http://dx.doi.org/10.3745/KIPSTB.2012.19B.2.135

술어-논항 튜플 기반 근사 정렬을 이용한 문장 단위 바꿔쓰기표현 유형 및 오류 분석

최 성 필[†] · 송 사 광[†] · 맹 성 현[††]

요 약

본 논문에서는 Predicate-Argument Tuple (PAT)를 기반으로 텍스트 간 심층적 근사 정렬(Approximate Alignment)을 통한 문장 단위 바꿔쓰기표현(sentential paraphrase) 식별 모델을 제안한다. 두 문장 간의 PAT 기반 근사 정렬 결과를 바탕으로, 두 문장의 의미적 연관성을 효과적으로 표현하는 다양한 정렬 자질(alignment feature)들을 정의함으로써, 바꿔쓰기표현 식별 문제를 지도 학습(supervised learning) 기반의 자동 분류 모델로 접근하였다. 실험을 통해서 제안 모델의 가능성을 확인할 수 있었으며, 시스템의 오류 분석을 통해 제안 방법이 아직 해결하지 못하는 다양한 바꿔쓰기표현 유형들을 식별함으로써 향후 시스템의 성능 개선 방향을 도출하였다.

키워드 : 바꿔쓰기표현 인식, 술어-논항 구조, 근사 정렬, 텍스트 함의, 텍스트 마이닝, 기계학습

Analysis of Sentential Paraphrase Patterns and Errors through Predicate-Argument Tuple-based Approximate Alignment

Sung-Pil Choi[†] · Sa-Kwang Song[†] · Sung-Hyon Myaeng[††]

ABSTRACT

This paper proposes a model for recognizing sentential paraphrases through Predicate-Argument Tuple (PAT)-based approximate alignment between two texts. We cast the paraphrase recognition problem as a binary classification by defining and applying various alignment features which could effectively express the semantic relatedness between two sentences. Experiment confirmed the potential of our approach and error analysis revealed various paraphrase patterns not being solved by our system, which can help us devise methods for further performance improvement.

Keywords : Paraphrase Recognition, Predicate-Argument Structure, Textual Entailment, Text Mining, Machine Learning

1. 서 론

기반 기술의 발전과 함께, 표면적 상이성을 가진 두 텍스트의 의미적 연관성 및 유사성을 식별하는 연구가 매우 활발하게 진행되어 왔다[1-11]. 바꿔쓰기표현(paraphrase)란 텍스트 내에서 "거의" 동일한 정보를 제공하기 위해 선택할 수 있는 다양한 표현 방법들이다. 예를 들어, "*나는 책을 본다*"와 "*나는 책을 읽는다*"는 동일한 의미로 인식된다. "*우리는 밥을 먹는다*"와 "*우리는 식사를 한다*"도 역시 동일한 의미를 나타낸다. 일반적으로 일상 언어 생활에서 혹은 텍스트 내에서 셀 수 없는 바꿔쓰기표현을 발견하거나 스스로 만들 수 있다.

언어처리 연구에서 이러한 바꿔쓰기표현에 대한 식별 및 처리는 매우 중요한데 그 이유는 정보검색(Information Retrieval), 질의응답(Question Answering) 그리고 문서요약(Summarization) 등에서 특정 의미에 대한 서로 다른 표층적 언어표현들이 성능 저하의 중요한 요인 중의 하나였기 때문이다[1,12,13]. 초기에는 특정 요소 분야에서 종속된 연구로서 개별 시스템의 성능을 높이는 요인으로 연구되었으나, 최근 들어, 그 중요성이나 난이도에 기인하여 규모가 큰 독립적인 연구 분야로 인식되고 있다[14].

기존 바꿔쓰기표현 분석 방법의 대부분은 두 텍스트 간의 세밀한 유사도 측정 모델에 주로 의존한다. 특히 지금까지 고안된 많은 방법론들이 병렬 말뭉치(parallel corpora)에서 단어 치환 테이블(word replacement table)이나 구절 치환

† 정 회 원 : 한국과학기술정보연구원 SW연구실 선임연구원
†† 종신회원 : 한국과학기술원 전산학과 교수(교신저자)
논문접수 : 2012년 2월 7일
수 정 일 : 1차 2012년 2월 28일
심사완료 : 2012년 3월 2일

테이블(phrase replacement table)을 생성하기 위한 어휘적 정렬(lexical alignment) 기법[1]에 근간을 둔 유사도 측정 모델을 많이 채용하였다[3,4,15-17]. 그러나 단일 언어로 표현된 두 텍스트 간의 단어 및 구절 정렬 성능은 현재까지 많은 노력을 기울여서 구축된 대용량 다국어 병렬 말뭉치에서의 정렬 성능을 따라가지 못하고 있다[4,18]. 가장 큰 이유는 단일 언어로 표현되고 동일한 의미를 나타내는 문장 쌍 집합으로 구성되는 비교 말뭉치(Comparable Corpora)의 구축 및 확보가 용이하지 않기 때문이다. 따라서 최근 연구들은 보다 면밀한 언어 처리를 통해서 도출될 수 있는 텍스트 내의 핵심적인 의미 요소(semantic element)를 기준으로 유사도를 측정하는 방법을 채택하고 있다[15,19,20].

본 논문에서는 Predicate-Argument Tuple (PAT)를 기반으로 한 텍스트 간 심층적 근사 정렬(Approximate Alignment)을 통한 문장 단위 바꿔쓰기표현 식별 모델을 제안한다. 술어-논항 구조(Predicate-Argument Structure, PAS)란 문장 내에서 단어 간의 구문적, 의미적 관계를 나타내는 그래프 구조(graph structure)이다[21]. 의미역 부착에서의 술어-논항 구조보다는 좀더 세부적인 관계를 표현한다. PAT은 이러한 술어-논항 구조를 구성하는 단위 요소를 의미하며, 문장의 전체적인 의미를 표현하는데 근본적인 역할을 수행한다. 두 문장 간의 PAT 기반 근사 정렬 결과를 바탕으로, 두 문장의 의미적 연관성을 효과적으로 표현하는 다양한 정렬 자질(alignment feature)들을 정의함으로써, 바꿔쓰기표현 식별 문제를 지도 학습(supervised learning) 기반의 자동 분류 모델로서 해결하고자 한다. 또한, 부가적으로 본 논문에서 제안하는 시스템의 성능 평가 및 오류 분석을 통해 제안 방법이 아직 해결하지 못하는 다양한 바꿔쓰기표현 유형들을 식별함으로써 향후 시스템의 성능 개선 방향을 도출한다.

논문의 구성은 다음과 같다. 우선 2장에서는 바꿔쓰기표현 식별에 대한 다양한 관련 연구에 대해서 소개한다. 이어서 3장에서는 본 논문에서 제안하는 PAT 기반 근사 정렬에 의한 바꿔쓰기표현 식별 접근 모델에 대해서 상세하게 소개한다. 이어서 4장에서는 성능 측정 실험 결과를 제시하고, 5장에서는 시스템의 오류에 대한 상세한 분석을 시도한다. 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

바꿔쓰기표현 식별은 텍스트에서 동일한 의미를 나타내는 상이한 형태의 문장, 구절 등을 찾아내는 작업이다. 주지한 바와 같이, 대부분의 연구가 심도 깊은 자연어 처리를 통한 두 텍스트 간의 유사도 측정 모델에 근간을 두고 있다. 지금까지 매우 활발한 연구가 이루어졌으며 많은 논문이 발표되었으나[19,22-25] 여기서는 이들 중 본 연구와 관련하여 가장 중요하다고 판단되는 연구 결과를 중심으로 소개한다.

