(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년08월27일
(11) 등록번호 10-2699233
(24) 등록일자 2024년08월22일

(51) 국제특허분류(Int. Cl.)
G06F 40/169 (2020.01) *G06F 40/106* (2020.01)
G06F 40/186 (2020.01) *G06F 40/205* (2020.01)
G06F 40/289 (2020.01) *G06F 40/30* (2020.01)
G06N 3/04 (2023.01) *G06N 3/08* (2023.01)
G06N 5/04 (2023.01)
(52) CPC특허분류
G06F 40/169 (2020.01)
G06F 40/106 (2020.01)
(21) 출원번호 10-2022-7011501
(22) 출원일자(국제) 2020년07월24일
심사청구일자 2023년07월21일
(85) 번역문제출일자 2022년04월06일
(65) 공개번호 10-2022-0059526
(43) 공개일자 2022년05월10일
(86) 국제출원번호 PCT/US2020/043606
(87) 국제공개번호 WO 2021/055102
국제공개일자 2021년03월25일
(30) 우선권주장
62/900,793 2019년09월16일 미국(US)
(56) 선행기술조사문헌
US08209278 B1*
(뒷면에 계속)

(73) 특허권자
도큐가미, 인크.
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218
(72) 발명자
비건, 앤드류
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
드로즈, 스티븐
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
(뒷면에 계속)
(74) 대리인
특허법인 무한

전체 청구항 수 : 총 20 항 심사관 : 김경완

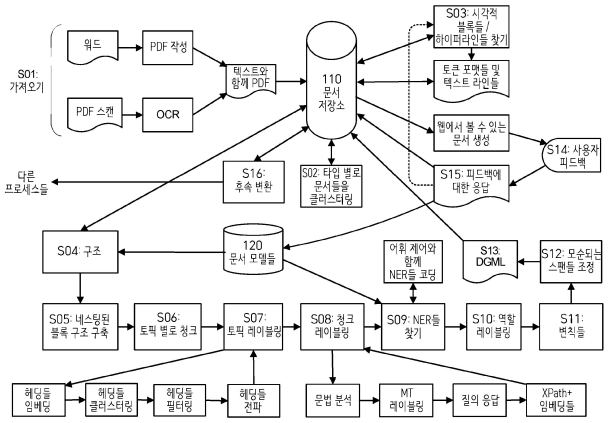
(54) 발명의 명칭 **문서 간 지능형 저작 및 처리 보조기**

(57) 요 약

기계 학습, 인공지능 및 다른 컴퓨터 구현 방법들이 문서들에서의 의미론적으로 중요한 다양한 청크들을 식별하고 이들을 적절한 데이터 타입들 및 의미론적 역할들로 자동으로 레이블링하고 이 향상된 정보를 이용하여 저작자들을 보조하고 후속 프로세스들을 지원하기 위해 사용된다. 청크 위치들, 데이터 타입들 및 의미론적 역할들

(뒷면에 계속)

대 표 도



이, 소위 '컨텍스트'라 불리는 것, 즉 좀 더 정확히 말해서 이들의 포맷팅, 구조 및 컨텍스트의 조합, 인접하거나 이웃하는 컨텐츠의 것들, 문서에서의 전반적인 출현 패턴들 및 문서들에 걸쳐서의 이 모든 것들의 유사성들(주로 동일 문서 세트에서의 문서들 중에서 이기는 하나 전적으로 그렇지는 않음)로부터 종종 자동으로 결정될 수 있다. 유사성은, 정확한 또는 퍼지 스트링 또는 속성 비교들에 한정되지 않고, 자연어의 문법적 구조, 워드, 청크 및 기타 임베딩들의 유사성을 측정하는 것과 같은 ML(machine learning) 기법들 및 이전에 식별된 청크들의 데이터 타입들 및 의미론적 역할들의 유사성을 포함할 수 있다.

(52) CPC특허분류
G06F 40/186 (2020.01)
G06F 40/205 (2020.01)
G06F 40/289 (2020.01)
G06F 40/30 (2020.01)
G06N 3/045 (2023.01)
G06N 3/088 (2023.01)
G06N 5/04 (2023.01)

(72) 발명자
자프리, 타키
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
마르티, 루이스
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
팔머, 마이클
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
파올리, 진
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
파블로포울로우, 크리스티나
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
프리코이우, 엘레나
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
사랑이, 스와가티카
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
사윅키, 마르신
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
쉐하데, 마나
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
타론, 마이클
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
토프라니, 바벤
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
와디아, 주빈, 루스톰
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
왓슨, 데이비드
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
화이트, 에릭
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.
팬, 조슈아, 용쉰
미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.

(56) 선행기술조사문헌
US09760634 B1*
US20070136221 A1*
US20120005686 A1*
US20150139610 A1*
US20170147910 A1*
*는 심사관에 의하여 인용된 문헌

굽타, 커쉬

미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.

호앙, 앤드류, 민

미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.

리우, 잔린

미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.

팔리악카라, 제롬, 조지

미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.

우, 자오펭

미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.

장, 유에

미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.

조우, 챠오콴

미국 워싱턴 98033 커클랜드 넘버 921 메인 스트리트 218 도큐가미 인크.

명 세 서

청구범위

청구항 1

문서들을 분석하고 개선하기 위한 명령어들을 실행하는 컴퓨터 시스템 상에서 구현되는 방법에 있어서,

복수의 문서들을 포함하는 문서 세트에 접근(access)하는 단계 - 상기 문서 세트는 또한 상기 문서 세트의 개별 문서들 내에서의 청크들을 식별함 -, 및

복수의 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계

를 포함하고,

상기 의미론적 역할 레이블들은,

상기 문서에 의하여 기술되는 거래(transaction) 내에서 청크들이 하는 의미론적 역할들을 나타내고,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

(a) 청크들에 대한 의미론적 역할들을 결정하기 위해 기계 학습 및/또는 자연어 처리 방법들을 사용하는 단계를 포함하고,

(b) 또한 상기 문서 세트 내 서로 다른 문서들에 있는 카운터파트 청크들(counterpart chunks)의 출현 패턴에 기초하고,

상기 카운터파트 청크들은,

다른 문서들에 있는 다른 청크들이면서, 각각의 문서들 내에서 동일한 의미론적 역할을 하는 것인,

상기 방법은,

상기 문서 세트에서의 문서들을 추가로 처리하는데 있어서, 상기 청크들 및 그 의미론적 역할 레이블들을 사용하는 단계

를 더 포함하는, 방법.

청구항 2

제1항에 있어서,

상기 문서 세트에서의 상기 복수의 문서들은 모두 동일 문서 유형인,

방법.

청구항 3

제1항에 있어서,

상기 문서 세트에서의 청크들은,

문서 템플릿들에서 필드들로서 사용하기에 적합한 상기 문서들 내에서의 컨텐츠를 포함하는 필드 청크들 - 상기 필드 청크들의 일부는 계층적이고 다른 청크들을 서브청크들로서 포함함 -, 및

상기 문서들의 레이아웃 내에서의 구조들을 포함하는 컨텐츠를 포함하는 구조 청크들

을 포함하는, 방법.

청구항 4

제1항에 있어서,

상기 문서 세트는,

법률 문서들을 포함하고,

상기 의미론적 역할들은,

(a) 상기 법률 문서들에 대한 당사자들이 하는 역할들, 및

(b) 날짜들, 시간 기간들(time periods) 또는 시간에 대한 다른 표현들이 하는 역할들

을 포함하는, 방법.

청구항 5

제1항에 있어서,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

상기 의미론적 역할 레이블들 중 일부를 청크들로부터 자동으로 추출하는 단계, 및

상기 추출된 의미론적 역할 레이블들을 청크들에 할당하는 단계

를 포함하는, 방법.

청구항 6

제1항에 있어서,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

(a) 개별 문서들에 있는 청크들의 컨텐츠, 레이아웃 및 컨텍스트들에 기초하고, (b) 상기 문서 세트에서의 문서들에 걸쳐 있는 청크들의 컨텐츠, 레이아웃 및 컨텍스트들의 패턴들에 기초하여, 그리고 (c) 청크들의 데이터 타입들에 기초하여, 청크들로부터 의미론적 역할 레이블들을 자동으로 추출하기 위해 기계 학습을 사용하는 단계, 및

상기 추출된 의미론적 역할 레이블들을 청크들에 할당하는 단계

를 포함하는, 방법.

청구항 7

제1항에 있어서,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

상기 의미론적 역할 레이블들의 일부를 자동으로 추출하기 위해 자동 인코더 기계 학습 기법들을 사용하는 단계, 및

상기 추출된 의미론적 역할 레이블들을 청크들에 할당하는 단계

를 포함하는, 방법.

청구항 8

제1항에 있어서,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

상기 청크들로부터 후보 의미론적 역할 레이블들을 자동으로 추출하는 단계,

상기 후보 의미론적 역할 레이블들을 정제(refine)하기 위해 기계 학습을 사용하는 단계, 및

상기 추출된 의미론적 역할 레이블들을 청크들에 할당하는 단계

를 포함하는, 방법,

청구항 9

제1항에 있어서,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

상기 문서 세트에서의 서로 다른 문서들로부터의 청크들의 컨텐츠, 레이아웃 및/또는 컨텍스트의 유사성에 기초하여 청크들로부터 상기 의미론적 역할 레이블들의 일부를 자동으로 추출하는 단계, 및

상기 추출된 의미론적 역할 레이블들을 청크들에 할당하는 단계

를 포함하는, 방법.

청구항 10

제1항에 있어서,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

후보 의미론적 역할 레이블들을 청크들에 할당하는 단계,

청크들을, 상기 청크들이 하는 의미론적 역할들의 유사성에 기초하여 클러스터들로 군집화하는 단계,

클러스터들에 있는 상기 청크들 사이에서 상기 후보 의미론적 역할 레이블들을 표준화하는 단계, 및

상기 표준화된 의미론적 역할 레이블들을 청크들에 할당하는 단계

를 포함하는, 방법.

청구항 11

제1항에 있어서,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

후보 의미론적 역할 레이블들을 청크들에 할당하는 단계,

청크들을, 상기 청크들의 크기 및 텍스트 임베딩(text embedding)의 유사성에 기초하여 청크 클러스터들로 군집화하는 단계,

후보 의미론적 역할 레이블들을, 상기 후보 의미론적 역할 레이블들의 텍스트 임베딩의 유사성에 기초하여 레이블 클러스터들로 군집화하는 단계, 및

상기 청크 클러스터들 및 상기 레이블 클러스터들에 기초하여 상기 후보 의미론적 역할 레이블들을 표준화하는 단계, 및

상기 표준화된 의미론적 역할 레이블들을 청크들에 할당하는 단계

를 포함하는, 방법.

청구항 12

제1항에 잇어서,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

문서들의 절들(sections)을 포함하는 청크들에 후보 의미론적 역할 레이블들을 할당하는 단계 - 상기 후보 의미론적 역할 레이블들은 상기 절들의 헤딩들에 기초함 -,

상기 절들에서의 컨텐츠의 유사성에 기초하여 상기 청크들을 클러스터들로 군집화하는 단계,

가장 공통의 후보 의미론적 역할 레이블(most common candidate semantic role label)을 클러스터에 있는 모든 청크들을 위한 의미론적 역할 레이블로서 선택함으로써 상기 후보 의미론적 역할 레이블들을 표준화하는 단계, 및

상기 표준화된 의미론적 역할 레이블들을 청크들에 할당하는 단계

를 포함하는, 방법.

청구항 13

제1항에 있어서,

상기 의미론적 역할 레이블들은,

미리 결정된 의미론적 역할 레이블들의 세트(predetermined set of semantic role labels)로부터 선택되는,

방법.

청구항 14

제1항에 있어서,

상기 의미론적 역할 레이블들은,

상기 문서 세트에 있는 문서들의 추가 처리를 위해 사용되는 소프트웨어 애플리케이션에 의해 인식되는 레이블들을 포함하는,

방법.

청구항 15

제11항에 있어서,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

(a) 바로 이웃하는 다른 청크들(nearby chunks)에 기초하여 또는 상기 청크들을 포함하는 함유 청크들(containing chunks)에 기초하여 청크들에 대한 의미론적 역할들을 결정하기 위해 기계 학습을 사용하는 단계, 또는

(b) 청크들에 대한 의미론적 역할들을 결정하기 위해 바로 이웃하는 청크들의 문법적 구조들에 기초하여 자연어 처리 방법들을 사용하는 단계

중 적어도 하나를 포함하는,

방법.

청구항 16

제11항에 있어서,

상기 청크들의 일부는,

개체명 참조들(Named Entity References)이고,

그러한 청크들은, 상기 문서들에서 그러한 청크들이 하는 의미론적 역할들을 위한 의미론적 역할 레이블들로 레이블링되고,

그러한 청크들은 또한, 상기 청크의 데이터 타입으로 레이블링되는,

방법.

청구항 17

제11항에 있어서,

상기 청크들의 일부는,

상기 문서들에서 다단락 구조들(multi-paragraph structures)이고,

그러한 청크들은, 상기 문서들에서 그러한 청크들이 하는 의미론적 역할들을 위한 의미론적 역할 레이블들로 레이블링되는,

방법.

청구항 18

제11항에 있어서,

상기 자동으로 할당되는 의미론적 역할 레이블들에 대한 신뢰 수준(confidence level)을 추정하는 단계,

상기 추정된 신뢰 수준에 기초하여 확인(confirmation)을 위해 사용자에게 약간의 과제들(some assignments)을 제시하는 단계,

상기 자동으로 할당된 의미론적 역할 레이블들에 대한 사용자 피드백을 수신하는 단계, 및

상기 사용자 피드백에 응답하여 상기 기계 학습 및/또는 자연어 처리 방법들을 개선하는 단계

를 포함하는, 방법.

청구항 19

비일시적 컴퓨터 판독가능 저장 매체에 있어서,

문서들을 분석하고 개선하기 위한 실행가능한 컴퓨터 프로그램 명령어들을 저장하고,

상기 명령어들은,

컴퓨터 시스템에 의해 실행가능하며, 상기 컴퓨터 시스템으로 하여금,

복수의 문서들을 포함하는 문서 세트에 접근하는 단계 - 상기 문서 세트는 또한 상기 문서 세트의 개별 문서들 내의 청크들을 식별함 -,

복수의 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계, 및

상기 문서 세트에 있는 문서들의 추가 처리를 위해 상기 청크들 및 그 의미론적 역할 레이블들을 이용가능하도록 하는 단계

를 포함하는 방법을 수행하도록 하고,

상기 의미론적 역할 레이블들은,

상기 문서에 의하여 기술되는 거래(transaction) 내에서 청크들이 하는 의미론적 역할들을 나타내고,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

(a) 청크들에 대한 의미론적 역할들을 결정하기 위해 기계 학습 및/또는 자연어 처리 방법들을 사용하는 단계를 포함하고,

(b) 또한 상기 문서 세트 내 서로 다른 문서들에 있는 카운터파트 청크들의 출현 패턴에 기초하고,

상기 카운터파트 청크들은,

다른 문서들에 있는 다른 청크들이면서, 각각의 문서들 내에서 동일한 의미론적 역할을 하는 것인,

비일시적 컴퓨터 판독가능 저장 매체.

청구항 20

문서들을 분석하고 개선하기 위한 컴퓨터 시스템에 있어서,

복수의 문서들을 포함하는 문서 세트를 수신하고 저장하기 위한 저장 매체 - 상기 문서 세트는 또한, 상기 문서 세트의 개별 문서들 내의 청크들을 식별함 -, 및

상기 저장 매체에 대한 접근을 가지고, 문서를 분석하고 개선하기 위한 애플리케이션 프로그램을 실행하는 프로세서 시스템

을 포함하고,

상기 애플리케이션 프로그램을 실행하는 프로세서 시스템은,

복수의 청크들에 의미론적 역할 레이블들을 자동으로 할당하고,

상기 문서 세트에 있는 문서들의 추가 처리를 위해 상기 청크들 및 그 의미론적 역할 레이블들을 이용가능하도

록 하고,

상기 의미론적 역할 레이블들은,

상기 문서에 의하여 기술되는 거래(transaction) 내에서 청크들이 하는 의미론적 역할들을 나타내고,

상기 청크들에 의미론적 역할 레이블들을 자동으로 할당하는 단계는,

(a) 청크들에 대한 의미론적 역할들을 결정하기 위해 기계 학습 및/또는 자연어 처리 방법들을 사용하는 단계를 포함하고,

(b) 또한 상기 문서 세트 내 서로 다른 문서들에 있는 카운터파트 청크들의 출현 패턴에 기초하고,

상기 카운터파트 청크들은,

다른 문서들에 있는 다른 청크들이면서, 각각의 문서들 내에서 동일한 의미론적 역할을 하는 것인,

컴퓨터 시스템.

청구항 21

삭제

청구항 22

삭제

청구항 23

삭제

청구항 24

삭제

청구항 25

삭제

청구항 26

삭제

청구항 27

삭제

청구항 28

삭제

청구항 29

삭제

청구항 30

삭제

청구항 31

삭제

청구항 32

삭제

청구항 33

삭제

청구항 34

삭제

청구항 35

삭제

청구항 36

삭제

청구항 37

삭제

청구항 38

삭제

청구항 39

삭제

청구항 40

삭제

청구항 41

삭제

청구항 42

삭제

청구항 43

삭제

청구항 44

삭제

청구항 45

삭제

청구항 46

삭제

청구항 47

삭제

청구항 48

삭제

청구항 49

삭제

청구항 50

삭제

청구항 51

삭제

청구항 52

삭제

청구항 53

삭제

청구항 54

삭제

청구항 55

삭제

청구항 56

삭제

청구항 57

삭제

청구항 58

삭제

청구항 59

삭제

청구항 60

삭제

청구항 61

삭제

청구항 62

삭제

청구항 63

삭제

청구항 64

삭제

청구항 65

삭제

청구항 66

삭제

청구항 67

삭제

청구항 68

삭제

청구항 69

삭제

청구항 70

삭제

청구항 71

삭제

청구항 72

삭제

청구항 73

삭제

청구항 74

삭제

청구항 75

삭제

청구항 76

삭제

청구항 77

삭제

청구항 78

삭제

청구항 79

삭제

청구항 80

삭제

청구항 81

삭제

청구항 82

삭제

청구항 83

삭제

청구항 84

삭제

청구항 85

삭제

청구항 86

삭제

청구항 87

삭제

청구항 88

삭제

청구항 89

삭제

청구항 90

삭제

청구항 91

삭제

청구항 92

삭제

청구항 93

삭제

청구항 94

삭제

청구항 95

삭제

청구항 96

삭제

청구항 97

삭제

청구항 98

삭제

청구항 99

삭제

청구항 100

삭제

발명의 설명

기 술 분 야

[0001] 본 출원은, 2019년 9월 16일 출원된, 발명의 명칭이 '교차 문서 지능형 저작 및 처리 보조기'(Cross-Document Intelligent Authoring and Processing Assistant)인 미국 가특허출원 제62/900,793호에 대해 미국 특허법(35 U.S.C.) 제119(e)조에 의거 우선권을 주장한다. 위 출원의 모든 사항의 주제는 전적으로 본 명세서에 참조로 병합된다.

[0002] 본 개시는 일반적으로 계층적으로 의미론적으로(semantically) 레이블링된(labeled) 서류의 AI 자가 지도(self-supervised) 생성 및/또는 그러한 서류의 보조 저작(assisted authoring)과 처리(processing)를 위한 방법 및 장치에 관한 것이다.

