인지과학, 제24권 제4호
Korean Journal of Cognitive Science
2013, Vol. 24, No. 4, 301～321.

한국어 텍스트 문장정렬을 위한 개체격자 접근법과 LSA 기반 접근법의 활용연구

김 영 삼† 김 홍 기 신 효 필‡
서울대학교 서울대학교 서울대학교
인지과학 협동과정 치과대학 언어학과

본 논문은 텍스트의 응집도 측정과 텍스트 자동생성 시스템을 위한 기초기술 중 하나인 문장정렬 과제에 대한 연구로, 개체기반적(entity-based) 접근의 한 유형인 개체격자 모형(Entity-Grid model)과 벡터공간 모형에 기반한 LSA(Latent Semantic Analysis)를 모두 시도하고 결과를 서로 비교하였다. 개체격자 모형에 대한 기존 연구들에서 논의된 명사들의 통사역(syntactic role) 정보가 한국어 텍스트 정렬과제에 미치는 영향을 실험하고자 하였으며, 기존 독일어권 응용연구 결과와는 달리 긍정적인 결과를 얻었다. 이 과정에서 한국어의 격조사를 활용하는 전략을 취했으며, 이는 한국어의 격표지 정보가 한국어 텍스트의 응집성을 측정하는 데에 유용할 수 있다는 점을 보인 것이다. 그리고 개체격자 모형을 통한 결과를 LSA 기반 모형결과와 비교하여 양 모형의 장단점과 향후 개선점을 아울러 논의하였다.

주제어 : 문장정렬, 응집성, Entity-Grid model, Latent Semantic Analysis, SVM

---

† 제1저자: 김영삼, 서울대학교 인지과학 협동과정, 서울시 관악구 관악로 1
연구분야: 인지과학, 전산언어학, E-mail: youngsamy@gmail.com
‡ 교신저자: 신효필, 서울대학교 언어학과, 서울시 관악구 관악로 1
연구분야: 전산언어학, E-mail: hpshin@snu.ac.kr

서 론

일반적인 글쓰기 과제에서 글의 응집성을 지키는 것은 글의 질을 평가함에 있어서 중요한 평가요소 중 하나이고, 때문에 자연어 처리에서 텍스트의 정합성 혹은 응집성(coherence)을 측정하는 것은 흥미로운 과제이다. 그리고 자연스럽게 자연어 처리 분야의 일부 연구자들은 텍스트의 응집성을 평가하는 이론을 통해 높은 수준의 자연어 텍스트 생성시스템을 구축하려 했다[1, 2]. 하지만 텍스트의 응집성을 평가하거나 생성하는 이와 같은 시스템들은 복잡한 규칙들로 이루어지고, 한정된 주제에 대해서만 적용가능하며 다양한 크기의 데이터에 대한 확장가능성(scalability)을 보장하지 못한다는 점이 주요 문제로 지적되었다[3].

본 연구는 위와 같은 문제를 피할 수 있도록, 텍스트의 국소적 응집성 측정을 위한 자동화된 문장정렬(sentence ordering) 과정에 초점을 맞춘다. 문장정렬 과제연구는 여러 자연어 처리 시스템에 다양하게 도움을 줄 수 있는데, 우선 첫째, 자연어로 쓰인 문서의 국소적 응집성(local coherence)에 영향을 미치는 요인은 문장 간의 자연스러운 연결이므로 문장 사이의 의미적 연관성을 구하는 것은 텍스트 문서의 국소적 응집도 측정과 직접적으로 연관된다. 그리고 텍스트의 국소적 응집성은 텍스트의 전체적 응집성에 필수적인 요소이므로 텍스트 문서의 응집성 계산에 있어 중요한 정보를 제공한다. 둘째, 자동으로 텍스트를 생성하는 텍스트 생성 시스템(natural language generation system)에 본 연구의 방법은 생성된 문장의 의미적 연결성을 최대로 하여 문장순서를 결정하는데 도움을 줄 수 있기에, 결과적으로 응집성이 높은 문서생성을 할 수 있게 한다. 셋째, 국소적 응집성은 글 읽기 도중 추론을 하는 데 있어서 중요한 요소이기에[4] 논술문과 같은 문서의 가독성(readability)을 측정하는데 도움을 줄 수 있다. 그리고 마지막으로 문장 정렬 기술은 문서의 자동요약(automatic summarization) 과제에도 쓰일 수 있다. 자동요약 시스템은 특정 문서 내에서 가장 중요하거나 핵심적인 정보를 담고 있는 단락이나 문장을 추출해주는 역할을 수행하는데, 본 연구의 기법은 문장 간의 응집성을 평가해줄 수 있으므로 문장들 중 가장 높은 응집성을 갖는 문장들을 후보문장으로 추천할 수 있게 해준다.