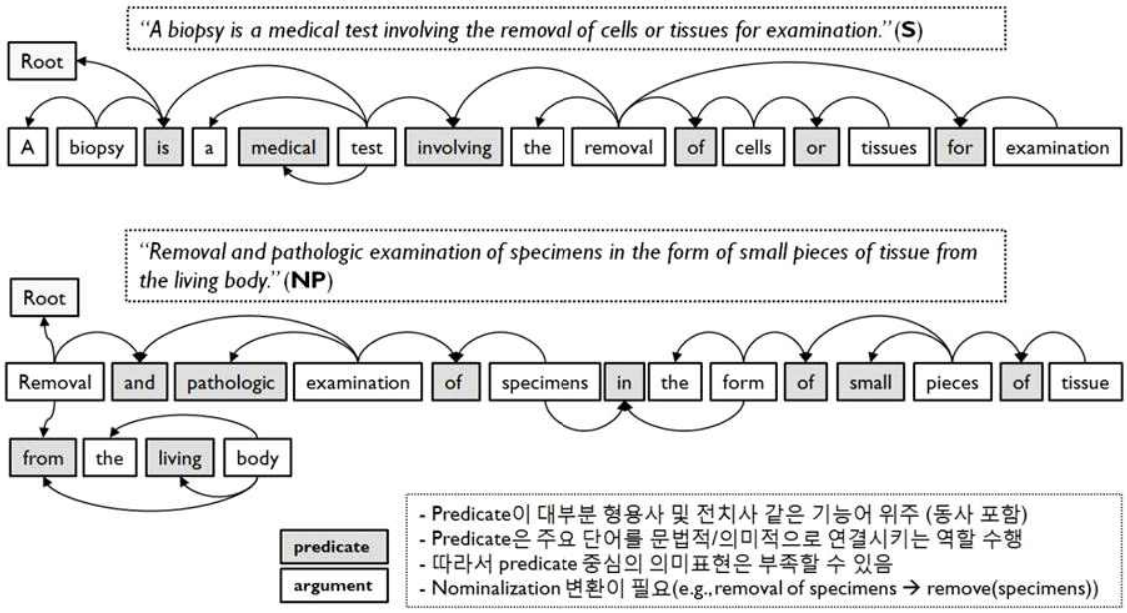
우선 [19]는 두 문장 간의 의미적 연관성을 측정함에 있어서, 유사점(similarity) 뿐만 아니라 차이점(dissimilarity)도 함께 고려되어야 한다고 주장한다. 문장 비교를 위해서 "정보 덩어리(Information Nugget)"라는 비교 요소를 정의하는데 이는 의미역 부착(Semantic Role Labeling, SRL) 결과로 도출되는 술어-논항 구조(Predicate-Argument Structure, PAS)이다. 예를 들어, "*Oswald killed Kennedy*"라는 문장에 대한 의미역 부착 결과는 "*killed(Oswald, Kennedy)*"이다. 유사도 비교는 이들 PAS 간의 중첩 정도를 바탕으로 이루어지는데, 이 때, 개별 PAS의 중요도를 계산하여 적용하게 된다. 즉 문장의 의미 표현에 직접적으로 연관되지 않은 PAS들은 아무리 중첩이 되어도 유사도가 높지 않게 된다. 예를 들어, "*It is said* ~"나 "*He told me that* ~" 등과 같은 구절은 PAS로는 구성될 수 있으나 문장 내에서 의미적인 관점에서의 역할이 매우 낮다. 문장을 구성하는 개별 PAS의 중요도 측정을 위해서 [19]에서는 문장 내에서의 구문적 위치 정보를 활용하여 자질화하고 기계학습을 이용한 이진 분류기를 구성하여 처리하였다. 접근 방법의 참신함에도 불구하고 이 연구는 몇 가지 한계점이 있는데, 그 중 하나는 전체 시스템의 성능이 의미역 부착 성능에 상당 부분 의지하고 있다는 점과, 개별 PAS의 중요도 측정 기법이 매우 단순한 형태의 자질 활용에 기인할 수 있는 오류 가능성을 가지고 있다는 점 등이다.

[7]는 보는 시각에 따라서 매우 단순하고 극단적인 방법을 사용하여 두 문장 간의 유사도를 측정한다. 이 논문의 핵심은 복잡한 자연어 처리나 의미 표현 기법 등을 사용하지 않고, 다양한 형태의 단순 유사도 기반 자질들을 추출하여 기계학습 모델을 활용한다면 충분히 기존 성능에 버금가는 바꿔쓰기표현 인식 시스템을 구성할 수 있다는 것이다. 이를 위해서 [7]에서는 두 문장에 대한 다양한 종류의 유사도 계산 수치로 구성되는 자질 벡터를 정의하였다. 원본 문장에 대한 토큰 기반 문자열 유사도, WordNet 기반의 동의어 치환 문자열 유사도, 의존 문법 관계 겹침 정도로 구성되는 세 가지 종류로 구성된 총 136가지의 유사도를 계산하고, 자질 선택 기법을 통해 걸러진 133가지 자질을 중심으로 지도학습 기반 바꿔쓰기표현 인식 시스템을 구성하였다. Microsoft Research Paraphrase Corpus(MSRPC)[26]를 이용한 실험에서 우수한 성능을 나타냈으나, 활용한 말뭉치 역시 기계적인 유사도 측정을 통해 구축된 말뭉치인 관계로, 표층적인 유사도 계산 방법이 효과를 나타낼 수 있기 때문에 접근 방법의 범용성 측면에서는 아직 고려할 사항이 많다.

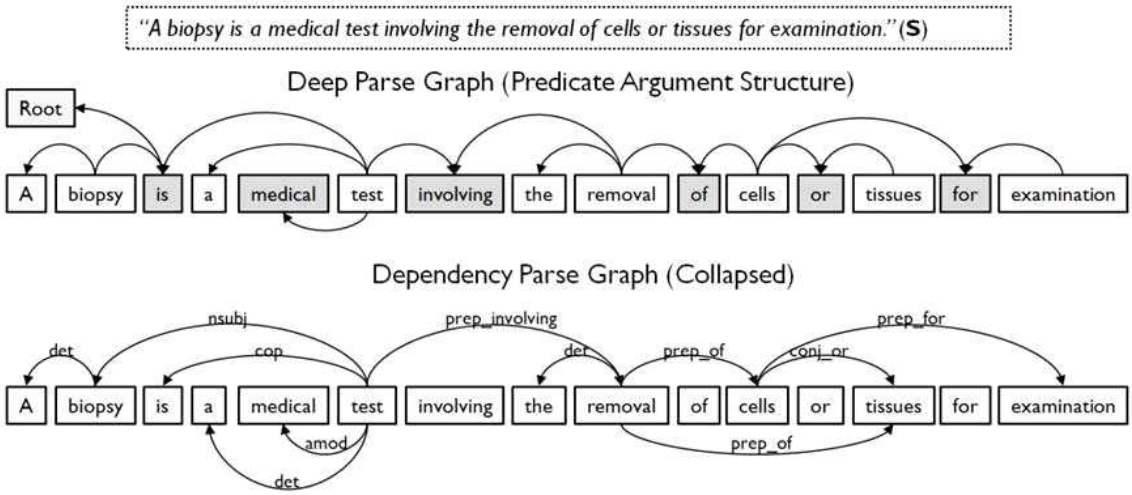
3. 제안 시스템

이 장에서는 본 논문에서 두 문장 간의 의미적 연관성을 측정하기 위해서 제안하는 PAT 기반 근사 정렬에 대해서 자세히 살펴본다. 그 전에 문장에 대한 의미 표현을 형식화하기 위한 단위 요소인 Predicate-Argument Tuple에 대해서 설명한다.

1) 통계적 기계 번역(Statistical Machine Translation)에서 디코딩(decoding)을 위한 기반 자원(변환 테이블)을 수집하기 위해 Giza++과 같은 단어 정렬 프로그램을 수행하여 서로 다른 언어로 쓰여진 두 문장을 연동한다.



(그림 1) 술어-논항 구조(PAS) 예제



(그림 2) 술어-논항 구조와 의존 구문 그래프

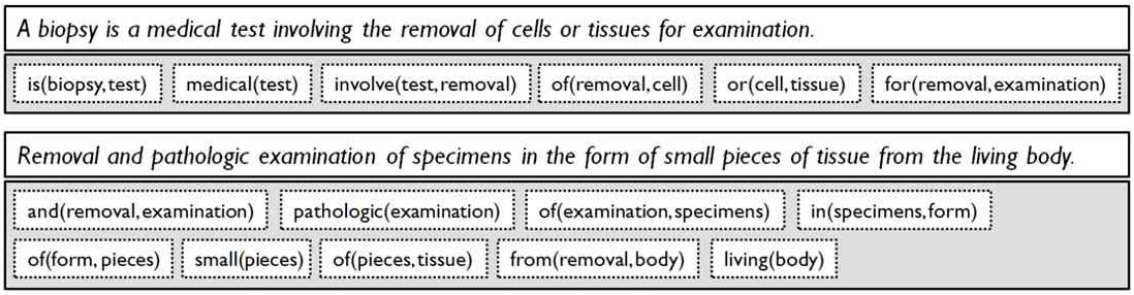
3.1 술어 논항 튜플 (Predicate-Argument Tuple, PAT)

술어-논항 구조(Predicate-Argument Structure, PAS)란 문장 내에서 단어 간의 구문적, 의미적 관계를 나타내는 그래프 구조(graph structure)이다[21]. 의미역 부착에서의 술어-논항 구조보다는 좀더 세부적인 관계를 표현한다. (그림 1)은 [21]에서 개발한 Enju parser[2]를 이용하여 특정 문장에 대한 술어-논항 구조를 분석한 예이다.

위 그림에서 진하게 표시된 사각형은 술어(predicate)를, 나머지는 논항(argument)을 나타낸다. 화살표는 술어와 논항 간의 문법적/의미적 관계를 표시하고 있다. 예를 들어, (그림 1)의 첫째 문장에서 "*involving*"이라는 술어는 "*test*"와 "*removal*"을 논항으로 가지고 있으며, 둘째 문장에서 "*pathologic*"이라는 형용사는 "*examination*"이라는 명사를 논항으로 소유함을 보여주고 있다. 부가적으로 등위접속사인 "*or*"나 "*and*"도 각각 개별적인 술어가 되며 이들 접속사를 통해 대등하게 연결된 단어들을 논항으로 가지고 있다 (e.g., "*cell or tissue*", "*removal and examination*").

앞에서도 잠시 언급하였으나, 통상적으로 술어는 대부분 동사, 형용사, 전치사 등과 같은 분야 독립적이거나 기능 중심적인 어휘들로 주로 구성되고 논항에는 명사를 중심으로 기술 용어 등과 같은 분야 의존적인 어휘들이 많이 출현한다. 술어는 그 기능 특성상 논항으로 대표되는 주요 핵심 단어들을 문장 내에서 문법적/의미적으로 연동시키는 역할을 주로 수행한다.

