2018년 8월

박사학위 논문

확장된 의미역 결정을 이용한 문서의 유사성 판단

조선대학교 대학원

컴퓨터공학과

이 은 지

확장된 의미역 결정을 이용한 문서의 유사성 판단

Determination of Document Similarity
Using Extended Semantic Role Labeling

2018년 8월 24일

조선대학교 대학원

컴퓨터공학과

이 은 지

확장된 의미역 결정을 이용한 문서의 유사성 판단

지도교수 김 판 구

이 논문을 공학박사학위신청 논문으로 제출함.

2018년 4월

조선대학교 대학원

컴퓨터공학과

이 은 지

이은지의 박사학위논문을 인준함

위원장 조선대학교 교수 정 일 용 (인)

위 원 조선대학교 교수 양 희 덕 (인)

위 원 숭실대학교 교수 홍 지 만 (인)

위 원 중앙대학교 교수 정 재 은 (인)

위 원 조선대학교 교수 김 판 구 (인)

2018년 6월

조선대학교 대학원

목 차

그 림 목 차

표 목 차

ABSTRACT

Determination of Document Similarity Using Extended Semantic Role Labeling

Eunji Lee
Advisor : Prof. Pankoo Kim, Ph.D
Department of Computer Engineering
Graduate School of Chosun University

Reusing documents is very common in the process of digitalizing information contents thanks to the Internet and the popularity of smartphone, and is in the complicated form of word insertion, deletion and replacement, and word order change. In particular, where a word in a document is replaced by a similar word semantically the same, it is not considered as an object of measuring similarity in the conventional method for measuring morphological similarity. Therefore, it has been studied to measure similarity to solve the aforementioned problem.

This study suggests a method for measuring semantic similarity, based on sentence structure analysis using semantic role labeling. Semantic role labeling is based on syntax analysis to analyze sentence elements in the Predicate-Argument structure, then determine and tag semantic roles of each sentence element in a sentence. It is used in various fields including machine translation or question-answering systems for semantic understanding of a document. Because the Predicate-Argument structure of a sentence is an important element showing the meaning thereof, and a predicate with a specific meaning requires essential argument information, common Predicate-Argument information is used for sentences with similar meaning.

In this study, semantic role labeling is used to improve detection performance fo r similar sentences having many transformations, for example, paraphrasing not de tected easily in the conventional similarity measurement methods. Conventional se mantic role labeling tools conduct document analysis based on language resources already constructed, and document analysis performance depends on the category of language resources.

In this study, FrameNet, one of conventional language resources for semantic rol e labeling, is used, which is manually constructed and very accurate, and to which the semantic information of 'predicative' and 'argument' is added. The process of extending FrameNet was conducted to address the issue of insufficient resources of FrameNet. The extended FrameNet is then used to select the predicative-argu ment information obtained through semantic role labeling for two documents as fe ature information for measuring document similarity and then measure similarity b etween two documents.

In this study, semantic role labeling information is used, which is obtained throu gh the extended FrameNet for measuring similarity between two documents. The result is then compared with the conventional methods for measuring similarity be tween documents by comparing it with cosine similarity and partial string similari ty used for measuring similarity between documents. Application of the suggested method for measuring similarity to the same experiment data reveals that the met hod suggested in this study does not show much difference from the conventional methods for plagiarized documents in which the documents are not modified much, but implements better results than the conventional methods for paraphrased docu ments with modified words and sentence structure.

Ⅰ. 서 론

1. 연구 배경

최근 표절과 관련한 사회적 이슈가 지속적으로 발생하고 있다. 표절은 “다른 사람 저작물의 전부나 일부를 그대로 또는 그 형태나 내용에 다소 변경을 가하여 자신의 것으로 제공 또는 제시하는 행위”라고 정의한다. 표절은 문학작품이나 학술논문 또는 기타 각종 글에만 국한되지 않으며, 음악, 영상물 등 모든 창작물에 해당된다[1].

1989년 팀 버너스 리에 의하여 웹(WWW, World Wide Web)의 개념[2]이 고안된 이래로 인터넷이라는 정보 공간을 통해 인간의 지식의 생산과 자유로운 공유가 가능하게 되었으며, 이러한 인터넷과 콘텐츠 공유기술의 발달과 더불어 폭발적인 인터넷 사용자의 증가와 함께 정보의 과부하시대가 도래하였다[3].

2017년 IBM Marketing Cloud Study(How Much Data is Created on the Internet Each Day)에 따르면 인터넷 상에 존재하고 있는 정보의 90%이상이 2016년 이후에 생산된 것으로 나타났으며, 사람, 기업 그리고 각종 기기들이 데이터 공장처럼 매일 엄청난 양의 정보를 생산하고 있다고 표현하였다[4].

이렇듯 정보의 양이 기하급수적으로 늘어나고 신문기사, 책, 학술논문과 같은 각종 정보들이 디지털화되어 온라인상에 존재함으로써 정보에 대한 접근성이 편리해짐과 동시에 무분별한 공유로인한 무단 도용이나 표절 등과 같은 사회적 문제가 대두되었다[5]. 표절은 정보화시대에서 정보의 바람직한 유통 및 활용에 악영향을 미치고 있으며, 이에 표절을 효과적으로 검출하기 위한 다양한 연구가 활발히 진행되고 있다.

문서 표절 검출을 위한 시스템은 문서 색인 단계와 유사성을 비교하는 단계로 구분된다. 먼저, 문서 표절 검사를 위한 문서 색인은 표절 검사 대상 문서와 비교되어질 문서를 색인하기 위하여 적용되며, 비교 문서를 얼마나 잘 색인하느냐가 문서 표절 검사 시스템의 정확성 및 속도 등 전체적인 성능에 영향을 미친다[6].

그리고, 문서 표절 검사 시스템의 가장 핵심적인 부분인 문서 유사성을 측정하는 방법은 크게 형태적 유사성을 비교하는 방법과 의미적 유사성을 비교하는 방법으로 구분할 수 있다.

형태적 유사성을 비교하는 대표적인 방법으로 문장 내의 인접한 n개의 단어들을 추출하여 비교하는 n-gram방식과 문장내 부분 문자열(Substring)을 비교하는 방식, 벡터공간 상에 문서를 나타내고 벡터간의 거리 측정을 통해 유사성을 판단하는 벡터 공간모델(VSM, Vector Space Model) 등이 있다[7, 8, 9]. 그러나, 이러한 형태적 유사성의 경우, 대부분의 표절검사 시스템에 적용되는 일반화된 방법이지만, 비교대상이 되는 두 문서에 포함된 단어만을 고려하기 때문에 단어의 교체, 다시쓰기, 고쳐쓰기 등 원문의 변형이 이뤄진 경우에는 유사성 측정에 고려되지 않는다는 단점을 가지고 있다[7, 8, 9].

이러한 문제를 개선하기 위해 의미적 유사성 측정 방법이 제안되었다. 의미적 유사성 측정 방법은 단어 상호간의 의미적 관계를 계층정보로 정의해 놓은 시소러스(Thesaurus), 지식베이스(Knowledge Base)를 이용한 방법이다. 이러한 의미적 유사성을 고려한 방법은 유사한 단어로의 교체나 변형을 고려할 수 있는 방법이지만, 문장 내 단어들 간의 유사성을 측정하는 방법일 뿐 단어가 속한 문장의 구조적 정보는 고려하지 못하기 때문에 다시쓰기 문장과 같은 다양한 형태로 진화하는 표절의 유형을 검출하는데에 한계가 존재한다[10, 11].

대부분의 사람들은 다른 저작의 표현을 복제하면 표절이지만, 해당 표현을 바꿔서 옮기는 다시쓰기(paraphrasing)는 표절이 아니라고 생각한다. 하지만, 표현을 바꾸어 쓰더라도 사실상 같은 내용으로 전체적인 요지와 구성이 같으면 표절이 된다.

리처드 앨런 포스너(Richard A. Posner)가 제시한 하버드 대학의 '비스와나탄 사례'는 다른 이의 표현을 그대로 복제하지 않아도 표절이 성립한다는 것을 잘 보여준다[12]. 다시쓰기와 같은 원문 형태의 변형이 많이 이루어진 문서의 표절 검사를 위해서는 문서의 구조 분석을 기반으로 의미 정보를 고려하여 내용적 유사성을 측정할 수 있는 방법이 요구된다.

본 논문에서는 이러한 다시쓰기 문장과 같이 원문의 변형이 많이 이루어진 의미적으로 유사한 문서의 유사성 측정을 위해 문장의 구조 정보를 기반으로 문서의 의미적 분석 방법인 의미역 결정(SRL, Semantic Role Labeling)을 통해 얻은 정보를 활용하여, 문서의 유사성 측정 방법에 대해 제안하고자 한다.

의미역 결정은 구문 분석을 기반으로 술어-논항정보를 추출하여, 각 문장 성분이 문장 내에서 맡고있는 의미 역할을 결정 및 태깅하는 작업으로, 문서 내용의 의미적 이해를 통해 기계번역, 질의응답시스템 등의 분야에서 다양하게 활용되고 있다 [13, 14]. 일반적으로 의미역 결정은 기 구축된 언어자원에 기반하여 문서를 분석하며, 언어자원의 범주에 따라서 문서 분석의 성능이 좌우된다. 본 논문에서는 의미역 결정의 성능 개선을 위해 기존의 의미역 결정 언어자원을 확장하고, 확장된 의미역 결정을 활용하여 문서의 유사성 측정을 수행한다.

2. 연구 내용 및 범위

본 연구는 단어의 변형과 유사단어를 이용한 다시쓰기 표절처럼 원문 변형이 많이 이뤄진 표절 문서에 대한 유사성 측정을 위해 의미역 결정을 이용하여 문장의 구조분석을 통한 의미적 유사성 측정방법에 대해 제안한다.

기존 의미역 결정의 언어자원 부족 문제를 해소하기 위해 기존의 언어자원과의 결합을 통한 확장 방법을 제안하였으며, 확장된 의미역 결정을 통해 얻은 술어와 논항 정보를 기반으로 문서의 의미적 유사성을 측정하는 방법을 소개한다.

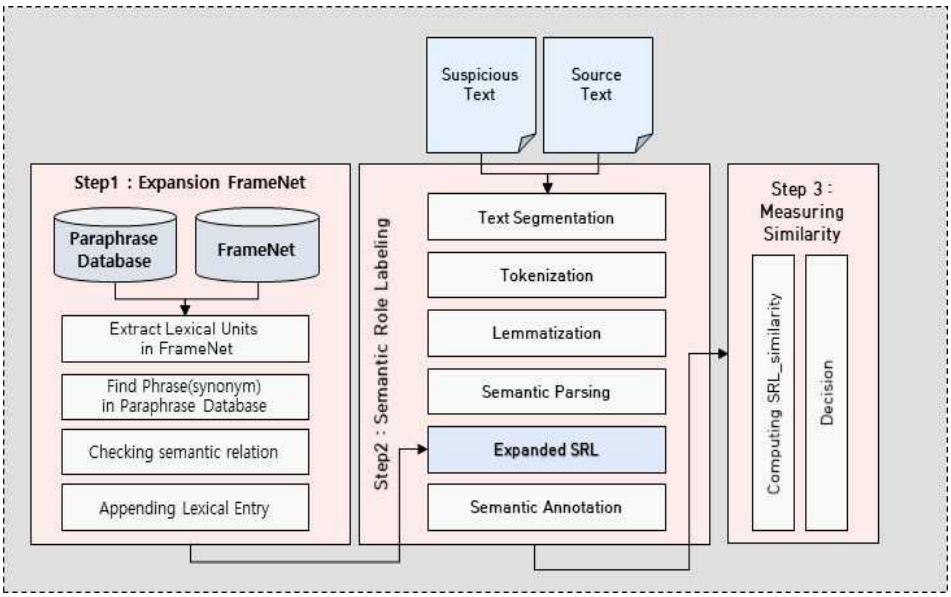

그림 1. 전체 구성도

그림 1은 본 연구의 전체 개요를 나타내고 있다. 본 연구는 문서 표절 검사 방법의 성능향상을 위한 기초 연구로써, 기존 FrameNet을 확장하는 과정과 확장된 FrameNet을 이용하여 두 문서에 의미역 결정을 수행하는 과정 그리고 의미역 결정을 통해 얻게 된 술어-논항 정보를 문서의 특징정보로 정의하여, 두 문서 간의 유사성을 측정하는 과정으로 나눈다.

문장의 의미역 결정을 통해 얻게 되는 술어-논항정보는 문장의 의미를 나타내는 중요한 성분이다. 예를 들어 “철수가 밥을 먹는다”라는 문장이 있다. “먹는다”라는 술어는 “철수가”라는 행위주 논항정보와 “밥을”이라는 대상 논항정보가 필요하다.

문장의 서술어가 하나의 문장을 이루기위해 필요한 정보를 논항정보라고 하며, 예문을 통해 확인한 것과 같이 특정 의미의 술어에는 필수적인 논항정보가 한정되기 때문에 유사한 의미의 문장에는 유사한 술어-논항정보가 나타나게 된다[15]. 본 논문에서는 이러한 이론을 기반으로 문장의 구조를 기반으로 추출한 의미 정보를 활용한 유사성 측정 방법에 대하여 제안한다.

본 논문의 구성은 다음과 같다.

본 장인 서론에 이어 2장에서는 본 연구의 이론적 배경인 문서의 유사성 측정 방법과 의미역 결정의 기본 개념과 관련 연구를 자세히 살펴본다. 그리고 연구 진행에 필요한 관련 연구들을 제시하여 3장부터 전개되는 연구 내용의 이해를 돕는다.

3장에서는 의미역 결정을 위한 기존 FrameNet의 범주 확장을 위한 방법에 대해 기술한다. 본 논문에서 FrameNet을 확장하기 위해 FrameNet의 구조와 Paraphrase Database의 구조에 대해서 소개하고, 확장을 위해 사용되는 정보와 확장 방법에 대해 언급한다. 또한, 확장된 FrameNet을 위한 학습 과정과 확장된 FrameNet의 성능 평가를 위해 기존의 FrameNet과 비교 평가를 수행한다.

4장에서는 본 논문에서 의미적 결정을 이용한 의미적 유사성 측정을 위해 문서의 전처리 과정과 의미역 결정 그리고 의미적 유사성 측정방법에 대해 기술한다. 또한, 본 연구에서 확장한 FrameNet을 이용한 문서의 의미역 결정에 대해 상세하게 설명하고, 확장한 FrameNet을 통해 문서 표절 검사를 위한 의미적 유사성 측정 방법을 제시한다.

5장에서는 실험 데이터를 이용하여 본 논문에서 제안한 방법론을 이용해 유사성 측정을 수행하고, 기존의 유사성 측정 방법과 비교평가를 수행한다. 또한, 확장된 FrameNet을 통해 수행한 유사성 검사에 대한 정확도와 재현율 평가 결과에 대해 기술하고, 결론과 향후 연구의 방향을 제시하며, 마무리한다.

Ⅱ. 관련연구

문서의 재사용과 표절문제는 인터넷과 스마트폰의 보급으로 인하여 정보 콘텐츠의 디지털화 과정에서 두드러지게 나타나고 있으며, 원본 문서 내 단어의 삽입, 삭제, 교체, 어순의 변경 등 복잡한 형태로 이뤄지고 있다[16]. 본 장에서는 연구의 배경이 되는 문서의 표절 유형과 기존의 문서 유사성 측정 방법, 문서의 의미적 분석 방법인 의미역 결정에 대한 관련 연구에 대해 살펴본다.

1. 문서 표절 유형

문서의 표절은 원본 문서의 내용을 그대로 복사해 사용하는 단순한 형태에서 표절 의혹을 벗어나기 위한 지능적이고 복잡한 형태로 변화하고 있기 때문에 표절 문서를 검출하기 위해서는 기존의 표절 유형에 대한 분석과 이해가 선행되어야 한다[17]. 표절 문서에서 발견되는 대표적인 표절 유형은 원문 복사(Copy and Paste), 단어치환(Replacement), 어순변경(Re-ordering), 다시쓰기(Paraphrasing)로 분류되며, 유형별 특징은 다음과 같다.

원문 복사 유형은 원문에서 존재하는 문장을 그대로 복제하여 표절 문서에서 사용한 유형이다. 학생들의 과제나 리포트, 블로그 등의 게시글을 작성시 수정없이 원문의 내용을 그대로 사용하는 형태로 나타나며, 새로운 연구나 논문에 자신의 이전 연구 내용을 다시 재사용하는 자가 표절의 경우도 원문 복사 유형에 포함된다. 또한, 출처가 있더라도 원문의 내용을 그대로 사용한 경우에도 원문 복사 유형으로 판단된다[18].

단어치환 유형은 문장의 전체적인 구조와 의미를 훼손하지 않는 범위에서 문장 내의 단어 또는 어절을 추가하거나 유사어로 교체하는 유형으로 문장의 주요 정보를 나타내는 단어를 유사한 의미를 갖는 다른 형태의 단어로 교체하는 것으로, 가장 많이 나타나는 표절 유형이다[19].