본 연구의 방법론은 말뭉치(corpus) 혹은 학습자료를 이용한 자동화된 접근법에 주제를 맞추는데 이것은 미리 정의된 지식체계나 수작업을 요하는 주석처리 과정을 요구하지 않는다는 장점이 있다. 그리고 문서의 응집성을 전체적으로(globally) 측정하는 시도를 취하지 않고 대신 국소적인(locally) 접근법을 취하는데, 이 접근은 앞 문장에서 다음 문장으로의 이행이 얼마나 의미적으로 연관이 있는지를 알아냄으로써 문서의 전반적인 응집성을 계산하려는 시도이다. 텍스트의 국소적 응집성에 대한 주요 접근으로 문장 내의 명사들 간의 이행관계를 포착하여 접근하는 개체기반 접근(entity-based approach)과 벡터공간 모델링에 의존하는 LSA(latent semantic analysis)를 이용한 전통적인 접근이 있는데, 본 연구에서는 한국어 자료에 대해서 이 두 접근법을 문장정렬 과제에 대해 적용하여 비교하고 장단점을 논의하는 자리를 마련하고자 한다.

본 연구에서 비교할 두 접근들 중 하나인 개체기반 접근법은 LSA 접근법과 마찬가지로 오랫동안 연구되어온 이론으로서 중심화 이론(centering theory)과 같이 대상 텍스트 문서를 개체기반적 표상(entity-based representation)으로 사상하여 이를 활용한다. 본 연구는 텍스트를 개체기반 표상으로 나타내는 모형들 중 가장 최근 모형에 속하는 '개체격자 모형(Entity-Grid model)'을 사용하였으며[5], LSA를 활용한 모형으로는 Foltz의 방법론을 이용하였다[6]. 본 연구의 실험은 인터넷 포탈업체가 제공하는 특정 섹션의 뉴스기사 데이터에 대해 행해졌으며, 그 수행은 서열화 과제(ranking task)를 통해 검증되었다.

관련연구

개체기반적(entity-based) 접근은 담화해석에 대한 언어학과 인지과학의 오랜 이론적 모형에 기초한다. 이 이론의 기본적 가정에 따르면 담화(discourse)의 응집성은 담화개체들(discourse entities)이 나타나고 언급되는 방식에 의해 결정된다[7, 8, 9]. 이 이론에 따르면 담화상의 어떤 시점의 특정 개체들은 다른 개체들보다 담화상의 기여가 더 두드러진다(salient)고 보는데, 중심화 이론(centering theory)은 개체기반적 접근이론을 계산모형화하여 응집성 측정 같은 연구과제에 활용하고자 하였다

[10, 11, 12]. 중심화 이론은 어떤 담화개체의 돋보임성(saliency)을, 그 개체가 발화(utterance, 텍스트의 경우에 문장)상에서 대용어(예, 대명사)를 통해 지시되었는지 아닌지를 통해 보고자 한다. 이 이론의 핵심적인 가정에 따르면, 담화상에서 중요한 대상, 즉 돋보이는 개체들(salient entities)은 대명사와 같은 문법적 장치를 통해 자주 지시되며, 문장내 특정 위치(주어 혹은 목적어의 위치)에 출현하는 경향성을 보인다. 텍스트의 국소적 응집성 측정문제에 대해 중심화 이론은 발화들에 대한 제약과 이전 발화요소(즉, 명사들)와 다음 발화요소 간의 전이규칙을 이용하여 모형화하고 이 모형은 국소적으로 응집적인(locally coherent) 텍스트에서 단어들의 분포는 규칙적인 패턴을 갖는다는 가정에 의존한다[10, 11, 12].