2) http://www-tsujii.is.s.u-tokyo.ac.jp/enju/



(그림 3) Predicate-Argument Tuples (PATs)

(그림 2)는 동일 문장에 대한 술어-논항 구조와 의존 구문 그래프를 비교하여 나타내고 있다. 의존 구문 분석을 위해서는 Stanford parser[3]를 사용하였다. 위 그림에서 제시된 분석 결과는 "*collapsing*"된 분석 결과이다. 이는 최종 분석 결과에 대해 후처리를 수행하여 전치사나 등위접속사 등에 대한 특수 처리를 통해서 불필요한 의존 관계를 정리하고 통합한 결과이다. 그림 상으로는 차이가 있으나 두 가지 분석 결과는 거의 동일하다. 예를 들어, PAS에서의 "*of*(*removal*, *cell*)"은 의존 그래프에서의 "*prep_of*(*removal*, *cells*)"와, "*medical*(*test*)"는 "*amod*(*medical*, *test*)"와 동일한 구조를 보여주고 있다. 그러나 본 논문에서는 생명과학 분야 기술 용어 및 문헌 집합을 다루고 있으므로, 이 분야에 최적화된 분석 모델을 지원하고, 술어-논항 구조를 추출하기에 편리한 출력 형태를 제공하는 Enju parser를 기본 구문 분석 모듈로 활용한다.

(그림 3)은 Enju parser를 이용하여 도출된 PAS에서 Predicate-Argument Tuple (PAT) 집합이 추출된 결과를 보여주고 있다. PAT란 단일 문장에 대한 PAS를 구성하는 요소를 나타내며, 문법적인 측면에서 크게 연결형(connection-oriented), 동사형(verbal), 형용사형(adjectival), 그리고 명사형(nominal) PAT 등으로 구분될 수 있다. 연결형 PAT은 전치사 및 등위접속사가 술어인 PAT로서, 여러 개의 단어(논항)들을 문법적으로 연결하는 역할을 수행한다 (e.g., "*of*(*removal*, *cell*)"). 동사형 PAT는 일반 동사를 술어로 가지며, 문장 내에서 핵심적인 의미 요소 역할을 수행할 가능성이 높다 (e.g., "*involve*(*test*, *removal*)"). 형용사형 PAT은 명사를 수식하는 형용사가 술어인 PAT이고, 명사형 PAT은 복합 명사(compound noun) 내에서 중심어(headword)를 수식하는 또 다른 명사가 술어인 PAT이다. 단일 문장에서 문법적 역할 특성에 따라 구분된 PAT 유형을 기반으로 문장 간 유사도 측정에 활용할 수 있다면 보다 향상된 결과를 얻을 수 있다.

3.2 PAT 기반 근사 정렬 (PAT-based Approximate Alignment)

본 절에서는 PAT 기반 근사 정렬 방법에 대해서 자세히 다룬다. 먼저 PAT 유사도 측정 모델을 설명하고, 이를 통한 근사 정렬 방법을 기술한다. 두 텍스트 간에 정렬된 PAT 집합을 중심으로 최종적인 유사도 측정 모델에 대해서 마지막에 소개한다.

3.2.1 PAT 유사도 (PAT Similarity)

PAT 기반 근사 정렬을 위해서는 우선 PAT 유사도 측정을 위한 기본 모델이 필요하다. 텍스트에서 추출된 의미 요소들 간의 유사도 측정을 위한 관련 연구 중에서 [19]는 의미역 부착의 결과로 도출되는 술어-논항 구조 (Predicate-Argument Structure)들 간의 유사도 측정 모델을 제안한 바 있다. 이를 기반으로 본 연구에서 사용하는 두 개의 PAT 간 유사도 측정을 위한 기본 모델을 다음과 같이 정의한다.

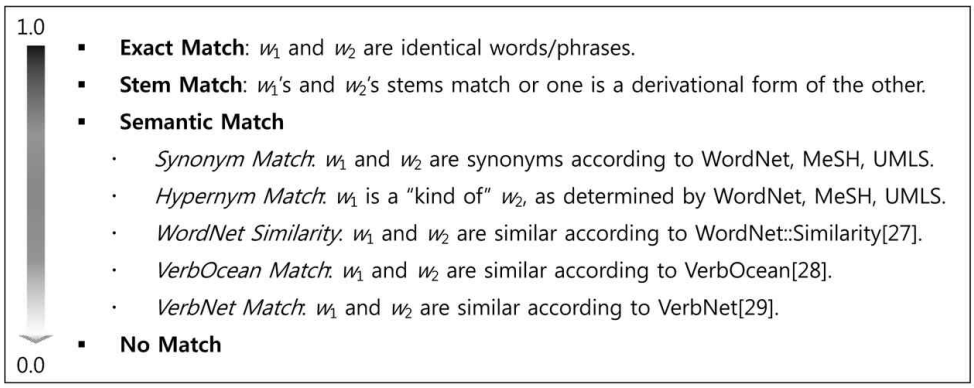
$$sim(P_1,P_2)=\frac{1}{\alpha}\left[\left(\sum_{a_1\in \arg(P_1),a_2\in \arg(P_2)} sim_a(a_1,a_2)\right)+w_p\times sim_p(p_1,p_2)\right] \quad (1)$$

P_i는 입력 PAT을 의미하고, a_i는 P_i의 특정 논항을, p_i는 P_i의 술어를 나타낸다. sim_a는 논항 간의 유사도(Argument Similarity) 계산 함수이고, sim_p는 술어 간의 유사도 (Predicate Similarity) 계산 함수이다. w_p는 술어 유사도 수치에 대한 중요도 가중치를 나타내고 마지막으로 α는 정규화 수치를 가리킨다.

식 (1)은 PAT를 구성하는 술어와 논항의 어휘적 특성에 적합하게 고안된 단어 기반 유사도 함수를 활용하며, 술어 유사도와 논항 유사도 간의 가중치 평균을 계산한다. 술어 및 논항 비교에 사용되는 단어간 유사도 측정 방법은 아래 (그림 4)와 같다.

(그림 4)는 [20]에서 두 개의 의존 그래프 간의 개별 노드에 대한 유사도 측정을 위해서 고안한 일치 방법 중심의 유사도 점수 배정표이다. 유사도 점수는 1.0인 "완전 일치 (Exact Match)"에서부터 시작하여 "어근 일치(Stem Match)", 언어자원을 활용한 "의미 기반 일치(Semantic Match)" 등으로 내려갈수록 점수가 낮아지게 된다. 술어 간 유사도 비교를 위해서는 특히 의미 기반 일치에서 "*VerbOcean*"이나 "*VerbNet*"을 이용한 동사 유사도 측정 기법을 사용한다. 일치 방법의 우선 순위에 따른 유사도 점수의 축소 정도(score cut degree)에 대한 최적화 결정은 표본 실험을 통해서 가능하다.

3) http://nlp.stanford.edu/software/lex-parser.shtml

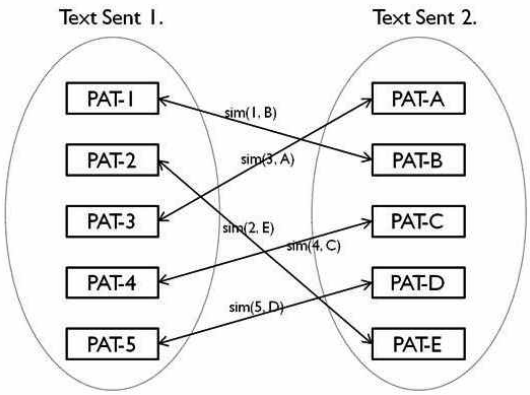


(그림 4) Argument/Predicate Similarity Measure[7,20]

3.2.2 PAT 기반 근사 정렬 알고리즘

일반적으로 자주 활용되는 GIZA++[4] [30], Berkeley Aligner[5] [31]과 같은 어휘 수준의 다국어 병렬 말뭉치 정렬 도구들의 성능이 매우 좋게 평가되고 있으므로 이들 시스템을 직접 활용하는 것이 고려될 수 있다. 그러나 본 논문에서의 정렬 기준은 단어가 아니라 PAT이고 다국어 기반이 아니라 단일 언어 기반의 정렬을 수행해야 하므로 적합하지 않다. 따라서 새로운 정렬 모델이 필요하다.

서로 다른 두 PAT 집합 간의 정렬 문제는 그래프 이론에서의 최소 비용 완전 이분 그래프 일치 모델(Minimum Cost Complete Bipartite Graph Matching Model)로서 설명될 수 있다. 이분 그래프(bipartite graph)란 전체 vertex 집합이 서로 소인 두 집합(two disjoint sets) U와 V로 분리되고, 모든 edge는 각각 U의 한 vertex와 V의 특정 vertex를 연결하는 그래프이다. 만일 두 vertex를 연결하는 모든 edge에 비용(cost)이 존재한다고 가정할 때, 최소 비용으로 두 vertex 집합 간에 완전 연결을 할 수 있는 방법을 찾는 문제는, 최대 유사도를 나타내는 두 PAT 간의 정렬 방법을 찾는 문제와 동일하다.



