creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

[UCI]I804:24011-200000265054
조선대학교 CHOSUN UNIVERSITY

2015년 8월
석사학위 논문

스토리텔링 저작을 위한 서사정보 추출 시스템 설계 및 구현

조선대학교 산업기술융합대학원

소프트웨어융합공학과

백 남 주

조선대학교
CHOSUN UNIVERSITY

스토리텔링 저작을 위한 서사정보 추출 시스템 설계 및 구현

Design and Implementation of Narrative Information Extraction System for Storytelling Authoring

2015년 8월 25일

조선대학교 산업기술융합대학원

소프트웨어융합공학과

백 남 주

조선대학교
CHOSUN UNIVERSITY

스토리텔링 저작을 위한 서사정보 추출 시스템 설계 및 구현

지도교수 김 판 구

이 논문을 공학석사학위신청 논문으로 제출함.

2015년 4월

조선대학교 산업기술융합대학원

소프트웨어융합공학과

백 남 주

조선대학교
CHOSUN UNIVERSITY

백남주의 석사학위논문을 인준함

위원장 조선대학교 교수 (인)

위 원 조선대학교 교수 (인)

위 원 조선대학교 교수 (인)

2015년 5월

조선대학교 산업기술융합대학원

조선대학교 CHOSUN UNIVERSITY

목 차

ABSTRACT

Ⅰ. 서론 ····· 1
A. 연구 배경 및 목적 ····· 1
B. 논문의 구성 ····· 2

Ⅱ. 관련 연구 ····· 3
A. 스토리텔링 저작 지원 기술 현황 ····· 3
1. 스토리텔링 저작을 위한 지능형 스토리 생성 기술 ····· 3
2. 스토리텔링 저작 지원 도구 ····· 5
3. 스토리텔링을 위한 서사구조 모델 ····· 9
B. 설화 데이터 분석을 위한 자연어 처리 시스템 ····· 13
1. 꼬꼬마 형태소 분석기 ····· 13
2. 한나눔 형태소 분석기 ····· 14
C. 토픽맵 ····· 17
1. 토픽맵 구조 및 구성요소 ····· 17
2. 토픽맵을 이용한 정보 시각화 ····· 20

Ⅲ. 서사정보 자동 추출 시스템 설계 및 구현 ····· 22
A. 서사정보 자동 추출 시스템 설계 및 인터페이스 정의 ····· 22
1. 시스템 구성도 ····· 22
2. 개발 환경 ····· 23
3. 사용자 인터페이스(User Interface) ····· 23
B. 서사정보 추출을 위한 전처리 및 명사 추출 방법 ····· 24
1. 특수문자 제거 및 POS(Part-Of-Speech) Tagging ····· 24
2. 서사정보 추출을 위한 명사 패턴 정의 및 추출 ····· 26
C. 서사정보 추출을 위한 TF 알고리즘 활용 방법 ····· 27
1. TF(Term Frequency) 값을 이용한 서사정보 후보 추출 ····· 27
2. 불용어 명사 목록 정의 ····· 29
3. 불용어 태그를 이용한 TF별 명사 재추출 ····· 30

조선대학교 CHOSUN UNIVERSITY

Ⅳ. 실험 및 평가 ····· 33
A. 실험 데이터 ····· 33
B. 실험 평가 및 결과 ····· 34
1. 실험 평가 방법 ····· 34
2. 실험 결과 ····· 34
C. 실험 결과에 따른 토픽맵 기반 시각화 응용방안 ····· 36
1. 서사정보를 이용한 토픽맵 구축 ····· 36
2. 토픽 및 관계 생성 ····· 39

Ⅴ. 결론 및 제언 ····· 42

참고문헌 ····· 43

조선대학교 CHOSUN UNIVERSITY

표 목 차

[표 2-1] 스토리텔링 저작 도구 개발에 요구되는 기술적 요소 ····· 6
[표 2-2] 서사구조의 단계별 내용 ····· 10
[표 2-3] 인물정보 분류 ····· 11
[표 2-4] 사건정보 분류 ····· 12
[표 2-5] 배경정보 분류 ····· 12
[표 2-6] 꼬꼬마 형태소 분석기의 태그 집합 예 ····· 14
[표 2-7] 분석 후보 생성 ····· 15
[표 2-8] 미등록어 처리기 ····· 15
[표 2-9] 한나눔 형태소 분석기의 태그 집합 예 ····· 16
[표 3-1] 서사정보 자동 추출 시스템 개발 환경 ····· 23
[표 3-2] 설화 텍스트 내 특수문자 제거 소스 ····· 24
[표 3-3] 설화 텍스트에 대한 POS Tagging 예 ····· 25
[표 3-4] 한나눔 형태소 분석기 내 정의된 명사 패턴 ····· 26
[표 3-5] 명사 추출 알고리즘 ····· 26
[표 3-6] 명사 패턴을 활용하여 추출된 명사 ····· 27
[표 3-7] TF값 추출 알고리즘 ····· 28
[표 3-8] 명사별 TF값 계산 결과 ····· 28
[표 3-9] 불용어 명사 목록 ····· 29
[표 3-10] 명사에 정의된 불용어의 태그 ····· 29
[표 3-11] 불용어 명사를 제거하기 위한 불용어 태그 정리 소스 ····· 30
[표 3-12] '지하국대적퇴치' 설화 내 서사정보 추출 결과 ····· 31
[표 3-13] '지하국대적퇴치' 설화 내 서사정보 분류 결과 ····· 31
[표 4-1] 실험데이터 정보 ····· 33
[표 4-2] 서사정보 자동 추출 시스템의 등장인물 추출에 대한 성능 ····· 34
[표 4-3] 설화 데이터 시각화를 위한 구성요소 도출 ····· 37

조선대학교 CHOSUN UNIVERSITY

그 림 목 차

[그림 2-1] 스토리 생성 시스템 전체 모델 ····· 3
[그림 2-2] 지능형 관리 모듈 ····· 4
[그림 2-3] 인터랙티브 스토리텔링 시스템 ····· 5
[그림 2-4] 드라마티카 프로의 초기 화면 ····· 7
[그림 2-5] 파이널 드래프트의 실행 화면 ····· 8
[그림 2-6] 스토리 헬퍼의 실행 화면 ····· 9
[그림 2-7] 한나눔 형태소 분석기의 워크플로우 ····· 15
[그림 2-8] 토픽 맵 구조 ····· 18
[그림 2-9] 토픽 ····· 18
[그림 2-10] 토픽 타입 ····· 19
[그림 2-11] 어커런스 ····· 19
[그림 2-12] 어소시에이션 ····· 20
[그림 2-13] 판소리 토픽맵 ····· 20
[그림 2-14] 이상한 나라의 앨리스 플롯 중심 토픽맵 ····· 21
[그림 3-1] 전체 시스템 구성도 ····· 22
[그림 3-2] 사용자 인터페이스 ····· 23
[그림 3-3] 추출된 서사정보 목록 결과 ····· 31
[그림 4-1] 각 설화별 등장인물 추출에 대한 정확률 ····· 35
[그림 4-2] '지하국대적퇴치' 설화 개념 모델링 ····· 36
[그림 4-3] 구현된 토픽맵 기반 설화 데이터 ····· 38
[그림 4-4] 인물 관계 생성 ····· 33
[그림 4-5] 배경 관계 생성 ····· 40
[그림 4-6] 사건 관계 생성 ····· 40
[그림 4-7] 서사정보 관계 생성 ····· 41

조선대학교 CHOSUN UNIVERSITY

ABSTRACT

Design and Implementation of Narrative Information Extraction System for Storytelling Authoring

NamJu Baek
Advisor : Prof. Pankoo Kim, Ph.D
Department of Software Convergence Engineering
Graduate School of Chosun University

Recently, the perceived value of stories has risen, which has created a greater focus on storytelling. Storytelling is method used for immersing an audience and prompting interest in the development of plot. This method has been receiving a lot of attention recently as an important technique in the cultural industry.

More specifically, there have many more cases of stories being created using authoring tool systems through Human-Computer Interaction (HCI). Compared to other countries, there have not been many studies regarding using such systems for storytelling in Korea.

This thesis proposes a system which automatically extracts narrative information, including key storytelling elements, that comprises the main components of a written story. Patterns for extracting narrative information nouns within the text are defined by applying a Korean natural language process technique to story data collection and analysis. Story data is then extracted using term frequency(TF).

조선대학교
CHOSUN UNIVERSITY

To assess this narrative information extraction system, experiments were done to analyze data from six stories. The results determined using the automated character extraction method. The results showed that approximately 71.83% of the average precision ratio and performance had been confirmed.

조선대학교 CHOSUN UNIVERSITY

Ⅰ. 서론

A. 연구 배경 및 목적

최근 스토리가 지닌 무한의 부가가치가 새로운 창출의 핵심적인 요인으로 떠오르면서 청중을 몰입시키고 내용전개에 흥미를 유발시키게 만드는 기법인 스토리텔링이 문화산업에 있어서 중요한 기술로 주목을 받고 있다[1]. 스토리텔링이란 스토리(Story)와 텔링(Telling)의 합성어로 상대방에게 알리고자 하는 바를 재미있고 생생한 이야기로 설득력 있게 전달하고자 하는 행위라고 할 수 있다[2]. 특히 인간과 컴퓨터간의 상호작용(HCI, Human-Computer Interaction)을 통한 스토리텔링 저작 도구 시스템을 활용하여 스토리를 창작하는 사례가 늘어나고 있다. 스토리텔링 저작 도구는 스토리를 창작하는데 있어서 사용되는 컴퓨터 소프트웨어를 가리키지만 단지 컴퓨터 프로그래밍 기술에 의해 구현된 것이 아니라 인문학적 이론인 서사학과 창작 이론의 결합을 통한 스토리 구조 분석 기반의 종합적인 연구에 의해 구현된다[3]. 주로 국외의 대표적인 스토리텔링 저작 도구로는 드라마티카 프로(Dramatica Pro)와 파이널 드래프트(Final Draft) 등이 존재하는 반면, 국내에서는 최근 스토리 헬퍼(Story Helper)가 개발된 사례 이외의 연구는 미흡한 실정이다.

이에 본 논문에서는 효율적인 스토리텔링 저작 지원을 위해 스토리 서사정보를 자동으로 추출하는 시스템을 설계 및 구현하고자 한다. 이를 위해 서사정보의 구성요소인 인물, 사건, 배경에 대한 단어 품사가 주로 명사에 분포됨을 이용한다. 먼저 한글로 작성된 설화 텍스트로부터 명사 추출을 위해 입력받은 텍스트 내 존재하는 특수문자를 제거하고 단어 별 품사에 대한 태깅(Tagging) 과정을 거친 후, 형태소 분석기 내 정의된 명사 태그(Tag)를 이용하여 명사 추출을 위한 패턴(Pattern)을 정의한다. 추출된 설화 내 명사로부터 TF(Term Frequency) 알고리즘을 이용하여 명사 내 서사정보 후보를 추출하며, 보다 정확한 서사정보 추출을 위해 불용어(Stopword) 명사 목록 정의하여 재필터링 된 서사정보를 추출한다. 또한, 추출된 서사정보를 중심으로 설화 데이터에 대한 시각화(Visualization)를 위해 토픽맵(Topic Maps) 기반의 응용방안을 제시하고자 한다.

조선대학교 CHOSUN UNIVERSITY

B. 논문의 구성

스토리텔링 저작을 위한 서사정보 자동 추출 시스템을 설계 및 구현하기 위해 본 논문은 다음과 같은 구성으로 작성되었다.

1장 서론에서는 연구의 배경 및 목적에 대해 간략하게 기술하고, 2장 관련연구에서는 스토리텔링 저작 지원 기술 현황과 설화 데이터 분석을 위한 자연어 처리 시스템에 대해 설명하며, 토픽맵을 이용한 정보 시각화에 대해 기술한다.

3장에서는 서사정보 자동 추출 시스템을 구현하기 위해 설화 텍스트로부터 서사정보 추출을 위한 전처리 과정과 추출된 명사로부터 TF 알고리즘 및 불용어 명사 목록을 활용한 서사정보 추출 방법에 대해 기술한다.

4장에서는 본 논문에서 제안한 서사정보 추출 시스템에 대한 성능 평가를 수행하며 추출된 서사정보 데이터를 중심으로 토픽맵 기반의 시각화 응용방안을 제시한다.

마지막으로 5장에서는 본 연구에 대한 전체적인 결과를 요약하고, 향후 연구 방향을 제시한다.

조선대학교 CHOSUN UNIVERSITY

Ⅱ. 관련 연구

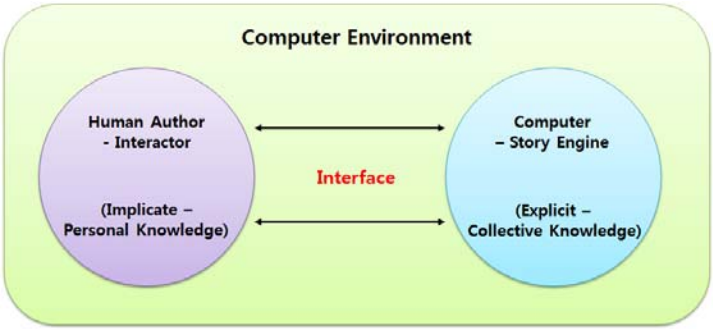
A. 스토리텔링 저작 지원 기술 현황

본 절에서는 스토리텔링 저작 지원 기술의 현황을 알아보기 위해 스토리텔링 저작을 위한 지능형 스토리 생성 기술과 스토리텔링 저작 지원 도구의 기술적 요소 및 사례에 대해 살펴보고 그 한계점에 대해 기술한다.

1. 스토리텔링 저작을 위한 지능형 스토리 생성 기술

지능형 스토리 생성 시스템 연구는 완성도 높은 스토리를 생성하기 위해 디지털 스토리텔링의 특성을 이용하여 스토리에 대한 서사(Narrative)성을 갖추며, 창작을 위한 보조수단으로써 어떻게 컴퓨터와 인간이 상호작용하는지에 대한 연구를 말한다. 여기서 스토리는 다중매체성, 네트워크성, 상호작용성, 비선형성 등으로 특징지어지는 새로운 스토리가 아닌 기존의 문자, 영상매체를 통해 만들어져 온 서사성을 갖춘 전통적이고 고적적인 스토리를 뜻한다.

특히, 지능형 스토리 생성 기술은 디지털 스토리텔링의 특성을 이용하여 최종적으로 만들어진 스토리에 대한 창작물이 아닌 창작의 과정 속에서 작가들의 보조수단으로써 활용을 위한 기술 연구라고 할 수 있다[4].



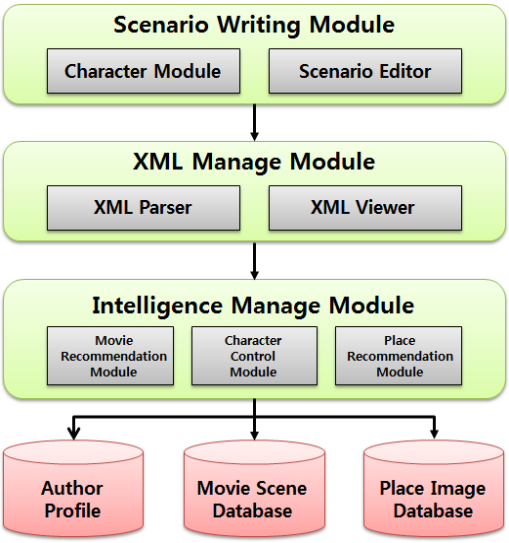
[그림 2-1] 스토리 생성 시스템 전체 모델[4]

조선대학교 CHOSUN UNIVERSITY

창작자는 이러한 기술이 바탕 되어지는 스토리 생성 시스템을 이용하여 창의적이고 다양한 스토리를 창작할 수 있다. [그림 2-1]은 스토리텔링 저작을 위해 인간과 컴퓨터간의 상호작용 (HCI, Human Computer Interaction)기술을 바탕으로 스토리 생성 시스템의 전체모델을 보여주고 있다. 이는 기존의 스토리 창작 소프트웨어인 워드프로세서 소프트웨어의 확장형과는 다른 스토리텔링 방식을 보여주고 있다.

스토리 생성 시스템은 컴퓨터가 스토리에 대한 정보에 대한 체계적이고 집합적인 지식(Domain Knowledge)을 갖추고 있으며 창작자는 컴퓨터와 함께 스토리 창작과정에 참여함으로써 새로운 스토리텔링 방식을 갖추고 있음을 보여준다[4].

라창현[5]은 스토리텔링 저작 도구 개발에 필요한 영화 및 드라마의 시나리오를 국내사례 중심의 분석을 토대로 메타데이터를 추출하였으며 사용자인 작가의 프로파일, 동영상 데이터베이스, 장소 이미지 데이터베이스를 기반으로 작가의 시나리오 작성을 위한 지능형 시나리오 작성 모듈을 제안하였다.

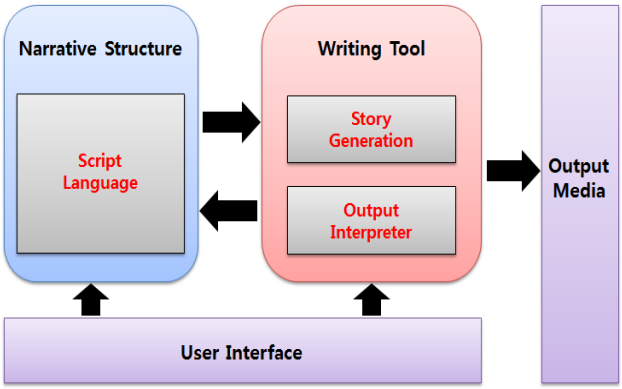


[그림 2-2] 지능형 관리 모듈[5]

김석규[6]는 인간과 컴퓨터간의 상호작용(HCI, Human-Computer Interaction)을 통한 다양한 스토리를 생성하는 기법인 인터랙티브 스토리텔링 시스템 구현방법에 대한 모델을 제안하였다. 시스템 구성요소인 서술구조를 표현하기 위한 방법을 제시하였으며 기존의 스크립트 언어 및 표현방식에 대한 분석을 통하여 스토리 생성

조선대학교 CHOSUN UNIVERSITY

을 위한 스크립트 언어 SML(Storytelling Mark up Language)를 제시하고 이를 통하여 [그림 2-3]과 같은 인터랙티브 스토리텔링 시스템을 구현하였다.



[그림 2-3] 인터랙티브 스토리텔링 시스템[6]

2. 스토리텔링 저작 지원 도구

스토리 생성을 위한 스토리텔링 시스템을 구성하기 위해 다양한 방법들이 제시되고 있다. 여기서 스토리텔링 시스템이란 스토리텔링 저작 지원 도구를 의미하며 이야기를 만들기 위해 사용되는 컴퓨터 소프트웨어를 말한다. 스토리텔링 시스템을 사용하는 창작자들은 스토리 저작을 위한 도구로써 워드프로세서나 스토리 데이터베이스, 상용 프로그램 등의 다양한 컴퓨터 기술을 이용하여 활용하고 있다. 이는 창작자에게 창작에 대한 보조 도구로서의 역할을 하고 일련의 스토리 생성의 자동화 과정을 통해 필요한 정보를 제공해줌으로써 영화, 게임, 소설, 드라마 등 다양한 문화산업에 전반적으로 활용되고 있다.

하지만 스토리텔링 저작 도구는 단지 컴퓨터 프로그래밍 기술에 의해 구현되는 것은 아니다. 컴퓨터 공학 이외에도 서사학이나 창작이론과 같은 인문적 지식도 함께 요구되므로 여러 분야에서의 학문적 요소에 대한 융합이 요구된다. 즉, 스토리텔링 저작 지원 도구를 설계 및 구현함에 있어서 스토리가 지닌 구조 분석을 통하여 [표 2-1]과 같은 기술적 요소들이 중요한 점으로 대두되고 있다.

현재 스토리 구조를 분석하여 컴퓨터 기술의 인공지능이나 알고리즘을 통해 스토리를 가공함으로써 컴퓨터가 스스로 이야기를 만드는 자동화 생성 기술에 대한 연구가 국내외적으로 진행되고 있다. 하지만 이에 따른 연구 성과물들은 기초적인

조선대학교 CHOSUN UNIVERSITY

수준에서 스토리 생성이 가능하다는 한계점이 있다.

본 절에서는 스토리 창작 활동을 지원하는 국내외 대표적 스토리텔링 저작 도구 사례에 대해 기술한다.

[표 2-1] 스토리텔링 저작 도구 개발에 요구되는 기술적 요소[3]

분야	기술	특징	활용 사례
서사학/ 창작이론	이야기 구조 분석	스토리텔링 저작 도구 구현에 가장 근본이 되는 이론적 토대를 제공	모티프, 플롯, 캐릭터 분석에 활용
	창작 과정 분석	창작자의 스토리 창작 과정에 대한 인지적, 기능적 특징 분석	창작 과정에 필요한 도움말, 가이드 제공
전산학/ 프로그래밍	데이터베이스	다양한 스토리 사례 ,데이터를 속성별로 데이터베이스 구축	소프트웨어에 포함되는 이야기 DB 구축
	UI 프로그래밍	소프트웨어의 UI, 메뉴화면, 기능 구현	소프트웨어 구현

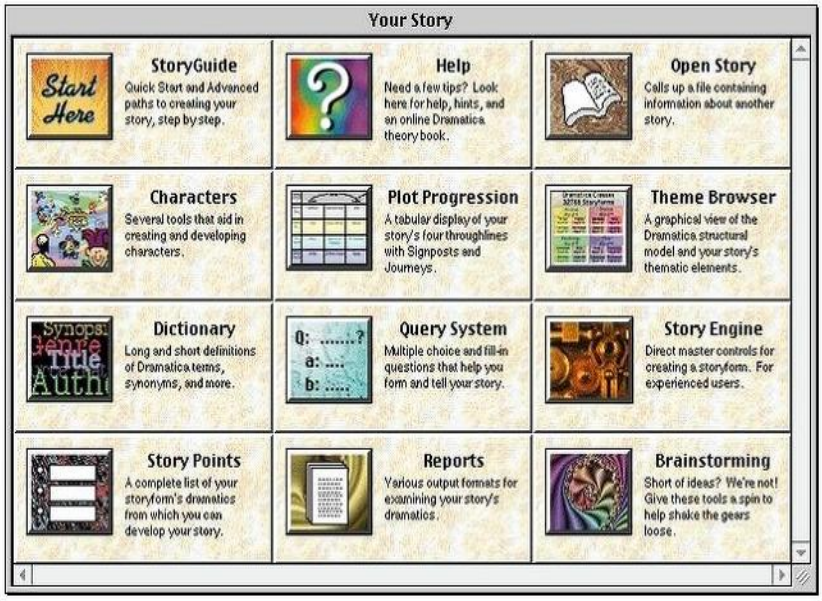
2.1 드라마티카 프로(Dramatica Pro)

드라마티카 프로는 미국의 라이트 브라더스(Write Brothers)에서 개발하였으며 영화·드라마 시나리오에 대한 스토리텔링 저작 소프트웨어이다. 드라마티카 프로는 캐릭터를 명확하게 설정하는 기능과 플롯(Flot)에 대한 체계화를 구성하고, 스토리 전개를 위한 전반적인 테마 레이어(Theme Layer)를 관리할 수 있는 스토리텔링 저작 지원 도구라고 할 수 있다. 이외에도 저작을 지원하기 위한 12가지 도구를 제공하고 있으며, 이에 산출된 작업을 스토리엔진(Story Engine) 모듈을 통하여 하나의 완결된 스토리를 생성할 수 있도록 제공하고 있다. 또한, 인간과 컴퓨터가 상호작용할 수 있는 쿼리 시스템(Dramatica Query System: DQS)을 통해 컴퓨터가 전달하는 주요 질문에 대한 창작자의 대답을 유도하는 스토리 저작 방식을 특징으로 가지고 있다.

드라마티카 프로는 스토리 이론(New Theory of Story)에 바탕을 두고 있으며, 기존의 작품에 대한 스토리 구조 분석을 토대로 새로운 형태로 이론화시킴으로써 캐릭터에 대한 유형과 플롯에 대한 갈등 양상을 구축할 수 있으며, 창작자는 그 틀에 맞춰 캐릭터, 플롯 속성, 각 속성들과의 관계를 설정할 수 있다.

조선대학교 CHOSUN UNIVERSITY

드라마티카 프로는 단순히 워드프로세서 기반의 시나리오 창작이 아닌 문예창작 이론을 컴퓨터 기술과 응용하여 실질적인 활용할 수 있다는 점에서 다른 스토리텔링 저작 도구와 차별성을 두고 있다.



[그림 2-4] 드라마티카 프로의 초기 화면

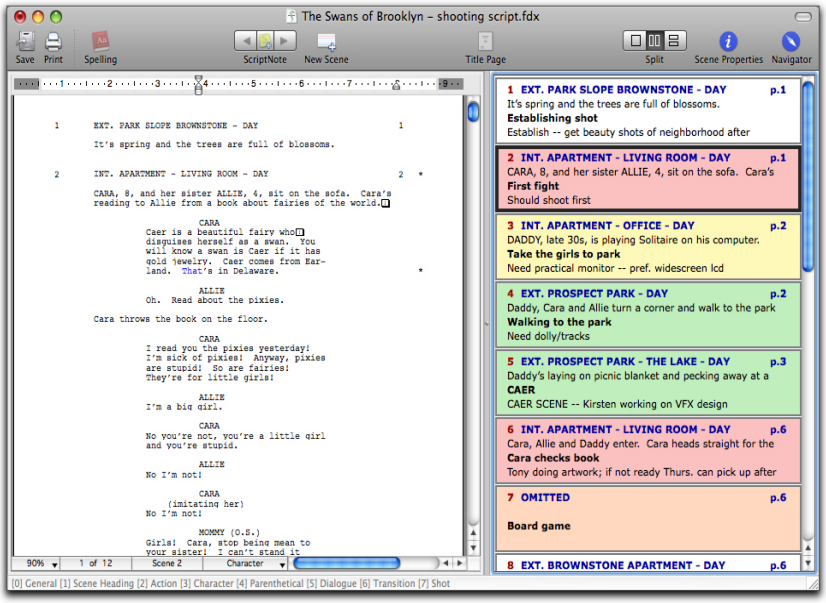
2.2 파이널 드래프트(Final Draft)

드라마티카 프로와 더불어 국외에서 많이 사용되고 있는 스토리텔링 저작 도구는 파이널 드래프트이다. 이는 연극과 TV 에피소드, 영화 스크립트 제작 등 시나리오 저작에 중점을 둔 저작 도구로써 다양한 유형의 시나리오 템플릿을 제공하여 스토리의 흐름을 검토할 수 있는 기능을 가지고 있다. 특히 인덱스 카드(Index Card Script) 기능을 통하여 여러 페이지를 한 화면에 볼 수 있도록 제공하고 있으며 시나리오 포맷에서 등장인물, 지문, 대사, 장면전환, 카메라 쇼트와 같은 문서서식에 대한 단축키를 제공하는 편리한 인터페이스를 가지고 있다.

특히, 새로운 문서를 생성할 때 영화, 방송, 소설을 작성할 수 있는 '스크립트', 'TV 템플릿', '텍스트 문서'의 3가지 양식을 제공함에 있어서 다양한 포맷을 선택할 수 있다. 하지만 시나리오 작성위주의 저작 도구로써 스토리와 캐릭터를 구조화할 수 있는 기능의 부재와 시나리오 작성 시 저장기능에 대한 오류로 인한 문제점이

조선대학교 CHOSUN UNIVERSITY

지적되고 있다.



