(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2012년08월23일
(11) 등록번호 10-1176772
(24) 등록일자 2012년08월18일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) *G06F 17/00* (2006.01)
(21) 출원번호 10-2009-0085995
(22) 출원일자 2009년09월11일
심사청구일자 2009년09월11일
(65) 공개번호 10-2011-0028077
(43) 공개일자 2011년03월17일
(56) 선행기술조사문헌
KR1020080029417 A
KR1020090078986 A
KR1020070007001 A
KR1020050032937 A
전체 청구항 수 : 총 8 항

(73) 특허권자
한국과학기술원
대전 유성구 구성동 373-1
(72) 발명자
최기선
대전광역시 유성구 대학로 291, 한국과학기술원 전산학과 (구성동)
최주원
대전광역시 유성구 대학로 291, 한국과학기술원 전산학과 (구성동)
(74) 대리인
이달로

심사관 : 석상문

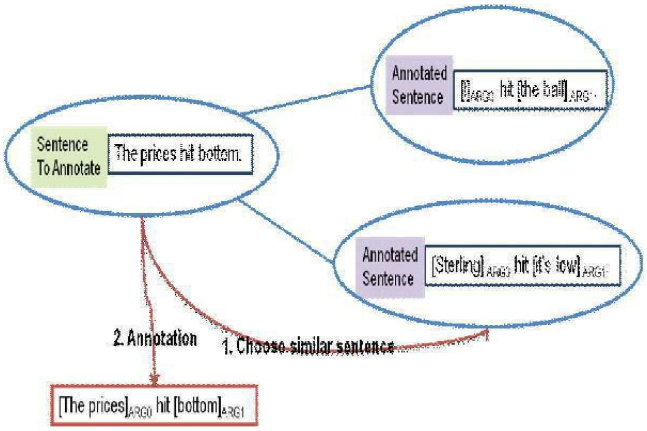
(54) 발명의 명칭 **문장 온톨로지 생성을 위한 자동 어노테이션 시스템 및 방법**

(57) 요 약

본 발명은 문장 온톨로지 생성을 위한 자동 어노테이션 시스템 및 방법에 관한 것으로, 본 발명에 따른 문장 온톨로지 생성을 위한 자동 어노테이션 시스템은 코퍼스의 어노테이션된 문장들을 적당한 단위로 분해해서 저장하는 분해 저장부, 입력 문장이 들어왔을 때, 상기 분해 저장부로부터 분해된 문장 조각중에 의미, 구조적으로 유사한 조각을 찾아 해당 조각에 어노테이션된 내용을 참조하여 입력문장을 어노테이션하는 어노테이션부를 포함한다.

또한, 본 발명에 따른 문장의 부분 유사도 기반 어노테이션 방법은 문장을 비교 가능한 단위로 분해하는 제1 과정, 상기 문장 조각의 구조적 유사도를 측정하는 제2 과정, 논항들이 의미상으로 얼마나 유사한지 알아보기 위해 상기 문장 조각의 의미적 유사도를 측정하는 제3 과정, 및 상기 문장 유사도를 기반으로 하여 어노테이션을 수행하는 제4 과정을 포함한다.

대 표 도 - 도6



특허청구의 범위

청구항 1

펜-트리뱅크(Penn-Treebank)를 포함한 코퍼스의 어노테이션된 문장들을 동사를 포함하는 문장 단위로 분해해서 저장하는 분해 저장부, 및

입력 문장이 들어왔을 때, 상기 분해 저장부로부터 분해된 문장 조각중에 의미, 구조적으로 유사한 조각을 찾아 해당 조각에 어노테이션된 내용을 참조하여 입력문장을 어노테이션하는 어노테이션부를 포함하는 문장 온톨로지 생성을 위한 자동 어노테이션 시스템.

청구항 2

제1항에 있어서,

상기 어노테이션부에서 동사를 중심으로 하는 술어-논항 구조의 부분만 유사도 비교가 되기 때문에, 한 문장에 여러 동사가 나오는 경우에는, 그 문장은 상기 분해 저장부에서의 첫번째 단계인 문장 분해 단계를 통해 각 동사를 중심으로 하는 조각들로 분해되는 것을 특징으로 하는 문장 온톨로지 생성을 위한 자동 어노테이션 시스템.

청구항 3

문장을 비교 가능한 단위로 분해하는 제1 과정,

상기 문장 조각의 구조적 유사도를 측정하는 제2 과정,

논항들이 의미적으로 얼마나 유사한지 알아보기 위해 상기 문장 조각의 의미적 유사도를 측정하는 제3 과정, 및

상기 문장 유사도를 기반으로 하여 어노테이션을 수행하는 제4 과정을 포함하는 문장 온톨로지 생성을 위한 자동 어노테이션 방법.

청구항 4

제3항에 있어서,

상기 제1 과정에서 분해된 상기 문장 조각은 두 가지 종류로 분류하되, 그중 하나는 다항 관계조각이고, 다른 하나는 이항 관계조각인 문장 온톨로지 생성을 위한 자동 어노테이션 방법.

청구항 5

제3항에 있어서,

상기 제1 과정에서 하나의 문장은 여러 개의 동사를 포함하고 있으므로, 동사의 개수만큼의 다항관계 조각으로 표현하되, 각 다항관계 조각은 논항의 개수만큼의 이항관계 조각으로 표현하는 문장 온톨로지 생성을 위한 자동 어노테이션 방법.

청구항 6

제3항에 있어서,

상기 제2 과정에서의 문장의 구조적 유사도 측정은 각 조각의 트리 경로에 대해서만 수행하며, Arg와 같은 요소들에 대해서는 수행되지 않는 문장 온톨로지 생성을 위한 자동 어노테이션 방법.

청구항 7

제3항에 있어서,

상기 제3 과정에서의 논항의 의미적 유사도 비교는 단어의 워드넷 상에서의 경로 유사도를 이용하며, 또한 품사의 순서 및 트리의 높이 등을 추가적으로 이용하여 논항의 유사도를 최종적으로 구하는 문장 온톨로지 생성을 위한 자동 어노테이션 방법.

청구항 8

제3항에 있어서,

상기 제4 과정에서의 어노테이션은 다항관계 조각, 이항관계 조각과 구조적 유사도, 의미적 유사도를 조합하여 수행하는 문장 온톨로지 생성을 위한 자동 어노테이션 방법.

명 세 서

발명의 상세한 설명

기 술 분 야

[0001] 본 발명은 문장 온톨로지 생성을 위한 자동 어노테이션 시스템 및 방법에 관한 것으로, 보다 상세하게는 특정 코퍼스 내에서 입력 문장과 유사한 문장을 찾고 해당 유사 문장의 술어-논항 구조를 그대로 가져와서 입력 문장을 어노테이션하는 문장 온톨로지 생성을 위한 자동 어노테이션 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 인간이 생활하며 얻는 지식은 여러 가지 형태로 표현되고 저장된다. 그것은 개인의 머릿속에만 존재하다 없어질 수도 있으며, 전래동화처럼 구전될 수도 있다.

[0003] 하지만, 많은 사람에게 유용한 지식은 대부분 후대에 전하기 위해 그림, 글의 형태로 남아 있다. 이중 글에 해당하는 것, 즉 문서가 담고 있는 지식을 컴퓨터가 처리 가능한 논리적 형태로 표현하는 것은 자연언어처리를 연구하는 학자들의 숙원이다.

[0004] 범위를 넓게 생각해 보면, 문서의 단어마다 품사를 부착하는 것, 구문 구조를 분석하는 것도 모두 컴퓨터가 문서의 내용을 이해할 수 있도록 만드는 방법의 하나의 예라고 할 수 있다.

[0005] 온톨로지는 컴퓨터가 이해할 수 있는 지식 표현의 한 형태로, 현재 방대한 정보를 저장하고 있는 월드와이드웹(World Wide Web)과 결합되고 있는 포맷이다.

[0006] 도 1은 온톨로지의 예를 도시한 도면이다.

[0007] 도 1에 도시된 바와 같이, 온톨로지는 기본적으로 개념(Concept), 및 개념 간의 관계(Relation)로 이루어진다. 도 1에서 타원형으로 된 도형들은 모두 개념(Concept)에 해당하는 것이고, 화살표는 모두 개념간의 관계(Relation)에 해당하는 것이다.

[0008] 예를 들어, (Person, wears, Apparel) -사람은 의복을 입는다.- 의 경우에는, 사람(person), 의복(apparel)이 각각 개념에 해당하고 사람과 의복 간에는 입다(wears)라는 관계가 있는 것이다.

[0009] 웹에는 수많은 문서가 쌓이고 있는데, 만약 이 문서들을 온톨로지로 표현하고 상위 온톨로지를 통해 서로 연결시킬 수 있다면 추론을 통해서 현재까지는 해 오지 못한 지능형 서비스를 제공할 수 있을 것이다. 이러한 측면에서 문서를 정확도 높게 온톨로지로 변환하는 일은 매우 의미 있는 일이다.

[0010] 그러나, 문장내의 개념을 찾고, 해당 개념들 간의 관계를 찾아내는 것은 매우 어려운 작업이다. 일단 문장 내에서 어떤 부분이 하나의 개념 단위를 이루는지를 찾는 작업도 어려운 문제이다.

[0011] 구절 묶음(Chunking, 표 1 참조), 전문용어 추출(Term Recognition, 표 2 참조) 등은 개념의 단위를 찾는 연구라고 볼 수 있는데, 각각이 하나의 어려운 문제를 형성하고 있다.

[0012] **표 1**

[0013]

```
(S    (NP-SBJ  Westinghouse Electric Corp))
   (VP  said
      (SBAR  (S  (NP-SBJ  it)    (VP  will  (VP  buy  (NP  Shaw-Walker  Co))))))    (.  .))
```

[0014] **표 2**

[0015]

[Westinghouse Electric Corp.]$_{Term}$ said it will buy [Shaw-Walker Co]$_{Term}$.

[0016] 개념 간의 관계를 찾는 것 또한 어려운 문제이다. 관계 추출(Relation Extraction)은 현재 정보 추출(Information Extraction)의 하나의 큰 분야를 차지하고 있을 정도이다.

[0017] 도 2는 자연언어 문서가 들어왔을 때, 문서의 각 문장들을 온톨로지 형태로 표현하는 것을 도시하고 있다. 즉, 문장마다 개념을 찾고, 개념간의 관계를 찾아내는 것을 도시하고 있다.

[0018] 술어는 고전 영문법에서는 주어와 함께 문장의 주요 두 성분 중의 하나로서, 주어를 제외한 문장의 모든 부분을 나타낸다. 하지만, 술어-논항 구조에서의 술어란, 고전적인 의미와는 다르게 대체로 동사를 칭하며, 논항은 술어에 대응되는 주어나 목적어 또는 수식어를 지칭한다. 즉, 술어-논항 구조란, 아래의 표 3에 나타낸 바와 같이, 특정 동사와 그 동사와 대응하는 주어, 목적어, 수식어를 묶어서 지칭하는 것이다.

[0019] **표 3**

[0020]

PRED: 술어, ARG: 논항
(1) [John]$_{ARG0}$ broke$_{PRED}$ [the window]$_{ARG1}$
(2) [The window]$_{ARG1}$ broke$_{PRED}$

[0021] 상기 표 3은 술어-논항 구조의 예를 나타낸 것으로, '부수다'라는 뜻을 가지는 'break'라는 동사는 보통 '부수는 사람'과 '부서지는 것'을 논항으로 가질 수 있다.