위와 같은 접근방식에 따라 국내에서 한국어 텍스트 문장정렬 과제를 연구한 경우는 많지 않았으나, 국내의 한 연구는 위의 기술된 중심화 이론에 기초하여 문장정렬 과정을 연구하였고, 수작업으로 명사와 대명사간의 지시 관계가 처리된 51개의 텍스트를 실험자료로 활용했다[13]. 그리고 그 결과로서 저자들은 중심화 기반 문장순서의 네 가지 평가척도 중에서 인접한 문장의 다음절에서 주제(topic)로 언급될 가능성이 가장 높은 개체의 수를 최대화하는 척도인 MAX.CPS를 한국어에 유리한 척도로서 제시하였다. 중심화 이론에 따르면 한 개의 발화는 Cf(forward looking center), Cb(backward-looking center), Cp(Preferred center)의 세 개의 중심구조를 가지며, MAX.CPS는 Preferred Center(즉, center of the preferred)를 최대화한다는 것을 의미한다.[1] 이는 부연하자면 현재 문장과 이전 문장의 가장 두드러진 명사가 같으면 그 문장 쌍은 국소적으로 응집되었다고 보는 것으로, 이러한 문장 쌍을 최대화하는 방식의 문장정렬을 선호하는 전략이다.

위 연구의[13] 경우와 같이 중심화 이론에 기초한 전통적인 개체기반 접근법은 미리 정교하게 주석된 언어자료를 필요로 하지만, Barzilay와 Lapata는 보다 자동화된 처리를 위해서 'Entity-Grid model'을 제안한다[5]. 이 모형은 텍스트에 등장하는 개체들(명사들)의 문법적 역할과 지시관계(coreference relation), 출현빈도 등의 정보를 이용하여 한 문서를 개체 자질간의 이행확률 벡터로 요약한다. 학습용 말뭉치로부터 얻은 이행확률 벡터들은 SVM_{rank}알고리즘[2]을 통해 분류기(classifier)를 학습시키게

1) 중심화 이론(Centering theory)에서 중심(centering)이란, 담화구조의 국소차원(local level)에 대한 주의적(attentional) 상태 요소를 가리킨다[11].

되는데, 이 분류기는 새로운 문서자료에 적용되어 해당 문서의 응집성을 평가할 수 있는 정보를 제공할 수 있다. 이 모형 역시 국소적 응집성을 통해 문서의 전체적 응집성에 접근하는 방법론을 따르며, 본 연구에서 개체기반 접근법의 전산모형으로 활용되었다. 기존 개체기반 모형과는 다르게 [5]의 모형은 수작업을 통해 처리되지 않은 입력자료를 자동화된 구문분석기(parser)를 이용하여 처리했음에도 두 개의 문서집합에 대해 각각 87%, 90%에 이르는 높은 정확도를 얻었다. 비교 실험의 하나로 선택한 Latent Semantic Analysis(LSA)를 통해서는 81%의 결과를 얻었다. 실험 코퍼스로 이용한 두 문서그룹은 각각 AP 통신의 지진 관련 기사와 국립수송안전국(National Transportation Safety Board)의 항공사고 데이터베이스로 이루어졌는데, 항공사고 문서는 문장구조가 단순하고 지시대명사 사용이 적어 상대적으로 높은 수행을 보였다.

'Entity-Grid' 모형은 영어가 아닌 독일어에도 Filippova와 Strube에 의해 응용되어 연구되었으나[14], 75%의 정확도를 보이는 선에서 그쳤으며 Barzilay와 Lapata의 연구와는 달리 명사들의 통사적 역할에 대한 정보(주어, 목적어, 그 외)가 실험결과의 향상에 긍정적인 영향을 미치지 못하고 오히려 분류기의 성능을 떨어트리는 것으로 나타났다(75%에서 72%로). Filippova와 Strube는 이에 대해 실험에 사용한 말뭉치의 크기 차이에서 온 효과일 것을 암시했으나[14], 이것이 실험적 상황과는 다른 언어적 특성의 차이에서 비롯된 효과일 가능성도 존재한다.

