한국어 텍스트의 논증 구조 내 담화 관계의 자동 분류 연구

이상아○, 신효필
서울대학교 언어학과
visualjan@snu.ac.kr, hpshin@snu.ac.kr

An Automatic Classification of Discourse Relations in the Arguing Structure of Korean Texts

Sana Lee○, Hyopil Shin
Seoul National University, Department of Linguistics

요 약

최근 온라인 텍스트 자료를 이용하여 대중의 의견을 분석하는 작업이 활발히 이루어지고 있다. 이러한 작업에는 주관적 방향성을 갖는 텍스트의 논증 구조와 중요 내용을 파악하는 과정이 필요하며, 자료의 양과 다양성이 급격히 증가하면서 그 과정의 자동화가 불가피해지고 있다. 본 연구에서는 정책에 대한 찬반 의견으로 구성된 한국어 텍스트 자료를 직접 구축하고, 글을 구성하는 기본 단위들 사이의 담화 관계를 정의하였다. 각 단위들 사이의 관계는 기계학습과 규칙 기반 방식을 이용하여 예측되고, 그 결과는 합성되어 하나의 글에 대응되는 트리 구조를 이룬다. 또한 텍스트의 구조상에서 주제문을 직접적으로 뒷받침하는 문장 혹은 절을 추출하여 글의 중요 내용을 얻고자 하였다.

주제어: 담화 관계 자동 분류, 논증 구조 분석, Argumentation Mining, 한국어 텍스트 분석

1. 서론

인터넷의 발달은 개개인의 의견 표출을 자유롭게 하였다. 사람들은 온라인상에서 게시글을 작성함으로써 자신의 의견을 불특정 다수의 독자에게 표현하고, 또 독자들이 자신의 의견을 따르도록 설득하기도 한다. 이러한 움직임에 따라, 온라인 텍스트 자료를 통해 대중의 의견을 파악하고자 하는 경우가 많아졌다. 예를 들면 기업은 상품에 대한 소비자의 반응을 살피고, 선거철 정계는 유권자들이 특정 후보를 얼마나 지지하는지 알고 싶어할 것이다.

하지만 온라인 게시글은 그 수가 매우 많고 주제와 형태도 다양하다. 따라서 단순히 게시글을 읽는 것으로는 많은 사람의 의견을 모으고 파악하기가 어렵다. 즉 데이터의 규모가 커지면서 자동화가 불가피해지는 것이다. 텍스트에서 자동적으로 중요한 정보를 추출해 취합하는 과정이 필요하다.

이 연구의 목표는 어떤 주관적인 주장을 포함한 글이 있을 때, 특정 주장을 개진하기 위해 글을 이루는 구조를 자동으로 파악하도록 하는 것이다. 특히 주제문의 바로 하위에서 주제문을 직접 뒷받침하는 문장 혹은 절을 찾아내, 주제와 그에 대한 주장의 방향성에 따라 저자가 제시하고자 하는 구조화된 근거를 얻는 데 도움이 되고자 한다. 이러한 정보는 차후에 주제별 입장 분류, 근거 추출, 텍스트 요약 등의 응용이 가능할 것이라 기대된다.

이를 위해 글을 이루는 문장 또는 절 사이의 관계를 네 가지 유형으로 정의하고, 그 관계를 이용해 글의 구성을 체계화하는 모듈을 개발하였다. 이는 기계 학습 방법을 기본으로 하며, 규칙 기반 방식을 통해 예측 결과를 보정하는 과정으로 이어진다. 또한 직접 구축한 한국어 데이터를 가지고 그 성능을 검증하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 텍스트의 구조를 자동으로 파악하고자 한 기존의 연구에 대해 간략히 기술한다. 3장에서는 앞서 언급했듯 기계 학습을 이용하여 텍스트를 구조화하는 과정을 설명하고, 4장에서 그 실험 결과를 서술할 것이다. 마지막으로 5장에서는 결론과 향후 과제에 대해 논의한다.

