서 측정한 논조분석의 정확률은 76.2%였다. 각 문장에 등장하는 대용어의 선행사를 언어학적 제약만을 이용하여 복원한 후 논조분석을 수행하면 논조분석의 정확률은 32%로 떨어진다. 그러나 기계학습을 통해 대용어해소를 수행한 후 논조분석을 수행할 경우에는 68.86%로 정확률이 36% 이상 상승함을 볼 수 있었다. 물론 본 연구에서는 주격 대명사만을 분석의 대상으로 하였기 때문에 Nikolov et al. (2008) 등의 연구결과와 직접적으로 비교하는 것은 어렵지만, 대용어의 선행사를 복원하고도 논조분석의 정확률이 68.86%로서 거의 70%에 육박한다는 점은 향후 양상기반의 논조분석을 수행할 때 대용어해소가 큰 역할을 할 수 있다는 점을 시사한다고 볼 수 있다.

Ⅴ. 맺는말

본 연구에서는 논조분석이 좀 더 실용적으로 활용될 수 있기 위해서는 구구조나 양상레벨에서의 논조분석이 수행되어야 함을 지적하였다. 이를 위해서는 대용어 해소가 반드시 필요하다는 것을 밝혔으며 이를 위한 방법론을 제안하였다. 기존의 언어학적 제약을 활용한 대용어해소 방법론은 결정론적 특징을 지니지만 다수개의 선행사 후보가 제약을 만족할 경우 최적합 후보를 선택하는데 어려움이 있었다. 이를 해결하고자 기계학습 기반의 방법론을 제안하였으며, 6개의 자질을 소개하였다.

본 연구에서 제안한 방법론의 성능을 평가한 결과 기존의 언어학적 제약기반의 방법론과 비교하여 월등히 높은 90.38%의 정확률을 보였으며, 논조분석에 적용할 경우 논조분석의 최종정확률이 68.86%로 측정되었다. 이 수치가 의미하는 것은 대략적으로 주격 대명사가 포함된 100개의 문장을 논조분석할 경우 주격 대명사의 선행사와 해당 문장의 논조를 정확하게 파악한 문장이 약 68개 정도 된다는 것이다.

그러나 본 연구의 범위는 주격대명사에만 한정되어 있었다. 주격대명사가 아닌 목적격이나 소유격 대명사의 선행사를 찾아내는 것도 매우 중요한 과제이며 해결방안이 본 연구에서 제안하는 방법과 다를 수도 있다. 또한 대명사만이 아니라 일반 한정의미의 명사구와 공지시 관계를 갖는 공지시관계 해소 또한 정확한 논조분석을 위해 반드시 해결되어야 할 과제라고 볼 수 있다. 이러한 주제들에 대한 연구를 본 연구의 후속주제로 다룰 예정이다.

참고문헌

박아름 & 홍문표(2015:a) 한-독 기계번역을 위한 한국어 영형 주어 처리연구, 독일문학 133집, 56권 1호, 197-223.

박아름 & 홍문표(2015:b) 한-독 기계번역을 위한 한국어 영형목적어 처리 연구, 독일언어문학 70집, 1-21.

홍문표(2013) 독일어 텍스트 논조자동분석, 독어학 28집, 361-383.

홍문표(2014) 바이그램을 활용한 텍스트 논조자동분석, 독일언어문학 65집, 27-46.

홍문표(2015) 독일어 트위터 문장의 규칙기반 논조분석 방안 연구, 독어학 32집, 153-173.

홍문표(2016) 논조분석을 위한 반어의미 문장 자동분류, 독어학 33집, 157-179.

Cambria, E., B. Schuller, Y. Xia & C. Havasi(2013) New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15-21.

Eberle, K.(2003) Anaphernresolution in flach analysierten Texten für Recherche und Übersetzung. In LDV Forum , Vol. 18, No. 1/2, 216-232.

Hinrichs, E., K. Filippova & H. Wunsch(2007) A data-driven approach to pronominal anaphora resolution for german. AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4, 292, 115.

Hurst, M. F. & K. Nigam(2003) Retrieving topical sentiments from online document collections. International Society for Optics and Photonics, 27-34.

Joshi, M. & C. Penstein-Rose(2009) Generalizing dependency features for opinion mining. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics, 313-316.

Liu, B.(2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Lukin, S. & M. Walker(2013) Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In Proceedings of the Workshop on Language Analysis in Social Media, 30-40.

Maynard, D. & M. A. Greenwood(2014) Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In LREC, 4238-4243.