배 경 기 술

[0003] 많은 비즈니스(many businesses)에서 비록 맞춤형이기는 하지만 상당히 유사한 다수의 문서가 생성된다. 예컨대, 보험회사는 특정 종류의 보험을 위해 많은 제안서들을 생성하나, 그 각각의 제안서는 특정 고객의 요구에 맞춰져야 한다. 이러한 문서들은 동일한 '유형'(type)이라 여겨질 수 있는데, 그 이유는 이들이 (유사한 목적들과 토픽들을 반영하는) 유사한 텍스트 컨텐츠(text content)를 (그리고 혹시 유사한 이미지 컨텐츠까지도) 가지고 있고, 절(sections)과 같은 대 단위들(large units)의 유사한 선집(選集, selections)과 배열(arrangements)을 가지고 있고, 그리고 이따금 유사한 기하학적 레이아웃(geometric layout)과 서식 특성(formatting characteristics)조차 가지고 있기 때문이다.

[0004] 몇몇 유형의 문서들이 널리 알려져 사용되고 있으나 많은 수의 문서들이 그러한 것은 아니다. 많은 수의 문서들이 특정의 사무, 시장 또는 응용분야에 특화되어 있고, 새로운 것들이 새로운 상황들을 위해 생성된다. '저작자'(authors) 또는 '편집자'(editors)로 불려지는 사용자들은, 대개 동일 유형의 초기 문서를 복제하고, 예컨대 컨텐츠의 특정 청크들(chunks)을 수작업으로 편집하거나 대체함으로써 필요한 변경을 가함에 의해 (이따끔 '타겟 문서들'이라 불리는) 특정 유형의 신규 문서들을 생성한다.

[0005] 현행의 실무에서는, 포맷팅(formatting)을 하기 위해 필요한 경우에만 워드 프로세싱(word processing)이 전형적으로 청크들을 식별한다. 즉, 예를 들어 헤딩들(headings), 풋노트들(footnotes) 및 그림들(figures)은 특수 포맷팅(special formatting)을 얻기 위해 명시적으로 마킹될 수도 있지만, 이름들(names), 주소들(addresses) 또는 날짜들(dates)은 드물게 명시적으로 마킹된다. 청크들은, 식별되는 경우에조차 유효한 정보인 (마진(margins), 폰트(fonts) 및 등등과 같은) 포맷팅 효과들(formatting effects)과만 대개 연관되어 있으나, 그들의 데이터 타입들(data types)이나 의미론적 역할들(semantic roles)의 어떠한 표시도 직접적으로 제공하지 않는다. 마찬가지로 워드 프로세서들(word processors)은 종종 계층적 포함관계(hierarchical containment)를 시각적으로만 나타낸다. 즉, 중첩 절들(nested sections) 자체의 명시적인 표현이 종종 없고 다르게 포맷팅된 헤딩들(headings)의 명시적인 표현만이 있을 뿐이다.

[0006] 이전의 문서들과 동일한 일반적인 종류의 신규 문서를 생성할 때에, 많은 경우에 있어서의 작업의 대부분은 특

정한 청크들의 삽입, 제거, 대체 또는 텍스트 편집인데, (구매자와 판매자의 주소들을 서로 바꾸는 것과 같은) 서로 다른 의미론적 목적들을 가지는 것들을 혼동하지 않도록 주의해야 한다. 이는 사람이 개입하는 것을 요하는데, 그 이유는 저작 시스템들(authoring systems)이 전형적으로 이러한 청크들, 특히 그들의 데이터 타입들 또는 의미론적 역할들에 대해 아무것도 알지 못하고 있고 따라서 매우 효과적으로 도움을 줄 수 없기 때문이다.

[0007] 몇몇 간단한 경우들에 있어서는 '서식들'(forms) 및 '템플릿들'(templates)을 사용할 수 있는데, 이들은 특정 청크들에 대해 컨텐츠를 기입해 넣을 명시적인 위치들(places)을 제공한다. 그러나 서식들은, 실질적으로 필요한 모든 청크들을 미리 셀 수 있고, 대형의, 반복가능한 또는 극히 구조화된 소수의 청크들이 있는, 간단한 경우들 만을 전형적으로 해결해 준다. 서식들은 또한 생성하기 위해 숙련된 노력을 요하고 변화하는 환경에 따라 조정하기가 어렵고 작가를 적극적으로 보조하지 않는다.

도면의 간단한 설명

[0008] 본 특허 또는 출원 파일은 컬러로 작성된 적어도 하나의 도면을 포함한다. 컬러 도면(들)을 가지는 이 특허 또는 특허 출원 공보의 복사본들은, 신청 및 필요한 비용 납부를 조건으로 청에 의해 제공될 것이다.

본 개시의 실시예들은, 이하의 상세한 설명과 첨부된 청구항들을 첨부된 도면들에서의 예들과 결부하여 읽을 때 더욱 명백해질 여타의 장점들과 특징들을 가진다.

도 1은 계층적으로 의미론적으로 레이블링된 문서들을 기계 학습 및 인공지능을 이용하여 생성하기 위한 시스템 및 프로세스의 일 구현예의 블록도를 도시한 도면이다.

도 2는 도 1의 시스템을 통해 서로 다른 문서 세트들을 처리하는 것을 추적하는 대시보드(dashboard)를 나타내는 스크린샷(screenshot)을 도시한 도면이다.

도 3은 사용자로부터 피드백을 수신하기 위한 사용자 인터페이스의 스크린샷을 도시한 도면이다.

도 4는 다른 소프트웨어 애플리케이션들과 통합한 것에 대한 스크린샷을 도시한 도면이다.

도 5는 본 발명과 함께 사용될 수 있는 컴퓨터 시스템의 일 실시예의 블록도를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0009] **개요**

[0010] 동일한 유형이라고 판정되는 일 군의 문서들은 '문서 세트'(document set) 또는 '문서 클러스터'(document cluster)를 구성한다. 예를 들어, 특정 부류의 고객들을 위한 특정 종류의 보험에 대한 보험회사의 제안서들은 동일한 유형이라고 여길 수 있고 문서 세트를 구성한다. 다른 종류의 보험에 대한 동일한 회사의 제안서들 또는 다르다고 여겨지는 고객들에 대한 제안서들은 다른 문서 세트에 속하는 다른 유형으로 여겨진다. 임대 계약서(rental agreements), 특정 부류의 환자에 해한 진료 기록들, 판매 제안서들, 계획서들, 미팅 기록들 등은 있을 수 있는 다른 유형의 문서들이고, 컨텐츠, 구조 및/또는 레이아웃의 상이한 패턴들을 공유하는 서브 타입들도 그러하다.

[0011] 문서 세트 내에서 신규 타겟 문서들을 생성하고 편집하는 것은, '의미론적으로 중요한'(semantically important) '청크들'(chunks)을 편집하고 대체하는 것을 수반한다. 이러한 청크는, 반드시 연속된 범위의 텍스트는 아니지만 특정의 데이터 타입(datatype) 및 의미론적 역할을 가지고 있고 비즈니스 또는 다른 프로세스 용으로 의미가 있고 중요한 문서의 특정 부분이다.

[0012] 이러한 청크들은, 많은 컴퓨터 시스템에서의 아토믹 데이터 타입(atomic datatypes)보다 훨씬 결이 고운(fine-grained) 다양한 데이터 타입들이다. 예를 들어, 주어진 청크는, 단지 스트링(string)뿐만이 아니라 개인 또는 기관 이름(personal or organizational name), 날짜, (날짜와 전혀 같은 것이 아닌) 시간 구간(duration of time), 통화량(current amount)을 나타낼 수 있다. 큰 청크들은, 약품들 또는 기타 물질들의 목록들, 여행 계획들, 따라야 할 절차들, 의료 처방전과 같은 정보 번들(bundles of information) 및 셀 수 없이 많은 것들을 포함할 수 있다.

[0013] 뿐만 아니라, 청크들은, 그들이 출현하는 문서와 관련하여 의미론적 역할들을 가질 수 있다. 예를 들어, 개인 이름은, 임대차 계약서(lease agreement)에서는 '차용자'(tenant)일 수 있고, 판매 제안서에서는 '판매자'일 수 있고 또는 다른 사람의 '대리인'(agent)일 수 있다. 날짜는 어떤 의무 또는 활동의 시작 또는 끝을 나타낼 수 있다. 달러 금액은, 정기 결제 금액 또는 특정 조건들에 관련된 위약금 또는 보너스 등일 수 있다. 이러한 의

미론적 역할들은 청크들에서의 정보를 적합하게 사용하는 데 있어서 중요하다. 의미론적 역할들에 대한 이름들을 '의미론적 역할 레이블들'(semantic role labels) 또는 간단히 '레이블들'이라 부른다.

[0014] 청크는, 그 위치, 데이터 타입, 의미론적 역할 및/또는 기타 데이터/메타 데이터를 포함하는 번들(bundle)로서 전형적으로 표현될 수 있다. 위치는, 삽입된 마커들(markers) 또는 바이트, 캐릭터 또는 (문서에 전역적이거나 아니면 설정된 ID들, 마커들 또는 기타 객체들에 대해 상대적인) 토큰 오프셋들(token offsets)과 같이 다양한 방식으로 표현될 수 있는 시작 및 종료 점으로서 대개 표현된다. 의미론적 역할들은, 레이블 또는 기타 식별자에 의해 표현된다. 청크들은, 임의의 크기일 수 있고, 몇몇 일부 청크들은 다른 청크들을 서브 청크들(sub-chunks)로서 포함할 수 있다. 청크들은, 텍스트뿐만 아니라 이미지들 또는 기타 미디어와 같은 텍스트가 아닌 데이터(non-text data) 및 테이블들(tables), 목록들(lists), 절들(sections) 등과 같은 '구조들'(structures)을 포함할 수 있다.

[0015] 본 명세서에 개시되는 기술들은, 문서들에서의 의미론적으로 중요한 다양한 청크들을 식별하고 그들에게 적절한 데이터 타입들과 의미론적 역할들을 자동으로 제공하고 이렇게 보강된 정보를 사용하여 저작자들을 보조하고 후속 프로세스들을 지원하기 위해 기계 학습(machine learning), 인공지능(artificial intelligence) 및 기타 컴퓨터 구현 방법들을 사용한다. 청크의 위치들, 데이터 타입들 및 의미론적 역할들은, 본 명세서에서 '컨텍스트'(context)라 부르는 것, 즉 좀 더 구체적으로 말하자면 그들의 포맷팅(formatting), 구조(structure) 및 컨텐츠(content)의 조합, 인접하는 또는 이웃하는 컨텐츠의 그것들, 문서에서의 전반적인 출현 패턴들(patterns of occurrence) 및 (주로 동일한 문서 세트에서의 문서들 중에서이나 전적으로 그런 것은 아닌) 문서들에 걸쳐서의 이 모든 것들의 유사성으로부터 종종 자동으로 결정될 수 있다. '이웃하는 컨텐츠'(nearby content)는 텍스트를 읽는 순서에 있어서 앞서고 뒤따르는 것과 같이 수평적으로 가까이 있는 컨텐츠뿐만 아니라 목록들(lists) 및 절들(sections) 및 그들의 마커들(markers), 헤딩들(headings), 레벨들(levels) 등과 같은 동일한 컨테이너 구조들(container structures) 내에 있는 것과 같이 수직적으로 가까이 있는 컨텐츠를 포함한다. 유사성은, 정확한 또는 퍼지 스트링 또는 속성 비교(exact or fuzzy string or property comparisons)에 한정되지 않고, 자연어의 문법적 구조, 워드, 청크 및 기타 임베딩들(embeddings)의 유사성을 측정하는 것과 같은 ML(machine learning) 기법들 및 이전에 식별된 청크들의 데이터 타입들 및 의미론적 역할들의 유사성을 포함할 수 있다.

[0016] 예를 들어, 개인 또는 기관의 이름은 대개 '판매자'와 같이 의미론적 역할을 가지는 것으로서 식별될 수 있는데, 그 이유는 문서가 몇몇 인간의 언어로된 문장을 종종 이용하여 그러나 더 큰 컨텍스트를 종종 수반하면서 그렇게 말해주고 있기 때문이다. 다른 예를 들면, 하나 또는 둘 이상의 단어들이, 종종 '약물 이름'(medication name)과 같은 데이터 타입을 쉽게 표현하는 것으로서 식별될 수 있으나, 그것이 처방(prescription)이라기 보다는 앨러지(allergy)의 의미론적 역할을 가지고 있다고 결정하기 위해서는 컨텍스트가 필요하다. 종종 의미론적 역할의 증거는 동일 문장에 있지 않고 ('알려진 앨러지들'(Known Allergies)이라는 절(section)과 같은) 더 큰 청크에 출현하는 청크에 의해서와 같이 다양한 다른 방식들로 표현된다. (타이포 에러(typos), 등사 에러(transcription errors) 등은 말할 것도 없이) 문법과 문서의 구조적 체계의 유연성과 다양성은, 데이터 타입을 식별하는 것을 덜 어렵게 하나, 의미론적 역할들, 특히 단일 문장보다 더 큰 범위(scope)를 갖는 것들을 식별하는 것을 매우 어렵게 한다.

[0017] 주어진 의미론적 역할은, 청크를 문서에 전체적으로 관련시킬 수 있고 다른 청크들에 관련시킬 수도 있다. 예를 들어, 출발 비행 시간(departure time of a flight)은, 여정에 있어서의 하나의 특정 여행 구간(leg)에 관련되어 있고 다른 것들에는 간접적으로만 관련되어 있다. 대체로 청크들의 계층적 구조는, 절들(sections), 테이블 부분들 등 내의 콜로케이션(co-location)과 같은 항목들을 적절히 함께 군집화(grouping)한다.

[0018] 더욱 상세히 설명하면, 의미론적으로 중요한 청크들의 예들은, 이름들, 주소들 및 계약에 대한 특정 당사자들의 기타 특성들, 의료 기록에 있어서의 처방된 약물 및 처방된 절차, 부동산 제안서들에서의 요구 사항들 (또는 배제 조항들(exclusions)), 여정에 있어서의 날짜들 및 항공편 번호들 등등을 포함한다. 이러한 것들은 모두 청크들의 의미론적 역할들로 여겨질 수 있다. 절들(sections) 및 소구분들(subsections)과 같이 다양한 타입들과 역할들을 가진 더 큰 청크들도 있다. 이러한 것들은, 아마도 그 안의 더 작은 청크의 변경들과 함께 전체적으로 삽입되거나 삭제된다. 청크들은 계층적이다. 즉, 더 큰 '포함적'(containing) 청크들은 다른 '서브 청크들'(sub-chunks)을 임의의 개수의 레벨들까지 포함할 수 있다.

[0019] 청크는 일반적으로, "John Doe"와 같이 문서 내에서의 일련의 연속된 단어들이다. 그러나 청크들은 부분적인 단어들을 포함할 수 있다. "john Doe's house"는 이름을 포함하나 그 이름은 (단어 중간의) 소유격 부호 이전

에 끝난다. 청크들은, 예를 들어 동일한 이름 "John ('Bill'이라고도 알려짐) Doe"에서와 같이 불연속적일 수도 있다. 레이아웃에 의해 청크들이 불연속적이 될 수 있는데, 예컨대 페이지 나누기(page break)가, (아마도, 어떤 목적상 무시될 수 있는 페이지 헤더들(headers), 푸터들(footers) 또는 풋노트들(footnotes)을 가진) 청크, 사이에 끼여있는 그림, 테이블, 챠트, 사이드 바 또는 기타 디스플레이 등의 중간에 출현할 수 있다.

[0020] 청크의 실제 위치와 컨텍스트가 중요할 수 있다 - 청크는, 서로 다른 경우들에 서로 다른 의미론적 역할들을 가지고 또는 아무런 의미론적 역할도 가지지 않고 여러 번 출현할 수도 있는 독립된 스트링(isolated string)이 아니다. 좀 더 최신식의 시스템들은, 다양한 레이블들과 기타 정보를 청크들과 지속적으로 연관시킬 수 있는, 때때로 '주석'(annotation)이라 불리는 인라인의(inline) 또는 스탠드오프(standoff) 마크업(markup)을 전형적으로 지원한다. 예를 들어, HTML은 일반적인 구조 청크들("div", "ol" 등)의 경계들 및 (전형적으로) 더 작은 청크들("acronym", "kbd", "dfn", "cite" 등)에 대한 몇몇 광범위한 타입들 및 역할들을 수작업으로 레이블링하기 위한 태그들(tags)을 제공한다. 기타 XML 스키마들(schemas)은 많은 다른 레이블들을 제공하고, 워드 프로세서들은 '스타일'(styles)을 통해 다소 유사한 레이블링을 가능하게 한다.

[0021] 몇몇 청크들은, 일반적으로 '필드'(fields)라 불리는 것을 나타낼 수 있다. 이들은 종종 작은 청크들이고 주어진 세트에서의 많은 또는 모든 문서들에서의 유사한 컨텍스트들 및 레이아웃들에서 종종 출현하나 각각이 서로 다른 텍스트 컨텐츠를 가진다. 이들은 단일 문서에서 동일한 또는 매우 유사한 컨텐츠를 가지고 복수회 출현할 수도 있다. 그러한 청크들은 '필드 청크'(field chunks)라 불릴 수 있다. 이들은 템플릿 기반의 시스템(template-based systems)에서 종종 '필드'(fields)로서 수작업으로 발견되고 취급될 수 있으나, 여기서는 문서들 내에서 그리고 문서들에 걸쳐 그들의 컨텍스트들과 출현 패턴들에 의해 발견되고 다른 청크들과 매우 유사한 방식으로 데이터 타입들과 의미론적 역할들이 할당(assign)된다. 그들은, 개인 이름들, 주소들, 날짜들 등과 같은 개체명들(Named Entities)을 나타낼 수도 있고 나타내지 않을 수도 있다.

[0022] 다른 흔한 청크 유형은 '구조'(structure) 또는 '구조적'(structural) 청크라 불려질 수 있다. 그러한 청크들은, 전형적으로 더 크고, 일부가 구조 청크들(structural chunks)일 수도 있는) 다른 많은 청크들을 종종 포함한다. 이들은, 구조 청크에 관한 이름, 번호, 설명 및/또는 기타 정보를 제공하는 '타이틀'(title) 또는 '헤딩'(heading)을 흔히 가진다. 구조 청크들의 예들은, 장들(chapters), 절들(sections), 테이블들, 그림들, 사이드바들(sidebars) 및 그외의 다수를 포함한다. 구조 청크들의 타입들과 의미론적 역할들은, 이웃하는 또는 포함된(contained) 청크들의 타입들과 의미론적 역할들을 결정하기 위해 종종 중요하다.