2. 관련 연구

2.1 텍스트 구조 분석

텍스트의 구조를 자동 분석하려는 시도는 꾸준히 있어 왔다. 영어 자료를 대상으로 하여 문맥 정보를 이용한 분석기도 여럿 제시되었다. Penn State Discourse Treebank (PDTB) 기반 분석기[1]는 접속사 중심으로 몇 가지의 자질을 이용하여 문장이나 구 단위 사이의 관계를 예측하도록 하였는데, 이 관계 안에는 두 단위 중 어느 것이 핵심적인 내용인지 나타나 있지는 않다. Rhetorical Structure Theory (RST) 기반 분석기[2]는 인접한 두 개의 문장 혹은 구 사이의 관계를 시작으로, 점점 상위 단계로 올라가면서 하나의 글을 이루는 방식을 사용했다. 이러한 분석기들은 한국어 자료에 대해서는 사용할 수 없으며, 주관적인 의견을 피력하는 글의 분석 목적을 충족시키기 어렵다.

[3]에서는 대화형으로 연결된 다이얼로그 데이터를 대상으로 하여, 해당 자료에서 중요하게 다루고 있는 내용을 얻고 비슷한 것끼리 묶는 클러스터링 작업을 하였다. 이 때 각 클러스터는 하나의 내용을 담은 facet이 된다. 또한 [4]는 온라인 토론 자료를 분석하였는데, 각 게시글이 특

정 주장을 지지하기 위해 제시하는 근거를 유형화하고 또 자동 분류하였다. 그런데 [3]의 중요 내용 부분과 [4]에서 추출한 근거는 모두 수동으로 요약, 주석한 것이다. 본 연구에서는 텍스트의 중요 내용을 찾는 부분까지도 최대한 자동으로 처리할 수 있도록 하는 것이 목적이다.

2.2 unit 사이의 관계

텍스트 내에서 문장이나 절이 모여서 구조를 이루기 위해 가지는 관계 역시 여러 가지로 제안되어 왔다. [5]는 관계를 설정하기 이전에 unit 자체를 major claim과 claim, premise의 세 가지로 분류하고, 이 분류에 따라 하나의 unit이 다른 하나의 unit을 뒷받침하는 관계를 연산하였다. 그런데 이러한 unit의 분류는, 정해진 형식이 거의 없어 다양하게 나타나는 본 연구의 자료에 적용하기에는 무리가 있다.

[6] 역시 unit의 claim과 premise 개념을 이용하였으며, 이러한 unit들이 모여 이루는 논증 구조의 유형을 제시하였다. 이 때 [6]에서 제안한 구조는 근거와 결론으로 이루어진 하나의 논증 구조이다. 본 연구에서 정의한 unit 간의 관계들은 이보다 간략하고 기본적인 단위에 해당하며, 이러한 기본 관계들이 모여서 [6]의 구조 유형을 포함한 여러 논증 구조를 형성하게 된다.

3. 논증 구조 구축

본 연구는 논증의 구조를 자동으로 찾아내는 것을 목적으로 한다. 이는 곧 기본 unit 단위를 정의하고, 그 unit 사이에 관계를 할당함으로써 한 편의 글이 어떻게 구성되어 있는지를 알 수 있도록 하는 과정을 말한다.

3.1 데이터

포털사이트의 온라인 토론글을 자료로 하였다. 포털사이트 '다음'의 '아고라'에서는 특정한 주제가 토론거리로 주어지고, 사람들이 그에 대해 자유롭게 게시글을 작성하여 의견을 개진하는 구조를 취한다. 그 중에서 서울시 정책에 대해 토론하고 있는 적절한 주제 두 가지를 선정하였다. 각각의 주제에 해당하는 게시글을 크롤링하여 코퍼스를 구성하고, 각 게시글이 주제에 대해 '찬성'과 '반대' 중 어느 입장에 속하는지를 주석하였다.