Latent Semantic Analysis(LSA)를 이용한 국소적 응집성 연구는 텍스트 간의 의미적 연관성 측정에서 인간의 직관과 매우 유사한 수행을 보여주는 LSA를 문장과 문장에 대해 적용한다[6]. 이 방법의 핵심적인 아이디어는 한 문서의 앞 문장과 후속 문장들 간의 의미적 연관성의 평균이 그 텍스트의 응집성을 반영한다고 보는 것이고, 때문에 국소적 응집성을 통해 문서의 전체 응집성에 접근하려는 방법의 연장선상에 있으므로 [5]와 함께 본 연구의 비교모형으로 채택되었다.3)

---

2) SVM(Support Vector Machine)은 분류알고리즘의 일종으로 자연어 처리를 비롯한 광범위한 분야에 널리 이용되고 있는데, 본 연구에서 활용한 SVM 알고리즘은 그것의 한 유형으로 SVM을 효율적으로 정렬과제에 이용하기 위한 것이다[22].

3) 기존연구에서는 은닉-마르코프 모형 기반의 Content 모형 역시 비교모형으로 이용되었으나, 이 모형은 텍스트 도메인에 의한 정보를 앞서 요구하며 텍스트 도메인 효과는 본 연

한국어 텍스트의 개체격자 표상(Entity-Grid representation)

개체격자 모형에서 각 텍스트는 개체격자로 사상된다. 이 격자는 2차원의 배열 형태를 취하는데 한 차원은 각 문장들을 가리키고 다른 한 차원은 어휘들을 가리킨다. 그리고 이 테이블의 각 셀에는 해당 어휘가 해당 문장에서 갖는 통사적 역할이 표시되는데, 이 역할은 주어(subject), 목적어(object) 혹은 그 외(others) 중 하나이다. 그림 1은 텍스트를 개체격자 형태로 표상하기 위한 예문을 보여준다. 그림 1의 텍스트는 문장으로 분해되어 분석의 기본단위를 이루는데, 이는 중심화 이론의 발화(utterance) 단위에 해당한다. 개체격자 표상법은 단어의 문법적 역할(syntactic role)에 대한 정보를 필요로 하므로, 각 문장은 제시된 것처럼 형태소 분석기를 통해 미리 처리되었다.

1	사람/NNG+의/JKG 소변/NNG+이/JKS 물고기/NNG+를 죽이/VV+다/EF+./SF
2	소변/NNG+이/JKS 물고기/NNG+에/JKB 직접/MAG 미치/VV+지는/EC 않/VX+지만/EC 일련/NNG+의/JKG 환경/NNG 변화/NNG+를/JKO 일으키/VV+어/EC 결국/MAG 물고기/NNG+를/JKO 질식/NNG+시키/XSV+ㄴ다는/ETM 것/NNB+이/VCP+다/EF+./SF
3	이/NP+에/JKB 대하/VV+아/EC 전문가/NNG+들/XSN+은/JX "/SS+충분히/MAG 그렇/VA+ㄹ/ETM 수/NNB 있/VA+다/EF+"/SS+는/JX 답변/NNG+을/JKO 제기/NNG+하/XSV+고/EC 나서/VV+었/EP+다/EF+./SF
4	소변/NNG+의/JKG 성분/NNG+이/JKS 부영양화/NNG+를/JKO 일으키/VV+었/EP+다/EF+./SF

그림 1. 개체격자 표상을 위해 형태소 분석된 텍스트 예문

그림 1과 같은 텍스트로부터 개체격자 표상을 구성하기 위해서는 단어-문장 행렬을 구축해야 하므로 텍스트로부터 단어 벡터를 추출해야 한다. 하지만 한국어는 영어와 달리 보통 한 단어가 공백을 통해 분리되어 있지 않고 두 명사가 공백을 두고도 함께 복합명사를 이루는 경우가 많기 때문에, 본 연구에서는 고유명사

구의 관심사가 아니므로 비교모형 후보에서 제외하였다.

(NNP) 혹은 일반명사(NNG)가 서로 연속해서 쓰이는 경우에 이를 한 단어로 취급하였다.[4)]

개체기반 접근은 선행사와 대명사 사이의 지시관계, 단어의 문법정보(syntactic role)를 활용하므로 엄밀하게는 텍스트 문장에 대한 대용어 해소(anaphora resolution)와 구문분석이 이루어져야 한다. 하지만 현재 한국어 텍스트에 대해 공개된 범용 대용어 해소도구를 찾기 힘든 실정이므로 본 연구에서는 응용되지 못했다. 그리고 마찬가지로 활용가능한 한국어 구문분석기를 찾을 수 없었으나 본 연구는 한국어에서 풍부하게 드러나는 격(case) 표지를 활용하여 이 문제를 우회하고자 하였다.