[0023] 데이터 타입들뿐만 아니라 특정의 의미론적 역할들은, 문서를 적절히 만들고 이용하기 위해 중요하다. 어떤 이름이 구매자와 판매자 중 누구를 나타내는지 또는 환자와 의사 중 누구를 나타내는지, 주어진 날짜가 요건의 시작 또는 끝 중 어느 것을 나타내는지 또는 비행의 출발 시간과 도착 시간 중 어느 것을 나타내는지, 숫자(number)가 원금(principal), 이자(interest), 투약량(dosage), 온도, 위약금(penalty) 또는 여타의 것 중 어느 것을 특정하는 것인지가 상당히 문제가 된다. 더 큰 청크들에 대해서는, 역할들은, '책임 제한'(limitations on liability)에 관한 스테이트먼트(statement) 대(versus) '준거법'(governing law) 명세 대 '정의'(definition) 및 그 밖의 무수히 많은 것들과 같은 것들을 포함한다. 청크들의 의미론적 역할들은, 종종 특정 영역들(domains) 또는 거래들(transactions)에 특유하고 어쩌면 문서들의 가장 중요한 특징들 중 하나이다. 많은 종류의 문서들에서 특정 데이터 타입들과 의미론적 역할들을 가지는 청크들이 요구되고 적어도 매우 흔하며, 청크들은 문서들에 걸쳐 대응하는 경우 '카운터파트들'(counterparts)이라 불려진다. 카운터파트 청크들(counterpart chunks)은, 특히 동일한 저작자 또는 기관에 의한 그리고 동일한 문서 세트의 문서들의 경우 유사한 순서와 패턴으로 출현될 수 있다. 카운터파트 청크들은 동일 또는 매우 유사한 역할들을 가지고, 대개 유사한 컨텍스트 및/또는 포맷팅을 가지고 있다. 따라서 청크의 데이터 타입들과 의미론적 역할들의 분포는, 문서들의 유형들을 구별할 뿐만 아니라 다른 문서들에서 카운터파트 청크들을 식별하는데 있어 도움을 주기 위한 귀중한 정보를 제공한다.

[0024] 많은 카운터파트 청크들은 유사한 컨텐츠를 가지나 다른 것들은 그렇지 않다. 예를 들어, 서로 다른 문서들에서의 동일한 당사자(동일한 의미론적 역할을 가지는 것)는 보통 다른 개인인데, 그렇지만 매우 유사한 컨텍스트들 및 사용 패턴들에서 등장한다. 이는, '필드 청크들'(field chunks)에 대해서만 전적으로 그런 것은 아니지만 하여간 필드 청크들에 대해서 특히 일반적이다.

[0025] 비즈니스 문서들에서 데이터 타입들과 의미론적 역할들을 가지는 계층적 의미론적 청크들은, 일단 발견되면 후속 비즈니스 프로세스들(downstream business processes)에서 사용될 수 있다. 예를 들어, 백오피스(back-office) 데이터베이스는, 특정 당사자들의 이름들, 특정 날짜들 및 기간 및 이율 등과 같은 숫자들이 주어지면

신규 저당권을 적절히 등록할 수 있다. 특히 그러한 용도로 의미론적 역할들이 매우 중요한데, (판매자 및 구매자의 이름들과 주소들을 서로 바꾸는 것과 같이) 정확한 데이터 타입을 틀린 데이터베이스 필드에 넣는 것은, 특히 정보를 후속 데이터베이스, 프로세스 또는 보고서로 이동시킬 때 큰 문제가 된다.

[0026] **몇몇 특징들과 이점들**

[0027] 본 명세서에서 설명되는 기술들은, 이하의 것들을 포함하는 다양한 특징들과 이점들을 가질 수 있다.

[0028] 몇몇 구현예들은, 비즈니스 목적상 유용한 의미론적 레이블링(semantic labelling)을 갖는 계층적으로 체계화된 청크들을 가지는 문서들을 생성하기 위한 더욱 쉽고 더욱 효율적이고 더욱 정확한 방법을 제공할 수 있다. 이는, 그러한 다양한 크기의 청크들을 식별하고, 데이터 타입들과 이들이 문서에서 하는 의미론적 역할들을 발견하고, 이들의 사용 패턴들, 특징적 컨텍스트들 등을 학습하기 위한 다양한 기술들을 이용하여 성취될 수 있다. 학습은, 현재 및 이전의 문서들의 컨텐츠, 구조 및 포맷팅의 분석, 저작자들 및 편집자들로부터의 피드백 및 복수의 문서, 특히 동일 문서 세트에 있는 문서들의 비교를 통해 이루어질 수 있다. 이러한 지식을 가지고 시스템은, 예컨대 높은 품질의 신규 문서들의 용이한 생성, 백오피스 데이터베이스에서 다른 소프트웨어 애플리케이션들과 함께 후속하여 사용할 원하는 정보를 추출하는 것, 파생 보고서(derived reports), 준수 검사(compliance checking) 등과 같이 사용자들에게 귀중한 지원을 제공할 수 있다. 이러한 학습은, 많은 양의 미리 레이블링된 또는 미리 분석된 데이터를 요하지 않으나 레이블링되지 않은 또는 최소한으로 레이블링된 데이터로부터 패턴들을 추론하는, 자율(unsupervised) 및 자가(self-supervised) 학습 기법들을 가지고 이루어질 수 있다.

[0029] 몇몇 구현예들은, 비즈니스 문서들 내에서 그리고 이들에 걸쳐 패턴들을 발견하고 사용하여 작가들로 하여금 이러한 많은 오류들을 피하도록 해주고 그리하여 주어진 품질 수준을 성취하기 위해 요구되는 시간을 줄임으로써 컴퓨터들로 하여금 저작 과정에서 도움을 주도록 할 수 있다.

[0030] 요즈음 전형적인 문서 시스템들은, 청크들 또는 특히 그들의 데이터 타입들 또는 의미론적 역할들을 식별하지 않는다. 이는 저작자들 또는 편집자들에게 시간과 비용을 부담시키고 문서들로부터의 데이터를 백엔드 데이터베이스(back-end databases), 데시보드 또는 여타 후속 비즈니스 프로세스들로 이동시키는데 추가의 시간과 비용을 들이게 한다. 예를 들어, 계약서들에서 데이터를 수작업으로 (청크 하나하나씩) 찾아 스프레드 시트(spreadsheets) 또는 데이터 기입 폼들(data-entry forms)로 복사하는 것이 일반적이다.

[0031] 몇몇 구현예들은, 저작 과정 동안 그러한 계층적 의미론적 청크들을 레이블링하여 그들을 명시적으로 나타내도록 지원해줌으로써 사람들 및/또는 컴퓨터들이 추출하기 쉽게 해주고 다양한 종류의 여타 비즈니스 프로세스들에 연결하는 데 있어서의 시간과 비용을 절약할 수 있게 된다.

[0032] 현재의 기술은, 신규 문서들에서 청크들을 더욱 신뢰성있게 식별하기 위해 또는 중요할 것 같은 차이들에 주의를 환기시키기 위해 이를 신호하기 위해 (특정 문서 세트에서 멤버쉽으로 나타낸) 동일 작가 또는 그룹에 의해 생성된 및/또는 동일 유형인 복수의 문서(plurality of documents) 간의 유사성들을 전형적으로 최대한으로 이용하지는 않는다. '잔존 규정'(severability)이라고 헤딩이 붙은 절(section)을 요구하는 것과 같은 명시적인 규칙들은, 분석가들이 쉽게 보고 설명하는 유사성들을 포괄할 뿐이고, 정적이고 종종 제약적이고(constraining)(예컨대, 리프레이징과 재구성을 가진 케이스들을 누락시키는 또는 상쇄적 조건들(countervail conditions)에 응답하지 못하는) 그리고 빠르게 구식이 되어간다. 소기업들은 종종 좀 더 민감한 기술을 개발하기 위해 필요한 자원들이 없고 비용을 정당화하기에는 너무 적은 수의 문서들을 가지고 있다. 반면에 더 작은 소기업들은, 이하에서 기술하는 것과 같은 자동화된 분석(automated analysis)에 더욱 적합화된, 덜 다양한 범위에 문서들을 종종 가지고 있다.

[0033] 몇몇 실시예들은, 작가들이 신규 문서들을 생성하는데 있어서 도움을 주기 위해 청크들과 그들의 컨텐츠 패턴, 컨텍스트 패턴, 레이아웃 패턴 및 문서들에 걸친 사용 패턴에 관한 추출된 정보를 이용할 수 있다. 예들은, 변경하거나 다시 포맷하거나 이동시킬 특정 컨텐츠, 유사한 문서들에는 공통으로 존재하지만 신규 문서에는 누락되어 있는 절들(clauses)('누락된'(missing) 또는 '생략되었을 수 있는'(possibly-omitted) 청크들 또는 컨텐츠라 불려짐), 유사한 문서들에는 공통으로 없으나 있는 절들(clauses)('이례적인'(unusual) 청크들 또는 컨텐츠), 특정 위치들에서 서로 다른 당사자들의 이름들과 역할들을 서로 바꾸는 것과 같은 변경 등을 적어도 제안하는 것을 포함한다.

[0034] 몇몇 구현예들은, 사용자가, 청크가 부정확한 범위, 데이터 타입 또는 의미론적 역할로 레이블링되었다거나 그들에게 관심거리가 아니거나 전혀 레이블링되지 못하였음을 나타내는 것 같이, 사용자 피드백을 수용하거나 유

지할 수 있다. 몇몇 피드백은, (추가의 학습이 특정 경우의 실수를 방지하지 못한다 할지라도) 사용자가 제안들을 거절한 경우에는 그러한 초기 제안들을 반복하지 않도록 기억하는 것뿐만 아니라 기계 학습 및 신경 모델들(neural models)을 개선하기 위해 특정 사용자 정정들을 사용할 수 있다. 특히 몇몇 실시예들은, 필요한 사용자 행동의 양을 최소화하기 위해, 퓨샷 학급 기법들(few-shot learning techniques) 및 어떤 피드백을 요청할 것인지에 대한 신중한 선택을 위해 많은 수의 검토 단계들 또는 정정들을 요구하는 것을 회피할 수 있다. 몇몇 현재의 기술은, 예컨대, 사용자가 스펠 검토기(spelling checker)에게 그 사전에 단어를 추가하라고 말할 때와 같이 극히 특정의 것들을 학습한다. 그러나 이는, 단순 로트 목록들(rote lists)을 수반하고, 정교한 이후 행동을 결정하기 위해 사용되는 모델들의 미세 조정(fine-tuning of models)이나 반복적 학습(iterative training)이 아니고, 따라서 본 명세세에 기술된 것과 같은 기능들을 전적으로 이용하지 않는다.

[0035] 몇몇 실시예들은, 개선되었으나 여전히 불완전한 모델들이 재적용되는 경우 반복된 제안들로 사용자들을 성가시게 하는 것을 피하면서 사용자들의 행동을 학습하고 개선시키기 위해 적은 개수의 사용자 정정들을 사용할 수 있다.

[0036] 많은 비즈니스들은, 그들의 프로세스들을 지원하는 다양한 종류의 데이터베이스들에 문서들로부터 획득한 특정 정보를 기록한다. 예컨대 많은 임대 재산(rental properties)을 소유한 회사는, 단지 임차인의 지불(renter's payments) 뿐만이 아니라 인정되는 애완동물, 임차인의 책임이 아닌 이전의 손상 부분(prior damage) 또는 기타 정보와 같은, 그들의 임대계약(rental agreements)에서 근원되는 특정 정보를 관리하기 위해 백엔드 시스템(back-end system)을 전형적으로 사용한다. 자동차 또는 도구 임대인들, 저당 회사들, 건강 관리 제공자들, 지방자치단체들(municipalities) 및 기타 단체들은 다른 정보를 사용한다. 많은 상품 및 서비스들은, 무수한 믹스 앤 매치(mix-and-match) 옵션들을 가지며, 관리자들은 그들의 인수(acceptance), 조합(combinations), 가격(pricing) 및 기타 요소들에 대한 통계를 검토한다. 비즈니스 정보 시스템들은, 분석을 제공하고, 일관성(consistency) 검사 또는 준법감시(compliance)를 수행하고, 보고서를 얻고 및/또는 기타 비즈니스 프로세스들을 지원하는데, 이들 모두는 본 명세서에서 기술하는 청크 정보의 사용을 통해서 용이하게 이루어질 수 있다.

[0037] 일반적으로 청크들 또는 그들이 제공하는 정보는 프로즈 텍스트(prose text)에 걸쳐 산재되어 있고, 수작업으로 추출되고, 스프레드시트, 데이터베이스 또는 기타 시스템들로 수작업으로 입력된다. 이전에는 수작업이 필요했는데, 그 이유는, 계약서들, 이메일들 및 그와 유사한 것들이 작성된 자연어들(natural human languages)의 유연성 및 유사한 가변적인 레이아웃 및 표현 규칙들(variable layout and representation conventions) 때문에 중요한 청크들이 무수히 다양한 방식으로 표현될 수 있기 때문이다. 그러한 문서들의 저변에 깔려있는 협상들(negotiations)은, 이메일들, 대화들로부터의 기록들(notes from conversations), 슬라이드 발표물들(slide presentations) 등을 포함하는 복수의 종류의 문서들에 걸쳐 또한 종종 산재되어 있다. 그러한 정보는 유용할 수 있으나 전형적으로 수작업으로 다루어진다. 몇몇 시스템은 그러한 정보 소스들을 문서들로서 취급할 수 있고, 이미 기술한 것과 동일한 장점을 얻는다.

[0038] 몇몇 구현예들은, 특정 문서가 계층적으로 의미론적으로 레이블링된 문서가 되기 위해 본 명세서에서 기술하는 것과 같이 변환되면, 컴퓨터로 하여금 이 특정 문서를 실행하는 것을 시작하도록 하는 수단을 제공할 수 있다. 문서의 계층적으로 레이블링된 구조를 텍스트의 벡터-의미론적 표현들(vector-semantic representations)을 제공하는 도구들과 결합함으로써, 특정의 청크들이 특정의 행동들을 요구하는 것으로서 식별될 수 있다. 예컨대 계약서는 송금(money transfers), 통지(notifications) 또는 여타 조치들(actions) 및 이들을 가능케하거나 트리거하는 조건들을 특정할 수 있다.

[0039] 몇몇 구현예들은, 문서 세트들로부터의 정보를 검토하고 '대시보드'(dashboards)와 같은 인터페이스에 해당 정보를 요약하고 식별된 정보를 고객의 백엔드 데이터베이스 또는 유사한 시스템들로 이동시킴으로써 더욱 효율적이고 비용이 덜 드는 비즈니스 데이터 흐름을 가능하게 하고 품질 보증(quality assurance), 일관성(consistency) 및 보고서 작성 작업(reporting)을 강화할 수 있는 쉬운 방법들을 제공할 수 있다. 청크들이 의미론적으로 레이블링되면, 문서들의 세트들에 걸쳐 카운터파트 청크들(counterpart chunks)을 포함하는 요약 보고서들을 생성하기가 더 쉬워진다. 몇몇 실시예들은, 사용자들로 하여금 포함될 청크들의 하나 또는 둘 이상의 예들을 단순히 클릭함으로써 그러한 보고서들을 생성하도록 하는 매우 쉬운 방법들을 제공할 수 있는데, 여기서 이들 청크들은 그 후 세트에 있는 모든 문서들에 걸쳐 찾아서 역할 또는 컨텍스트 별로 추출된다. 몇몇 실시예들은, 예상되는 카운터파트 청크들이 없는 문서들을 찾아서 이들이 그러한 청크들을 포함하거나 식별하도록 정정하거나 이들이 그러한 청크들을 포함하지 않는 것이 올바르다는 것을 확인하는 데에 있어서 사용자들을 보조할 수 있다.

[0040] 다른 특징에 있어서, 회사 또는 부서와 같은 주어진 그룹에 대한 성과가, 청크의 의미론적 역할들, 출현 패턴들(patterns of occurrence) 및 그들의 문서들의 기타 특성들 및 그들의 사용자들의 피드백과 같은 정보를 시스템의 학습 프로세스들로 합체시킴으로써 그리고 향후의 문서들을 보강하고 및/또는 체크하기 위한 결과적으로 개선된 모델들을 이용하여 보강될 수 있다. 그러나 많은 고객들은 다른 고객들과 공유되는 그러한 정보를 원하지 않고, 많은 고객들은 구속력 있는 비밀유지 요구조건들(binding confidentiality requirements)을 가진다. 한편 공공의 비밀이 아닌 소스들로부터 유도된 일반 정보 및 학습이 사용되고 자유롭게 공유될 수 있다.

[0041] 몇몇 실시예들은, 각 고객의 데이터와 그로부터 유도된 임의의 모델 정보를 각 고객에게 개별적이고 사적으로 유지하면서 그리고 비밀이 아닌 공공의 데이터에 기초한 일반적인 정보를 여전히 공유하면서 피드백과 학습의 장점들을 제공할 수 있다. 그러한 데이터 프로세스들을 개별적으로 유지함으로써 한 고객으로부터 다른 고객으로 정보가 통계적으로조차 누설될 수 없게 된다.

[0042] **예시적 구현예에 대한 소개**

[0043] 다음은 예시적 시스템에 관한 설명이다. 도 1을 참조하기 바란다. 본 시스템은, 일반적으로, 계층적으로 의미론적으로 레이블링된 문서들의 AI 자가지도 생성(self-supervised creation)을 위한 및/또는 그러한 문서들의 보조된 저작 및 처리를 위한 방법들 및 장치에 관한 것이다. 이는, 문서들을 만들고(composing), 구조화하고(structuring), 주석 처리하고(annotating), 수정하고, 검토하고, 그로부터 데이터를 추출하고 및/또는 그러한 정보를 후속 비즈니스 프로세스들(downstream business processes)에서 이용하는 것과 같은 프로세스들을 포함한다. 더욱 구체적으로, 제각기의 역할들과 연관된, 의미론적으로 의미있는 많은 청크들로 구성된 문서들의 상세한 계층적 구조를 발견하기 위해 주로 자율 및 자가 기계 학습 기법들을 비교적 작은 세트들을 포함하는 문서들의 세트들에 걸쳐 사용함으로써 이전 문서들에 유사한 문서들과 그러한 고도로 향상된 문서들의 비즈니스 프로세스들에서의 사용을 중점적으로 다룬다.

[0044] 본 예시적 시스템의 동작은, 이하의 절들(sections)에서 더욱 상세히 설명되는 이하의 프로세스들을 이용한다. 이는 단지 예일뿐이다. 다른 구현예들은, 단계들을 생략하는 것, 다른 단계들을 추가하는 것 및 몇몇 단계들의 순서를 변경하는 것을 포함하는 단계들의 서로 다른 조합들을 사용할 수 있다. 이들은, 각 단계 하에 기술되는 기법들의 서로 다른 조합들을 포함하는, 이하에서 나열되는 단계들의 서로 다른 구현들을 사용할 수도 있다. 도 1에서 단계들 앞에는 "S"가 있고, 따라서 단계 1에는 S01이라는 레이블을 붙이고 나머지에 대해서도 이러한 식으로 레이블을 붙인다.

[0045] 1) 가져오기(import): 사용자들의 문서들의 군들을 데이터 저장소(110)로 가져온다.

[0046] 2) 편성하기(organize): 문서들을, 임대 대(vs) 판매 계약들 또는 병력(medical histories) 대 현재의 임상 기록(current clinical notes)과 같이 타입에 의해 문서 세트들(document sets)로 나눈다.

[0047] 3) 시각적으로 추출하기(visual extraction): 상이한 텍스트 및 기타 영역들, 그들의 시작 및 종료 위치들, 포맷들(formats) 및 컨텐츠들에 관한 한정된 정보를 비롯하여 그 제각기의 컨텐츠 및 시각적 레이아웃에 적어도 기초하여 각각의 문서로부터 선형적 텍스트 스트림(들)을 추출한다. 추출된 데이터는, 기하학적 레이아웃에 의해 구분되는 문단들과 같은, '시각적 라인들'(visual lines) 또는 '시각적 블록들'(visual blocks)('하이퍼라인들'(hyperlines) 또는 '시각적'(visual) 청크들이라고도 함)로서 편성될 수 있다.