[그림 2-5] 파이널 드래프트의 시스템 실행 화면

2.3 스토리 헬퍼(Story Helper)

국외 상용되고 있는 저작 도구와 더불어 영화 시나리오 및 방송 대본 작성의 저작 지원을 위해 국내 스토리텔링 저작 도구인 스토리 헬퍼가 이화여자대학교 연구진과 NC SOFT간의 공동연구로 개발되었다. 스토리 헬퍼는 드라마티카 프로와 파이널 드래프트처럼 스토리의 창작을 도와주는 저작 도구로써 1500편의 영화 및 애니메이션을 검토하여 3만 여개 요소의 데이터베이스를 생성하고, 205개의 모티프와 36개의 에피소드로 재정리하여 작가들의 스토리 구상에 따라 자유롭게 검색, 대조, 추출, 검토, 재구성 할 수 있도록 개발된 저작 도구이다. 하지만 영화 시나리오 중심의 스토리텔링 저작 지원도구로써 스토리 창작을 위한 작업의 초기단계에서 많은 인력과 시간이 소모되는 자료의 수집 및 분류에 대한 필요한 내용 전개에 큰 도움을 주기 어렵다는 문제점이 존재한다.

조선대학교
CHOSUN UNIVERSITY

[그림 2-6] 스토리 헬퍼의 실행 화면

3. 스토리텔링을 위한 서사구조 모델

본 장에서는 스토리의 서사구조에 따른 서사정보 및 인물관계에 대한 패턴을 분석하기 위해 본 논문에서 적용하고자 하는 서사구조 모델에 대한 분석을 기술한다.

3.1 스토리 전개에 따른 기승전결 서사구조

일반적으로 스토리텔링의 대표적인 서사구조는 기승전결의 형식을 갖추며 스토리 흐름이 분류되고 각 단계별로 스토리 내 인물간의 관계가 특정 사건으로 인해 변화하거나 지속된다. 기승전결을 분류하는 방법에는 전반부, 중반부, 후반부로 3단계로 나누거나 발단, 발전, 절정, 해결의 4단계 또는 발단, 전개, 위기, 절정, 결말의 5단계로 나눌 수 있으며 [표 2-2]는 서사구조의 단계별 내용을 나타낸다. 먼저, 발단부분에 해당하는 '기' 단계는 전체적인 스토리 전개를 위해 기본적인 정보를 제공하는 단계로 스토리에서 등장하는 인물이 소개되거나 사건의 실마리, 갈등이 암시되는 부분이다. '승' 단계는 등장인물의 성격이 변화하거나 다양한 사건들에 대한 복선이 암시되는 단계로 인물간의 관계가 뚜렷해지며, '전' 단계에서는 스토리의 절정부분에 속하며 인물간의 갈등 및 새로운 사건이 발생하고 인물간의 관계가 적대적 관계에서 우호적 관계 또는 우호적 관계에서 적대적 관계로 변할 수 있는 단계로 인물간의 관계에 따른 갈등의 최고조에 이르는 단계이다. 마지막으로 '결' 단계

조선대학교 CHOSUN UNIVERSITY

에서는 인물간의 관계가 마무리되는 단계로써 스토리의 주체가 되는 인물의 운명이 결정된다.

[표 2-2] 서사구조의 단계별 내용[7]

3단	4단	5단	내용
전반부	발단	발단	등장인물의 소개 배경의 제시 사건의 실마리 제시 갈등의 암시
중반부	발전	전개	등장인물 성격의 변화·발전 다양한 복선과 암시
			사건의 본격적 전개 갈등과 분규의 제시
	절정	위기	갈등과 분규의 상승 새로운 사건의 발생
		절정	갈등의 최고조 결말을 향한 분기점 해결의 실마리
후반부	해결	결말	인물의 운명 결정 갈등과 분규의 해결

3.2 조셉 캠벨의 영웅 서사구조

미국의 신화학자 조셉 켐벨(J.Cambell)은 시대와 지역을 막론하고 모든 신화의 근본적인 구조를 영웅들의 모험담 중심으로 사건정보를 3단계로 나누고 12가지의 모티프로 세분화 하였다[8]. 이러한 영웅 서사구조는 단순한 여행 스토리가 아닌 인간의 보편적인 성장 스토리의 일부이며 기존 공간에서의 이탈, 그로 인해 겪는 고난, 그 과정에서 우정과 사랑으로 인한 아픔을 극복하기 위한 영웅의 활약, 보조자의 도움, 그리고 최후 갈등과의 싸움 속에서 행복한 결말이라는 과정은 할리우드 영화 시나리오 구조에 전반적인 영향을 주었다. 본 논문에서는 설화 텍스트에서 추출된 서사정보에 대한 관계 설정을 위한 방법으로 영웅 서사구조의 인물, 사건, 배경 관계모델을 활용한다.

조선대학교 CHOSUN UNIVERSITY

① 인물(Character)

스토리에서 등장하는 인물의 원형 없이는 스토리를 전개시킬 수 없다. 따라서 설화 속에서 등장하는 인물(Character)의 유형을 역할에 따라 분류할 수 있다. 캠벨의 영웅 서사구조에서는 인물의 역할에 따라 영웅, 정신적 스승, 전령관, 관문 수호자, 변신 자재자, 그림자, 장난꾸러기의 7가지로 분류하였다.

[표 2-3] 인물정보 분류

인물	역할
영웅 (Hero)	- 특별한 세계로 모험을 떠남 - 영웅적인 임무 수행
그림자 (Shadow)	- 영웅과 반대되는 입장 - 주인공을 위기에 빠뜨림
전령관 (Herald)	- 주인공에게 모험을 떠나도록 경고, 자극 - 동기부여
정신적 스승 (Mentor)	- 영웅에게 능력 및 물건 등의 권한 부여 - 조언
관문 수호자 (Threshold Guardian)	- 영웅이 가는 길목에 지켜섬 - 영웅을 방해, 시험
변신 자재자 (Shape Shifter)	- 영웅을 속임 - 아군이 적군으로, 적군이 아군으로 변모함
장난 꾸러기 (Trickster)	- 극에 혼란을 일으킴

② 사건(Event)

스토리 창작에 있어 등장인물간의 조력적, 적대적 관계에 따라 특정 사건이 전개되고 이를 바탕으로 스토리의 내용은 더욱 풍성해질 수 있다. 영웅 서사구조에서는 시대와 지역을 막론하고 모든 신화의 근본적인 구조와 단계는 동일하다고 정의한다. 따라서 영웅 서사구조 기반의 설화 속에서 진행되는 사건을 사건이 시작하는 출발 단계와, 전개 단계, 귀환단계의 3단계로 분류할 수 있다. 캠벨의 영웅 서사구조에서는 영웅은 평범한 일상의 세계에서 초자연적인 신비의 땅으로 뛰어드는 출발(Departure), 미지의 세계에서 다양한 인물들과 조우하며 결정적인 승리를 쟁취하는 전개(Initiation), 모험을 끝으로 현실세계로 돌아오는 귀환(Return)의 3단계로 분류하고, 이에 12가지의 사건을 소분류로 나누었다. [표 2-4]는 사건정보에 따른 분류를 나타낸다.

조선대학교 CHOSUN UNIVERSITY

[표 2-4] 사건정보 분류

대분류	중분류	소분류
Event	출발 (Departure)	일상세계
		모험에의 소명
		소명의 거부
		정신적 스승과 만남
		첫 관문의 통과
	전개 (Initiation)	시험, 협력자, 적대자
		동굴 깊은 곳으로의 진입
		시련
		보상
	귀환 (Return)	귀환의 길
		부활
		영약을 가지고 귀환

③ 배경(Setting)

영웅 서사구조에서는 영웅이 모험 하는 과정에 대한 배경 요소로 인물과 사건이 존재하는 공간인 현실세계와 초현실세계로 구분한다. 주로 초현실세계의 배경은 오지의 땅, 숲, 지하 왕국, 물밑 세계, 하늘 위와 같은 미지의 공간으로 분류하였고, 현실세계의 배경은 산, 강, 바다와 같은 자연적인 공간으로 분류하였다. 이는 설화 스토리의 다양한 공간적, 시간적인 배경을 포함하고 있다. [표 2-5]는 배경정보에 따른 분류를 나타낸다.

[표 2-5] 배경정보 분류

공간	상세
초현실세계 (Adventure)	- 사건의 배경이 되는 시대와 공간
현실세계 (Real-word)	

조선대학교 CHOSUN UNIVERSITY

B. 설화 데이터 분석을 위한 자연어 처리 시스템

한국어의 특성상 복잡한 형태소들로 이루어져 있으며, 동음이의어가 많이 존재함에 따라 이를 해결하기 위한 많은 연구가 진행되고 있다. 본 절에서는 한국어로 작성된 설화 데이터를 처리하기 위한 방법으로 한글 형태소 분석기에 대해 기술한다. 영어는 굴절어로써 문장에 대한 구조 분석 방법이 문장을 구성하는 단어들에 대한 품사 태깅(Part of Speech Tagging)을 통해 해당 단어에 대한 품사를 추출하고 단어의 의미 분석을 통해 문장이 지닌 의미를 파악할 수 있다. 하지만 한국어는 교착어로써 복잡한 문장 구조를 가지고 있기 때문에 영어와 비교하였을 때 정확한 의미 파악이 힘들다고 할 수 있다[9][10].

예를 들면, 영어는 일반적인 자연어 처리(NLP, Natural Language Processing)과정에서 전처리(불용어 및 특수문자 제거)를 통해 일반명사, 고유명사, 동사, 형용사 등의 단어에 대한 품사를 POS Tagger를 통해 추출이 가능하지만, 한국어는 전처리 과정에서 조사, 어미, 접사 등의 영어에서 찾아보기 힘든 품사들이 존재하기 때문에 기존에 개발된 한글 형태소 분석기를 사용하여 데이터를 처리한다.

1. 꼬꼬마 형태소 분석기

한국어 자연어 처리를 위한 시스템으로써 서울대학교 IDS(Intelligent Data System) 연구실에서 개발된 꼬꼬마 한글 형태소 분석기[11]가 있다. 꼬꼬마 형태소 분석기는 Java 기반의 오픈소스(Open Source)로 배포되고 있으며 한국어 처리를 위한 연구를 수행할 때 사용되는 라이브러리와 사전 데이터를 포함하고 있다. 특히, 대량의 코퍼스(Corpus)가 필요할 경우 '세종 말뭉치'를 이용할 수 있는 시스템을 구축하여 품사, 형태소, 문어 및 구어 등의 정보를 확인할 수 있는 다양한 용도로 활용되고 있다. 꼬꼬마 형태소 분석기는 띄어쓰기 오류에 덜 민감하게 구현된 한글 형태소 분석기로 검증되었으며 휴리스틱(Heuristic)과 히든 마르코프 모델(HMM:Hidden Markov Model)에 기반한 확률 모델(Probabilistic Model)을 이용하여 상용 형태소 분석기와 성능을 비교하였을 때 뒤지지 않는 성능을 보이고 있다. 또한, 태그 집합으로써 [표 2-6]과 같이 '체언, 용언, 관형사, 부사, 감탄사' 등의 품사를 나누고 각 세부 품사를 구분한다. 하지만 한국어 처리를 위한 실행 효율성 및

조선대학교 CHOSUN UNIVERSITY

정확성이 높은 반면에 접근성 및 확정성이 떨어지고 다양한 요구에 대해 유연하게 대처할 수 없다는 단점이 존재하고 있다[12].

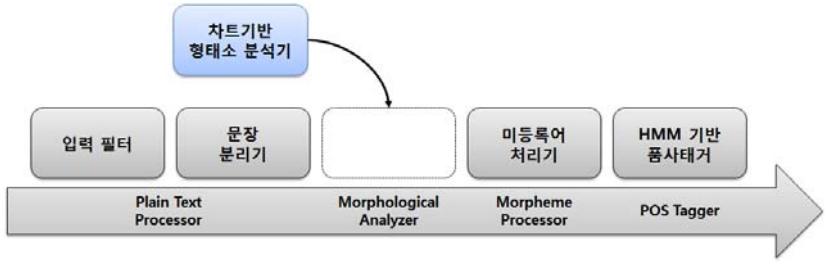
[표 2-6] 꼬꼬마 형태소 분석기의 태그 집합 예[11]

대분류	세종 품사 태그		대분류	세종 품사 태그	
	태그	설명		태그	설명
체언	NNG	일반명사	용언	VV	동사
	NNP	고유명사		VA	형용사
	NNB	의존명사		VX	보조 용언
	NR	수사		VCP	긍정 지정사
	NP	대명사		VCN	부정 지정사
관형사	MM	관형사	조사	JKS	주격 조사
부사	MAG	일반 부사		JKC	보격 조사
	MAJ	접속 부사		JKG	관형격 조사
감탄사	IC	감탄사		JKO	목적격 조사
어말 어미	EF	종결 어미		JKB	부사격 조사
	EC	연결 어미		JKV	호격 조사
	ETN	명사형 전성 어미		JKQ	인용격 조사
	ETM	관형형 전성 어미		JX	보조사

2. 한나눔 형태소 분석기

한나눔 형태소 분석기[13]는 기존 형태소 분석기의 접근성과 유연성의 단점을 보완하기 위해 한국과학기술원(KAIST) 시맨틱 웹 첨단연구센터에서 개발하고 관리하고 있는 오픈소스(Open Source)기반의 형태소 분석기이다. 한나눔 형태소 분석기는 한국어 처리 목적에 따른 플러그인 컴포넌트 아키텍처 구조를 지니고 있으며 필요에 따라 다양한 플러그인을 지원한다. 또한, UI(User Interface)기능을 통해 이용자의 필요에 따라 플러그인 형태를 추가하는 방식으로 사용이 가능하며 다양한 필터를 추가함으로써 확장이 가능하도록 한다. 따라서 유연하게 한국어 데이터를 처리하기 위한 워크플로우를 구성할 수 있으며 다양한 목적에 맞게 활용할 수 있다[13][14]는 장점을 지니고 있다. 이에 본 논문에서는 한국어로 작성된 설화 데이터를 처리하기 위해 한나눔 형태소 분석기를 활용하였다.

조선대학교 CHOSUN UNIVERSITY



[그림 2-7] 한나눔 형태소 분석기의 워크플로우[13]

[그림 2-7]는 한나눔 형태소 분석기의 워크플로우를 나타낸다. Plain Text Processor 단계에서는 문장 경계 인식 및 필터링, 자동 띄어쓰기 등의 형태소 분석을 위한 전처리 작업을 수행한다. 문장 경계 인식 방법으로는 문장의 구분자 역할을 할 수 있는 마침표 '.', 물음표 '?', 느낌표 '!'를 기준으로 문장 구분을 결정한다.

Morphological Analyzer단계에서는 입력된 문장에 대한 어절 단위로 발생 가능한 모든 형태소 분석 결과를 생성한다. 차트 기반 형태소 분석기는 형태소 분석을 위해 내부 저장 공간을 말하며 Morpheme Chart, Segment Position, Inverse Segment Position으로 구성되는 Lattice 형태의 차트를 사용하고 있다.

Morpheme Processor 단계는 미등록어를 처리하는 단계로써 형태소 분석 과정에서 사전에 등록되지 않는 단어를 처리하기 위해 'unk' 태그를 사용한다. 예를 들어, "나는 지금 불닭을 먹고 있다." 라는 문장에서의 '불닭'은 형태소 사전에 등록되지 않은 단어로써 [표 2-7]과 같은 분석 후보를 생성한다.

[표 2-7] 분석 후보 생성

불닭을	
	불닭/unk+을/jco
	불닭을/unk

이에 대해 한나눔 형태소 분석기의 미등록어 처리기는 "불닭/unk"에 대해 비서술성명사(ncn)와 고유명사(nq)를 태깅하며 [표 2-8]과 같은 결과를 생성한다.

[표 2-8] 미등록어 처리기

@불닭을	
	불닭/ncn+을/jco
	불닭/nq+을/jco

조선대학교 CHOSUN UNIVERSITY

마지막으로 POS Tagging 단계는 [표 2-9]와 같이 형태소 태그 집합을 통해 입력된 데이터로부터 최종 품사 태깅 결과를 반환하게 된다. 이는 어절에 대한 토큰화를 통해 해당 어절에 대한 품사를 HMM 기반의 품사 태거를 사용하여 한국어 처리를 가능하게 한다. 또한, 명사 추출기(Noun Extractor)를 통해 형태소 분석 및 품사 태깅 결과로 명사로 인식된 형태소만 추출하여 제공하는 기능이 있다.

[표 2-9] 한나눔 형태소 분석기의 태그 집합 예[13]

상위 분류			태그			
기호 s			sp 쉼표 sl 여는 따옴표 sd 이음표 su 단위기호		sf 마침표 sr 닫는 따옴표 se 줄임표 sy 기타기호	
외국어 f			f 외국어			
체언 n	보통명사 nc	서술성명사 np	ncpa 동작성 명사		ncps 상태성 명사	
		비서술성명사 ncn	ncn 비서술성 명사		ncr 직위 명사	
	고유명사 nq		ncpa 성 nqpc 성+이름		nqpb 이름 nqq 기타 -일반	
	의존명사 nb		nbu 단위성 의존명사 nbn 비단위성 의존명사			
	대명사 np		npp 인칭대명사		npd 지시대명사	
	수사 nn		nnc 양수사		nno 서수사	
용언 p	동사 pv		pvd 지시 동사		pvg 일반 동사	
	형용사 pa		pad 지시형용사		paa 성상형용사	
	보조용언 px		px 보조용언			

조선대학교 CHOSUN UNIVERSITY

C. 토픽맵

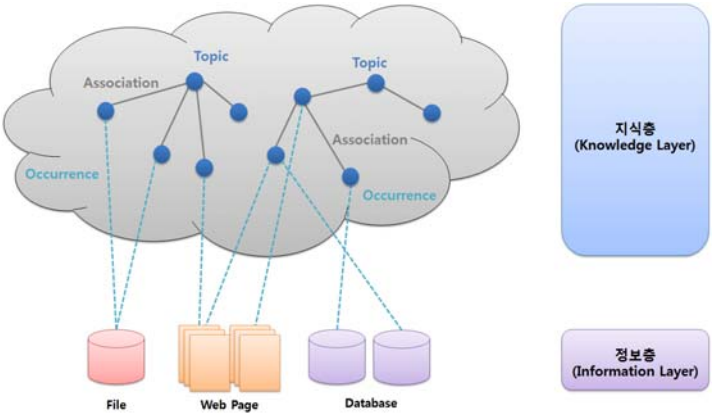
토픽맵은 국제표준화기구(ISO)에서 시맨틱 웹을 구현하기 위한 온톨로지 생성 언어 중 하나이며 W3C(Word Wide Web Consortium)의 OWL과 상호 보완적이고 경쟁 관계에 있으며 용어의 의미적 구조를 나타내는 용어집, 색인집, 시소러스를 다루는 목적이 있다[15]. 현재 토픽맵은 시맨틱 웹의 핵심기술로 인정받으며 지식을 표현하는 방법론으로써 정의된 구조와 지식 자원에 대한 의미적 관계를 지도(Map)로 표현하고 검색이 가능하도록 하며, 정보에 대한 시각화(Visualization)에 목적이 있다. 또한, W3C의 RDF, OWL 보다 많은 정보를 가지며 정보를 관리함에 있어서 GPS(Global Positioning System)와 같은 역할을 하는 토픽맵은 지식표현과 정보관리의 각 응용분야간의 다리역할을 하며 매우 간단하고 복잡한 것에서 단순하고 모호한 정보까지 지식표현 스키마의 전체적인 범위를 처리할 수 있다[16].

본 절에서는 서사정보 추출 시스템으로부터 추출된 스토리 데이터에 대한 관리 및 시각화를 위해 토픽맵 기반의 응용방안을 제시하기 위해 토픽맵의 구조와 그에 대한 구성요소를 살펴보고, 토픽맵을 이용한 정보 시각화 사례에 대해 기술한다.

1. 토픽맵의 구조 및 구성요소

토픽맵은 지식을 관련된 정보와 연계를 위한 추상적 구조를 말하며 2개의 레이어로 구성되어 있다. [그림2-7]는 토픽맵 모델을 나타내는 것으로 지식 레이아웃층(Knowledge Layout)과 정보 레이아웃층(Information Layout)을 나타내고 있다. 상위(Upper)계층인 지식 레이어는 색인의 역할을 하며 주제를 실체화하는 토픽(Topic)을 기본요소로 토픽간의 관계를 설정하는 어소시에이션(Association), 토픽의 연관 정보자원의 위치를 나타내는 어커런스(Occurrence)로 구성된다[17]. 토픽에 다양한 명칭을 부여하기 위한 명명법과 토픽의 개념에 대한 범위를 한정하는 Scope, 토픽의 역할을 규정하는 Role 등 3가지의 보조 개념도 사용된다[18]. 하위(Lower)계층인 정보 레이어는 정보의 콘텐츠 자원영역을 나타내는 역할을 하며 텍스트나 그래픽, 비디오, 오디오, 웹페이지 등 어떠한 포맷도 가능하다.

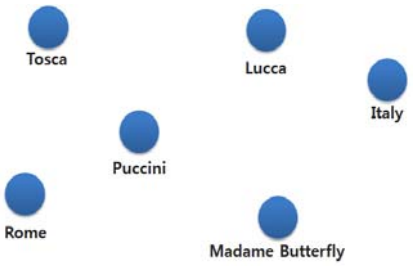
조선대학교
CHOSUN UNIVERSITY



[그림 2-8] 토픽맵의 구조[19]

① 토픽(Topic)

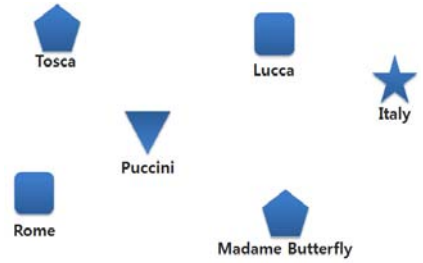
토픽[20]은 일반적으로 사람, 개체(Entity), 주제(Subject), 단어 등 어떠한 것이든 될 수 있으며 특정 속성이나 어떤 의미 등을 기계적으로 이해하고 처리할 수 있도록 객체(Object)로 표현 되는 것을 말한다. 주로 명사형으로 표현되며 각각의 토픽은 하나의 이름(name)을 갖는다.



[그림 2-9] 토픽

다양한 토픽들을 서로 연결하기 위해 토픽 타입(Topic Type)을 가질 수 있는데, 토픽 타입은 토픽의 유형을 말하며 토픽을 유형에 따라 분류시켜 주는 역할을 한다. 또한, 토픽은 하나 이상의 토픽 타입에 속할 수 있고 토픽 타입 자체가 또 하나의 토픽으로 정의될 수 있으며 경우에 따라 토픽 타입이 없는 토픽이 존재할 수 있다.

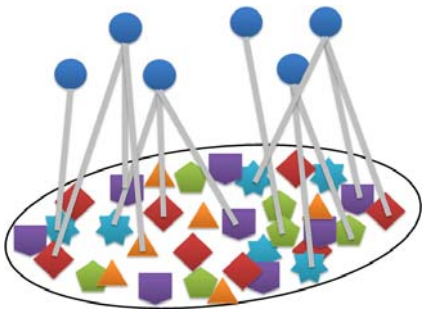
조선대학교 CHOSUN UNIVERSITY



[그림 2-10] 토픽 타입

② 어커런스(Occurrence)

어커런스는 토픽을 하나 이상의 정보 리소스와 연결 시켜주는 기능을 하고 음악 파일, 이미지파일, 문서파일, 데이터베이스 내 특정 레코드 등의 외부 정보 리소스에 대한 참조를 나타낼 수 있으며 웹 페이지에 존재하는 정보 또한 URL을 통해 어커런스가 될 수 있다. 즉, 토픽맵에 존재하는 모든 토픽은 HyTime 어드레싱 또는 Xpointer와 같은 기법들을 사용하여 자신이 참조하는 리소스를 가르키며, 어커런스에 의해 참조되는 리소스는 일반적으로 토픽맵과 별도로 저장된다.

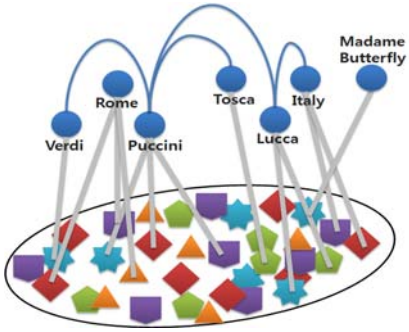


[그림 2-11] 어커런스

③ 어소시에이션(Association)

연관관계는 둘 이상의 토픽에 대한 관계를 정의하는 것이며 스토리 도메인의 네트워크를 보여주기 위해 사용되는 것으로 토픽간의 상하관계가 아닌 의미적 관계를 정의한다. 아래 [그림 2-11]은 토픽과 어커런스, 그들 간의 관계를 정의한 것이다.

조선대학교
CHOSUN UNIVERSITY

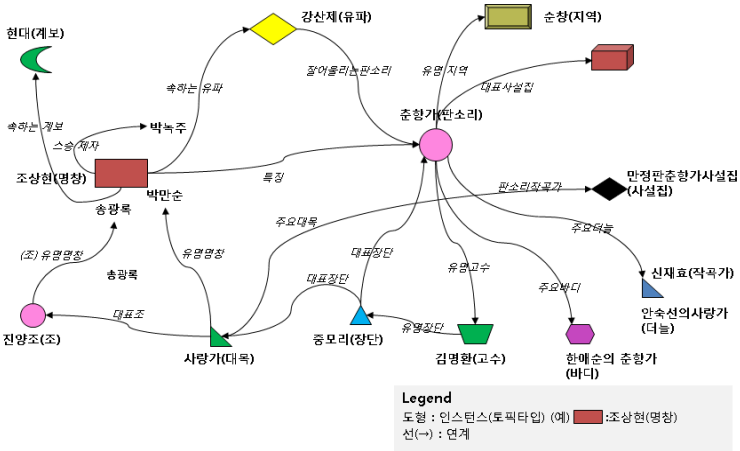


[그림 2-12] 어소시에이션

이러한 토픽맵의 특징으로 인해 기존의 시소러스 체계인 상하 관계(Broad Term/Narrow Term)나 유사용어 관계(Related Terms)체계와 비교하면 아주 우수한 장점이다. 다시 말해, 다양한 형태의 연계를 자연어(Natural language)형식으로 표현하며 토픽 간의 고정된 관계를 설정한 것이 아니므로 기존 분류시스템과 다른 유연성 측면에서 큰 차이를 보인다고 할 수 있다[19].

2. 토픽맵을 이용한 정보 시각화

오삼균[21]은 판소리에 대한 정보 시각화를 위해 토픽맵 기반 판소리 검색 시스템을 구축하였다. 판소리와 관련된 명창, 유파, 대표 장단 등의 요소 간 관계를 토픽맵으로 시각화하였으며, 기존 판소리 검색시스템과 질의별 실험을 수행하였다.

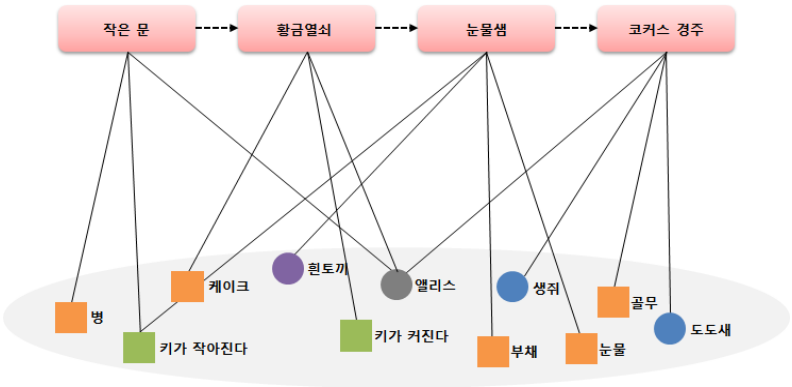


[그림 2-13] 판소리 토픽맵[21]

조선대학교 CHOSUN UNIVERSITY

권진만[22]은 토픽맵을 이용하여 뮤지션에 대한 정보를 시각화하기 위해 가수, 작곡가, 노래를 토픽타입(Topic Type)으로 분류하고 관련된 정보를 동영상과 음악파일, 웹 페이지의 노래정보와 연결하였다. 토픽간의 연계타입(Association Type)으로는 '부르다', '작사하다', '작곡하다'로 연결하였으며 최종적으로 뮤지션 검색시스템 구축을 통해 기존의 음악 검색사이트와 정보검색에 대한 시각화 측면의 비교실험을 수행하였다.