한국어에서 명사는 후속하는 격표지에 의해 그 어휘의 통사적 역할을 나타내는 경우가 많다. 일반적으로 '이/가'는 주어역할을 하는 명사 뒤에 붙으며, '은/는'은 주제(topic)를 가리키는 격표지로 흔히 사용된다. 그리고 '을/를'은 목적어인 명사 뒤에 주로 붙는다. 본 연구는 이와 같은 한국어 화자의 일반 언어지식을 활용하여 '이/가' 표지를 가진 명사는 주어 역할을, '은/는'의 경우에는 주제어 역할을, '을/를' 표지를 가진 어휘는 목적어 역할을 갖는다고 보았다. 표 1은 그림 1의 예문을 이와 같은 방식에 따라 개체격자로 변형한 결과를 보여준다.[5)]

표 1. 텍스트의 개체격자 표상 예제. 각 셀의 입력값은 해당 열의 명사가 각 문장에서 갖는 문법적 역할을 가리킨다: subj(주어), topic(주제어), obj(목적어), etc(그 외), x(없음).

	사람	소변	물고기	일련	환경 변화	질식	전문가	답변	제기	성분	부 영양화
S1	etc	subj	obj	x	x	x	x	x	x	x	x
S2	x	subj	obj	etc	obj	etc	x	x	x	x	x
S3	x	x	x	x	x	x	topic	obj	etc	x	x
S4	x	etc	x	x	x	x	x	x	x	subj	obj

4) 따라서 "세계/NNG+평화/NNG는 멀었다." 혹은 "런던/NNP+탑/NNG"과 같이, 두 일반명사 혹은 고유명사가 연속한 경우에는 이를 한 단어로 취급하였다('세계평화', '런던탑').

5) 주격표지와 주제격 표지의 구분은 주지하다시피 항상 명확한 것은 아니며 의미적으로 주어역할을 하는 명사에 대해서도 주제격 표지는 곧잘 쓰인다. 본 연구에서는 실험조건을 주격/주제격 표지 구분조건과 병합조건으로 비교하여 각 결과를 따로 관찰하였다.

표 1의 행은 그림 1의 문장들을 순차적으로 나타낸 것이므로 총 네 개이며, 각 열은 텍스트에 등장한 명사들을 가리킨다. 이 표는 개체기반적 접근에서 필요로 하는 정보를 효과적으로 집약해서 보여주는데, 개체들이 n번째 문장에서 n+1번째 문장으로 이행시 그들의 통사적 자질변화를 잘 시각화한다. 중심화 이론을 이용한 접근에서는 규칙들을 이용해 담화개체들의 국소적 이행과정에 제약을 부여하는 방법을 쓰는 것에 반해, 개체격자 모형은 개체자질간의 이행확률을 계산하여 문서전체를 국소적 이행패턴의 분포로써 요약하는 방법을 제안한다. 따라서 자질집합 {subject, topic, object, etc, x}의 원소들 사이의 이행확률(transition probability) 분포가 하나의 문서에 대해 주어지게 된다. 예를 들어, 표 1의 subj에서 subj로 이어지는 연쇄는 한 경우만 존재하므로(문장 1에서 문장 2로) 이 텍스트의 subj-subj 이행확률은 0.03(1/33)이 된다. 반면에 x-x 이행패턴은 전부 13개이므로 이것을 이항 이행연쇄 전체 가능 개수인 33((문장개수-1)×단어수)으로 나누면 0.39가 된다. 이와 같은 방식으로 모든 자질과 자질간의 이행확률을 한 문서에 대해 구성할 수 있기 때문에, 한 문서로부터 총 25개 길이의 벡터가 추출될 수 있다(전부 5개의 자질이 있으므로, 이항연쇄가 가능한 경우의 수는 5^2이므로 25개). 그리고 [5]의 통사역 랭킹방식을 따라 만일 한 개체가 같은 문장에서 다른 문법적 역할을 갖고 두 번 이상 출현하면 다음과 같은 순위법칙에 따라 개체의 문법적 역할을 결정한다(주어 > 주제어 > 목적어 > 그 외).