[0048] 4) 구조(structure): 문서들에서의 헤딩들(headings), 리스트 항목들(list items) 및 다른 광범위한 부류의 구조 청크들을 식별한다.

[0049] 5) 리네스팅(re-nesting): 절들(sections) 및 리스트들(lists) 및 각 텍스트의 범위의 네스팅 관계들(nesting relationships)을 결정한다.

[0050] 6) 토픽 청킹(topic chucking): 각 문서의 토픽 컨텐츠(topical content)를 분석하고 유사한 토픽의 영역들을 둘러싸는 청크들(토픽 레벨 청크들)을 생성한다.

[0051] 7) 토픽 레이블링(topic labelling):

[0052] i) 코퍼스(corpus)에서의 각각의 헤딩에 대한 후보 데이터 타입 및 의미론적 역할 레이블들을 생성하기 위해 임베딩(embeddings)과 클러스터링(clustering)을 사용한다.

[0053] ii) 청크들에 대한 후보 데이터 타입과 의미론적 역할 레이블들을 생성하기 위해 키 프레이즈 추출 기법들(key phrase extraction techniques)을 사용한다.

[0054] 8) 청크 레이블링(chunck labelling): 예를 들어 신경망(neural networks), 단어 및 캐릭터 임베딩들(word and character embeddings), 문법적 분석 및 패턴 매칭, 정규 표현들(regular expressions), 유사 메트릭(similarity metrics) 및/또는 다른 방법들과 같은 복수의 방법들을 이용하여, 문사들에 걸쳐 (아마도 복수인) 데이터 타입들과 의미론적 역할 후보들을 식별하고 다른 청크들에 할당한다. 특정 실시예들에 대해 특별히 관심있는 것들은 다음과 같은 것들이다.

[0055] i) 결과적 구조들에 대한 문법적 파싱(grammatical parsing) 및 패턴 매칭

[0056] ii) 작은 청크들을 문서들에서 그들이 하는 특정 의미론적 역할들과 연결시키기 위해 질문 응답 기술들(question-answering technologies)을 사용한다.

[0057] iii) 프레이징(phrasing)과 단어 선택에 있어서 있을 수 있는 광법위한 차이들에도 불구하고 구조 및 문법적 트리들(trees)에 있어서의 패턴들을 매칭하기 위해 XPath 트리 매칭(XPath tree-matching)을 워드 임베딩 기술(word-embedding technology)과 결합한다.

[0058] 9) 개체명 인식(Named Entity Recognition: NER): 데이터 타입들을 식별하고 문서들을 통해 개체명들(Named Entities)로서 검출되는 청크들에 이들을 할당한다.

[0059] 10) 역할 레이블링(role labeling), 추출성의 레이블링(extractive labeling): 이름이 계약에 대한 '판매자'(seller) 측을 구성한다거나 약(drug)이 앨러지(allergy) 대(vs.) 처방(prescription)으로서 언급된다고 나타내는 것과 같이 의미론적 역할 레이블들을 청크들에 할당한다.

[0060] 11) 변칙(anomalies): 고려 대상의 문서 세트의 문서들에 일반적으로 있거나 없으나 현재의 문서에서는 그렇지 않은 (또는 그 반대인) 의미론적 역할들을 식별한다.

[0061] 12) 조정(arbitration): 대안적 범위들, 데이터 타입들 및 청크들에 대한 의미론적 역할 레이블들 가운데서 조절하고 및/또는 선택하여 XML과 같은 포맷들로 쉽게 표현가능한 잘 형성된 구조들(Well-Formed structures)을 생성한다.

[0062] 13) DGML: 청크 위치들, 데이터 타입들 및 의미론적 롤 레이블들 및 각각의 식별된 청크의 신뢰 수준(confidence level), (날짜, 날짜 범위, 개인 이름 등과 같은) 유사한 청크들에서 예측되는 데이터 타입 등등과 같은 아마도 추가적일 수 있는 정보의 명시적인 식별을 포함하는 문서의 향상된 버전을 생성한다. 향상된 버전은 DGML이라 불리는 XML 기반의 마크업 언어를 이용하여 생성된다.

[0063] 14) 피드백(feedback): 사용자(들)에서 향상된 버전을 표시하고 청크들 (및 생략되었을 수 있는 청크들(possibly-omitted chunks)에 대한 잠재적 위치들)을 선택하여, 확인하고 부인하고 그리고 다른 변경을 할 사용자의 선택을 수집한다. 사용자들은 그들 자신의 읽기 및 검토 순서를 자유롭게 선택할 수도 있다. 피드백은, 문서들을 단계(2)에서 기술한 문서 세트들로 편성하는 것과 같이, 시스템이 한 임의의 다른 해석들(interpretations)에 적용될 수도 있다.

[0064] i) 생략되었을 수 있는 청크들의 경우에, 조사되고 및/또는 원하는 대로 현재의 문서로 복사될 수 있는 그리고 더 작은 네스팅된 청크들(nested chunks)에 대해 타겟 문서 값들(target-document values)을 적용함으로써 자동으로 주문형으로 맞춰질 수 있는 다른 문서들로부터 우선되는 예들(prioritized examples)을 제공한다.

[0065] 15) 피드백 응답(feedback response): 이러한 상호작용들에 대한 사용자들의 응답들을 추적하고 그 정보를 이용하여 모델들(120)을 미세 조정할 뿐만 아니라 동일 또는 유사한 에러들을 후에 반복하는 것을 방지한다.

[0066] 16) 하향 통신, 전송(downstream communication, transmit): 타입 및/또는 역할에 따라 청크들을 선택하고 이들을 이용하여 문서 세트들에 걸쳐 보고서들을 생성하고 및/또는 이들을 백엔드 계약 데이터베이스들, 준법감시 체커들(regulatory compliance checkers), 관리 보고서 생성기들(management report generators) 등등과 같은 기능을 더하는 하향 시스템들에 보낸다.

[0067] 도 2는 상술한 프로세스를 통해 서로 다른 문서 세트들을 처리하는 것을 추적하는 대시보드를 나타내는 스크린샷을 도시한 도면이다. 이 대시보드에서 프로세스는 다음의 스테이지들(stages)로 분할된다.

[0068] - 업로딩(uploading)

[0069] - 전처리(preprocessing)

[0070] - 큰 청크들 검토(review large chunks)

[0071] - 작은 청크들 검토(review small chunks)

[0072] - 사용 준비(ready to use)

[0073] 컬러 코딩은 완성도(degree of completion)를 나타낸다. 초록 스테이지는 완성된 것이고, 붉은 스테이지는 처리중인 것이고, 검은 스테이지는 시작되지 않은 것이다.

[0074] 위에 나열된 단계들의 각각은 이하에서 더욱 상세히 기술된다.

[0075] **예시적 구현예의 추가적 설명**

[0076] 본 명세서에서의 번호붙이기(numbering)는 이 특정 예에 대한 분석의 일반적인 순서를 반영한다. 그러나 모든 단계가 모든 이전 단계에 의존하지는 않으며, 결과적으로 다른 구현예들에서는 많은 구성요소들의 순서를 바꿀 수 있고 이들을 병렬화(parallelize)시킬 수 있다. 구성요소들은, 다른 구성요소들과 추가의 정보를 교환하기 위해 시프트되거나 반복되기 조차 할 수 있고, 또는 구성요소들은, 별도의 프로세스들 또는 머신들에서와 같이 독립적으로 실행될 수 있다.

[0077] 1) 가져오기(import)

[0078] 시스템은 (MS 워드와 같은) 전형적인 워드 프로세서 문서들 및 (PDF 또는 .png 파일들과 같은) 페이지 레이아웃 문서들을 받아들인다. 각각의 경우에 헤딩들, 단락들, 테이블 셀들, 테이블, 이미지들 및 그 밖의 것들과 같은 시각적으로 연속적인 영역들이, 이들의 상대적 위치들, 주위의 여백, 폰트 및 레이아웃 특성들 등등의 조합을 이용하여 식별되고 청크들로서 표현된다. 이러한 특징들은 설계자들에 의해 부분적으로 선택되고 많은 수의 문서들에서의 이미지 및 패턴 분석에 의해 부분적으로 학습된다. 기계가 판독할 수 있는 텍스트 컨텐츠를 이미 가지지 않는 인입 문서에 대해 OCR이 적용될 수도 있다.

[0079] 이러한 청크들은, 선택된 레이아웃 정보와 함께 시스템에서의 후단 모듈들에 제시된다.

[0080] 2) 편성하기(organize)

[0081] 사용자들은 그들이 검사하는 문서들을 시스템으로 편성해야 할 필요는 없다. 시스템은, 문서들을, 예컨대 임대계약 대 임대차 계약 대 판매와 같은 특정 유형들의 문서들에 대한 '세트들'(sets)로 군집화하기 위해 텍스트 컨텐츠, 레이아웃 정보 및 (몇몇 헤딩들의 식별과 같은) 이미 검출된 구조적 정보에 대해 작동하는 클러스터링 방법들을 사용한다. 찾은 특정 문서 세트들은 사용자에 의해 검토되어 자동으로 또는 사용자에 의해 이름이 붙여질 수 있다. 이들 문서 세트들은, 일단 설정되면 포맷, 컨텐츠, 의미론적 역할들 및 이들 내에서의 차이들에 관한 후속 기계 학습 및 추론을 용이하게 한다. 예를 들어, 시스템은, 주어진 세트에서의 대부분의 모든 문서들이, 특정 역할들의 3개의 특정 서브 청크들 및 그 하나가 다섯개의 다른 절들에서 되풀이되는 개인 이름의 데이터 타입들을 가진 특정 절을 가지고 있음을 발견할 수 있다. 그러한 패턴들은, 사용자에게 검토 또는 변경을 제안하고 동일 (또는 다를 수 있는) 세트들에서의 다른 문서들에서 재사용하기 위한 예시적 텍스트를 제공하기 위해 다른 문서들의 유사한 (그리고 비유사한) 부분들을 식별하도록 돕기위해 사용된다.

[0082] 문서들의 문서 세트들로의 클러스터링은, 문서 구조(다양한 크기들, 데이터 타입들 및 역할들의 청크들 간의 순서 및 포함 관계들(order and containment relations)) 및 레이아웃뿐만 아니라 텍스트 컨텐츠로부터의 특징들을 사용할 수 있다. 일단 몇몇 청크들 및/또는 역할들이 적어도 몇몇 문서들에서 식별되면, 그 정보를 사용하여 완전히 다시 클러스터링하거나 더 작은 조정들을 함으로써 클러스터링을 향상시킬 수 있다. 예를 들어, 판매자 및 구매자 이름들, 주소들 등과 같이 동일 역할을 가진 청크들의 특정 컨테츠를 무시하거나, 예를 들어 하나의 이름(즉, 판매자의 이름)이 특정 위치들에 나타나는 한편 다른 이름(즉 구매자의 이름)이 특정의 다른 위치들에 나타난다는 것과 같이 서로 다른 청크들이 나타나는 패턴이 동일하다는 것을 검사한다면, 유사한 문서들이 거의 또는 완전하게 조차 동일해 질 수 있다.

[0083] 시스템은, 디렉토리들이 있다면 디렉토리들로 업로드된 파일들의 원래의 편성(organization) 및 이들의 세트들로의 그 고유 편성을 모두 유지한다. 따라서 사용자들은 두 편성들 모두를 볼 수 있고, 학습 알고리즘들은 둘 모두를 정보로서 이용할 수 있다. 예를 들어, 몇몇 사용자들은, 다양한 규칙들에 따라 문서들에 이름을 붙이고 및/또는 고객, 문서 종류 또는 (공통의 청크 위치들 및 역할들을 갖는 것과 같은) 유사 패턴들 및 문서들 간의 관계들을 이해하는데 거의 항상 유용한 다른 특징들에 의해 문서를 편성한다.

[0084] 3) 시각적으로 추출하기(visual extraction)

[0085] i) 영역 찾기(area finding)

[0086] 시스템은, 기하학적 패턴들에 기초하여 문서들에서의 영역들을 식별하기 위해 발견법(heuristics) 및 기계 학습을 사용한다. 예를 들어, 많은 문서들에 있어서, 의미있는 청크들은, 서명란(signature block), 초록(abstract), 정의 목록(list of definitions), 테이블들 등과 같은 특수 레이아웃을 가진다. 그러한 패턴들은, 기하학적 및/또는 레이아웃 특징들, 독창성 또는 희소성 및/또는 동일 문서 내에서의 또는 문서들에 걸쳐서의, 특히 동일 문서 세트 내에서의 일치성(correspondence)을 고려하여 자동으로 학습될 수 있다.

[0087] 인입되는 문서의 포맷에 따라 접근법들을 선택한다. 예를 들어, 워드 프로세서 문서들은 일반적으로 문단 경계들에 대한 명시적인 정보를 제공하나, PDF 또는 스캔된 페이지들의 경우에는, 시스템이, 시각적 라인들로부터 이들을 모아서 (다중 컬럼 문서들에서와 같이) 시각적 라인들에 문자들을 할당하기 위해 여백 치수들을 분석하는 것이 요구된다.

[0088] ii) 서명 찾기(signature finding)

[0089] 시스템은, 문서 부분들에 대해서 ('적요'(digests)라고도 알려진) 서명들(signatures)을 생성하고 이들을 사용하여 '흥미있는'(interesting) 추가의 청크들을 식별하고 분류하고 이들의 경계들을 찾는다. 서명들은 텍스트 컨텐츠에 기초할 뿐만 아니라 컨텍스트의 다양한 특징들에도 기초하며, 더 작은 포함된 청크들(예를 들어, 상대들(counterparts)에 있어서의 컨텐츠가 변하는 필드 청크들)의 컨텐츠를 무시할 수 있다.

[0090] 서명은, 청크의 픽셀 표현 조차 사용할 수 있다. 텍스트 레이아웃의 비트맵 이미지(bitmap image)는, 바람직하게는 (스캔 해상도를 위해 조정되는) 24 평방 픽셀(pixels square) 단위의 크기의 타일들(tiles)로 분할되고, 이 타일들은 클러스터링된다. 이들의 이웃 관계들을 비롯한 이들의 오토인코더 및 신경망 처리(autoencoders and neural network processing)는, 텍스트 및 규칙들 간의 경계들, 텍스트 블록들의 모서리들 및 코너들, 균일한 들여쓰기 변경들(even indentation changes) 및 실질적인 폰트/스타일 변경들과 같은 유사한 시각적 사건들(similar visual events)을 드러낸다. 더욱이 신경망들은 그 후 이 클러스터링을 사용하여 중요한 청크들을 자주 지시하거나 특성화하는 유사한 레이아웃 객체들을 공통으로 식별한다.

[0091] 여기서의 접근법은, 문서 청크에서의 문자들, 청크의 크기, 문서에서의 그 위치 등뿐만 아니라 픽셀들에 기초하여 문서 청크 임베딩들(document chunk embeddings)을 생성하기 위한 자율 접근법들(unsupervised approaches)을 사용할 수 있다(이미지들은 청크들일 수 있음). 많은 후속 태스크들을 위해 이 임베딩들에 대해 클러스터링 및 비교 기법들을 사용할 수 있다.

[0092] iii) 추출(extraction)

[0093] 이 특징은, 포스트 레이아웃 문서(예를 들어, PDF 또는 스캔된 프린트된 페이지)를 취해서 문서에서의 인식된 문자 이미지들("glyphs")을 glyphs의 정확한 문서 순서를 나타내는 텍스트 스트림으로 변환한다(스트림은, 적절한 경우 그림 또는 이미지 객체들을 포함할 수 있고, 읽는 순서에 있어 전형적인 위치들을 가지지 않는 풋노트들(footnotes) 또는 페이지 헤더들(page headers)과 같은 복수의 스트림이 있을 수 있다). 몇몇 문서들에서는, 읽는 순서의 불완전한 명시적 표현이 잇다. 잘 알려진 예는, 다중 컬럼 레이아웃이 임의의 주어진 지점에서 유효하고 따라서 첫번째 '라인'(line)이 끝까지라기 보다는 중간까지만 (또는 그에 못미치게) 연장된다는 표시는 전형적으로 없다는 것이다. 그러나, 텍스트의 순서가 복잡할 수 있거나 명백하지 않을 수 있는 많은 추가의 예들이 있다. 예를 들어, 몇몇 레이아웃 프로그램들은 각각의 문자를 독립하여 묘사하여 문자 경계들이 뚜렷하지 않게 한다. 테이블 셀들(table cells), 사이드바들(side-bars), 그림들, 풋노트들 및 기타 디스플레이들은 텍스트 순서에 있어서 명백한 위치를 가지지 않을 수 있다. (라인 끝 하이픈들(end-of-line hyphens) 뿐만 아니라) 페이지 헤더들 및 풋터들에 있는 것과 같은 몇몇 텍스트는 텍스트 순서에서의 위치를 전혀 요하지 않을 수 있다. 많은 포맷들은, 무엇인가가 그러한 특별한 카테고리들에 있다는 명시적인 표시를 제공하지 않는다.

[0094] 시스템은, glyphs의 시각적 정보(위치, 스타일 등)을 텍스트 스트림을 구축하기 위해 문서에서 사용되는 작성 언어(written language)의 특성들을 이해하는 심층 신경망과 결합함으로써 이 태스크를 다룬다. 뿐만 아니라, 이는, 라인(line), 블록, 컬럼, 이미지, 인라인 폰트 변경들(inline font-changes) 및 헤더/풋터 객체들에 대한 것과 같은 많은 기본적 텍스트 경계들을 검출한다.

[0095] iv) 표현하기(represent)

[0096] 텍스트 시퀀스(text sequence)와 몇몇 가정된 구조 청크들(hypothesizedstructure chunks)이 추출되면, 시스템은, 시각적 특성들(폰트들, 컬러들, 크기들 등)에 관한 정보뿐만 아니라 그것들을 포함하는 (일 실시예에서

"DGML"이라고 알려진) 문서의 표현(representation)을 생성한다. 청크들의 위치, 타입 및 역할과 같은 정보를 포함하는 청크들의 표현들(representations of chunks)은, '주석들'(annotations)이라고 불려진다. 그 후, 결합된 데이터가 자연어 처리(natural language processing: NLP) 및 심층 신경망들(deep neural networks: DNN)에 의해 사용될 수 있다. 심층 신경망들은, 이 시각적 정보를 포함하고 있어, 문서를 계층으로 구조화하여 헤더들/본문들, 리스트들/리스트 항목들 등과 같은 청크들을 포함하는 문서 구조를 표현하는 것을 도와준다.

[0097] 충분한 정보가 포함될 수 있어서, 이후의 특징들(later aspects)이 원래의 소스(original source)와 근접하게 닮은 편집가능한 워드 프로세서 문서를 구축할 수 있다. 이는, 다른 구조, 컨텐츠 및 청크 정보와 함께 DGML 또는 유사한 표현에 포함될 수 있다. 많은 경우에 있어서, 상이한 포맷팅과 레이아웃을 가진 문서의 부분들 또한 유용한 청크들이다. 그러나, 그렇지 않으면 필요한 청크들과 일치하지 않는 (및 그 반대의 경우의) 포맷팅 특성들이, 여전히 특수 유형의 청크, 스탠드오프 주석(standoff annotation) 또는 다른 방법들을 통해 나타내질 수 있다.