이지수[23]는 일련의 흐름을 갖는 스토리 라인의 형상화를 위해 스토리 내 존재하는 등장인물과 그들 주변에서 발생하는 사건관계를 토픽맵 기반의 시각화 방법을 제안하였다. 스토리 대상으로는 루이스 캐롤(Lewis Carroll)의 <이상한 나라의 앨리스>를 중심으로 수행하였으며, 스토리 내 요소들을 프로프의 '민담 형태론'에 근거하여 토픽맵 기반의 시각화 방법을 제시하였다.



[그림 2-14] 이상한 나라의 앨리스의 플롯 중심 토픽맵[22]

[그림 2-14]는 토픽맵을 이용하여 정보 레이아웃에 스토리 요소인 인물, 장소, 키워드의 토픽을 배열하고, 지식 레이아웃에는 정보 레이아웃의 토픽들에 대한 관계를 나타내는 이벤트를 나타냄으로써 플롯 내부에서의 인과관계를 표현하였다.

조선대학교 CHOSUN UNIVERSITY

Ⅲ. 서사정보 자동 추출 시스템 설계 및 구현

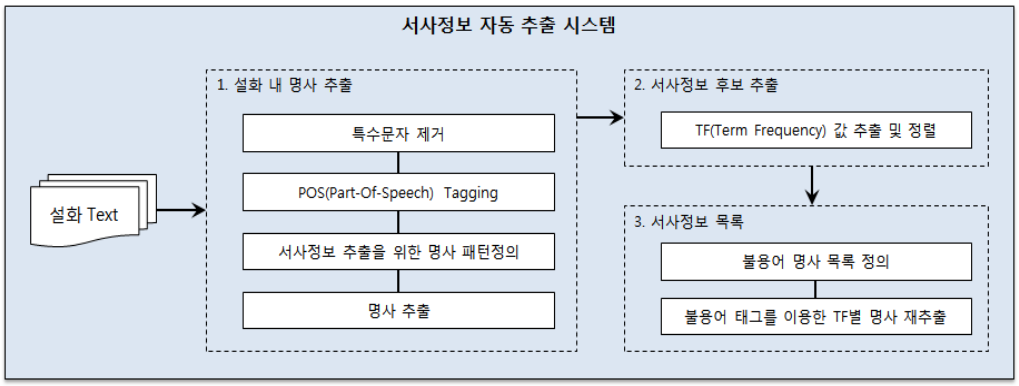
본 논문에서 제안하는 스토리텔링 저작 지원을 위 설화 내 서사정보를 자동으로 추출하기 위한 시스템을 설명하고, 그에 대한 방법에 대해 기술한다.

A. 서사정보 자동 추출 시스템 설계 및 인터페이스 정의

본 절에서는 한국어로 작성된 설화 텍스트를 자연어 처리(NLP, Natural Language Process) 기법을 활용한 서사정보 자동 추출 시스템 설계 및 구현을 위해 전체적인 시스템 구성도와 개발 환경에 대해 기술하고, 사용자 인터페이스를 보여준다.

1. 시스템 구성도

[그림 3-1]은 제안하는 설화 내 서사정보 자동 추출 시스템의 시스템 구성도이다. 시스템 구성도는 크게 입력받은 설화 텍스트로부터 한국어 자연어 처리 기법을 활용하여 명사를 추출하는 단계, 추출된 명사로부터 서사정보 후보명사를 추출하기 위해 TF 알고리즘을 적용하는 단계, 마지막으로 입력된 설화 텍스트로부터 인물, 사건, 배경의 추출된 서사정보 목록을 보여주는 단계로 이루어진다.



[그림 3-1] 시스템 구성도

조선대학교 CHOSUN UNIVERSITY

2. 개발 환경

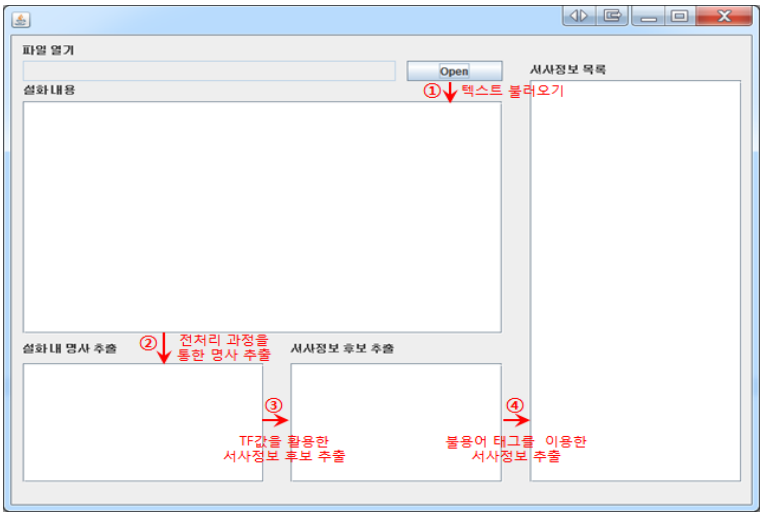
서사정보 자동 추출 시스템의 구축을 위한 개발환경은 [표 3-1]과 같다. 컴퓨터 CPU는 Intel(R) Core(TM) i7 2.93GHz 환경에서 Java 기반으로 구현하였으며, 인터페이스 설계를 위해 Eclipse 플러그인 Window Builder를 이용하여 레이아웃을 구성하였고, 텍스트 처리를 위해 한나눔 형태소 분석기의 품사 태그를 이용하였다.

[표 3-1] 서사정보 자동 추출 시스템 개발 환경

항목	사양
CPU	Intel(R) Core(TM) i7 2.93GHz
RAM	4.00GB
운영체제	윈도우7 Enterprise K
개발 툴	Java, Eclipse, Window Builder
분석기	한나눔 형태소 분석기

3. 사용자 인터페이스(User Interface)

[그림 3-2]는 구현된 서사정보 자동 추출 시스템의 사용자 인터페이스를 보여준다. 사용자 인터페이스는 시스템 구성도의 흐름과 같이 입력받은 설화 텍스트로부터 명사를 추출하는 부분과 추출된 명사 중에서 서사정보의 후보들을 추출하는 부분과 서사정보에 대한 목록을 보여주는 부분으로 구성된다.



[그림 3-2] 사용자 인터페이스

조선대학교 CHOSUN UNIVERSITY

B. 서사정보 추출을 위한 전처리 및 명사 추출 방법

본 절에서는 입력받은 설화 텍스트로부터 서사정보를 추출하기 위한 전처리(Pre-processing) 과정으로 [그림 3-2]의 사용자 인터페이스 화면에서 설화 내 명사를 추출하는 방법에 대해 기술한다. 명사를 추출하는 이유는 다양한 품사 중에서 서사정보를 구성하고 있는 단어에 대한 품사가 주로 명사에 분포되어 있기 때문이다. 이에 설화 텍스트 내 명사 추출을 위해 모든 자연어 처리 기법에서 이루어지는 특수문자 제거 방법과 해당 단어별 품사 처리를 위한 POS(Part-Of-Speech) Tagging 과정에 대해 설명하고 서사정보 추출을 위한 명사 패턴을 정의하여 추출한다.

1. 특수문자 제거 및 POS(Part-Of-Speech) Tagging

모든 언어에서 사용되는 특수문자의 종류는 매우 다양하다. 대화 내용을 나타내기 위해 큰따옴표(“ ”)를 사용하고, 문장의 맥락을 구성하기 위해 마침표(.), 쉼표(,) 등을 사용한다. 하지만, 이러한 특수기호들은 자연어 처리에서 큰 걸림돌이 되기 때문에 본 논문에서도 본격적인 한국어 처리를 하기 위해 특수문자를 제거한다. 제거할 특수 문자는 “*, +, $, |, !, #, %, &, @, ,, ?, ", ', \ , [,], /, <, >, ^, {, }”이며, 특수문자 제거 방법은 [표 3-2]과 같다. 특수문자 제거 시 주의할 사항은 제거할 특수문자를 정의할 때 대부분 대괄호([ ]) 안에 제거할 문자를 정의하지만 JAVA 함수에서 사용되는 대, 중, 소괄호와 큰따옴표, 작은따옴표는 반드시 앞에 역슬러쉬(\)를 포함하여 그 특수성을 지정해주어야 한다.

[표 3-2] 설화 텍스트 내 특수문자 제거 소스

```
public class specialchar {
        public static void main(String[] args){
                String input_text = "*+$|!#%&@,?\\[\\]\\/\\<\\>\\^\\{\\}";
                String output =
        input_text.replaceAll("[*+$|!#%&@,?\\[\\]\\/\\<\\>\\^\\{\\}]", "");
                System.out.println(output);

        }
}
```

조선대학교 CHOSUN UNIVERSITY

전체 텍스트 내에 특수문자를 제거 후 각 단어별 품사를 파악하기 위한 POS Tagging 과정을 진행한다. 한국어 자연어 처리에 있어서 매우 다양한 품사들이 존재하는데, 문맥적으로 특정 품사를 보조해주는 역할을 하는 품사들이 존재하기 때문에 영어와 달리 매우 어려운 점이 존재한다.

따라서, 본 논문에서는 기존에 한국과학기술원(KAIST) 시맨틱 웹 첨단연구센터에서 개발 및 관리를 하고 있는 한나눔 형태소 분석기를 통해 본 과정을 진행한다. 한나눔 형태소 분석기를 이용하는 이유는 타 분석기와 달리 특정 명사에 대한 품사 종류를 나타낼 수 있는 경우에 수에 따라 최대한 많이 보여줌으로써 연구자들이 사용할 수 있는 범위를 확대시켜 놓았기 때문이다. 또한 체계적으로 정리된 단어-품사 관계를 제공함으로써 본 논문에서 사용하고자 하는 불용어 처리에 큰 역할을 할 수 있을 것으로 사료되어 한나눔 형태소 분석기를 사용한다.

설화 텍스트에 대한 POS Tagging 과정은 먼저 설화 텍스트 원문을 입력받을 때, 모든 기호(“\n”)를 제거하여 하나의 문장형태로 데이터를 수정한다. 그리고 앞서 정의한 특수문자를 제거하는데, 마침표는 문장의 끝을 구분하여야 함으로 제거하지 않는다. 최종적인 POS Tagging은 한나눔 분석기 함수에 기반하여 데이터가 가공되는데 각각의 품사 별로 작성된 기반데이터를 통해 Tagging과정을 거치게 되며 본 논문에서 처리하는 품사 태깅 과정은 [표 3-3]와 같다.

[표 3-3] 설화 텍스트에 대한 POS Tagging 예

지하국대적퇴치 설화 원문	옛날 아귀 귀신이라는 큰 도적이 있었다. 그는 종종 이 세상에서 나와서 세상을 요란하게 하고 예쁜 여자를 납치해 가기도 하였다.…	
특수문자 제거	옛날 아귀 귀신이라는 큰 도적이 있었다. 그는 종종 이 세상에서 나와서 세상을 요란하게 하고 예쁜 여자를 납치해 가기도 하였다.…	
POS Tagging	옛날/**ncn** 아귀/**ncn** 귀신/**ncn**+이/**jp**+라/**ef**+는/**etm** 크/**paa**+ㄴ/**etm** 도적/**ncn**+이/**jcc** 있/**px**+었/**ep**+다/**ef** 그/**npp**+는/**jxc** 종종/**mag** 이/**mmd**	세상/**ncn**+에서/**jca** 나오/**pvg**+아/**ecx** 세상/**ncn**+을/**jco** 요란/**ncps**+하/**xsms**+게/**ecs** 하/**pvg**+고/**ecc** 예쁘/**paa**+ㄴ/**etm** 여자/**ncn**+를/**jco** 납치/**ncpa**+하/**xsva**+어/**ecx** 가/**px**+기/**etn**+도/**jxc** 하/**pvg**+었/**ep**+다/**ef**

조선대학교 CHOSUN UNIVERSITY

2. 서사정보 추출을 위한 명사 패턴 정의 및 추출

앞서 설명한 POS Tagging까지의 과정이 완료되면, 설화 텍스트 내 명사를 추출하기 위해 명사를 나타내는 품사의 종류를 파악하여 해당 품사만 선별해 내는 과정을 거친다. 현재 한나눔 형태소 분석기에서 정의하고 있는 명사 형태소의 구성은 ncpa(동작성 명사), ncps(상태성 명사), ncn(비서술성 명사), ncr(직위 명사), ncpa(성), nqpb(이름), nqpc(성+이름), nqq(기타 일반), nbu(단위성 의존명사), nbn(비단위성 의존명사)와 같다. 위 명사 형태소 구성에서 동일 패턴을 분석한 결과 다음과 같은 패턴을 갖는 것을 확인할 수 있었다.

[표 3-4] 한나눔 형태소 분석기 내 정의된 명사 패턴

○ nc + n, r, pa, ps - 일반 명사 및 동작, 상태성 명사
○ nq + pb, pc, q - 기타 명사 및 고유명사의 성 또는 이름
○ nb + n, u - 단위성, 비단위성 의존명사

따라서 [표 3-4]에서 정의한 3가지 패턴을 기반으로 명사만을 추출하며, 명사 추출 알고리즘은 [표 3-5]와 같다.

[표 3-5] 명사 추출 알고리즘

```
try {
        /* 쓰레드 모드 내 Workflow 알고리즘 수행 */
        workflow.activateWorkflow(true);

        /* Analysis using the workflow알고리즘을 사용한 분석 시작 */
        String document = "옛날 아귀 귀신이라는 큰 도적이 있었다.
                          그는 종종 이 세상에서 나와서 세상을 요란하게 하고…\n";
        workflow.analyze(document);
        str_test = workflow.getResultOfDocument();
        /* workflow 알고리즘 분석 결과 반환 */
        String li_str[] = str_test.split("\n");
        /* 정의한 명사 추출 시작 */
        for(int i=1; i < li_str.length; i++){
         if (li_str[i].contains("nc") || li_str[i].contains("nq") li_str[i].contains("nb")) {
                        System.out.println(li_str[i]);
                }
}
```

조선대학교 CHOSUN UNIVERSITY

[표 3-5]의 과정을 통해 추출한 명사 총 382개이며, [표 3-6]과 같다.

[표 3-6] 명사 패턴을 활용하여 추출된 명사

옛날, 아귀, 귀신, 도적, 그, 세상, 세상, 요란, 여자, 납치, 때, 아귀, 귀신, 임금님, 세, 공주, 납치, 임금님, 신하, 귀, 계획, 신통, 계책, 말, 사람, 사람, 무신, 나, 자신, 일, 임금님, 저, 집, 대, 국록, 저, 생명, 은혜, 공주, 님, 임금님, 이, 허락, 세, 공주, 중, 막내, 공주, 결혼, 무신, 하인, 아귀, 귀신, 소굴, 출발, 천하, 귀신, 소굴, 곳, 수, 하루, 산모퉁이, 피곤, 몸, 잡시, 동안, 잠, 꿈, 머리, 노인, 다음, 말, 나, 산, 산신령, 너, 아귀, 귀신, 소굴, 산, 쪽, 중, 산, 이상, 한, 바위
...

C. 서사정보 추출을 위한 TF 알고리즘 활용 방법

본 절에서는 설화 내 서사정보 목록을 추출하기 위해 [그림 3-2]의 서사정보 후보를 추출하는 알고리즘 활용 방법에 대해 기술한다. 서사정보의 후보 명사를 추출하는 방법으로는 TF(Term Frequency)를 활용한다. 이때 추출된 서사정보의 후보에서 해당 명사들 중 한가지만의 품사만을 갖지 않는 명사가 있기 때문에 불용어 명사 목록을 정의하여 입력된 설화 텍스트로부터 서사정보 목록을 추출한다.

1. TF(Term Frequency) 값을 이용한 서사정보 후보 추출

본 논문에서는 자연어 처리 분야에서 많이 사용되는 TF(Term Frequency)값을 활용하여 서사정보 후보를 추출하고자 한다. TF값은 일반적인 문서 내 출현하는 특정단어의 빈도수를 통해 해당 단어의 중요성을 판별하는데 많이 사용되는 기법으로 이를 사용하여 서사정보 데이터를 추출한다. 이러한 서사정보 추출을 위해 TF값을 사용하는 이유는 스토리의 핵심요소인 서사정보를 구성하고 있는 단어는 명사 품사를 가지고 있으며, 스토리의 주체가 되는 인물, 흐름을 나타내는 사건 및 배경은 타 명사보다 그 출현빈도가 크기 때문에 TF값을 활용한다.

[표 3-7]은 TF값 추출 알고리즘의 소스를 나타내고 있다. TF값을 추출하는 알고리즘에 대한 설명으로는 배열 내 들어있는 단어들끼리 비교하여 기존에 측정했던 단어일 경우에는 다음 단어로 넘어가서 재비교 후 새로운 단어일 경우만 Count값

조선대학교 CHOSUN UNIVERSITY

을 늘려주면서 단어에 대한 TF값을 계산하게 된다.

[표 3-7] TF값 추출 알고리즘

```
private Word[] getFrequentWords(String words[]){
    HashMap<String,Word> map = new HashMap<String,Word>();
      for(String s:words){ //map 배열에 들어있는 단어들 간 비교
          Word w = map.get(s);
            if(w==null)  w = new Word(s,1);
            else  w.count++;
            map.put(s, w);
        } //map 배열 내 단어가 있을 경우 count값을 1씩 증가
        Word[] list = map.values().toArray(new Word[]{});
        Arrays.sort(list); return list;
    }
    // 리스트 형태로 저장된 값을 배열로 전환
    public static String[] main(String[] args) {
        words = args;
        int i = 0;
        Word[] frequency = new tf().getFrequentWords(words);
        tf_word = new String[words.length];
        System.out.println(words.length);
        for(Word w:frequency){
                tf_word[i] = w.word+"="+w.count;
                i++;
        } return tf_word;          }
```

[표 3-7]에서 나타낸 TF값 계산 알고리즘을 통하여 추출해 낸 단어별 TF값은 다음 [표 3-8]과 같으며, 기존 382개의 명사 중 115개의 명사를 서사정보 후보로 추출하였다. 최종적으로 중복 처리된 명사를 바탕으로 재필터링 과정을 거치게 된다. 재필터링 하는 이유는 명사만을 추출하였더라도 해당 명사가 한가지의 품사만을 갖지 않는 단어가 존재하기 때문에 재필터링 과정을 수행하여 설화 내 서사정보를 추출하기 위함이다.

[표 3-8] 명사별 TF값 계산 결과

명사	TF값	명사	TF값	명사	TF값	명사	TF값
공주	17	아귀	7	술	4	광주리	3
무신	15	그	6	위	4	그것	3
도적	10	하인들	6	줄	4	나	3
귀신	9	산	5	구멍	4	다음	3
말	9	세상	5	노인	4	때	3
사람	9	수	5	막내	4	머리	3
임금님	9	일	5	바위	4	몸	3
공주들	8	저	5	집	4	물 …	3 …

조선대학교 CHOSUN UNIVERSITY

2. 불용어 명사 목록 정의

앞서 추출된 서사정보의 후보 명사 중에서 한 가지 품사만을 갖지 않는 명사들을 제거하는 과정을 위해 [표 3-9]와 같이 불용어 명사 목록을 정의한다.

[표 3-9] 불용어 명사 목록

번호	불용어 명사
1	가구
2	가닥
3	가락
4	가량
5	가리
6	가마
904	힝

추출된 명사와 함께 정의된 태그를 포함한 불용어 명사를 제거함으로써 원하는 서사정보 추출이 가능하게 된다. [표 3-10]는 명사에 정의된 불용어의 태그와 단어의 개수를 보여준다. 예를 들어 [표 3-8]에서 '그것'의 경우 명사 품사 이외에 지시대명사를 나타내는 태그인 'npd'의 품사를 지니므로 [표 3-10]의 정의된 태그를 이용하여 불용어 명사를 제거하는 과정이 필요하다.

[표 3-10] 명사에 정의된 불용어의 태그

태그명(단어개수)	설명	태그명(단어개수)	설명
xsnu(25)	명사 파생 접미사	xsms(2)	형용사 파생 접미사
xsnca(1)	명사 파생 접미사	xsmn(6)	형용사 파생 접미사
xsncc(26)	명사 파생 접미사	xsam(1)	부사 파생 접미사
xsna(8)	명사 파생 접미사	xsas(1)	부사 파생 접미사
xsns(2)	명사 파생 접미사	ii(208)	감탄사
xsnp(7)	명사 파생 접미사	nbn(70)	비단위성 의존명사
xsnx(14)	명사 파생 접미사	nbu(388)	단위성 의존명사
xsvv(7)	동사 파생 접미사	nbs(10)	비단위성 의존명사
xsva(5)	동사 파생 접미사	npp(78)	인칭대명사
xsvn(1)	동사 파생 접미사	npd(44)	지시대명사

조선대학교 CHOSUN UNIVERSITY

3. 불용어 태그를 이용한 TF별 명사 재추출

본 장에서는 [표 3-10]에서 정의된 불용어 태그를 바탕으로 TF값을 계산하여 추출된 단어들을 재분류함으로써 서사정보를 구성하는 명사에 대한 목록을 추출한다.

[표 3-11]은 TF별 명사 재추출을 위해 불용어 명사를 제거하기 위한 불용어 태그 정리 소스를 나타낸다.

[표 3-11] 불용어 명사를 제거하기 위한 불용어 태그 정리 소스

```
public static void main(String[] args) {
 try {
        outputStream = new FileWriter("C:/my.txt");
        BufferedReader br = new BufferedReader(
        new InputStreamReader( new FileInputStream(dir1), "utf-8"));
       while((var_text = br.readLine()) != null){
       if(var_text.contains("xsnu") || var_text.contains("xsnca") || var_te
       xt.contains("xsncc") || var_text.contains("xsna") || var_text.contai
       ns("xsns") || var_text.contains("xsnp") || var_text.contains("xsnx
       ") || var_text.contains("xsvv") || var_text.contains("xsva") || var_
       text.contains("xsvn") || var_text.contains("xsms") || var_text.cont
       ains("xsmn") || var_text.contains("xsam") || var_text.contains("xs
       as") || var_text.contains("ii")|| var_text.contains("nbn") || var_tex
       t.contains("nbu") || var_text.contains("nbs") || var_text.contains("
       npp") || var_text.contains("npd")) {

       outputStream.write(var_text);}
          br.close();
   } catch(Exception e2){
      JOptionPane.showMessageDialog(null, e2);
  }
 }
}
```

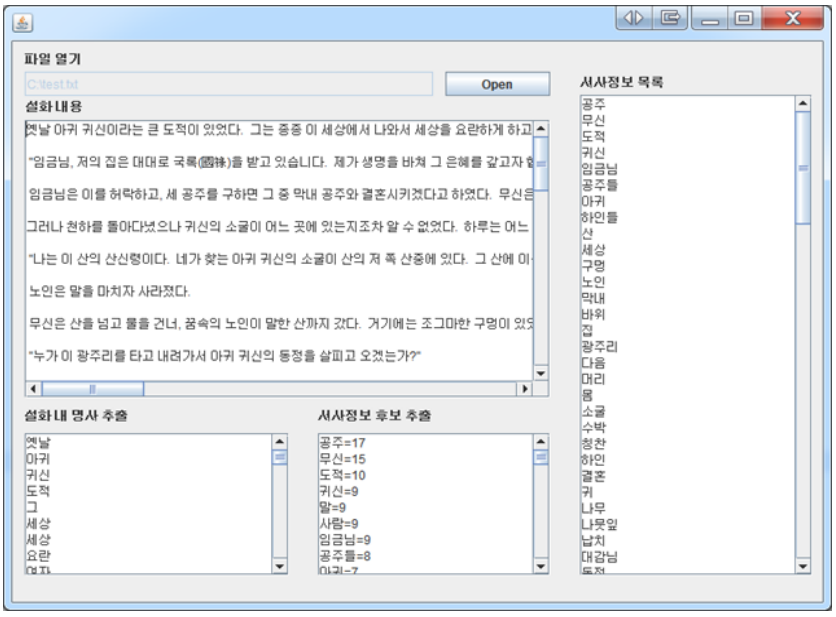
이와 같이 불용어 제거 단계를 통해 최종 추출된 서사정보 명사는 [표 3-12]와 같으며, 추출된 서사정보 명사는 총 41개의 명사를 확인할 수 있고, 단순 출현 명사를 제거하기 위해 TF값이 1이하인 데이터는 삭제하였다.

조선대학교 CHOSUN UNIVERSITY

[표 3-12] '지하국대적퇴치' 설화 내 서사정보 추출 결과

명사	TF	명사	TF	명사	TF
공주	17	집	4	대감님	2
무신	15	광주리	3	동정	2
도적	10	다음	3	땅	2
귀신	9	머리	3	마음	2
임금님	9	몸	3	목	2
공주들	8	소굴	3	물동이	2
아귀	7	수박	3	비늘	2
하인들	6	칭찬	3	소원	2
산	5	하인	3	아가씨	2
세상	5	결혼	2	옆구리	2
구멍	4	귀	2	의심	2
노인	4	나무	2	잠	2
막내	4	나뭇잎	2	치맛자락	2
바위	4	납치	2		

[그림 3-3]는 본 논문에서 제안한 방법을 통해 설화 텍스트 중 하나인 '지하국대적퇴치' 설화를 바탕으로 서사정보 자동 추출 시스템의 결과를 보여준다. 좌측의 입력된 설화 텍스트에 대한 내용을 기반으로 설화 내 명사를 추출하는 단계, 서사정보 후보를 추출하는 단계를 통하여 우측의 서사정보 목록을 나타낸다.



[그림 3-3] 추출된 서사정보 목록 결과

조선대학교 CHOSUN UNIVERSITY

[표 3-13]은 본 논문에서 구현한 서사정보 자동 추출 시스템의 처리 과정을 통해 얻은 결과를 나타내며 서사정보인 인물, 사건, 배경의 3요소로 '지하국대적퇴치' 설화 데이터를 분류한 결과를 나타낸다.

[표 3-13] '지하국대적퇴치' 설화 내 서사정보 분류 결과

인물	배경	사건	
공주	산	칭찬	애기
무신	세상	결혼	요란
도적	소굴	납치	은혜
귀신	공중	동정	응답
임금님	굴	소원	이야기
공주들	꿈	의심	조사
아귀	꿈속	계책	죽음
하인들	산모퉁이	계획	출발
노인	세계	대답	피곤
막내	옛날	명령	허락
하인	오랜만	방법	혼인
대감님	우물곁	병환	약속
아가씨	천하	아양	야단

이와 같이 추출된 서사정보 내 분류 결과를 통해 서사정보 자동 추출 시스템의 성능을 증명하기 위해 등장인물을 대상으로 4장에서 정확률을 측정하고자 한다.

조선대학교 CHOSUN UNIVERSITY

Ⅳ. 실험 및 평가

본 장에서는 설화 내 서사정보 자동 추출 시스템에 대한 성능 평가를 위해 추가적인 설화 데이터를 수집하여 실험 데이터를 구축한다. 성능을 평가하는 방법으로는 서사정보 자동 추출 시스템의 등장인물에 대한 정확률(Precision)을 측정하였다.

또한 실험 결과에 따른 응용방안으로 토픽맵 기반의 서사정보 대한 시각화 방법을 기술함으로써 설화 텍스트로부터 추출된 서사정보간의 관계를 통해 스토리텔링 저작 지원에 활용될 수 있음을 설명한다.

A. 실험 데이터

본 논문의 실험데이터로 3장에서 사용된 '지하국대적퇴치' 설화 이외에 추가적인 수집을 통해 총 6개의 설화 데이터를 수집한 후 각각의 설화에 나오는 등장인물을 분석하여 실험 데이터를 구축한다. [표 4-1]은 각각의 설화에 대해 등장인물과 수를 나타낸다.