Mitkov, R.(2014) Anaphora resolution. Routledge.

Moilanen, K. & S. Pulman(2007) Sentiment composition. In Proceedings of RANLP, 378-382.

Nicolov, N., F. Salvetti, & S. Ivanova(2008). Sentiment analysis: Does coreference matter? In AISB 2008 Convention Communication, Interaction and Social Intelligence, Vol.1, 37.

Pang, B., L. Lee & Vaithyanathan, S.(2002) Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing − Volumn 10.

Pang, B. & L. Lee(2004) A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04), 271-278.

Park, A. & Hong, M.(2014) Hybrid Approach to Zero Subject Resolution for multilingual MT-Spanish-to-Korean Cases. In Proceedings of the 28th Pacific Asia Conference On Language Information and Computing, 254-261.

Pontiki, M., D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos & S. Manandhar(2014) Semeval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th international workshop on semantic evaluation, 27-35.

Remus, R. U. Quasthoff & G. Heyer(2010) SentiWS - a Publicly Available German-language Resource for Sentiment Analysis. Proceedings of the 7th International Language Ressources and Evaluation(LREC'10), 1168-1171.

Schwarz-Friesel, M.(2009) Ironie als indirekter expressiver Sprechakt: Zur Funktion emotionsbasierter Implikaturen bei kognitiver Simulation. Perspektiven auf Wort, Satz und Text. Semantisierungsprozesse auf unterschiedlichen Ebenen des Sprachsystems, 223-232.

Waltinger, U.(2010) German Polarity Clues : A Lexical Resource for German Sentiment Analysis, Proceedings of the Seventh International Conference on Language Resources and Evaluation(LREC), 1638-1642.

Wiebe, J., Wilson, T. & C. Cardie(2005) Annotating Expressions of Opinions and Emotions in Language, Language Resources and Evaluation, 39(2/3), 164-210.

Wilson, T., J. Wiebe, & P. Hoffmann(2005) Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing, 347-354.

Zusammenfassung

Anaphernresolution für die Sentimentanalyse deutscher Texte

Hong, Munpyo (Sungkyunkwan Univ.)

Das Ziel der Sentimentanalyse liegt darin, die Stimmungen des Sprechers oder des Autors eines Textes automatisch zu bestimmen. Die Stimmungen werden normalerweise im Text durch die Verwendung bestimmter Wörter übertragen. Aber die konnotative Bedeutung eines Wortes wird sowohl vom syntaktischen als auch vom semantischen Kontext beeinflusst. Um dies in der Sentimentanalyse zu berücksichtigen, werden verschiedene Ansätze vorgeschlagen, u.a., die von Wiebe et al. (2005) und von Joshi & Pennstein-Rose (2009). In diesen Arbeiten wurden hauptsächlich die Methoden für die Berücksichtigung der syntaktischen Struktur in einer Sentimentanalyse vorgeschlagen.

In der vorliegenden Arbeit wurde ein Ansatz für die sogenannte Anaphernresolution vorgestellt. Anaphernresolution ist ein sehr wichtiges Thema in der Sprachverarbeitung und stellt auch ein sehr schwieriges Problem für die Sentimentanalyse dar. Die meisten bisherigen Ansätze für Anaphernresolution sind regelbasiert und deshalb deterministisch. Dies bedeutet, dass wenn ein Kandidatnomen alle linguistischen Constraints für die Resolution eines Pronomens erfüllt, es automatisch für das Antezedens eines Pronomens gehalten wird. Aber wenn gleichzeitig mehrere

Kandidaten die Constraints erfüllen, fällt in diesem Ansatz die Entscheidung sehr schwer.

Um das zu vermeiden, wird hier ein anderer Ansatz, der sich auf maschinelles Lernen stützt, vorgeschlagen. Für das maschinelle Lernen werden sechs Merkmale verwendet. Das Experiment zeigt, dass der vorgeschlagene Ansatz für die Anaphernresolution 90.38% Genauigkeit aufweist. Versucht man die Sentimentanalyse mit Anaphernresolution zu verbinden, weist er 68.86% Genauigkeit der Sentimentanalyse auf.

핵심어 : 대용어해소 Anaphernresolution, 논조분석 Sentimentanalyse,
기계학습 maschinelles Lernen, 담화분석 Diskursanalyse,
데이터 마이닝 Data Mining

필자 E-mail : skkhmp@skku.edu

논문투고일 : 2016. 9. 1 / 심사일 : 2016. 9. 10 / 게재확정일 : 2016. 9. 19