[표1] 게시글 분포

주제	PRO	CON	합
길거리 쓰레기통 설치 찬/반	190	70	260
한강변 바베큐 파티 허용 찬/반	42	137	179

본 연구는 총 439개의 한국어 게시글을 대상으로 하였고, 주제와 찬반 입장에 따른 게시글의 분포는 [표1]에서 보는 바와 같다.

'아고라'의 게시글은 제목과 본문으로 이루어져 있다. 기본적으로는 사용자가 게시글에 댓글과 답글을 달 수 있게 되어 있으나, 이 연구에서는 댓글은 고려하지 않았고 답글은 별개의 게시물로 간주하였다.

3.2 트리 구조 구축

3.2.1 기본 unit 정의

하나의 의미를 일관적으로 나타내는 가장 작은 단위를 얻기 위해, 기본 단위가 되는 'unit'의 경계를 수동으로 정의 및 분리하는 작업을 수행하였다. 하나의 단위는 문장을 기본으로 하되 절(clause) 경계에 따라 분리하였다. 따라서 하나의 unit은 문장 전체가 되기도 하고, 하나의 문장을 이루는 여러 개의 절 중 하나가 되기도 한다. 이렇게 분리된 unit은 띄어쓰기나 오타 수정 등의 전처리를 거쳐, 형태소 분석 결과와 의존 구조 분석 결과를 각각 추가적인 정보로 가지게 된다.

[표2] 하나의 게시글을 이루는 unit 목록

0 쓰레기통이 있어야 합니다.
1 길 가다가 편의점이나 슈퍼 같은 데서 아이스크림이나 음료 과자 같은 거 사 먹고 나면 쓰레기가 발생하는데
2 쓰레기통이 없으니 많이 불편하죠.
3 예전에 그 많던 쓰레기통 그 쓰레기통 주변이 더럽다고 치워 버렸는데
4 쓰레기통을 즉각 즉각 비우면 더럽혀질 일이 별로 없을 겁니다.
5 도심의 길거리 쓰레기통 설치하고 관리하는 미화원 배치하면 일자리도 생기고 거리도 깨끗해지고 좋을 것 같네요.

[표3]에 따르면, 주제별, 입장별 분류에 따라 게시글의 평균 길이에 차이가 없다는 것을 알 수 있다. 즉 게시글의 구조를 파악할 때 게시글의 길이 자체가 변이 요소로 작용하는 경우는 없을 것이라고 예상된다.

[표3] 주제별, 입장별 게시글의 평균 길이 (unit 개수)

	PRO	CON
길거리 쓰레기통 설치 찬/반	7.8684	7.9714
한강변 바베큐 파티 허용 찬/반	8.0238	7.8394

3.2.2 Unit 간 관계 설정

Root Node 처리

'아고라' 게시글의 경우는 글의 제목이 글 전체의 내용을 압축적으로 나타내므로 제목을 곧 주제문이라고 볼 수 있었다. 글의 unit들을 각각의 node로 하여 하나의 트리 구조를 만들면, 주제문이 root node에 위치하게 되는 것이다.

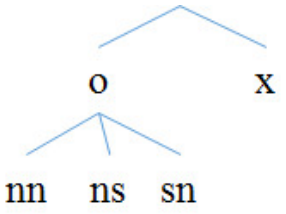
본 연구에서는 게시글의 root 정보가 주어졌다고 가정하였다. 대부분의 게시글에서 첫 번째 unit 즉 제목을 root로 이용하였고, 이는 곧 게시글이 두괄식 논증의 형태를 갖추었다고 볼 수 있다. 이렇게 두괄식으로 쓰인 게시글은 모두 289개로, 전체 데이터의 약 65.8%를 차지한다. 또한 그렇지 않은 일부의 게시글에 대해서도 root 정보를 할당한 뒤 연구를 진행하였다. 따라서 root node는 다음과 같이 세 가지의 유형을 갖는다.