(그림 5) 두 문장에 대한 PAT 기반 근사 정렬

(그림 5)와 같이 두 PAT 집합 간의 정렬 문제는 3.2.1절에서 제시한 PAT 유사도를 기반으로 하여 각 연결 edge에 비용(1 / 유사도) 수치를 지정한다면, 위에서 언급한 최소 비용 완전 이분 그래프 일치 문제로 귀결될 수 있다. 이 때, 부분 정렬(partial alignment)[6]에 대한 부분은 특정 연결 edge의 비용(cost)이 무한대, 다시 말해서 두 PAT의 유사도가 0임을 가정한다면 가능하다. 이러한 최소 비용 일치 방법을 찾는 문제는 예전부터 연구가 되었으며 다음 식과 같은 선형 계획법(linear programming) 기반의 헝가리안 방법(Hungarian Method)[32]으로 해결이 가능하다.

$$\begin{aligned} &\text{Mimimize} && \sum_{i,j} c_{ij} x_{ij} && \\ &\text{subject to :} && \sum_{j} x_{ij} = 1 && i \in S_1 \\ &&& \sum_{i} x_{ij} = 1 && j \in S_2 \\ &&& x_{ij} \geq 0 && i \in S_1, j \in S_2 \end{aligned} \quad (2)$$

여기서 S_i는 문장 i에 대한 PAT 집합을 나타내며, c_{ij}는 edge (i, j)의 비용(cost)으로서 연결되는 두 PAT의 유사도의 역수이다[7]. x는 발생 벡터(incidence vector)로서 특정 정렬 방법 M과 이 벡터의 개별 항목 x_{ij}에 대해서 다음과 같은 조건이 성립된다.

$$x_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathrm{M}, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

위와 같이 구성된 선형 계획법에 따라서 최소 비용을 가지는 완전 정렬 방법 M^*를 구할 수 있으며 다음 절에서 설명되는 유사도 계산 모델에 따라서 최종적인 근사 정렬 기반 유사도가 도출된다.

4) http://code.google.com/p/giza-pp/
5) http://code.google.com/p/berkeleyaligner/
6) 정렬 대상의 특정 부분은 상호 간의 유사도가 낮은 관계로 서로 정렬이 되지 않고 남아 있는 경우를 가리킨다.
7) c_{ij} = 1 / $sim(P_i, P_j)$

3.2.3 PAT 기반 근사 정렬을 이용한 문장 유사도 측정

앞 절에서 설명한 정렬 결과로 도출되는 정렬된 PAT 쌍 집합(paired PAT set)을 기반으로 근사 정렬 유사도(approximate alignment similarity)를 다음과 같이 계산할 수 있다.

$$sim_{agn}(S_1, S_2) = \frac{1}{N} \sum_{(i,j) \in M^*} sim(p_i^1, p_j^2) \quad (4)$$

이 식에서 S_i는 i번째 문장을 나타내고, p_i^j는 S_j의 특정 PAT을 가리킨다. 또한 M^*는 3.2.2절에서 근사 정렬 결과로 도출된 최소 비용 정렬 방법이며 $sim(\cdot,\cdot)$는 3.2.1절에서 정의한 PAT 유사도를 나타낸다. 마지막으로 N은 정규화 요소이다.

(식 4)에서도 알 수 있듯이 본 논문에서의 근사 정렬 유사도는 모든 정렬된 PAT 간의 유사도의 정규화된 합이다. 이 모델의 특징은 두 텍스트의 의미 요소에 대한 중복 정도를 계산하거나, 순차적 유사도 비교 기반의 단순 정렬 기법 등을 사용하는 기존의 방법과는 달리, 두 의미 요소의 일치 형태에 대한 최적의 방법을 결정하고 이에 따른 의미 요소 간의 유사도를 전체적으로 적용함으로써 보다 엄밀한 의미적 연관성을 발굴할 수 있는 장점이 있다.

4. 실험 및 분석

4.1 Microsoft Research Paraphrase Corpus (MSRPC)

텍스트 간 심층적 유사도 측정이나 구절 혹은 문장 단위 바꿔쓰기표현 식별 연구에 대한 많은 관심과는 달리, 현재 이러한 연구를 뒷받침하고 개발되는 다양한 접근 방법을 객관적으로 검증할 수 있는 평가 컬렉션은 매우 부족한 실정이다. 이러한 가운데, 2005년에 [5]에서는 웹에서 수집된 대량의 신문 기사를 바탕으로 이를 문장 단위로 가공하여 쌍방간 유사도(단어 중첩도 및 편집 거리 등) 자동 계산을 통한 바꿔쓰기표현 관계일 가능성이 높은 문장 쌍 집합을 구축하였다. 구축된 다량의 데이터 중에서 수작업 검증을 통해서 총 5,801 쌍의 문장 집합을 선별하였으며 세부적인 통계 정보는 다음 표와 같다.

〈표 1〉 MSRPC 데이터 구성

	학습집합	검증집합	합계
Positive Instances	2,753	1,147	3,900
Negative Instances	1,323	578	1,901
합계	4,076	1,725	5,801

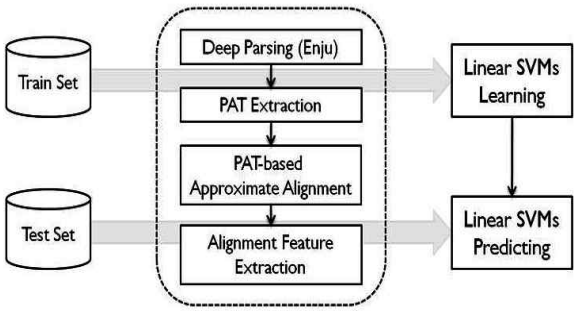
위 표에서 보는 바와 같이, MSRPC에는 서로 바꿔쓰기표현 관계인 문장 쌍과 더불어 의미적으로 관계가 없는 문장 쌍도 존재한다. 따라서 입력 문장 쌍에 대해서 바꿔쓰기표현 여부를 판단하는 이진 분류로서 시스템 평가가 가능하도록 구성되어 있다. 다음 표는 MSRPC 데이터의 예를 나타낸다.

〈표 2〉 MSRPC 예시

Sentence Pairs	label
• He said the foodservice pie business doesn't fit the company's long-term growth strategy. • The foodservice pie business does not fit our long-term growth strategy.	Positive
• Around 0335 GMT, Tab shares were up 19 cents, or 4.4%, at A$4.56, having earlier set a record high of A$4.57. • Tab shares jumped 20 cents, or 4.6%, to set a record closing high at A$4.57.	Negative
• But he added group performance would improve in the second half of the year and beyond. • De Sole said in the results statement that group performance would improve in the second half of the year and beyond.	Positive

4.2 실험 시스템의 구성

MSRPC 기반의 최신 연구들과 마찬가지로 본 연구에서도 기계학습 기법을 활용한다. 다음 그림은 전체적인 실험 시스템의 실행 흐름도를 보여준다.



(그림 6) 실험 시스템의 구성 및 흐름도

우선 [21]에서 개발한 Enju parser를 이용하여 입력 문장에 대한 구문분석을 수행하였다. PAT 기반 근사 정렬을 위해서는 통합 그래프 분석 라이브러리인 LEMON[8]을 활용하였으며, LibSVM[9]을 이용한 학습 및 이진 분류를 수행하였다. 개별 입력 문장 쌍에 대한 PAT 기반 근사 정렬의 결과를 바탕으로 두 문장의 정렬 유사도, 정렬 형태, 정렬되지 않은 PAT 집합 등의 정보를 활용하여 <표 3>과 같이 5 가지의 자질 정보를 추출할 수 있었으며, 이에 대한 세부적인 설명은 다음 표와 같다.

8) Bipartite Graph Matching Optimizer (http://lemon.cs.elte.hu/trac/lemon)
9) http://www.csie.ntu.edu.tw/~cjlin/libsvm/

〈표 3〉 근사 정렬 기반 자질 집합

자질 이름	설명
paas	PAT-based Approximate Alignment Similarity
root_pred_sim	두 문장의 root 술어간 어휘 유사도
root_pat_sim	두 문장의 root PAT간 PAT 유사도
pat_sim_one_ratio	정렬된 PAT 집합에서 유사도가 1.0인 PAT의 비율
min_pat_sim	정렬된 PAT 집합에서 최저 유사도

본 실험 시스템에는 대상 데이터의 특성이나 성능 최적화를 위한 다양한 매개변수가 존재한다. 이들 각각에 대한 설명과 본 논문에서 사용한 매개변수 수치가 아래 표에 나타나있다.