[0098] 4) 구조(structure)

[0099] 구조 파이프라인(structure pipeline)은, 평면적인 텍스트 파일(flat text file)을 계층적 구조로 변환하는데, 이 계층적 구조에서는 문서의 절들(sections), 소구분들(subsections) 및 다른 부분들이 당업자에게 알려진 구조인, 콘텐츠 기반 객체의 순서 계층(Ordered Hierarchy of Content-Based Objects)을 형성한다. 이러한 변환은, 자율 기계 학습 기법들을 이용하여 이루어진다. 이 방법은 여러 스테이지들을 갖는다.

[0100] i) 하이퍼라이닝(hyperlinking)

[0101] 이는 텍스트를 '하이퍼라인들'(hyperlines)로 분할하는 것을 수반하는데, 하이퍼라인들은, 시각적 라인들 보다 큰 그룹들이고 문단들, 헤딩들 또는 이와 유사한 것들과 같은 (시각적인 것에 반대되는) 더욱 의미있는 논리적인 유니트들을 포함한다. 이는, 토큰들(특히, 선행 및 후행 토큰들(leading and trailing tokens))의 '단어 모양'(word shape)과 같은 특징들, 폰트 및 스페이싱(spacing) 특성들과 같은 레이아웃 정보 및 유사한 특징들을 고려하는 미리 학습된 신경망을 이용하여 바람직하게 이루어질 수 있다. 몇몇 하이퍼라인들은, (입력 문서의 포맷에 따라서) 이전 단계들에 의해 제공되었을 수도 있다.

[0102] ii) 문서 언어 모델(document language model)

[0103] 이는, 문서 언어 모델(Document Language Model)을 바람직하게 사용하는데, 이 문서 언어 모델은, 단지 텍스트에 기반하는 언어 모델(Language Model) 대신에 텍스트 컨텐츠, 포맷팅 및 어떤 구조가 되었든 이제까지 발견된 구조에 대한 정보를 또한 포함한다. 이는, 포맷팅된 페이지들로부터 의미있는 청크들과 그 출현 패턴들을 인식하는 것을 학습함으로 인해 청크들 및 (헤더들/본문들, 리스트/리스트 항목들 등과 같은) 청크들의 계층을 더 잘 검출할 수 있도록 해준다.

[0104] 이는, 텍스트적인 컨텐츠(textual content)와 시각적인 특성들(기하학적 배열(geometry), 폰트들, 컬러들, 크기들 등) 모두를 포함하는 문서의 표현을 생성한다. 그 후, 심층 신경망들 및 NLP 프로세스들은, 문서 구조를 나타내는 다양한 크기의 청크들의 범위들 및/또는 경계들을 찾음으로써 문서를 데이터 타입들과 의미론적 역할 레이블들을 가지는 청크들의 계층으로 구조화하는 작업에 있어서 그러한 정보를 이용한다. 이 스테이지에서는, 발견한 청크들은, 주로 헤딩들, 절들, 리스트들 및 항목들, 테이블들, 그림들 및 기타 상대적으로 큰 유니트들이다.

[0105] iii) 하이퍼라인 클러스터링(hyperline clustering)

[0106] 이는, 단어 모양 구조(word-shape structure)에 기초하여 문서 세트에 걸쳐 하이퍼라인들을 클러스터링하여 각각의 하이퍼라인을 레이아웃, 시작 및 종료 컨텐츠 및 다른 특성들의 면에서 유사한 하이퍼라인들의 클러스터로 할당하기 위해 오토인코더(autoencoder)를 사용한다. - 여기서 각각의 클러스터는 '클러스터 ID'에 의해 식별됨 - (이를 문서 세트들의 생성 또는 식별과 혼동해서는 안된다).

[0107] iv) 인라인 헤딩들(inline headings)

[0108] 특별한 관심이 있는 특수한 경우는 '인라인 헤딩들'(inline headings)인데, 여기서는 (가끔 청크의 의미론적 역할을 제공하는) 청크의 헤딩이 그 자체로 별도의 시각적 라인(들)에 있는 것이 아니라 다음 텍스트의 시작과 같은 라인 상에 있다. 대체로 인라인 헤딩들은, 볼드체(boldface), 밑줄(underlining), 다른 폰트, 뒤따르는 콜론(following colon) 또는 다른 효과들에 의해서와 같이 활자체 면에서(typographically) 구분된다. 별도의

발견법(heuristics)과 신경 알고리즘들(neural algorithms)이 이러한 청크들을 식별한다.

[0109] v) 퓨샷 구조 학습(few-shot structure learning)

[0110] 상술한 진보된 구조화 방법들에도 불구하고, 생성되는 구조가 몇몇 결점들을 가지고 있고 사용자의 선험적인 기대를 만족시키지 않을 것으로 예상될 수 있다. 퓨샷 구조 학습은, 단계(14) 및 단계(15)에서 기술된 바와 같이 사용자에 의해 제공되는 피드백에 의존하여 기계 학습 모델을 생성하는 처리를 한다. 그 후 이 모델은, 구조에 관한 사용자 피드백을 시스템에 의해 이미 생성된 (그리고 아마도 이전의 피드백에 의해 반복적으로 향상된) 것과 결합하는 구조를 생성하기 위해 사용된다.

[0111] 이 경우에 적용되는 주요한 원칙은 기계 번역(machine translation: MT) 방법들로부터 유도되는데, 이 기계 번역 방법들의 경우에는 시퀀스(sequence)가 다른 것으로 변환된다. 이 경우에, 하이퍼라인들을 기술하는 하나의 시퀀스는, 계층을 인코딩하는 시작/종료 마커들(start/end markers)을 또한 포함하는 다른 것으로 변환된다.

[0112] 이 프로세스는 서로 다른 페이스들(phases) 또는 단계들에서 일어난다.

[0113] (a) 첫째로, 기계 번역 모델이 공공연히 입수가능한 데이터 세트를 이용하여 미리 학습된다.

[0114] (b) '디스패처'(dispatcher)(설명을 위한 '피드백 응답'(Feedback Response) 절 참조)가 사용자 피드백을 필터링한다.

[0115] (c) 새로운 구조 파일들이 사용자 피드백 및 생성된 미세 조정 기계 번역 데이터 세트로부터 생성된다.

[0116] (d) 미리 학습된 모델이 퓨샷 학습 원리를 이용하여 더욱 학습된다.

[0117] 5) 리네스팅(re-nesting)

[0118] 이 특징은, 클러스터 ID들의 플랫 리스트(flat list)가 주어지면 푸시다운 오토메이션(pushdown automation)을 사용하여 바람직하게는 하이퍼라인(Hyperline) 클러스터링 단계로부터 네스팅된 구조(nested structure)를 반복적으로 생성하는 '코퍼스 리네스트'(corpus re-nest) 알고리즘을 이용한다. 이웃하는 하이퍼라인들의 서명들을 비교함으로써, 시스템은, 주어진 헤딩 또는 리스트 항목이 더 높은 수준으로, 동등하게 또는 더 낮은 수준으로 네스팅된 레벨에 속하는지의 여부를 결정할 수 있다. 이로 인해 (장들(chapters), 절들(sections), 소구분(subsection), 구절(clauses), 리스트들 등과 같은) 많은 문서들의 다중으로 네스팅된 계층적 구조(multiply-nested hierarchical structure)를 재구성할 수 있게 된다.

[0119] 리네스팅(re-nesting)에서 고려되는 특징들은, 첫번째와 마지막, 이전의 라인에서 끝나는 구두점 마크가 있다면 그 구두점 마크의 특정 클래스, 대문자로 쓰기(capitalization), 앞서는 여백, 들여쓰기, 볼드 및 밑줄과 같은 포맷팅 정보, 라인의 시작에 있는 열거형 스트링(enumeration string)(예를 들어, "IV(A)(1)" 또는 "iv)"와 같은 패턴들)의 존재와 형태 또는 특정의 큰 점(bullet) 또는 다른 장식 문자(dingbat character), 그 열거자(enumerator)의 값, 동일 종류의 앞서는 열거자들의 존재, 레벨들 및 값들 등등을 특히 고려하여, 하이퍼라인들에서의 토큰들(tokens)의 (NLP 기술에서 알려진 바의) '모양'(shape)을 포함한다.

[0120] 6) 토픽 청킹(topic chucking)

[0121] 이 특징은, 어디서 토픽들이 이동하는지를 찾기 위해 문서의 연속되는 청크들에 걸쳐 사전적인 통계(lexical statistics)와 다른 학습 기법들을 사용한다. 이는, 주어진 토픽들에 대한 전체 절들과 같은 큰 청크들에 대한 경계들의 식별을 향상시키는데, 그 이유는 (어떤 레벨이든지간에) 절(section)은, 이웃하는 절들과의 관계에서보다는 그 내부에서 토픽, 어휘 및 스타일의 일체성(uniformity)을 더 갖기 때문이다.

[0122] 7) 토픽 레이블링(topic labeling)

[0123] i) 헤딩 레이블러(heading labeler)

[0124] 도 1에 도시된 코퍼스(corpus)에서의 각각의 헤더에 대해, 이 단계는,

[0125] - 각각의 헤딩에 대해 '임베딩'(embedding)이라고 알려진 수치적 표현을 생성하고,

[0126] - 그러한 임베딩들에 적어도 기초하여 헤딩들을 클러스터링하고,

[0127] - 밀도(density), 애리티(arity) 및 유사도와 같은 척도들에 적어도 기초하여 '불량'(bad) 클러스터들을 걸러내고,

[0128] - 각각의 남아있는 클러스터에서 가장 일반적인 의미론적 역할 레이블을 상기 클러스터에서의 모든 헤딩들에 전파시킨다.

[0129] ii) 핵심 구 레이블러(keyphrase labeler)

[0130] 각각의 청크에 대해, 이 단계는, 텍스트에 대한 후보의 의미론적 역할 레이블들을 생성하기 위해 (룰 기반의 언어 기법들(Rule-based linguistic techniques), ML, 통계(Statistical), 베이지언(Bayesian) 및/또는 그 밖의 것들과 같은) 핵심 구 추출 기법들(key phrase extraction techniques)의 총체를 이용한다.

[0131] 8) 청크 레이블링(chunk labeling)

[0132] i) 문법(Grammar)

[0133] 시스템의 이 특징은, 품사 태깅(Part of Speech (POS) Tagging), 의존 구문분석(Dependency Parsing), 구성 구문분석(Constituent Parsing) 및 그 밖의 것들을 포함하는 자연어 처리 태스크들과 같은, 텍스트의 언어적 분석으로 시작된다. 이 시스템은, 그 후 다른 도메인으로부터의 트리 매칭 메커니즘들(tree-matching mechanisms)을 적용하여 NLP를 통해 발견된 트리들 또는 트리와 같은 구조들 내에서 문법적이고 다른 구조들을 찾는다. 이들은, XPath, GATE 및 다른 것들과 같은 도구들에 의해 예시되는 바의 트리 문법들(tree grammars) 및 트리 패턴 매칭(tree pattern matching)과 같은 문서 구조화 방법들(document structuring methods)을 포함한다.

[0134] 문장들에서 문법적 현상을 식별하기 위해 그러한 패턴들을 사용하게 되면, 시스템이 텍스트 자체로부터 의미론적 역할 레이블들을 추출할 수 있게 되는데, 이 의미론적 역할 레이블들은 그 후 이웃하는 청크들을 주석 처리하기 위해 사용된다. 예를 들어, 문장의 구성 구조(constituency structure of the sentence)에 기초하여 "The following are the terms of our agreement"라는 문장 (및 유사한 문법적 구조를 가지는 다른 문장들)과 정합되는 서치 패턴(search pattern)을 구성할 수 있고, 명사구(본 예에서는, "the terms")를 추출하여 이를 이 문장 뒤에 오는 그리고 그러한 "terms"를 포함하는 컨텐츠에서의 하나 또는 둘 이상의 청크들에 대한 의미론적 역할 레이블로서 사용한다.

[0135] ii) 질의 응답(question-answering)

[0136] 질의 응답(Question Answering)을 위한 BERT를 포함하는 질의 응답 기법들은, (예를 들어, 날짜들, 개인 이름들, 달러 액수인) 후보 청크들을 위한 의미론적 역할 레이블들을 식별하도록 특별히 조정된다. 대조적으로, 대부분의 전통적인 질의 응답 모델들은, "What is the Effective Date?"와 같은 질의들에 응답하는 쪽을 지향한다. 이 시스템은, 대신 "What is July 8, 2018?"과 같은 질의들에 응답하기 위한 모델들을 학습시키고 "Effective Date" 또는 "effective date of X"를 예측하는 쪽을 지향한다 - 여기서 X는 텍스트에서의 (의미론적 역할이라기 보다는 데이터 타입인 단순한 "Date"가 아닌) 또 다른 청크를 나타냄 -.

[0137] 이 시스템은, 또한 응답되었을 때 텍스트에서의 관련 정보를 지적할 수 있는 종합적 질의들(synthetic questions)을 발견할 수 있다. 이는, 질의 응답에 의해 사용되기 위한 질의들을 자동으로 제기하는 기능을 제공한다.

[0138] iii) 임베딩들과 통합된 XPath 형 규칙들(XPath-like rules integrated with embeddings)

[0139] 여기서는, '문법'(Grammar)이라는 제목 하에 논의되는 영역들에서의 도구들은, word2vec, char2vec 및 많은 관련된 방법들과 같은 텍스트의 벡터-의미론적 표현들(vector-semantic representations)을 제공하는 도구들과 통합된다. 이 시스템은, 분석가들로 하여금 (XML- 또는 DOM- 호환 형식들로 표현되는 청크 데이터를 포함할 수 있는) XPath 또는 유사한 도구들에 의해 잘 다루어지는 구조 정보 및 벡터 모델들에 의해 잘 다루어지는 '퍼지'(fuzzy) 또는 '의미론적'(semantic) 유사성 정보를 모두 포함하는 패턴들을 표현하고 그에 대해 질의하도록 해준다.

[0140] 9) NER (레이블링되지 않은 작은 청크(unlabeled small chunk))

[0141] 기술은, 개인 또는 기업체 이름들, 주소들 등등과 같은 데이터 타입 별의 몇몇 청크들을 식별할 수 있다(이는, '개체명 인식'(Named Entity Recognition) 또는 "NER"이라 알려져 있다). 그러나 NER은 문서에서 그러한 개체들의 의미론적 역할들을 식별하는 능력이 부족하다. 현재의 기술은 또한, 전체 구절들(clauses) 또는 절들(sections)과 같은 더 큰 청크들 또는 의미있고 유용한 더 큰 청크들을 포함하는 청크들의 그룹들(groups of chunks)을 식별하지 못한다.

[0142] 시스템의 이러한 특징은, 주의를 끄는 작은 청크들(interesting small chunks)에 필연적으로 역할들을 할당하지

않고도 그러한 청크들을 검출한다. 텍스트에서 NER들을 식별하기 위한 많은 방법들과 도구들이 존재한다. 이 시스템은, 복수의 방법을 사용하는데, 그 예들이 아래에 나열되어 있다. 이러한 혁신들(innovations)은 주로 자율적이다.

[0143] i) 정착된 NER 방법들

[0144] ii) 예상되는 단어들

[0145] 위키피디아(Wikipedia)와 같은 방대한 일반적 텍스트를 사용하여 n-그램 언어 모델(language model of n-grams)을 학습시킴으로써 '정상적 영어를 위한 문맥에 있어서의 예상되는 단어들'("expected words, in context, for normal English")의 모델을 구축한다. 특정 문서를 주목할 때, 시스템은, 그 일반적 모델에 맞지 않고 따라서 처리되는 문서에 특수한 경향이 있는 n-그램들(n-grams)을 식별하는 수단을 제공한다.

[0146] iv) TF-IDF

[0147] 이는 TF-IDF 기반의 접근법('단어 빈도 대 역문서 빈도'(term frequency vs. inverse document frequency))이고, 레이블 전파 및 문맥적 의미 레이블링(Label Propagation and Contextual Semantic Labeling)과 결부하여 사용된다.

[0148] v) 시퀀스 클러스터링(sequence clustering)

[0149] 작은 단어 또는 n-그램들(n-grams)과 같은 문자 시퀀스들을 추출하고, 이들을 문맥적 임베딩들(contextual embeddings)(예를 들어, BERT의 것들)을 이용하여 클러스터링한다. 예상되는 결과는, 의미론적 의미(semantic meaning)를 공유하는 n-그램들이 같이 클러스터링을 시작할 것이라는 것이다. 조합확산(combinatorial explosion)의 비용은, (구문 트리(syntax tree)에 대한 것을 포함하는) 발견법(heuristics)을 사용하여 클러스터링 이전에 몇몇 n-그램들을 걸러냄으로써 해결된다. 다양하고 폭넓은 클러스터링 알고리즘들이 적용될 수 있다. 이 예에서는 "hdbscan" 알고리즘이 '비'(none) 클러스터에 랜덤 노이즈를 할당하면서 효율적인 클러스터링을 성취할 수 있다.

[0150] vi) 퓨샷 NER(few-shot NER)

[0151] 시스템은, 작은 수의 레이블링된 인스턴스들(labelled instances)(예를 들어, 선택적 사용자 피드백)으로부터 좀 더 폭 넓게 적용가능한 규칙 또는 학습된 변수들의 조정(adjustment of learned parameters)으로 일반화하기 위해 퓨샷 학습 기법들을 사용한다. 이는, 사용자들이 피드백을 요청받아야 하는 회수를 상당히 줄이고 시스템 성능을 급진적으로 향상시킨다.

[0152] 10) 추출성의 레이블링(extractive labeling)

[0153] 시스템의 이 특징은, 청크 주위의 문장(들)에 직접적으로 나타나는 작은 청크들에 대한 의미론적 역할 레이블들을 찾는다. 의미있는 청크들의 경우 그들의 역할이 컨텍스트에 의해 몇몇 형식으로 종종 특정된다. 예를 들면 다음과 같다.

[0154] Jane Doe (the "Seller"), resides at....

[0155] Rent of $999 must be paid by the end of each month.

[0156] i) 문맥적 의미 레이블링(Contextual Semantic Labeling: CSL)

[0157] 이 프로세스는, 텍스트의 어느 부분들이 다양한 청크들에 대한 의미론적 역할 레이블들일 가능성이 많은지를 학습하기 위해 문장 파스(sentence parses)를 포함하는 이전에 구축된 구조들에 대해 작동하는 신경망들을 사용한다. 많은 청크들은, 다양한 소스들과 신뢰 수준들을 갖는 그러한 레이블들을 이미 가질 수 있으나, 이는, 신규 레이블들뿐만 아니라 그러한 것들에 대한 또는 그러한 것들에 반하는 추가의 증거를 제공한다. 여기서 패턴들의 몇몇은 문법을 수반한다. 예를 들어, "Doe shall pay a rent of $1000 by the last business day of each month"에서 선두 동사(head verb)는, 통화량의 역할이 무엇인지를, 즉 그것이 지불해야 할 임대료 금액이다라는 것을 분명히 해준다. 다른 패턴들은, 문맥상으로 이용가능한 구조, 청킹(chunking), 레이블링 및 컨텐츠의 특징들을 이용하여 지도 및/또는 자율적 방법들에 의해 자동으로 학습된다. 괄호들(parentheses), 테이블 레이아웃, 핵심 구들(key phrases) 및 단어들과 같은 포맷팅 및 다른 특징들은 신경망에 대한 특징들을 또한 제공한다.