- 한 개의 unit만이 root에 해당하는 경우: 게시글의 제목이 대부분이나 미괄식 구조 등의 다른 예도 존재한다.
- 두 개 이상의 unit이 root에 해당하는 경우: 글의 구

조가 양괄식이거나 글의 주제가 반복적으로 언급될 때, 주제가 되는 여러 개의 unit은 모두 같은 내용을 담고 있으므로 결국 같은 node로 간주해도 무방하다.

- root이 explicit하게 드러나지 않는 경우: 어느 입장을 지지하는지는 알 수 있으나 확실한 하나의 주제문이 없이, 해당 입장을 지지하는 근거들만 열거된 경우이다. 이때는 임시로 zero-form root를 만들고, 이것을 parent로 해서 근거 node들이 매달리는 구조로 처리한다.

기본 담화 관계 정의



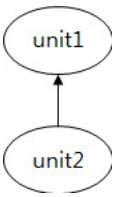
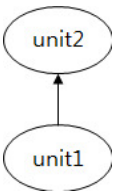
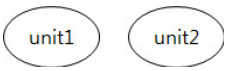
<그림1> 담화 관계 분류 구조도

두 개의 unit 사이의 담화 관계는 두 개의 층위에 걸쳐서 정의된다. <그림1>과 같이, 먼저 두 unit이 서로 내용적으로 관련이 있는지 없는지를 결정한다. 그리고 서로 관련이 있는, 즉 'o'로 분류된 관계들에 대해서만 'nn', 'ns', 'sn'의 세부 분류를 수행한다.

두 unit 중 앞에 나오는 것을 unit1, 뒤에 나오는 것을 unit2라 하고 이 사이의 관계를 정의하면 다음과 같다.

- 'o'는 두 unit이 의미적으로 관계를 가지고 있는 경우. 같은 내용을 나타내고 있거나, 어느 한 쪽이 다른 한 쪽의 하위 내용이 되는 경우를 포함한다.
- 'x'는 같은 입장을 뒷받침하는 unit이기는 하지만 직접적인 관계를 가지고 있지 않을 때 'x'로 분류된다. 내용적으로는 새로운 subtree를 구성할 수 있고, 두 unit은 서로 동등한 층위에 있게 된다.
- 'nn'은 nucleus-nucleus라는 의미로, unit1과 unit2가 동등한 층위에 있으면서 같은 내용을 나타내는 경우를 말한다. 이 때 두 unit이 하나의 unit으로 합쳐진 것으로 간주된다고 해도 내용적으로는 무리가 없다. 만약 게시글에서 여러 개의 root unit이 주어졌다면, 이 unit들 간에는 모두 'nn' 관계가 성립한다.
- 'ns'와 'sn'은 모두 수직적 관계를 나타낸다. 따라서 두 unit의 층위에는 차이가 발생하게 된다. 'ns'와 'sn' 여부는 상위 노드가 되는 unit이 무엇이냐에 따라 달라진다. 'ns'는 nucleus-satellite으로 unit2가 unit1의 하위 노드로 들어가면서 unit1을 뒷받침하는 관계를, 'sn'은 반대로 satellite-nucleus를 뜻하면서 unit1이 unit2의 하위 노드로 들어가면서 unit2를 뒷받침하는 관계를 나타낸다. 결국 이 두 경우 모두 하위 노드가 상위 노드를 뒷받침하는 관계에 속하며, 이 관계는 곧 부연설명, 이유 제시, 세부 사항 서술, 논평, 의견 등의 세부 내용으로 나뉘게 된다.

위의 관계들은 <그림2>와 같이 시각화된다.

관계	그림	예시
ns		unit1: 한류 열풍 어쩌고 하는 게 낯 뜨거울 정도로 명동이나 종로 거리는 온갖 쓰레기의 전시장입니다. unit2: 이런저런 이유를 대 가면서 미화요원들 잘라낸 때문이지요.
sn		unit1: 길 가다가 편의점이나 슈퍼 같은 데서 아이스크림이나 음료 과자 같은 거 사 먹고 나면 쓰레기가 발생하는데 unit2: 쓰레기통이 없으니 많이 불편하죠.
nn		unit1: 길거리 휴지통 있어야 합니다. unit2: 길거리에 휴지통이 있어야 합니다.
x		unit1: 설치하는 게 좋다고 생각함. unit2: 60대 중반입니다.