〈표 4〉 시스템 매개변수에 대한 설명과 실험을 위한 설정 수치

매개변수	설명	설정 수치
lex_sim_thresh	어휘유사도 기반 논항정렬에서 정렬이 될 최소 유사도	0.2
pred_weight	PAT 유사도 계산 시에 술어에 대한 가중치	1.4
min_wordnet_sim	WordNet 유사도의 최소 임계치	0.2
min_mesh_sim	MeSH 유사도의 최소 임계치	0.2
min_patc_sim	PAT 클러스터링 기반 유사도의 최소 임계치	0.2
patc_sim_check	PAT 클러스터링 기반 유사도 계산 여부	False
ppr_sim_thresh	근사 정렬 기반 바꿔쓰기표현 판정에서의 기준 유사도	0.5

시스템의 성능을 높이기 위한 전면적인 최적화 작업은 수행하지 않았다. 대신, 학습 집합 일부를 활용하여 작은 규모의 자료에서의 근사 정렬 결과를 바탕으로 매개변수에 대한 경험적 수치를 선택하여 지정하였다. 향후 연구에서는 이 부분에 대한 보다 엄밀하고 체계적인 접근 방안이 요구된다.

4.3 실험 결과

아래 표에 본 논문에서 제안한 시스템에 대한 성능 실험 결과와 기존 연구 결과와의 비교 수치를 제시하였다. 앞에서도 지적하였듯이 MSRPC는 학습 집합과 검증 집합으로 나뉘는데, 학습 집합을 통해서 학습을 수행하고 검증 집합을 이용하여 시스템의 최종 성능을 판정하였다. 본 논문에서의 제안 시스템을 포함하여 다음의 총 6 개의 시스템 성능을 세 가지 평가 기준으로 비교하였다.

비록 전체적인 성능은 정확도(accuracy) 기준으로 볼 때, 그 순위가 낮게 나타나고 있으나, (1)성능의 차이가 미미한 점(2.6% 미만), (2) SVM 및 근사 정렬 매개변수 등의 설정

〈표 5〉 성능 실험 결과 및 기존 시스템과의 비교

	Accuracy	Precision	Recall
Mihalcea et al., 2006 [33]	70.30	69.60	97.70
Qiu et al., 2006 [19]	72.00	72.50	93.40
Lintean & Rus, 2009	72.06	74.04	89.28
Fernando & Stevenson, 2008 [24]	74.10	75.20	91.30
Das & Smith, 2009 [34]	73.86	79.57	86.05
Our Method (5 feats, linear)	71.53	73.42	89.62

에 따른 최적화를 수행하지 않은 점, 마지막으로 (3) 근사 정렬의 결과로 도출되는 5가지의 자질만을 활용하였다는 점에서 고무적인 결과로 판단할 수 있다. 위에서 제시한 대부분의 기존 방법론들이 MSRPC 말뭉치를 표본 추출하여 이를 분석하고 두 문장의 의미적 동질성 식별에 필요한 단순한 규칙 집합을 사용하여 접근하고 있는데 반하여 본 논문은 문장에서 의미요소를 식별하고 이를 자동으로 정렬함으로써 얻을 수 있는 다양한 유사도 측정치를 자질로 활용한 점에서 그 장점이 있다. 본 논문은 이러한 확장성 측면에서 향후 지금보다 높은 성능치를 확보할 수 있는 기반을 구축했다으며, 추후 연구를 위한 시작점으로서 심층적인 오류 분석을 수행하였다.

추가적으로 세부적인 오류 분석 내용에서도 언급하였으나, MSRPC가 전 세계적으로 가장 많이 활용되고 있는 말뭉치임에도 불구하고 일관성 측면에서 많은 오류를 내포하고 있다. 이러한 점은 기존의 접근 방법들에서 추구한 데이터 표본 추출 방식의 단순한 규칙기반 방법에 의한 성능 측정 결과가 절대적인 시스템의 고유 성능이라고 보기 어렵다는 사항을 간접적으로 알려주고 있다. 본 논문에서는 동 연구 분야에서 가장 널리 알려진 MSRPC를 이용하여 본 논문의 제안시스템의 한계점과 더불어 바꿔쓰기표현 식별의 체계적인 방법론에 대한 기반 정보를 독자들에게 알려주기 위한 목표를 명확하게 하고 있다. 논문의 제목, 초록 및 본문 내에서도 언급하였으나 본 논문에서 제안하는 시스템의 가장 중요한 특징 중의 하나가 두 문장 내에 존재하는 서로 유사한 의미요소를 자동으로 정렬하는 능력이며 이를 통해서 오류 분석이 보다 원활하게 이루어질 수 있음을 강조하고 있다.

5. 바꿔쓰기표현 식별 오류 분석

문장 단위 바꿔쓰기표현 식별을 위한 새로운 접근 모델을 제안하고, 이에 대한 초기 성능 실험과 더불어 본 장에서는 시스템이 도출한 두 가지 오류인 거짓 양성(false positive) 오류와 거짓 음성(false negative) 오류에 대한 심층 분석을 시도한다. 이를 통해서, 바꿔쓰기표현 식별 과정에 있어서 제안 시스템의 한계점을 파악하여 향후 시스템의 성능 개선을 도모할 수 있으며, 문장 간의 바꿔쓰기표현 양상 분석을 통해서 보다 포괄적이고 정확한 접근 모델을 도출할 수 있다.

(그림 7) PAT 기반 근사 정렬 결과

5.1 오류 분석 방법

실험에 사용된 검증 집합(총 1,725 개) 내에서 본 시스템이 잘못 판정한 총 496개의 문장 쌍에 대해서 수동으로 오류 유형 및 구문 구조 분석을 수행하였다. 이 중 거짓 양성 오류는 총 383 개, 거짓 음성 오류는 113개로 파악되었다. 1차 분석 완료된 데이터에 대해서 다시 검증을 수행함으로써 결과의 객관성을 유지하려고 노력하였다.

제안 시스템의 특징이자 장점 중의 하나는 (그림 7)과 같이 두 문장에 대한 근사 정렬 결과를 바탕으로 시스템의 동작이나 특성을 직관적으로 파악할 수 있다는 것이다. 그림에서 보듯이, 개별 문장을 구성하는 PAT들 간의 근사 정렬 과정이 알기 쉽게 묘사되어 있다. 본 논문에서의 오류 분석 과정에서도 위 정보를 활용하였으며 세부적인 분석에 많은 도움이 되었다.

5.2 세부 오류 유형별 분석

5.2.1 추가 구절 삽입에 의한 의미 변화 감지 실패 (49.5%, 246개)

제안 시스템이 가장 많은 오류를 범한 유형으로서 두 문장의 핵심적인 의미나 내용은 거의 동일하지만 문장 내에서 부가적인 역할을 수행하는 구절의 추가에 의한 의미 변화를 제대로 감지하지 못하여 발생하는 오류이다. 다음의 예를 보자.

정답:음성, 시스템:양성	
• Hilsenrath was also indicted on 33 counts of wire fraud, which each carry a maximum penalty of five years in prison, a $250,000 fine and restitution. • Each of those counts carries a maximum penalty of five years in prison and a $250,000 fine plus restitution.	(Ex.01)

위의 예는 MSRPC에서 음성(바꿔쓰기표현이 아님)으로 지정하였지만 본 논문의 시스템이 양성(바꿔쓰기표현임)으로 분류한 거짓 양성 오류이다. 예에서 밑줄 친 부분은 거의 일치하지만, 첫째 문장에서 “Hilsenrath was also indicted on 33 counts of wire fraud”가 추가됨으로써 문장의 의미가 변화되었다. 반대의 경우는 다음과 같다.

정답:양성, 시스템:음성	
• The industry group's non-manufacturing index rose to 50.7 during the month, up from March's reading of 47.9. • The Arizona-based ISM reported Monday that its non-manufacturing index rose to 50.7 last month, from 47.9 in March.	(Ex.02)

위 문장 쌍은 MSRPC에서 양성으로 지정되었으나, 시스템이 음성으로 분류한 거짓 음성 오류인 경우이다. 앞의 예에서와 같이 “The Arizona-based ISM reported Monday that”이라는 절이 둘째 문장에 추가되었음에도 불구하고 정답은 양성인 이유는 위 절이 문장의 전체적인 의미 변화에는 크게 영향을 미치지 않는 내용이기 때문이다. 다시 말해서 위 절은 “that” 이하 절이 특정한 출처가 있는 인용 구절임을 표현하는데, 이는 문장의 핵심적인 내용과는 크게 상관이 없다. 그러나 뒤에서 언급하겠지만, 데이터 분석 결과 MSRPC에서의 양성/음성 판정 기준 적용이 매우 모호한 경우가 다수 존재한다.

5.2.2 숫자 비교 실패 (12.5%, 62)

MSRPC가 신문기사를 중심으로 구축된 말뭉치인 관계로 기사에서 출현할 수 있는 다양한 문장 패턴이 존재한다. 특히 경제 분야 기사에서 많이 나타나는 주가 및 환율 변동 등에 관한 문장이 상당수 출현하는데, 이 때, 두 문장이 거의 동일한 내용을 표현하고 있음에도 불구하고 숫자가 달라

서 음성으로 분류된 말뭉치가 상당수 존재한다. MSRPC는 후보 문장 쌍을 추출하기 위한 유사도 비교 이전에 숫자에 대한 단일화 작업을 수행하였고 따라서 문장 간의 숫자의 차이가 바꿔쓰기표현 관계 판정에 영향을 미치지 않는다고 알려져 있다[5]. 하지만 본 논문에서의 오류 분석 과정에서 이러한 주장이 사실과 다름을 알 수 있었다. 다음의 예를 살펴보자.