[0022] 위의 예제에서는 '부수는 사람'을 ARG0라고 표시하였고, '부서지는 것'을ARG1이라고 표시하였다. 표 3의 (1)번 예제에서는 'break'가 술어가 되고, 'John'이 '부수는 사람', 'the window'가 '부서지는 것'이 됨을 알 수 있다. 표 3의 (2)번 예제의 경우는 'broke'가 자동사로 쓰여 '부서지다'의 뜻으로 쓰였고, 이에 'the window'가 주어의 위치에 오게 되었고, 의미상으로는 '부서지는 것'에 대응하므로 ARG1으로 표시되었다.

[0023] 술어-논항 구조는 문서 기반 온톨로지 생성에 기초가 되는 자료로 사용될 수 있다. 기본적으로 논항은 온톨로지에서 사용되는 개념과 비슷한 속성을 가지고 있고, 술어의 경우는 온톨로지의 관계와 연관 지어 생각해 볼 수 있다.

[0024] 이렇게 술어, 논항을 각각 개념, 관계와 연관시킴으로써 문서에 나타난 각 문장을 온톨로지 형태로 변환할 수 있는 토대를 마련할 수 있다. 또한, 술어-논항 구조는 특히 이벤트 구조의 어노테이션을 하기 위한 부트스트랩(Bootstrap)자료로서 의미가 있다[P. Kingsbury et al., "PropBank as a Bootstrap for Richer Annotation Schemes" The 6th Workshop on Interlinguas: Annotations and Translations,, in conjunction with MT Summit 2003, New Orleans, Louisiana, 9 pages, 2003].

[0025] 술어-논항 구조를 자동으로 어노테이션할 수 있다면, 웹에서 만들어지고 있는 수많은 자연언어 문서들을 이벤트 구조로 어노테이션할 수 있는 기초적 형태가 갖추어 지게 되는 것이다. 이는 현재 쏟아지는 각종 뉴스 또한 일정부분 자동으로 의미적 분석을 할 수 있다는 것을 말한다. 즉, 사건 간의 원인-결과 분석, 이벤트 어노테이션을 통한 각종 사건의 검색 등의 성능의 향상에 술어-논항 구조가 사용될 수 있다.

[0026] 프롭뱅크(Propbank)는 술어-논항 구조 어노테이션되어 있는 코퍼스 중 하나로, 펜-트리뱅크(Penn-Treebank)의 월스트리트(Wall-Street) 저널 기사에 대해 술어-논항을 어노테이션하였다[M. Palmer et al., "The Proposition Bank: An Annotated of Semantic Roles", Computational Linguistics, Vol. 31. No. 1. pp. 71-

106, 2005].

[0027] 텍스트 기반 온톨로지 생성에도 두 가지 다른 관점이 있다.

[0028] 첫번째 방법은 특정 분야의 텍스트로부터 잘 정련된 하나의 온톨로지를 얻어내고자 하는 관점이다. 만약 '스포츠' 도메인에 해당하는 온톨로지를 만들고자 하는 경우, '스포츠' 도메인에 해당하는 문서를 수집하고, 해당 문서에서 전문 용어의 계층구조를 얻은 후, 전문 용어 간의 관계를 추출하는 과정을 거친다.

[0029] 이 방법론은 전형적인 도메인 온톨로지 구축과정[N. Noy and D. L. McGuiness, "Ontology development 101: A guide to creating your first ontology", Technical Report KSL-01-05 and SMI-2001-0880, Stanford Knowledge Systems Laboratory and Stanford Medical Informatics, 2001]을 컴퓨터로 그대로 재현하고자 하는 방법으로, 문장의 의미적 손실 여부와 관계없이, 하나의 일관된 형태의 온톨로지를 만들고자 하는 방법이다.

[0030] 두번째 방법은 문장 그 자체의 구조 및 의미를 표현하는 데 집중하는 온톨로지 구축 방법이다. 이는 문장을 논리화하는 방법의 하나로써 온톨로지를 이해한다. 이 방법론에서는 문장의 의미를 손실하지 않고 최대한 그대로 나타내는 것을 목적으로 하며, 상위 온톨로지와의 매핑 과정을 통해 문장을 이해할 수 있도록 한다. 오메가 온톨로지(Omega Ontology)는 워드넷을 확장하고 재구성한 온톨로지로써 문장을 이해할 수 있는 문법적 요소들을 같이 제공한다[A. Philpot et al, "The omega ontology", In Proceedings of the ONTOLEX Workshop at the International Conference on Natural Language Processing, pp. 59-66, 2005].

[0031] TextToOntto는 특정 문서 집합으로부터 도메인 온톨로지를 생성하는 프로그램이다[D. Oberle, R. Volz, B. Motik, and S. Staab, "An extensible ontology software environment", In S. Staab and R. Studer, editors, Handbook on Ontologies, International Handbooks on Information Systems, chapter III, pp. 311-333, 2004].

[0032] 텍스트 마이닝(mining) 기술을 이용하여 온톨로지 공학(engineering)에 정의되어 있는 단계를 도와주는 역할을 한다. 전형적인 온톨로지 구축 방법론에 따르면, 온톨로지 구축과정은 다음과 같이 분류된다.

[0033] 첫째, 도메인을 확정한다. 둘째, 현재 존재하는 온톨로지를 찾아본다. 셋째, 도메인에서 자주 나오는 전문용어를 수집한다. 넷째, 용어의 계층구조(hierarchy)를 생성한다. 다섯째, 클래스의 속성(Property)을 정의한다. 여섯째, 클래스 간의 관계를 설정한다.

[0034] TextToOnto는 위의 과정들을 반 자동화하여 온톨로지 구축의 시간을 단축시키고자 하는 툴이다. 이 경우에는 문서 집합에 담겨있는 내용 중 도메인에 적합하게 사용될 수 있는 부분만 온톨로지 형태로 변환되게 된다.

[0035] 도 3에서 볼 수 있듯이, 'CheesyPizza is a kind of pizza.'라는 문장은 피자 온톨로지의 계층구조에 반영되었다. 두 번째 문장인 'Joo-won likes it very much.'라는 문장은 개별적 사실로서, 피자 도메인 온톨로지에는 영향을 미치지 않았다.

[0036] 오메가 온톨로지는 워드넷[Christiane Fellbaum, editor., "WordNet: An Electronic Lexical Database", MIT Press, 1998.]을 재구성하고 확장한 온톨로지이다. 워드넷 뿐만 아니라, FrameNet[C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet Project", In proceedings of COLLING/ACL, pp 86-90, 1998], Propbank 등의 동사의 골격(frame)을 가지고 있으며, Semcor[M. Sundeanu et al., Using Predicate-Argument Structures for Information Extraction", In Proceedings of the 41st Annual Conference of the Association for Computational Linguistics (ACL-03), pp. 8-15, 200] 등의 단어 의미 모호성해소(Word Sense Disambiguation) 결과 역시 오메가 온톨로지와 연결되어 있다. 오메가 온톨로지에 포함되어 있는 풍부한 개념은 자연언어를 해석할 수 있는 기초 자원이 될 수 있다.

[0037] 수많은 코퍼스중, 자동 온톨로지 구축에 가장 도움이 되는 것들은 Penn-Treebank와 Propbank이다. Penn-Treebank란 구문 분석된 문장의 코퍼스이다.

[0038] 트리뱅크(Treebank)라는 이름에서 알 수 있듯이, 구문구조는 트리(Tree)형태로 표현된다. 이들 구문 분석 결과는 자동 분석 이후, 두 명 이상의 어노테이터들이 수작업으로 검증하여 정확성 여부를 검증하였다.

[0039] 펜 트리뱅크(Penn-Treebank)는 현재 단순한 구문구조뿐만 아니라, 주어(Sub -ject) 등이나 시공간(Temporal, Location) 등에 대한 의미적 정보 또한 같이 담고 있다. 상기 펜 트리뱅크(Penn-Treebank)는 품사 부착, 구문 분석 등을 위한 학습 데이터로서 널리 이용되고 있다.

[0040] 프롭뱅크(Propbank)는 펜 트리뱅크(Penn-Treebank)의 월스트리트 저널문을 술어-논항 어노테이션해 놓은

것이다. 각 동사마다 골격(Frame)을 만들어 놓고, 해당 동사의 골격과 일치하는 논항들을 찾아 표시하였다. 프롭뱅크(Propbank)의 논항에 해당하는 것은 온톨로지의 클래스나 인스턴스로 변환할 수 있는 하나의 기초 개념이 된다.

[0041] 문장의 유사도 측정 방법을 크게 분류하여 보면, 두 가지 갈래로 나누어 볼 수 있다. 하나는 문장의 의미적 유사도에 초점을 둔 경우이고, 다른 하나는 문장의 구조적 유사도에 초점을 둔 것이다.

[0042] "It is a dog", "That must be a dog.", "It is a log." 라는 세 문장이 있을 때, 문장의 의미적 유사도에 중점을 두는 방법은 "dog" 라는 단어가 함께 출현하는 "It is a dog", "That must be a dog." 의 유사도를 높게 판단한다.

[0043] 반면에, 문장의 구조를 중히 여기는 입장은 "It is a dog"와 "It is a log." 가 더욱 비슷하다고 판단하게 된다. 문장의 유사도는 크게는 이러한 두 가지 측면으로 나누어 볼 수 있으며, 각각의 경우 역시 세부적으로는 목적에 맞게 다른 양상을 보인다.

[0044] 문장의 구조적 유사도 측정은 트리 유사도를 이용한다.

[0045] 도 4는 트리 유사도 측정 방법 중 편집거리(Edit Distance)를 이용한 방법에서 사용되는 세가지 연산을 설명하고 있다. 해당 방법에서는 트리를 변경하는 방법으로 "insert", "delete", "change"의 세가지 연산을 제시하고 있으며, 각 연산을 사용할 때마다 페널티를 줌으로써 트리의 유사도를 구하게 된다.

[0046] 문장의 구문구조 분석 트리를 해당 편집거리(Edit Distance)를 이용해서 비교할 경우에는 문장의 구조적 유사도를 측정할 수 있다.

[0047] 문장의 의미적 유사도 측정은 단어의 유사도를 문장 유사도로 확장시키는 방법을 취하고 있다. 단어간의 유사도로는 단어간의 의미적 계층구조를 이용한다.

[0048] 도 5는 워드넷(WordNet)의 일부로서, 이러한 계층 구조하에서 단어간의 거리 등을 측정함으로써, 단어간의 유사도를 쉽사리 측정할 수 있다. 이를 문장의 유사도로 확장시키기 위해, 문장을 의미 벡터(Semantic Vector) 형태로 표현하게 된다.

[0049] 이때, 유사한 단어를 같은 축에 위치시킴으로써 문장의 의미적 유사도를 표현할 수 있는 의미 벡터를 구성하게 된다. 또한, 의미 벡터에서 고려하지 못하는 단어간의 순서 등도 최종적으로 고려하여 문장의 의미적 유사도를 측정하게 된다.

[0050] 자동 술어-논항 어노테이션은 논항 단위로 학습 데이터를 모아 특정 학습 모델을 통해 학습하는 경우가 대표적이다. 이 경우에는, 술어에서 논항까지의 트리의 경로, 동사의 어조, 구나 절의 태그 등이 자질로써 사용된다.