어순변경 유형은 문장 내의 단어의 위치를 재배열 하여 다른 문장인 것처럼 표절 검출을 회피하는 유형이다[18, 19]. 어순변경 유형은 문장의 전체적인 의미나 흐름을 그대로 사용하기 때문에 문장을 만들 때 저자의 독창적인 아이디어가 적용되었다고 볼 수 없어 표절 유형으로 분류된다. 그림 2는 어순 변경 유형을 보여주는 예시이다.

그림 2. 어순 변경 문장 예

다시쓰기 유형은 글의 내용의 구조와 단어를 새롭게 재구성하여, 같은 의미의 글을 전혀 다른 표현으로 나타낸 것을 나타낸 것이다[20]. 문장을 이루는 핵심적인 단어와 구조를 모두 변형하기 때문에 원문과의 유사성이 낮게 판단된다. 그림 3은 다시쓰기 문장의 예시이다.

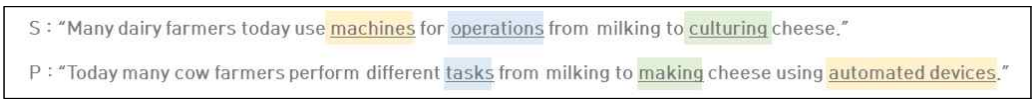

그림 3. 다시쓰기 문장 예

문서의 표절을 검출하기 위한 표절 검사 시스템은 두 개 또는 그 이상의 문서를 비교하는 것이 일반적이다. 표절 검사를 위한 검사 대상 문서가 입력되면, 비교 대상 문서 집단에서 검사 대상을 색인하고, 색인된 비교 문서들과 검사 대상 문서 간에 일정한 특징 정보를 기반으로 유사한 정도를 측정하고, 유사성을 판단하는 기준에 따라서 표절 여부를 판별한다. 문서의 유사성 검사 방법은 문서를 특징화하는 방법에 따라 분류할 수 있으며, 단어의 형태적 특징, 의미적 특징 그리고 구조적 특징 등 문서를 나타낼 수 있는 여러 특징 정보에 따라 문서 간의 유사성을 측정할 수 있다[21].

2. 문서 유사성 측정

본 장에서는 문서의 유사성 측정 연구를 위해 기존의 문서의 유사성 측정 방법에 대한 선행연구를 이해하고, 취약점을 분석하여 개선사항을 도출함으로써, 한계점을 극복할 수 있는 방안을 모색한다. 이런 내용들을 통해 전개되는 연구 내용의 이해를 높이고 문서의 유사성 측정 연구의 필요성과 중요성을 살펴본다.

문서의 유사성을 측정하는 연구는 정보 검색과 같이 컴퓨터가 자연어를 이해하고 처리하는 분야에서 큰 관심의 대상으로 연구되고 있다[22]. 자연어 처리 기술은 인간의 언어를 이해하고 사용자의 의도를 분석하기 위하여 더욱 다양하게 연구되고 있으며, 그 자체가 최종 목표인 기술이 아니라 하나의 서비스를 지능적으로 처리하기 위한 과정이다.

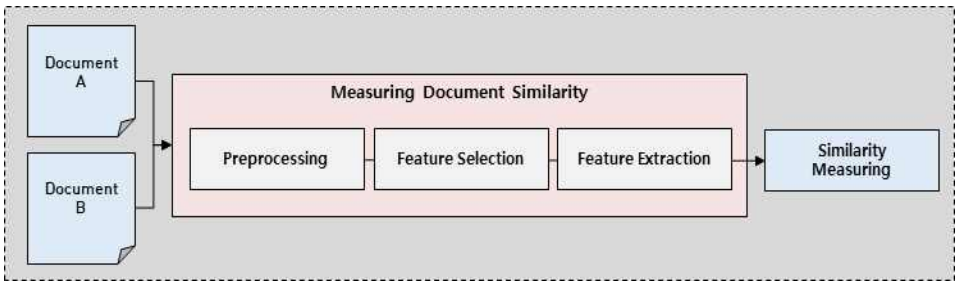

그림 4. 문서간 유사성 측정 과정

일반적인 문서간 유사성을 측정하는 과정은 그림 4와 같이 도식화할 수 있다[23]. 문서 유사성 측정 과정은 먼저 비교하고자 하는 두 문서에 대해 문서를 대표할 수 있는 특징(Feature)정보를 선정 및 추출하며, 추출한 정보를 결합하여 유사성을 산출하는 과정을 거친다[24]. 이 때, 특징 정보는 문자, 문자열, 단어, 구문 분석 정보 등으로 문서 정보를 대표적으로 나타낼 수 있도록 설정한다.

문서 표절 검사 시스템의 가장 핵심적인 부분인 문서 유사성을 측정하는 방법은 크게 형태적 유사성을 비교하는 방법과 의미적 유사성을 비교하는 방법으로 구분하며, 형태적 유사성을 비교하는 대표적인 방법으로 문장 내의 인접한 n개의 단어들을 추출하여 비교하는 n-gram방식과 문장내 부분 문자열(Substring)을 비교하는 방식, 벡터공간 상에 문서를 나타내고 벡터간의 거리 측정을 통해 유사성을 판단하는 벡터공간모델(VSM, Vector Space Model) 방식 등이 있다.

의미적 유사성을 비교하기 위한 방식은 단어 상호간의 의미적 관계를 계층 정보로 정의해 놓은 지식베이스(Knowledge Base)를 이용하여, 문서 내 형태적 유사성 측정 방식의 한계를 극복하고자 하였다. 이어지는 장에서 문서의 기존에 연구된 다양한 유사성 측정 방법에 대하여 상세하게 기술한다.

1) 문자 기반(Character-based) 유사성 측정

문자 기반 유사성 측정 방법은 주로 문장을 이루는 단어의 형태를 비교하는 형태적 유사성 측정 방법에 속하며, 현존하는 대부분의 표절 시스템에서 문자 기반 유사성 측정 방법을 기본으로 채택하고 있다. 문자 기반의 유사성 측정 방법은 단어의 순서를 고려하는 정확한 매칭 방법과 근사 매칭 방법으로 구분할 수 있다.

정확한 매칭 방법은 두 문자열의 모든 단어는 같은 순서로 일치해야 같은 유사성 측정 대상으로 판단한다. 대표적으로 n-gram 방식이 정확한 매칭 방법이다. n-gram 방식은 비교하고자 하는 두 문장에서 n-gram생성하고 추출하여, 전체에서 일치하는 n-gram의 비율로 문장의 유사 여부를 판단하는 방법이다[18].

다음은 n-gram 생성 및 추출 방법에 대하여 설명한다. 먼저 문서 내 문장들에 대해 공백이나 쉼표, 마침표 등의 구분자를 기준으로 어절을 나누고, 구분자를 기준으로 나뉜 어절들에 대해 n-gram을 생성한다. 예를 들면 “Hello”이라는 단어의 bi-gram은 “He”,“el”,“ll”,“lo”이며, tri-gram은 “Hel”,“ell”,“llo”이다. 어절을 이루는 음절 수가 n보다 크면, 하나의 어절이 여러 개의 n-gram으로 분리되고, n 보다 작으면, 하나의 n-gram으로 생성되기 때문에 철자 오류가 있더라도 유사성을 고려할 수 있다.

그러나 이 방식은 문서의 양이 많아질수록 기하급수적으로 많은 n-gram이 생성됨에 따라 저장 공간과 계산량이 증가하며, 문서 내의 불용어 때문에 전혀 다른 문장임에도 불구하고 일치하다고 판단하는 부적합 현상이 생길 수 있다[18].

근사 매칭 방법은 정확한 매칭 방법의 성능 개선을 위한 방법으로 두 문장 간의 차이를 허용하여 문자열을 비교하는 방식으로, 전체 문장 중 두 문장 간에 일치하는 부분의 비율로로 문장의 유사여부를 판단하는 방식[25]이다. 대표적인 근사 매칭 방법에는 부분문자열을 이용한 LCS(Longest Common Subsequence)알고리즘과 GST(Greedy String Tiling)알고리즘이 있다.

LCS 알고리즘은 두 문장을 순차적으로 비교하여, 제일 긴 Subsequence(연속되지 않은 부분 문자열)을 값으로 유사성을 판단하는 방법으로 어순의 변형이 이뤄진 경우 Subsequence로 고려되지 않는다는 단점이 있다.

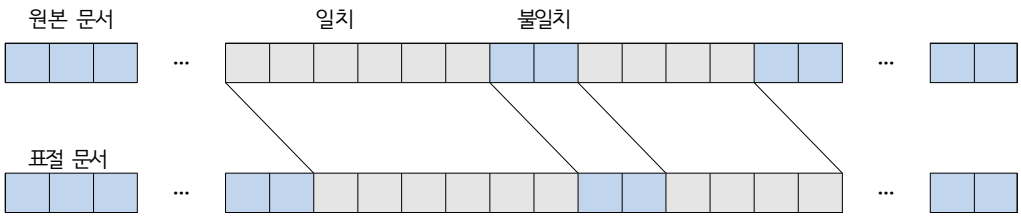

그림 5. 부분 문자열을 이용한 문서 유사성 측정

GST 알고리즘은 최소 매칭 길이를 설정하고, 이에 해당되는 Substring(연속된 부분 문자열)의 길이의 합으로 유사성을 판단한다. 이 때, 매칭된 문자열은 순서가 바뀌어도 유사성 측정의 대상이 된다는 점에서 정확한 매칭 방법과 차이점이 있다.

문자열 비교를 통한 문서의 유사성 측정 방법은 기존의 표절 시스템에서 가장 대표적으로 사용되는 방식으로 원문의 변형이 많이 이루어지지 않은 경우에서는 높은 신뢰도를 가지지만, 변형이 많이 이뤄진 문장의 경우 성능이 견고하지 못하다는 문제점이 있다.

2) 벡터 공간 모델 기반 (Vector Space Model-based) 유사성 측정

벡터 공간 모델 방식은 문서를 이루고 있는 단어를 추출하여 벡터 공간 상의 벡터로 표현하고 벡터 간의 거리로 유사성을 계산하여, 문서와 문서 사이의 유사성을 측정하는 방식이다[26].

$$d_j = (w_{1,j}, w_{2,j}, w_{3,j}, \cdots, w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, w_{3,q}, \cdots, w_{t,q}) \quad (1)$$

문서를 이루고 있는 각 단어를 하나의 차원으로 정의하고, 만약 문서 내 단어가 포함되지 않는 경우에는 0의 값으로 표현하고, 단어가 포함되지 않는 경우에는 0이 아닌 가중치의 값으로 표현하게 된다. 가중치를 부여하여, 문서를 n차원의 벡터 값으로 나타낸 후, 벡터 간의 거리를 측정함으로써, 문서와 문서 간의 유사성을 측정한다.

벡터 공간 모델 방식에서 유사도 계산 방법은 다이스 유사도, 자카드 유사도, 코사인 유사도 등이 이용된다[27]. 코사인 유사도 측정 방법의 경우 일반적으로 문서 유사도 계산시 가장 많이 쓰이는 방법으로 각각의 문서에 해당하는 열을 벡터로 놓고 두 벡터를 내적할 때, 두 벡터가 이루는 각도를 계산함으로써 유사성을 판단한다.

$$\cos(d_1, d_2) = \frac{\overrightarrow{d_1} \cdot \overrightarrow{d_2}}{|\overrightarrow{d_1}| \cdot |\overrightarrow{d_2}|} \quad (2)$$

벡터 공간 모델에서 유사도 계산 방식은 비교적 단순한 계산에 비해 정규화 된 값을 얻을 수 있다는 장점을 가지고 있지만 단어가 정확히 일치할 때에만 유사성을 비교하는 대상에 포함되며, 같은 의미의 다른 단어로 교체되었을 경우 유사도 계산에 고려할 수 없다.

3) 의미 기반 (Semantic-based) 유사성 측정

문서의 유사성을 측정할 때 앞서 소개했던 n-gram방식과 부분 문자열 기반 유사성 측정 방식 그리고 벡터 공간 모델 방식은 문서의 형태적 유사성 측정 방법에 속한다. 문장이 정렬된 단어 그룹이라고 정의한다면, 두 문장은 같은 의미를 가질 수 있지만, 단어의 순서가 다르거나 같은 의미의 다른 단어를 사용하여 표현할 수 있으며, 이러한 경우 형태적 유사성 측정 방법으로는 유사한 의미의 문장으로 측정할 수 없게 된다.

이러한 문제점을 개선하기 위하여 지식베이스(Knowledge Base)에 정의된 개념간의 의미관계를 이용하여 지식정보를 활용한 유사성 측정 연구가 수행되었고, 다양한 연구를 통해 활용가능성과 중요성을 입증하였다. 본 장에서는 대표적인 영어의 지식베이스인 워드넷을 기반으로 유사성을 측정하는 방법[28]과 잠재 의미 분석(LSA, Latent Semantic Analysis)방법에 대하여 소개한다[29].

워드넷은 프린스턴 대학에서 구축한 대용량 어휘 데이터베이스이다. 그림 6에서 보이는 것처럼 워드넷은 개념의 정의문과 개념간의 계층관계를 정의하여, 사용자가 열람하고 이를 활용할 수 있도록 제공하고 있으며, 이러한 단어 간 계층정보는 두 단어 간의 의미가 얼만큼 밀접한가를 측정하는데 매우 중요한 척도로 사용된다[30].

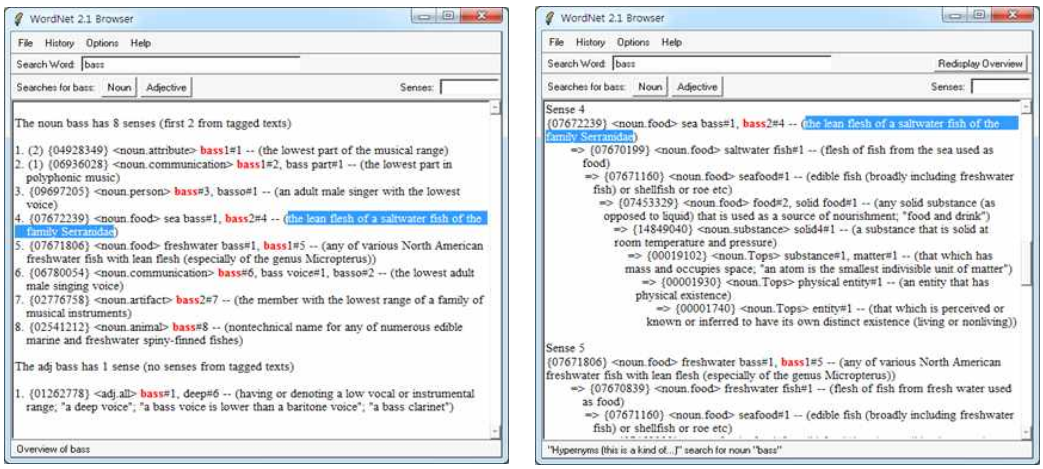

그림 6. 워드넷에 정의된 계층과 관계 정보

워드넷은 단어의 관계를 기술하기 위해 계층적인 형태를 이루고 있으며, 상위 개념으로 올라갈수록 포괄적인 의미정보를 나타내고, 하위 개념으로 내려갈수록 구체적인 의미정보를 나타낸다. 이러한 계층정보를 사용하여 의미적 유사성 평가 방법은 다양하게 연구되었으며, 본 논문의 관련연구에서는 워드넷 계층정보의 간선기반 유사성 측정방식과 정보량기반 유사성 측정방식으로 구분하여 소개한다[31].

워드넷에 정의된 계층정보를 이용한 유사성 평가 방법은 정의된 두 개념 사이의 최소거리를 계산하거나, 깊이정보를 사용하여 측정한다[31]. 수식 3은 간선기반 유사성 측정방식으로 두 개념 간의 최소거리와 계층구조의 깊이를 고려하여 유사성을 측정한다. $length(c_1, c_2)$는 두 개념(c_1, c_2)을 연결하는 최소 간선의 개수이고, $Depth$는 두 개념을 포함된 전체 계층구조의 깊이이며, 깊이가 깊을수록 유사성이 높게 나타난다.

$$Similarity_{link}(c_1,c_2) = -\log\frac{length(c_1,c_2)}{2\,Depth} \quad (3)$$

그러나 간선기반 유사성 측정방식은 동일한 상위개념을 갖는 개념들에 대해서만 고려되기 때문에 비교에 있어서는 사실상 의미가 없다고 볼 수가 있으며, 단순히 최단 거리가 짧은 두 개념 간의 유사도가 더 높은 결과를 보인다. 또한 어떤 인접한 개념사이의 거리는 동등하지가 않기 때문에 두 개념 간의 노드를 연결하는 간선에 가중치를 주어 정보량을 고려하는 방식이 필요하다.