[0158] 유용한 정보는, 종종 절 또는 소구분과 같은 포함적 청크(containing chuck) 또는 그 헤딩들에 상주한다. 예를

들어, 주어진 약물(medication)이 처방(prescription)으로서 아니면 앨러지(allergy)로서 관련이 있는지는, 포함적 절(containing section)의 헤딩을 봄으로써 검출가능할 수 있다(이는, 절들의 정확한 계층적 네스팅(nesting)을 검출하는 것이 왜 중요한지를 설명해 주는 또 다른 예이다). 기계 학습 기법들에 의해 학습될 수 있고 다양한 청크들에 대해 적용가능한 역할들을 발견하기 위해 적용될 수 있는 많은 여타의 단서들이 있다. 유사한 컨텍스트들에 대해 발견되었으나 분리된 문서(document in isolation)에 대해서는 발견될 수 없을 수도 있는 의미론적 역할들을 연관시키기 위해, 특히 동일 문서 세트에서의 문서들의 경우 문서 간 유사성(cross-document similarities)이 또한 이용될 수 있다.

[0159] ii) 레이블 전파(label propagation)

[0160] 이 프로세스는, 문서들의 코퍼스(corpus)에서의 유사한 텍스트 청크들에 걸쳐 레이블들을 표준화한다. 이는, 컨텍스트로부터 추출된 레이블들 및 이전 단계들로부터 이용가능한 레이블들에 모두 적용된다. 이 알고리즘은, 청크들을 그 임베딩들에 기초하여 클러스터링하기 위해 응집 군집화(Agglomerative Clustering)를 사용하고, (초기 노드 가중치들로서 레이블들의 빈도/신뢰도를 사용하는) 가중 페이지 랭크 알고리즘(weighted PageRank algorithm)을 사용하여 각각의 청크 클러스터(cluster of chunks)에 대한 후보 레이블들에 대해 등급을 매기고, 레이블들이 서로서로 얼마나 유사한지를 결정하기 위해 동시 출현(co-occurrence) 및 임베딩 유사도(embedding similarity)를 사용한다. 이는, 청크에 그들의 클러스터 레벨 점수(cluster-level score) 그리고 우리가 레이블링하는 청크가 (컨텐츠, 임베딩들, 구조, 데이터 타입, 의미론적 역할 및/또는 컨텍스트의 면에서) 레이블이 나온 청크와 얼마나 유사한지에 기초하여 레이블들을 할당한다. 유사한 컨텍스트들에 걸쳐 레이블들을 전파시키기 위해 그리고 레이블들이 문서들의 세트에 걸쳐 더욱 일관성이 있게 만들기 위해 응집 군집화(Agglomerative Clustering) 및 페이지 랭크(PageRank) 알고리즘들이 적용된다.

[0161] 11) 변칙들(anomalies)

[0162] 시스템의 이 특징은, 단계(2)에서 생성된 것과 같은 문서 세트 내의 복수의 문서들을 조사하고, 현재의 문서에는 출현하나 동일 세트의 다른 문서들에서 카운터파트 청크들(counterpart chunks)을 공통으로 가지지 않는 또는 그 반대의 경우에 해당하는 청크들을 식별한다. 카운터파트 청크들은, 동일한 컨텐츠, 구조, 포맷팅, 컨텍스트, 데이터 타입들 및 의미론적 역할들을 가질 필요가 없으나, 문서 마다 변형들(variations)을 가질 수 있다. 그럼에도 불구하고, 이들은 그런 방식으로 다른 식별된 청크들과 실질적으로 유사한 것으로서 인식될 수 있다.

[0163] 신규 문서가 동일 세트의 다른 문서들에 전형적으로 존재하지 않는 청크들을 포함하는 경우, 사용자는, 이들이 실제로 의도된 것이라고 확인하기 위해 이들의 일부 또는 모두에 관해 질의를 받을 수 있다. 이 예시적인 시스템에서, 의문의 청크가 신규 문서 및 그것이 기초로 삼은 것(만일 있다면)에 공통인 경우 그러한 질의하기(querying)는 더욱 널리 퍼지게 될 것이다.

[0164] 신규 문서에, 동일 세트의 다른 문서들 또는 특히 관련있는 외부의 소스들(예를 들어, 하우스 스타일 매뉴얼(manual of house style), 준수 요건(compliance requirement) 등)에서조차 전형적으로 존재하는 카운터파트 청크들이 없는 경우, 몇몇의 또는 모든 그러한 청크들의 예들이, 다른 문서들로부터 끌어낸 컨텐츠와 함께 사용자에게 제안된다. 이 제안들은, 사용 빈도, 이용가능한 대안들의 가장 전형적인 것(중심(centroid))인지 또는 신규 문서에 있는 다른 청크들과의 동시 출현(co-occurrence)의 확률이 높은지와 같은 요소들에 따라 사용자를 위해 등급이 매겨질 수 있다. 청크 제안들은, 예를 들어 이름들, 날짜들 및 예들을 도출한 문서들에 특유한 다른 서브 청크들을 신규 문서로부터 도출한 값들로 대체하기 위해 자동으로 갱신될 수 있다.

[0165] 더욱이, 추가나 삭제를 위해 제안할 청크들의 선택은, 다양한 저작자들, 편집자들 또는 여타 스태프의 실무에 실질적으로 의존할 수 있다. 예를 들어, 현재의 저작자의 문서들이 다른 저작자의 것들과 특별한 방식으로 빈번히 다른 경우, 이는, 그 차이가 고려된 선택이고 오류가 아님을 나타낼 수 있다. 한편, 같은 감독자 아래서 작업하는 모든 저작자들이 무언가를 어떤 방식으로 하나 현재의 저작자는 그와 다른 경우, 이는, 적어도 처음 인지되었을 때 더 큰 검토의 필요성을 나타낸다.

[0166] 변칙들의 모델링은, 텍스트, 컨텐츠 및 포맷뿐만 아니라 구조 및 청크 데이터 타입들 및 의미론적 역할들을 고려한다. 예를 들어, 청크들의 어떤 데이터 타입들과 의미론적 역할들이 다른 것들 안에, 다른 것들에 인접하여 또는 다른 것들 가까이에서 출현하는 것인지의 패턴들을 모델링한다. 잘 정착된 패턴들의 위반(violations)은 변칙들로서 분류될 수 있고 임의의 다른 변칙들과 같이 사용자 피드백을 위해 제시될 수 있다.

[0167] 12) 조정(arbitration)

[0168] 많은 이전의 단계들은, 단계(3)에서 생성된 선형 시퀀스(들) 내의 문자들, 토큰들 및/또는 텍스트가 아닌(non-text) 객체들의 (전형적이긴 하나 필연적으로는 연속되지 않는) 범위들로서 정의된 문서의 청크들을 생성하고 이에 대해 동작한다.

[0169] 임의의 지점에서 고려 대상인 청크들은, 마크업(markup)과 같은 인라인(inline) 메터 정보에 의해서 또는 다양한 종류의 포인터들에 의해 텍스트에서의 위치들을 지시하는 '스탠드오프'(standoff) 표현들에 의해 표현될 수 있다. 이 예에서, 스탠드오프 표현들이 대부분의 처리를 위해 사용되나, 인라인 표현들(inline representations)이, 종종 이를 선호하는 외부 도구들과의 통신과 같은 몇몇 목적을 위해 사용된다. 이러한 표현들과 다른 것들은, 기능적으로 서로 교체할 수 있고, 이들 간의 선택은 성능, 편리성 등등의 관심사에 의해 좌우된다.

[0170] 청크들의 표현은, 어떤 단계들 또는 구현들이 이들을 생성하였는지, 이들이 얼마나 확실한지('신뢰도'(confidence level)) 그리고 이들의 특정 데이터 타입들 및/또는 의미론적 역할 레이블들에 관한 정보를 포함한다. 여분의, 불확실한, 모순되는 또는 부분적으로 겹치는 청크들이 빈번히 나타날 수 있는데, 이러한 청크들을 여기서는 '최적이 아닌'(non-optimal) 것이라고 부른다. 예를 들어, 두개 또는 세개 이상의 서로 다른 프로세스들이, 동일 범위의 텍스트(same span of text)에 (또는, 예를 들어, 이름 전에 "Dr."를 포함하는 것 및 그렇지 않은 것과 같이 거의 동일 범위의 텍스트에) 의미론적 역할 레이블들을 붙일 수 있다. 청크들은, 가끔은 깊게 네스팅될 수 있으나, 임의로 겹칠 수도 있다(즉, 겹치는 청크들의 각각은 다른 것에도 있는 몇몇 컨텐츠 및 그렇지 않은 몇몇을 포함한다). 전술한 단계들을 통해서, 시스템은, 겹치는 또는 동위치에 배치된(co-located) 것들을 포함하는 많은 수의 주석들을 나타낼 수 있는 표현들을 유지할 수 있다.

[0171] 그러한 최적이 아닌 청크들은, 적어도 문서가 사용자에게 제시될 때 보통 바람직하지 않다. 뿐만 아니라, 최신의 많은 NLP 도구들이, XML, JSON, SQL 및 다른 표현 시스템들(representational systems)과 같은 본 기술 분야에 숙련된 자들에게 친숙한 문서 도구들 및 방법들이 그러한 것 같이, 겹치지 않는 구조들(non-overlapping structures)을 선호한다. 보통 선호되는 좀 더 제한된 구조들은, 종종 '계층적인'(hierarchical) 또는 '잘 형성된'(well-formed) 것이라고 불려지고, 부분적으로 겹치는 청크들을 피한다.

[0172] 시스템의 이러한 특징은, 청크들의 모음(collection of chucks)이 엄격하게 계층적이되도록 그리고 최적이 아닌 청크들을 피하도록 수정한다. 이는 다양한 방식들로 성취될 수 있다. 첫째로, 청크들이 완전히 삭제될 수 있다(즉, 청크들 그 자체, 그들이 식별한 문서 컨텐츠는 삭제되지 않는다). 둘째로, 청크 범위(chunk scope)가, 다른 청크(들)과의 중첩을 방지하기 위해 (예를 들어, 하나 또는 둘 이상의 문자들 또는 토큰들을 포함시키거나 양쪽으로부터(from either end) 제외시킴으로써) 수정될 수 있다. 셋째로, 청크들은, 남아도는(redundant) 것으로 그리고 병합된(merged) 것으로 결정될 수 있다. 넷째로, 청크들은 모순된(contradictory) 것으로 드러날 수 있고(예를 들어, 하나의 도구가 "Essex"가 장소라고 생각한다면, 하나는 사람이다), 선택이 이루어진다.

[0173] 이 프로세스는, 부분적 및/또는 완전한 겹침의 경우들을 신속하게 찾고, 타입, 역할 및 신뢰도에 의해 청크들을 비교하고, 청크들과 그들에 연관된 데이터를 수정함으로써 최적이 아닌 경우들을 해결하기 위한 수단을 포함한다. 어떤 청크들을 수정하고 병합하고 또는 삭제할 것인지를 선택하는 것은, 신뢰도 수준들, 주어진 청크의 데이터 타입, 의미론적 역할 및 컨텐츠의 선험적인 확률, 의미론적 역할 레이블들 간의 하위성(hyponymy), 주어진 컨텍스트에서의 출현(occurrence)에 대한 조건부 확률들, 현재의 그리고 다른 유사한 문서들에서의 다른 청크들의 이름, 역할들 및 분포, 그 당시의 프로세스에 대한 우선순위들, 유사한 케이스들에 대한 고객 피드백 및/또는 다른 방법들과 같은 여러 요소들을 고려한다.

[0174] 수정을 함으로써 청크 신뢰 수준들 역시 변경할 수 있다. 예를 들어, 본 시스템의 많은 특징들은, 유사한 또는 동일한 의미론적 역할 레이블들을 문서의 동일한 또는 거의 동일한 부분에 적용할 수 있다. 그 경우, 레이블들이 전형적으로 병합되고, 결과적인 청크가 그것이 포괄하는 개개의 청크들 보다 더 높은 신뢰도를 할당받는다. 다른 경우들에 있어서, 모순되는 청크 할당들 간의 선택이 이루어지나, 선택된 청크는, 어느 정도의 반증(counter-evidence)이 있었다는 것을 반영하기 위해 결과적으로 그 신뢰도가 줄어들 수 있다.

[0175] 이 프로세스는, 청크 식별 및 레이블링의 품질과 일관성을 개선하고, 정보가 다양한 범위의 도구들과 상호 작용하도록 해주며, 결과가 더욱 쉽고 신뢰성 있게 분석될 수 있도록 해준다. 이제까지 기술한 동작들은 단지 끝부분이 아니라 언제라도 적용될 수 있다. 예를 들어, 이전 단계가 몇몇 서브태스크(subtask)를 위한 외부 도구를 사용하는 경우, 이는 정형성(well-formedness)으로의 변형(reduction)을 요구할 수 있다. 제거된 또는 수정된 청크들은 대신에 보류될 수 있는데, 이는 그들이 더 이상 처리과정에 영향을 미치지 않으나 요청이 있을 시 다시 도입될 수 있다는 것을 의미한다. 이는, 그 후에 무에서부터 이전의 작업을 재생성해야 할 필요없이 겹치

지 않는 지원 도구들(non-overlap-supporting tools)의 사용을 가능하게 해주고 처리의 유연성과 속도를 증가시킨다.

[0176] 한 접근 방식에서는, 겹치는 모든 및/또는 최적이 아닌 모든 청크들이, 사용자에게 보여지는 문서를 생성하기 이전에 해결되어, 결과가, 많은 신식의 워드 프로세서들 및 다른 도구들에 의해 사용되는 XML 포맷들과 같은 계층적 포맷들로 쉽게 인코딩될 수 있다. 그러나, 사용자 피드백 또는 향상된 알고리즘적 학습에 의해서와 같이 추후에 있을 수 있는 해결을 위해 특정 위치들에서 겹칠 가능성이 있는 복수의 대안(multiple possibly overlapping alternatives)을 유지하는 것이 (XML에서조차) 또한 가능하다.

[0177] 13) DGML(DocuGami Markup Language)

[0178] 문서의 향상된 버전은 문서의 구조, 포맷, 컨텐츠 및 식별된 청크들을 나타내고, 프로세스의 어느 단계들이 어느 청크들을 어느 수준의 신뢰도로 식별하였는지를 식별할 수 있다. 폭 넓은 범위의 표현들(representations)이 다른 XML 스키마들(schemas), JSON, 다양한 데이터베이스들, 커스텀 텍스트(custom text) 및 이진 포맷들 등등과 같은 동일한 정보를 실질적으로 포함할 수 있지만, 몇몇 실시예들은, 이 표현(representation)의 구문(syntax)으로서 XML을 사용한다.

[0179] 이 단계에서는, 문서와 발견된 청크들에 관한 정보가, 다른 프로세스들, 특히 피드백, 편집 및 검토에 사용되는 프론트엔드 사용자 인터페이스로 쉽게 전달될 수 있는 XML 형식으로 그리고 그룹 매니져들, 품질 관리 직원 등과 같은 다른 사용자들에게 개요, 통계 및 준수(compliance) 정보를 제공하는 '대시보드'(dashboard) 애플리케이션들 용으로 유용한 포맷들로 변한된다 (또는 직렬화된다(serialized)).

[0180] DGML(DocuGami Markup Language)은, 하나의 패키지에 기술되는 정보의 모두를 수용하는, 이러한 용도로의 특별한 XML 스키마이다. 대부분의 이전의 스키마들은 구조, 컨텐츠 그리고 가끔은 레이아웃을 다룰 수 있으나, 여기에 기술되는 초록(abstract)에서는 '청크들'(chunks)을 주석 처리하지 않는다. 이전의 많은 스키마들은, 청크들이 자동으로 검출되고 표현될 수 있는 일반화된 매카니즘을, 특히 신뢰 수준들 및 출처(provenance) 정보와 함께 제공하지 않는다.

[0181] 몇몇 워드 프로세서와 다른 도구들의 파일 포맷들로, 동일 정보를 그 포맷에 투명한 형식들로 나타냄으로써 그 동일 정보를 터널링("tunnel")하는 것이 가능하다. 예를 들어, 도구가, 임베딩된 커멘트 또는 메타데이터, '보이지 않는'(invisible) 텍스트, '무시할 수 있는'(ignorable) 속성들(attributes) 또는 다른 유사한 특징들을 지원하는 경우, 여기에 기술된 정보가 그들 내에 감춰질 수 있어서, 결과적인 문서가 그 도구에서 사용되고 어쩌면 수정될 수 있게도 해주고 또한 그 결과적인 문서가 터널링된 정보가 여전히 이용가능한 상태로 시스템으로 반환될 수 있도록 해준다.

[0182] 14) 피드백 모델 프론트엔드(Feedback Model Front End)

[0183] 이미 기술한 방법들에 의해 문서들과 그 발견된 청크들에 붙여진 방대한 주석과 분석은, 사용자들을 편집 샘플들(editing samples), 템플릿들(templates) 또는 이전의 문서들로 안내하여 현재의 필요에 맞춰진 유사하나 신규한 문서들을 생성하도록 하는 것을 가능하게 한다. 예컨대, 이 시스템은, 계약에 연루된 당사자들 및 재산, 병력(medical history)이라는 제목 하에 나열된 약물(medications) 또는 질병(conditions), 현재의 소견(current findings) 및 의료 기록의 다른 특별한 절들(other specific sections of clinical notes), 관련된 날짜들 등등을 전형적으로 식별한다. 동일한 문서 세트의 다른 문서를 조사함으로써, 이 시스템은, 어떤 것들이 일반적이 아니고, 일반적이고 또는 필요한 것인지를 학습하고, 따라서 무엇을 검토하고 및/또는 갱신할지에 관해 사용자에게 더욱 유용한 제안을 할 수 있다. 예를 들어, 문서 세트에서의 거의 모든 계약서에 유효 일자(effective date)가 존재할 수 있으나, 그 값은 각각에 있어서 다를 수 있다. 마찬가지로, 당사자들이 달라질 수 있으나 당사자들의 부류는 더욱 일관되어 있다.

[0184] i) 비유도 피드백(unguided feedback)

[0185] 사용자와 상호 작용함에 있어서, 시스템은, 몇몇 문서에서 발견된 (또는 발견되지 않았을 수도 있는) 청크들에 관한 피드백을 먼저 요청한다. 피드백을 위해 제시된 처음 몇몇 문서들은, 문서 세트에 대한 '클러스터 중심들'(cluster centroids)일 것이다. 마지막 몇 개는 문서 세트에서 '아웃라이어들'(outliers)일 것이다.

[0186] ii) 유도 피드백(guided feedback)

[0187] 이것 이후에 시스템은 사용자들에게 문서의 선택된 부분들을 보여주고 이들에 대한 현재 있는 또는 있을 가능성이 있는 레이블들, 이들의 범위(extent) 등에 관해 질의함으로써 사용자로 하여금 피드백을 제공하라고 유도한

다.

[0188] a. '관심을 끄는 레이블들(interesting labels)이 페이지랭크(PageRank) 기반의 알고리즘과 문법적 및 구조 모델에 의해 결정된다. 그러한 레이블들 중에서 낮은 신뢰도의 인스턴스들(instances)의 세트가 검토를 위해 선택된다.