[0051] 특정 자질의 사용 여부에 따라 정확률과 재현률이 반비례하는 현상이 있었으며, 이 연구는 다른 술어-논항 어노테이션의 기초가 되었다. 그 이후의 자동 술어-논항 어노테이션은 보통 논항 경계 분석과 역할 표시(Role Labeling)로 나뉘고, 각 과정을 기계학습을 통하여 해결하였다. 이중 트리커널(TreeKernel)은 트리의 경로를 학습하는 대표적인 예이다.

[0052] 온톨로지는 개념, 개념간의 관계를 기술한 포맷이다. 문장을 온톨로지 형태로 나타내기 위해서는 결국 개념의 경계를 확실하게 포착할 수 있어야 하며, 또한 포착된 개념들 간에 관계를 맺어 줄 수 있어야 한다. 하지만, 종래의 복잡한 문장의 경우에는 위의 작업들을 하는 것이 쉽지 않다는 문제점이 있었다.

[0053] 또한, 본 발명은 문장을 쪼개서 자질로 사용한다는 측면에서 기존의 머신러닝(machine-learning)과도 비슷한 성격을 띤다. 다만, 머신러닝의 경우에는 예제를 역추적하기가 어렵다는 문제점이 있었다.

발명의 내용

해결 하고자하는 과제

[0054] 상기 문제점을 해소하기 위해 안출된 본 발명은 코퍼스(예를 들어, 펜-트리뱅크)를 이용하여, 자연언어문장을 술어-논항 구조 어노테이션하는 문장 온톨로지 생성을 위한 자동 어노테이션 시스템 및 방법을 제공함을 그 목적으로 한다.

과제 해결수단

[0055] 상기 목적을 달성하기 위한 본 발명에 따른 문장 온톨로지 생성을 위한 자동 어노테이션 시스템은 코퍼스(예를 들어, 펜-트리뱅크)의 어노테이션된 문장들을 적당한 단위로 분해해서 저장하는 분해 저장부, 입력 문장이 들어왔을 때 상기 분해 저장부로부터 분해된 문장 조각중에 의미, 구조적으로 유사한 조각을 찾아 해당 조각에 어노테이션된 내용을 참조하여 입력문장을 어노테이션하는 어노테이션부를 포함한다.

[0056] 또한, 본 발명에 따른 문장의 부분 유사도 기반 어노테이션 방법은 문장을 비교 가능한 단위로 분해하는 제1 과정, 상기 문장 조각의 구조적 유사도를 측정하는 제2 과정, 논항들이 의미상으로 얼마나 유사한지 알아보기 위해 상기 문장 조각의 의미적 유사도를 측정하는 제3 과정, 및 상기 문장 유사도를 기반으로 하여 어노테이션을 수행하는 제4 과정을 포함한다.

효 과

[0057] 본 발명에 따른 시스템은 예제를 추적하고 피드백을 알리기 쉽다는 장점이 있다.

[0058] 현재의 문장 비교는 시간이 비교적 많이 걸리는 작업인데, 본 발명에 따라 클러스터링을 통해 술어-논항 구조의 핵심적인 부분만 뽑아내어 비교한다면, 좀 더 정확하고, 빠른 술어-논항 구조 어노테이션이 가능하다는 효과가 있다.

[0059] 본 발명에서 제안한 시스템은 다른 추가적인 문제들이 풀릴 때 더욱 빛을 발한다. 우선 온톨로지 매핑 과제가 가장 중요한 부분이라고 할 수 있다. 기본적으로 온톨로지 매핑이라고 하면 수정거리(Edit Distance)를 비롯한 많은 어휘적 비교를 통해 이루어지는데, 본 발명에서 제안하는 시스템에 의하면 이러한 비교는 중요하지 않다.

발명의 실시를 위한 구체적인 내용

[0060] 이하, 본 발명의 바람직한 실시예를 첨부된 도면들을 참조하여 상세히 설명한다.

[0061] 도 6은 본 발명에 따른 유사한 문장을 이용한 어노테이션 예제를 도시한 도면이다.

[0062] 본 발명은 프롭뱅크(Propbank)를 이용하여, 자연언어문장을 술어-논항 구조 어노테이션하는 것이다. 기본적인 접근 방식은 프롭뱅크(Propbank)내에서 입력 문장과 유사한 문장을 찾아, 해당 유사 문장의 술어-논항 구조를 그대로 가져와 입력 문장을 어노테이션하는 것이다.

[0063] 상기 도 6의 예제에서는 'The prices hit bottom.'이라는 문장을 술어-논항 어노테이션하고자 한다. 이와 유사한 문장을 찾아 보았더니 오른쪽의 두 가지 문장 중에서도 특히 아래의 문장과 유사함을 알 수 있었다.

[0064] 따라서, 해당 유사 문장과 같이, "hit"의 주어 부분을 ARG0로 어노테이션하고, 목적어 부분을 ARG1으로 어노테이션하였다. 위와 같이 문장의 유사도에 기반한 어노테이션이 본 발명에서 제시하는 방법이다.

[0065] 도 7은 본 발명에 따른 문장 온톨로지 생성을 위한 자동 어노테이션 시스템의 구성을 도시한 도면이다.

[0066] 도 7에 도시된 바와 같이, 본 발명에 따른 문장 온톨로지 생성을 위한 자동 어노테이션 시스템은 펜-트리뱅크(Penn-Treebank)의 어노테이션된 문장들을 적당한 단위로 분해해서 저장하는 분해 저장부, 입력 문장이 들어왔을 때, 상기 분해 저장부로부터 분해된 문장 조각중에 의미, 구조적으로 유사한 조각을 찾아 해당 조각에 어노테이션된 내용을 참조하여 입력문장을 어노테이션하는 어노테이션부를 포함한다.

[0067] 또한, 상기 어노테이션부에서 동사를 중심으로 하는 술어-논항 구조의 부분만 유사도 비교가 되기 때문에, 한 문장에 여러 동사가 나오는 경우에는, 그 문장은 상기 분해 저장부에서의 첫번째 단계인 문장 분해 단계를 통해 각 동사를 중심으로 하는 조각들로 분해되는 것을 특징으로 한다.

[0068] 다음으론, 본 발명에 따른 문장의 부분 유사도 기반 어노테이션 방법에 관해 상세히 설명한다.

[0069] 본 발명에 따른 문장의 부분 유사도 기반 어노테이션 방법은 문장을 비교 가능한 단위로 분해하는 제1 과정, 상기 문장 조각의 구조적 유사도를 측정하는 제2 과정, 논항들이 의미적으로 얼마나 유사한지 알아보기 위해 상기 문장 조각의 의미적 유사도를 측정하는 제3 과정, 및 상기 문장 유사도를 기반으로 하여 어노테이션을 수행하는

제4 과정을 포함한다.

[0070] 더욱 상세히 설명하자면 다음과 같다.

[0071] 문장 분해과정

[0072] 도 8은 본 발명에 따른 문장 조각 생성을 도시한 도면이다.

[0073] 본 발명에서는 문장 유사도 기반 어노테이션을 위해, 프롭뱅크(Propbank)의 모든 문장들에 대해서 문장이 더 작은 단위로 분해되었다. 도 8에서 볼 수 있듯이, 분해된 조각에는 두가지 종류가 있는데, 하나를 다항관계 조각(Multinomial Fragm -ent)라고 하고, 다른 하나를 이항관계 조각(Binomial Fragment)이라고 한다.

[0074] 도 8에 도시된 예제는 "The market hits bottom" 이라는 문장의 술어-논항 관계를 문장 조각으로 나타내는 것이다. 각 조각은 술어로부터 논항에 이르는 경로의 트리를 문장의 조각으로 갖게 되는데, 다항관계 조각은 한 동사에 대한 술어-논항 구조를 모두 보존하고자 하는 형태이다.

[0075] 즉, ([The market]$_{ARG0}$, hit, [bottom]$_{ARG1}$)은 더 이상 분해할 수 없는 하나의 단위로 취급해야 한다는 시각이다. 이와는 다르게, 이항관계 조각은 ([The marke -t]$_{ARG0}$, hit, [bottom]$_{ARG1}$)을 더 잘게 분해하여 ([The market]$_{ARG0}$, hit), (hit, [bottom]$_{ARG1}$)의 더 작은 단위로 분해한다.

[0076] 이때, 분해의 기준은, 동사로부터 각 논항에 이르는 경로를 하나의 조각으로 취급하는 것이다.

[0077] 따라서, 하나의 문장은 여러 개의 동사를 포함하고 있으므로, 동사의 개수만큼의 다항관계 조각으로 표현된다. 또한, 각 다항관계 조각은 논항의 개수만큼의 이항관계 조각으로 표현된다. 이러한 문장 분해 과정을 알고리즘으로 나타내면 아래의 표 4와 같다.

[0078] **표 4 : 문장 분해 알고리즘**

[0079]

TreeDecomposition(*ParseAndPropTree T*)
1. Find the *predicate verb*(*VBD* in Figure 6)
2. Loop every argument, Arg_i
    a. Find the parent node of Arg_i and the ancestor of *VBZ*
    b. Decide the direction of the branch to Arg_i
    c. Record the trace from *VBZ* to Arg_i and the internal of Arg_i
    d. Produce one decomposed element from the record
3. Return the set of the decomposed elements

[0080] 또한, 분해된 문장 조각은 다음과 같이 표현할 수 있다.

[0081]

$$td_i^k = \{arg, n_1 \ldots n_b, dir\}$$

= (Tree decomposit ion branch)

arg = (arguement from PropBank)

n_i = (skeleton node from TreeBank)

dir = (direction from tracing)

$$TD^k = \{td_1^k, \ldots, td_n^k\}$$

= (k_{th} Tree Decomposit ion Set with n elements)

[0082] td는 각 이항관계 조각이 되며, TD는 다항관계 조각이다. TD는 이항관계 조각을 모아서 역으로도 생성 가능하며, td는 arg, n, dir 등으로 이루어진다. n은 동사로부터 논항에 이르기까의 경로를 나타내며, dir은 경로의 방향전환이 생기는 경우 그것을 기록하는 것이며, arg는 논항의 트리 그 자체를 나타낸다. 이렇게 분해된

문장 조각은 유사도 비교 과정에 사용되게 된다.

[0083] 문장 조각의 구조적 유사도 측정

[0084] 도 9는 본 발명에 따른 다항 관계조각의 구조적 유사도를 도시한 도면이다.

[0085] 문장 조각의 구조적 유사도 측정은 트리 유사도 측정 방식을 그대로 따른다. 이중에서 'change' 연산의 점수를 트리뱅크(Treebank)에 맞도록 수정하였다. 문장의 구조적 유사도 측정은 각 조각의 트리 경로에 대해서만 하게 되며, Arg와 같은 요소들에 대해서는 하지 않는다. 다항관계 조각의 경우에는 도 9에 나타낸 바와 같이 유사도를 측정한다.

[0086] 도 9에서 알 수 있듯이, 왼쪽의 트리를 오른쪽 트리와 일치하도록 변경하기 위해서는 한번의 'Change'와 한번의 'Delete' 그리고 한번의 'Insert'가 필요했다. 'VBG'를 'VBD'로 변경하는데는 50의 페널티가 부과되며, 'Insert', 'Delete' 의 경우에는 일괄적으로 200의 페널티가 부과되었다.

[0087] 따라서, 트리를 비교하는데, 총 7번의 연산이 필요한데, 그 중 450의 페널티가 부과되었으므로, (200*7 - 450)/(200*7)=0.68로, 즉, 0.68의 유사도를 갖게 된다.

[0088] 이항관계 조각의 경우도 하나의 트리라 가정하고 위와 같은 방식으로 비교를 할 수 있다. 하지만, 본질적으로 이항관계 조각은 하나의 문자열로 볼 수 있고, 일반적인 레벤시타인 거리(Levenshtein Distance)와 같게 된다.