정보량을 고려한 방식은 두 개념이 얼마나 많은 정보를 공유하는지의 정도를 가지고 의미적 유사성을 측정하는 방법으로 Resnik[32]에 의해 제안되었다. 정보량을 고려한 의미적 유사성 측정 방식에서 워드넷의 개념(c)에 대해 IC(Information Content)로 정보량을 나타내고, 수식 4를 통해 정량화한다. 이 때, $P(c)$는 개념이 나타날 확률값을 나타낸다.

$$IC(c) = \log^{-1}P(c) \quad (4)$$

정보량을 고려한 유사성 측정은 수식 5와 같이 정의되며, 포섭자(Subsumer)는 두 개념(c_1, c_2)을 모두 포함하는 집합이다. 두 개념을 모두 포함하는 집합이 갖는 정보량을 계산하여 두 개념간의 유사성을 측정하게 된다.

$$Similarity_{IC}(c_1,c_2) = \max([IC(c)]),\ c \in subsumer(c_1,c_2) \quad (5)$$

워드넷 이외에도 지식베이스를 이용한 문서의 의미적 유사성 측정 방법인 잠재 의미 분석 (LSA, Latent Semantic Analysis)방식은 문장의 문맥을 고려하기 위해 문장 내 단어의 의미관계를 분석하여 유사성을 측정하고, 의미관계를 분석하기 위해 통계정보와 선형대수를 이용한다[29].

LSA는 각 문서를 단어벡터와 문서벡터로 이루어진 행렬로 구성하고, 특이값 분해(SVD, Singular Vector Decomposition)를 통해 형성된 행렬을 기반으로 단어와 문맥간의 내재적인 의미를 분석하는 방법이다. LSA는 문서에 나타난 단어간 의미적 관계를 고려한다는 장점을 가지고 있는 반면, 의미적 유사성을 측정할 때 문서에 나타난 단어의 정보만을 이용하기 때문에 유사성 측정에 사용하게되는 정보의 양이 한정적이므로 정확도가 낮아지는 단점이 있다[29].

4) 구문정보 기반 (Syntax-based) 유사성 측정

구문 정보 기반 유사성 측정 방식은 문장을 이루는 단어 정보와 의미적 정보를 기반으로 측정하는 다른 유사성 측정 방식과는 다르게 문장의 구성 성분을 특징으로 선택하여 문맥상의 유사성을 비교하는 방식이다[33].

문장에서 주어와 술어의 위치를 바꾸어 구조적 정보가 조금 수정된 경우가 있다고 할 때 유사성 측정 결과에 큰 차이를 가져올 수 있다. 예를 들어, “A cat chases a mouse” 문장의 구조적 정보를 수정하여 “A mouse is chased by a cat”라는 문장으로 변형하였을 때, 기존의 형태적 유사성으로 측정하였을 경우 “A mouse chases a cat”. 이라는 문장보다 더 낮은 유사성으로 측정된다. 하지만, 구문정보 기반의 유사성을 측정하는 방법은 문장을 이루는 구성 성분을 추출하여 비교하게 되며, 문장이 변형되더라도 동일한 구성 성분으로 유사한 의미를 나타내는 문장일 경우 유사성을 높게 측정할 수 있는 문맥 상의 유사성을 비교하는 방식이다.

표 1은 앞서 살펴본 기존 문서의 유사성 측정방법들의 표절 유형별 대응능력에 대한 비교를 나타낸다. 를 기존의 연구를 바탕으로 분석한 결과, 형태적 유사성 측정 방법의 경우 다시쓰기 문장과 같이 변형이 많이 이뤄진 문장에 대해 유사성 측정을 수행할 수 없다는 문제점을 가지고 있다. 이에 본 논문에서는 이를 개선하기 위한 연구를 수행하고자 한다.

표 1. 기존 문서 유사성 측정 방법 비교

	문서의 표절 유형				
	복사	삽입,삭제	어순변경	단어교체	다시쓰기
n-gram, 문자열	○	△	×	×	×
벡터공간모델	○	○	○	×	×
의미 정보	○	○	○	○	△
구문 정보	○	○	○	○	○

3. 의미역 결정

의미역 결정(Semantic role labeling)이란 구문 분석을 통해 문장 성분의 의미역을 결정하는 것을 말한다[13, 14]. 의미역 결정은 자연어 문장의 서술어를 중심으로 하여 서술어가 하나의 문장을 완성하기 위해 필수적으로 요구되는 논항(agument) 정보를 식별하고, 서술어와 각 논항들 간의 의미 관계를 결정하는 것이다. 의미역 논항들을 행위주, 경험주, 대상 등으로 의미역 정보를 맵핑함으로써, 의미역 결정을 수행한다[34]. 문장의 각 성분이 다른 구조로 배열된다 하더라도 같은 의미역을 가질 수 있으며, 같은 구조로 배열된 구문이 다른 의미역을 지닐 수 있다[15].

그림 7. 의미역 결정 예

그림 7은 문장의 의미역 결정을 수행한 예시이다. 문장이 의미를 전달하기 위해서는 문장 내의 단어들은 각자의 의미 역할을 적절하게 담당해야 한다. 예를들어, “철수가 책을 산다”라는 문장에서 ‘철수’는 행위의 주체 역할을, ‘책’은 행위의 대상이다.

그림 8은 의미역 결정을 수행 과정을 보여주고 있으며, 이 과정은 서술어 인식(PI: Predicate Identification)단계, 서술어 분류(PC: Predicate Classification)단계, 논항 인식(AI: Argument Identification)단계 그리고 논항 분류(AC: Argument Classification)단계인 총 4단계로 구분된다[35].

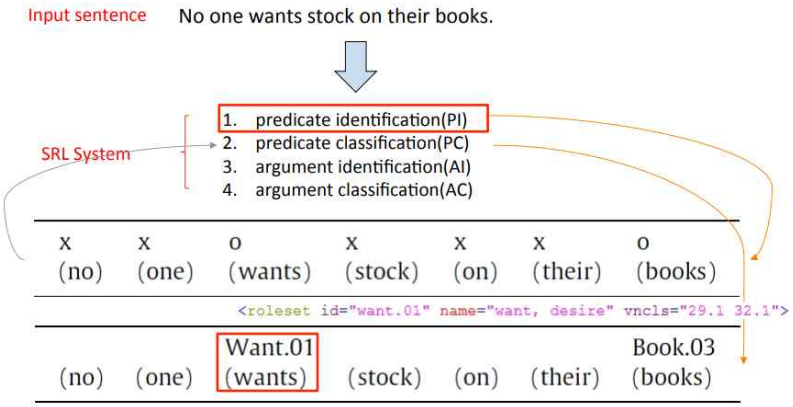

그림 8. 의미역 결정 수행 과정

입력 문장에 대해 구문 분석을 통해 서술어를 인식하는 서술어 인식 단계와 서술어의 중의성을 해소하는 서술어 분류단계가 선행되며, 문장 내 특정 의미를 갖는 서술어가 하나의 문장을 이루기 위해 필요한 정보를 인식하는 논항 인식단계와 인식된 논항의 의미역을 결정하는 논항 분류단계로 수행된다[35].

의미역 결정 방법은 사전에 정의된 정보를 기반으로한 방법과 말뭉치에 기반한 방법으로 나눌 수 있다. 사전을 이용하는 방법은 분석하고자 하는 문장에 나타난 술어-논항정보를 사전에서 찾아 찾아 의미역을 결정하는 방법[36]으로 문장의 술어-논항 정보와 사전 정보 사이의 유사도 계산을 통해 의미역을 결정하기 때문에 높은 정확률을 보이지만, 사전의 구축이 어렵고 사전에 기술되지 않는 정보는 고려할 수 없는 문제가 있다[36]. 말뭉치를 이용하는 방법은 말뭉치에 의미역 정보를 태깅하여 의미역 결정 학습 데이터를 생성한 후, 기계학습(machine learning)을 적용하여 입력 문장의 의미역을 결정하는 방법이다[37].

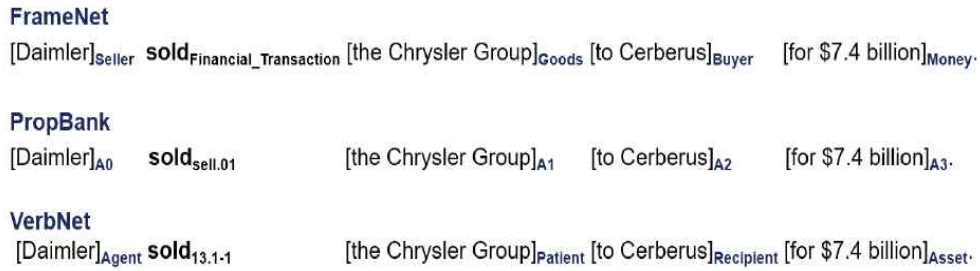

그림 9. FrameNet, PropBank, VerbNet 비교

영어의 대표적인 의미역 결정을 위한 언어 자원으로 Propbank과 FrameNet이 있다. 이들은 의미역을 부착한 말뭉치 형태로 제공하고 있다[38]. Propbank(Martha Palmer, 2005)는 Penn Tree bank의 통사 구조 분석 자료에 논항별로 의미역을 추가하는 방식으로 구축되었다. Propbank는 서술어와 논항관계 구축을 위해 서술어로써 동사만을 고려한다[39].

반면, FrameNet은 BNC(British natioanl Corpus), ANC(Amerian National Corpus) 등의 말뭉치를 대상으로 술어-논항정보를 구축하였으며, Frame이라는 분류개념을 적용하여, 동사 뿐만아니라 명사, 부사, 형용사 등의 논항관계를 고려하여 프레임-논항 정보체계로 구축하였다[40, 41]. FrameNet의 경우 동사 이외의 품사에 대한 어휘 정보를 고려하였고, 프레임과 논항에 대해 의미 정보를 부착하여 기존의 propbank보다 많은 의미적 정보를 포함하고 있다는 장점을 가지고 있어 FrameNet의 의미 정보는 다양한 연구에 활용되고 있다.

하지만, FrameNet은 전적으로 언어 전문가에 의해 수작업으로 직접 구축되었기 때문에, 상당수의 단어들이 FrameNet에 나타나지 않는 경우가 많다[40, 41]. 예를 들어, blatant라는 단어는 Obviousness 프레임 정보를 나타내지만 현재 FrameNet의 프레임 내의 단어 목록인 Lexical Units에 정의되어 있지 않다. 이 발견은 Palmer and Sporleder[42]의 연구에서 FrameNet의 자원의 한계에 대해 언급하고 있는 것과 일치한다. 이러한 낮은 어휘 범위는 FrameNet이 실제 응용 프로그램에 활용하는데 어려움을 갖게된다.

따라서, FrameNet의 자원의 한계를 극복하려는 시도가 계속되고 있다. FrameNet과 워드넷의 단어의 수반 관계를 이용해서 확장하려는 연구[43]가 있었으며, 기존의 FrameNet을 온톨로지로 구축하여 다른 언어들과 매핑하는 연구도 있었다[44].

본 논문에서는 FrameNet의 자원의 범위를 확장하는 연구에 대하여 수행하고자하며, FrameNet을 문서의 의미적 유사성 측정에 활용하기 위해 언어 자원과 확장하는 방법을 제안한다. 또한, 본 연구에서는 문서의 유사성 측정을 위해 의미역 결정을 적용하는 방법에 주안점을 두었다. 문장을 이루는 구조인 술어-논항정보는 하나의 문장 내 두 개체 간의 연관관계를 표현하고 있기 때문에 문장별 술어-논항정보를 고려하여 문서의 의미적 유사성을 측정한다면, 문서 유사성 측정의 정확성 향상이 가능하고, 같은 의미를 다른 단어를 사용하여 다시쓰기(Paraphrasing)문장에 대해서도 의미적 유사성을 측정해 낼 수 있을것으로 사료된다.

III. 언어자원을 이용한 의미역 결정 확장

1. FrameNet 확장 방법

본 논문에서는 의미역 결정을 수행함으로써 문장의 구성요소를 분석하고, 구성요소를 특징으로 한 문장의 유사성 비교를 통하여 문서의 표절 검사를 수행한다. 이를 위해 기존의 의미역 결정을 위한 대표적인 언어자원인 FrameNet을 기반으로 Paraphrase Database와 결합을 통해 다른 단어로 표현된 같은 의미의 문장을 찾아내어 문서의 유사성 측정 결과의 정확성 향상을 도모하고자한다. 본 장에서는 FrameNet을 기반으로 Paraphrase Database와의 결합을 통해 확장하는 방법에 대해 소개한다.

1) FrameNet 구조

FrameNet은 전문가가 구축한 프레임-논항정보에 관한 언어 자원으로 Frame(틀)이라는 인지 모형을 사용하여 단어에 대한 지식을 의미 관계로 구성하는 틀의미론(Frame Semantics)을 중심으로 구축되었다[37]. FrameNet은 전문가에 의해 수작업으로 구축되어진 만큼 높은 정확률을 가지고 있으나 데이터의 양이 자동으로 구축된 언어자원에 비해 현저하게 낮아 도메인 의존적인 한계를 가지고 있다. 이러한 FrameNet의 자원의 한계를 극복하려는 다양한 시도가 계속되어지고 있다[37, 43, 44].

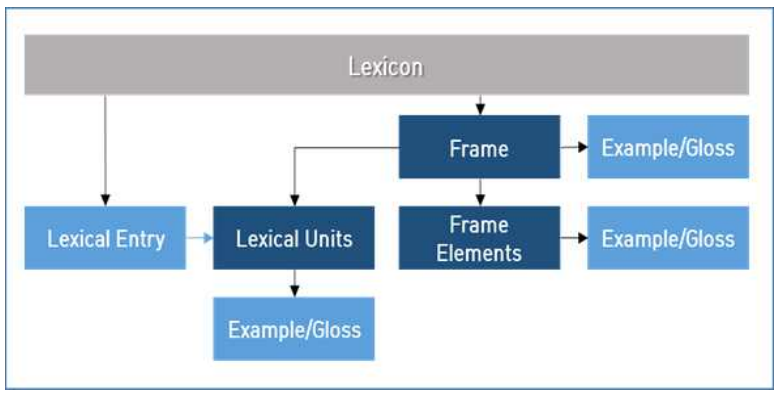

그림 10. FrameNet 구조

그림 10은 FrameNet의 구조도이다. FrameNet은 Frame과 Frame Elements 그리고 Lexical Entry로 구분지을 수 있다.

■ Frame은 단어의미 그룹으로 특정 이벤트이나 상태에 대한 서술 정보를 담고 있다. 예를 들어, Activity_finish 프레임은 행위를 끝낸다는 의미를 갖고 있는 프레임이며, Completion(명사), Complete(동사) 등의 모든 품사가 프레임으로 정의 될 수 있다.

■ 해당 Frame의 논항 정보는 FrameNet에서는 Frame Elements라 불리며, 프레임과 의미 관계에 있는 논항 정보 즉, 의미역 정보를 나타낸다. 논항 정보는 문장을 구성하는데 중요한 역할을 하는 필수 논항인지 아닌지에 따라 core role과 non-core role로 구분되어진다. 예를들어, 행위를 나타내는 Activity 프레임의 경우 core role으로 Agent(행위자), Activity(행위)를 갖으며, non-core role으로 Place(장소), Duration(기간), Manner(태도) 등의 기타 요소를 갖을 수 있다.

■ 각 Frame에 속하는 어휘 정보를 Lexical Units으로 표기하고, Lexical Units의 리스트를 Lexical Entry라고 한다. Lexical Entry 내에는 Lexical Units에 대한 lemma(단어의 원형), Part-of-Speech(품사)정보 등의 어휘 정보를 정의하고 있다. 그림 11은 FrameNet의 프레임-논항 정보의 예시로써, Frame에 대한 정의문과 Frame에 속하는 Lexical Units(술어정보) 그리고 Frame Elements(논항정보)를 보여주고 있다.

Cutting

Definition: An [Agent] cuts an [Item] into [Pieces] using an [Instrument] (which may or may not be expressed).

Core frame elements:

Agent	The [Agent] is the person cutting the [Item] into [Pieces].
Item	The item which is being cut into [Pieces].
Pieces	The [Pieces] are the parts of the original [Item] which are the result of the slicing.

Non-core frame elements:

Instrument	The [Instrument] with which the [Item] is being cut into [Pieces].
Manner	[Manner] in which the [Item] is being cut into [Pieces].
Result	The [Result] of the [Item] being sliced into [Pieces]. (extrathematic)
In addition: Means, Purpose, Place, Time	

Lexical units: *carve, chop, cube, cut, dice, fillet, mince, pare, slice*

그림 11. FrameNet 프레임-논항정보 예

2) Paraphrase Database 구조

'Para-'라는 접두어는 그리스어로 beside, near, similar의 뜻을 갖으며, paraphrase는 본래의 말(phrase)을 유사어로 바꾸어 말하는 것을 말한다[45]. Ganitkevitch et al. (2013)은 Paraphrase Database(PPDB)라고하는 단어의 Paraphrase 정보를 담고 있는 언어 자원을 구축하였다[46]. Paraphrase Database 2.0은 paraphrase 문장을 분석하여, paraphrase가 이뤄질 때의 단어들의 조건부 확률을 계산하여 단어와 단어 쌍을 구축한 언어 자원이다. 16개의 언어에 대하여 paraphrase 단어 정보를 제공하고 있어 기계 번역에 주로 활용되고 있는 언어자원이다[46, 47].