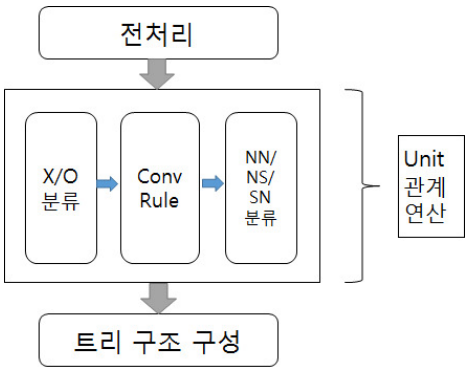
<그림2> 관계 시각화 및 예시

위 <그림2>에서 각 unit은 타원으로 표시하였고, 두 개의 타원을 잇는 직선은 두 개의 unit 사이의 관계를 뜻한다. 이 때 화살표는 수직적인 관계에서 하나의 unit이 다른 하나의 unit을 뒷받침하는 경우를 뜻한다. 화살표가 시작되는 방향, 즉 아래에 있는 unit이 다른 하나의 하위로 들어가면서 세부 사항 서술, 근거 제시 등의 역할을 하게 된다. 화살표가 아닌 직선은 두 unit이 동등함을 나타낸다.

3.2.3 Unit 간 관계 자동 분류

3.2.2에서 정의한 대로 'o'와 'x'의 분류, 그리고 'o' 하위의 'nn', 'ns', 'sn'의 분류, 이렇게 두 단계로 나누어 분류 작업을 진행하였다.

<그림3>에 나타나듯, 두 단계의 학습 및 예측 과정을 진행하였고, 특히 x/o 분류의 예측 뒤에는 규칙에 기반한 예측 결과 보정 과정을 거치도록 하였다.



<그림3> 분류 관계 연산 과정

기계학습모델

Training 및 predicting에는 Scikit-learn 패키지의 Support Vector Machine을 이용하였고, 이 때 사용된 feature는 다음과 같다.

- 관계 유형에 따른 어휘적 자질

관계의 세부적인 유형에는 이유, 예시, 보완책, 대책, 결과, 세부 사항, 화제 전환, 열거, 반박 등이 있다. 두 unit 사이의 관계가 어떤 세부 유형에 해당하느냐에 따라 각 unit은 다른 종류의 explicit connective를 포함할 것이다. 이러한 어휘 자질은 [13]에서 구축한 한국어 사전과 [14]의 코퍼스를 참고하였고, 각 관계 및 세부 분류에 맞는 것을 선정하여 이용하였다. 어휘적 자질의 구체적인 예시는 [표4]와 같다.

또한 위와 같은 세부 분류는 unit 간의 'nn', 'ns', 'sn', 'x'와 같은 관계의 대분류에 각각 속함으로써 그 예측에 도움을 줄 것이라 가정하였다.

[표4] 어휘적 자질 예시

관계	어휘적 자질 예시
nn	또한/MAJ, 게다가/MAG, 더욱이/MAJ, …
ns	이유: 때문/NNB, 덕분/NNG, 왜냐하면/MAG … 예시: 예컨대/MAG, 특히/MAG, … 기타: 가령/MAG, 이를테면/MAG, …
sn	이유: 아서/EC, 으니/EC, 때문/NNB, … 반대 입장: 지만/EC, 아도/EC, 그러나/MAJ…
o	따라서/MAJ, 그러니까/MAJ, 그러므로/MAJ, …
x	한편/MAG, 아무튼/MAG, 어쨌든/MAG, …

- unit2가 의존적인 문장 형태인가?

관계를 이루는 인접한 두 개의 unit 중 뒤에 나오는 것이 적절한 주어나 목적어, 서술어 등을 가지지 않는 경우, 독립적인 단위일 가능성이 낮다. 또한 생략된 문장 성분은 앞에 나오는 unit의 정보에 의존하고 있을 것이다.