개체격자 표상의 자질벡터 추출

중심화 이론과 같은 개체기반적 접근은 한 문장에서 다음 문장으로 이행시 개체들에 대한 담화초점의 이동에 제약을 부여하므로 강하게 응집된 텍스트에서 담화초점이 맞춰진 명사들은(salient entities) 문서내 발화들(문장들)에서 중요한 문법적 역할을 가지고 자주 출현하게 된다. 이는 개체격자 표상에서 쉽게 관찰되는데 표 1의 단어 '소변'은 다른 단어들에 비해 잦은 빈도(3회)와 중요한 역할(두 번을 주어로서)을 띠고 여러 문장들에 걸쳐 등장한다. [5]는 개체기반적 접근의 일반적인 가정, 즉 응집성이 강한 텍스트의 개체들의 분포는 그렇지 않은 텍스트에 비해 규칙

적이라는 가정에 따라 특정 중요 어휘는 이행패턴이 주어 혹은 목적어의 연쇄로서 나타날 확률이 높다고 가정한다(표 1의 '소변'과 '물고기'의 경우처럼).

위와 같은 방식으로 얻은 이행확률들의 벡터는 각 문서에 대해 하나씩 추출될 수 있으며 이 벡터는 모두 고정된 크기를 갖고 있다. Barzilay와 Lapata는 이 벡터들이 학습 알고리즘에 이용하기 적합한 형태라는 것을 지적하고, SVM(Support Vector Machine) 알고리즘 사용을 제안한다.[5] 본 연구도 이와 같은 방법론에 의해 수행되었으며 자연어 생성 연구 방법론의 생성후-서열화(generate-and-rank) 패러다임에 따라 학습자료의 문장들을 무작위로 뒤섞은 무작위 자료(random permutations)와 함께 학습시킨 이후에 분류시스템의 성능을 테스트 자료에 대해 검증하였다.

LSA에 기반한 텍스트 응집성 측정

LSA(Latent Semantic Analysis)는 큰 규모의 말뭉치의 어휘들에 대한 통계적 계산에 의해 단어들 간의 맥락적 사용의미를 분석해준다[15]. 이 분석은 어휘들이 사용되는 맥락적 양상을 정량적으로 분석할 수 있는 길을 제공해주고 문서간의 의미적 거리를 측정하는데 매우 효과적인 모형이라는 것이 기존의 연구들에 의해 알려졌다[15, 16]. 이 모형은 여러 형태의 변이들이 있지만 일반적으로 벡터공간 모형에 기초하여 텍스트를 표상하고, 로그함수를 이용해 가중치를 부여하며 SVD(Singular Value Decomposition)를 통해 데이터 노이즈를 줄이고, 최종적으로 문서 벡터끼리의 거리를 계산하여 의미적 유사도를 추출하는 방식을 가리킨다[17].

[6]은 문서의 문장과 문장간의 의미적 유사도를 순차적으로 구하고 합산한 결과의 평균을 문서 응집도를 계산하는 한 가지 방식으로 제안하였는데, 이는 앞서 기술한 개체기반적 접근법과 마찬가지로 이전 문장과 이후 문장과의 응집성을 이용하여 문서의 전체 응집성에 접근하는 방법이므로 비교하기 적절한 측면이 있다. 하지만 한 가지 중요한 차이점은 개체격자 모형은 문서를 어휘 자질간의 이행패턴의 분포로서 요약하여 개별어휘의 정보에는 무관심한 것에 반해 LSA를 이용한 방법은 어휘들이 구체적으로 사용되는 정보(lexical information)에 주로 의존한다는 것이다. 본 연구에서도 [6]과 [5]와 같이 코사인 유사도를 이용하여 문장 간의 코사인

유사도를 구하여 국소적 응집도를 측정하려 시도했으며 이는 공식 (1)과 같다. 각 문장은 동일한 길이의 단어벡터로서 표현되었으며 벡터의 값들은 특정 단어 x가 문장 S_i에서 출현한 빈도값에 tf-idf 변환을 적용한 값이 이용되었다.