[표 4-1] 실험데이터 정보

설화명	등장인물명(등장인물 수)
지하국대적퇴치 설화	아귀, 귀신, 도적, 여자, 임금님, 공주, 신하, 무신, 막내, 하인, 노인, 산신령, 아가씨, 사람, 문지기, 대감님, 우두머리 (17명)
거타지 설화	진성여왕, 막내아들, 사신, 양패, 거타지, 노인, 임금, 궁사, 중, 용, 그녀 (11명)
아기장수 우투리 설화	임금, 벼슬아치, 백성, 영웅, 부부, 아기, 우투리, 장수, 군사, 부모, 어머니 (11명)
지하국의 도적 설화	한량, 부자, 대적, 딸, 무사, 장수, 부모, 처녀 (8명)
노랑대구리 설화	부부, 아이, 부모, 노랑두대구리, 동네사람, 코샘생, 오종소티, 주인, 노인, 아들, 괴물, 삼형제 (12명)
황금 돼지 설화	선비, 딸, 사위, 부자, 처녀, 까치, 황금돼지, 괴물 (8명)

조선대학교 CHOSUN UNIVERSITY

B. 실험 평가 및 결과

1. 실험 평가 방법

본 논문의 실험은 구현한 설화 내 서사정보 자동 추출 시스템에 대한 성능을 평가하기 위해 등장인물 추출에 대한 정확률(Precision)을 이용하였으며 다음 식 (1)과 같다. 정확률은 정보 검색에서 검색 결과에 대한 정확성을 평가하는 측도이며 본 논문에서는 설화 내에서 등장하는 전체 인물의 수 중 구현된 시스템에 의해 추출된 등장인물의 수로 정확률을 측정한다.

$$\text{정확률}(\text{Precision}) = \frac{\text{구현된 시스템에 의해 추출된 등장인물 수}}{\text{설화 내 전체 등장인물 수}} \quad (1)$$

2. 실험 결과

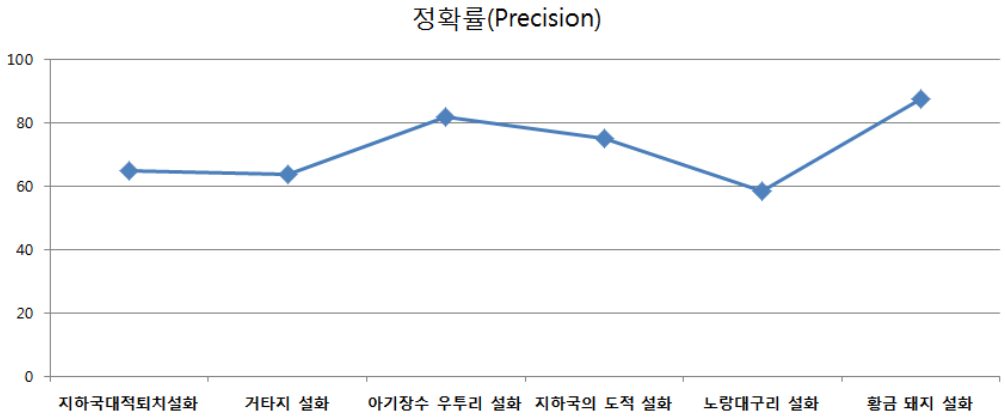
[표 4-2]는 식 (1)을 이용하여 구현된 시스템에 대한 성능을 평가한 결과를 보여준다.

[표 4-2] 서사정보 자동 추출 시스템의 등장인물 추출에 대한 성능

구분 \ 설화명	지하국대적 퇴치 설화	거타지 설화	아기장수 우투리 설화	지하국의 도적 설화	노랑대구리 설화	황금 돼지 설화
설화 내 전체 등장인물 수	17	11	11	8	12	8
시스템에 의해 추출된 등장인물 수	11	7	9	6	7	7
정확률	64.71%	63.64%	81.82%	75%	58.33%	87.5%
평균	71.83%					

조선대학교 CHOSUN UNIVERSITY

서사정보 자동 추출 시스템의 등장인물 추출에 대한 평균 정확률은 71.83%로 기대 이상의 성능을 보여준다. 이는 영어와는 달리 복잡한 형태소와 동음이의어를 가지는 한국어의 특성상 특정 품사를 추출하는데 어려움이 있는 반면, 각 설화별 정확률은 50% 이상으로 TF값 기반 등장인물 추출이 잘 되는 것을 보였다는 점은 매우 흥미로운 결과를 보여준다. [그림 4-1]은 [표 4-2]의 결과를 그래프로 표현하였다.



[그림 4-1] 각 설화별 등장인물 추출에 대한 정확률

각 설화별 등장인물 추출에 대한 정확률 그래프에서는 황금 돼지 설화에 대한 정확률이 87.5%로 가장 높은 정확률을 보여주었다. 이는 구현한 서사정보 자동 추출 시스템을 통해 주인공에 해당하는 등장인물들이 정확히 추출되었고, 해당 등장인물의 빈도수가 타 설화에 비해 월등히 높은 빈도를 보여주었음을 알 수 있다. 그러나 노랑대구리 설화에서는 58.33%의 가장 낮은 정확률을 보여준다. 이는 단어에 대한 의미 분석이 제대로 되지 않아 같은 등장인물을 나타내는 단어의 중복 검출에 대한 문제로 분석된다. 또한 정확한 비교분석을 위해 타 실험과 비교를 해야 하나, 현재까지 한국어 자연어 처리 기반의 스토리 내 서사정보 추출을 위한 연구가 활발히 진행되지 않고, 단지 등장인물의 속성이나 성격 파악 위주의 연구가 진행되다보니 정확한 비교 실험이 어려웠다. 또한, 설화 자체의 데이터양이 대부분 500단어 내외로 매우 적다. 따라서 데이터양에 비례함에 따라 등장인물 추출에 대한 정확률이 설화에 따라 상이함을 실험 결과로써 나타났다.

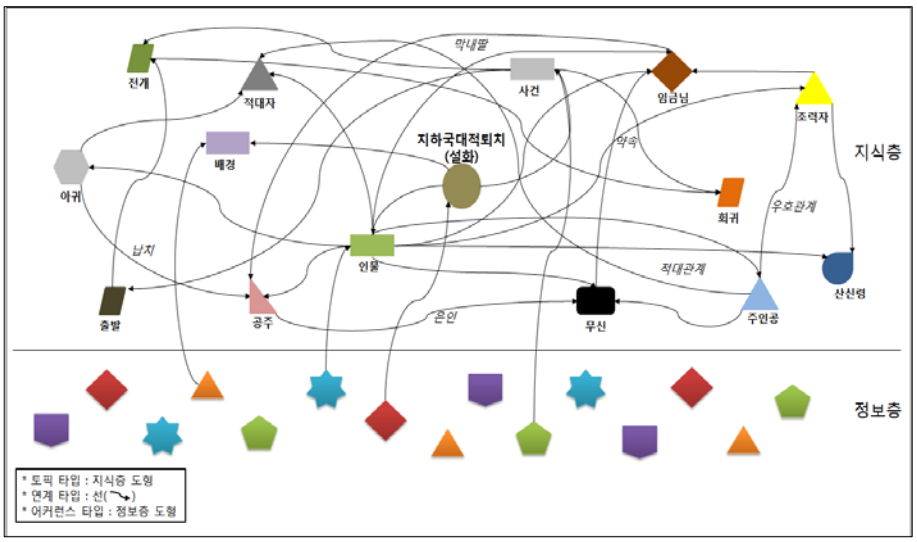
조선대학교 CHOSUN UNIVERSITY

C. 실험 결과에 따른 토픽맵 기반 시각화 응용방안

본 절에서는 설화 텍스트로부터 추출된 서사정보를 중심으로 인물간의 관계 및 사건, 배경에 대한 정보를 토픽맵으로 표현하고자 한다. 이를 위해 영웅 서사구조 기반의 관계 모델을 활용하여 토픽맵으로 구축하였으며, 설화 텍스트 정보를 시각화함으로써 스토리텔링 저작 지원을 위한 효율적인 정보를 제공하고자 그 응용방안을 제시하고자 한다.

1. 서사정보를 이용한 토픽맵 구축

본 논문에서 추출된 서사정보에 대한 시각화를 표현하기 위해 [그림 4-2]와 같은 개념 모델링을 도출하였다. 지식층에는 인물, 사건, 배경 토픽 타입이 존재하며, 인물 토픽 타입 중심으로 주인공, 조력자, 적대자 토픽간의 관계가 설정되어 있고, 사건 토픽 타입 중심으로 영웅 서사구조의 사건정보인 출발, 전개, 귀환 토픽간의 관계가 설정되어 있다. 또한, 배경 토픽 타입 중심으로 현실세계, 초현실세계 토픽간의 관계가 설정되어 있으며, 추출된 서사정보 인스턴스는 각 토픽에 따라 연계 타입을 갖고 있다.



[그림 4-2] '지하국대적퇴치' 설화 개념 모델링

조선대학교 CHOSUN UNIVERSITY

이러한 토픽은 각각의 정보층에 있는 이미지 파일, URL, 특정 데이터베이스 내 레코드를 참조하며 토픽 및 인스턴스에 대한 부가적인 정보를 확장할 수 있다. 예를 들어 인물에 대한 구체적인 속성인 이름, 나이, 성별, 인물 이미지 등의 정보를 인물 데이터베이스 및 웹 상에 존재하는 해당 정보를 참조함으로써 구체적인 설화 데이터에 대한 시각화를 표현할 수 있다.

[표 4-3]은 개념 모델링을 토대로 토픽맵 구현을 위한 구성요소를 도출하였다. 토픽 타입으로는 인물을 포함하는 요소로써 주인공, 조력자, 적대자등과 같은 토픽으로 구성요소를 도출하였으며 인물간의 관계에 따라 적대관계, 조력관계 등의 연계타입을 도출하였다. 또한, 정보층으로부터 설화에 대한 줄거리 및 인물 이미지, 배경 이미지 등의 외부 리소스를 어커런스 타입으로 도출하였다.

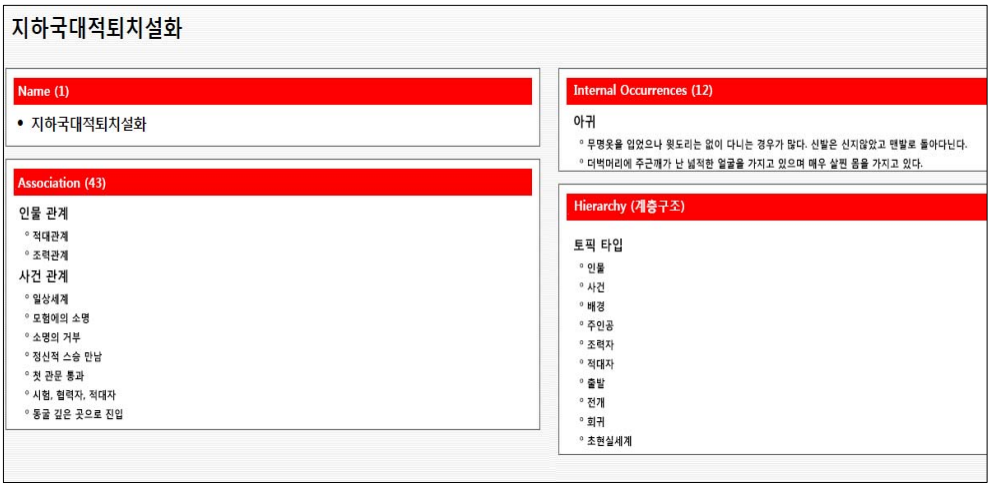
[표 4-3] 설화 데이터 시각화를 위한 구성요소 도출

토픽타입	인스턴스	연계타입	인스턴스
인물	무신	적대관계	주인공 - 적대자
사건	하인	조력관계	주인공 - 조력자
배경	아귀	일상세계	출발 - 일상세계
주인공	산신령	모험에의 소명	출발 - 모험에의 소명
조력자	문지기	소명의 거부	출발 - 정신적 스승 만남
적대자	공주	정신적 스승 만남	출발 - 첫 관문의 통과
출발	산	첫 관문 통과	전개 - 시험, 협력자, 적대자
전개	산모퉁이	시험, 협력자, 적대자	전개 - 시련
회귀	바위	동굴 깊은 곳으로 진입	전개 - 보상
현실세계	땅 속	시련	회귀 - 귀환의 길
초현실세계	…	보상	회귀 - 부활
		귀환의 길	회귀 - 영약을 가지고 귀환
		영약을 가지고 귀환	…

조선대학교
CHOSUN UNIVERSITY

본 논문에서는 설화 텍스트에 대한 시각화를 표현함으로써 실험 결과에 따른 응용방안 제시를 위해 최종적으로 토픽맵 편집기 도구인 온토피아에서 개발한 옴니게이터(Omnigator)도구[24]를 이용하여 토픽, 어소시에이션, 어커런스에 대한 입력 및 편집함으로써 디자인을 구현하였다.

[그림 4-3]은 구현된 설화 데이터 기반의 토픽맵을 나타내며 [표 4-3]의 설화 데이터 구성요소를 기반으로 구축하였다. 설화 데이터의 토픽 타입인 인물, 사건, 배경과 주인공, 조력자, 적대자, 출발, 전개, 귀환, 현실세계, 초현실세계의 토픽은 계층적 구조를 가지고 있으며, 인물 관계 및 배경 관계, 사건 관계를 통하여 각 요소들 간의 의미적 관계를 형성하고 있다. 또한 인물의 구체적인 속성을 표현하기 위해 각 인물에 따른 내부 및 외부 어커런스 참조를 지정하여 시각화 된 설화 데이터를 표현할 수 있다. 예를 들어 '지하국대적퇴치' 설화에서 아귀라는 인물은 단지 주인공의 적대자로써 스토리의 사건을 전개하지만 단지 표면적인 텍스트상의 정보뿐만 아니라 스토리의 새로운 창작에 대한 발상을 위해 아귀귀신에 대한 구체적인 설명을 어커런스를 통해 표현할 수 있도록 하며, 이러한 설화 데이터의 시각화를 토픽맵과 접목하여 스토리텔링 저작 지원을 위한 수단으로 표현할 수 있다.

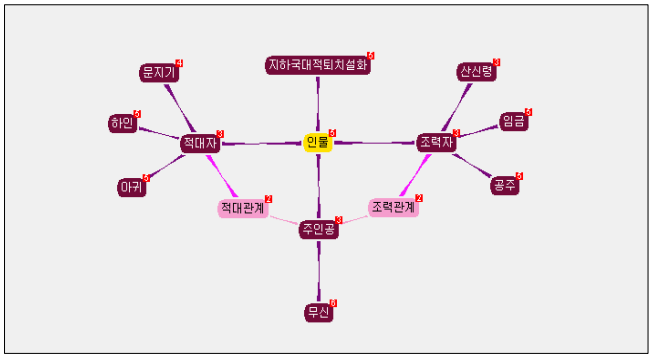


[그림 4-3] 구현된 토픽맵 기반의 설화 데이터

조선대학교 CHOSUN UNIVERSITY

2. 토픽 및 관계 생성

구현된 서사정보 자동 추출 시스템의 서사정보 추출 결과와 [표 4-3]의 시각화 구성요소를 기반으로 스토리텔링 서사구조 모델의 영웅 서사구조 관계모델을 활용하여 실질적인 토픽간의 관계설정을 기술한다. 인물 데이터는 인물의 유형에 따라 적대관계, 조력관계로 관계성을 구축하였으며, 인물과 사건의 관계는 출발, 전개, 귀환 단계의 하위구조에서 인물이 등장하는 사건과 관계성을 구축하였다. 또한, 인물과 사건이 존재하는 공간적 장소인 배경과 관계성을 구축함으로써 설화 텍스트에 대한 서사정보를 시각화 할 수 있었다. [그림 4-4], [그림 4-5], [그림 4-6]는 영웅 서사구조에서 정의한 관계 모델을 바탕으로 '지하국대적퇴치' 설화에서 추출된 데이터를 바탕으로 관계를 생성한 그림이다.



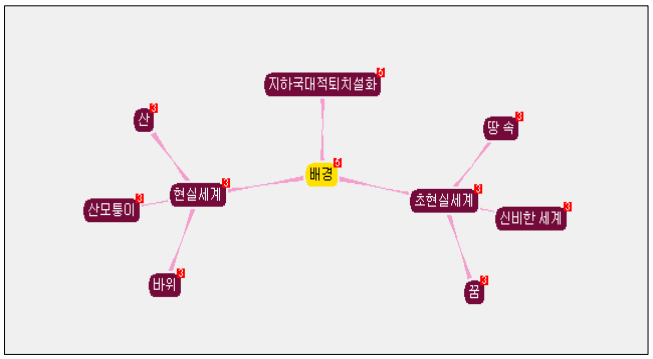
[그림 4-4] 인물 관계 생성

[그림 4-4]는 토픽맵을 이용하여 인물 관계를 표현한 그림이다. 인물 토픽 타입에는 인물의 유형에 따라 주인공, 조력자, 적대자 토픽이 존재하며 주인공을 중심으로 인물간의 관계를 적대관계, 조력관계로 설정 하였다, 주인공 토픽에는 지하국대적퇴치 설화에서 주인공으로 분류된 무신 인스턴스가 존재하고, 적대자 토픽에는 문지기, 하인, 아귀 인스턴스가 존재한다. 또한, 조력자 토픽에는 산신령, 임금, 공주 인스턴스를 갖는다.

배경 토픽 타입은 [그림 4-5]와 같이 현실세계 토픽과 초현실세계 토픽으로 생성하였다. 현실세계 토픽으로는 산, 산모퉁이, 바위 등과 같이 설화 내 등장하는 현실공간에 대한 인스턴스를 갖는다. 초현실세계 토픽으로는 꿈, 아귀소굴, 땅 속등과

조선대학교 CHOSUN UNIVERSITY

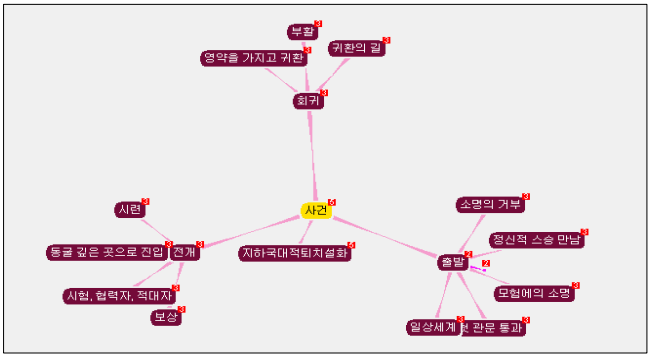
같이 설화에서 등장하는 가상적인 공간으로써 주인공이 초현실세계로 여행을 하면서 인물들이 존재하는 공간적인 요소에 대한 인스턴스를 갖는다.



[그림 4-5] 배경 관계 생성

사건 토픽 타입은 크게 출발, 전개, 귀환의 토픽으로 이루어져 있다. 본 논문에서 대상으로 삼은 '지하국대적퇴치' 설화뿐만 아니라 모든 영웅 서사구조 기반의 설화에 대한 스토리의 구조적 접근이 가능하기 위함이다. 먼저, 출발 토픽은 일상세계, 모험에의 소명, 소명의 거부, 정신적 스승과 만남, 첫 관문 통과의 토픽을 가지고 있으며, 각 단계마다 등장하는 인물과 인물이 존재하는 공간적 요소인 배경 토픽의 인스턴스와 관계가 설정되어 있다.

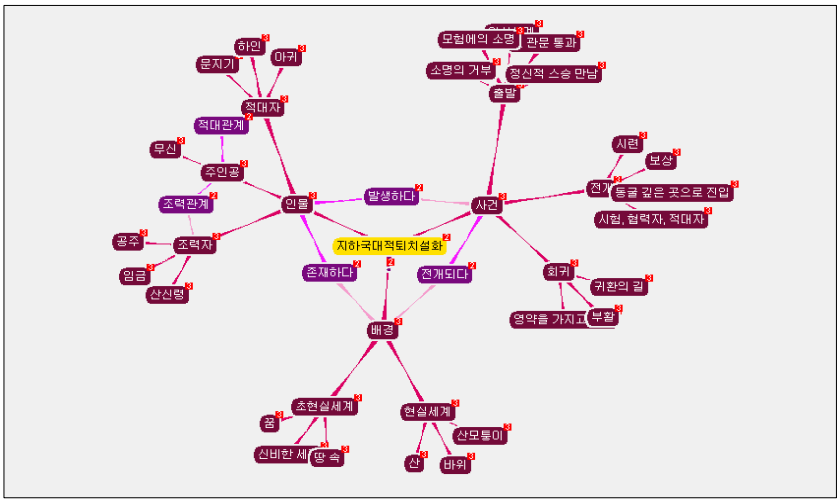
또한, 전개 토픽 타입은 시련, 동굴 깊은 곳으로 진입, 보상 토픽을 가지고 있으며, 귀환 토픽 타입은 귀환의 길, 부활, 영약을 가지고 귀환의 토픽을 갖으며 인물 및 배경 인스턴스를 모두 포함하고 있다.



[그림 4-6] 사건 관계 생성

조선대학교 CHOSUN UNIVERSITY

또한, 최종적으로 서사정보간의 관계를 생성함으로써 스토리 내 핵심요소인 인물, 사건, 배경간의 유기적인 결합을 위한 연계타입(Association)을 정의하도록 한다. 먼저 인물과 사건 토픽 타입의 연계타입을 '발생하다', 사건과 배경의 연계타입은 '전개하다', 인물과 배경의 연계타입을 '존재하다'로 관계성을 구축하였다.



[그림 4-7] 서사정보 관계 생성

이러한 토픽맵 기반의 관계설정을 이용한 시각화는 설화 텍스트로부터 추출된 스토리 내 서사정보 데이터를 가공할 수 있으며, 스토리를 구성하는 각 요소(인물, 사건, 배경)간의 의미적 관계를 생성할 수 있다. 이에 스토리 저작을 위한 작가들에게 보다 효율적인 스토리텔링 저작에 대한 응용방안으로 활용할 수 있음을 나타내고 있다.

조선대학교 CHOSUN UNIVERSITY

Ⅴ. 결론 및 제언

최근 문화콘텐츠에 대한 관심이 증가됨으로써 영화, 공연 관광, 드라마 등 다양한 문화산업에 걸쳐 스토리텔링 기법을 활용하여 부가가치 창출을 위한 노력이 지속적으로 증대되고 있다. 이에 양질의 스토리텔링 콘텐츠 생산을 위해 스토리텔링 저작을 위한 연구가 국외를 중심으로 수행되고 있는 반면 국내에서는 상대적으로 미흡한 실정이다.

이에 본 연구에서는 스토리텔링 저작을 위해 스토리 내 서사정보를 자동적으로 추출하기 위한 시스템을 설계 및 구현하였다. 제안하는 서사정보 자동 추출 시스템은 한국어로 작성된 설화 텍스트 내 명사를 추출하는 단계, 서사정보 후보를 추출하는 단계, 서사정보 목록을 제공하는 단계로 구성하였다. 본 시스템에서 설화 텍스트 내 명사를 추출하는 이유는 서사정보 요소인 인물, 사건, 배경에 대한 품사가 주로 명사에 분포되어 있기 때문이며, 입력된 설화 텍스트에 대한 품사를 파악하기 위해 POS Tagging 과정을 수행하였다. 또한, 기존 형태소 분석기에서 정의하고 있는 명사 형태소에 대한 동일 패턴을 정의하며, TF(Term Frequency) 알고리즘을 이용한 서사정보 후보를 추출하였다. 추출된 서사정보 후보의 품사가 지닌 모호성을 해소하기 위해 불용어 명사 목록을 정의하여 추출된 명사들에 대한 재필터링 과정을 수행함으로써 재분류된 서사정보를 추출할 수 있었다.

구현된 서사정보 자동 추출 시스템의 성능을 평가하기 위해 서사정보 중 등장인물 추출에 대한 정확률을 측정하였다. 실험 데이터는 본문에서 설명한 '지하국대적퇴치' 설화를 포함한 총 6개의 설화 데이터로 진행하였으며 71.83%의 평균 정확률을 보여주었다.

향후 연구로는 자연어 처리 기법의 텍스트 분석을 통한 단어 추출뿐만 아니라 스토리 내 서사정보간의 유기적인 결합을 위해 의미적 관계 생성을 위한 온톨로지 기반의 연구가 필요할 것으로 판단된다. 또한, 추출된 서사정보 데이터에 대한 시각화를 통해 스토리텔링 저작을 위한 작가들에게 창작을 위한 자동화 시스템을 활용함으로써 스토리 구성을 위한 시간적, 인력적 부담을 최소화할 것으로 기대된다.

조선대학교 CHOSUN UNIVERSITY

참고문헌

[1] 나성준, 김우중, 최이권, 전현택, "디지털 스토리텔링의 스토리 핵심 요소를 이용한 기획 지원 온톨로지 검색 시스템 연구", 한국지능정보시스템학회, pp.338-242, 2010.

[2] 안성혜, 허영, "비주얼 스토리텔링의 재미요소 설계", 한국콘텐츠학회 춘계학술대회논문집, 제7권 제1호(상), pp.250-257, 2009.

[3] 문화기술 (CT) 심층리포트, "스토리텔링 저작 도구 연구 동향과 사례분석", 2012.

[4] 권호창, "디지털 스토리텔링에 있어 스토리 생성 시스템 연구", 한국예술종합학교 석사학위논문, 2009.

[5] 라창현, 하얀, "지능형 스토리텔링을 위한 저작에 관한 연구", 한국지능시스템학회 학술발표 논문집, pp.456-459, 2005.

[6] 김석규, 문성현, 박준, 장준호, 한상영, "이야기 생성을 위한 인터랙티브 스토리텔링 스크립트 언어에 관한 연구", 멀티미디어학회논문지 제12권 제2호, pp313-322, 2009.

[7] 박승보, 이예지, 유은순, 정재은 "등장인물 누적에 따른 스토리 발단 부분의 경계 판단", 한국지능정보시스템학회 추계학술대회, pp135-140, 2011.

[8] J.Cambell, The Hero with a Thousand Face, Bollingen, 1972.

[9] 구민호, 민덕기, "반복적 추출 단어 분리를 이용한 핵심 키워드 추출 기법에 관한 연구", 한국정보과학회 가을학술발표집 제36권 제2호(B), pp.329-334, 2009.

[10] 백남주, 고병규, 이은지, 김판구, "연관 검색어 추출을 위한 Triple-Data Pattren 정의에 관한 연구", 한국스마트미디어학회 춘계학술대회, 2014.

[11] http://kkma.snu.ac.kr/

[12] 최주호, 이재영, "전자기록물의 메타데이터 추출 및 비교 검증 기술 연구", 한국기록관리학회지, 제12권 제1호, pp.7-32, 2012.

[13] http://kldp.net/projects/hannanum/

[14] 유정훈, "워드넷 기반의 클러스터를 이용한 문서 요약 시스템 : 스포츠 뉴스 기사 헤드라인 추출", 단국대학교 석사학위논문, 2013.

[15] 오삼균, "A Comparison of Ontology Language : Focusing on W3C OWL and ISO Topic Maps"

[16] Ontopia.net "The TAO of Topic Maps." http://www.ontopia.net/topicmaps/mat

조선대학교 CHOSUN UNIVERSITY

erials/ta.html.

[17] 이재호, “시맨틱 웹의 온톨로지 언어”, 정보과학회지, 제21권 제3호, pp.18-27, 2003.

[18] Pepper, Steve and Garshol, Lars Marius, The XML Papers : Learned on Applying Topic Maps, 2002.

[19] 박여삼, 장옥배, 한성국, "X-TOP : 레거시 시스템상에서 온톨로지 구축을 위한 토픽맵 플랫폼의 설계와 구현", 정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 14권 제2호, pp.130-142, 2008.

[20] Jack Park and Sam Hunting, XML Topic Maps : Creating and Using Topic Maps for the Web, Addision-Wesley, 2003.

[20] 오삼균, 박옥남, “토픽맵 기반 판소리 검색시스템 구축 및 평가에 관한 연구”, 한국도서관·정보학회지, 제36권 제4호, pp.77-98, 2005.

[21] 권진만, 정명범, 성보경, 김정수, 고일주, “토픽맵기반의 뮤지션 검색시스템 구축”, 한국HCI학회 학술대회, pp.760-765, 2008.

[22] 이지수, 정겨운, 이경원, “스토리텔링의 시각화를 위한 개념적 맵들의 특성분석”, 한국HCI학회, pp1214-1219, 2008.