정답:음성, 시스템:양성	
• The two-year note US2YT=RR fell 5/32 in price, taking its yield to 1.23 percent from 1.16 percent late on Monday. • The benchmark 10-year note US10YT=RR lost 11/32 in price, taking its yield to 3.21 percent from 3.17 percent late on Monday.	(Ex.03)

위 두 문장은 환율 변동을 나타내는 거의 동일한 내용 템플릿을 사용하고 있지만 변동 비율을 나타내는 숫자가 완전히 달라서 음성으로 지정되었고 제안 시스템에서는 양성으로 분류한 예이다. 현재 시스템이 PAT기반의 근사 정렬을 수행하는 관계로 숫자에 대한 엄밀한 완전 일치 기능을 제공하지 못하고 있으나 이러한 기능은 쉽게 구현이 가능하다. 부가적으로 위의 예는 두 개의 주어, 즉 “The two-year note US2YT=RR”과 “The benchmark 10-year note US10YT=RR”이 서로 상이하므로 다음 절에서 제시되는 “주어 불일치 판정 오류”와도 연계가 되어 있다. 다음의 예를 보자.

정답:양성, 시스템:음성	
• The Standard & Poor's 500 stock index pulled back by nearly 4 points to 1,066.62. • The broad Standard & Poor's 500 Index <.SPX> fell 0.70 points, or 0.07 percent, to 1,069.42.	(Ex.04)

위의 예는 MSRPC에서는 양성으로 판정하였으나 시스템에서 음성으로 분류한 경우이다. 그러나 위의 예는 그 전의 예와 비교해 볼 때, 두 가지 측면에서 문제가 있어 보인다. 우선 위의 문장들의 주된 내용은 주가의 변동폭과 최종 주가에 대한 정보 제공이므로 출현한 숫자가 문장의 전체 의미를 결정짓는다고 볼 수 있는데, 모든 수치가 완벽하게 다름에도 불구하고 바꿔쓰기표현 관계라고 지정하는 것은 무리가 있다. 또한 (Ex.03)을 음성으로 분류한 이유 중의 하나인 숫자 불일치가 (Ex.04)에서도 그대로 나타나는데 이를 양성으로 판정하는 것은 문제가 있다고 본다.

5.2.3 주어 불일치 판정 오류 (8%, 40개)

문장 내의 다른 부분은 거의 동일한 의미를 나타내지만 주어가 서로 달라서 다른 의미를 가지게 되는 경우를 나타낸다. 다음의 예에서 (Ex.05)는 거짓 양성 오류이고 (Ex.06)은 거짓 음성 오류이다.

정답:음성, 시스템:양성	
• Shiites make up 20 percent of the country's population. • Sunnis make up 77 percent of Pakistan's population, Shiites 20 percent.	(Ex.05)

(Ex.05)는 숫자 비교 실패와 함께, 근사 정렬 과정에서 두 문장의 주어인 “Shiites”와 “Sunnis”의 차이를 제대로 식별하지 못하여 시스템에 의해 두 문장이 동일하다고 판단을 한 경우이다. 그러나 (Ex.05)의 둘째 문장의 마지막에 “Shiites 20 percent”라는 구절이 포함되었으므로 완전히 다른 문장은 아닌 것으로 볼 수 있다.

정답:양성, 시스템:음성	
• A Washington County man may have the county's first human case of West Nile virus, the health department said Friday. • The county's first and only human case of West Nile this year was confirmed by health officials on Sept. 8.	(Ex.06)

(Ex.06)은 비교적 바꿔쓰기표현 정도가 매우 심한 경우에 속한다. 구문적으로 분석해 보면, 우선 둘째 문장의 주어인 “The county's first and only human case of West Nile this year”가 첫째 문장에서 주어인 “A Washington County man”과 목적어인 “the county's first human case of West Nile virus”로 분리되고 있고, 첫째, 둘째 문장의 부가어구인 “the health department said Friday”와 “was confirmed by health officials on Sept. 8”가 암묵적인 의미적 관계로 연결되어 있다. 주어가 서로 완전히 상이함을 인지하지 못하는 시스템으로서는 PAT 기반의 근사 정렬이 거의 이루어지지 않으므로 유사도를 매우 낮게 책정함으로써 음성으로 판단하였다.

5.2.4 수식 구절의 내용 차이에 따른 의미 변화 탐지 실패 (7%, 37)

문장 내에서 명사 등을 수식하는 형용사절, 관계사절 등의 존재 유무 및 의미 차이 등에 따른 전체 문장의 의미 변화를 감지하지 못하는 경우이다. 다음의 예를 살펴보자.

정답:음성, 시스템:양성	
• “I would like the FCC to start all over,” said Sen. Kay Bailey Hutchison, R-Texas, who supported reversing the rules. • “I would like the FCC to start all over again,” said Sen. Kay Bailey Hutchison, R-Texas, who expressed concern about “potentially dangerous” newspaper-broadcast combinations.	(Ex.07)

주절의 문장은 동일하지만 주어인 “Sen. Kay Bailey Hutchison, R-Texas”를 수식하는 관계사절이 다른 의미를

나타내고 있어서 MSRPC는 이를 음성으로 판정하였고, 시스템은 이에 대한 감지를 실패하여 이를 양성으로 분류한 경우이다.

정답:음성, 시스템:양성	
• Talabani told him the Governing Council would "need UN assistance and advice in implementing the new decisions which have been taken." • Talabani told him Iraqi leaders would "need U.N. assistance and advice in implementing the new decisions which have been taken" on organising an interim Iraqi government by June.	(Ex.08)

이 경우도 위의 경우와 마찬가지로 거짓 음성 오류에 속하는데, (Ex.07)과는 조금 다른 양상을 나타낸다. 즉, 둘째 문장에서 "the new decisions"를 수식하는 관계사절에 전치사구가 삽입되어 더 구체적인 의미를 나타내는 경우이다. MSRPC 내에서 양성으로 지정된 다른 데이터들을 전체적으로 살펴볼 때 이 경우는 논란의 여지가 있다.

5.2.5 길이가 긴 개체명 등으로 구성된 단일 문장성분에 의한 과도 정렬(6%, 32)

신문 기사의 특성상 MSRPC 내에는 다양한 개체명(인명, 지명, 상호명, 기관명 등)이 출현한다. 특히 직함이 포함된 인명(appositive), 길이가 매우 긴 상호명, 수식어구가 붙은 명사 등이 동시에 두 문장에 출현하여 PAT 기반 근사정렬 과정에서 과도한 정렬이 발생하는 경우가 있었다.

정답:음성, 시스템:양성	
• Actor Arnold Schwarzenegger is leaving backers in suspense, and former Los Angeles Mayor Richard Riordan will consider if the Terminator balks. • Arnold Schwarzenegger and former Los Angeles Mayor Richard Riordan may jump in by the Aug. 9 deadline to file.	(Ex.09)

(Ex.09)의 각 문장에 나타나는 두 개의 주어인 "Actor Arnold Schwarzenegger"와 "former Los Angeles Mayor Richard Riordan"는 단어 개수 기준으로 문장 전체의 50% 정도를 차지하고 서로 일치하고 있으나, 그 외의 부분에서는 의미가 완전히 다르다. 시스템의 주된 판단 기준이 근사정렬의 정도임을 고려할 때, 이러한 오류 유형에 대한 특수처리가 필요하다.

5.2.6 루트 PAT 과도 일치 (5%, 26)

주로 be-동사나 인용을 나타내는 "say", "announce", "tell" 등과 같이 문장 내에서 그 비중이 비교적 낮은 동사를 술어로 가지는 루트 PAT이 과도하게 정렬되는 현상을 의미한다. 아래의 예를 보면서 루트 PAT의 의미와 오류 유형에 대한 설명을 제시한다.

정답:음성, 시스템:양성	
• Against the Japanese currency, the euro was at 135.92/6.04 yen against the late New York level of 136.03/14. • The dollar was at 117.85 yen against the Japanese currency, up 0.1 percent.	(Ex.10)

두 문장 모두 be-동사를 취하는 2형식 문장이다. Be-동사를 술어로 가지는 PAT을 구성해 보면 각각 "be(euro, yen)"과 "be(dollar, yen)"이고, 이 두 PAT의 유사도는 매우 높다. 그러나 각 문장의 다른 부분을 살펴보면 문장 간 의미적 연관성은 매우 떨어짐을 알 수 있다. 문장 내에서 주요 동사의 의미나 역할이 상대적으로 낮음에도 불구하고 일률적으로 이에 대한 일치 여부에 대한 중요도를 높게 책정하여 시스템이 발생시키는 오류 유형도 존재한다.

5.2.7 문장의 구체성 판정 오류 (5%, 26)

요약해서 보면 두 문장의 주요 테마가 동일하지만 특정 문장이 더 세부적이고 구체적으로 의미를 기술한 경우 이를 제대로 판정하지 못한 경우이다.