$$Sim(S_1, S_2) = \cos(\overrightarrow{S_1}, \overrightarrow{S_2}) \tag{1}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

$$= \frac{S_1 \cdot S_2}{\sqrt{S_1^2} \cdot \sqrt{S_2^2}}$$

공식 (2)는 공식 (1)을 통해 얻은 선행문장과 후속문장에 대한 코사인 유사도의 단순 평균을 구하는 것으로 LSA 기반 문서 응집도 측정은 이 공식을 통해 이루어진다.

$$Coherence(Doc) = \frac{\sum_{i=1}^{n-1} \cos(S_i, S_{i+1})}{n-1} \tag{2}$$

실 험

데이터

학습 데이터

포털업체 네이버(news.naver.com)의 뉴스사이트 사회 부문의 '환경' 섹션에서 100개의 뉴스기사들을 수집하였다. 뉴스기사는 다양한 언론사에 의해 작성되었으며 동일한 내용의 기사를 배제하였고, 기사당 평균 문장 수는 13.2개였다. 21세기 세종계획에서 공개한 지능형 형태소 분석기를 사용하여 기사를 분석 하였으며[6], 각

기사마다 무작위로 재배열한 20개의 무작위-배열 기사들을 생성하였다. 데이터 수집은 수집기간 내(2012년 6월~7월)에 시간 순으로 제시된 기사들을 텍스트 파일로 저장하는 방식으로 이루어졌다.

실험 데이터

위와 마찬가지의 방식으로 다른 신문기사 50개를 수집하였다. 기사당 평균 문장 수는 14.3개였으며, 학습용 데이터와 마찬가지로 각 기사당 20개의 무작위-배열 문서를 생성하였다.

말뭉치 자료

21세기 세종계획 말뭉치 자료에서 사전에 정교하게 형태소 분석된 279개의 문서자료를 활용하였다. 이 말뭉치는 생활, 환경, 정치 및 경제 등 다양한 주제를 포괄하며 보도기사, 사설, 시사잡지 등 다양한 출처로부터 수집되었다.

방법

실험과제는 다음과 같다. 실험데이터인 50개의 신문기사와 각각 생성된 20개의 무작위-배열 문서 쌍에 대해, 원래의 신문기사를 무작위-배열 문서보다 높은 응집도를 가진 문서로서 올바르게 판정하는 것이다. 따라서 두 모형 모두 총 1000개의 {본래 문서, 무작위-배열문서} 쌍에서 원래 문서를 대립항(무작위-배열 문서)보다 응집도가 높은 것으로 평가하는 것이 과제목표가 된다. 이리한 실험방법은 본래 사람에 의해 작성된 문서의 응집도는 그 문서의 문장들이 무작위 순서로 배열된 문서의 응집도보다 높다는 가정 위에서 설계되며 본 연구에서 참고한 기존연구들의 실험방법과 동일하다[5, 14]. 유의할 점은 개체격자 접근법과 LSA 기반 접근법은 모두 위와 같은 실험과제를 공유하지만 같은 학습데이터를 공유하지는 않는다는 것이다. LSA 기반 접근법은 그 특성상 의미 공간 구축을 위한 학습 데이터를 별도로 필요로 하는데, 먼저 큰 규모의 말뭉치 자료로부터 추출된 단어-문서 행렬

6) http://www.sejong.or.kr/

구축이 필요하다. 따라서 LSA가 사용하는 학습데이터는 개체격자 모형이 취하는 학습자료와 다르다. 하지만 두 접근의 실험목표는 모두 기술된 바와 같이 동일하다.

LSA 기반 모형 실험방법

세종 말뭉치 데이터로부터 명사와 동사에 해당하는 총 153,861개 단어를 추출하였고 이들 단어들을 행으로 하고 279개의 문서를 열로 배열하여 153,861×279 크기의 단어-문서 행렬을 구성하였다. 이 행렬의 각 셀에는 해당 단어가 해당 문서열에서 출현한 빈도값을 로그함수를 통해 변환한 값이 입력되었는데 이는 Salton의 전형적인 가중치부여 방식을 따른 것이다(tf-idf weighting)[18, 19].