[24] http://www.ontopia.net/

조선대학교 CHOSUN UNIVERSITY

creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조선대학교
CHOSUN UNIVERSITY

2015년 8월
석사학위 논문

스토리텔링 저작을 위한 서사정보 추출 시스템 설계 및 구현

조선대학교 산업기술융합대학원

소프트웨어융합공학과

백 남 주

조선대학교
CHOSUN UNIVERSITY

스토리텔링 저작을 위한 서사정보 추출 시스템 설계 및 구현

Design and Implementation of Narrative Information Extraction System for Storytelling Authoring

2015년 8월 25일

조선대학교 산업기술융합대학원

소프트웨어융합공학과

백 남 주

조선대학교
CHOSUN UNIVERSITY

스토리텔링 저작을 위한 서사정보 추출 시스템 설계 및 구현

지도교수 김 판 구

이 논문을 공학석사학위신청 논문으로 제출함.

2015년 4월

조선대학교 산업기술융합대학원

소프트웨어융합공학과

백 남 주

조선대학교
CHOSUN UNIVERSITY

백남주의 석사학위논문을 인준함

위원장 조선대학교 교수 (인)

위 원 조선대학교 교수 (인)

위 원 조선대학교 교수 (인)

2015년 5월

조선대학교 산업기술융합대학원

조선대학교 CHOSUN UNIVERSITY

목 차

ABSTRACT

Ⅰ. 서론 ····· 1
A. 연구 배경 및 목적 ····· 1
B. 논문의 구성 ····· 2

Ⅱ. 관련 연구 ····· 3
A. 스토리텔링 저작 지원 기술 현황 ····· 3
1. 스토리텔링 저작을 위한 지능형 스토리 생성 기술 ····· 3
2. 스토리텔링 저작 지원 도구 ····· 5
3. 스토리텔링을 위한 서사구조 모델 ····· 9
B. 설화 데이터 분석을 위한 자연어 처리 시스템 ····· 13
1. 꼬꼬마 형태소 분석기 ····· 13
2. 한나눔 형태소 분석기 ····· 14
C. 토픽맵 ····· 17
1. 토픽맵 구조 및 구성요소 ····· 17
2. 토픽맵을 이용한 정보 시각화 ····· 20

Ⅲ. 서사정보 자동 추출 시스템 설계 및 구현 ····· 22
A. 서사정보 자동 추출 시스템 설계 및 인터페이스 정의 ····· 22
1. 시스템 구성도 ····· 22
2. 개발 환경 ····· 23
3. 사용자 인터페이스(User Interface) ····· 23
B. 서사정보 추출을 위한 전처리 및 명사 추출 방법 ····· 24
1. 특수문자 제거 및 POS(Part-Of-Speech) Tagging ····· 24
2. 서사정보 추출을 위한 명사 패턴 정의 및 추출 ····· 26
C. 서사정보 추출을 위한 TF 알고리즘 활용 방법 ····· 27
1. TF(Term Frequency) 값을 이용한 서사정보 후보 추출 ····· 27
2. 불용어 명사 목록 정의 ····· 29
3. 불용어 태그를 이용한 TF별 명사 재추출 ····· 30

조선대학교 CHOSUN UNIVERSITY

Ⅳ. 실험 및 평가 ····· 33
A. 실험 데이터 ····· 33
B. 실험 평가 및 결과 ····· 34
1. 실험 평가 방법 ····· 34
2. 실험 결과 ····· 34
C. 실험 결과에 따른 토픽맵 기반 시각화 응용방안 ····· 36
1. 서사정보를 이용한 토픽맵 구축 ····· 36
2. 토픽 및 관계 생성 ····· 39

Ⅴ. 결론 및 제언 ····· 42

참고문헌 ····· 43

조선대학교 CHOSUN UNIVERSITY

표 목 차

[표 2-1] 스토리텔링 저작 도구 개발에 요구되는 기술적 요소 ····· 6
[표 2-2] 서사구조의 단계별 내용 ····· 10
[표 2-3] 인물정보 분류 ····· 11
[표 2-4] 사건정보 분류 ····· 12
[표 2-5] 배경정보 분류 ····· 12
[표 2-6] 꼬꼬마 형태소 분석기의 태그 집합 예 ····· 14
[표 2-7] 분석 후보 생성 ····· 15
[표 2-8] 미등록어 처리기 ····· 15
[표 2-9] 한나눔 형태소 분석기의 태그 집합 예 ····· 16
[표 3-1] 서사정보 자동 추출 시스템 개발 환경 ····· 23
[표 3-2] 설화 텍스트 내 특수문자 제거 소스 ····· 24
[표 3-3] 설화 텍스트에 대한 POS Tagging 예 ····· 25
[표 3-4] 한나눔 형태소 분석기 내 정의된 명사 패턴 ····· 26
[표 3-5] 명사 추출 알고리즘 ····· 26
[표 3-6] 명사 패턴을 활용하여 추출된 명사 ····· 27
[표 3-7] TF값 추출 알고리즘 ····· 28
[표 3-8] 명사별 TF값 계산 결과 ····· 28
[표 3-9] 불용어 명사 목록 ····· 29
[표 3-10] 명사에 정의된 불용어의 태그 ····· 29
[표 3-11] 불용어 명사를 제거하기 위한 불용어 태그 정리 소스 ····· 30
[표 3-12] '지하국대적퇴치' 설화 내 서사정보 추출 결과 ····· 31
[표 3-13] '지하국대적퇴치' 설화 내 서사정보 분류 결과 ····· 31
[표 4-1] 실험데이터 정보 ····· 33
[표 4-2] 서사정보 자동 추출 시스템의 등장인물 추출에 대한 성능 ····· 34
[표 4-3] 설화 데이터 시각화를 위한 구성요소 도출 ····· 37

조선대학교 CHOSUN UNIVERSITY

그 림 목 차

[그림 2-1] 스토리 생성 시스템 전체 모델 ····· 3
[그림 2-2] 지능형 관리 모듈 ····· 4
[그림 2-3] 인터랙티브 스토리텔링 시스템 ····· 5
[그림 2-4] 드라마티카 프로의 초기 화면 ····· 7
[그림 2-5] 파이널 드래프트의 실행 화면 ····· 8
[그림 2-6] 스토리 헬퍼의 실행 화면 ····· 9
[그림 2-7] 한나눔 형태소 분석기의 워크플로우 ····· 15
[그림 2-8] 토픽 맵 구조 ····· 18
[그림 2-9] 토픽 ····· 18
[그림 2-10] 토픽 타입 ····· 19
[그림 2-11] 어커런스 ····· 19
[그림 2-12] 어소시에이션 ····· 20
[그림 2-13] 판소리 토픽맵 ····· 20
[그림 2-14] 이상한 나라의 앨리스 플롯 중심 토픽맵 ····· 21
[그림 3-1] 전체 시스템 구성도 ····· 22
[그림 3-2] 사용자 인터페이스 ····· 23
[그림 3-3] 추출된 서사정보 목록 결과 ····· 31
[그림 4-1] 각 설화별 등장인물 추출에 대한 정확률 ····· 35
[그림 4-2] '지하국대적퇴치' 설화 개념 모델링 ····· 36
[그림 4-3] 구현된 토픽맵 기반 설화 데이터 ····· 38
[그림 4-4] 인물 관계 생성 ····· 33
[그림 4-5] 배경 관계 생성 ····· 40
[그림 4-6] 사건 관계 생성 ····· 40
[그림 4-7] 서사정보 관계 생성 ····· 41

조선대학교 CHOSUN UNIVERSITY

ABSTRACT

Design and Implementation of Narrative Information Extraction System for Storytelling Authoring

NamJu Baek
Advisor : Prof. Pankoo Kim, Ph.D
Department of Software Convergence Engineering
Graduate School of Chosun University

Recently, the perceived value of stories has risen, which has created a greater focus on storytelling. Storytelling is method used for immersing an audience and prompting interest in the development of plot. This method has been receiving a lot of attention recently as an important technique in the cultural industry.

More specifically, there have many more cases of stories being created using authoring tool systems through Human-Computer Interaction (HCI). Compared to other countries, there have not been many studies regarding using such systems for storytelling in Korea.

This thesis proposes a system which automatically extracts narrative information, including key storytelling elements, that comprises the main components of a written story. Patterns for extracting narrative information nouns within the text are defined by applying a Korean natural language process technique to story data collection and analysis. Story data is then extracted using term frequency(TF).

조선대학교
CHOSUN UNIVERSITY

To assess this narrative information extraction system, experiments were done to analyze data from six stories. The results determined using the automated character extraction method. The results showed that approximately 71.83% of the average precision ratio and performance had been confirmed.

조선대학교 CHOSUN UNIVERSITY

Ⅰ. 서론

A. 연구 배경 및 목적

최근 스토리가 지닌 무한의 부가가치가 새로운 창출의 핵심적인 요인으로 떠오르면서 청중을 몰입시키고 내용전개에 흥미를 유발시키게 만드는 기법인 스토리텔링이 문화산업에 있어서 중요한 기술로 주목을 받고 있다[1]. 스토리텔링이란 스토리(Story)와 텔링(Telling)의 합성어로 상대방에게 알리고자 하는 바를 재미있고 생생한 이야기로 설득력 있게 전달하고자 하는 행위라고 할 수 있다[2]. 특히 인간과 컴퓨터간의 상호작용(HCI, Human-Computer Interaction)을 통한 스토리텔링 저작 도구 시스템을 활용하여 스토리를 창작하는 사례가 늘어나고 있다. 스토리텔링 저작 도구는 스토리를 창작하는데 있어서 사용되는 컴퓨터 소프트웨어를 가리키지만 단지 컴퓨터 프로그래밍 기술에 의해 구현된 것이 아니라 인문학적 이론인 서사학과 창작 이론의 결합을 통한 스토리 구조 분석 기반의 종합적인 연구에 의해 구현된다[3]. 주로 국외의 대표적인 스토리텔링 저작 도구로는 드라마티카 프로(Dramatica Pro)와 파이널 드래프트(Final Draft) 등이 존재하는 반면, 국내에서는 최근 스토리 헬퍼(Story Helper)가 개발된 사례 이외의 연구는 미흡한 실정이다.

이에 본 논문에서는 효율적인 스토리텔링 저작 지원을 위해 스토리 서사정보를 자동으로 추출하는 시스템을 설계 및 구현하고자 한다. 이를 위해 서사정보의 구성요소인 인물, 사건, 배경에 대한 단어 품사가 주로 명사에 분포됨을 이용한다. 먼저 한글로 작성된 설화 텍스트로부터 명사 추출을 위해 입력받은 텍스트 내 존재하는 특수문자를 제거하고 단어 별 품사에 대한 태깅(Tagging) 과정을 거친 후, 형태소 분석기 내 정의된 명사 태그(Tag)를 이용하여 명사 추출을 위한 패턴(Pattern)을 정의한다. 추출된 설화 내 명사로부터 TF(Term Frequency) 알고리즘을 이용하여 명사 내 서사정보 후보를 추출하며, 보다 정확한 서사정보 추출을 위해 불용어(Stopword) 명사 목록 정의하여 재필터링 된 서사정보를 추출한다. 또한, 추출된 서사정보를 중심으로 설화 데이터에 대한 시각화(Visualization)를 위해 토픽맵(Topic Maps) 기반의 응용방안을 제시하고자 한다.

조선대학교 CHOSUN UNIVERSITY

B. 논문의 구성

스토리텔링 저작을 위한 서사정보 자동 추출 시스템을 설계 및 구현하기 위해 본 논문은 다음과 같은 구성으로 작성되었다.

1장 서론에서는 연구의 배경 및 목적에 대해 간략하게 기술하고, 2장 관련연구에서는 스토리텔링 저작 지원 기술 현황과 설화 데이터 분석을 위한 자연어 처리 시스템에 대해 설명하며, 토픽맵을 이용한 정보 시각화에 대해 기술한다.

3장에서는 서사정보 자동 추출 시스템을 구현하기 위해 설화 텍스트로부터 서사정보 추출을 위한 전처리 과정과 추출된 명사로부터 TF 알고리즘 및 불용어 명사 목록을 활용한 서사정보 추출 방법에 대해 기술한다.

4장에서는 본 논문에서 제안한 서사정보 추출 시스템에 대한 성능 평가를 수행하며 추출된 서사정보 데이터를 중심으로 토픽맵 기반의 시각화 응용방안을 제시한다.

마지막으로 5장에서는 본 연구에 대한 전체적인 결과를 요약하고, 향후 연구 방향을 제시한다.

조선대학교 CHOSUN UNIVERSITY

Ⅱ. 관련 연구

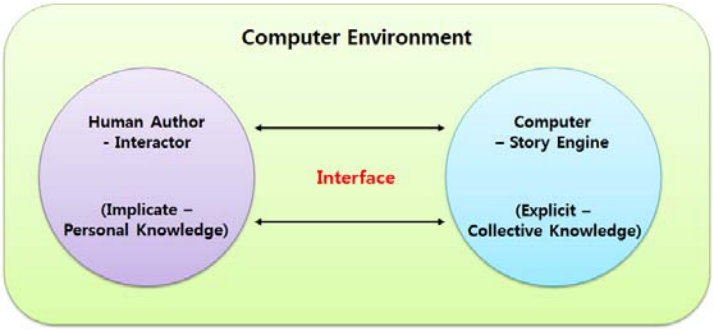
A. 스토리텔링 저작 지원 기술 현황

본 절에서는 스토리텔링 저작 지원 기술의 현황을 알아보기 위해 스토리텔링 저작을 위한 지능형 스토리 생성 기술과 스토리텔링 저작 지원 도구의 기술적 요소 및 사례에 대해 살펴보고 그 한계점에 대해 기술한다.

1. 스토리텔링 저작을 위한 지능형 스토리 생성 기술

지능형 스토리 생성 시스템 연구는 완성도 높은 스토리를 생성하기 위해 디지털 스토리텔링의 특성을 이용하여 스토리에 대한 서사(Narrative)성을 갖추며, 창작을 위한 보조수단으로써 어떻게 컴퓨터와 인간이 상호작용하는지에 대한 연구를 말한다. 여기서 스토리는 다중매체성, 네트워크성, 상호작용성, 비선형성 등으로 특징지어지는 새로운 스토리가 아닌 기존의 문자, 영상매체를 통해 만들어져 온 서사성을 갖춘 전통적이고 고적적인 스토리를 뜻한다.

특히, 지능형 스토리 생성 기술은 디지털 스토리텔링의 특성을 이용하여 최종적으로 만들어진 스토리에 대한 창작물이 아닌 창작의 과정 속에서 작가들의 보조수단으로써 활용을 위한 기술 연구라고 할 수 있다[4].



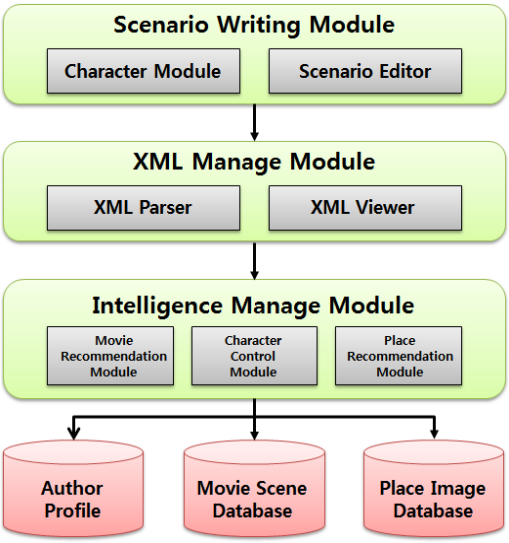
[그림 2-1] 스토리 생성 시스템 전체 모델[4]

조선대학교 CHOSUN UNIVERSITY

창작자는 이러한 기술이 바탕 되어지는 스토리 생성 시스템을 이용하여 창의적이고 다양한 스토리를 창작할 수 있다. [그림 2-1]은 스토리텔링 저작을 위해 인간과 컴퓨터간의 상호작용 (HCI, Human Computer Interaction)기술을 바탕으로 스토리 생성 시스템의 전체모델을 보여주고 있다. 이는 기존의 스토리 창작 소프트웨어인 워드프로세서 소프트웨어의 확장형과는 다른 스토리텔링 방식을 보여주고 있다.

스토리 생성 시스템은 컴퓨터가 스토리에 대한 정보에 대한 체계적이고 집합적인 지식(Domain Knowledge)을 갖추고 있으며 창작자는 컴퓨터와 함께 스토리 창작과정에 참여함으로써 새로운 스토리텔링 방식을 갖추고 있음을 보여준다[4].

라창현[5]은 스토리텔링 저작 도구 개발에 필요한 영화 및 드라마의 시나리오를 국내사례 중심의 분석을 토대로 메타데이터를 추출하였으며 사용자인 작가의 프로파일, 동영상 데이터베이스, 장소 이미지 데이터베이스를 기반으로 작가의 시나리오 작성을 위한 지능형 시나리오 작성 모듈을 제안하였다.

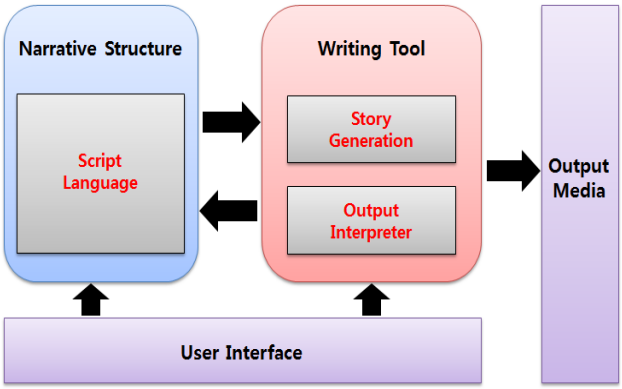


[그림 2-2] 지능형 관리 모듈[5]

김석규[6]는 인간과 컴퓨터간의 상호작용(HCI, Human-Computer Interaction)을 통한 다양한 스토리를 생성하는 기법인 인터랙티브 스토리텔링 시스템 구현방법에 대한 모델을 제안하였다. 시스템 구성요소인 서술구조를 표현하기 위한 방법을 제시하였으며 기존의 스크립트 언어 및 표현방식에 대한 분석을 통하여 스토리 생성

조선대학교 CHOSUN UNIVERSITY

을 위한 스크립트 언어 SML(Storytelling Mark up Language)를 제시하고 이를 통하여 [그림 2-3]과 같은 인터랙티브 스토리텔링 시스템을 구현하였다.



[그림 2-3] 인터랙티브 스토리텔링 시스템[6]

2. 스토리텔링 저작 지원 도구

스토리 생성을 위한 스토리텔링 시스템을 구성하기 위해 다양한 방법들이 제시되고 있다. 여기서 스토리텔링 시스템이란 스토리텔링 저작 지원 도구를 의미하며 이야기를 만들기 위해 사용되는 컴퓨터 소프트웨어를 말한다. 스토리텔링 시스템을 사용하는 창작자들은 스토리 저작을 위한 도구로써 워드프로세서나 스토리 데이터베이스, 상용 프로그램 등의 다양한 컴퓨터 기술을 이용하여 활용하고 있다. 이는 창작자에게 창작에 대한 보조 도구로서의 역할을 하고 일련의 스토리 생성의 자동화 과정을 통해 필요한 정보를 제공해줌으로써 영화, 게임, 소설, 드라마 등 다양한 문화산업에 전반적으로 활용되고 있다.

하지만 스토리텔링 저작 도구는 단지 컴퓨터 프로그래밍 기술에 의해 구현되는 것은 아니다. 컴퓨터 공학 이외에도 서사학이나 창작이론과 같은 인문적 지식도 함께 요구되므로 여러 분야에서의 학문적 요소에 대한 융합이 요구된다. 즉, 스토리텔링 저작 지원 도구를 설계 및 구현함에 있어서 스토리가 지닌 구조 분석을 통하여 [표 2-1]과 같은 기술적 요소들이 중요한 점으로 대두되고 있다.

현재 스토리 구조를 분석하여 컴퓨터 기술의 인공지능이나 알고리즘을 통해 스토리를 가공함으로써 컴퓨터가 스스로 이야기를 만드는 자동화 생성 기술에 대한 연구가 국내외적으로 진행되고 있다. 하지만 이에 따른 연구 성과물들은 기초적인

조선대학교 CHOSUN UNIVERSITY

수준에서 스토리 생성이 가능하다는 한계점이 있다.

본 절에서는 스토리 창작 활동을 지원하는 국내외 대표적 스토리텔링 저작 도구 사례에 대해 기술한다.

[표 2-1] 스토리텔링 저작 도구 개발에 요구되는 기술적 요소[3]

분야	기술	특징	활용 사례
서사학/ 창작이론	이야기 구조 분석	스토리텔링 저작 도구 구현에 가장 근본이 되는 이론적 토대를 제공	모티프, 플롯, 캐릭터 분석에 활용
	창작 과정 분석	창작자의 스토리 창작 과정에 대한 인지적, 기능적 특징 분석	창작 과정에 필요한 도움말, 가이드 제공
전산학/ 프로그래밍	데이터베이스	다양한 스토리 사례 ,데이터를 속성별로 데이터베이스 구축	소프트웨어에 포함되는 이야기 DB 구축
	UI 프로그래밍	소프트웨어의 UI, 메뉴화면, 기능 구현	소프트웨어 구현

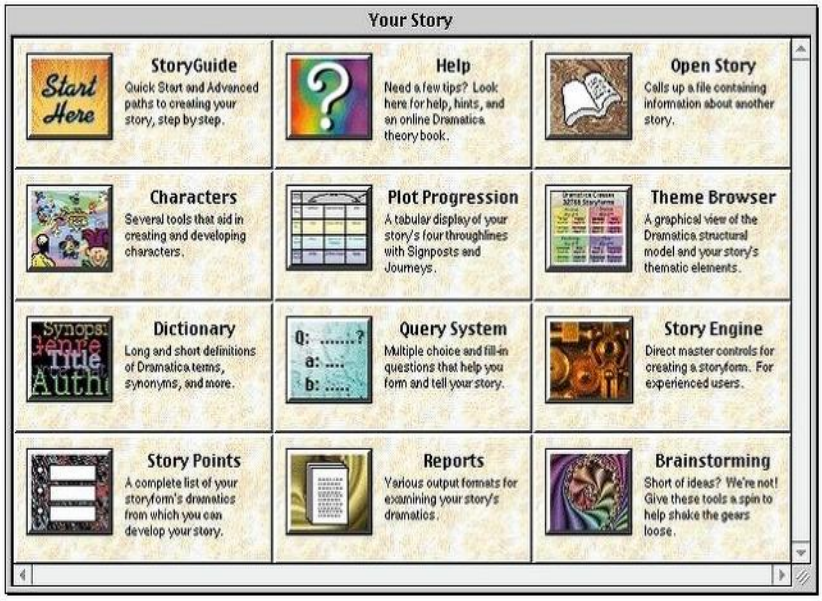
2.1 드라마티카 프로(Dramatica Pro)

드라마티카 프로는 미국의 라이트 브라더스(Write Brothers)에서 개발하였으며 영화·드라마 시나리오에 대한 스토리텔링 저작 소프트웨어이다. 드라마티카 프로는 캐릭터를 명확하게 설정하는 기능과 플롯(Flot)에 대한 체계화를 구성하고, 스토리 전개를 위한 전반적인 테마 레이어(Theme Layer)를 관리할 수 있는 스토리텔링 저작 지원 도구라고 할 수 있다. 이외에도 저작을 지원하기 위한 12가지 도구를 제공하고 있으며, 이에 산출된 작업을 스토리엔진(Story Engine) 모듈을 통하여 하나의 완결된 스토리를 생성할 수 있도록 제공하고 있다. 또한, 인간과 컴퓨터가 상호작용할 수 있는 쿼리 시스템(Dramatica Query System: DQS)을 통해 컴퓨터가 전달하는 주요 질문에 대한 창작자의 대답을 유도하는 스토리 저작 방식을 특징으로 가지고 있다.

드라마티카 프로는 스토리 이론(New Theory of Story)에 바탕을 두고 있으며, 기존의 작품에 대한 스토리 구조 분석을 토대로 새로운 형태로 이론화시킴으로써 캐릭터에 대한 유형과 플롯에 대한 갈등 양상을 구축할 수 있으며, 창작자는 그 틀에 맞춰 캐릭터, 플롯 속성, 각 속성들과의 관계를 설정할 수 있다.

조선대학교 CHOSUN UNIVERSITY

드라마티카 프로는 단순히 워드프로세서 기반의 시나리오 창작이 아닌 문예창작 이론을 컴퓨터 기술과 응용하여 실질적인 활용할 수 있다는 점에서 다른 스토리텔링 저작 도구와 차별성을 두고 있다.



[그림 2-4] 드라마티카 프로의 초기 화면

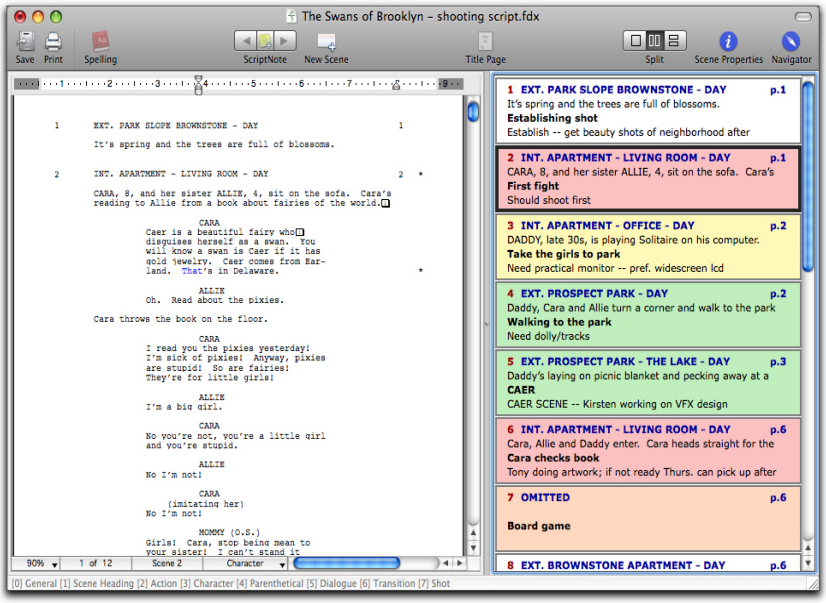
2.2 파이널 드래프트(Final Draft)

드라마티카 프로와 더불어 국외에서 많이 사용되고 있는 스토리텔링 저작 도구는 파이널 드래프트이다. 이는 연극과 TV 에피소드, 영화 스크립트 제작 등 시나리오 저작에 중점을 둔 저작 도구로써 다양한 유형의 시나리오 템플릿을 제공하여 스토리의 흐름을 검토할 수 있는 기능을 가지고 있다. 특히 인덱스 카드(Index Card Script) 기능을 통하여 여러 페이지를 한 화면에 볼 수 있도록 제공하고 있으며 시나리오 포맷에서 등장인물, 지문, 대사, 장면전환, 카메라 쇼트와 같은 문서서식에 대한 단축키를 제공하는 편리한 인터페이스를 가지고 있다.

특히, 새로운 문서를 생성할 때 영화, 방송, 소설을 작성할 수 있는 '스크립트', 'TV 템플릿', '텍스트 문서'의 3가지 양식을 제공함에 있어서 다양한 포맷을 선택할 수 있다. 하지만 시나리오 작성위주의 저작 도구로써 스토리와 캐릭터를 구조화할 수 있는 기능의 부재와 시나리오 작성 시 저장기능에 대한 오류로 인한 문제점이

조선대학교 CHOSUN UNIVERSITY

지적되고 있다.