- Unit2가 대명사를 포함하고 있는가?

대명사는 reference를 가지므로, 앞에 나오는 unit의 정보에 의존할 수밖에 없다.

- 두 unit이 같은 문장에 포함되어 있는가?

한 문장에 포함된 unit들은 서로 연관되어 있을 가능성이 높다.

- 두 unit이 모두 주제문인가?

두 unit이 모두 주제문일 경우, 이 두 unit은 서로 동등하며 같은 내용을 나타낸다.

- Word pairs

unit1의 단어와 unit2의 단어를 하나씩 짝지어 pair 형태로 만든다. 이 때 기능 형태소는 제외하며, 이렇게 만들어진 pair의 목록을 bag of words와 같이 사용한다. 실험의 기본 baseline으로 사용되어 비교의 기준으로 쓰인다.

규칙에 기반한 예측값 조정

SVM을 통해 예측한 결과에 후처리 규칙을 적용시켜서 정확도를 높이도록 한다.

- Convergent Relation Rule:

앞에 있는 unit들끼리의 관계를 먼저 예측하고, 먼저 예측해 둔 관계들을 다른 unit들 사이의 관계 예측에 이용한다. 위치에 따라 반드시 'x'의 관계, 즉 직접 연관성이 없는 관계에 있을 수밖에 없는 경우를 규칙으로 정의하는 과정이다. 이러한 관계 보정 규칙은 <그림4>와 같이 pseudocode로 나타낼 수 있다.

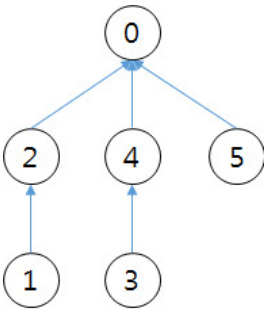
```
for unit_j in <unit list>:
    for unit_i in <unit list>:  (i<j)
        for unit in <unit list>:  (x<i)
            if relation(unit_x, unit_i) is 'x':
            then relation(unit_x, unit_j) = 'x'
```
<그림4> Convergent Relation Rule

3.2.4 트리 구조 구축

글에 있는 모든 unit pair에 대해서 관계 할당이 완료되면, 이들을 모아서 하나의 트리 구조로 합성한다. 즉 최소 단위의 담화 관계가 모여서 글 한 편에 대응되는 트리 구조를 만드는 것이다. 이 과정은 <그림3>의 마지막 단계에 표시되어 있다.

위의 관계에 기반하여 unit들을 연결하고, 그렇게 완성된 구조를 그림으로 나타내면 <그림5>와 같다.

<그림5> 텍스트의 트리 구조 예시



또한 이렇게 구성한 트리 구조에서, root 바로 하위에 매달리는 unit만을 따로 추출한다. <그림5>에서는 (2), (4), (5)번 unit이 여기에 해당되며, 그 unit들은 아래

[표5]에 열거하였다. 이는 주제문인 root를 뒷받침하는 주요 근거가 될 것이다. 이보다 하위에 있는 unit은 생략 가능한 세부 사항일 가능성이 높다. 즉 이 단계는 텍스트 내에서 중요한 근거로 쓰인 unit을 찾고, 내용을 파악하는 데 필요하지 않은 것을 제외하고자 하는 과정이 된다.

[표5] root 하위의 중요 근거 unit 예시

2 쓰레기통이 없으니 많이 불편하죠.
4 쓰레기통을 즉각 즉각 비우면 더럽혀질 일이 별로 없을 겁니다.
5 도심의 길거리 쓰레기통 설치하고 관리하는 미화원 배치하면 일자리도 생기고 거리도 깨끗해지고 좋을 것 같네요.

4. 실험 결과

각 주제별 게시글이 지지하는 입장과 주제문, 그리고 모든 unit pair가 갖는 관계성을 수동 주석하였다. 이러한 관계를 자동으로 예측한 뒤 규칙을 이용해 보정한 결과 및 그 평가는 다음과 같다.