[0089] 문장 조각의 의미적 유사도 측정

[0090] 도 10은 본 발명에 따른 논항의 의미적 유사도 측정을 도시한 도면이다.

[0091] 도 10에 도시된 것과 같이, 의미적 유사도 측정은 논항들이 의미적으로 얼마나 유사한지 알아보기 위한 것으로, 워드넷(WordNet)의 단어 유사도에 기반한다. 워드넷의 단어 유사도는 보통 경로의 길이를 이용하여 측정한다[T. Pedersen et al., "WordNet::Similarity - Measuring the Relatedness of Concepts" Proceedin -gs of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), 2004. 7]. 동일한 단어이거나, 같은 Synset에 포함된 단어는 1.0의 점수를 주며, 보통의 경우에는 네트워크를 탐색해가면서, 경로의 개수를 구한 다음, 경로의 개수의 역수에 반비례하여 점수를 매기게 된다.

[0092] 상기 식에서 "dog"과 "fox"는 워드넷 경로 유사도가 0.33점을 받게 되고, 두 논항의 의미적 유사도는 최종적으로 0.8325라는 값을 갖게 된다. 논항의 유사도 비교에는 이와 같이 단어의 워드넷 상에서의 경로 유사도를 이용하게 되며, 또한 품사의 순서 및 트리의 높이 등을 추가적으로 이용하여 논항의 유사도를 최종적으로 구하게 된다.

[0093] 문장 유사도 기반 어노테이션

[0094] 문장 유사도 기반 어노테이션은 앞에서 말한 다항관계 조각, 이항관계 조각과 구조적 유사도, 의미적 유사도를 조합하는 방식이다. 다만 어노테이션 대상 문장은 문장이 조각으로 나뉘어져 있지 않으므로, 문장 조각을 유추한 후 조각간의 유사도를 계산하게 된다. 어노테이션하는 기본 과정은 도 11에 나타낸 바와 같다.

[0095] "The prices severely depressed levels"라는 어노테이션할 문장에는 비교할 문장 조각이 존재하지 않는다. 따라서 왼쪽의 붉은색 부분의 다항관계 조각의 경계노드인 'NP', 'S', 'hit', 'NP'를 이용하여 오른쪽의 구문분석 트리에서 다항 관계 조각을 찾게 된다. 또한 찾은 조각의 각 경계점의 부분을 논항이라고 가정한 후 두 개의 조각간의 유사도를 계산하게 된다. 위와 같은 경우는 조각의 구조도가 완전히 일치하는 경우로 이런 경우에는 왼쪽의 논항 정보를 이용하여 입력 문장을 어노테이션 할 수 있다.

[0096] 문장 유사도 기반 어노테이션 방법에는 총 6가지 종류가 있다. 문장의 구조적 유사도만 이용하는 경우, 구조적 유사도와 의미적 유사도를 이용하는 경우의 두 가지 조합이 그 첫째이다. 또한, 문장 조각을 이용함에 있어서는 다항관계 조각만 이용하는 경우, 이항관계 조각만 이용하는 경우, 다항관계 조각과 이항관계 조각을 동시에 이용하는 경우의 세 가지가 있을 수 있다. 이러한 두 가지 요소의 조합으로 총 여섯 가지의 문장 유사도를 생각해 볼 수 있다.

[0097] 도 12는 다항관계 조각과 문장의 구조적 유사도를 이용하여 어노테이션하는 과정을 보여준다. 우선, 어노테이션 할 문장의 조각의 경계를 결정하고 구조적 유사도를 측정한 결과 유사함을 알 수 있어, 적합한 술어-논항 구조를 어노테이션 할 수 있다.

[0098] 하지만, 이러한 문장의 구조적 유사도와 다항관계 조각에만 의존한 도 12상의 어노테이션 방법에는 한계점이 있다.

[0099] 도 13의 예제는 그 첫번째 문제점을 말해준다. "The prices hit bottom"은 구조적으로는 "I hit the ball"과 유사하다. 그러나, 앞의 두 예제의 "hit"은 서로 다른 의미로 사용되었으며, 오히려 "Sterling plunged and hit its low"와 같은 뜻으로 사용되었다.

[0100] 이런 경우에는, "Sterling"과 "Prices" 그리고 "low"와 "bottom"이 비슷한 경우에 사용되는 용어임을 알 수 있다면 위의 문장과 더 유사함을 알 수 있다. 즉, 문장의 의미적 유사도가 꼭 필요한 경우이다.

[0101] 도 14는 다항관계 조각의 한계를 말하고 있다.

[0102] "abort" 동사에 대한 예제 문장이 다음과 같이 두 문장밖에 없을 때, "The company aborted its bid" 라는 문장이 입력으로 들어오면, 두개의 예제 문장을 모두 사용해야 술어-논항 구조 어노테이션을 할 수 있다. 즉, 예제의 논항 부족 현상을 다른 예제의 일부분으로서 보완해 주어야 한다. 다항관계 조각뿐만 아니라, 이항관계 조각까지 이용하면 이러한 부족한 부분을 채워줄 수 있다.

[0103] 위에서 설명한 구조적 유사도의 한계와 다항관계 조각의 한계로 인해, 문장의 의미적 유사도와 이항관계 조각이 필요해지며, 이를 이용할 때의 어노테이션은 다음과 같다.

[0104] 도 15는 본 발명에 따른 의미적 유사도를 이용한 술어-논항 구조 어노테이션 을 도시한 도면이다.

[0105] 도 15에서 볼 수 있듯이, 도 14의 한계점을 논항의 의미적 유사도를 이용하여 해결할 수 있었다. "The prices hit bottom."과 "I hit the ball"의 유사도는 1.0 + 0.01 = 1.01이 되고, "Sterling plunged and hit its low."의 경우에는 0.83 + 0.365=1.195가 되어 올바른 문장을 선택하게 됨을 알 수 있었다.

[0106] 도 16은 다항관계 조각과 이항관계 조각을 동시에 이용한 어노테이션으로, 다항관계 조각으로, "She completely aborted"를 선택하여 ARG0를 어노테이션하고, 비어있는 부분을 "Aborts command"의 이항관계 조각을 이용하여 어노테이션하는 과정이다. 이렇게 두 가지를 동시에 이용하는 것은 예제의 성김(sparseness) 문제를 해결할 수 있도록 해 준다.

[0107] 또한, 다항관계 조각을 이용하지 않고 이항관계 조각만 이용하는 경우, 문장의 부분적 일치도를 가장 높게 가져갈 수 있다. 다만, 술어-논항 구조의 전체를 잃어버리게 된다는 측면이 있다. 이항관계 조각만 이용하는 경우는 도 17과 같다.

[0108] 이항 조각을 이용한 어노테이션에서는 "The prices hit bottom" 어노테이션 하기 위해, "The prices hit severely depressed levels"의 "The prices" 부분과"The market didn't hit bottom yet"의 "hit bottom" 부분만 가져와서 조합하게 된다. 이러한 방식은 구조, 의미적으로 가장 유사한 문장 조각을 찾을 수 있도록 해 준다.

[0109] 본 발명의 성능적 측면에서, 정확률은 기존의 연구보다 높게 나왔고, 재현률은 낮게 나오는 현상을 발견하였다. 재현률이 낮게 나오는 문제는 기존 연구와의 접목을 통해 각종 머신러닝을 부분적으로 사용함으로써 해결 가능할 것이다.

[0110] 자체 평가에서는 문장의 유사도를 구조적 유사도, 의미적 유사도로 나누어 성능을 측정해 보았다.

[0111] 구조적 유사도만 이용하는 경우에는 약 60%의 확률로 술어-논항 구조를 어노테이션 할 수 있었다. 논항의 의미적 유사도를 같이 이용하는 경우에는 약 3%의 향상을 가져왔으며, 술어-논항 구조를 더 분해하여 이항관계 조각으로 이용할 경우에는 약 15% 정도의 성능 향상을 보여주었다.

[0112] 또한, 다항관계 조각과 이항관계 조각을 동시에 이용하는 것보다, 이항관계 조각만을 이용하는 경우가 4%가량 높은 성능을 보여주었는데, 이는 술어-논항 구조를 다항 관계로 유지하는 것보다 분해하는 것이 더 이점이 있음을 보여 주고 있다.

[0113] 프롭뱅크(Propbank)를 이용한 문장의 클러스터링 및 네트워크 구성은 좀 더 빠르고, 정확한 어노테이션을 위한 하나의 개선 방안이다.

[0114] 문장의 단어는 워드넷(WordNet)으로 매핑이 되기 때문에 대부분의 단어가 워드넷에 존재하는 한 매핑할 단어가 존재하지 않는 경우는 별로 없을 것이다. 그보다도 워드넷 역시 단어의 의미(Sense)로 이루어져 있기 때문에 문장 온톨로지의 각 단어의 의미(sense)를 파악하는 것이 중요한 문제가 된다.

[0115] 즉, WSD(Word Sense Disambiguation)이 온톨로지 매핑 문제를 치환하게 되는 것이다. 다만, WSD의 조건이 각 개념이 평평한 단어가 아닌 주제역할(Thematic Role)등의 구체적인 정보를 더 많이 가지고 있는 상황이다.

[0116] 이러한 특수한 상황에서 WSD를 하는 것도 하나의 주제가 될 수 있을 것이다. 또한, 현재 온톨로지 구축 과정은 담화 수준의 정보를 담고 있지는 못하다. 문장간의 관계(Sentimental Analysis)를 찾는 연구가 잘 진행될 경우 문장 온톨로지를 더욱 풍부하게 하고 더 나아가 문서 전체를 잘 파악할 수 있는 온톨로지로 발전할 수 있도록 만들어 줄 것이다.

[0117] 마지막으로 문장 온톨로지의 응용을 찾는 문제 또한 흥미로운 문제라고 할 수 있다. 각 신문기사를 문장 온톨로지 형태로 만든 다음에, 똑같은 주제의 기사에 대하여 공통되는 온톨로지는 무엇인지, 다른 부분은 무엇인지 분석함으로써 공통된 의견을 요약하여 제시할 수 있고, 신문간의 의견차를 부각시킬 수도 있을 것이다.