표 2. Parapharase Database 구조

LHS	PHRASE	PARAPHRASE	(FEATURE=VALUE)	ENTAILMENT
[NN]	legalisation	legalization	PPDB2.0Score=4.80262	Equivalence
[JJ]	south-east	south-eastern	PPDB2.0Score=4.79356	Equivalence
[JJ]	israel-palestinian	israeli-palestinian	PPDB2.0Score=4.79011	Equivalence
[NNS]	handled	handles	PPDB2.0Score=4.78854	Equivalence
[NNS]	beginnings	begins	PPDB2.0Score=4.70006	Entailment
[VBZ]	verify	investigate	PPDB2.0Score=3.75207	Independent
[VBD]	trip	visit	PPDB2.0Score=3.67524	Entailment
[VBN]	support	assist	PPDB2.0Score=3.57730	Independent
[VBZ]	reform	change	PPDB2.0Score=3.53699	Entailment

Paraphrase Database의 구조는 표 2와 같다. LHS는 통사 범주(Syntactic Category) 정보를 나타내며, 문법성을 잃지 않고 서로 대신해서 사용할 수 있는 표현으로 품사(Part-of-speech)라고 할 수 있다. Phrase는 단어 또는 구를 표현하며, Paraphrase는 같은 의미의 다른 표현으로 사용되는 단어 또는 구를 나타낸다. Feature value는 Phrase가 Paraphrase로 사용되는 조건부 확률을 나타낸다. entailment는 phrase-paraphrase 단어 간의 관계를 기술하고 있는 정보이다.

Paraphrase Database에서 사용되는 관계는 Equivalence, Entailment, Exclusion, Other relation, Unrelated가 있다[48].

Equivalence는 동등한 의미로 동의어를 나타내며 Entailment는 수반어로써 이는 대체적으로 상하위어 관계에 속한다. A단어로 인해 B가 야기될 때 A와 B는 수반어 관계에 있다고 말한다. Other relation에는 특정 지을 수 없는 기타 관계를 나타내고 있고, Exclusion은 서로 배제되는 공통부분이 없는 관계를 의미한다. 마지막으로 Unrelated관계는 연관관계가 없음을 나타낸다. 본 논문의 FrameNet의 확장을 위해서는 Equivalence, Entailment, other relation 관계를 갖는 단어들을 고려하여, 기존의 FrameNet의 Lexical Units와 비교를 통해 확장한다.

3) Paraphrase Database를 이용한 FrameNet 확장

본 논문에서 FrameNet 확장은 다음과 같은 순서를 따른다.

1. FrameNet xml로부터 Frame별 Lexical Units 추출을 통한 Lexical Entry 생성
2. Lexical Entry 내 Lexical Units과 Paraphrase Database의 Phrase 비교
3. 일치하는 Phrase에 대해 Paraphrase Database의 관계정보를 확인하여, Equivalence, Entailment, other relation를 갖는 Paraphrase를 Lexical Units 확장을 위한 Lexical Entry에 추가함.

FrameNet을 확장하기 위해 FrameNet의 Frame이 갖는 Lexical Entry 내의 Lexical Units과 Paraphrase Database의 Phrase와 비교를 통해 일치하는 Phrase의 Paraphrase를 추출하여 FrameNet의 Lexical Units의 확장 후보로 선정한다. 예를 들어, abandonment라는 Frame을 확장하고자 할 때, 해당 Frame의 Lexical Entry 내의 Lexical Units은 abandon, abandoned, abadonment, forget, laeve이며, 확장을 위해 모든 Lexical Units을 Paraphrase Database 내의 phase와 비교하여, 일치하는 단어를 확장을 위한 후보로 선정한다. 그 다음 Paraphrase Database 내의 단어 간의 의미 관계를 확인하여, 동의어, 상하위어, 반의어 등의 관계를 갖는 단어를 Lexical Entry에 추가하여 확장한다.

표 3. FrameNet Lexical Units과 Paraphrase Database 결합

Frame	Lexical Units = phrase	paraphrase
Abandonment	abandon.v	ceases, relinquishes, surrenders, renounces, leaves, forgoes, quits, drops
	abandoned.a	withdrew, abdicated, discontinued, aborted, relinquished, surrendered, surrender, renounced, renounced, left, abrogated, discarded
	abandonment.n	discontinuations, discontinuances departures
	forget.v	messed, left, omitted
	leave.v	authorizations

표 4는 확장된 FrameNet의 일부를 나타내고 있다. 본 논문에서 FrameNet의 확장의 범위는 Frame 정보의 Lexical Entry에만 한정하였으며, 추가된 Lexical Units의 모든 정보는 해당되는 Frame의 Frame Elements를 상속한다.

표 4. 확장된 FrameNet의 Lexical Entry

ID	Frame	new_Lexical Units
2031	Abandonment	abandoned.a , abandonment.n , abdicate.v , discard.v , discontinue.v, drop.v , forget.v , forgo.v , leave.v , neglect.v , quit.v , relinquish.v , renounce.v , strand.v , surrender.v
1903	Accuracy	accurate.a , accurately.adv , careful.a, clarify.a , clear.a , detailed.a erroneous.a , exact.a , false.a imprecise.a , inaccuracy.n , inaccurate.a , inaccurately.adv , incorrect.a , inexact.a , off.prep , on.prep , prcis.a , precise.a , precision.n, reliable.a , right.a , specific.a , true.a , unclear.a , wrong.a
403	Achieving_first	coinage.n , concoct.v , discover.v , discoverer.n , discovery.n, fabrication.n , forerunner.n , frontier.n , invent.v , invention.n , inventive.n , inventiveness.n , inventor.n , originate.v , originator.n , pioneer.n , pioneer.v , pioneering.a, vanguard.n
404	Word_relations	antonym.n, collocate.n, collocate.v, contrary.n, holonym.n, homograph.n, homonym.n, homophone.n, homophonous.a, hypernym.n, hyponym.n, meronym.n, synonym.n, synonymous.a
1892	Work	peg away.v, plug away.v, action.n, action.v, active.v, activity.n, activity.v, business.n, business.v, co-operated.v, collaborate.n, collaborate.v, collaboration.n, collaboration.v, cooperate.n, cooperate.v, discussion.n, discussion.v, do.v, duty.n, duty.v, effort.n, effort.v, employ.v, factor.v, force.n, force.v, group.v, job.n, job.v, labor.n, labor.v, occupational.n, occupational.v, operation.n, operation.v, party.v, professional.n, professional.v, serve.v, strive.v, task.n, task.v, travail.n, travail.v, walk.n, walk.v, work.n, work.v, worker.v, working-level.v

기존 배포되고 있는 FrameNet 1.5 에서는 1,019개의 Frame과, 9,633개의 FrameElements, 그리고 11,942개의 Lexical Units을 가지고 있다[41].

본 논문에서는 기존 FrameNet의 Lexical Units과 Paraphrase Database를 이용하여 FrameNet 확장을 통해 제공하는 1,019개의 Frame 중 894개의 Frame에 대해서 확장을 수행하였으며, 그 결과 11,942개의 Lexical Units을 31,336개로 확장할 수 있었다.

본 논문에서는 Frame과 Frame Elements는 추가적으로 확장하지 않고, 기존의 Frame이 갖는 Frame Elements를 상속하여 의미역 결정을 수행하는 것으로 가정하였다. 확장된 FrameNet을 기반으로 의미역 결정을 수행하기 위해서는 확장된 FrameNet 정보를 학습데이터로 구축하여, 학습을 수행한 후 의미역 결정을 수행해야한다. FrameNet의 구조가 서로 연결되어 있고, 모든 Frame, Lexcial Units, Frame Elements 정보가 계층정보로 구성되어 있기 때문에 FrameNet을 새롭게 구축하는데는 어려움이 있다.

2. 확장된 FrameNet 학습 및 성능평가

본 논문에서는 확장된 FrameNet을 위한 학습데이터 구축을 위해 FrameNet에서 제공하는 말뭉치인 FrameNet Fulltext에 기존의 Lexical Units과 새롭게 생성된 Lexical Units을 교체하는 방식으로 Fulltext를 생성하고, 이 정보를 FrameNet 확장을 위한 학습 데이터로 사용한다. 표 5은 확장된 FrameNet의 Fulltext의 일부이다.

표 5. 확장된 FrameNet의 Fulltext

LU	new_LU	FULLTEXT
task.n	work.n	iran has begun preliminary **[[work]]** to install thousands of centrifuges at its uranium enrichment facility as a dispute roiled over whether the international atomic energy agency was receiving the access it wants to monitor the site , wire services reported today .
rocket.n	missile.n	china has provided technology and expertise to the **[[missile]]** programs of several countries , including pakistan , iran , and north korea .

본 논문에서 학습 데이터의 학습과정과 의미역결정을 수행하기위해 자동 의미역결정 시스템인 SEMAFOR를 사용하였다. SEMAFOR는 FrameNet이론을 기반으로 구현된 Semantic Parser이면서, 학습기능과 의미역 태깅을 수행할 수 있다[49]. SEMAFOR의 학습모델은 Frame Identification 모델과 Argument Identification 모델로 나뉘며, 해당 모델을 학습을 위한 학습 데이터를 구축하였다[35].

Frame Identification 모델은 FrameNet에서 의미역 결정을 수행하기 위한 문장 내 프레임을 인식하는 단계로써, Frame Identification 모델을 위한 데이터 포맷 (*.all.lemma.tag)은 표 6과 같다.

표 6. Frame Identification(*.all.lemma.tag) 학습 데이터

# token	token	pos-tag	parsing-tag	lemma
8	That is the way the system work .	DT VBZ DT NN DT NN VBZ .	nsubj cop det null det nsubj rcmod punct	that be the way the system work .

Fulltext의 문장 한 개는 표 6과 같이 형태로 정형화된 정보로 분석하여 표현한다. 첫 번째 열은 문장을 이루는 단어의 수(n)이다. 두 번째 열은 (n)개에 해당하는 단어가 나열된다. 세 번째 열은 (n)개의 단어에 pos-tag정보이고, 네 번째 열은 구문 분석 태그가 나열된다. 마지막으로 n개의 단어에 대한 원형정보를 나타낸다.

표 7. Argument Identification(*.all.frame.elements) 학습 데이터

# token	Frame	Lexical Units	LU span	lemma	sentence ID	role span
4	Means	way.n	3	way	131	Agent(0) Means (4:6) purpose(3)
2	System	system.n	5	system	131	Complex (5)

Argument Identification 모델은 FrameNet에서 의미역 결정을 수행하기 위한 술어와 의미적 관계가 있는 논항을 인식하는 단계로써, 학습데이터를 통해 결정된다. Argument Identification 모델을 위한 데이터 포맷은 표 7과 같다. 각 행은 하나의 술어-논항정보를 나타낸다. 첫 번째 열은 해당 술어와 논항의 개수의 합을 나타낸다.

예를 들어, 하나의 서술어에 3개의 논항이 있을 경우 값이 4가 된다. 두 번째 열은 해당 술어-논항정보가 속한 프레임을 나타내고, 세 번째 열은 Lexical Units(술어)을 나타낸다. 네 번째와 일곱번째 열은 문장 내에서 단어의 위치(span)를 나타낸다. 이는 구문 분석을 통해 얻을 수 있다. 다섯 번째 열의 lemma는 Lexical Units의 원형을 나타내고, 여섯 번째 열은 해당 술어-논항정보가 추출된 Fulltext 문장 번호를 의미한다.

본 논문에서 제안한 확장된 FrameNet의 언어의 범주가 개선됨을 증명하기 위해 FrameNet 의미역 정보가 기부착된 MASC I (Manually Annotated Sub-Corpus First Release) 말뭉치를 이용하여 의미역 결정 성능을 확인하였다[50]. MASC I 말뭉치를 구성하는 Token의 개수, Frame의 수 그리고 Frame Elements의 개수를 추출한 결과는 표 8과 같다. 이를 본 연구를 통해 확장한 FrameNet과 비교를 위한 정답 집단으로 선정하였다.

본 연구를 통해 기존의 FrameNet과 확장한 FrameNet을 이용하여, MASC I 말뭉치에 대한 의미역 결정 통해 추출한 Token의 개수, Frame의 수 그리고 Frame Elements의 개수를 표 8과 표 9를 통해 나타내었다.

기존의 FrameNet과 확장된 FrameNet을 이용해 의미역 결정을 수행한 결과를 비교해 보면 기존의 FrameNet을 통해 태깅된 의미역 결정의 결과는 10,043개의 Token에 대하여 2,762개의 Frame과 2,909개의 Frame Element를 태깅하였으나, 본 연구를 통해 확장한 FrameNet을 통해 태깅된 의미역 결정의 결과는 같은 데이터에 대하여 3,363개의 프레임과 4,477개의 의미역 정보를 포함함으로써, 더 많은 의미역 정보를 포함하였음을 확인 할 수 있었다.

표 8. MACS I – 기존 FrameNet 태깅된 의미역 정보

구분	#Token	#Frame	#FrameElements
1	509	172	174
2	224	77	79
3	165	60	61
4	353	99	99
5	284	105	106
6	49	12	13
7	204	58	59
8	243	60	63
9	85	15	16
10	418	108	108
11	3025	913	946
12	3041	677	675
13	576	162	165
14	867	244	345
합	10,043	2,762	2,909

표 9. MACS I – 확장된 FrameNet 태깅된 의미역 정보

구분	#Token	#Frame	#FrameElements
1	509	194	240
2	224	87	93
3	165	65	88
4	353	125	134
5	284	94	146
6	49	13	20
7	204	70	90
8	243	81	88
9	85	21	16
10	418	132	172
11	3025	1038	1486
12	3041	966	1298
13	576	183	225
14	867	294	381
합	10,043	3,363	4,477

확장된 FrameNet의 정확성 평가를 위해 정확률과 재현율을 측정한 결과는 그림 12와 같다. 재현율의 경우 확장된 FrameNet은 더 많은 정보를 포함시킬 수 있었기 때문에 기존의 방식보다 향상된 결과를 보였으나, 정확률의 경우 기존 FrameNet으로 태그된 의미역을 정답으로 판단하고, 확장을 통해 새롭게 나타나는 부분은 실패로 간주되기 때문에 기존의 방식의 경우 1의 정확률로 계산되며, 확장된 FrameNet의 경우 0.79의 결과를 보였다.

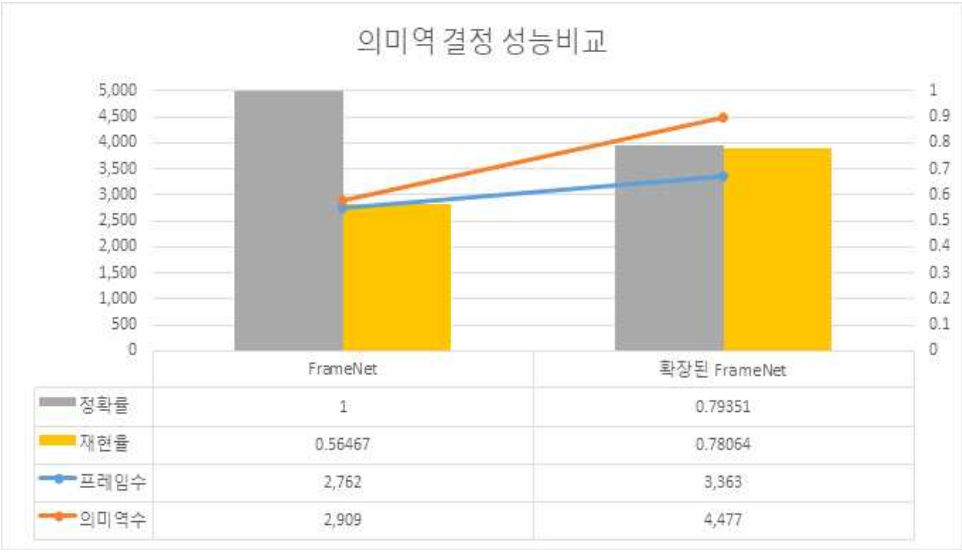

그림 12. 확장된 FrameNet 성능평가

IV. 확장된 의미역 결정을 이용한 문서 유사성 측정

본 장은 Paraphrase Database를 이용해 확장한 FrameNet을 이용하여 문서의 의미역 결정을 수행하고, 문장 내 부착된 의미역 정보를 비교함으로써, 의미적 유사성을 측정하는 방법에 관하여 기술한다.