[표6] O/X 분류 결과

	precision	recall	f1-score	support
o	0.608	0.457	0.52	360.3
x	0.882	0.935	0.909	1720.7
avg	0.835	0.846	0.837	2081

[표7] O/X를 feature만을 이용해 분류한 결과

	precision	recall	f1-score	support
o	0.571	0.497	0.53	360.3
x	0.888	0.917	0.902	1720.7
avg	0.833	0.84	0.836	2081

[표8] O/X 분류 baseline

	precision	recall	f1-score	support
o	0.344	0.113	0.164	360.3
x	0.827	0.954	0.883	1720.7
avg	0.742	0.8	0.754	2081

[표6]와 [표7], [표8]은 관계의 첫 층위인 'o'와 'x'의 분류 결과를 나타낸 것이다. 이 단계에서는 기계 학습과 rule을 이용한 결과 보정 과정이 있었으므로 세 가지 경우에 대해 결과를 비교하였다. [표6]는 3.2.3에서 정의한 자질 목록과 Convergent Relation Rule 모두를 사용해 실험한 결과를 나타내고, [표7]은 규칙의 적용 없이 자질 목록만을 이용해 기계 학습을 한 결과를 표시한다. 또한 [표8]은 가장 기본적인 자질인 Word Pair 목록만을 가지고 실험한 결과를 나타낸 것이다. 이들 결과에 따르면, 본 연구를 위해 정의한 자질 목록을 사용했을 때 성능은 0.754(F1)에서 0.836(F1)로 향상된다.

Convergent Relation Rule을 적용하면 0.837(F1)의 성능을 보이게 된다. 이는 f1-score의 평균값을 보아서는 의미 있는 향상은 아니나, 'o'로 분류되는 경우의 precision이 0.571에서 0.608로 올라가는 것을 볼 수 있다. 본 연구에서는 이 단계에서 'o'로 분류된 관계에 대해서만 그 다음 층위인 'nn', 'ns', 'sn'의 분류를 수행하므로, 'o'의 precision이 향상된다는 점이 중요하다.

[표9] nn/ns/sn 분류 결과

	precision	recall	f1-score	support
nn	0.644	0.205	0.301	82.8
ns	0.644	0.92	0.756	202.7
sn	0.583	0.359	0.436	74.8
avg	0.637	0.633	0.584	360.3

[표10] nn/ns/sn 분류 baseline

	precision	recall	f1-score	support
nn	0.315	0.059	0.095	82.8
ns	0.573	0.945	0.71	202.7
sn	0.398	0.066	0.108	74.8
avg	0.486	0.553	0.444	360.3

[표9], [표10]의 분류 결과는 상위 단계에서 'o'로 주석했던 unit pair만을 대상으로 한 것이다.

[표9]은 3.2.3에서 정의한 자질 모두를 사용한 결과로, Word Pair 자질만을 이용한 [표10]의 결과를 기준으로 비교할 수 있다. 위의 결과에 따르면, 가장 기본적인 자질을 이용해서 실험했을 때보다 추가적인 자질을 사용했을 때 성능이 0.486(F1)에서 0.584(F1)로 향상되었다.

위의 두 단계의 실험에서는 결과의 비교 대상을 [표8]과 [표10]에서와 같이 Word Pair 자질만을 사용한 경우로 설정하였다. 이는 본 연구가 한국어 자료를 대상으로 하였고 기존 연구와 다른 형태의 Annotation Scheme 및 작업 형태를 설정하여, 기존 연구와의 직접적인 비교가 불가능하기 때문이다. 따라서 가장 기본적인 자질인 Bag of Words를 응용하여 Word Pair 자질을 설정하고 이를 baseline으로 하였다.

또한 수동으로 주석한 구조에서 주제문 바로 하위에 있는 unit들을 Gold Standard로 하고, 기계 학습을 통해 얻은 unit들과 비교하여 [표11]과 같은 성능을 얻었다.