[그림 2-5] 파이널 드래프트의 시스템 실행 화면

2.3 스토리 헬퍼(Story Helper)

국외 상용되고 있는 저작 도구와 더불어 영화 시나리오 및 방송 대본 작성의 저작 지원을 위해 국내 스토리텔링 저작 도구인 스토리 헬퍼가 이화여자대학교 연구진과 NC SOFT간의 공동연구로 개발되었다. 스토리 헬퍼는 드라마티카 프로와 파이널 드래프트처럼 스토리의 창작을 도와주는 저작 도구로써 1500편의 영화 및 애니메이션을 검토하여 3만 여개 요소의 데이터베이스를 생성하고, 205개의 모티프와 36개의 에피소드로 재정리하여 작가들의 스토리 구상에 따라 자유롭게 검색, 대조, 추출, 검토, 재구성 할 수 있도록 개발된 저작 도구이다. 하지만 영화 시나리오 중심의 스토리텔링 저작 지원도구로써 스토리 창작을 위한 작업의 초기단계에서 많은 인력과 시간이 소모되는 자료의 수집 및 분류에 대한 필요한 내용 전개에 큰 도움을 주기 어렵다는 문제점이 존재한다.

조선대학교
CHOSUN UNIVERSITY

[그림 2-6] 스토리 헬퍼의 실행 화면

3. 스토리텔링을 위한 서사구조 모델

본 장에서는 스토리의 서사구조에 따른 서사정보 및 인물관계에 대한 패턴을 분석하기 위해 본 논문에서 적용하고자 하는 서사구조 모델에 대한 분석을 기술한다.

3.1 스토리 전개에 따른 기승전결 서사구조

일반적으로 스토리텔링의 대표적인 서사구조는 기승전결의 형식을 갖추며 스토리 흐름이 분류되고 각 단계별로 스토리 내 인물간의 관계가 특정 사건으로 인해 변화하거나 지속된다. 기승전결을 분류하는 방법에는 전반부, 중반부, 후반부로 3단계로 나누거나 발단, 발전, 절정, 해결의 4단계 또는 발단, 전개, 위기, 절정, 결말의 5단계로 나눌 수 있으며 [표 2-2]는 서사구조의 단계별 내용을 나타낸다. 먼저, 발단부분에 해당하는 '기' 단계는 전체적인 스토리 전개를 위해 기본적인 정보를 제공하는 단계로 스토리에서 등장하는 인물이 소개되거나 사건의 실마리, 갈등이 암시되는 부분이다. '승' 단계는 등장인물의 성격이 변화하거나 다양한 사건들에 대한 복선이 암시되는 단계로 인물간의 관계가 뚜렷해지며, '전' 단계에서는 스토리의 절정부분에 속하며 인물간의 갈등 및 새로운 사건이 발생하고 인물간의 관계가 적대적 관계에서 우호적 관계 또는 우호적 관계에서 적대적 관계로 변할 수 있는 단계로 인물간의 관계에 따른 갈등의 최고조에 이르는 단계이다. 마지막으로 '결' 단계

조선대학교 CHOSUN UNIVERSITY

에서는 인물간의 관계가 마무리되는 단계로써 스토리의 주체가 되는 인물의 운명이 결정된다.

[표 2-2] 서사구조의 단계별 내용[7]

3단	4단	5단	내용
전반부	발단	발단	등장인물의 소개 배경의 제시 사건의 실마리 제시 갈등의 암시
중반부	발전	전개	등장인물 성격의 변화·발전 다양한 복선과 암시
			사건의 본격적 전개 갈등과 분규의 제시
	절정	위기	갈등과 분규의 상승 새로운 사건의 발생
		절정	갈등의 최고조 결말을 향한 분기점 해결의 실마리
후반부	해결	결말	인물의 운명 결정 갈등과 분규의 해결

3.2 조셉 캠벨의 영웅 서사구조

미국의 신화학자 조셉 켐벨(J.Cambell)은 시대와 지역을 막론하고 모든 신화의 근본적인 구조를 영웅들의 모험담 중심으로 사건정보를 3단계로 나누고 12가지의 모티프로 세분화 하였다[8]. 이러한 영웅 서사구조는 단순한 여행 스토리가 아닌 인간의 보편적인 성장 스토리의 일부이며 기존 공간에서의 이탈, 그로 인해 겪는 고난, 그 과정에서 우정과 사랑으로 인한 아픔을 극복하기 위한 영웅의 활약, 보조자의 도움, 그리고 최후 갈등과의 싸움 속에서 행복한 결말이라는 과정은 할리우드 영화 시나리오 구조에 전반적인 영향을 주었다. 본 논문에서는 설화 텍스트에서 추출된 서사정보에 대한 관계 설정을 위한 방법으로 영웅 서사구조의 인물, 사건, 배경 관계모델을 활용한다.

조선대학교 CHOSUN UNIVERSITY

① 인물(Character)

스토리에서 등장하는 인물의 원형 없이는 스토리를 전개시킬 수 없다. 따라서 설화 속에서 등장하는 인물(Character)의 유형을 역할에 따라 분류할 수 있다. 캠벨의 영웅 서사구조에서는 인물의 역할에 따라 영웅, 정신적 스승, 전령관, 관문 수호자, 변신 자재자, 그림자, 장난꾸러기의 7가지로 분류하였다.

[표 2-3] 인물정보 분류

인물	역할
영웅 (Hero)	- 특별한 세계로 모험을 떠남 - 영웅적인 임무 수행
그림자 (Shadow)	- 영웅과 반대되는 입장 - 주인공을 위기에 빠뜨림
전령관 (Herald)	- 주인공에게 모험을 떠나도록 경고, 자극 - 동기부여
정신적 스승 (Mentor)	- 영웅에게 능력 및 물건 등의 권한 부여 - 조언
관문 수호자 (Threshold Guardian)	- 영웅이 가는 길목에 지켜섬 - 영웅을 방해, 시험
변신 자재자 (Shape Shifter)	- 영웅을 속임 - 아군이 적군으로, 적군이 아군으로 변모함
장난 꾸러기 (Trickster)	- 극에 혼란을 일으킴

② 사건(Event)

스토리 창작에 있어 등장인물간의 조력적, 적대적 관계에 따라 특정 사건이 전개되고 이를 바탕으로 스토리의 내용은 더욱 풍성해질 수 있다. 영웅 서사구조에서는 시대와 지역을 막론하고 모든 신화의 근본적인 구조와 단계는 동일하다고 정의한다. 따라서 영웅 서사구조 기반의 설화 속에서 진행되는 사건을 사건이 시작하는 출발 단계와, 전개 단계, 귀환단계의 3단계로 분류할 수 있다. 캠벨의 영웅 서사구조에서는 영웅은 평범한 일상의 세계에서 초자연적인 신비의 땅으로 뛰어드는 출발(Departure), 미지의 세계에서 다양한 인물들과 조우하며 결정적인 승리를 쟁취하는 전개(Initiation), 모험을 끝으로 현실세계로 돌아오는 귀환(Return)의 3단계로 분류하고, 이에 12가지의 사건을 소분류로 나누었다. [표 2-4]는 사건정보에 따른 분류를 나타낸다.

조선대학교 CHOSUN UNIVERSITY

[표 2-4] 사건정보 분류

대분류	중분류	소분류
Event	출발 (Departure)	일상세계
		모험에의 소명
		소명의 거부
		정신적 스승과 만남
		첫 관문의 통과
	전개 (Initiation)	시험, 협력자, 적대자
		동굴 깊은 곳으로의 진입
		시련
		보상
	귀환 (Return)	귀환의 길
		부활
		영약을 가지고 귀환

③ 배경(Setting)

영웅 서사구조에서는 영웅이 모험 하는 과정에 대한 배경 요소로 인물과 사건이 존재하는 공간인 현실세계와 초현실세계로 구분한다. 주로 초현실세계의 배경은 오지의 땅, 숲, 지하 왕국, 물밑 세계, 하늘 위와 같은 미지의 공간으로 분류하였고, 현실세계의 배경은 산, 강, 바다와 같은 자연적인 공간으로 분류하였다. 이는 설화 스토리의 다양한 공간적, 시간적인 배경을 포함하고 있다. [표 2-5]는 배경정보에 따른 분류를 나타낸다.

[표 2-5] 배경정보 분류

공간	상세
초현실세계 (Adventure)	- 사건의 배경이 되는 시대와 공간
현실세계 (Real-word)	

조선대학교 CHOSUN UNIVERSITY

B. 설화 데이터 분석을 위한 자연어 처리 시스템

한국어의 특성상 복잡한 형태소들로 이루어져 있으며, 동음이의어가 많이 존재함에 따라 이를 해결하기 위한 많은 연구가 진행되고 있다. 본 절에서는 한국어로 작성된 설화 데이터를 처리하기 위한 방법으로 한글 형태소 분석기에 대해 기술한다. 영어는 굴절어로써 문장에 대한 구조 분석 방법이 문장을 구성하는 단어들에 대한 품사 태깅(Part of Speech Tagging)을 통해 해당 단어에 대한 품사를 추출하고 단어의 의미 분석을 통해 문장이 지닌 의미를 파악할 수 있다. 하지만 한국어는 교착어로써 복잡한 문장 구조를 가지고 있기 때문에 영어와 비교하였을 때 정확한 의미 파악이 힘들다고 할 수 있다[9][10].

예를 들면, 영어는 일반적인 자연어 처리(NLP, Natural Language Processing)과정에서 전처리(불용어 및 특수문자 제거)를 통해 일반명사, 고유명사, 동사, 형용사 등의 단어에 대한 품사를 POS Tagger를 통해 추출이 가능하지만, 한국어는 전처리 과정에서 조사, 어미, 접사 등의 영어에서 찾아보기 힘든 품사들이 존재하기 때문에 기존에 개발된 한글 형태소 분석기를 사용하여 데이터를 처리한다.

1. 꼬꼬마 형태소 분석기

한국어 자연어 처리를 위한 시스템으로써 서울대학교 IDS(Intelligent Data System) 연구실에서 개발된 꼬꼬마 한글 형태소 분석기[11]가 있다. 꼬꼬마 형태소 분석기는 Java 기반의 오픈소스(Open Source)로 배포되고 있으며 한국어 처리를 위한 연구를 수행할 때 사용되는 라이브러리와 사전 데이터를 포함하고 있다. 특히, 대량의 코퍼스(Corpus)가 필요할 경우 '세종 말뭉치'를 이용할 수 있는 시스템을 구축하여 품사, 형태소, 문어 및 구어 등의 정보를 확인할 수 있는 다양한 용도로 활용되고 있다. 꼬꼬마 형태소 분석기는 띄어쓰기 오류에 덜 민감하게 구현된 한글 형태소 분석기로 검증되었으며 휴리스틱(Heuristic)과 히든 마르코프 모델(HMM:Hidden Markov Model)에 기반한 확률 모델(Probabilistic Model)을 이용하여 상용 형태소 분석기와 성능을 비교하였을 때 뒤지지 않는 성능을 보이고 있다. 또한, 태그 집합으로써 [표 2-6]과 같이 '체언, 용언, 관형사, 부사, 감탄사' 등의 품사를 나누고 각 세부 품사를 구분한다. 하지만 한국어 처리를 위한 실행 효율성 및

조선대학교 CHOSUN UNIVERSITY

정확성이 높은 반면에 접근성 및 확정성이 떨어지고 다양한 요구에 대해 유연하게 대처할 수 없다는 단점이 존재하고 있다[12].

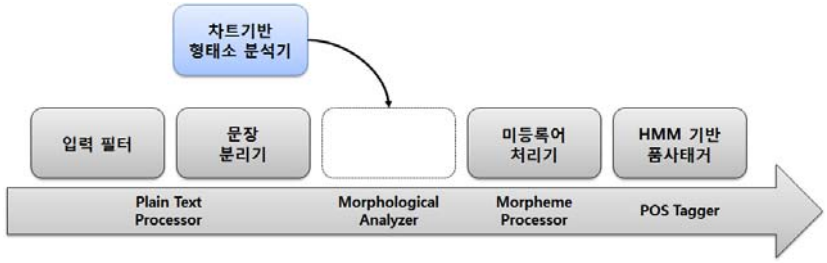
[표 2-6] 꼬꼬마 형태소 분석기의 태그 집합 예[11]

대분류	세종 품사 태그		대분류	세종 품사 태그	
	태그	설명		태그	설명
체언	NNG	일반명사	용언	VV	동사
	NNP	고유명사		VA	형용사
	NNB	의존명사		VX	보조 용언
	NR	수사		VCP	긍정 지정사
	NP	대명사		VCN	부정 지정사
관형사	MM	관형사	조사	JKS	주격 조사
부사	MAG	일반 부사		JKC	보격 조사
	MAJ	접속 부사		JKG	관형격 조사
감탄사	IC	감탄사		JKO	목적격 조사
어말 어미	EF	종결 어미		JKB	부사격 조사
	EC	연결 어미		JKV	호격 조사
	ETN	명사형 전성 어미		JKQ	인용격 조사
	ETM	관형형 전성 어미		JX	보조사

2. 한나눔 형태소 분석기

한나눔 형태소 분석기[13]는 기존 형태소 분석기의 접근성과 유연성의 단점을 보완하기 위해 한국과학기술원(KAIST) 시맨틱 웹 첨단연구센터에서 개발하고 관리하고 있는 오픈소스(Open Source)기반의 형태소 분석기이다. 한나눔 형태소 분석기는 한국어 처리 목적에 따른 플러그인 컴포넌트 아키텍처 구조를 지니고 있으며 필요에 따라 다양한 플러그인을 지원한다. 또한, UI(User Interface)기능을 통해 이용자의 필요에 따라 플러그인 형태를 추가하는 방식으로 사용이 가능하며 다양한 필터를 추가함으로써 확장이 가능하도록 한다. 따라서 유연하게 한국어 데이터를 처리하기 위한 워크플로우를 구성할 수 있으며 다양한 목적에 맞게 활용할 수 있다[13][14]는 장점을 지니고 있다. 이에 본 논문에서는 한국어로 작성된 설화 데이터를 처리하기 위해 한나눔 형태소 분석기를 활용하였다.

조선대학교 CHOSUN UNIVERSITY



[그림 2-7] 한나눔 형태소 분석기의 워크플로우[13]

[그림 2-7]는 한나눔 형태소 분석기의 워크플로우를 나타낸다. Plain Text Processor 단계에서는 문장 경계 인식 및 필터링, 자동 띄어쓰기 등의 형태소 분석을 위한 전처리 작업을 수행한다. 문장 경계 인식 방법으로는 문장의 구분자 역할을 할 수 있는 마침표 '.', 물음표 '?', 느낌표 '!'를 기준으로 문장 구분을 결정한다.

Morphological Analyzer단계에서는 입력된 문장에 대한 어절 단위로 발생 가능한 모든 형태소 분석 결과를 생성한다. 차트 기반 형태소 분석기는 형태소 분석을 위해 내부 저장 공간을 말하며 Morpheme Chart, Segment Position, Inverse Segment Position으로 구성되는 Lattice 형태의 차트를 사용하고 있다.

Morpheme Processor 단계는 미등록어를 처리하는 단계로써 형태소 분석 과정에서 사전에 등록되지 않는 단어를 처리하기 위해 'unk' 태그를 사용한다. 예를 들어, "나는 지금 불닭을 먹고 있다." 라는 문장에서의 '불닭'은 형태소 사전에 등록되지 않은 단어로써 [표 2-7]과 같은 분석 후보를 생성한다.

[표 2-7] 분석 후보 생성

불닭을	
	불닭/unk+을/jco
	불닭을/unk

이에 대해 한나눔 형태소 분석기의 미등록어 처리기는 "불닭/unk"에 대해 비서술성명사(ncn)와 고유명사(nq)를 태깅하며 [표 2-8]과 같은 결과를 생성한다.

[표 2-8] 미등록어 처리기

@불닭을	
	불닭/ncn+을/jco
	불닭/nq+을/jco

조선대학교 CHOSUN UNIVERSITY

마지막으로 POS Tagging 단계는 [표 2-9]와 같이 형태소 태그 집합을 통해 입력된 데이터로부터 최종 품사 태깅 결과를 반환하게 된다. 이는 어절에 대한 토큰화를 통해 해당 어절에 대한 품사를 HMM 기반의 품사 태거를 사용하여 한국어 처리를 가능하게 한다. 또한, 명사 추출기(Noun Extractor)를 통해 형태소 분석 및 품사 태깅 결과로 명사로 인식된 형태소만 추출하여 제공하는 기능이 있다.

[표 2-9] 한나눔 형태소 분석기의 태그 집합 예[13]

상위 분류			태그			
기호 s			sp 쉼표 sl 여는 따옴표 sd 이음표 su 단위기호		sf 마침표 sr 닫는 따옴표 se 줄임표 sy 기타기호	
외국어 f			f 외국어			
체언 n	보통명사 nc	서술성명사 np	ncpa 동작성 명사		ncps 상태성 명사	
		비서술성명사 ncn	ncn 비서술성 명사		ncr 직위 명사	
	고유명사 nq		ncpa 성 nqpc 성+이름		nqpb 이름 nqq 기타 -일반	
	의존명사 nb		nbu 단위성 의존명사 nbn 비단위성 의존명사			
	대명사 np		npp 인칭대명사		npd 지시대명사	
	수사 nn		nnc 양수사		nno 서수사	
용언 p	동사 pv		pvd 지시 동사		pvg 일반 동사	
	형용사 pa		pad 지시형용사		paa 성상형용사	
	보조용언 px		px 보조용언			

조선대학교 CHOSUN UNIVERSITY

C. 토픽맵

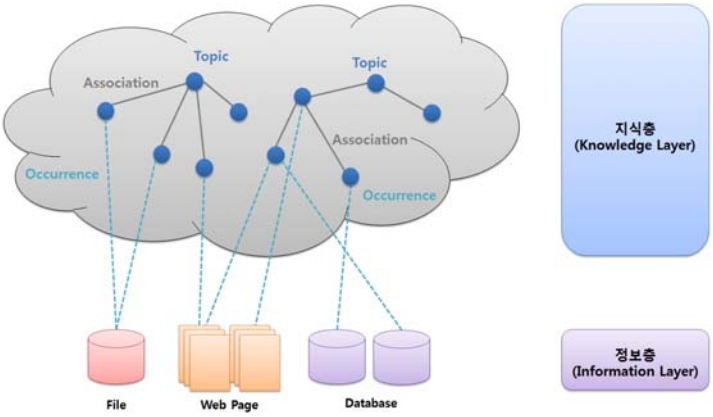
토픽맵은 국제표준화기구(ISO)에서 시맨틱 웹을 구현하기 위한 온톨로지 생성 언어 중 하나이며 W3C(Word Wide Web Consortium)의 OWL과 상호 보완적이고 경쟁 관계에 있으며 용어의 의미적 구조를 나타내는 용어집, 색인집, 시소러스를 다루는 목적이 있다[15]. 현재 토픽맵은 시맨틱 웹의 핵심기술로 인정받으며 지식을 표현하는 방법론으로써 정의된 구조와 지식 자원에 대한 의미적 관계를 지도(Map)로 표현하고 검색이 가능하도록 하며, 정보에 대한 시각화(Visualization)에 목적이 있다. 또한, W3C의 RDF, OWL 보다 많은 정보를 가지며 정보를 관리함에 있어서 GPS(Global Positioning System)와 같은 역할을 하는 토픽맵은 지식표현과 정보관리의 각 응용분야간의 다리역할을 하며 매우 간단하고 복잡한 것에서 단순하고 모호한 정보까지 지식표현 스키마의 전체적인 범위를 처리할 수 있다[16].

본 절에서는 서사정보 추출 시스템으로부터 추출된 스토리 데이터에 대한 관리 및 시각화를 위해 토픽맵 기반의 응용방안을 제시하기 위해 토픽맵의 구조와 그에 대한 구성요소를 살펴보고, 토픽맵을 이용한 정보 시각화 사례에 대해 기술한다.

1. 토픽맵의 구조 및 구성요소

토픽맵은 지식을 관련된 정보와 연계를 위한 추상적 구조를 말하며 2개의 레이어로 구성되어 있다. [그림2-7]는 토픽맵 모델을 나타내는 것으로 지식 레이아웃층(Knowledge Layout)과 정보 레이아웃층(Information Layout)을 나타내고 있다. 상위(Upper)계층인 지식 레이어는 색인의 역할을 하며 주제를 실체화하는 토픽(Topic)을 기본요소로 토픽간의 관계를 설정하는 어소시에이션(Association), 토픽의 연관 정보자원의 위치를 나타내는 어커런스(Occurrence)로 구성된다[17]. 토픽에 다양한 명칭을 부여하기 위한 명명법과 토픽의 개념에 대한 범위를 한정하는 Scope, 토픽의 역할을 규정하는 Role 등 3가지의 보조 개념도 사용된다[18]. 하위(Lower)계층인 정보 레이어는 정보의 콘텐츠 자원영역을 나타내는 역할을 하며 텍스트나 그래픽, 비디오, 오디오, 웹페이지 등 어떠한 포맷도 가능하다.

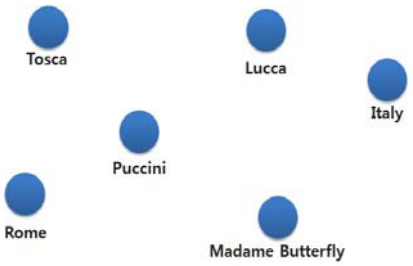
조선대학교
CHOSUN UNIVERSITY



[그림 2-8] 토픽맵의 구조[19]

① 토픽(Topic)

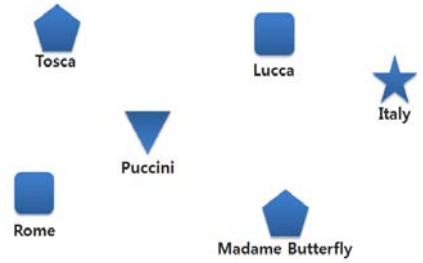
토픽[20]은 일반적으로 사람, 개체(Entity), 주제(Subject), 단어 등 어떠한 것이든 될 수 있으며 특정 속성이나 어떤 의미 등을 기계적으로 이해하고 처리할 수 있도록 객체(Object)로 표현 되는 것을 말한다. 주로 명사형으로 표현되며 각각의 토픽은 하나의 이름(name)을 갖는다.



[그림 2-9] 토픽

다양한 토픽들을 서로 연결하기 위해 토픽 타입(Topic Type)을 가질 수 있는데, 토픽 타입은 토픽의 유형을 말하며 토픽을 유형에 따라 분류시켜 주는 역할을 한다. 또한, 토픽은 하나 이상의 토픽 타입에 속할 수 있고 토픽 타입 자체가 또 하나의 토픽으로 정의될 수 있으며 경우에 따라 토픽 타입이 없는 토픽이 존재할 수 있다.

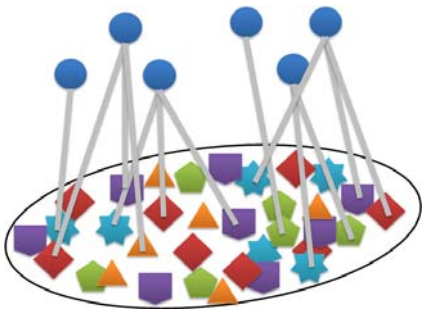
조선대학교 CHOSUN UNIVERSITY



[그림 2-10] 토픽 타입

② 어커런스(Occurrence)

어커런스는 토픽을 하나 이상의 정보 리소스와 연결 시켜주는 기능을 하고 음악 파일, 이미지파일, 문서파일, 데이터베이스 내 특정 레코드 등의 외부 정보 리소스에 대한 참조를 나타낼 수 있으며 웹 페이지에 존재하는 정보 또한 URL을 통해 어커런스가 될 수 있다. 즉, 토픽맵에 존재하는 모든 토픽은 HyTime 어드레싱 또는 Xpointer와 같은 기법들을 사용하여 자신이 참조하는 리소스를 가르키며, 어커런스에 의해 참조되는 리소스는 일반적으로 토픽맵과 별도로 저장된다.

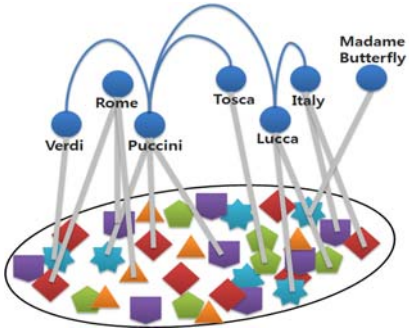


[그림 2-11] 어커런스

③ 어소시에이션(Association)

연관관계는 둘 이상의 토픽에 대한 관계를 정의하는 것이며 스토리 도메인의 네트워크를 보여주기 위해 사용되는 것으로 토픽간의 상하관계가 아닌 의미적 관계를 정의한다. 아래 [그림 2-11]은 토픽과 어커런스, 그들 간의 관계를 정의한 것이다.

조선대학교
CHOSUN UNIVERSITY

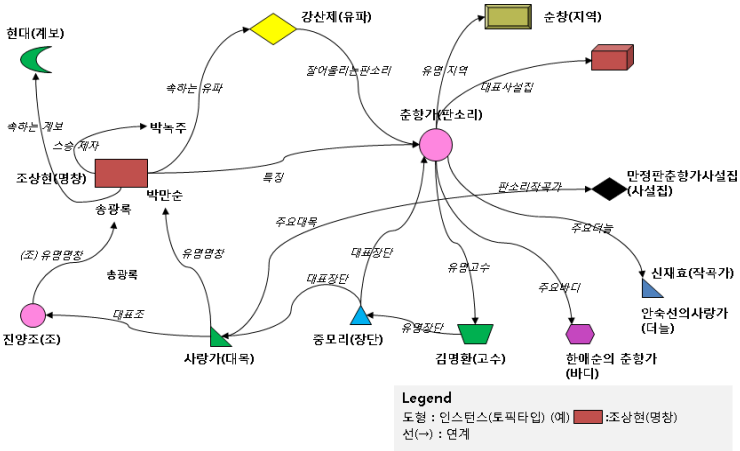


[그림 2-12] 어소시에이션

이러한 토픽맵의 특징으로 인해 기존의 시소러스 체계인 상하 관계(Broad Term/Narrow Term)나 유사용어 관계(Related Terms)체계와 비교하면 아주 우수한 장점이다. 다시 말해, 다양한 형태의 연계를 자연어(Natural language)형식으로 표현하며 토픽 간의 고정된 관계를 설정한 것이 아니므로 기존 분류시스템과 다른 유연성 측면에서 큰 차이를 보인다고 할 수 있다[19].

2. 토픽맵을 이용한 정보 시각화

오삼균[21]은 판소리에 대한 정보 시각화를 위해 토픽맵 기반 판소리 검색 시스템을 구축하였다. 판소리와 관련된 명창, 유파, 대표 장단 등의 요소 간 관계를 토픽맵으로 시각화하였으며, 기존 판소리 검색시스템과 질의별 실험을 수행하였다.

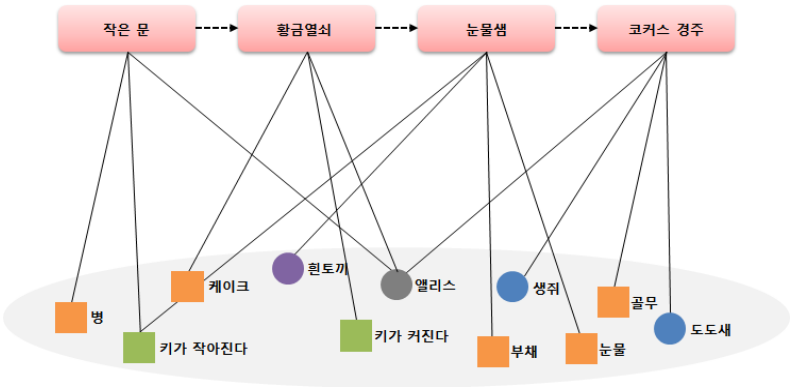


[그림 2-13] 판소리 토픽맵[21]

조선대학교 CHOSUN UNIVERSITY

권진만[22]은 토픽맵을 이용하여 뮤지션에 대한 정보를 시각화하기 위해 가수, 작곡가, 노래를 토픽타입(Topic Type)으로 분류하고 관련된 정보를 동영상과 음악파일, 웹 페이지의 노래정보와 연결하였다. 토픽간의 연계타입(Association Type)으로는 '부르다', '작사하다', '작곡하다'로 연결하였으며 최종적으로 뮤지션 검색시스템 구축을 통해 기존의 음악 검색사이트와 정보검색에 대한 시각화 측면의 비교실험을 수행하였다.