[0189] b. 현재의 문서에 낮은 신뢰도의 레이블들이 더 이상 없는 경우, 동일한 프로세스가 추가의 문서들에 대해 반복될 수 있다. 몇몇 실시예들에서, 사용자가 제공하는 피드백에 기초하여 모델들이 연속적으로 갱신된다. 그러나 피드백은, 대신에 축적되어 배치로(in batches) 및/또는 오프라인으로(offline) 후에 적용될 수 있다. 모델들에 대한 조정은, 피드백을 위해 그 후 제시되는 청크들과 레이블들의 선택에 영향을 미칠 수 있고, 몇몇 문서들의 재분석을 유발할 수 있다.

[0190] c. 시스템은 실질적으로 동일한 매카니즘을 이용하여 필드 및 구조 청크들에 관한 피드백을 요청한다. 하나의 접근 방식에 있어서, 모든 청크 검출기들(chunk detectors)이, 피드백을 위한 후보들을 선택하기 위해 다른 정보와 함께 사용될 수 있는 신뢰도 추청치들(estimates of confidence)을 제공한다. 더 작은 대 더 큰 청크들, 필드 대 구조 청크들에 대한 서로 다른 패스들에서(in different passes for smaller vs. larger chunks, field vs. structure chunks) 또는 다른 순서들로 피드백이 요청될 수 있다. 사용자 피드백에 대한 예시적인 사용자 인터페이스에 대해서는 도 3을 참조하기 바란다. 이는 몇몇 또는 모든 청크들을 표시하고, 사용자로 하여금 할당된 타입 및/또는 역할 그리고 선택적으로 대안들을 보면서 검사할 특정의 것들을 선택하도록 해준다. 사용자는 청크 경계들을 이동시키고 레이블들을 선택하거나 편집하는 등등을 할 수 있다. 바람직하게, 사용자들은 레이블에 대한 것과 같은 특정 변경이 모든 대응하는 또는 동일 유형의 청크들에 적용될 것을 또한 요청할 수 있다.

[0191] 15) 피드백 응답(feedback response)

[0192] i) 플리트 쿼링(fleet querying)은 시스템으로 하여금 전형적으로는 복수의 사용자로부터의 사용자 피드백에 기초하여 사적인 데이터 및 공적인 데이터 모두를 쿼리하도록 해주는 방법이다. 선택된 예들은, 의미론적으로나 구문적으로(syntactically) 이전의 실패 케이스들(failure cases)에 유사한데, 이는 피드백의 가치를 증가시킨다.

[0193] ii) 디스패쳐(dispatcher). 디스패쳐는, 복수의 ML 모델과 비 ML(non-ML) 알고리즘들의 결합된 출력에 대한 사용자 피드백을, 피드백으로부터 학습할 수 있는 특정 학습 모델들(120)로 다시 연결시키기 위한 방법론이다.

[0194] 본 시스템은, 모델들로 하여금 그 출력에 관한 사용자 피드백으로부터 그리고 다른 학습 및 비학습(non-learning) 모델들의 출력에 대한 사용자 피드백으로부터 개선하도록 해준다. 이는, 기술된 여러 수치적 및 신경적 모델들(numerical and neural models)에 대한 ('미세 조정'(fine-tuning)이라고도 불리는) 증분의(incremental) 학습 데이터로서 피드백을 사용함으로써 성취될 수 있다. 모델들을 개선하기 위해 피드백이 사용된 이후, 특정 문서가 재평가될 뿐만 아니라 세트에 있는 모든 문서들 또는 심지어 사용자를 위한 모든 문서들이 재평가된다. 따라서, 각각의 문서에 대한 피드백은, 모든 문서들에 대한 청크 식별, 역할 할당, 구조 발견 및 사용자 보조(user assistance)를 개선할 수 있다. 이러한 재학습(retraining)은, 도 1에서 단계(15)로부터 단계(3)까지의 점으로 된 연결자(dotted connector)에 의해 표현된다.

[0195] 문서 및 연관된 모든 정보는, 학습과 문서들의 세트들의 분석(특별히이긴 하나 전적으로 특정 문서 세트들 내에서는 아님)에 기여하고, 따라서 향후의 문서들(future documents)에 대한 성능을 개선한다. 예를 들어, 신규 청크가 세트에서의 하나 또는 둘 이상의 문서들에 추가되면, 이는 향후의 문서들(또는 오래된 것들의 개정판들(revisions))에서 사용가능하고 향후의 문서들 용으로 제안될 수 있다. 어느 시점에서, 최근에 도입된 청크의 역할이 없다거나 최근에 덜 사용되는 청크의 역할이 있는 것은, 변칙이 될 수 있다. 이 시점은, 사용자가 솔선수범하여 스스로 선택할 수도 있고 피드백 질의에 응답하여 선택될 수도 있고 시간에 따른 카운터파트 청크들의 사용 커브(usage curve)에 기초하여 자동으로 선택될 수도 있다. 예를 들어, 어느 시점 이전에 작성된 하나의 세트에서의 극소의 문서들만이 주어진 역할 및/또는 컨텍스트(즉, '배제'(Exclusions) 절)을 가지거나 어떠한 문서도 이를 가지지 않으나 그 후에 작성된 대부분의 또는 모든 문서들이 그것을 가지는 경우, 카운터파트 청크가 없다는 것은 신규 문서들에 있어서는 변칙적인 것일 가능성이 있고 사용자에게 유용하게 그렇게 제안될 수 있다.

[0196] 16) 후속 통신(downstream communication)

[0197] 기술된 바와 같이 청크 정보로 문서(들)을 주석 처리하였으면, 선택된 정보가 데이터베이스, 분석 도구

(analytics tools) 등과 같은 외부 비즈니스 정보 시스템에 의해 요구되는 특정 포맷들로 변환되고 이들 시스템들로 직접적으로 또는 자동 및/또는 수동 검토 단계들을 통해서 전달된다. 예를 들어, 특정 당사자들의 이름들 및 주소는 데이터베이스에서의 올바른 필드들(fields)에 복사될 수 있는데, 이는, 이들이 '이름들'(names) 및 '주소들'(addresses) 그 자체로서만 식별되었다면 자동으로 이루어질 수 없었을 것이다. 후속 소프트웨어 애플리케이션과의 통합의 예에 대해서는 도 4를 참조하기 바란다. 이 예에서는, 당사자가 찬성할 것이라고 예상되는 조건들(terms)을 나타내는 청크들이 추출되었고, 이들은 기입하고 서명할 Docusign과 유사하게 후속 애플리케이션으로 전달된다.

[0198] 도 5는 본 발명과 함께 사용될 수 있는 컴퓨터 시스템(510)의 일 실시예의 블록도를 나타낸 도면이다. 전술한 단계들은 그러한 컴퓨터 시스템 상에서 실행되는 소프트웨어에 의해 구현될 수 있다. 컴퓨터 시스템(510)은 버스 서브시스템(512)을 통해 주변 장치들과 통신하는 적어도 하나의 컴퓨터 또는 프로세서(514)를 전형적으로 포함한다. 전형적으로, 컴퓨터는, 마이크로프로세서, 그래픽스 프로세싱 유니트 또는 디지털 신호 처리기 및 ASIC(Application Specific Integrated Circuit) 또는 FPGA(Field Programmable Gate Array)와 같은 위의 것들의 전자적 처리 동등물들(electronic processing equivalents) 중 임의의 것을 포함할 수 있고, 프로세서도 그런 것일 수 있다. 이러한 주변 장치들은, 메모리 서브시스템(526) 및 파일 저장 서브시스템(528)을 포함하는 저장 서브시스템(524), 사용자 인터페이스 입력 장치들(522), 사용자 인터페이스 출력 장치들(520) 및 네트워크 인터페이스 서브시스템(516)을 포함할 수 있다. 입력 및 출력 장치들은 컴퓨터 시스템(510)과의 사용자 상호작용을 가능케 한다.

[0199] 컴퓨터 시스템은, 서버 컴퓨터, 클라이언트 컴퓨터, 워크스테이션, 메인프레임, 퍼스널 컴퓨터(PC), 태블릿 PC, 랙 탑재형 '블레이드'(rack-mounted "blade") 또는 임의의 데이터 프로세싱 머신일 수 있다 - 여기서 임의의 데이터 프로세싱 머신은, 해당 머신이 취할 동작들을 규정하는 (순차적 또는 그렇지 않은) 명령어들을 실행할 수 있음 -.

[0200] 컴퓨터 시스템은, 마이크로소프트 사의 윈도우즈, 선 마이크로시스템 사의 솔라리스(Solaris), 애플 컴퓨터 사의 MacOS, 리눅스 또는 유닉스와 같은 운영체제를 전형적으로 포함한다. 컴퓨터 시스템은, 또한 BIOS(Basic Input/Output System) 및 프로세서 펌웨어를 전형적으로 포함할 수 있다. 운영체제, BIOS 및 펌웨어는 프로세서에 연결된 서브시스템들 및 인터페이스들을 제어하기 위해 프로세서에 의해 사용된다. 이러한 운영체제들과 호환가능한 전형적인 프로세서들은, 인텔 사의 펜티엄(Pentium) 및 이타니엄(Itanium), AMD(Advanced Micro Devices) 사의 옵테론(Opteron) 및 애슬론(Athlon) 및 ARM 홀딩스 사의 ARM 프로세서를 포함한다.

[0201] 청구된 발명의 혁신들, 실시예들 및/또는 예들은, 전통적인 컴퓨터 애플리케이션들로 제한되는 것이 아니고 이들을 실행하는 프로그램가능한 장치로 제한되는 것도 아니다. 예를 들어, 청구된 것의 혁신들, 실시예들 및/또는 예들은, 광 컴퓨터, 양자 컴퓨터(quantum computer), 애널로그 컴퓨터 또는 이와 유사한 것을 포함할 수 있다. 컴퓨터 시스템은, 멀티 프로세서 또는 멀티 코어 시스템일 수 있고, 분산 또는 원격 시스템을 사용하거나 그에서 구현될 수 있다. 여기서, '프로세서'라는 용어는, 단일 처리기(singular processor) 및 그래픽 프로세싱 유니트들, 디지털 신호 처리기들, 디지털 프로세서들 및 이들 장치들의 조합들을 포함하는 멀티 코어 또는 멀티 프로세서 어레이들을 포괄하도록 가장 넓은 의미로 사용된다. 더욱이, 단지 하나의 컴퓨터 시스템 또는 하나의 머신이 예시될 수 있지만, 그러한 용어들의 단수형의 사용은, 여기서 논의된 임의의 하나 또는 둘 이상의 동작들을 수행하기 위한 명령어들을 개별적으로 또는 연대하여 실행하는 컴퓨터 시스템들 또는 머신들의 집합을 뜻하는 것일 것이다. 컴퓨터들 및 네트워크들의 끊임없이 변화하는 성향으로 인해, 도 5에 표시된 컴퓨터 시스템(510)의 설명은, 바람직한 실시예들을 예시할 목적의 하나의 예로서만 의도된 것이다. 도 5에 도시된 컴퓨터 시스템 보다 많거나 적은 개수의 소자들을 가지는, 컴퓨터 시스템(510)의 다른 많은 구성들이 가능하다.

[0202] 네트워크 인터페이스 서브시스템(516)은, 통신 네트워크(518)에 대한 인터페이스를 비롯한, 외부 네트워크들에 대한 인터페이스를 제공하고, 통신 네트워크(518)를 통해 다른 컴퓨터 시스템들 또는 머신들에서의 대응하는 인터페이스 장치들에 결합된다. 통신 네트워크(518)는 다양한 상호 접속된 컴퓨터 시스템들, 머신들 및 통신 링크들을 포함할 수 있다. 이러한 통신 링크들은, 유선 링크들(wireline links), 광 링크들, 무선 링크들 또는 정보 통신을 위한 기타 임의의 장치들일 수 있다. 통신 네트워크(518)는, 임의의 적절한 컴퓨터 네트워크(예를 들어, 인터넷과 같은 광역망(wide area network) 및/또는 이더넷과 같은 근거리 통신망)일 수 있다. 통신망은, 유선 및/또는 무선일 수 있고, 통신망은, 가상사설망(virtual private network)에서 사용가능한 것과 같은 암호화 및 복호화 방법들을 사용할 수 있다. 통신 네트워크는, 다른 시스템들로부터 데이터를 수신하고 그로 데이터를 전송할 수 있는 하나 또는 둘 이상의 통신 인터페이스들(communications interfaces)을 사용한다. 통신 인터페이스들의 실시예들은, 이더넷 카드, 모뎀(예를 들어, 전화기, 위성, 케이블 또는 ISDN), (비동기식) 디지

털 가입자 회선(digital subscriber line: DSL) 유니트, 파이어와이어(Firewire) 인터페이스, USB 인터페이스 및 그와 유사한 것들을 전형적으로 포함한다. HTTP, TCP/IP, RTP/RTSP, IPX 및/또는 UDP와 같은 하나 또는 둘 이상의 통신 프로토콜들이 사용될 수 있다.

[0203] 사용자 인터페이스 입력 장치들(522)은, 알파뉴머릭(alphanumeric) 키보드, 키패드, 마우스, 트랙볼, 터치패드, 스틸러스(stylus) 또는 그래픽스 태블릿과 같은 포인팅 장치들(pointing devices), 스캐너, 디스플레이에 결합된 터치스크린, 음성 인식 시스템들 또는 마이크로폰들과 같은 오디오 입력 장치들, 시선 인식기(eye-gaze recognition), 뇌파 패턴 인식기(brainwave pattern recognition) 및 다른 유형의 입력 장치들을 포함할 수 있다. 이러한 장치들은, 컴퓨터 시스템에 유선으로 또는 무선으로 연결될 수 있다. 일반적으로, '입력 장치들'(input devices)이라는 용어를 사용한 것은, 정보를 컴퓨터 시스템(510)으로 또는 통신 네트워크(518)로 입력하기 위한 모든 가능한 유형의 장치들 및 방식들을 포괄하기 위한 의도에서이다. 사용자 인터페이스 입력 장치들은, 사용자로 하여금, 예를 들어 디스플레이 서브시스템과 같은 몇몇 유형의 사용자 인터페이스 출력 장치들 상에 나타나는, 객체들, 아이콘들, 텍스트 및 그와 유사한 것들을 선택하도록 해준다.

[0204] 사용자 인터페이스 출력 장치들(520)은, 디스플레이 서브시스템, 프린터 또는 오디오 출력 장치들과 같은 비 시각적(non-visual) 디스플레이들을 포함할 수 있다. 디스플레이 서브시스템은, 액정 디스플레이(liquid crystal display: LCD)와 같은 평판 디스플레이(flat-panel display), 프로젝션 장치 또는 가상현실 시스템과 같은 시각적 이미지를 생성하기 위한 여타의 장치를 포함할 수 있다. 디스플레이 서브시스템은, 오디오 출력 또는 촉각(tactile) 출력(예컨대, 진동) 장치들과 같은 비 시각적 장치를 또한 제공할 수 있다. 일반적으로, '출력 장치'(output device)라는 용어를 사용한 것은, 정보를 컴퓨터 시스템(510)으로부터 사용자로 또는 다른 머신 또는 컴퓨터 시스템으로 출력하기 위한 모든 가능한 유형의 장치들 및 방식들을 포괄하기 위한 의도에서이다.

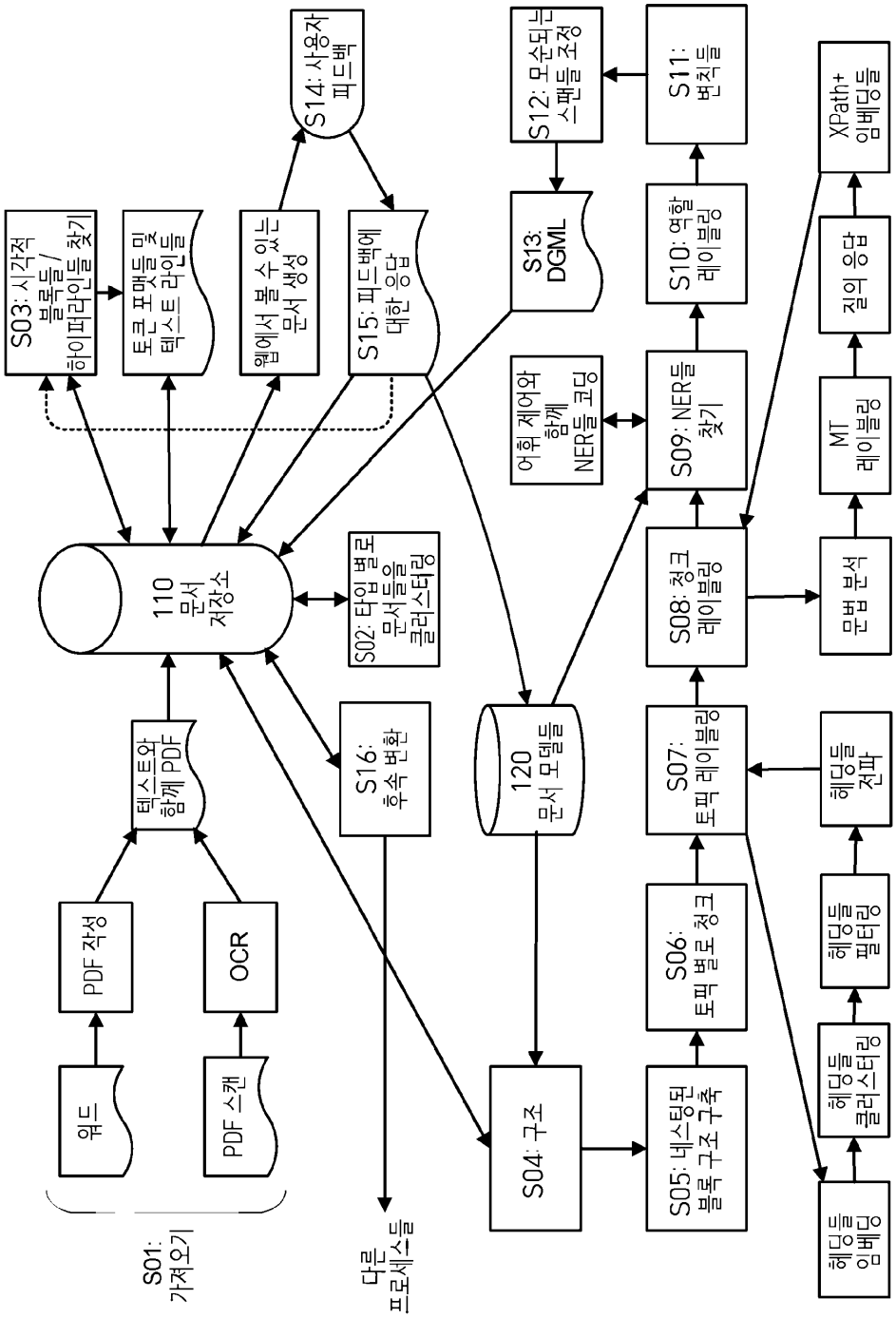
[0205] 메모리 서브시스템(526)은, 전형적으로 프로그램을 실행하는 동안 명령어들과 데이터를 저장하기 위한 메인 랜덤 액세스 메모리(RAM)(530) (또는 기타의 휘발성의 저장 장치) 및 고정된 명령어들이 저장되는 판독 전용 메모리(ROM)(532)를 비롯한 복수의 메모리를 전형적으로 포함한다. 파일 저장 서브시스템(528)은, 프로그램 및 데이터 파일들을 위한 지속적 저장 기능을 제공하고, 하드 디스크 드라이브, 연관된 이동식 매체(removable media)와 함께의 플로피 디스크 드라이브, CD-ROM 드라이브, 광 드라이브, 플래시 메모리 또는 이동식 매체 카트리지(removable media cartridges)를 포함할 수 있다. 어떤 실시예들의 기능을 구현하는 데이터베이스들 및 모듈들은 파일 저장 서브시스템(528)에 의해 저장될 수 있다.