확장된 의미역 결정을 이용한 문서 유사성 측정을 위한 과정은 전처리과정을 통해 문장분할과 토큰화(Tokenization), 원형화(Lemmatization)를 수행하고, 확장된 FrameNet을 이용한 의미역 결정을 수행한다. 마지막으로 문장 내 결정된 의미역 정보를 기반으로 비교 대상 문장과의 의미적 유사성 계산한다. 그림 13는 확장된 FrameNet을 이용한 문서 유사성 측정 방법의 구조도이다.

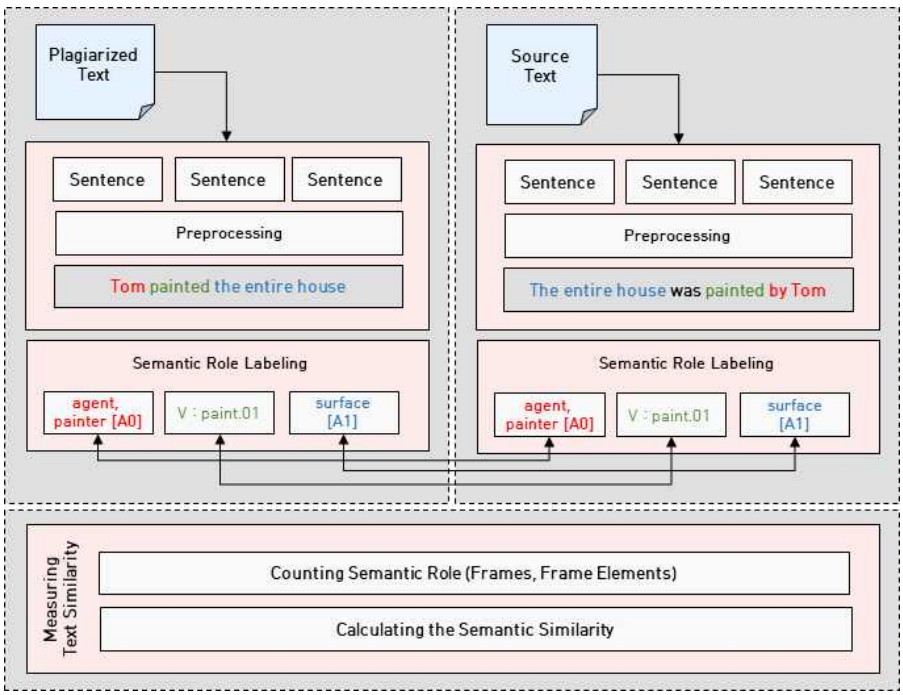

그림 13. 확장된 FrameNet을 이용한 문서 유사성 측정

1. 전처리 과정

■ 텍스트 분할(Text Segmentation) : 전처리 과정은 자연어 처리의 핵심 단계 중 하나의 과정으로 텍스트를 의미있는 단위로 나누는 과정을 텍스트 분할이라하며, 문장 분할은 형태소 분석이나 구문 분석을 하기 위한 전처리 과정으로 문서를 하나하나의 문장으로 분할하는 과정이다. 본 논문에서 제안한 방법은 구문 분석을 기반으로 문서의 유사성 측정을 위한 의미역 결정을 수행하기 때문에 입력된 문서에 대하여 문장 단위로 나누는 문장 분할 과정을 수행하게 된다.

■ 불용어 제거(Removing Stopwords) : 불용어 제거는 문장 내 무의미한 단어를 삭제하는 단계이다. 불용어는 the, and와 같이 문서에 자주 등장하지만 문서의 의미에 영향을 주지 않는 단어를 말하며, 따라서 색인 용어 집합에서 제외된다. 일반적으로 불용어는 문서에서 절반을 차지하며, 정보 검색의 인덱싱에서 불용어를 제거하면 시스템 처리 속도가 빨라지고 엄청난 양의 저장공간을 절약할 수 있다[51]. 본 논문에서는 불용어 제거를 위해 기존에 정의된 불용어 리스트를 활용하여, 텍스트 내의 모든 불용어를 제거하여 시스템 처리 속도를 높이고자 하였다.

■ 원형화(Lemmatization)이란 문장 내 다양한 형태로 변형된 단어의 표제어(lemma)를 찾는 과정이다. 표제어란 사전에서 단어의 뜻을 찾을 때 쓰는 기본형을 뜻한다. 예를 들어, '아름다운'이란 단어는 원형화 과정을 거치면 '아름답다'가 되며, 원형화 과정을 통해 해당 단어가 문장 속에서 어떤 품사로 쓰였는지까지 판단한다. 원형화는 문서 분석에서 단어 형태의 변형으로 같은 의미의 단어가 다른 정보로 인식되는 문제를 해결하는 효과적인 방법으로 평가된다[52]. 본 논문에서도 원형화를 통해 구문 분석의 성능을 향상시키고자 하였다.

표 10. 전처리 과정 예

Step	Description Example
Raw data	The cry of, "Pig out!" and the consequent rush of children in pursuit, at last reached such a pitch that both Miss Grey and the much-tried Andrew made complaint to the vicar. Miss Grey declared that discipline was becoming impossible, and Andrew that there would not be a "martal vegetable in the garden if Master David's pig got out so often." Then the vicar made a rule to this effect:
Text Segmentation	The cry of, Pig out! and the consequent rush of children in pursuit, at last reached such a pitch that both Miss Grey and the much-tried Andrew made complaint to the vicar. Miss Grey declared that discipline was becoming impossible, and Andrew that there would not be a martal vegetable in the garden if Master David's pig got out so often. Then the vicar made a rule to this effect:
Tokenization	The cry of , Pig out ! and the consequent rush of children in pursuit , at last reached such a pitch that both Miss Grey and the much-tried Andrew made complaint to the vicar . Miss Grey declared that discipline was becoming impossible , and Andrew that there would not be a martal vegetable in the garden if Master David 's pig got out so often . Then the vicar made a rule to this effect
Lemmatization	The cry of , Pig out ! and the consequent rush of **child** in pursuit , at last **reach** such a pitch that both Miss Grey and the much-tried Andrew **make** complaint to the vicar . Miss Grey **declare** that discipline **be become** impossible , and Andrew that there would not be a martal vegetable in the garden if Master David 's pig **get** out so often . Then the vicar **make** a rule to this effect :

2. 확장된 FrameNet을 이용한 의미역 결정

의미역 결정은 입력 문장의 서술어를 기준으로 술어와 논항관계를 행위주, 경험주, 대상 등의 의미 관계로 사상하는 문제로 볼 수 있다[13, 14]. 일반적으로 의미역 결정을 위해서는 입력 문장에 대해 구문 분석을 수행하는 서술어 인식단계와 서술어의 중의성을 해소하는 서술어 분류단계가 선행되며, 문장 내 특정 의미를 갖는 서술어가 하나의 문장을 이루기 위해 필요한 정보를 인식하는 논항 인식단계와 인식된 논항의 의미역을 결정하는 논항 분류단계로 수행된다[35]. 본 장에서는 확장된 FrameNet을 적용한 의미역 결정 과정에 대해 기술한다.

1) 서술어 인식(Predicate Indentification)

서술어 인식은 어떤 단어가 문장에서 서술부에 해당되는지 식별하는 단계이다. ProBank에서는 문장의 동사를 서술부로 정의하고 있어 구별하기 쉽지만 FrameNet에서는 동사, 명사, 형용사, 심지어 전치사조차도 때에 따라서는 Frame역할을 하기도 한다. 따라서, 어휘분석기(Parser)를 통해 구문 분석을 수행하여, 해당 문장의 서술부를 식별하는 과정을 거친다. 일반적으로 구문 분석은 입력문장(S)을 일차적으로 주어부(Subject)와 술어부(Predicate)로 나누고, 의미해석 단계를 거쳐 그 구성 성분들을 세분화하는 단계로 진행된다. 구문 분석의 결과는 그림 14와 같이 그래프 형태로 표현할 수 있다.

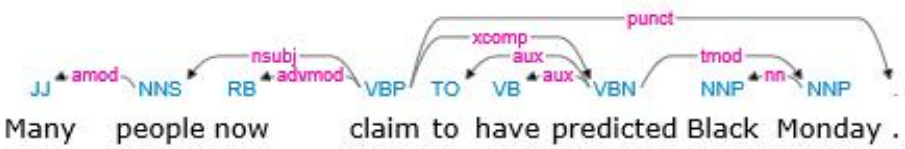

그림 14. 구문 분석 결과 예

의존 구문 분석은 토큰 간의 이진 관계로 문장의 구문 구조를 나타낸다. 문장 내 동사는 종속관계에 있는 논항에 연결되며, 따라서 의존 구문 분석기는 노드와 종속 관계 유형을 라벨링한 결과를 반환하기 때문에 대부분의 의미역 결정의 서술어 인식을 위해 의존 구문 분석기를 사용한다. 구문 분석을 통해 서술어가 식별되면, 해당 서술어가 문장에서 어떤 의미적 정보를 표현하는지 판단하는 서술어 분류단계를 수행하게 된다. 표 11은 의존 구문 분석기의 태그의 정보를 나타낸다[53].

표 11. 구문 분석 결과 태그 정보

태그	내용	
sbj	subject	주어
obj	object	목적어
mod	modifier	관형어
conj	conjunct	접속사
nsubj	nominal subject	주격 명사구
dobj	direct object	직접목적어
iobj	indirect object	간접목적어
aux	auxiliary verb	조동사
prep	prepositional phrase	전치사구
pobj	object of preposition	전치사의 목적어
det	determiner	한정사
nn	noun compound modifier	명사 수식어
amod	adjectival modifier	형용사 수식어
advmod	adverb modifier	명사 수식어
punct	punctuation	구두점

2) 서술어 분류(Predicate Classification)

본 연구에서는 의미역 결정을 위해 FrameNet을 사용하였기 때문에 구문 분석을 통해 식별된 서술부를 FrameNet의 Lexcial Units으로 간주하고, Lexical Units가 속한 Frame을 검색한다. Lexical Units이 여러 의미를 갖는 동음이의어일 경우 하나의 Lexical Units가 여러 Frame에 연관 될 수 있다. 예를들어, claim.v은 총 11개의 Frame과 연관이 있다.

$$
\begin{aligned}
f &= \{f_1, f_2, \cdots, f_n\} \\
&= \{\ \textit{claim\_ownership, communication, have\_as\_requirement,} \\
&\quad \textit{predicting, questioning, reasoning, request, seeking,} \\
&\quad \textit{seeking\_to\_achieve, statement, submitting\_documents}\ \}
\end{aligned}
$$

이렇게 Lexical Units은 의미적 모호성을 가지고 있다고 할 수 있기 때문에 Lexical Units의 Frame을 결정하기 위한 서술어 분류 과정을 수행하며, 수식 6을 통해 분류할 수 있다[54].

$$
\begin{aligned}
f_i \leftarrow \arg max_{f_i \in f} Score_i(frame, predicate, sentence) \\
Score_i = P(frame | predicate, sentence)
\end{aligned}
\tag{6}
$$

Score는 Lexical Units의 Frame을 분류하기 위해 문장 내 등장한 서술어가 해당 Frame으로 분류될 확률값으로, 모든 Frame에 대해 계산된 Score 중 가장 큰 값을 갖는 Frame으로 분류된다.

표 12의 Score는 서술어 분류과정의 이해를 돕기위한 예시로 입력문장과 Frame의 Description과의 워드넷의 유사도를 적용하여 Score를 산출하였다. 실제 본 논문의 실험에서는 학습된 의미역 결정 시스템으로 SEMAFOR를 이용하여 서술어 분류를 수행한다.

표 12. 서술어 인식 및 분류 과정

입력문장	Many people now claim to have predicted Black Monday.		
구문 분석 결과	Many[DT], people[NN], now[RB], claim[VBP], to[TO], have[VB], predicted[VBN], black monday[NNP].		
Lexical Units	Frame	Description	Score
claim.v	CLAIM_OWNERSHIP	A Claimant asserts his or her right to possession of a piece of Property.	0.7094
	COMMUNICATION	A Communicator conveys a Message to an Addressee.	0.7573
	HAVE_AS_REQUIREMENT	The obtaining of a Requirement state of affairs or the presence of a Required_entity is profiled as a prerequisite for the obtaining or occurring of a Dependent state-of-affairs.	0.3204
	PREDICTING	A Speaker states or makes known a future Eventuality on the basis of some Evidence .	0.7081
	QUESTIONING	The words in this frame have to do with a Speaker asking an Addressee a question which calls for a reply.	0.7550
	REASONING	An Arguer presents a Content , along with Support , to an Addressee .	0.6155
	REQUEST	In this frame a Speaker asks an Addressee for something, or to carry out some action.	0.4279
	SEEKING	A Cognizer_agent attempts to find some Sought_entity by examining some Ground.	0.018
	SEEKING_TO_ACHIEVE	An Agent intends and takes steps towards bringing about a State_of_affairs or, metonymically, towards acquiring a Sought_entity.	0.2286
	STATEMENT	**This frame contains verbs and nouns that communicate the act of a Speaker to address a Message to some Addressee using language.**	**0.8047**
	SUBMITTING_DOCUMENTS	A Submittor gives Documents to an Authority so that they can be processed as part of an application, evaluation or other official or bureaucratic process.	0.5190

3) 논항 인식 및 분류 (Argument Identification and Classification)

본 논문에서는 문장의 서술어가 결정된 후에 서술어가 문장을 이루기 위해 필요한 논항에 대해 서술어와 의미관계를 갖는 논항인지, 그렇지 않은지 판단하는 과정인 논항 인식단계를 거친다.

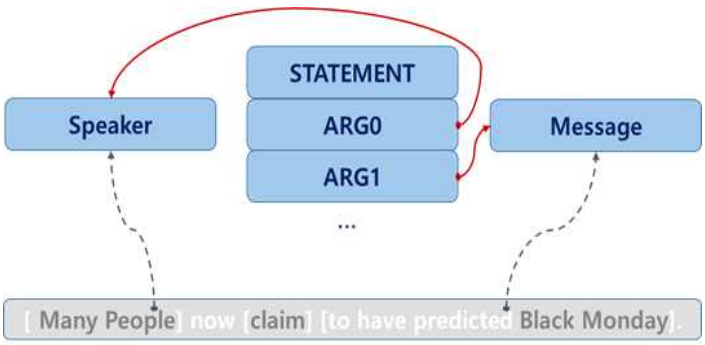

그림 15. 논항 인식

그림 15는 논항을 인식하는 과정을 도식화한 것이다. 서술어 인식 및 분류과정을 통해 결정된 서술어인 “claim.v”과 의존관계에 있는 논항들을 찾게되며, 예시 문장에서는 서술어 “claim.v”에 대한 논항으로 “Many people”, “now”, “to have predicted Black Monday”가 인식되었다.

논항으로 인식된 노드들에 대해서는 논항 분류단계를 거치게 된다. 앞선 서술어 인식 및 분류단계에서 문장 내 “claim” 이라는 서술어의 Frame으로 STATEMENT가 선정되었다. 해당 Frame이 갖을 수 있는 의미역은 Speaker, Addressee, Messeage, Topic, Manner, Mean, Internal_cause, Time으로 총 8개이며, 해당 의미역 중 문장 내 논항에 적합한 의미역을 결정하는 과정을 거친다.

각 논항들에 적합한 의미역을 부여하기 위해 해당 논항이 Frame이 갖을 수 있는 의미역으로 분류될 확률을 구하게 된다. 이러한 확률값은 각 논항이 해당 의미역을 갖을 확률정보를 나타내며, 수식 7에서 Score에 해당된다.

$$Score_i = P(role|\, span, frame) \tag{7}$$

의미역 결정을 위해 수식 8을 통해 모든 논항과 의미역의 점수를 계산하고, 확률값이 최대가 되는 조합을 찾아 의미역결정을 수행한다. 논항이 분류되지 않은 의미역 값은 0이 되어 합계에 계산되지 않는다.

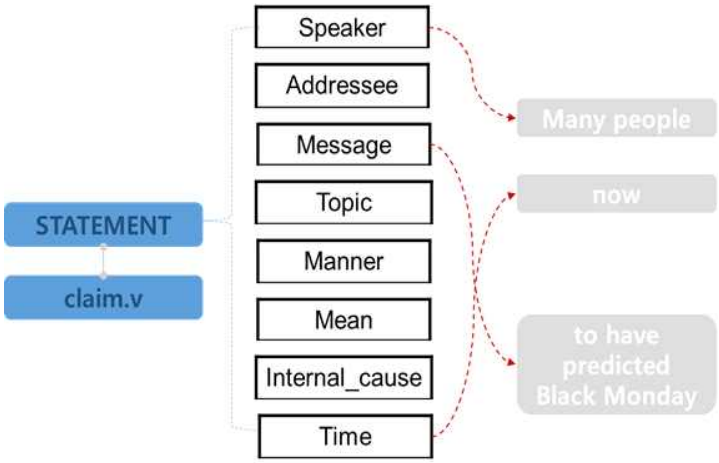

그림 16. 논항 분류

$$Semantic\,Role\,Labeling = \max imize \sum_{role,\,spans} z \bullet Score_i(role, span), \tag{8}$$

$$z = \begin{cases} 1 : \text{if } span\,has\,role \\ 0 : otherwise \end{cases}$$

그 결과 해당 Frame의 행동주 의미역인 [Speaker]으로 “Many People”을 분류하였고, 대상 논항정보를 나타내는 의미역인 [Message]로 “to have predicted Black Monday”을 분류하였다.