정답:양성, 시스템:음성	
• The launch marks the start of a new golden age in Mars exploration. • The launch marks the start of a race to find life on another planet.	(Ex.11)

(Ex.11)은 MSRPC에서 양성으로 판정되었으나 시스템이 음성으로 판단한 경우이다. 이를 양성으로 판단할 것인가에 대해서는 논란의 여지가 있으나, 위의 밑줄 그은 부분을 살펴보면, 두 문장 모두 구체성이라는 측면에서 서로 상보적인 관계에 있음을 알 수 있다. 다시 말해서, 우선 첫째 문장에서는 "Mars exploration"이라는 구절이 둘째 문장의 "another planet"이라는 문장보다 더 구체적이다. 더불어 첫째 문장의 "a race to find life"는 둘째 문장의 "a new golden age"라는 구절보다 더 구체적이라고 볼 수 있다. 논란의 여지가 있는 또 다른 예를 아래에서 살펴보자.

정답:음성, 시스템:양성	
• Apple Computer's new online music service sold more than 1 million songs during its first week of operation, the company said Monday. • Apple Computer Inc. said Monday it exceeded record industry expectations by selling more than 1 million songs since the launch of its online music store a week ago.	(Ex.12)

이 예에서 밑줄 그은 부분은 두 문장의 핵심 이벤트를 나타내며 서로 일치하고 있다. 그러나 둘째 문장에서 "it exceeded record industry expectations"라는 문구가 삽입됨

으로써 그 이벤트의 결과를 나타내며 더 구체적인 내용을 전달하고 있다. 그러나 (Ex.11)을 양성으로 판단한 기준에 비하면 이 예를 음성으로 판정한 것은 무리가 있어 보인다. 바꿔쓰기표현 본연의 의미인 동질성이라는 측면에서 접근한다면 (Ex.11)보다 오히려 (Ex.12)가 더 양성에 가깝다.

5.2.8 단어 중첩도가 높은 두 문장의 의미적 상이성 식별 실패 (3%, 18)

두 문장이 동일하거나 유사한 어휘들을 많이 사용하고 있으나, 각각의 주된 의미는 완전히 다른 경우 이를 제대로 식별하지 못하는 경우를 말한다.

정답:음성, 시스템:양성	
• The fines are part of failed Republican efforts to force or entice the Democrats to return. • Perry said he backs the Senate's efforts, including the fines, to force the Democrats to return.	(Ex.13)

위와 같이 상당한 수의 단어가 서로 중복되고 있으나 두 문장의 의미는 다르다. 이 오류 유형은 앞에서 설명한 길이가 긴 단일 문장성분에 의한 과도 정렬과도 연관되어 있다. 이에 대한 처리를 위해서는 PAT 기반 근사 정렬의 엄밀성을 더 높여야 한다. 또한 현재까지는 대부분 양성(바꿔쓰기표현 관계)을 제대로 식별하기 위한 연구가 대부분 시도되었으나, 위의 예에서 보듯이, 음성, 즉 문장 간의 의미적 상이성 식별에 관한 연구도 함께 진행되어야 한다.

5.2.9 주요 동사의 의미 구별 실패 (3%, 18)

루트 동사구(root verb phrase)가 전체적인 의미 전달에 핵심적인 역할을 수행하는 두 문장에 대해서 이들 동사구의 의미적 연관성 파악이 제대로 되지 않아서 발생하는 오류를 말한다. 예를 들어, 두 동사구가 서로 반의어 및 대립어 관계에 있거나 숙어일 경우에도 두 루트 동사 간의 의미 판별이 매우 어렵다. 예를 살펴보자.

정답:음성, 시스템:양성	
• Families stuck on the highway remained in their cars, and used their cell phones to call home. • Families stuck on the highway were being urged to remain in their cars, and to use their cell phones only in case of emergency.	(Ex.14)

위의 경우는 행위자의 의도 여부나 행위의 결과에 대한 상이성으로 인해서 MSRPC에서 음성으로 분류한 데이터이다. 첫째 문장에서는 가족이 차에 머물러 있었으며, 둘째 문장에서는 차에 머물러 있도록 강요당하였으나 실제로 머물러 있었는지에 대한 여부는 모른다. 따라서 의미적인 측면에서 상당한 괴리가 있으나 시스템에서는 이를 감지하지 못하고 있다.

5.2.10 이벤트 발생 시간이 다름 (3%, 17)

두 문장이 동일한 테마와 내용 및 이벤트를 나타내고 있으나 그 발생 시간이 달라서 바꿔쓰기표현이 아닌 경우, 이에 대한 식별을 실패한 오류 유형이다.

정답:음성, 시스템:양성	
• He was tracked to Atlanta where he was arrested on Tuesday night. • He was arrested in Atlanta, Georgia, on Monday night by police acting on a tip-off.	(Ex.15)

5.2.11 기타 오류 유형

그 외에도 특정 문장 전체가 다른 쪽 문장의 일부 구절로 포함된 경우, 각 문장에서 서로 다른 역할(서로 다른 문장성분)을 수행하는 동일한 구절에 의한 과도 정렬, 한쪽은 가주어, 나머지는 일반 주어로 구성된 문장 쌍, 한쪽은 행위자이고 나머지는 행위자의 소유물인 경우 등이 있었으나 대부분 그 비율이 낮았다.

특히 다음의 예에서 보듯이, 대부분의 기존 연구에서 바꿔쓰기표현의 전형적인 형태로 다루어 왔던 어휘 다양성이나 구문 패턴 다양성에 관련한 오류 유형에 대한 비율도 대부분 낮았다.

“~ added to pressure on ~ by downgrading ~”	“~ downgraded ~”
“Median household income declined~ ”	“The survey found the median household~”
“The conference, sponsored by the U.S. ”	“The conference was sponsored by the U.S. ”
“Mount Airy native”	“ ~ grew up in Mount Airy ~”
“The vulnerability”	“The security hole”
“Microsoft Corp.”	“MS”

5.3 오류 분석 결과 정리

아래 표에서 지금까지 앞 절에서 제시한 다양한 오류 분석 결과를 정리하였다. 특히 추가 구절 삽입에 따른 두 문장 간의 의미 변화에 대한 탐지 실패 오류가 가장 많았으며, 이를 숫자 비교 실패 유형과 합치면 50%가 넘는다. 따라서 MSRPC에서 강조한 주요 바꿔쓰기표현 유형에 잘 대처하고 높은 성능을 나타내는 시스템을 구현하고자 한다면, 이 두 가지 오류 유형에 대한 처리가 필요하다.

앞에서도 수시로 언급하였고, 위 표에서도 나타난 바와 같이 본 논문에서 제안하는 시스템이 오류를 범한 총 486개의 인스턴스에는 상당수의 오류가 포함되어 있음을 알 수 있었다. 전체 데이터의 9%에 해당하는 47개 정도가 데이터 오류로 의심되었는데, 대부분 판정 일관성 부족에 기인하였다. 추후 연구에서 이에 대한 심층적인 분석이 필요하다.

〈표 6〉 오류 유형에 따른 발생 비율

오류 유형	빈도 (비율)
(1) 추가 구절 삽입에 의한 의미 변화 감지 실패	246 (49.5%)
(2) 숫자 비교 실패	62 (12.5%)
(3) 주어 불일치 판정 오류	40 (8%)
(4) 수식 구절의 내용 변화에 따른 의미 변화 탐지 실패	37 (7%)
(5) 길이가 긴 단일 문장성분에 의한 과도 정렬	32 (6%)
(6) Root PAT 과도 일치	26 (5%)
(7) 문장의 구체성 판정 오류	26 (5%)
(8) 개별 어휘 중복 정도는 높으나 의미상 다른 문장 쌍	18 (3%)
(9) 주요 동사의 의미 구별 실패	18 (3%)
(10) 이벤트 발생 시간이 다름	17 (3%)
(11) 대부분이 동일하나 극히 작은 일부분에서 결정적으로 다름	14 (2%)
(12) 기타 오류	75 (15%)
(13) 데이터 오류 의심	47 (9%)

6. 결 론

본 논문에서는 문장 단위 바꿔쓰기표현 식별을 위한 새로운 접근 모델인 PAT 기반 근사 정렬(PAT-based Approximate Alignment Similarity) 유사도 기법을 제안하였다. 이를 기반으로 이 모델에 의해서 도출된 총 5가지의 정렬 자질만을 활용하여 MSRPC에 대한 바꿔쓰기표현 식별 실험을 수행하였다. 실험 결과, (1)성능의 차이가 미미한 점(2.6% 미만), (2)SVM 및 근사 정렬 매개변수 등의 설정에 따른 최적화를 수행하지 않은 점, 마지막으로 (3)근사 정렬의 결과로 도출되는 5가지의 자질만을 활용하였다는 점에서 볼 때, 최근 발표된 연구 결과와 거의 동등한 수준을 견지할 수 있었다.

제안 시스템의 특성(정렬 결과 덤프, 사용 자질들의 직관성 등)을 최대한 활용하여, 시스템 출력 오류를 분석하였으며, 다양한 오류 분석을 통해서 시스템의 성능 향상을 위한 기반 지식과 문장 단위 바꿔쓰기표현 양상에 대한 정형적 접근을 수행할 수 있었다.