도면의 간단한 설명

[0118] 도 1은 온톨로지의 예를 도시한 도면,

[0119] 도 2는 문장 온톨로지의 예제를 도시한 도면,

[0120] 도 3은 TextToOnto의 계층구조 생성 예제를 도시한 도면,

[0121] 도 4는 트리 유사도 측정을 위한 세 가지 편집 거리(Edit Distance) 연산을 도시한 도면,

[0122] 도 5는 단어의 의미적 계층 구조(WordNet의 일부)를 도시한 도면,

[0123] 도 6은 본 발명에 따른 유사한 문장을 이용한 어노테이션 예제를 도시한 도면,

[0124] 도 7은 본 발명에 따른 문장 온톨로지 생성을 위한 자동 어노테이션을 도시한 도면,

[0125] 도 8은 본 발명에 따른 문장 조각 생성을 도시한 도면,

[0126] 도 9는 본 발명에 따른 다항 관계조각의 구조적 유사도를 도시한 도면,

[0127] 도 10은 본 발명에 따른 논항의 의미적 유사도 측정을 도시한 도면,

[0128] 도 11은 본 발명에 따른 베이스라인(Baseline) - 문장 유사도 기반 어노테이션을 도시한 도면,

[0129] 도 12는 본 발명에 따른 다항 관계조각과 구조적 유사도를 이용한 어노테이션을 도시한 도면,

[0130] 도 13은 본 발명에 따른 구조적 유사도의 한계를 도시한 도면,

[0131] 도 14는 본 발명에 따른 다항 관계조각의 한계를 도시한 도면,

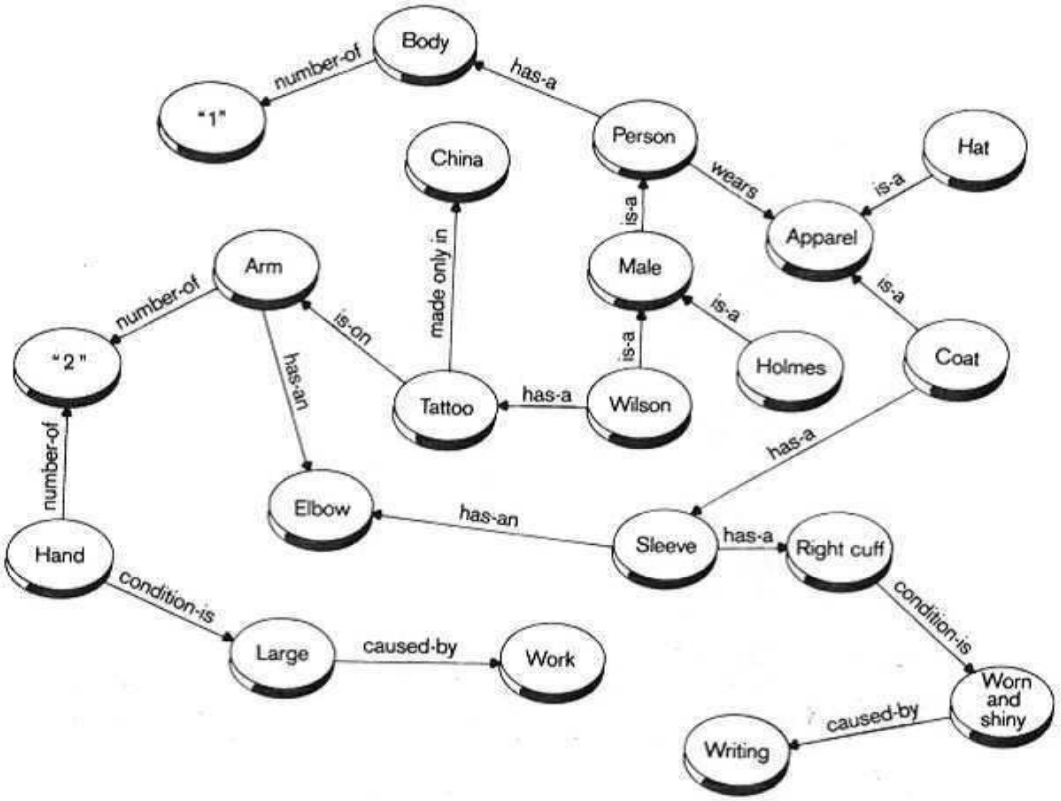
[0132] 도 15는 본 발명에 따른 의미적 유사도를 이용한 술어-논항 구조 어노테이션 을 도시한 도면,

[0133] 도 16은 본 발명에 따른 다항 관계조각과 이항 관계조각을 동시에 이용한 어노테이션을 도시한 도면, 및

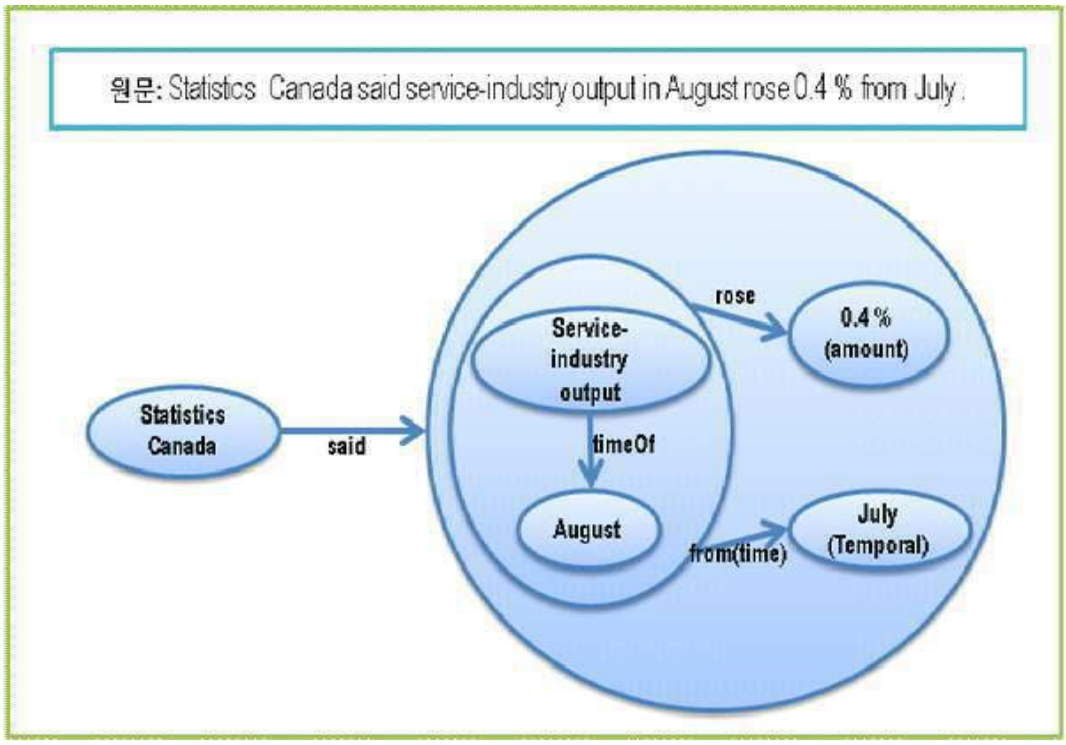
[0134] 도 17은 본 발명에 따른 이항 관계조각을 이용한 어노테이션을 도시한 도면이다.

도면

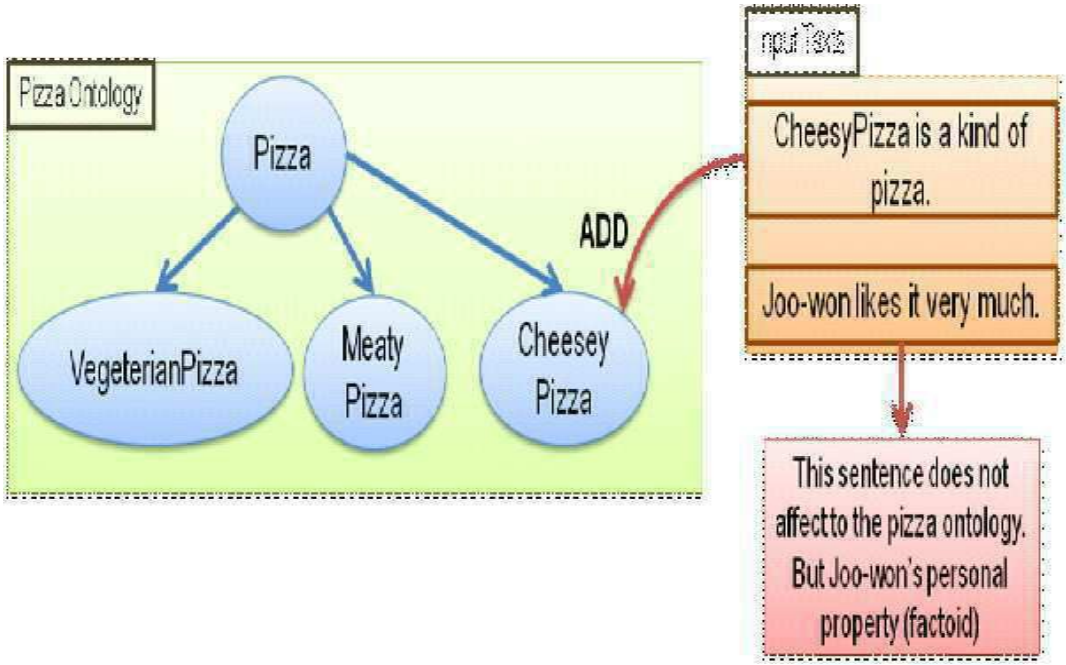
도면1


Body
number-of
"1"
has-a
Person
China
wears
Hat
is-a
Apparel
is-a
made only in
is-a
Male
Arm
number-of
"2"
is-on
has-an
Tattoo
has-a
Wilson
is-a
Holmes
is-a
Coat
has-a
number-of
Elbow
has-an
Sleeve
has-a
Right cuff
Hand
condition-is
Large
caused-by
Work
condition-is
Worn and shiny
Writing
caused-by

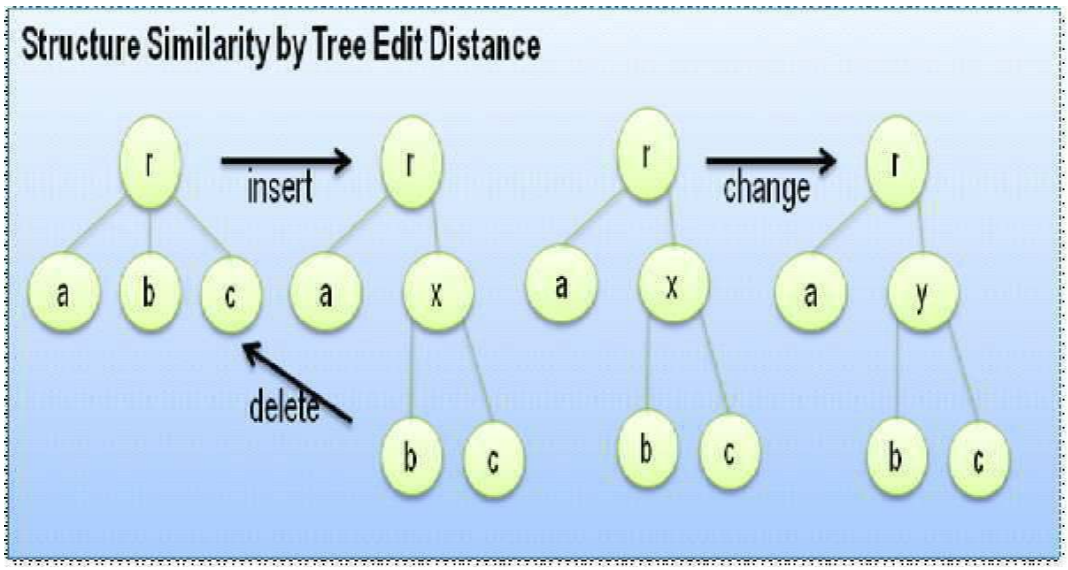
도면2


원문: Statistics Canada said service-industry output in August rose 0.4 % from July .
Statistics Canada
said
Service-industry output
rose
0.4 % (amount)
timeOf
August
from(time)
July (Temporal)