이렇게 문장 내의 술어와 의미적으로 관계있는 논항 정보를 인식하고, 논항의 의미역할을 분류하기 위해 본 논문에서는 FrameNet 의미역 논항정보가 부착된 학습데이터를 기반으로 논항 분류를 수행하며, 논항 분류의 결과는 표 13과 같이 나타낼 수 있다.

표 13. 논항 분류 결과

Frame	Argument Identification	Frame Elements (Semantic Role Labeling)	
STATEMENT	Many people	Speaker	Speaker
		Addressee	
		Message	
	now	Topic	Time
		Manner	
		Mean	
	to have predicted Black Monday	Internal_cause	Message
		Medium	
		Time	

학습 데이터에는 각 프레임의 예문과 의미역 정보를 담고 있다. 동일한 논항을 갖는 프레임이 거의 없기 때문에 학습데이터를 이용해서 논항 분류하는 것이 가능하다. 또한, 동일한 논항을 갖는 Frame은 비슷한 의미적 속성을 갖게된다. 예를 들어 Perpetrator라는 논항을 갖는 ARSON과 THEFT 두 프레임의 의미적 속성은 유사하다. 이러한 이론을 기반으로 본 논문은 의미역 결정의 결과를 바탕으로 문서의 의미적 유사성을 측정한다.

3. 문서 유사성 측정

문서 유사성 측정은 검사 대상 문서와 비교 문서와의 비교를 통해 문서의 유사성을 측정하며, 일반적으로 문장 내 색인어를 이용하여 비교 문서 집합에서 후보 문장을 선별하고, 두 문장 간의 유사성 측정을 통해 문서의 유사성을 측정하는 방법을 통해 이뤄진다. n-gram이나 부분 문자열 매칭과 같은 유사성 측정에서 사용 된 비교 방식은 두 문장 또는 문서간의 등장하는 단어를 비교하게 된다. 하지만, 문장을 이루는 모든 단어들 간의 유사성을 측정하는 방법은 유사성 측정을 위한 시간을 소요시킬 뿐만 아니라 정확성을 저해할 수 있다.

제안된 방법은 문장내 단어에 의미역을 부착함으로써 두 문장 간의 의미역 간의 유사성 측정에 초점을 맞추는 것으로 의미역은 서술어로부터 의미관계를 갖는 논항들이 문장 내 분포하게 되기 때문에 유사한 내용의 문장에서는 유사한 의미의 논항들이 나타난다. 문장 내 의미역을 기반으로 유사성을 비교하는 것은 문장의 유사성 측정을 위해 소요되는 시간적인 비용을 감소시킬 수 있으며, 문장의 변형이 이루어졌으나 같은 의미를 담고 있는 다시쓰기 유형을 검출 할 수 있다는 장점이 있다.

두 문장의 유사도는 부분 문자열(substring)을 활용한 비교 방법인 Greedy String Tiling 유사도[25]를 기반으로 계산한다. 이 때, 부분 문자열 대신 의미역 결정을 통해 얻게되는 문장 내 단어와 의미역 정보를 특징정보로 정의하여, 두 문장 간에 공통으로 등장하는 특징정보의 비율로써 두 문장의 유사도를 계산하게 된다. 유사성은 수식 9을 통해 계산한다.

$$Semantic\,Role\,Labeling\,Similarity = \frac{2 * \sum_{i \in srl} length_i}{|Srl_{sus} + Srl_{src}|} \quad (9)$$

Srl_{sus}는 의심 문서의 의미역, Srl_{src}는 원본 문서의 의미역이며, $length_i$는 두 문서에서 공통으로 발견되는 의미역과 단어의 개수이다. $|Srl_{sus} + Srl_{src}|$는 의심 문서와 Srl_{sus}와 원본 문서 Srl_{src}를 구성하는 전체 의미역의 개수이고, 두 문서에서 공통으로 등장하는 의미역이 차지하는 $length_i$의 비율로 두 문장 간의 유사도를 계산한다.

표 14. 유사성 측정을 위한 python 코드

```
import pandas as pd
import numpy as np
from collections import Counter

for i in range(range(00000,02999)):
    SRC=pd.read_excel('SourceDocument'+str(i)+'.csv') #source file 읽기
    SUS=pd.read_excel('suspiciousDocument'+str(i)+'.csv') #suspicious file 읽기

    Notnull_Index=np.where(pd.notnull(SRC)) # SRC 데이터가 있는 곳의 인덱스
    count1=np.shape(list(Notnull_Index))[1]  # SRC 데이터 개수

    Notnull_Index2=np.where(pd.notnull(SUS)) #SUS 데이터가 있는 곳의 인덱스
    count2=np.shape(list(Notnull_Index2))[1]  # SUS 데이터 개수

    SRC_list1=SRC.iloc[np.where(pd.notnull(SRC.iloc[:,0]))[0],0]
    # SRC 1열 데이터들 보기
    SUS_list1=SUS.iloc[np.where(pd.notnull(SUS.iloc[:,0]))[0],0]
    # SUS 1열 데이터들 보기
    sum1 = len(list((Counter(SRC_list1) & Counter(SUS_list1)).elements()))
    # 1열 일치하는 데이터 개수

    SRC_list2=SRC.iloc[np.where(pd.notnull(SRC.iloc[:,1]))[0],1]
    # SRC 2열 데이터들 보기
    SUS_list2=SUS.iloc[np.where(pd.notnull(SUS.iloc[:,1]))[0],1]
    # SUS 2열 데이터들 보기
    sum2 = len(list((Counter(SRC_list2) & Counter(SUS_list2)).elements()))
    # 2열 일치하는 데이터 갯수

    Sum_Result = sum1 + sum2  # 두 문서 일치하는 데이터의 개수의 합
    similarity = 2 * Sum_Result / (count1+count2)  # 유사성 계산

    print(str(i)+'의 유사도: '+ str(similarity))
```

V. 실험 및 성능평가

1. 실험 데이터

본 논문에서 제안한 확장된 의미역 결정을 이용한 문서의 유사성 측정방법의 적용 및 성능평가를 위해 PAN 2012 말뭉치(pan12-text-alignment-test-corpus)를 사용하였다. PAN 2012 말뭉치는 표절 검사 시스템을 위해 구축된 말뭉치로 표절 검사 관련 알고리즘의 개발 및 성능 평가를 위해 제공하고 있는 말뭉치이다[55].

PAN 2012 말뭉치는 3,500개의 원본 문서(Source text)와 3,000개의 의심 문서(suspicious text)로 구성되어있다. 원본문서와 의심문서는 표절의 유형에 따라 6개의 카테고리로 분류되어 있으며, 카테고리 정보는 표 15와 같다. 이 때, 본 논문은 영문을 대상으로 유사성 측정을 수행하기 위한 연구이기 때문에 영문을 번역한 형태의 표절 검출을 위한 Translation 카테고리에 해당되는 문서는 본 실험에서 제외하였다.

표 15. PAN 2012(text alignment text corpus) 정보

카테고리	정의
01. no_plagiarism	어떠한 표절도 없는 500개의 문서 쌍
02. no_obfuscation	의심 문서에 원본 문서의 수정 없이 표절된 내용을 포함하는 500개의 문서 쌍
03. artificial_low	의심 문서에 원본 문서의 수정이 이루어진 표절된 내용을 포함하는 500개의 문서 쌍이며, 04 카테고리보다 수정된 정도가 낮음.
04. artificial_high	의심 문서에 원본 문서의 수정이 이루어진 표절된 내용을 포함하는 500개의 문서 쌍이며, 03 카테고리보다 수정된 정도가 높음.
05. translation	유럽언어에서 영문으로 번역하여 표절된 내용을 포함하는 500개의 문서 쌍이며, 영문이 아닌 원본 문서 500개를 포함하고 있음.
06. simulated_paraphrase	아마존 메카닉 터크(Amazon Mechanical Turk)를 통해 다시쓰기 표절을 수행한 500개의 문서 쌍을 포함하고 있음.

표 16. PAN 2012 말뭉치 정보

```
<document reference="...">     <!-- file name of the suspicious document          -->
<feature
  name="detected-plagiarism"   <!-- type of the plagiarism annotation             -->
  this_offset="5"              <!-- char offset within the suspicious document   -->
  this_length="1000"           <!-- number of chars beginning at the offset      -->

  source_reference="..."       <!-- file name of the source document             -->
  source_offset="100"          <!-- char offset within the source document       -->
  source_length="1000"         <!-- number of chars beginning at the offset      -->

/>
...                            <!-- more detections in this suspicious document -->
</document>
```

표 16은 PAN 2012 말뭉치의 문서 정보를 나타내고 있는 xml 파일의 구조를 나타낸 것이다. 해당 정보는 문서 유사성 측정 결과에 대한 정확률 및 재현율을 평가를 위해 사용한다.

각 카테고리 내에는 xml 파일을 통해 의심 문서와 원본 문서 쌍 내부에 표절이 수행된 위치 및 범위를 기술하고 있다. this_offset은 텍스트 내 표절이 이뤄진 위치를 나타내고 있으며, this_length는 표절이 이뤄진 내용의 길이를 나타내고 있다. source_offset은 원본 문서 내 표절에 사용된 문장의 위치를 나타내고, source_length는 원본 문서 내 표절에 사용된 문장의 길이를 나타내며, 각 카테고리 별 500개의 xml 파일로 제공되고 있다.

2. 문서 유사성 측정

1) 부분 문자열 기반 유사성 측정

본 논문에서 제안한 유사성 측정을 위해 5개의 카테고리 (no_plargiarism, no_obfuscation, artificial_low, artificial_high, simulated_paraphrase)의 데이터를 사용하였으며, 본 연구에서 제안한 방법과 비교하기 위해 Greedy Srting Tiling 유사도[25]를 이용하여 원본 문서와 의심 문서의 유사성을 측정하였다.

GST 유사도는 단어를 기반으로 문서 내 공통된 부분 문자열의 개수를 측정하는 방법으로 문서의 유사성을 측정하는 표절검사 시스템에서 일반적으로 적용하고 있는 방법이다. GST 유사도는 수식 10으로 계산한다.

$$Greedy\,String\,Tiling\,Similarity = \frac{2 * \sum_{i \in tiles} length_i}{|Src_a + Sus_b|} \quad (10)$$

Sus_a는 의심 문서, Src_b는 원본 문서를 이루는 전체 단어의 수를 나타내고 Tiles은 두 문서 간에 공통으로 등장하는 최소 부분 문자열의 크기를 나타내며, 본 논문에서는 세 단어이상 공통으로 등장하는 부분 문자열을 Tile의 크기로 정의하여 유사성 측정을 수행하였다.

실험 데이터인 5개의 카테고리의 2500개의 문서쌍에 대한 의미역 결정 정보 기반의 유사성을 측정한 실험결과는 표 17과 같다. 이 방법은 기존에 알려진 바와 같이 변형이 많이 이뤄지지 않은 표절에 대해서는 높은 유사성을 보였으나, 다시쓰기와 같은 비슷한 단어로 변형된 표절의 경우 유사성을 측정하지 못하는 것으로 확인 할 수 있었다.

표 17. 부분 문자열 정보 기반 유사성 측정 결과

pair	# source text	#suspicous text	# match word	GST Similarity
no_plagiarism(00000~00499)				
00000	23174	645	0	0
00001	5823	23374	0	0
00002	630	6684	3	0.00082
…	…	…	…	
00497	694	3853	0	0
00498	554	7198	7	0.001806
00499	7030	745	0	0
no_obfuscation(00500~00999)				
00500	788	785	785	0.9980928
00501	568	568	567	0.9982394
00502	64	64	64	1
…	…	…	…	…
00997	646	646	646	1
00998	482	482	482	1
00999	186	186	186	1
artificial_low(01000~01499)				
01000	318	320	161	0.5047022
01001	3212	3118	719	0.2271722
01002	75	73	28	0.3783784
…			…	
01497	425	431	196	0.4579439
01498	402	391	201	0.5069357
01499	103	145	40	0.3225806
artificial_high(01500~01999)				
01500	800	827	54	0.0663798
01501	140	141	10	0.0711744
01502	142	139	24	0.1708185
…			…	
01997	523	507	10	0.0194175
01998	4079	1226	69	0.0260132
01999	310	314	38	0.1217949
simulated_paraphrase(02500~02999)				
02500	888	784	172	0.2057416
02501	93	94	70	0.7486631
02502	164	135	42	0.2809365
…	…	…	…	
02997	324	331	199	0.6076336
02998	42	32	7	0.1891892
02999	213	195	39	0.1911765

2) 확장된 FrameNet을 이용한 유사성 측정

본 절에서는 확장된 FrameNet을 이용하여 표절 검사를 수행하기위한 실험 데이터에 대하여 의미역 결정을 수행한다. 확장된 FrameNet의 의미역을 부착하기 위한 과정은 앞선 실험과 동일하게 문서의 의미적 분석을 위해 개발된 오픈소스 API인 SEMAFOR 시스템을 이용하였으며, 3장에서 구축한 frame identification 학습 모델과 Argument identification 학습 모델의 학습과정을 수행한 후 실험데이터의 의미역 결정을 수행하였다.

SEMAFOR 시스템은 분석을 위해 입력된 문서 내 문장을 분할하고, 원형화하는 과정을 거친다. 그 다음 품사태깅 후 구문 분석을 수행하는 전처리과정을 시작으로 문장 내에서 서술어를 나타내는 단어를 식별하게 된다. 이 단어를 FrameNet에서는 Target 또는 Lexical Units이라고 하며, 문장 내 술어 역할을 하는 단어라고 판단하면 된다. 이 과정은 4장에서 설명한 Frame Identification 과정에 해당된다. 서술어인 Lexical Units을 식별한 후에는 해당 Lexical Units를 갖는 Frame 중 해당 문장에서 나타내고 있는 Frame을 선정하게 된다.

이렇게 Lexical Units의 Frame 분류를 위해 SEMAFOR 시스템에서는 확장된 FrameNet의 학습데이터를 이용하여 해당 Lexical Units이 Frame에 분류될 확률을 계산하여, 가장 높은 값을 갖는 Frame을 Lexical Units의 Frame으로 선정하게 된다. 그 후 동일한 데이터에 대해 학습된 확장된 FrameNet은 Frame이 갖을 수 있는 모든 의미역을 고려하여 문장 내 논항의 단어 범위(span)를 채우는 과정으로 의미역 분류를 수행한다. 학습된 SEMAFOR 시스템을 통해 수행된 의미역 결정의 결과는 프레임-논항 정보가 부착된 형태인 JSON 파일로 생성된다. 표 18은 문서에 의미역 결정 정보가 부착된 결과를 나타낸 JSON파일의 일부이다.

표 18. SOURCE_DOCUMENT01000 의미역 결정 결과 (JSON포맷)

{"frames":[{"target":{"name":"Vocalizations","spans":[{"start":1,"end":2,"text":"cry"}]},"annotationSets":[{"rank":0,"score":19.707551234628124,"frameElements":[]}]},{"target":{"name":"Locative_relation","spans":[{"start":5,"end":6,"text":"out"}]},"annotationSets":[{"rank":0,"score":17.381475530760007,"frameElements":[]}]},{"target":{"name":"Self_motion","spans":[{"start":10,"end":11,"text":"rush"}]},"annotationSets":[{"rank":0,"score":102.09564263824811,"frameElements":[]}]},{"target":{"name":"Kinship","spans":[{"start":12,"end":13,"text":"children"}]},"annotationSets":[{"rank":0,"score":24.12012421488906,"frameElements":[{"name":"Alter","spans":[{"start":12,"end":13,"text":"children"}]}]}]},{"target":{"name":"Seeking_to_achieve","spans":[{"start":14,"end":15,"text":"pursuit"}]},"annotationSets":[{"rank":0,"score":61.66296317454807,"frameElements":[]}]},{"target":{"name":"Relative_time","spans":[{"start":17,"end":18,"text":"last"}]},"annotationSets":[{"rank":0,"score":18.669719299367696,"frameElements":[{"name":"Focal_occasion","spans":[{"start":19,"end":35,"text":"such a pitch that both Miss Grey and the much-tried Andrew made complaint to the vicar"}]}]}]},{"target":{"name":"Arriving","spans":[{"start":18,"end":19,"text":"reached"}]},"annotationSets":[{"rank":0,"score":84.36351438387071,"frameElements":[{"name":"Goal","spans":[{"start":19,"end":35,"text":"such a pitch that both Miss Grey and the much-tried Andrew made complaint to the vicar"}]},{"name":"Theme","spans":[{"start":17,"end":18,"text":"last"}]}]}]},{"target":{"name":"Quantity","spans":[{"start":23,"end":24,"text":"both"}]},"annotationSets":[{"rank":0,"score":21.737565116511191,"frameElements":[{"name":"Quantity","spans":[{"start":23,"end":24,"text":"both"}]}]}]},{"target":{"name":"Success_or_failure","spans":[{"start":24,"end":25,"text":"Miss"}]},"annotationSets":[{"rank":0,"score":50.670001671927636,"frameElements":[]}]},{"target":{"name":"Causation","spans":[{"start":30,"end":31,"text":"made"}]},"annotationSets":[{"rank":0,"score":64.1656535078296,"frameElements":[{"name":"Cause","spans":[{"start":23,"end":30,"text":"both Miss Grey and the much-tried Andrew"}]},{"name":"Effect","spans":[{"start":32,"end":35,"text":"to the vicar"}]}]}]},{"target":{"name":"Complaining","spans":[{"start":31,"end":32,"text":"complaint"}]},"annotationSets":[{"rank":0,"score":42.16990099049123,"frameElements":[]}]}],"tokens":[" The","cry","of",",","Pig","out","!","and","the","consequent","rush","of","children","in","pursuit",",","at","last","reached","such","a","pitch","that","both","Miss","Grey","and","the","much-tried","Andrew","made","complaint","to","the","vicar","."]} ....