이지수[23]는 일련의 흐름을 갖는 스토리 라인의 형상화를 위해 스토리 내 존재하는 등장인물과 그들 주변에서 발생하는 사건관계를 토픽맵 기반의 시각화 방법을 제안하였다. 스토리 대상으로는 루이스 캐롤(Lewis Carroll)의 <이상한 나라의 앨리스>를 중심으로 수행하였으며, 스토리 내 요소들을 프로프의 '민담 형태론'에 근거하여 토픽맵 기반의 시각화 방법을 제시하였다.



[그림 2-14] 이상한 나라의 앨리스의 플롯 중심 토픽맵[22]

[그림 2-14]는 토픽맵을 이용하여 정보 레이아웃에 스토리 요소인 인물, 장소, 키워드의 토픽을 배열하고, 지식 레이아웃에는 정보 레이아웃의 토픽들에 대한 관계를 나타내는 이벤트를 나타냄으로써 플롯 내부에서의 인과관계를 표현하였다.

조선대학교 CHOSUN UNIVERSITY

Ⅲ. 서사정보 자동 추출 시스템 설계 및 구현

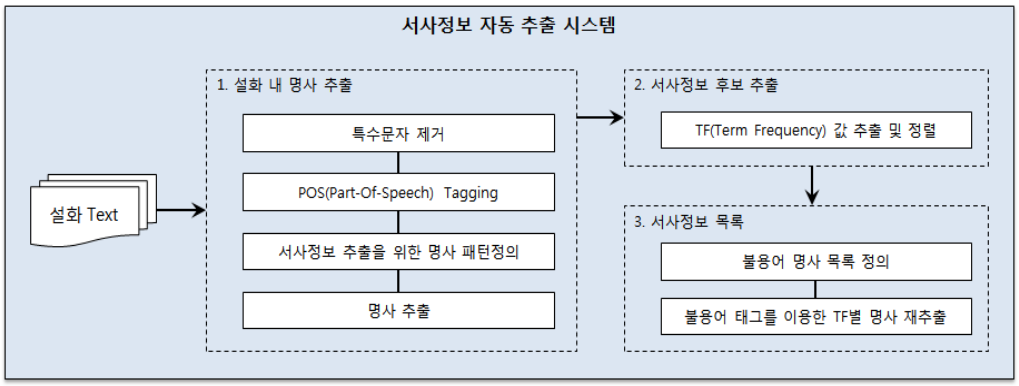
본 논문에서 제안하는 스토리텔링 저작 지원을 위 설화 내 서사정보를 자동으로 추출하기 위한 시스템을 설명하고, 그에 대한 방법에 대해 기술한다.

A. 서사정보 자동 추출 시스템 설계 및 인터페이스 정의

본 절에서는 한국어로 작성된 설화 텍스트를 자연어 처리(NLP, Natural Language Process) 기법을 활용한 서사정보 자동 추출 시스템 설계 및 구현을 위해 전체적인 시스템 구성도와 개발 환경에 대해 기술하고, 사용자 인터페이스를 보여준다.

1. 시스템 구성도

[그림 3-1]은 제안하는 설화 내 서사정보 자동 추출 시스템의 시스템 구성도이다. 시스템 구성도는 크게 입력받은 설화 텍스트로부터 한국어 자연어 처리 기법을 활용하여 명사를 추출하는 단계, 추출된 명사로부터 서사정보 후보명사를 추출하기 위해 TF 알고리즘을 적용하는 단계, 마지막으로 입력된 설화 텍스트로부터 인물, 사건, 배경의 추출된 서사정보 목록을 보여주는 단계로 이루어진다.



[그림 3-1] 시스템 구성도

조선대학교 CHOSUN UNIVERSITY

2. 개발 환경

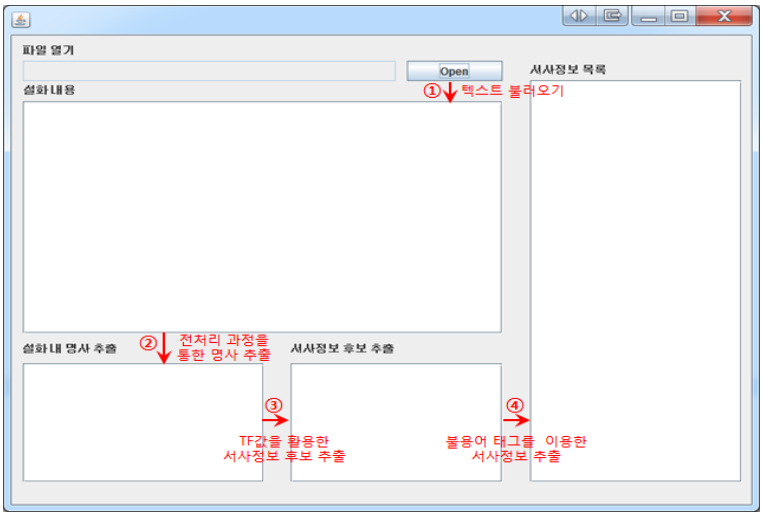
서사정보 자동 추출 시스템의 구축을 위한 개발환경은 [표 3-1]과 같다. 컴퓨터 CPU는 Intel(R) Core(TM) i7 2.93GHz 환경에서 Java 기반으로 구현하였으며, 인터페이스 설계를 위해 Eclipse 플러그인 Window Builder를 이용하여 레이아웃을 구성하였고, 텍스트 처리를 위해 한나눔 형태소 분석기의 품사 태그를 이용하였다.

[표 3-1] 서사정보 자동 추출 시스템 개발 환경

항목	사양
CPU	Intel(R) Core(TM) i7 2.93GHz
RAM	4.00GB
운영체제	윈도우7 Enterprise K
개발 툴	Java, Eclipse, Window Builder
분석기	한나눔 형태소 분석기

3. 사용자 인터페이스(User Interface)

[그림 3-2]는 구현된 서사정보 자동 추출 시스템의 사용자 인터페이스를 보여준다. 사용자 인터페이스는 시스템 구성도의 흐름과 같이 입력받은 설화 텍스트로부터 명사를 추출하는 부분과 추출된 명사 중에서 서사정보의 후보들을 추출하는 부분과 서사정보에 대한 목록을 보여주는 부분으로 구성된다.



[그림 3-2] 사용자 인터페이스

조선대학교 CHOSUN UNIVERSITY

B. 서사정보 추출을 위한 전처리 및 명사 추출 방법

본 절에서는 입력받은 설화 텍스트로부터 서사정보를 추출하기 위한 전처리(Pre-processing) 과정으로 [그림 3-2]의 사용자 인터페이스 화면에서 설화 내 명사를 추출하는 방법에 대해 기술한다. 명사를 추출하는 이유는 다양한 품사 중에서 서사정보를 구성하고 있는 단어에 대한 품사가 주로 명사에 분포되어 있기 때문이다. 이에 설화 텍스트 내 명사 추출을 위해 모든 자연어 처리 기법에서 이루어지는 특수문자 제거 방법과 해당 단어별 품사 처리를 위한 POS(Part-Of-Speech) Tagging 과정에 대해 설명하고 서사정보 추출을 위한 명사 패턴을 정의하여 추출한다.

1. 특수문자 제거 및 POS(Part-Of-Speech) Tagging

모든 언어에서 사용되는 특수문자의 종류는 매우 다양하다. 대화 내용을 나타내기 위해 큰따옴표(“ ”)를 사용하고, 문장의 맥락을 구성하기 위해 마침표(.), 쉼표(,) 등을 사용한다. 하지만, 이러한 특수기호들은 자연어 처리에서 큰 걸림돌이 되기 때문에 본 논문에서도 본격적인 한국어 처리를 하기 위해 특수문자를 제거한다. 제거할 특수 문자는 “*, +, $, |, !, #, %, &, @, ,, ?, ", ', \ , [,], /, <, >, ^, {, }”이며, 특수문자 제거 방법은 [표 3-2]과 같다. 특수문자 제거 시 주의할 사항은 제거할 특수문자를 정의할 때 대부분 대괄호([ ]) 안에 제거할 문자를 정의하지만 JAVA 함수에서 사용되는 대, 중, 소괄호와 큰따옴표, 작은따옴표는 반드시 앞에 역슬러쉬(\)를 포함하여 그 특수성을 지정해주어야 한다.

[표 3-2] 설화 텍스트 내 특수문자 제거 소스

```
public class specialchar {
        public static void main(String[] args){
                String input_text = "*+$|!#%&@,?\\[\\]\\/\\<\\>\\^\\{\\}";
                String output =
        input_text.replaceAll("[*+$|!#%&@,?\\[\\]\\/\\<\\>\\^\\{\\}]", "");
                System.out.println(output);

        }
}
```

조선대학교 CHOSUN UNIVERSITY

전체 텍스트 내에 특수문자를 제거 후 각 단어별 품사를 파악하기 위한 POS Tagging 과정을 진행한다. 한국어 자연어 처리에 있어서 매우 다양한 품사들이 존재하는데, 문맥적으로 특정 품사를 보조해주는 역할을 하는 품사들이 존재하기 때문에 영어와 달리 매우 어려운 점이 존재한다.

따라서, 본 논문에서는 기존에 한국과학기술원(KAIST) 시맨틱 웹 첨단연구센터에서 개발 및 관리를 하고 있는 한나눔 형태소 분석기를 통해 본 과정을 진행한다. 한나눔 형태소 분석기를 이용하는 이유는 타 분석기와 달리 특정 명사에 대한 품사 종류를 나타낼 수 있는 경우에 수에 따라 최대한 많이 보여줌으로써 연구자들이 사용할 수 있는 범위를 확대시켜 놓았기 때문이다. 또한 체계적으로 정리된 단어-품사 관계를 제공함으로써 본 논문에서 사용하고자 하는 불용어 처리에 큰 역할을 할 수 있을 것으로 사료되어 한나눔 형태소 분석기를 사용한다.

설화 텍스트에 대한 POS Tagging 과정은 먼저 설화 텍스트 원문을 입력받을 때, 모든 기호(“\n”)를 제거하여 하나의 문장형태로 데이터를 수정한다. 그리고 앞서 정의한 특수문자를 제거하는데, 마침표는 문장의 끝을 구분하여야 함으로 제거하지 않는다. 최종적인 POS Tagging은 한나눔 분석기 함수에 기반하여 데이터가 가공되는데 각각의 품사 별로 작성된 기반데이터를 통해 Tagging과정을 거치게 되며 본 논문에서 처리하는 품사 태깅 과정은 [표 3-3]와 같다.

[표 3-3] 설화 텍스트에 대한 POS Tagging 예

지하국대적퇴치 설화 원문	옛날 아귀 귀신이라는 큰 도적이 있었다. 그는 종종 이 세상에서 나와서 세상을 요란하게 하고 예쁜 여자를 납치해 가기도 하였다.…	
특수문자 제거	옛날 아귀 귀신이라는 큰 도적이 있었다. 그는 종종 이 세상에서 나와서 세상을 요란하게 하고 예쁜 여자를 납치해 가기도 하였다.…	
POS Tagging	옛날/**ncn** 아귀/**ncn** 귀신/**ncn**+이/**jp**+라/**ef**+는/**etm** 크/**paa**+ㄴ/**etm** 도적/**ncn**+이/**jcc** 있/**px**+었/**ep**+다/**ef** 그/**npp**+는/**jxc** 종종/**mag** 이/**mmd**	세상/**ncn**+에서/**jca** 나오/**pvg**+아/**ecx** 세상/**ncn**+을/**jco** 요란/**ncps**+하/**xsms**+게/**ecs** 하/**pvg**+고/**ecc** 예쁘/**paa**+ㄴ/**etm** 여자/**ncn**+를/**jco** 납치/**ncpa**+하/**xsva**+어/**ecx** 가/**px**+기/**etn**+도/**jxc** 하/**pvg**+었/**ep**+다/**ef**

조선대학교 CHOSUN UNIVERSITY

2. 서사정보 추출을 위한 명사 패턴 정의 및 추출

앞서 설명한 POS Tagging까지의 과정이 완료되면, 설화 텍스트 내 명사를 추출하기 위해 명사를 나타내는 품사의 종류를 파악하여 해당 품사만 선별해 내는 과정을 거친다. 현재 한나눔 형태소 분석기에서 정의하고 있는 명사 형태소의 구성은 ncpa(동작성 명사), ncps(상태성 명사), ncn(비서술성 명사), ncr(직위 명사), ncpa(성), nqpb(이름), nqpc(성+이름), nqq(기타 일반), nbu(단위성 의존명사), nbn(비단위성 의존명사)와 같다. 위 명사 형태소 구성에서 동일 패턴을 분석한 결과 다음과 같은 패턴을 갖는 것을 확인할 수 있었다.

[표 3-4] 한나눔 형태소 분석기 내 정의된 명사 패턴

○ nc + n, r, pa, ps - 일반 명사 및 동작, 상태성 명사
○ nq + pb, pc, q - 기타 명사 및 고유명사의 성 또는 이름
○ nb + n, u - 단위성, 비단위성 의존명사

따라서 [표 3-4]에서 정의한 3가지 패턴을 기반으로 명사만을 추출하며, 명사 추출 알고리즘은 [표 3-5]와 같다.

[표 3-5] 명사 추출 알고리즘

```
try {
        /* 쓰레드 모드 내 Workflow 알고리즘 수행 */
        workflow.activateWorkflow(true);

        /* Analysis using the workflow알고리즘을 사용한 분석 시작 */
        String document = "옛날 아귀 귀신이라는 큰 도적이 있었다.
                          그는 종종 이 세상에서 나와서 세상을 요란하게 하고…\n";
        workflow.analyze(document);
        str_test = workflow.getResultOfDocument();
        /* workflow 알고리즘 분석 결과 반환 */
        String li_str[] = str_test.split("\n");
        /* 정의한 명사 추출 시작 */
        for(int i=1; i < li_str.length; i++){
         if (li_str[i].contains("nc") || li_str[i].contains("nq") li_str[i].contains("nb")) {
                        System.out.println(li_str[i]);
                }
}
```

조선대학교 CHOSUN UNIVERSITY

[표 3-5]의 과정을 통해 추출한 명사 총 382개이며, [표 3-6]과 같다.

[표 3-6] 명사 패턴을 활용하여 추출된 명사

옛날, 아귀, 귀신, 도적, 그, 세상, 세상, 요란, 여자, 납치, 때, 아귀, 귀신, 임금님, 세, 공주, 납치, 임금님, 신하, 귀, 계획, 신통, 계책, 말, 사람, 사람, 무신, 나, 자신, 일, 임금님, 저, 집, 대, 국록, 저, 생명, 은혜, 공주, 님, 임금님, 이, 허락, 세, 공주, 중, 막내, 공주, 결혼, 무신, 하인, 아귀, 귀신, 소굴, 출발, 천하, 귀신, 소굴, 곳, 수, 하루, 산모퉁이, 피곤, 몸, 잡시, 동안, 잠, 꿈, 머리, 노인, 다음, 말, 나, 산, 산신령, 너, 아귀, 귀신, 소굴, 산, 쪽, 중, 산, 이상, 한, 바위
...

C. 서사정보 추출을 위한 TF 알고리즘 활용 방법

본 절에서는 설화 내 서사정보 목록을 추출하기 위해 [그림 3-2]의 서사정보 후보를 추출하는 알고리즘 활용 방법에 대해 기술한다. 서사정보의 후보 명사를 추출하는 방법으로는 TF(Term Frequency)를 활용한다. 이때 추출된 서사정보의 후보에서 해당 명사들 중 한가지만의 품사만을 갖지 않는 명사가 있기 때문에 불용어 명사 목록을 정의하여 입력된 설화 텍스트로부터 서사정보 목록을 추출한다.

1. TF(Term Frequency) 값을 이용한 서사정보 후보 추출

본 논문에서는 자연어 처리 분야에서 많이 사용되는 TF(Term Frequency)값을 활용하여 서사정보 후보를 추출하고자 한다. TF값은 일반적인 문서 내 출현하는 특정단어의 빈도수를 통해 해당 단어의 중요성을 판별하는데 많이 사용되는 기법으로 이를 사용하여 서사정보 데이터를 추출한다. 이러한 서사정보 추출을 위해 TF값을 사용하는 이유는 스토리의 핵심요소인 서사정보를 구성하고 있는 단어는 명사 품사를 가지고 있으며, 스토리의 주체가 되는 인물, 흐름을 나타내는 사건 및 배경은 타 명사보다 그 출현빈도가 크기 때문에 TF값을 활용한다.

[표 3-7]은 TF값 추출 알고리즘의 소스를 나타내고 있다. TF값을 추출하는 알고리즘에 대한 설명으로는 배열 내 들어있는 단어들끼리 비교하여 기존에 측정했던 단어일 경우에는 다음 단어로 넘어가서 재비교 후 새로운 단어일 경우만 Count값

조선대학교 CHOSUN UNIVERSITY

을 늘려주면서 단어에 대한 TF값을 계산하게 된다.

[표 3-7] TF값 추출 알고리즘

```
private Word[] getFrequentWords(String words[]){
    HashMap<String,Word> map = new HashMap<String,Word>();
      for(String s:words){ //map 배열에 들어있는 단어들 간 비교
          Word w = map.get(s);
            if(w==null)  w = new Word(s,1);
            else  w.count++;
            map.put(s, w);
        } //map 배열 내 단어가 있을 경우 count값을 1씩 증가
        Word[] list = map.values().toArray(new Word[]{});
        Arrays.sort(list); return list;
    }
    // 리스트 형태로 저장된 값을 배열로 전환
    public static String[] main(String[] args) {
        words = args;
        int i = 0;
        Word[] frequency = new tf().getFrequentWords(words);
        tf_word = new String[words.length];
        System.out.println(words.length);
        for(Word w:frequency){
                tf_word[i] = w.word+"="+w.count;
                i++;
        } return tf_word;          }
```

[표 3-7]에서 나타낸 TF값 계산 알고리즘을 통하여 추출해 낸 단어별 TF값은 다음 [표 3-8]과 같으며, 기존 382개의 명사 중 115개의 명사를 서사정보 후보로 추출하였다. 최종적으로 중복 처리된 명사를 바탕으로 재필터링 과정을 거치게 된다. 재필터링 하는 이유는 명사만을 추출하였더라도 해당 명사가 한가지의 품사만을 갖지 않는 단어가 존재하기 때문에 재필터링 과정을 수행하여 설화 내 서사정보를 추출하기 위함이다.

[표 3-8] 명사별 TF값 계산 결과

명사	TF값	명사	TF값	명사	TF값	명사	TF값
공주	17	아귀	7	술	4	광주리	3
무신	15	그	6	위	4	그것	3
도적	10	하인들	6	줄	4	나	3
귀신	9	산	5	구멍	4	다음	3
말	9	세상	5	노인	4	때	3
사람	9	수	5	막내	4	머리	3
임금님	9	일	5	바위	4	몸	3
공주들	8	저	5	집	4	물 …	3 …

조선대학교 CHOSUN UNIVERSITY

2. 불용어 명사 목록 정의

앞서 추출된 서사정보의 후보 명사 중에서 한 가지 품사만을 갖지 않는 명사들을 제거하는 과정을 위해 [표 3-9]와 같이 불용어 명사 목록을 정의한다.

[표 3-9] 불용어 명사 목록

번호	불용어 명사
1	가구
2	가닥
3	가락
4	가량
5	가리
6	가마
904	힝

추출된 명사와 함께 정의된 태그를 포함한 불용어 명사를 제거함으로써 원하는 서사정보 추출이 가능하게 된다. [표 3-10]는 명사에 정의된 불용어의 태그와 단어의 개수를 보여준다. 예를 들어 [표 3-8]에서 '그것'의 경우 명사 품사 이외에 지시대명사를 나타내는 태그인 'npd'의 품사를 지니므로 [표 3-10]의 정의된 태그를 이용하여 불용어 명사를 제거하는 과정이 필요하다.

[표 3-10] 명사에 정의된 불용어의 태그

태그명(단어개수)	설명	태그명(단어개수)	설명
xsnu(25)	명사 파생 접미사	xsms(2)	형용사 파생 접미사
xsnca(1)	명사 파생 접미사	xsmn(6)	형용사 파생 접미사
xsncc(26)	명사 파생 접미사	xsam(1)	부사 파생 접미사
xsna(8)	명사 파생 접미사	xsas(1)	부사 파생 접미사
xsns(2)	명사 파생 접미사	ii(208)	감탄사
xsnp(7)	명사 파생 접미사	nbn(70)	비단위성 의존명사
xsnx(14)	명사 파생 접미사	nbu(388)	단위성 의존명사
xsvv(7)	동사 파생 접미사	nbs(10)	비단위성 의존명사
xsva(5)	동사 파생 접미사	npp(78)	인칭대명사
xsvn(1)	동사 파생 접미사	npd(44)	지시대명사

조선대학교 CHOSUN UNIVERSITY

3. 불용어 태그를 이용한 TF별 명사 재추출

본 장에서는 [표 3-10]에서 정의된 불용어 태그를 바탕으로 TF값을 계산하여 추출된 단어들을 재분류함으로써 서사정보를 구성하는 명사에 대한 목록을 추출한다.

[표 3-11]은 TF별 명사 재추출을 위해 불용어 명사를 제거하기 위한 불용어 태그 정리 소스를 나타낸다.

[표 3-11] 불용어 명사를 제거하기 위한 불용어 태그 정리 소스

```
public static void main(String[] args) {
 try {
        outputStream = new FileWriter("C:/my.txt");
        BufferedReader br = new BufferedReader(
        new InputStreamReader( new FileInputStream(dir1), "utf-8"));
       while((var_text = br.readLine()) != null){
       if(var_text.contains("xsnu") || var_text.contains("xsnca") || var_te
       xt.contains("xsncc") || var_text.contains("xsna") || var_text.contai
       ns("xsns") || var_text.contains("xsnp") || var_text.contains("xsnx
       ") || var_text.contains("xsvv") || var_text.contains("xsva") || var_
       text.contains("xsvn") || var_text.contains("xsms") || var_text.cont
       ains("xsmn") || var_text.contains("xsam") || var_text.contains("xs
       as") || var_text.contains("ii")|| var_text.contains("nbn") || var_tex
       t.contains("nbu") || var_text.contains("nbs") || var_text.contains("
       npp") || var_text.contains("npd")) {

       outputStream.write(var_text);}
          br.close();
   } catch(Exception e2){
      JOptionPane.showMessageDialog(null, e2);
  }
 }
}
```

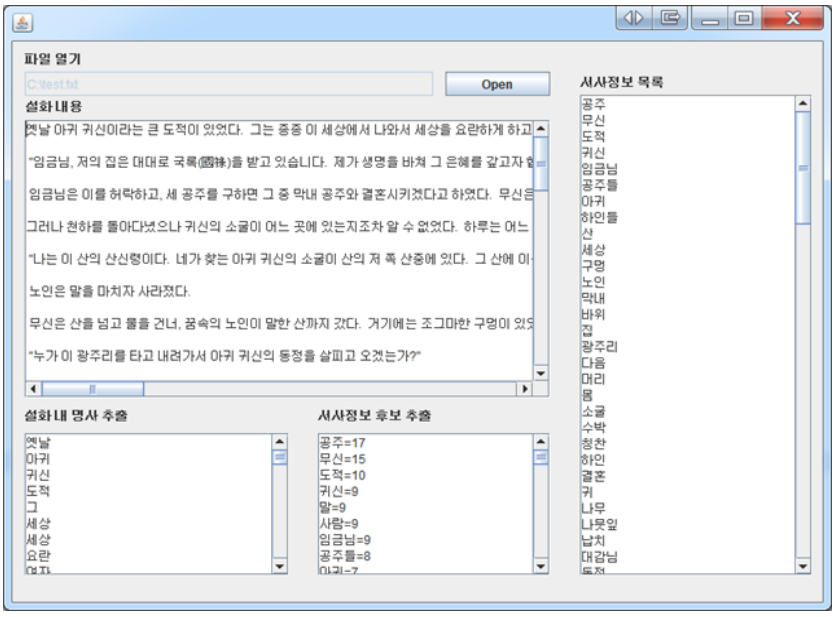
이와 같이 불용어 제거 단계를 통해 최종 추출된 서사정보 명사는 [표 3-12]와 같으며, 추출된 서사정보 명사는 총 41개의 명사를 확인할 수 있고, 단순 출현 명사를 제거하기 위해 TF값이 1이하인 데이터는 삭제하였다.

조선대학교 CHOSUN UNIVERSITY

[표 3-12] '지하국대적퇴치' 설화 내 서사정보 추출 결과

명사	TF	명사	TF	명사	TF
공주	17	집	4	대감님	2
무신	15	광주리	3	동정	2
도적	10	다음	3	땅	2
귀신	9	머리	3	마음	2
임금님	9	몸	3	목	2
공주들	8	소굴	3	물동이	2
아귀	7	수박	3	비늘	2
하인들	6	칭찬	3	소원	2
산	5	하인	3	아가씨	2
세상	5	결혼	2	옆구리	2
구멍	4	귀	2	의심	2
노인	4	나무	2	잠	2
막내	4	나뭇잎	2	치맛자락	2
바위	4	납치	2		

[그림 3-3]는 본 논문에서 제안한 방법을 통해 설화 텍스트 중 하나인 '지하국대적퇴치' 설화를 바탕으로 서사정보 자동 추출 시스템의 결과를 보여준다. 좌측의 입력된 설화 텍스트에 대한 내용을 기반으로 설화 내 명사를 추출하는 단계, 서사정보 후보를 추출하는 단계를 통하여 우측의 서사정보 목록을 나타낸다.



[그림 3-3] 추출된 서사정보 목록 결과

조선대학교 CHOSUN UNIVERSITY

[표 3-13]은 본 논문에서 구현한 서사정보 자동 추출 시스템의 처리 과정을 통해 얻은 결과를 나타내며 서사정보인 인물, 사건, 배경의 3요소로 '지하국대적퇴치' 설화 데이터를 분류한 결과를 나타낸다.

[표 3-13] '지하국대적퇴치' 설화 내 서사정보 분류 결과

인물	배경	사건	
공주	산	칭찬	애기
무신	세상	결혼	요란
도적	소굴	납치	은혜
귀신	공중	동정	응답
임금님	굴	소원	이야기
공주들	꿈	의심	조사
아귀	꿈속	계책	죽음
하인들	산모퉁이	계획	출발
노인	세계	대답	피곤
막내	옛날	명령	허락
하인	오랜만	방법	혼인
대감님	우물곁	병환	약속
아가씨	천하	아양	야단

이와 같이 추출된 서사정보 내 분류 결과를 통해 서사정보 자동 추출 시스템의 성능을 증명하기 위해 등장인물을 대상으로 4장에서 정확률을 측정하고자 한다.

조선대학교 CHOSUN UNIVERSITY

Ⅳ. 실험 및 평가

본 장에서는 설화 내 서사정보 자동 추출 시스템에 대한 성능 평가를 위해 추가적인 설화 데이터를 수집하여 실험 데이터를 구축한다. 성능을 평가하는 방법으로는 서사정보 자동 추출 시스템의 등장인물에 대한 정확률(Precision)을 측정하였다.

또한 실험 결과에 따른 응용방안으로 토픽맵 기반의 서사정보 대한 시각화 방법을 기술함으로써 설화 텍스트로부터 추출된 서사정보간의 관계를 통해 스토리텔링 저작 지원에 활용될 수 있음을 설명한다.

A. 실험 데이터

본 논문의 실험데이터로 3장에서 사용된 '지하국대적퇴치' 설화 이외에 추가적인 수집을 통해 총 6개의 설화 데이터를 수집한 후 각각의 설화에 나오는 등장인물을 분석하여 실험 데이터를 구축한다. [표 4-1]은 각각의 설화에 대해 등장인물과 수를 나타낸다.