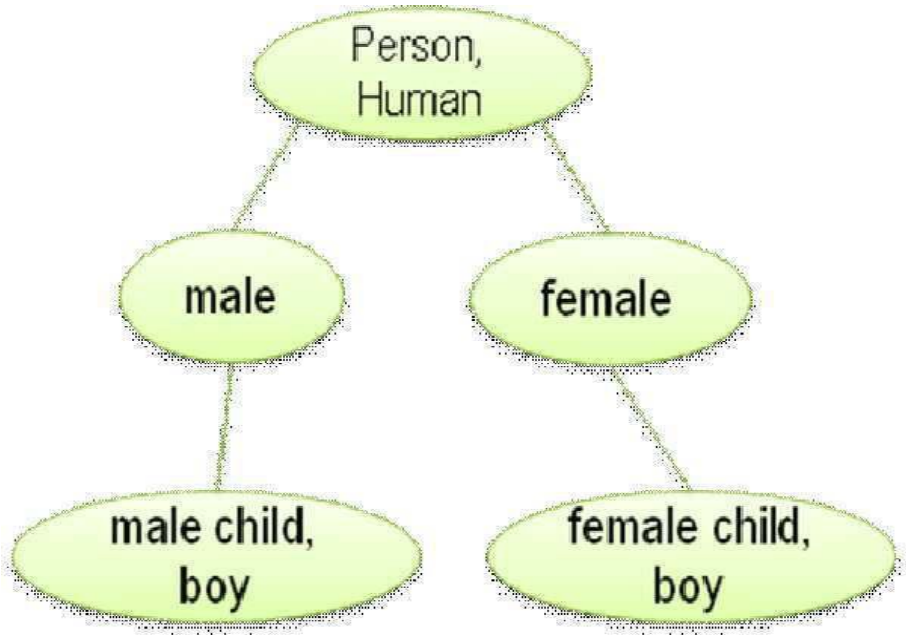
도면3


Pizza Ontology
Pizza
VegeterianPizza
Meaty Pizza
Cheesey Pizza
ADD
CheesyPizza is a kind of pizza.
Joo-won likes it very much.
This sentence does not affect to the pizza ontology. But Joo-won's personal property (factoid)

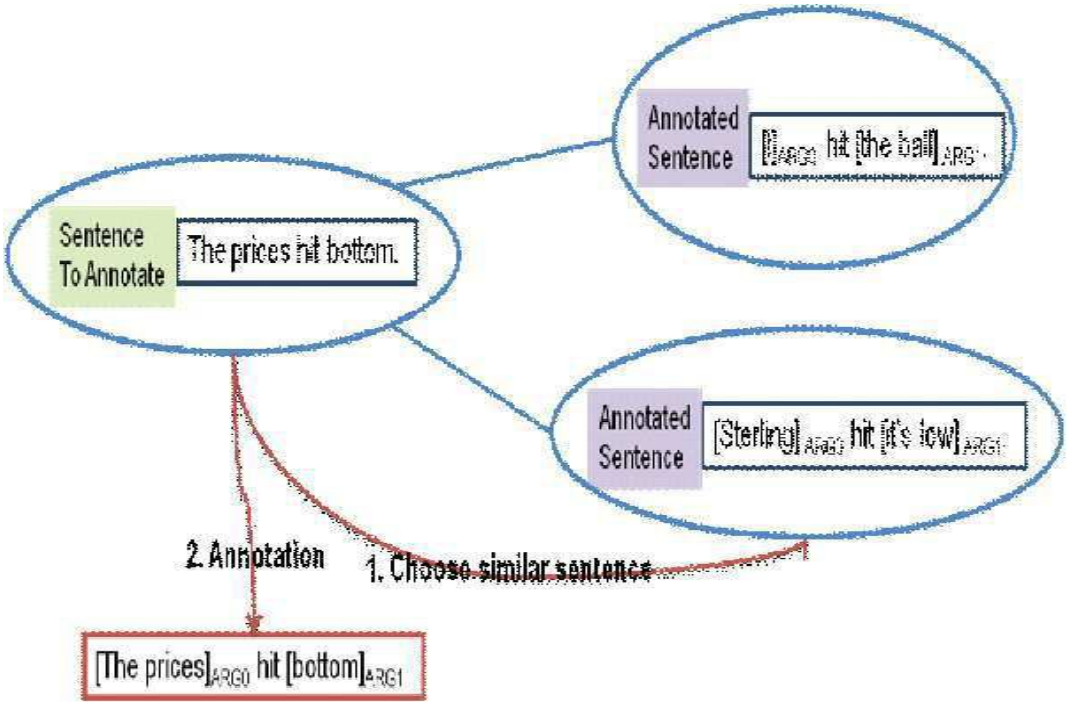
도면4


Structure Similarity by Tree Edit Distance
r
a
b
c
insert
x
delete
change
y

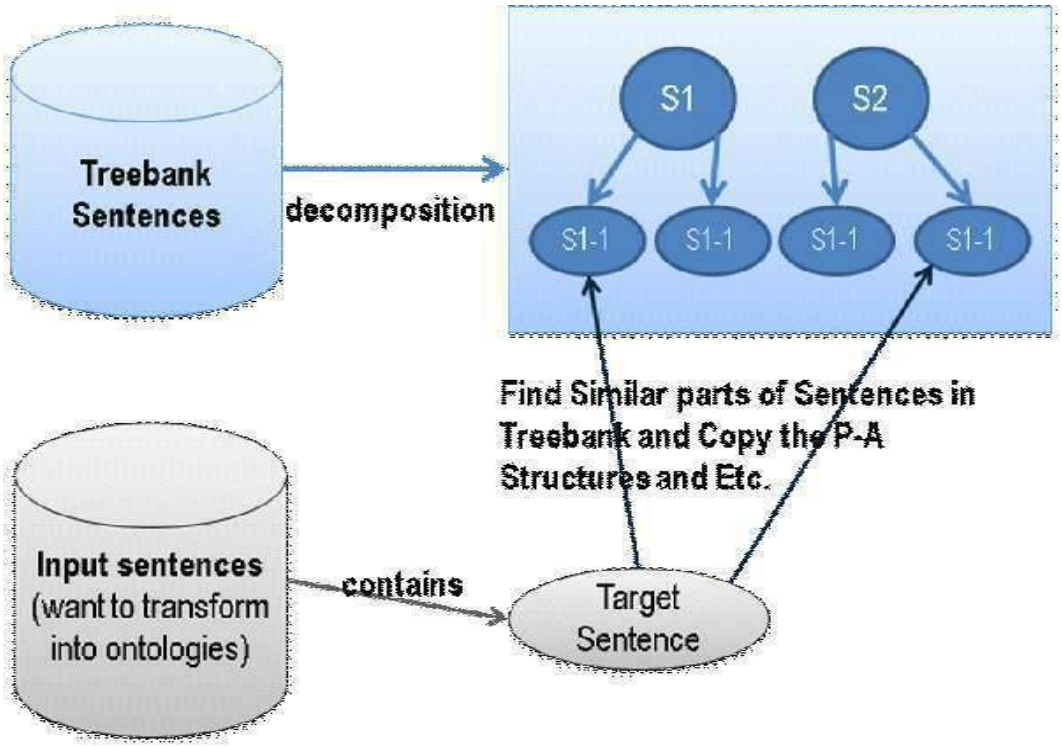
도면5


Person,
Human
male
female
male child,
boy
female child,
boy

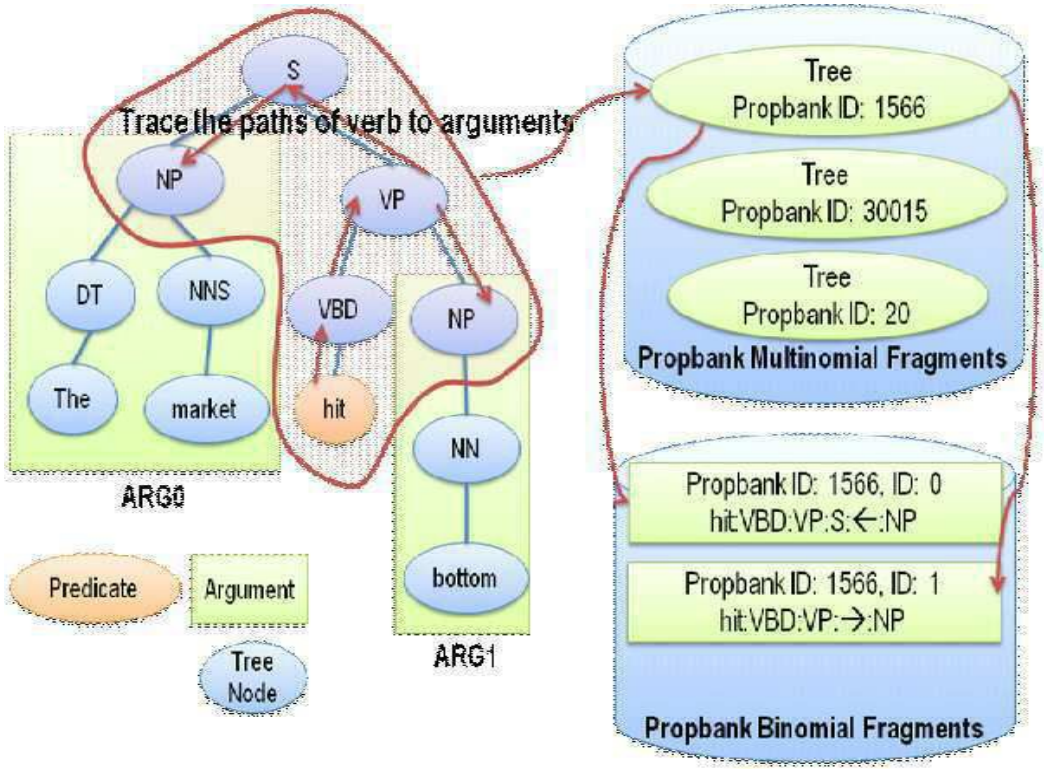
도면6


Annotated
Sentence
Sentence
To Annotate
The prices hit bottom.
Annotated
Sentence
2. Annotation
1. Choose similar sentence
[The prices]ARG0 hit [bottom]ARG1

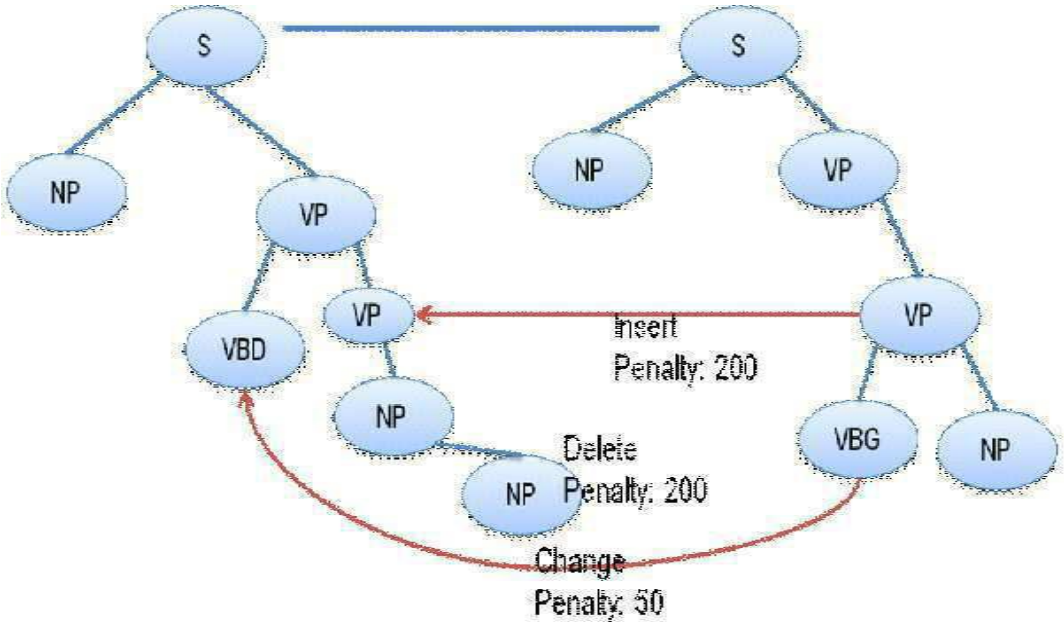
도면7


Treebank Sentences
decomposition
S1
S2
S1-1
S1-1
S1-1
S1-1
Find Similar parts of Sentences in Treebank and Copy the P-A Structures and Etc.
Input sentences (want to transform into ontologies)
contains
Target Sentence