향후 연구 계획으로서는 본 연구에서 파악된 문장 단위 바꿔쓰기표현 유형을 시스템에 적용하여 성능을 개선하는 작업이 필수적이다. 두 문장 간의 PAT 비교(유사도 측정)의 엄밀성을 확장하기 위해서 (그림 4)에서 제시한 개별 단어 매칭 임계치를 자동으로 조정하는 기계학습 기법을 도입해야 한다. 또한 정렬되지 않는 PAT들의 의미적 중요도(구문적, 통계적 중요도)를 파악하여 이를 유사도에 반영하는 방법도 반영되어야 한다. 현재 가장 많은 빈도를 차지하고 있는 오류 유형이 “추가 구절 삽입에 의한 의미 변화 감지 실패 오류”이고 정렬되지 않는 PAT들이 바로 이 추가 구절과 연관되어 있으므로 이에 대한 중요도를 식별함으로써 중요한 PAT들이 정렬되지 않은 경우에는 유사도를 낮추거나 특정 자질 값을 마이너스로 표기하는 기법들을 생각할 수 있다.

추가적으로, 근사 정렬의 대상 단위인 PAT의 표현력을 향상시키는 문제도 남아 있다. 현 시스템은 다중 어절로 구성된 개체명이나 용어를 비롯한 구절 동사(phrasal verb)를 제대로 인식하지 못하여 이를 PAT 표현에 제대로 반영하지 못하고 있다. 더불어, PAT 간의 유사도를 보다 심층적으로 측정하기 위한 시맨틱 자원의 확충도 성능 향상에 필수적일 것으로 보인다.

참 고 문 헌

[1] R. Barzilay and K. R. McKeown (2001), “Extracting paraphrases from a parallel corpus,” in *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, pp.50-57.

[2] F. Rinaldi, J. Dowdall, K. Kaljurand, M. Hess, and Moll₩′a, Diego (2003), “Exploiting paraphrases in a Question Answering system,” in *Proceedings of the second international workshop on Paraphrasing*, Morristown, NJ, USA, pp.25-32.

[3] R. Barzilay and L. Lee (2003), “Learning to paraphrase: an unsupervised approach using multiple-sequence alignment,” in *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Morristown, NJ, USA, pp.16-23.

[4] C. Quirk, C. Brockett, and W. Dolan (2004), “Monolingual machine translation for paraphrase generation,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp.142-149.

[5] C. Brockett and W. B. Dolan (2005), “Support Vector Machines for Paraphrase Identification and Corpus Construction,” in *Third International Workshop on Paraphrasing (IWP2005)*, pp.1-9.

[6] X. Wang, D. Lo, J. Jiang, L. Zhang, and H. Mei (2009), “Extracting paraphrases of technical terms from noisy parallel software corpora,” in *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Morristown, NJ, USA, pp.197-200.

[7] P. Malakasiotis (2009), “Paraphrase recognition using machine learning to combine similarity measures,” in *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, Morristown, NJ, USA, pp.27-35.

[8] F. Keshtkar and D. Inkpen (2010), “A Corpus-based Method for Extracting Paraphrases of Emotion Terms,” in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, CA, pp.35-44.

[9] 조정현, 정현기, 김유섭 (2009), "웹 검색과 문서 유사도를 활용한 2 단계 신문 기사 표절 탐지 시스템," *정보처리학회논문지B*, Vol.16B, pp.181-194,

[10] 박경미, 문영성 (2010), "부분 구문 분석 결과에 기반한 두 단계 부분 의미 분석 시스템," *정보처리학회논문지B*, Vol.17B, pp.85-92.

[11] 이공주, 윤보현 (2006), "정렬된 성경 코퍼스로부터 바꿔쓰기표현(paraphrase)의 자동 추출," *인지과학*, Vol.17, pp.323-336.

[12] B. Pang, K. Knight, and D. Marcu (2003), "Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences," in *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Morristown, NJ, USA, pp.102-109.

[13] Y. Shinyama and S. Sekine (2003), "Paraphrase acquisition for information extraction," in *Proceedings of the second international workshop on Paraphrasing*, Morristown, NJ, USA, pp.65-71.

[14] I. Androutsopoulos and P. Malakasiotis (2010), "A Survey of Paraphrasing and Textual Entailment Methods," *Journal of Artificial Intelligence Research*, Vol.38, pp.135-187,

[15] A. Hickl and J. Bensley (2007), "A discourse commitment-based framework for recognizing textual entailment," in *RTE '07: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Morristown, NJ, USA, pp.171-176.

[16] S. Zhao, H. Wang, T. Liu, and S. Li (2008), "Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora," in *Proceedings of ACL-08: HLT*, Columbus, Ohio, pp.780-788.

[17] A. Finch, Y.-S. Hwang, and E. Sumita (2005), "Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence," in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

[18] C. Bannard and C. Callison-Burch (2005), "Paraphrasing with bilingual parallel corpora," in *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, pp.597-604.

[19] L. Qiu, M.-Y. Kan, and T.-S. Chua (2006), "Paraphrase recognition via dissimilarity significance classification," in *EMNLP '06: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Morristown, NJ, USA, pp.18-26.

[20] A. D. Haghighi, A. Y. Ng, and C. D. Manning (2005), "Robust textual inference via graph matching," in *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Morristown, NJ, USA, pp.387-394.

[21] Y. Miyao and J. i. Tsujii (2008), "Feature Forest Models for Probabilistic HPSG Parsing," *Computational Linguistics*, Vol.34, pp.35 - 80.

[22] Y. Zhang and J. Patrick (2005), "Paraphrase Identification by Text Canonicalization," in *Proceedings of the Australasian Language Technology Workshop 2005*, Sydney, Australia, pp.160-166.

[23] Z. Kozareva and A. Montoyo, "Paraphrase Identification on the Basis of Supervised Machine Learning Techniques," in *Advances in Natural Language Processing*. Vol.4139, T. Salakoski, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2006, pp.524-533.

[24] S. Fernando and M. Stevenson (2008), "A Semantic Similarity Approach to Paraphrase Detection," in *Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloquium*.

[25] V. Rus, P. M. McCarthy, M. C. Lintean, D. S. McNamara, and A. C. Graesser (2008), "Paraphrase Identification with Lexico-Syntactic Graph Subsumption," in *FLAIRS Conference*, pp.201-206.

[26] B. Dolan, C. Quirk, and C. Brockett (2004), "Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources," in *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, Morristown, NJ, USA, pp.350.

[27] T. Pedersen, S. Patwardhan, and J. Michelizzi (2004), "WordNet::Similarity - Measuring the Relatedness of Concepts," in *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, San Jose, CA pp.1024-1025.

[28] T. Chklovski and P. Pantel (2004), "VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations," in *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain.

[29] K. K. Schuler, A. Korhonen, and S. Brown (2009), "VerbNet overview, extensions, mappings and applications," in *HLT-NAACL*, pp.13-14.

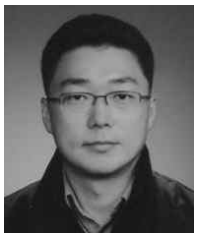
[30] F. J. Och and H. Ney (2003), "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, Vol.29, pp.19-51.

[31] P. Liang, B. Taskar, and D. Klein (2006), "Alignment by Agreement," in *Proceedings of NAACL 2006*, New York City, USA, pp.104-111.

[32] H. W. Kuhn (1955), "The Hungarian Method for the assignment problem," *Naval Research Logistics Quarterly*, Vol.2, pp.83-97.

[33] R. Mihalcea, C. Corley, and C. Strapparava (2006), "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI'06: Proceedings of the 21st national conference on Artificial intelligence*, pp.775-780.

[34] D. Das and N. A. Smith (2009), "Paraphrase identification as probabilistic quasi-synchronous recognition," in *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, Morristown, NJ, USA, pp.468-476.



최 성 필

e-mail : spchoi@kisti.re.kr
1996년 부산대학교 전자계산학과(학사)
1998년 부산대학교 전자계산학과 (이학석사)
2012년 한국과학기술원 정보통신공학과 (공학박사)
1998년~2001년 연구개발정보센터 선임연구원
2001년~현 재 한국과학기술정보연구원 SW연구실 선임연구원
관심분야 : 텍스트마이닝, 기계학습, 자연어처리, 텍스트 추론

송 사 광

e-mail : esmallj@kisti.re.kr
1997년 충남대학교 통계학과(학사)
1999년 충남대학교 컴퓨터과학과(석사)
2011년 KAIST 전산학과(박사)
2005년~2010년 한국전자통신연구원 바이오인포매틱스팀 연구원
2010년~현 재 한국과학기술정보연구원 SW연구실 선임연구원
관심분야 : 텍스트 마이닝, 자연어처리, 정보검색, 시맨틱 웹

맹 성 현

e-mail : myaeng@kaist.ac.kr
1983년 미국 캘리포니아 주립대학(학사)
1985년 미국 Southern Methodist University (SMU)(석사)
1987년 미국 Southern Methodist University (SMU)(박사)
1987년~1988년 미국 Temple University 교수
1988년~1994년 미국 Syracuse University 교수(tenured)
1994년~2003년 충남대학교 컴퓨터과학과 교수
2003년~2009년 한국정보통신대학교 교수
2009년~현 재 한국과학기술원 전산학과 교수
관심분야 : 정보 검색, 텍스트 마이닝, HCI, 상황인지 컴퓨팅 등