[표11] root 하위의 중요 근거 unit 추출 성능

	precision	recall	f1-score
score	0.8098	0.6640	0.7297

5. 결론

본 논문에서는 주관적인 방향성을 포함한 텍스트의 구조 내 담화 관계를 자동으로 분류하는 방법을 제안하였다. 먼저 특정 정책에 대한 주관적 논평으로 이루어진

한국어 텍스트 자료를 구축하였다. 또한 하나의 글을 이루는 문장이나 절의 단위들이 갖는 관계성에 기초해, 기계 학습과 규칙 기반 방식을 이용한 구조화 모듈을 제안하였다. 이 때 어휘 차원의 실마리를 포함한 몇 가지 언어적 특징들이 자질로서 기능한다. 이들 자질과 규칙의 유용성은 실험을 통해 확인되었다.

이렇게 얻은 텍스트의 구조로부터는 주제문을 직접 뒷받침하는 중요 근거를 추출할 수 있다. 이들 근거가 되는 unit이 아직은 raw text의 형태로 되어 있으므로, 이 unit 안에서 다시 핵심적인 정보 혹은 키워드를 추려내는 과정을 고려하고 있다.

또한 텍스트 내의 중요 근거는 다양한 방향으로의 응용이 가능할 것이다. 텍스트 자동 요약의 한 방법이 될 수도 있을 것이며, 텍스트 내의 중요 내용에 따라 그 주장의 입장을 자동 분류하는 작업도 생각해 볼 수 있다.

현재는 정책에 대한 입장을 표현한 한국어 텍스트 자료만을 분석 대상으로 하고 있다. 향후에는 영어 텍스트 자료로도 연구를 확장하여, 언어 의존적이지 않은 모듈을 완성하고자 한다. 또한 최종적으로는 신문 기사나 사설과 같은 다양한 장르의 자료도 분석 대상에 포함시킬 계획이다.

참고문헌

[1] Lin, Z., Ng, H., & Kan, M., “A PDTB-styled end-to-end discourse parser”, Natural Language Engineering Nat. Lang. Eng., 151-184, 2012.

[2] Feng, V., & Hirst, G., “A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing”, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014.

[3] Misra, A., Anand, P., Tree, J., & Walker, M., “Using Summarization to Discover Argument Facets in Online Idealogical Dialog”, Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015.

[4] Hasan, K., & Ng, V., “Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[5] Stab, C., & Gurevych, I., “Identifying Argumentative Discourse Structures in Persuasive Essays”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[6] Palau, R., & Moens, M., “Argumentation mining”, Proceedings of the 12th International Conference on Artificial Intelligence and Law - ICAIL '09, 2009.

[7] Govier, T., “A practical study of argument”, Belmont, Calif.: Wadsworth Pub, 1985.

[8] Hitchcock, D., “The Linked-Convergent Distinction”, Argumentation Library Reflections on Theoretical Issues in Argumentation Theory, 83-91, 2015.

[9] Potter, A., “A Discourse Approach to Explanation Aware Knowledge Representation”, In ExaCt (pp. 56-63), 2007.

[10] Potter, A., “Linked and Convergent Structures in Discourse-Based Reasoning”, In ExaCt (pp. 72-84), 2008.

[11] Polanyi, L., Van Den Berg, M., & Ahn, D., “Discourse structure and sentential information structure”, An initial proposal. Journal of Logic, Language and Information, 12(3), 337-350, 2003

[12] Villena Román, J., Collada Pérez, S., Lana Serrano, S., & González Cristóbal, J. C., “Hybrid approach combining machine learning and a rule-based expert system for text categorization”, AAAI, 2011.

[13] Jang, H., & Shin, H., “Effective Use of Linguistic Features for Sentiment Analysis of Korean”, In PACLIC (pp. 173-182), 2010.

[14] Jang, H., Kim, M., & Shin, H., “KOSAC: A Full-fledged Korean Sentiment Analysis Corpus”, In Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation, 2013.