의미역 결정을 통해 분석된 정보는 다음과 같다. frame_target_name은 문장에 있는 서술어 역할을 하는 단어인 Lexical Units(target)를 구문 분석을 통해 선정하고, 해당 단어가 속한 frame 정보를 나타내고 있으며, frame_target_span_start와 frame_target_span_end는 frame는 영향을 미치는 범위인 span의 시작하는 부분과 끝나는 부분의 위치정보를 나타낸다. frame_target_spans_text는 span start와 end 위치에 해당되는 텍스트를 나타낸다. frames_annotationSets_score는 frame 선정시 계산된 확률 정보이다. frame_annotationSets_frameElements_name는 선정된 frame과 의미적 관계를 가지고 있는 논항정보를 보여준다. frames_annotationSets_frameElements_span_start와 frame_annotationSets_frameElements_span_end는 논항정보의 범위인 span의 시작하는 부분과 끝나는 부분의 위치정보를 나타내고 있다. frames_annotationSets_frameElements_spans_text는 span start와 end 위치에 해당되는 텍스트를 나타내고, 마지막 부분에는 분석된 문장 전체 tokens을 출력한다. 표 19는 의미역 결정이 수행된 JSON 파일에 대한 정보 추출의 편의를 위해 csv포맷으로 변환하는 과정을 수행한 결과이다.

표 19. SOURCE_DOCUMENT01000 의미역 정보 추출

frames name	frames spans text	frame score	frame elements name	frame elements spans text
Vocalizations	cry	19.70755	**-**	-
Locative_relation	out	17.38148	**-**	-
Self_motion	rush	102.0956	**-**	-
Kinship	children	24.12012	**Alter**	children
Seeking_to_achieve	pursuit	61.66296	**-**	-
Relative_time	last	18.66972	**Focal_occasion**	such a pitch that both Miss Grey and the much-tried Andrew made complaint to the vicar
Arriving	reached	84.36351	**Goal**	such a pitch that both Miss Grey and the much-tried Andrew made complaint to the vicar
-	-	-	**Theme**	last
Quantity	both	21.73757	**Quantity**	both
Success_or_failure	Miss	50.67	**-**	-
Causation	made	64.16565	**Cause**	both Miss Grey and the much-tried Andrew
-	-	-	**Effect**	to the vicar
Complaining	complaint	42.1699	-	-

본 장에서는 문서의 의미적 유사성 측정과정에 대해 서술한다. 이를 위해 원분문서와 의심문서를 입력받아 문장단위로 분할하고, 각 문장에 대해 의미역결정을 통해 얻은 결과물을 유사성 측정을 위한 정보로 활용하여, 4장에서 제시한 의미역 결정 기반 문서 유사성 측정을 평가한다. 표 20는 유사성을 측정하고자 하는 문서의 문장에 대한 의미역 결정을 수행하고, 유사성 측정을 위해 필요한 데이터를 추출한 결과의 예이다. 문장 내 Lexical Units 분석을 통해 선정된 frame 정보를 나타내고 있는 frame_target_name과 frame과 의미적 관계를 가지고 있는 논항정보인 frames_annotationSets_frameElements_name을 유사성 측정을 위한 정보로 사용한다.

의미역 결정을 수행한 두 문서에서 태깅된 프레임과 논항정보 전체에 대해 두 문서에서 공통으로 등장하는 프레임과 논항정보의 비율로 유사성을 측정한다. 표 20과 표 21은 의심 문서에서 원본 문서의 표절된 문장 각각에 대해서 태깅된 프레임과 논항정보를 비교하여 일치하는 정보를 표시하였다. 두 문장은 같은 의미이지만 약간 다른 단어로 수정된 전형적인 표절 유형의 문장이다(02. artificial low 카테고리 문서). 해당 문장의 유사성을 본 논문에서 제안하는 방식으로 측정한 결과 유사한 의미로 변형되고, 구조를 변형한 문장에 대해서도 유사성을 측정할 수 있었다. 실험데이터인 5개의 카테고리의 2500개의 문서쌍에 대한 의미역 결정 정보 기반의 유사성을 측정한 실험 결과는 표 22와 같다.

표 20. SOURCE_DOCUMENT01000 내 문장의 의미역 정보

Sentence #1	The cry of, Pig out! and the consequent rush of children in pursuit, at last reached such a pitch that both Miss Grey and the much-tried Andrew made complaint to the vicar.	
Index	frames_target_name	frames_annotationSets_ frameElements_name
1	Vocalizations	-
2	Locative_relation	-
3	**Self_motion**	-
4	Kinship	Alter
5	**Seeking_to_achieve**	-
6	**Relative_time**	**Focal_occasion**
7	Arriving	Goal
8	-	Theme
9	**Quantity**	**Quantity**
10	Success_or_failure	-
11	**Causation**	**Cause**
12	-	**Effect**
13	Complaining	-

표 21. SUSPICIOUS_DOCUMENT01000 내 문장의 의미역 정보

Sentence #1	The call of, "Pig away!" and the dash of bairn in the pursuit, at last make such a soprano that both attend grey and the much-try Andrew make disorder to the vicar.	
Index	frames_target_name	frames_annotationSets_ frameElements_name
1	Request	-
2	**Self_motion**	-
3	**Seeking_to_achieve**	-
4	**Relative_time**	**Focal_occasion**
5	Type	Subtype
6	-	Category
7	-	Type_Property
8	**Quantity**	**Quantity**
9	Attending	Agent
10	-	Event
11	**Causation**	**Cause**
12	-	**Effect**
13	Medical_conditions	-

표 22. 확장된 의미역 결정 정보 기반 유사성 측정 결과

pair	# roles in source text	# roles in suspcious text	# common roles	ex SRL _ similarity
no_plagiarism(00000~00499)				
00000	18476	481	0	0
00001	4619	21828	0	0
00002	466	4671	4	0.0015573
…	…	…	…	…
00497	546	2695	0	0
00498	359	4869	2	0.0007651
00499	4464	472	0	0
no_obfuscation(00500~00999)				
00500	562	562	561	1
501	482	486	482	1
502	31	31	31	1
…	…	…	…	…
997	68	68	68	1
998	351	351	351	1
999	122	122	122	1
artificial_low(01000~01499)				
1000	249	208	120	0.525164
1001	2146	1736	1146	0.590417
1002	58	57	33	0.573913
…	…	…	…	
1497	372	320	153	0.442197
1498	240	224	138	0.594828
1499	91	87	29	0.325843
artificial_high(01500~01999)				
1500	587	459	156	0.298279
1501	75	82	34	0.433121
1502	101	81	26	0.285714
…	…	…	…	…
1997	377	309	115	0.335277
1998	2916	670	461	0.257111
1999	223	207	148	0.688372
simulated_paraphrase(02500~02999)				
2500	667	613	281	0.439063
2501	57	58	37	0.643478
2502	86	72	27	0.341772
…	…	…	…	…
2997	232	236	140	0.598291
2998	41	37	25	0.641026
2999	131	140	76	0.560886

3. 실험 결과 및 성능 평가

1) 문서 유사성 측정 결과

PAN 2012 말뭉치의 2500개의 문서쌍에 대해 본 논문에서 제안한 방법과 기존의 유사성 측정 방법들과 비교 평가를 수행하였다. PAN 2012 말뭉치에서는 카테고리별 코사인 유사도 값의 평균을 제공하고 있다[55]. 코사인 유사도 측정 방법의 경우 일반적으로 정보 검색 등 문서 유사도 계산시 자주 사용되는 방법이며, 부분 문자열 기반 유사성 측정 방법은 대부분의 표절 검사 시스템에서 사용하고 있는 형태적 유사성 측정 방식이다. 또한 본 연구를 위해 확장한 FrameNet을 이용한 유사성 측정방법과 기존 FrameNet 의미역 결정 정보를 기반으로 유사성 측정 결과를 포함하여 총 4가지의 문서 유사성 측정 결과를 비교하였다.

그림 17은 PAN 2012 말뭉치의 5개의 카테고리에 대해서 코사인 유사도와 부분 문자열 유사도, 기존 의미역 결정 유사도와 확장된 의미역 결정 유사도를 측정한 결과를 나타낸다.

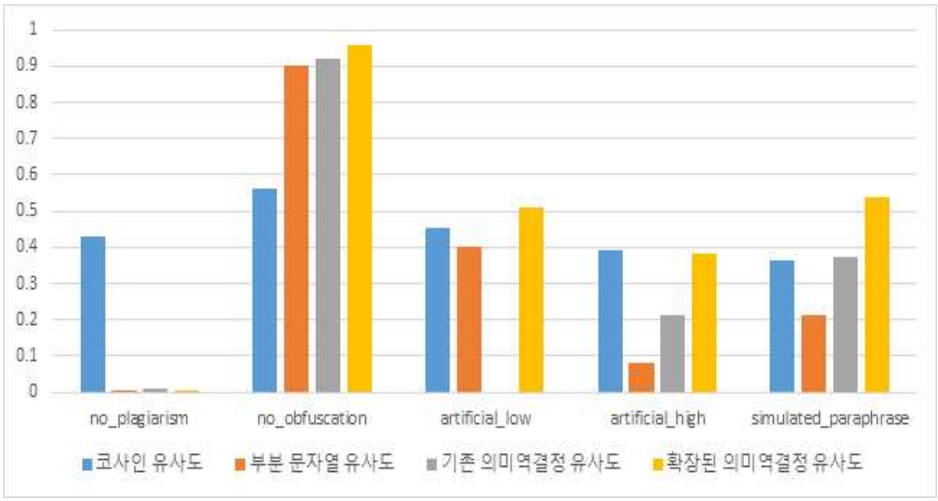

그림 17. 두 문서간 유사성 측정 결과 비교

코사인 유사도는 실험을 수행한 모든 카테고리에서 유사성을 계산해 내고 있다. 특히, 표절이 수행되지 않은 문서에 대해서도 0.43이라는 유사성을 측정하고 있기 때문에 코사인 유사도 방식은 표절 검사를 위한 시스템에는 적합하지 않다고 사료된다.

또한, 부분 문자열 유사성 측정 방법의 경우 전혀 수정이 없는 표절 유형인 no_obfuscation 문서집합과 수정된 부분이 많지 않은 artifical_low 문서집합에는 다른 유사성 측정 방식과 유사한 정도의 유사성을 계산할 수 있었지만, 다른 의미의 단어로 수정을 많이 가해진 artificial_high 문서집합과 수정이 이루어진 simulated_paraphrase 문서집합에는 현저하게 낮은 점수를 산출하였다.

제안한 의미역 결정 정보 기반의 유사성 측정 방법의 경우 형태적 유사성 측정에서도 우수한 성능을 보였으며, 수정이 가해진 artificial_low, artificial_high 문서 집합과 다시쓰기 문장으로 수정이 이루어진 simulated_paraphrase 문서집합에서도 개선된 성능을 나타내었다.

표 23. 두 문서간 유사성 측정 결과 비교

구분	코사인 유사도	부분문자열 유사도	기존 의미역결정 유사도	확장된 의미역결정 유사도
no_plagiarism	0.431	0.00262	0.00789	0.00387
no_obfuscation	0.56	0.89983	0.92121	0.95822
artificial_low	0.455	0.39961	0.43262	0.50873
artificial_high	0.392	0.07926	0.21156	0.38298
simulated_paraphrase	0.364	0.21067	0.37235	0.53742

FrameNet의 의미역 결정 정보를 이용한 유사성 측정은 FrameNet의 의미역 결정을 수행한 결과 중 Frame 정보와 부착된 의미역(Frame Elements)정보를 추출하여, 두 문서간에 공통으로 나타난 Frame정보와 의미역 정보의 비율로써, 유사성을 계산하였으며, 기존의 FrameNet을 이용한 유사성 측정 결과와 확장된 FrameNet을 이용한 유사성 측정결과는 그림 18과 같다. 본 논문에서 확장한 FrameNet은 기존의 FrameNet의 구조를 유지한 상태에서 언어 자원을 결합하는 방식으로 확장하였기 때문에 결과적으로 유사한 양태를 보이지만, 변형이 많이 이루어진 표절 문서에 대해 개선된 성능을 확인할 수 있었다.

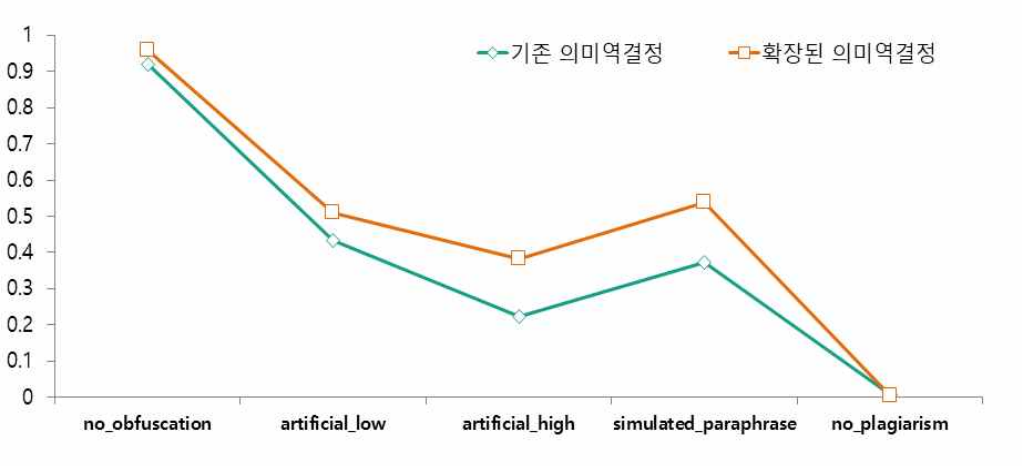

그림 18. 기존 의미역 결정과 제안된 방법의 비교

마지막으로 본 논문에서 비교한 4가지 유사성 측정 결과에 대해 타당도를 평가하였다. 타당도(gold standard)란 측정 결과가 사실을 반영하느냐를 내포하는 개념으로, 일반적으로 표절이 없는 문서 간의 유사성은 0의 값을 나타내며, 완벽하게 표절된 문서의 유사성은 1의 값을 나타낸다고 전제하여, 판단하였으며, 수정이 이루어진 문서의 경우 유사성을 0.5로 하여, 타당도를 기준으로 유사성 측정결과와 비교하였다.

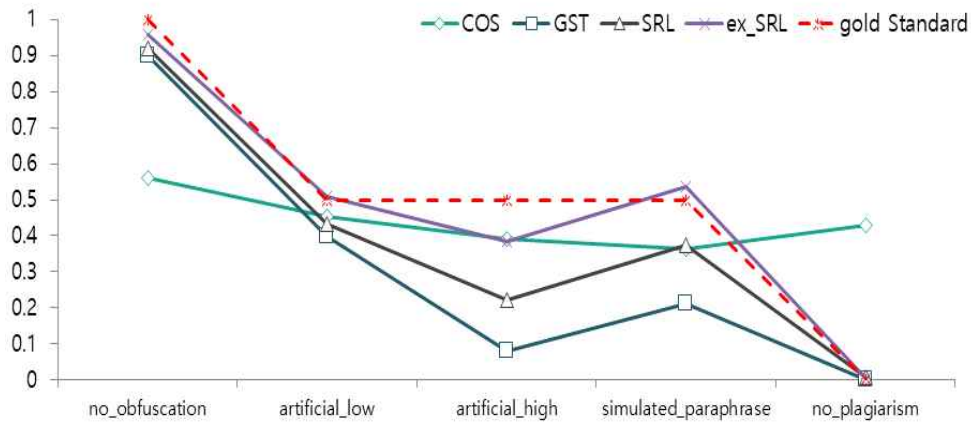

그림 19. 타당도

gold standard 값과 비교한 유사성 측정결과는 그림 19와 같으며, 각 유사성 측정 방법과 상관계수를 측정함으로써, 타당도와의 차이를 측정하였다. 상관계수는 비교하고자 하는 두 개념 사이의 상관 관계의 정도를 나타내는 수치이다. 상관 관계가 있을수록 1과 가까운 값을 갖으며, 무상관일 때는 0과 가까운 값을 갖는다. 표 24는 유사성 측정 결과와 타당도와의 상관성 분석 결과를 나타낸다.