[0206] 버스 서브시스템(512)은, 컴퓨터 시스템(510)의 다양한 소자들 및 서브시스템들이 의도한 바대로 서로 통신하도록 해주기 위한 장치를 제공한다. 버스 서브시스템(512)이 단일 버스로서 개괄적으로 도시되어 있지만, 이 버스 서브시스템의 대안적 실시예들은 복수의 버스를 사용할 수 있다. 예를 들어, RAM 기반의 메인 메모리가 DMA(Direct Memory Access) 시스템들을 이용하여 파일 저장 시스템들과 직접적으로 통신할 수 있다.

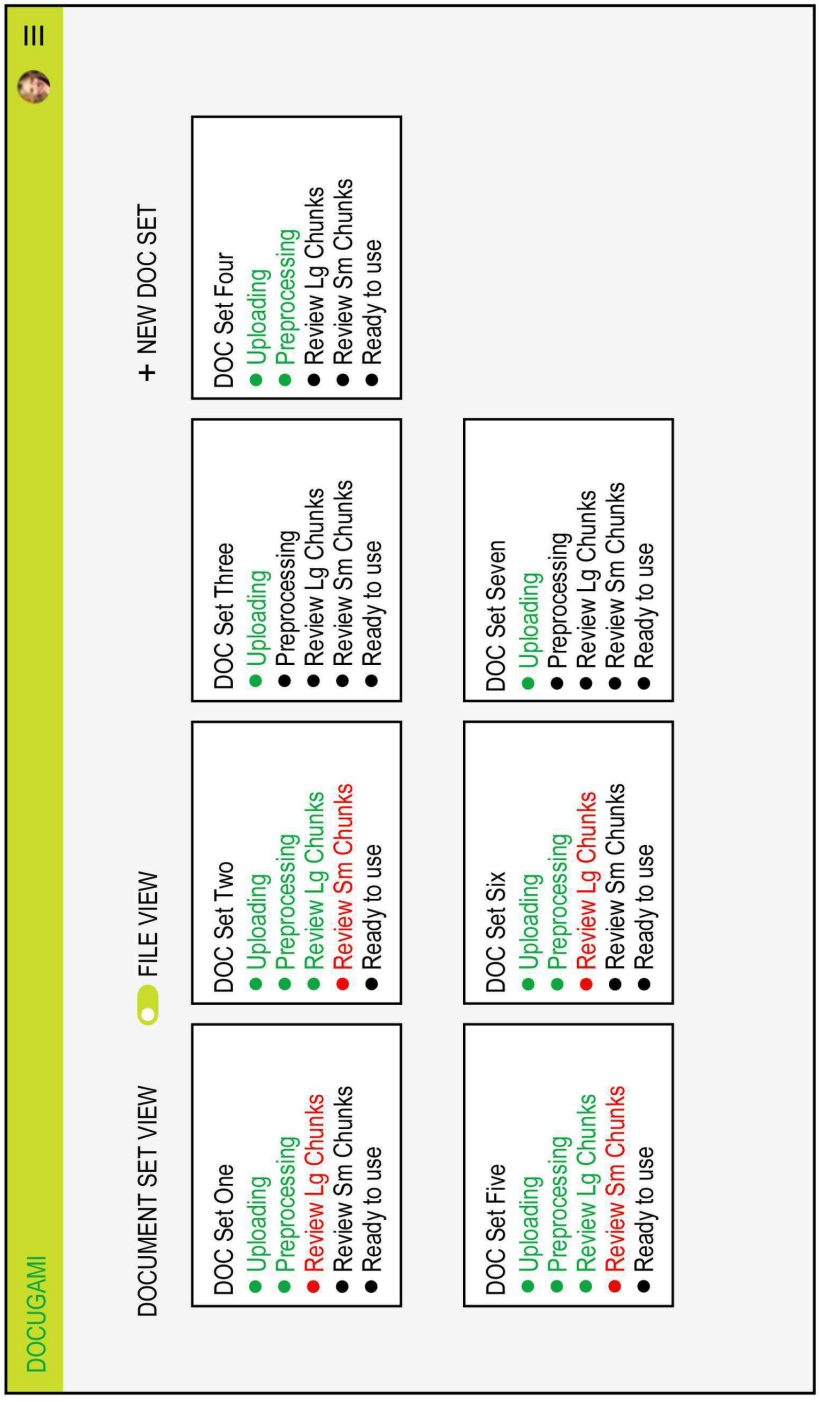
[0207] 본 상세한 설명은 많은 특별한 것들을 포함하고 있지만, 이러한 것들이 본 발명의 범위를 제한하는 것으로 해석되어서는 아니되고 단지 여러가지 예들을 예시하는 것으로서 해석되어야 한다. 본 개시의 범위는 앞서 상세히 논의되지 않은 다른 실시예들을 포함하는 것으로 이해되어야 한다. 본 기술 분야에 숙련된 자들에게 명백해질 다양한 여타의 수정들, 변경들 및 변형들이, 첨부된 청구항들에 정의된 정신과 범위에서 벗어나지 않고 여기에 개시된 방법 및 장치의 구성, 동작 및 상세한 사항들 내에서 이루어질 수 있다. 따라서 본 발명의 범위는 첨부된 창구항들과 그들의 법적 균등물들에 의해 결정되어야 한다.

도면

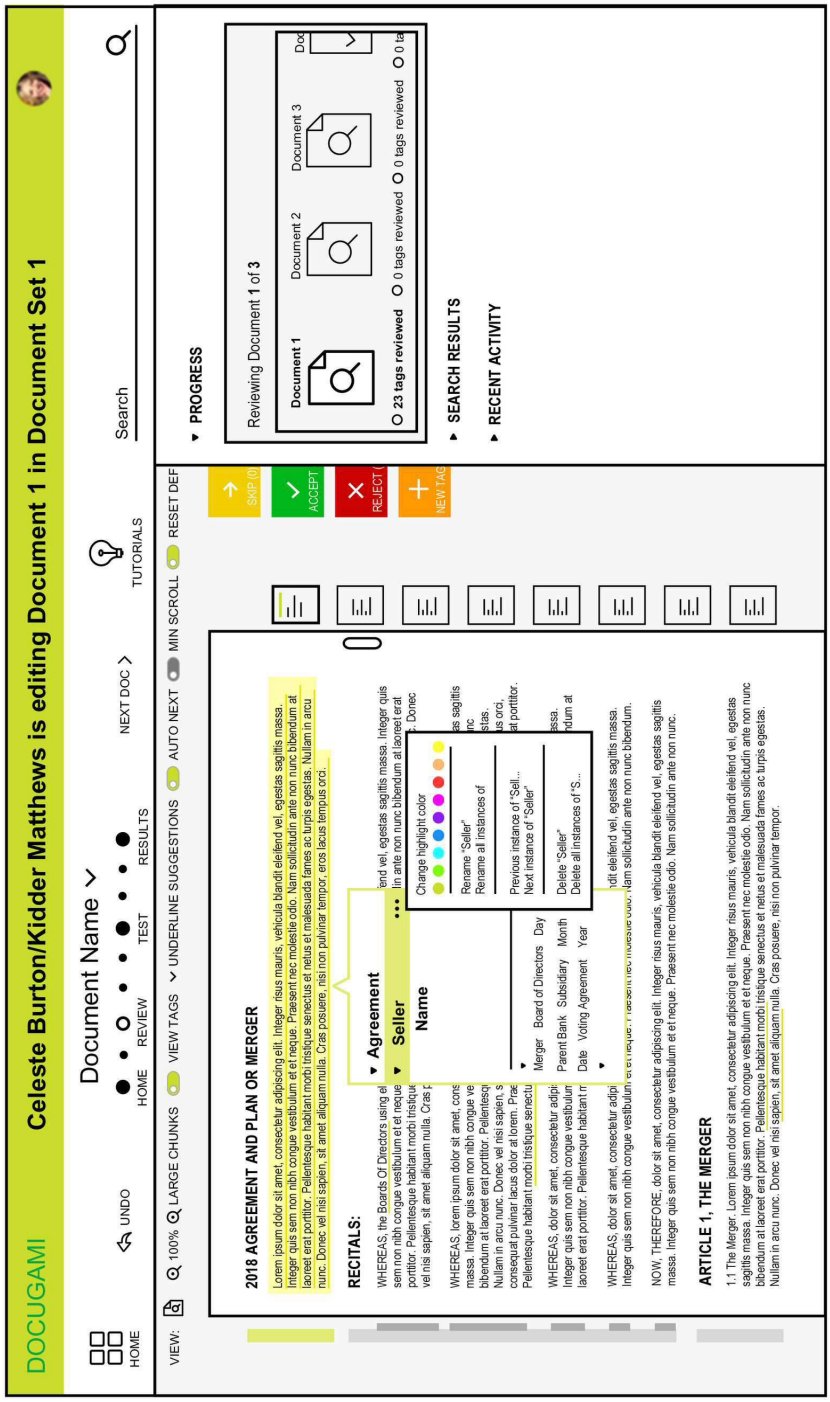
도면1


S01: 가져오기
워드
PDF 스캔
PDF 작성
OCR
텍스트와 함께 PDF
110 문서 저장소
S02: 타입 별로 문서들을 클러스터링
S03: 시각적 블록들 / 하이퍼라인들 찾기
토큰 포맷들 및 텍스트 라인들
웹에서 볼 수 있는 문서 생성
S14: 사용자 피드백
S15: 피드백에 대한 응답
S16: 후속 변환
다른 프로세스들
120 문서 모델들
S04: 구조
S05: 네스팅된 블록 구조 구축
S06: 토픽 별로 청크
S07: 토픽 레이블링
S08: 청크 레이블링
S09: NER들 찾기
어휘 제어와 함께 NER들 코딩
S10: 역할 레이블링
S11: 변칙들
S12: 모순되는 스팬들 조정
S13: DGML
헤딩들 임베딩
헤딩들 클러스터링
헤딩들 필터링
헤딩들 전파
문법 분석
MT 레이블링
질의 응답
XPath+ 임베딩들

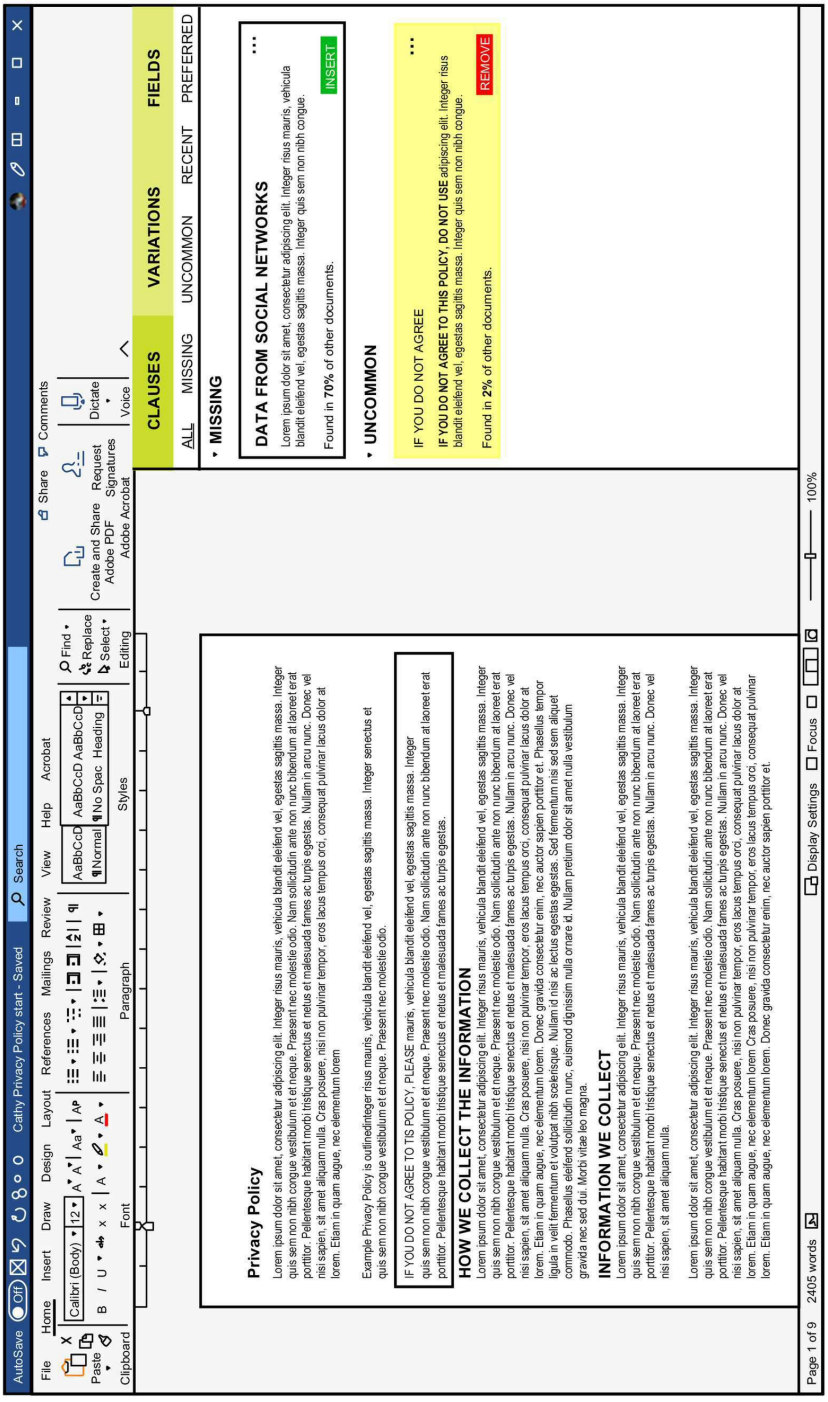
도면2


DOCUGAMI
DOCUMENT SET VIEW
FILE VIEW
+ NEW DOC SET
DOC Set One
Uploading
Preprocessing
Review Lg Chunks
Review Sm Chunks
Ready to use
DOC Set Two
Uploading
Preprocessing
Review Lg Chunks
Review Sm Chunks
Ready to use
DOC Set Three
Uploading
Preprocessing
Review Lg Chunks
Review Sm Chunks
Ready to use
DOC Set Four
Uploading
Preprocessing
Review Lg Chunks
Review Sm Chunks
Ready to use
DOC Set Five
Uploading
Preprocessing
Review Lg Chunks
Review Sm Chunks
Ready to use
DOC Set Six
Uploading
Preprocessing
Review Lg Chunks
Review Sm Chunks
Ready to use
DOC Set Seven
Uploading
Preprocessing
Review Lg Chunks
Review Sm Chunks
Ready to use

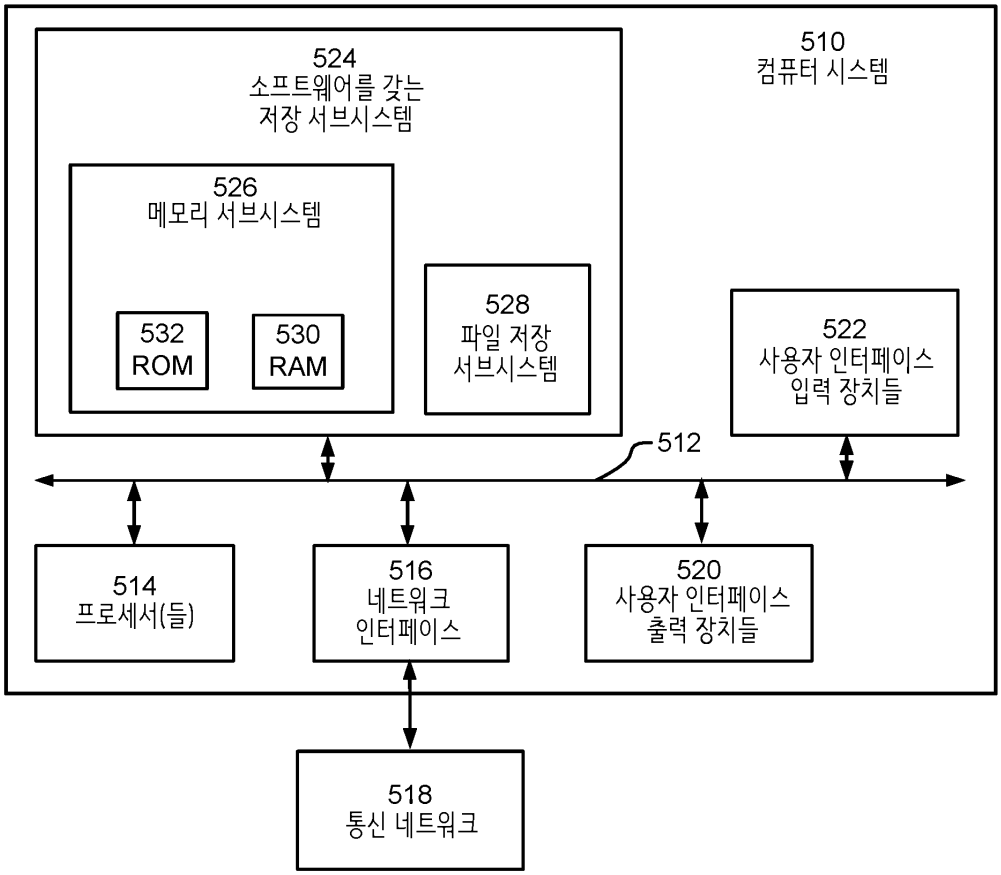
도면3


DOCUGAMI
Celeste Burton/Kidder Matthews is editing Document 1 in Document Set 1
HOME
UNDO
Document Name
HOME
REVIEW
TEST
RESULTS
NEXT DOC
TUTORIALS
Search
VIEW:
100%
LARGE CHUNKS
VIEW TAGS
UNDERLINE SUGGESTIONS
AUTO NEXT
MIN SCROLL
RESET DEF
2018 AGREEMENT AND PLAN OR MERGER
RECITALS:
Agreement
Seller
Name
ARTICLE 1. THE MERGER
SKIP (0
ACCEPT
REJECT (
NEW TAG
PROGRESS
Reviewing Document 1 of 3
Document 1
Document 2
Document 3
23 tags reviewed
0 tags reviewed
0 tags reviewed
SEARCH RESULTS
RECENT ACTIVITY

도면4


Privacy Policy
HOW WE COLLECT THE INFORMATION
INFORMATION WE COLLECT
CLAUSES
VARIATIONS
FIELDS
ALL
MISSING
UNCOMMON
RECENT
PREFERRED
MISSING
DATA FROM SOCIAL NETWORKS
INSERT
Found in 70% of other documents.
UNCOMMON
IF YOU DO NOT AGREE
IF YOU DO NOT AGREE TO THIS POLICY, DO NOT USE
REMOVE
Found in 2% of other documents.

도면5


524
소프트웨어를 갖는
저장 서브시스템
510
컴퓨터 시스템
526
메모리 서브시스템
532
ROM
530
RAM
528
파일 저장
서브시스템
522
사용자 인터페이스
입력 장치들
512
514
프로세서(들)
516
네트워크
인터페이스
520
사용자 인터페이스
출력 장치들
518
통신 네트워크