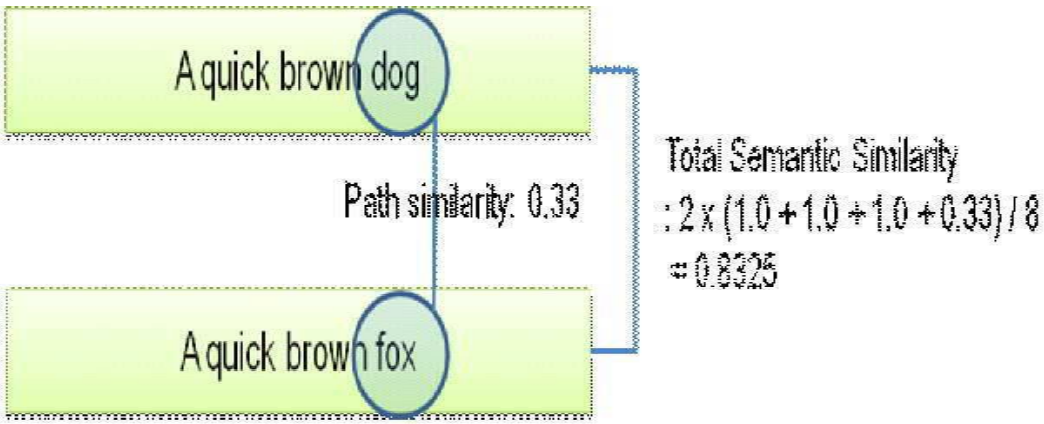
도면8


S
Trace the paths of verb to arguments
NP
VP
DT
NNS
VBD
NP
The
market
hit
NN
bottom
ARG0
ARG1
Predicate
Argument
Tree Node
Tree
Propbank ID: 1566
Tree
Propbank ID: 30015
Tree
Propbank ID: 20
Propbank Multinomial Fragments
Propbank ID: 1566, ID: 0
hit:VBD:VP:S:←:NP
Propbank ID: 1566, ID: 1
hit:VBD:VP:→:NP
Propbank Binomial Fragments

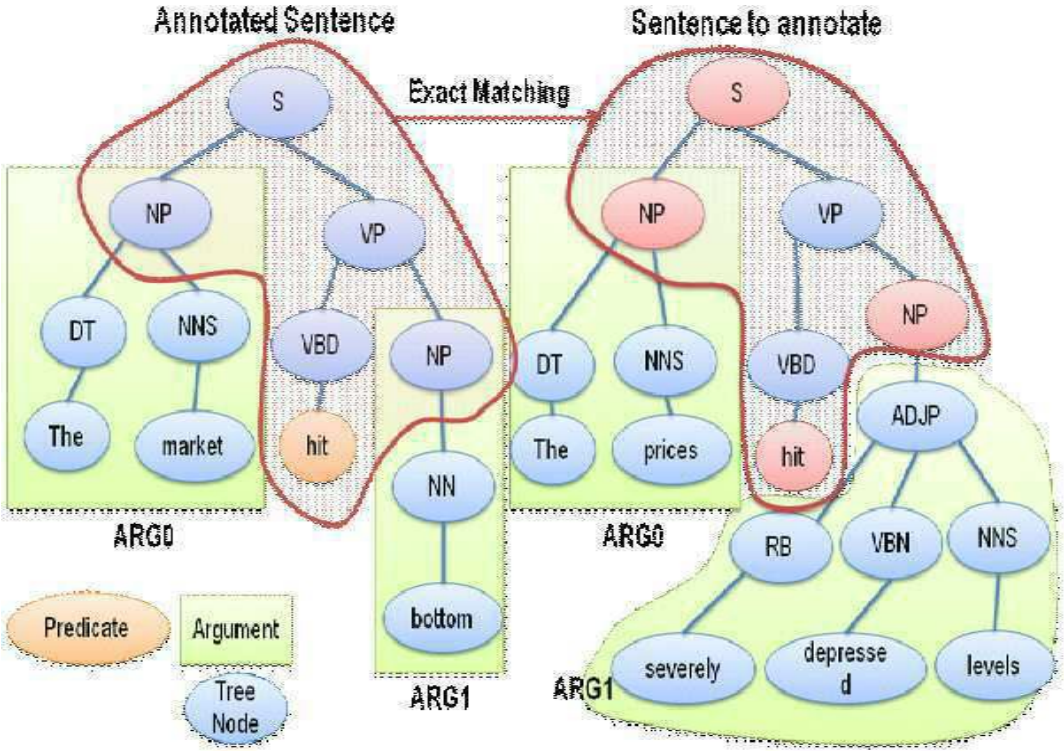
도면9


S
S
NP
NP
VP
VP
VP
VP
VBD
Insert
Penalty: 200
NP
VBG
NP
Delete
NP
Penalty: 200
Change
Penalty: 50

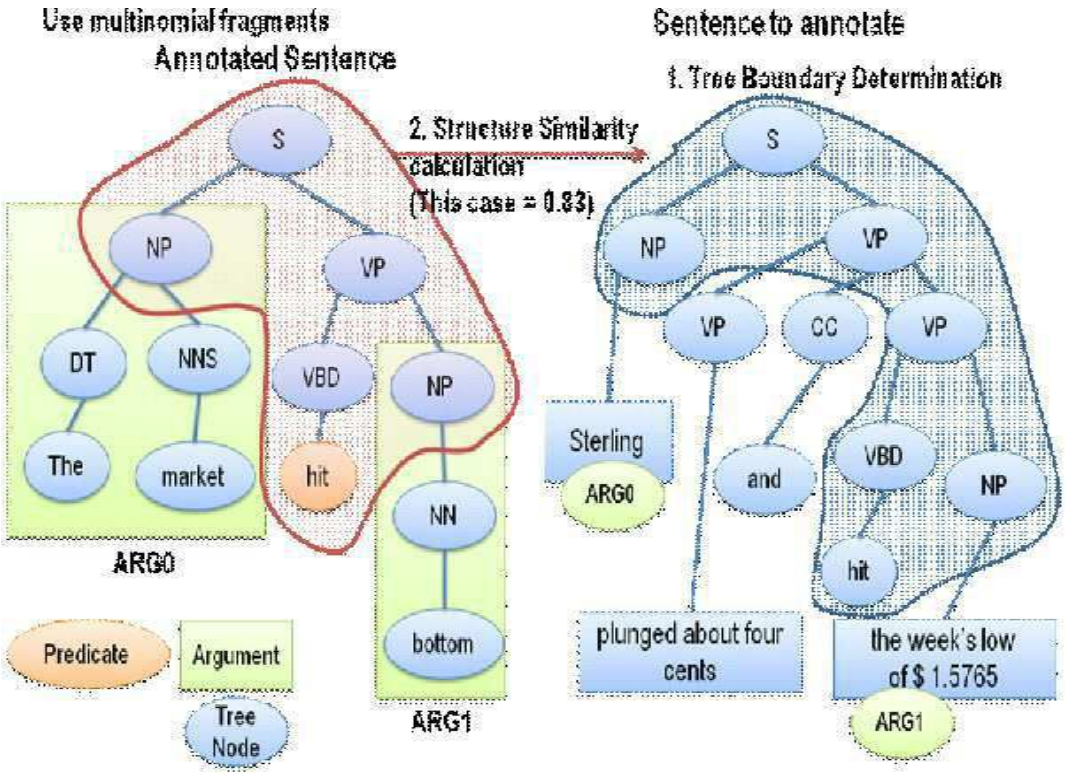
도면10


A quick brown dog
Path similarity: 0.33
A quick brown fox
Total Semantic Similarity
: 2 x (1.0 + 1.0 + 1.0 + 0.33) / 8
= 0.8325

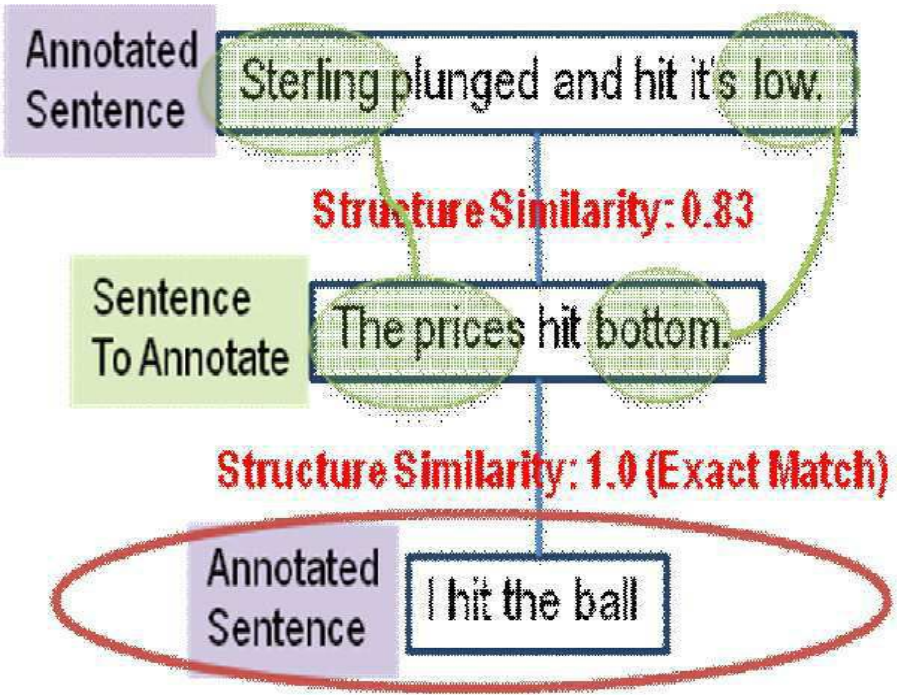
도면11


Annotated Sentence
Sentence to annotate
Exact Matching
S
NP
VP
DT
NNS
VBD
NP
The
market
hit
NN
bottom
ARG0
ARG1
Predicate
Argument
Tree Node
ADJP
RB
VBN
NNS
severely
depresse d
levels