표 24. 유사성 측정 결과-타당도 상관성 분석

구분	코사인 유사도	부분문자열 유사도	기존 의미역결정 유사도	확장된 의미역결정 유사도
상관계수	0.604294	0.885625	0.953565	0.985476

2) 성능 평가

본 논문의 실험을 위해 사용된 PAN 2012 말뭉치의 성능평가를 위한 정확률과 재현율의 개념은 PAN 2012 에서 제공하는 개념을 사용하였다[57]. 본 논문에서는 해당 정확률과 재현율을 사용하여 제안한 유사성 측정 방법을 평가하였다. 정확률은 표절이 나타난 문서의 범위의 의미역 중 제안한 방법으로 찾아낸 의미역의 비율이다. 재현율은 제안한 방법으로 찾아낸 두 문서간 일치한 의미역 중 표절이 나타난 범위의 의미역의 비율이다. 정확률은 시스템이 해결한 문제가 적합한지의 여부를 판단하는 정확성을 의미하고, 재현율은 풀어야할 문제를 어느정도 해결 할 수 있는지를 판단하는 완전성을 의미한다.

정확률과 재현율은 수식 11과 12를 통해 계산한다. 'S'는 표절이 나타난 범위로 제안한 방법을 평가하기 위한 정답 집단에 해당하고, 'R'은 제안한 방법으로 찾은 표절 범위을 의미한다. 'r'는 제안한 방법으로 추출한 두 문서간 일치한 의미역 정보의 개수, 's'은 표절 범위의 의미역의 개수, 'r∩s' 두 문서간 일치한 의미역 정보의 개수 중 표절 범위에 속하는 의미역의 개수로 계산할 수 있다.

$$pre(S,R) = \frac{1}{|R|}\sum_{r\in R}\frac{\cup_{s\in S}(s\cap r)}{|r|} \tag{11}$$

$$rec(S,R) = \frac{1}{|S|}\sum_{s\in S}\frac{\cup_{r\in R}(s\cap r)}{|s|} \tag{12}$$

표 25은 확장된 FrameNet을 통해 부착한 의미역 결정 정보를 기반으로 문서간 유사성을 측정한 정확률과 재현율을 나타낸다. 본 논문에서 제안한 방법은 표절된 부분이 없는 문서 집합인 no_plagiarism을 제외한 artificial_low, artificial_high, simulated _paraphrase 카테고리에 대해서 정확률과 재현율을 평가하였다.

표 25. 제안한 방법론의 성능 평가

구분	표절정답 의미역의 개수(A)	추출한 의미역 개수(B)	표절정답과 일치하는 의미역의 개수(C)	정확률 (C/B)	재현율 (C/A)
no_ofuscation	154872	157740	154393	0.97878	0.99691
artificial_low	31673	30943	27634	0.89306	0.87247
aritficial_High	28627	18732	14883	0.79454	0.51991
simulated_ paraphrase	18303	9278	5724	0.61692	0.31273

본 연구에서 제안한 의미역 결정 기반의 유사성 측정 방법의 경우는 유사한 단어로 교체된 단어에 대해서도 상위어의 역할을 하는 Frame과 Frame과 의미적 관계를 갖고 있는 논항정보를 비교함으로써, 의미적 유사성을 고려하는 유사성 측정방법을 제안하였으며, 특히 단어의 의미 뿐만아니라 문서의 구조까지 고려하는 방법으로 다시 쓰기 방법으로 simulated_paraphrase에서 기존의 방법보다 개선된 성능을 확인할 수 있었다.

VI. 결론 및 향후연구

문서의 재사용은 인터넷과 스마트폰의 보급으로 인하여 정보 콘텐츠의 디지털화 과정에서 두드러지게 나타나고 있으며, 단어의 삽입, 삭제 교체, 어순의 변경 등 복잡한 형태로 이뤄지고 있다. 특히 문서 내의 단어가 같은 의미의 다른 형태를 갖는 유사 단어로 교체되었을 때, 기존의 형태적 유사성 측정의 경우 유사성 측정의 대상으로 고려되지 않는다는 문제점이 있다. 이러한 문제를 해소하기 위해 유사도 측정에 대한 다양한 연구가 수행되어 왔다.

본 연구는 의미역 결정을 이용하여 문장의 구조분석을 통한 의미적 유사성 측정방법을 제안하였으며, 의미역 결정은 문장의 서술어를 중심으로 서술어가 하나의 문장을 이룰 때 필요로하는 정보인 논항정보가 문장에서 어떤 역할을 하는지 결정하여 문서를 분석하는 방법이다. 이 과정을 통해 얻게되는 의미역 결정 정보를 문서의 특징으로 삼아 문서의 유사성 측정을 위해 사용하였다.

기존 의미역 결정 도구들은 기 구축된 언어자원에 기반하여 문서 분석을 수행하고 있기 때문에 언어자원의 범주에 따라서 문서 분석의 성능이 좌우되는 문제를 개선하기 위하여 기존 의미역 결정 도구인 FrameNet의 확장을 수행하였으며, 기존의 FrameNet과 비교를 통해 언어 범주 확장에 대한 검증과정을 수행한 결과 같은 데이터셋에 대하여 더 많은 의미역 정보를 나타낼 수 있음을 확인할 수 있었다.

본 논문에서는 두 문서간 유사성을 측정하기 위해 확장된 FrameNet을 통해 얻게된 의미역 결정 정보를 문서를 나타내는 특징 정보로 이용하였으며, 제안된 방법의 성능평가를 위해 기존 문서간 유사성 측정에 활용되고 있는 코사인 유사도 측정 방법과 부분 문자열 비교를 통해 얻게된 결과와 비교하였다.

동일한 표절 검사를 위한 말뭉치에 대해 각각의 유사성 측정 방식을 적용한 결과를 비교하였다. 본 논문에서 제안한 방법은 문서의 수정을 많이 가하지 않는 표절 문서에 대해서는 기존의 방법과 유사한 수준의 성능을 보였으나, 유사어로의 교체, 문장 구조의 변형이 이루어진 다시쓰기 문장의 경우에는 기존의 방법들 보다 개선된 성능을 확인 할 수 있었다.

본 논문은 문서의 의미적 유사성 측정을 통해 다시쓰기 유형과 같이 변형이 많이 이루어진 표절에 대한 검출 성능을 향상시키기 위한 연구로써, 본 논문에서 제안한 방법의 경우 현재 사용되고 있는 형태적 유사성 위주의 표절 시스템에 추가적인 기능으로 적용하여, 내용적 유사성 판단을 위해 적용한다면, 형태적 유사성과 의미적 유사성을 고려한 표절 시스템을 제공할 수 있을 것으로 판단된다.

참고문헌

[1] 저작권심의조정위원회, “저작권표준용어집”, 1993.

[2] https://en.wikipedia.org/wiki/World_Wide_Web

[3] Speier, C., Valacich, J., Vessey, I. “The Influence of Task Interruption on Individual Decision Making: An Information Overload Perspective,” Decision Science, Vol. 30, No. 2, pp.337-360, 1999.

[4] https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

[5] Leuf, B., Cunningham, W. “The Wiki Way: Collaboration and Sharing on the Internet,” 2001.

[6] Chowdhury, H. A., Bhattacharyya. D. K. “Plagiarism: Taxonomy, Tools and Detection Techniques,” 2018.

[7] Daniel, B., Torsten, Z. Iryna, G. “Text Reuse detection using a composition of text similarity measures,″ Proceeding of COLING 2012, pp.167-184, 2012.

[8] 황인수, “연관분석을 이용한 효과적인 표절검사 및 문서분류에 관한 연구,” 정보시스템연구, 제23권, 제3호, pp. 127-142, 2014.

[9] 조준희, “한국어 문서 표절 검사를 위한 LSA와 N-gram 기반의 유사 문장 판별” 고려대학교 대학원 석사학위 논문, 2009.

[10] Mihalcea, R., Corley, C., Strapparava, C. “Corpus-based and Knowledge-based Measures of Text Semantic Similarity,″ In Proceedings of the 21st National Conference on Artificial Intelligence, pp. 775 - 780, 2006.

[11] Sumathy. K. L., Chidambaram, “A Hybrid Approach for Measuring Semantic Similarity between Documents,″ Journal of Advanced Computer Science and Application, Vol. 7, No. 8, pp.231-237, 2016.

[11] S. Brin, J. Davis, H. Garcia-Molina, “Copy detection mechanisms for digital documents,” ACM SIGMOD Record, Vol. 24, pp.398-409, 1995.

[12] Richard. A. P., “A Little Book of Plagiarism”, 2007.

[13] Liu, D., Gildea, D. “Semantic Role Features for Machine Translation,” Proceeding of COLING 2010, pp. 716-724, 2010.

[14] Woodsend, K., Lapata, M. “Text Rewriting Improves Semantic Role Labeling”, Proceeding of Artificial Intelligence, pp. 5095-5099, 2017.

[15] Boas, H. C. “From Theory to Practice : Frame Semantics and the Design of FrameNet”, In. S. Langer & D. Schnorbusch, pp.1-29, 2005.

[16] Maurer, H. A., Zaka, B. “Plagiarism-A Survey,” Journal of Universial Computer Science, Vol. 12, pp. 1050-1084, 2006.

[17] 이준웅, “표절의 이해” , 2015

[18] Clough, P., Gaizauskas, R., Piao, S. S. L., Wilks, Y. “METER : MEasuring TExt Reuse,” Proceeding of the Association for Computational Linguistics, pp. 152-159, 2002.

[19] Pera, M. S., Ng, Y. K. “SimPaD : word similarity sentece-based plagiarism detection tool on web document,” Journal of Web Intelligence and Agent Systems, Vol. 1, pp.1-15, 2009.

[20] Smadi, M., S., Jaradat, Z., Ayyoub, M., Jararweh, Y. “Paraphrase identification and Semantic Text Similarity Analysis in Arabic News Tweets Using Lexical, Syntatic, and Semantic Features,” Information Processing and Management, Vol. 53, pp.640-652, 2017.

[21] White, D., Joy, M. “Sentence-based Natural Language Plagiarism Detection,″ ACM Journal on Educational Resources in Computing, Vol. 4, No. 4, pp. 1-20, 2004.

[22] Koberstein, J., Ng. Y. K, "Using Word Cluster to Detect Word Similar Web Documents," Proceedings of the Knowlege Science, Engineering and Management, LNAI 4092, pp.215-225, 2006.

[23] 최동진, "지능적 문서 분석을 위한 개선된 WSD 방법 연구" 조선대학교 대학원 박사학위논문, 2015.

[24] Alzahrani, S., Salim, N. "Fuzzy Semantic-based String Similarity for Extrinsic Plagiarism Detection," CLEF 2010, pp.1-7, 2010.

[25] Wise, M. "Running Karp-Rabin Matching and Greedy String Tiling," 1993.

[26] Narayanan, S. "SCAM : A Copy Detection Mechanism for digital Documents," Proceedings of Theory and Practice of Digital Libraries, pp.1-12, 1995.

[26] Xiao, C., Wang, W., Lin, X., Yu, X. "Efficient similarity joins for near-duplicate detection," Proceeding of the World Wide Web, pp. 131-140, 2008.

[27] 임마누, 김종익, "유사도 검색을 위한 데이터 재배열을 이용한 공간 효율적인 역 색인 기법", 정보과학회논문지, 제42권, 제 10호, pp.1247-1253, 2015.

[28] Liu, Y., Sun, C., Lin, L., Zhao, Y., Wang, X. "Computing Semantic Text Similarity Using Rich Features," Proceeding of the Language, Information and Computation, pp. 44-52, 2015.

[29] Georgina, C., Mike, J. "Evaluating the Performance of LSA for Source-code Plagiarism Detection," Informatica, Vol. 36, pp. 409-424, 2012.

[30] Miller, G. A. "Wordnet : a lexical database for English," Communications of the ACM, Vol. 38, No. 11, pp.39-41, 1995.

[31] Pedersen, T., Patwardhan, S., Michelizzi, J. "WordNet :: Similarity : measuring the relatedness of concepts," paper presented at the Demonstration paper at hlt-naacl 2004.

[32] Resnik, P. “Semantic Similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language,” JAIR, pp.95-130, 1999.

[33] Uzuner, O., Katz, B., Nahnsen, T.“Using Syntatic Information to Identifying Plagiarism,” Proceeding of the ACL Workshop on Educational Applications, pp.37-44, 2005.

[34] Yakushiji, A., Miyao, Y., Ohta, T., Tateisi, Y., Tsujii, J. "Automatic Construction of Predicate-argument Structure Patterns for Biomedical Information Extraction," Proceeding of the Empirical Methods in Natural Language Processing, pp.284-292, 2006.

[35] Das, D., Chen, D., Martins, A. F. T., Schneider, N., Smith, N. A., “Frame-Semantic Parsing,” Journal of Computational Liguistics, Vol. 40, Issue, 1, pp. 9-56, 2014.

[36] Che, W., Liu, T., Li, Y. “Improving Semantic Role Labeling with Word Sense,” Human Language Technologies: Proceeding of the North America Chapter of the ACL, pp.246-249, 2010.

[37] Hartmann, S. “Knowlege-based Supervision for Domain-adaptive Semantic Role Labeling”, Dissertation, 2017.

[38] Shi, L., Mihalcea, R. “Putting Pieces Together : Combining FrameNet, VerbNet, and WordNet for Robust Semantic Parsing,” LNCS 3406, pp.100-111, 2005.

[39] Palmer, M., Gildea, D. “The Proposition Bank : An Annotated Corpus of Semantic Roles,” Journal of Computational Linguistics, Vol.31, Issue. 1, pp.71-106, 2005.

[40] Fillmore, C. J., Baker, C. “A frames approach to semantic analysis,” In Bernd Heine and Heiko Narrog, editors, The Oxford Handbook of Linguistic Analysis, pp. 791-816, 2009.

[41] https://framenet.icsi.berkeley.edu/fndrupal/frameIndex.

[42] Palmer, A., Sporleder, C. “Evaluating FrameNet-style semantic parsing: the role o

f coverage gaps in FrameNet," Proceeding of COLING 2010, pp.928-936, 2010.

[43] Johansson, R., Nugues, P. "Using WordNet to Extend FrameNet Coverage," Buildi ng Frame Semantics Resources for Scandinavian and Baltic Languages, pp.1-4, 20 07.

[44] Rouces, J., Melo, G., Hose, K. "FrameBase : Enabling Integration of Heterogeneou s Knowledge," Journal of Semantic Web, Vol. 8, No. 6, pp.817-850, 2017.

[45] https://en.wikipedia.org/wiki/Paraphrase

[46] Pavlick, E., Rastoogi, P., Ganitkevitch, J., Durme, B. V., Callison-Burch, J. "PPDB 2.0 : Better paraphrase ranking, fine-grained entailment relations, word embeddin g, and style classification," Association for Computational Linguistics, pp.425-430, 2015.

[47] http://paraphrase.org/#/download

[48] Pavlick, E., Bos, J., Nissim, M., Beller, C., Durme, B. V., Callison-Burch, C. "Addi ng semantics to data-driven paraphrasing", Association for Computational Linguis tics, pp. 1512-1522, 2015.

[49] https://github.com/Noahs-ARK/semafor

[50] http://www.anc.org/data/masc/downloads/data-download/

[51] Rajaraman, A., Ullman, J. D. "Mining of Massive Datasets", 2011.

[52] Manning, C. D., Raghavan, P., Schütze, H. "Introduction to Information Retrieval", Cambridge University Press. 2008.

[53] Marneffe, M., Manning, C. D. "Stanford Typed Dependecies munual," Stanford Pa rser, 2010.

[54] Das, D. "Statistical Models for Frame-Semantic Parsing", Proceeding of Frame Se mantics in NLP, Association for Computational Linguistics, pp.26-29, 2014.

[55] Potthast, M. et al. “Overviews of the 4th International Competition on Plagiarism Detection,” CLEF 2012 Evaluation Labs and Workshop, pp.1-28, 2012.

[56] http://pan.webis.de

[57] Potthast, M., Stein, B., Cedeno, A. B., Rosso, P. “An evaluation framework for pla giarism detection,” Proceeding of the COLING 2010, pp. 1-9, 2010.