[표 4-1] 실험데이터 정보

설화명	등장인물명(등장인물 수)
지하국대적퇴치 설화	아귀, 귀신, 도적, 여자, 임금님, 공주, 신하, 무신, 막내, 하인, 노인, 산신령, 아가씨, 사람, 문지기, 대감님, 우두머리 (17명)
거타지 설화	진성여왕, 막내아들, 사신, 양패, 거타지, 노인, 임금, 궁사, 중, 용, 그녀 (11명)
아기장수 우투리 설화	임금, 벼슬아치, 백성, 영웅, 부부, 아기, 우투리, 장수, 군사, 부모, 어머니 (11명)
지하국의 도적 설화	한량, 부자, 대적, 딸, 무사, 장수, 부모, 처녀 (8명)
노랑대구리 설화	부부, 아이, 부모, 노랑두대구리, 동네사람, 코샘생, 오종소티, 주인, 노인, 아들, 괴물, 삼형제 (12명)
황금 돼지 설화	선비, 딸, 사위, 부자, 처녀, 까치, 황금돼지, 괴물 (8명)

조선대학교 CHOSUN UNIVERSITY

B. 실험 평가 및 결과

1. 실험 평가 방법

본 논문의 실험은 구현한 설화 내 서사정보 자동 추출 시스템에 대한 성능을 평가하기 위해 등장인물 추출에 대한 정확률(Precision)을 이용하였으며 다음 식 (1)과 같다. 정확률은 정보 검색에서 검색 결과에 대한 정확성을 평가하는 측도이며 본 논문에서는 설화 내에서 등장하는 전체 인물의 수 중 구현된 시스템에 의해 추출된 등장인물의 수로 정확률을 측정한다.

$$\text{정확률}(\text{Precision}) = \frac{\text{구현된 시스템에 의해 추출된 등장인물 수}}{\text{설화 내 전체 등장인물 수}} \quad (1)$$

2. 실험 결과

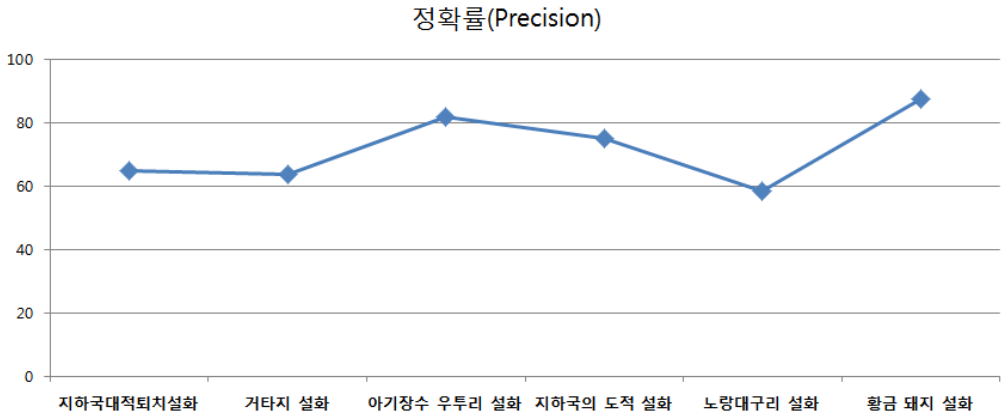
[표 4-2]는 식 (1)을 이용하여 구현된 시스템에 대한 성능을 평가한 결과를 보여준다.

[표 4-2] 서사정보 자동 추출 시스템의 등장인물 추출에 대한 성능

구분 \ 설화명	지하국대적 퇴치 설화	거타지 설화	아기장수 우투리 설화	지하국의 도적 설화	노랑대구리 설화	황금 돼지 설화
설화 내 전체 등장인물 수	17	11	11	8	12	8
시스템에 의해 추출된 등장인물 수	11	7	9	6	7	7
정확률	64.71%	63.64%	81.82%	75%	58.33%	87.5%
평균	71.83%					

조선대학교 CHOSUN UNIVERSITY

서사정보 자동 추출 시스템의 등장인물 추출에 대한 평균 정확률은 71.83%로 기대 이상의 성능을 보여준다. 이는 영어와는 달리 복잡한 형태소와 동음이의어를 가지는 한국어의 특성상 특정 품사를 추출하는데 어려움이 있는 반면, 각 설화별 정확률은 50% 이상으로 TF값 기반 등장인물 추출이 잘 되는 것을 보였다는 점은 매우 흥미로운 결과를 보여준다. [그림 4-1]은 [표 4-2]의 결과를 그래프로 표현하였다.



[그림 4-1] 각 설화별 등장인물 추출에 대한 정확률

각 설화별 등장인물 추출에 대한 정확률 그래프에서는 황금 돼지 설화에 대한 정확률이 87.5%로 가장 높은 정확률을 보여주었다. 이는 구현한 서사정보 자동 추출 시스템을 통해 주인공에 해당하는 등장인물들이 정확히 추출되었고, 해당 등장인물의 빈도수가 타 설화에 비해 월등히 높은 빈도를 보여주었음을 알 수 있다. 그러나 노랑대구리 설화에서는 58.33%의 가장 낮은 정확률을 보여준다. 이는 단어에 대한 의미 분석이 제대로 되지 않아 같은 등장인물을 나타내는 단어의 중복 검출에 대한 문제로 분석된다. 또한 정확한 비교분석을 위해 타 실험과 비교를 해야 하나, 현재까지 한국어 자연어 처리 기반의 스토리 내 서사정보 추출을 위한 연구가 활발히 진행되지 않고, 단지 등장인물의 속성이나 성격 파악 위주의 연구가 진행되다보니 정확한 비교 실험이 어려웠다. 또한, 설화 자체의 데이터양이 대부분 500단어 내외로 매우 적다. 따라서 데이터양에 비례함에 따라 등장인물 추출에 대한 정확률이 설화에 따라 상이함을 실험 결과로써 나타났다.

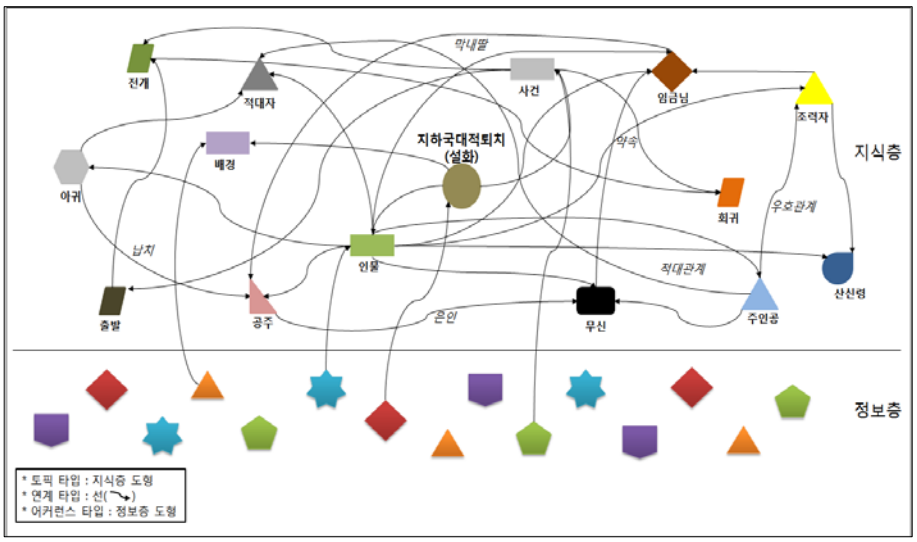
조선대학교 CHOSUN UNIVERSITY

C. 실험 결과에 따른 토픽맵 기반 시각화 응용방안

본 절에서는 설화 텍스트로부터 추출된 서사정보를 중심으로 인물간의 관계 및 사건, 배경에 대한 정보를 토픽맵으로 표현하고자 한다. 이를 위해 영웅 서사구조 기반의 관계 모델을 활용하여 토픽맵으로 구축하였으며, 설화 텍스트 정보를 시각화함으로써 스토리텔링 저작 지원을 위한 효율적인 정보를 제공하고자 그 응용방안을 제시하고자 한다.

1. 서사정보를 이용한 토픽맵 구축

본 논문에서 추출된 서사정보에 대한 시각화를 표현하기 위해 [그림 4-2]와 같은 개념 모델링을 도출하였다. 지식층에는 인물, 사건, 배경 토픽 타입이 존재하며, 인물 토픽 타입 중심으로 주인공, 조력자, 적대자 토픽간의 관계가 설정되어 있고, 사건 토픽 타입 중심으로 영웅 서사구조의 사건정보인 출발, 전개, 귀환 토픽간의 관계가 설정되어 있다. 또한, 배경 토픽 타입 중심으로 현실세계, 초현실세계 토픽간의 관계가 설정되어 있으며, 추출된 서사정보 인스턴스는 각 토픽에 따라 연계 타입을 갖고 있다.



[그림 4-2] '지하국대적퇴치' 설화 개념 모델링

조선대학교 CHOSUN UNIVERSITY

이러한 토픽은 각각의 정보층에 있는 이미지 파일, URL, 특정 데이터베이스 내 레코드를 참조하며 토픽 및 인스턴스에 대한 부가적인 정보를 확장할 수 있다. 예를 들어 인물에 대한 구체적인 속성인 이름, 나이, 성별, 인물 이미지 등의 정보를 인물 데이터베이스 및 웹 상에 존재하는 해당 정보를 참조함으로써 구체적인 설화 데이터에 대한 시각화를 표현할 수 있다.

[표 4-3]은 개념 모델링을 토대로 토픽맵 구현을 위한 구성요소를 도출하였다. 토픽 타입으로는 인물을 포함하는 요소로써 주인공, 조력자, 적대자등과 같은 토픽으로 구성요소를 도출하였으며 인물간의 관계에 따라 적대관계, 조력관계 등의 연계타입을 도출하였다. 또한, 정보층으로부터 설화에 대한 줄거리 및 인물 이미지, 배경 이미지 등의 외부 리소스를 어커런스 타입으로 도출하였다.

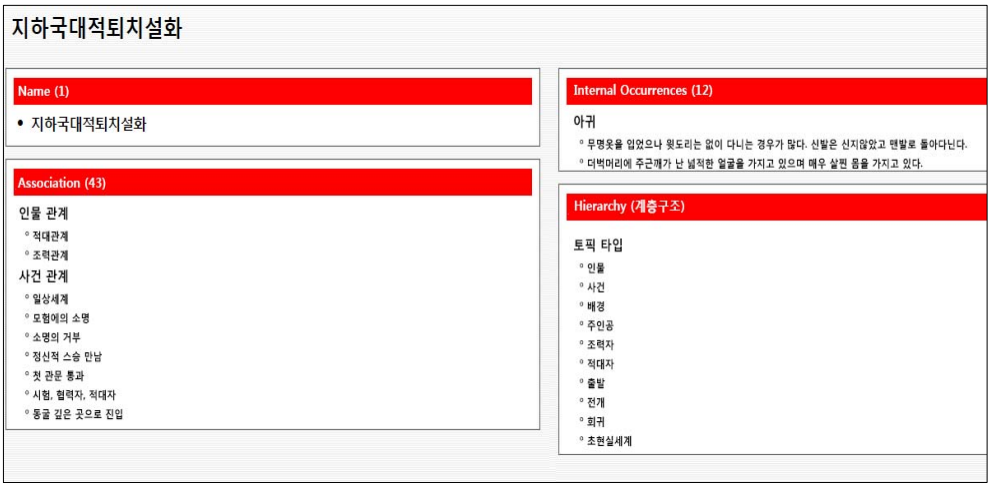
[표 4-3] 설화 데이터 시각화를 위한 구성요소 도출

토픽타입	인스턴스	연계타입	인스턴스
인물	무신	적대관계	주인공 - 적대자
사건	하인	조력관계	주인공 - 조력자
배경	아귀	일상세계	출발 - 일상세계
주인공	산신령	모험에의 소명	출발 - 모험에의 소명
조력자	문지기	소명의 거부	출발 - 정신적 스승 만남
적대자	공주	정신적 스승 만남	출발 - 첫 관문의 통과
출발	산	첫 관문 통과	전개 - 시험, 협력자, 적대자
전개	산모퉁이	시험, 협력자, 적대자	전개 - 시련
회귀	바위	동굴 깊은 곳으로 진입	전개 - 보상
현실세계	땅 속	시련	회귀 - 귀환의 길
초현실세계	…	보상	회귀 - 부활
		귀환의 길	회귀 - 영약을 가지고 귀환
		영약을 가지고 귀환	…

조선대학교
CHOSUN UNIVERSITY

본 논문에서는 설화 텍스트에 대한 시각화를 표현함으로써 실험 결과에 따른 응용방안 제시를 위해 최종적으로 토픽맵 편집기 도구인 온토피아에서 개발한 옴니게이터(Omnigator)도구[24]를 이용하여 토픽, 어소시에이션, 어커런스에 대한 입력 및 편집함으로써 디자인을 구현하였다.

[그림 4-3]은 구현된 설화 데이터 기반의 토픽맵을 나타내며 [표 4-3]의 설화 데이터 구성요소를 기반으로 구축하였다. 설화 데이터의 토픽 타입인 인물, 사건, 배경과 주인공, 조력자, 적대자, 출발, 전개, 귀환, 현실세계, 초현실세계의 토픽은 계층적 구조를 가지고 있으며, 인물 관계 및 배경 관계, 사건 관계를 통하여 각 요소들 간의 의미적 관계를 형성하고 있다. 또한 인물의 구체적인 속성을 표현하기 위해 각 인물에 따른 내부 및 외부 어커런스 참조를 지정하여 시각화 된 설화 데이터를 표현할 수 있다. 예를 들어 '지하국대적퇴치' 설화에서 아귀라는 인물은 단지 주인공의 적대자로써 스토리의 사건을 전개하지만 단지 표면적인 텍스트상의 정보뿐만 아니라 스토리의 새로운 창작에 대한 발상을 위해 아귀귀신에 대한 구체적인 설명을 어커런스를 통해 표현할 수 있도록 하며, 이러한 설화 데이터의 시각화를 토픽맵과 접목하여 스토리텔링 저작 지원을 위한 수단으로 표현할 수 있다.

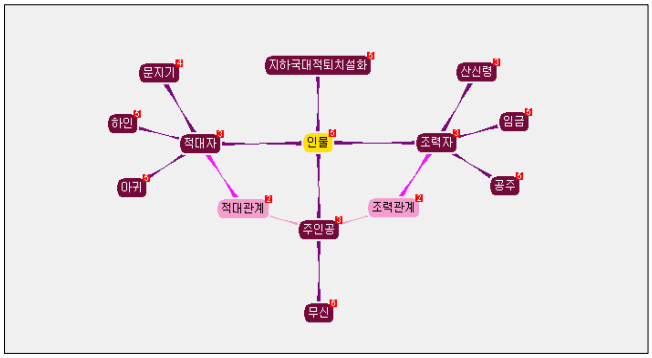


[그림 4-3] 구현된 토픽맵 기반의 설화 데이터

조선대학교 CHOSUN UNIVERSITY

2. 토픽 및 관계 생성

구현된 서사정보 자동 추출 시스템의 서사정보 추출 결과와 [표 4-3]의 시각화 구성요소를 기반으로 스토리텔링 서사구조 모델의 영웅 서사구조 관계모델을 활용하여 실질적인 토픽간의 관계설정을 기술한다. 인물 데이터는 인물의 유형에 따라 적대관계, 조력관계로 관계성을 구축하였으며, 인물과 사건의 관계는 출발, 전개, 귀환 단계의 하위구조에서 인물이 등장하는 사건과 관계성을 구축하였다. 또한, 인물과 사건이 존재하는 공간적 장소인 배경과 관계성을 구축함으로써 설화 텍스트에 대한 서사정보를 시각화 할 수 있었다. [그림 4-4], [그림 4-5], [그림 4-6]는 영웅 서사구조에서 정의한 관계 모델을 바탕으로 '지하국대적퇴치' 설화에서 추출된 데이터를 바탕으로 관계를 생성한 그림이다.



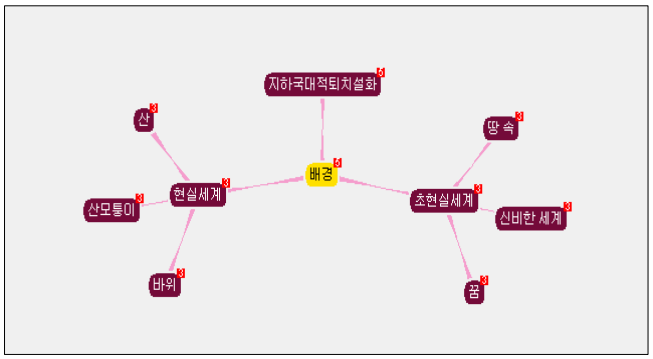
[그림 4-4] 인물 관계 생성

[그림 4-4]는 토픽맵을 이용하여 인물 관계를 표현한 그림이다. 인물 토픽 타입에는 인물의 유형에 따라 주인공, 조력자, 적대자 토픽이 존재하며 주인공을 중심으로 인물간의 관계를 적대관계, 조력관계로 설정 하였다, 주인공 토픽에는 지하국대적퇴치 설화에서 주인공으로 분류된 무신 인스턴스가 존재하고, 적대자 토픽에는 문지기, 하인, 아귀 인스턴스가 존재한다. 또한, 조력자 토픽에는 산신령, 임금, 공주 인스턴스를 갖는다.

배경 토픽 타입은 [그림 4-5]와 같이 현실세계 토픽과 초현실세계 토픽으로 생성하였다. 현실세계 토픽으로는 산, 산모퉁이, 바위 등과 같이 설화 내 등장하는 현실공간에 대한 인스턴스를 갖는다. 초현실세계 토픽으로는 꿈, 아귀소굴, 땅 속등과

조선대학교 CHOSUN UNIVERSITY

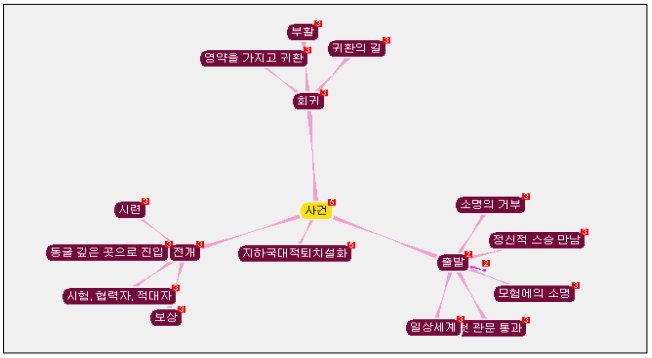
같이 설화에서 등장하는 가상적인 공간으로써 주인공이 초현실세계로 여행을 하면서 인물들이 존재하는 공간적인 요소에 대한 인스턴스를 갖는다.



[그림 4-5] 배경 관계 생성

사건 토픽 타입은 크게 출발, 전개, 귀환의 토픽으로 이루어져 있다. 본 논문에서 대상으로 삼은 '지하국대적퇴치' 설화뿐만 아니라 모든 영웅 서사구조 기반의 설화에 대한 스토리의 구조적 접근이 가능하기 위함이다. 먼저, 출발 토픽은 일상세계, 모험에의 소명, 소명의 거부, 정신적 스승과 만남, 첫 관문 통과의 토픽을 가지고 있으며, 각 단계마다 등장하는 인물과 인물이 존재하는 공간적 요소인 배경 토픽의 인스턴스와 관계가 설정되어 있다.

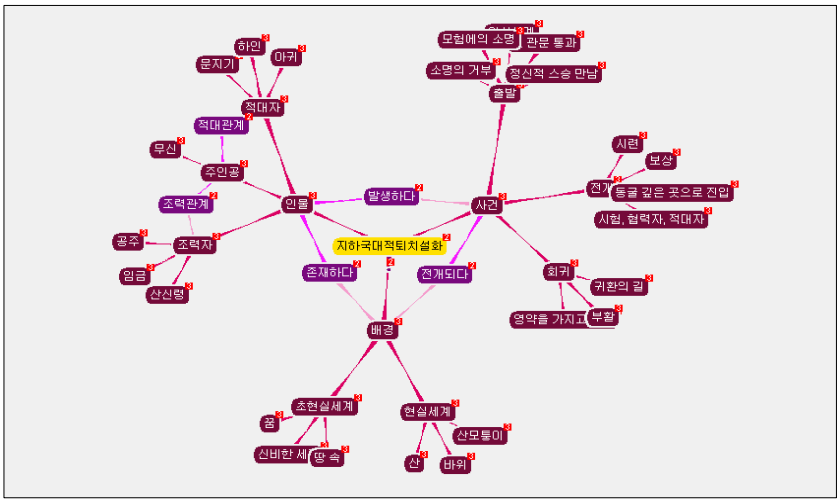
또한, 전개 토픽 타입은 시련, 동굴 깊은 곳으로 진입, 보상 토픽을 가지고 있으며, 귀환 토픽 타입은 귀환의 길, 부활, 영약을 가지고 귀환의 토픽을 갖으며 인물 및 배경 인스턴스를 모두 포함하고 있다.



[그림 4-6] 사건 관계 생성

조선대학교 CHOSUN UNIVERSITY

또한, 최종적으로 서사정보간의 관계를 생성함으로써 스토리 내 핵심요소인 인물, 사건, 배경간의 유기적인 결합을 위한 연계타입(Association)을 정의하도록 한다. 먼저 인물과 사건 토픽 타입의 연계타입을 '발생하다', 사건과 배경의 연계타입은 '전개하다', 인물과 배경의 연계타입을 '존재하다'로 관계성을 구축하였다.



[그림 4-7] 서사정보 관계 생성

이러한 토픽맵 기반의 관계설정을 이용한 시각화는 설화 텍스트로부터 추출된 스토리 내 서사정보 데이터를 가공할 수 있으며, 스토리를 구성하는 각 요소(인물, 사건, 배경)간의 의미적 관계를 생성할 수 있다. 이에 스토리 저작을 위한 작가들에게 보다 효율적인 스토리텔링 저작에 대한 응용방안으로 활용할 수 있음을 나타내고 있다.

조선대학교 CHOSUN UNIVERSITY

Ⅴ. 결론 및 제언

최근 문화콘텐츠에 대한 관심이 증가됨으로써 영화, 공연 관광, 드라마 등 다양한 문화산업에 걸쳐 스토리텔링 기법을 활용하여 부가가치 창출을 위한 노력이 지속적으로 증대되고 있다. 이에 양질의 스토리텔링 콘텐츠 생산을 위해 스토리텔링 저작을 위한 연구가 국외를 중심으로 수행되고 있는 반면 국내에서는 상대적으로 미흡한 실정이다.

이에 본 연구에서는 스토리텔링 저작을 위해 스토리 내 서사정보를 자동적으로 추출하기 위한 시스템을 설계 및 구현하였다. 제안하는 서사정보 자동 추출 시스템은 한국어로 작성된 설화 텍스트 내 명사를 추출하는 단계, 서사정보 후보를 추출하는 단계, 서사정보 목록을 제공하는 단계로 구성하였다. 본 시스템에서 설화 텍스트 내 명사를 추출하는 이유는 서사정보 요소인 인물, 사건, 배경에 대한 품사가 주로 명사에 분포되어 있기 때문이며, 입력된 설화 텍스트에 대한 품사를 파악하기 위해 POS Tagging 과정을 수행하였다. 또한, 기존 형태소 분석기에서 정의하고 있는 명사 형태소에 대한 동일 패턴을 정의하며, TF(Term Frequency) 알고리즘을 이용한 서사정보 후보를 추출하였다. 추출된 서사정보 후보의 품사가 지닌 모호성을 해소하기 위해 불용어 명사 목록을 정의하여 추출된 명사들에 대한 재필터링 과정을 수행함으로써 재분류된 서사정보를 추출할 수 있었다.

구현된 서사정보 자동 추출 시스템의 성능을 평가하기 위해 서사정보 중 등장인물 추출에 대한 정확률을 측정하였다. 실험 데이터는 본문에서 설명한 '지하국대적퇴치' 설화를 포함한 총 6개의 설화 데이터로 진행하였으며 71.83%의 평균 정확률을 보여주었다.

향후 연구로는 자연어 처리 기법의 텍스트 분석을 통한 단어 추출뿐만 아니라 스토리 내 서사정보간의 유기적인 결합을 위해 의미적 관계 생성을 위한 온톨로지 기반의 연구가 필요할 것으로 판단된다. 또한, 추출된 서사정보 데이터에 대한 시각화를 통해 스토리텔링 저작을 위한 작가들에게 창작을 위한 자동화 시스템을 활용함으로써 스토리 구성을 위한 시간적, 인력적 부담을 최소화할 것으로 기대된다.

조선대학교 CHOSUN UNIVERSITY

참고문헌

[1] 나성준, 김우중, 최이권, 전현택, "디지털 스토리텔링의 스토리 핵심 요소를 이용한 기획 지원 온톨로지 검색 시스템 연구", 한국지능정보시스템학회, pp.338-242, 2010.

[2] 안성혜, 허영, "비주얼 스토리텔링의 재미요소 설계", 한국콘텐츠학회 춘계학술대회논문집, 제7권 제1호(상), pp.250-257, 2009.

[3] 문화기술 (CT) 심층리포트, "스토리텔링 저작 도구 연구 동향과 사례분석", 2012.

[4] 권호창, "디지털 스토리텔링에 있어 스토리 생성 시스템 연구", 한국예술종합학교 석사학위논문, 2009.

[5] 라창현, 하얀, "지능형 스토리텔링을 위한 저작에 관한 연구", 한국지능시스템학회 학술발표 논문집, pp.456-459, 2005.

[6] 김석규, 문성현, 박준, 장준호, 한상영, "이야기 생성을 위한 인터랙티브 스토리텔링 스크립트 언어에 관한 연구", 멀티미디어학회논문지 제12권 제2호, pp313-322, 2009.

[7] 박승보, 이예지, 유은순, 정재은 "등장인물 누적에 따른 스토리 발단 부분의 경계 판단", 한국지능정보시스템학회 추계학술대회, pp135-140, 2011.

[8] J.Cambell, The Hero with a Thousand Face, Bollingen, 1972.

[9] 구민호, 민덕기, "반복적 추출 단어 분리를 이용한 핵심 키워드 추출 기법에 관한 연구", 한국정보과학회 가을학술발표집 제36권 제2호(B), pp.329-334, 2009.

[10] 백남주, 고병규, 이은지, 김판구, "연관 검색어 추출을 위한 Triple-Data Pattren 정의에 관한 연구", 한국스마트미디어학회 춘계학술대회, 2014.

[11] http://kkma.snu.ac.kr/

[12] 최주호, 이재영, "전자기록물의 메타데이터 추출 및 비교 검증 기술 연구", 한국기록관리학회지, 제12권 제1호, pp.7-32, 2012.

[13] http://kldp.net/projects/hannanum/

[14] 유정훈, "워드넷 기반의 클러스터를 이용한 문서 요약 시스템 : 스포츠 뉴스 기사 헤드라인 추출", 단국대학교 석사학위논문, 2013.

[15] 오삼균, "A Comparison of Ontology Language : Focusing on W3C OWL and ISO Topic Maps"

[16] Ontopia.net "The TAO of Topic Maps." http://www.ontopia.net/topicmaps/mat

조선대학교 CHOSUN UNIVERSITY

erials/ta.html.

[17] 이재호, “시맨틱 웹의 온톨로지 언어”, 정보과학회지, 제21권 제3호, pp.18-27, 2003.

[18] Pepper, Steve and Garshol, Lars Marius, The XML Papers : Learned on Applying Topic Maps, 2002.

[19] 박여삼, 장옥배, 한성국, "X-TOP : 레거시 시스템상에서 온톨로지 구축을 위한 토픽맵 플랫폼의 설계와 구현", 정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 14권 제2호, pp.130-142, 2008.

[20] Jack Park and Sam Hunting, XML Topic Maps : Creating and Using Topic Maps for the Web, Addision-Wesley, 2003.

[20] 오삼균, 박옥남, “토픽맵 기반 판소리 검색시스템 구축 및 평가에 관한 연구”, 한국도서관·정보학회지, 제36권 제4호, pp.77-98, 2005.

[21] 권진만, 정명범, 성보경, 김정수, 고일주, “토픽맵기반의 뮤지션 검색시스템 구축”, 한국HCI학회 학술대회, pp.760-765, 2008.

[22] 이지수, 정겨운, 이경원, “스토리텔링의 시각화를 위한 개념적 맵들의 특성분석”, 한국HCI학회, pp1214-1219, 2008.

[24] http://www.ontopia.net/