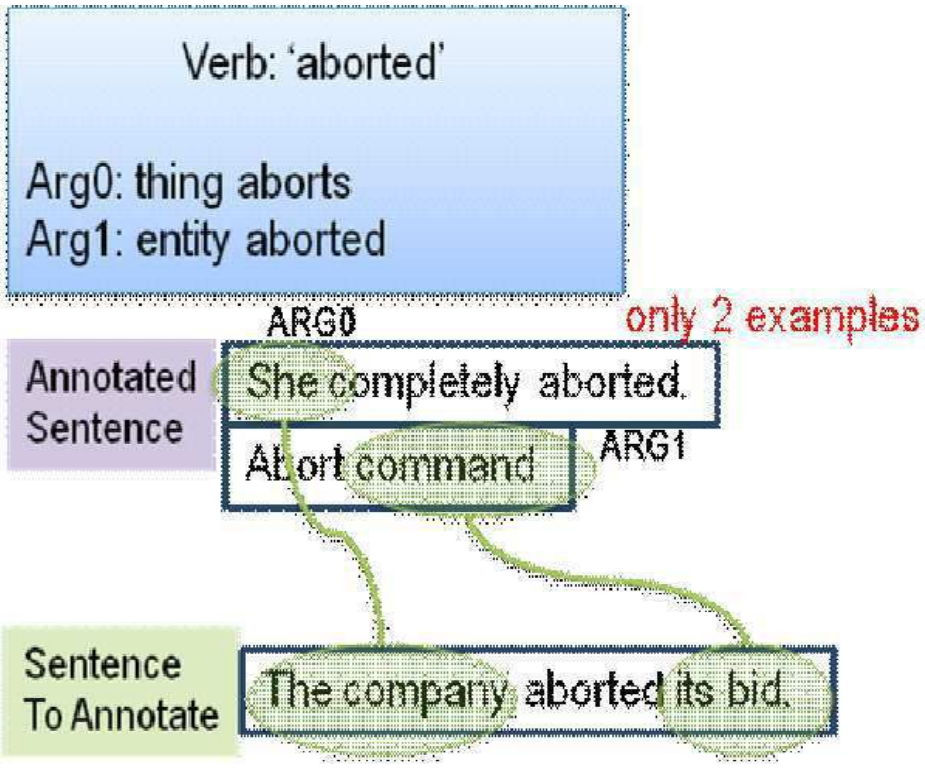
도면12


Use multinomial fragments
Annotated Sentence
Sentence to annotate
1. Tree Boundary Determination
2. Structure Similarity calculation (This case = 0.83)
S
NP
VP
DT
NNS
VBD
NP
The
market
hit
NN
bottom
ARG0
ARG1
Predicate
Argument
Tree Node
CC
and
Sterling
plunged about four cents
hit
the week's low of $ 1.5765

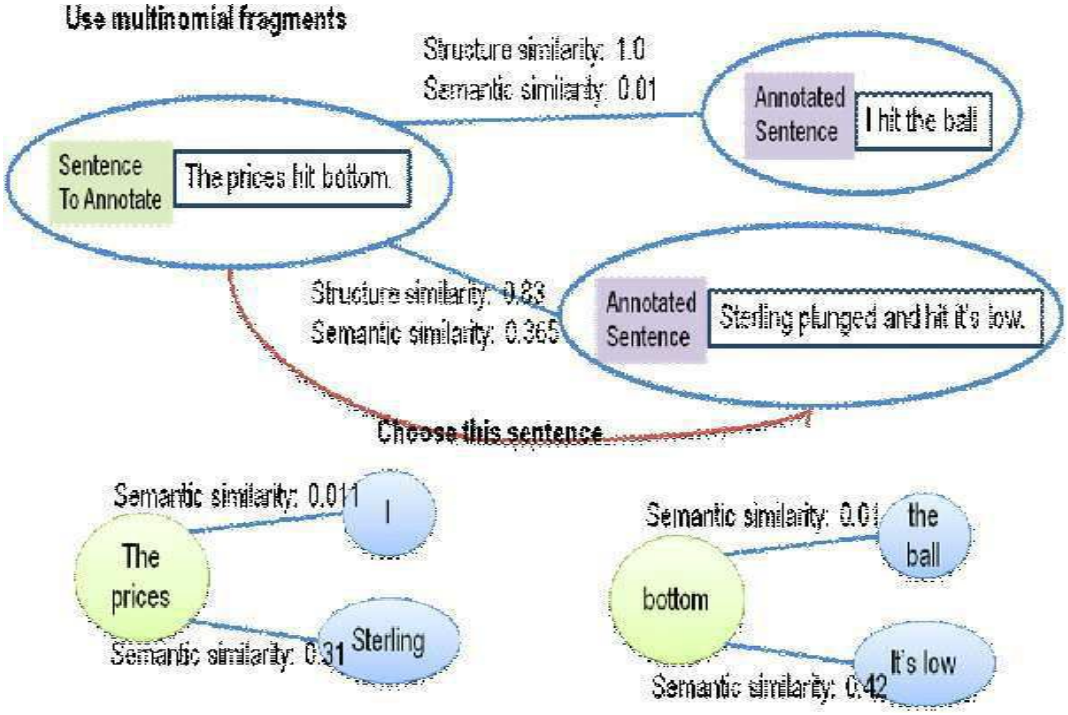
도면13


Annotated Sentence
Sterling plunged and hit it's low.
Structure Similarity: 0.83
Sentence To Annotate
The prices hit bottom.
Structure Similarity: 1.0 (Exact Match)
Annotated Sentence
I hit the ball

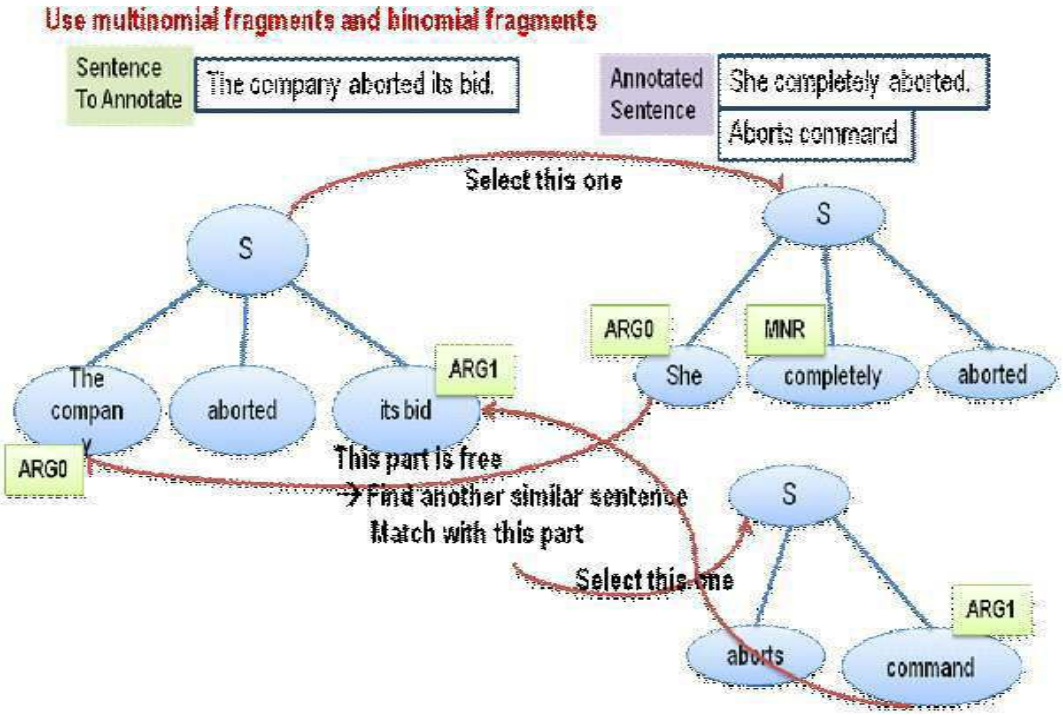
도면14


Verb: 'aborted'
Arg0: thing aborts
Arg1: entity aborted
ARG0
only 2 examples
Annotated Sentence
She completely aborted
ARG1
Abort command
Sentence To Annotate
The company aborted its bid.

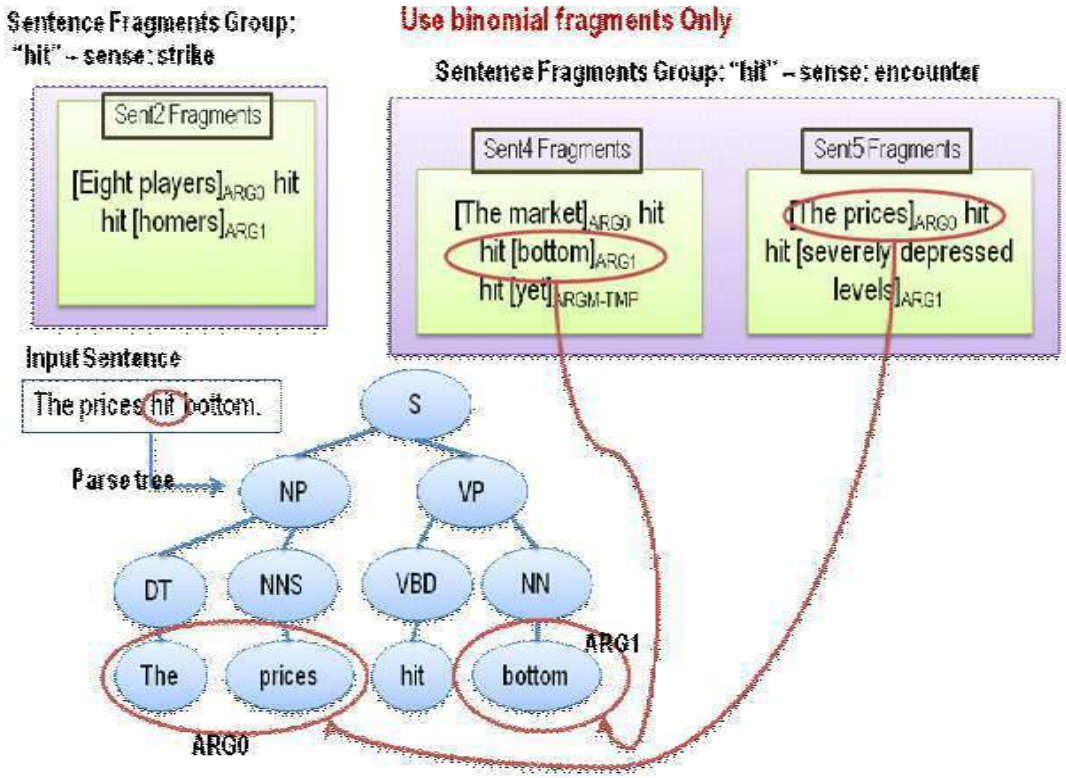
도면15


Use multinomial fragments
Structure similarity: 1.0
Semantic similarity: 0.01
Annotated Sentence
I hit the ball
Sentence To Annotate
The prices hit bottom.
Structure similarity: 0.83
Semantic similarity: 0.365
Annotated Sentence
Sterling plunged and hit it's low.
Choose this sentence
Semantic similarity: 0.011
I
The prices
Semantic similarity: 0.31
Sterling
Semantic similarity: 0.01
the ball
bottom
Semantic similarity: 0.42
It's low

도면16


Use multinomial fragments and binomial fragments
Sentence To Annotate
The company aborted its bid.
Annotated Sentence
She completely aborted.
Aborts command
Select this one
S
S
ARG0
MNR
ARG1
The compan
aborted
its bid
She
completely
aborted
ARG0
This part is free
Find another similar sentence
Match with this part
S
Select this one
ARG1
aborts
command

도면17


Sentence Fragments Group: "hit" – sense: strike
Sent2 Fragments
[Eight players]ARG0 hit
hit [homers]ARG1
Use binomial fragments Only
Sentence Fragments Group: "hit" – sense: encounter
Sent4 Fragments
[The market]ARG0 hit
hit [bottom]ARG1
hit [yet]ARGM-TMP
Sent5 Fragments
[The prices]ARG0 hit
hit [severely depressed
levels]ARG1
Input Sentence
The prices hit bottom.
Parse tree
S
NP
VP
DT
NNS
VBD
NN
The
prices
hit
bottom
ARG1
ARG0