그리고 보다 효율적인 분석을 위해 일반적으로 응용되는 SVD(Singular Value Decomposition)를 이용하여 행렬자료의 차원을 축소하였고 이 벡터공간에 실험 텍스트 자료의 문장들을 벡터로 사상하여 이들 벡터간의 코사인 유사도를 앞에서 기술한 공식에 의해 구하였다.[7] 그리고 주어진 행렬구조의 길이가 크기 때문에 보다 효율적인 연산을 위해서, 유사도를 구할 때마다 벡터공간을 재구축하지 않고 이미 구축된 벡터공간에 새로운 입력벡터를 끼워넣는 방식인 'Folding-in' 기법을 사용하였다[20, 21]. 차원감소(dimension reduction)는 효과적인 LSA에서 중요한 부분이지만 유감스럽게도 이 문제는 경험적인 주제에 가까운 문제로서 기존연구에서는 100개의 차원으로 축소하여 분석하였다[5, 6]. 본 연구의 실험에서는 50, 100, 그리고 150개의 차원으로 축소하여 분석 후 각각 그 결과를 관찰하였다.

개체격자 기반 모형 실험방법

실험방법의 절차는 크게 학습절차와 실험절차로 이루어진다. 우선 학습데이터로서 수집된 100개의 신문기사와 각 신문기사에 대해 생성된 20개의 무작위-배열 문서들을 합친, 총 2000개의 텍스트 쌍에 대해 개체격자 표상을 구축하고 이로부터 3과 4에서 기술된 방식을 통해 자질벡터 쌍을 추출한다. 이 2000개의 자질벡터 쌍

7) SVD와 벡터공간 재구축을 위해서 Python 기반 수치해석 모듈인 Scipy 패키지에서 제공하는 선형대수 함수를 이용하였다.(http:/bj/docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.linalg.svds.html)

은 각각 <V_o, V_p>의 형태를 가지는데 V_o는 원래 문서로부터 추출된 자질벡터이고 V_p는 그 문서의 무작위-배열에 의해 생성된 문서로부터 추출된 벡터를 가리킨다.

이 벡터 쌍들은 SVM 기반 학습 알고리즘을 통해 자동 분류기를 학습시키는데, 이 학습과정은 입력된 벡터 쌍(pairwise vectors)에서 무작위-배열문서 벡터가 그렇지 않은 문서벡터에 대해 상위로 판정되는 경우를 최소화하는 최적화된 서열화 함수(ranking function)를 제공하는 모수 벡터(parameter vector)를 찾는 과정이다[5]. 본 실험에서 이 학습과정은 공개된 패키지인 SVM_{rank}를 통해 구현되었다[22].

위 학습과정을 통해 학습된 자동 분류기는 실험데이터인 50개의 신문기사×20개의 무작위-배열문서에서 추출된 1000개의 벡터 쌍을 분류하도록 하였다. 관찰하고자 한 실험 요소는 두 가지로, 격표지를 이용하여 부여한 문법적 역할정보가 개체격자 기반 시스템의 성능을 향상시키는지, 그리고 그 중에서도 주격 표지와 주제격 표지의 구별이 의미를 갖는지를 보는 것이었다.

실험의 성공적인 수행을 평가하기 위한 지표로는 [5]와 [14]에서 사용한 것과 마찬가지로 실험데이터로 주어진 1000개의 벡터 쌍 중에서 정확하게 무작위-배열 문서의 벡터를 하위로 판정한 개수의 비(ratio)를 정확도로 해석하고 이용하였다. 이 지표의 기저율은 50%이고 공식은 (3)과 같다.

$$\text{Accuracy} = \frac{CorrectN((V_o, V_p))}{Total((V_0, V_p))} \tag{3}$$

결 과

실험을 통해 관찰하고자 하는 첫 번째 요소는 언어적 요소의 기여도이다. 본 연구의 개체격자 모형은 주어와 주제격 조사를 모두 활용할 경우에 25개의 자질을 갖는 자질벡터가 생성되지만, 주어와 주제격 조사를 같은 자질로 통합할 경우에는 전부 16개 자질의 자질벡터가 만들어진다(subj-subj, subj-obj, subj-etc, subj-x, obj-subj, obj-obj, obj-etc, obj-x, etc-subj, etc-obj, etc-etc, etc-x, x-subj, x-obj, x-etc, x-x). 그리고 가장 단순하게 문법적 역할에 대한 정보를 무시하고 해당 어휘가 문장에 있음 혹은

없음으로 처리하면 단지 네 개 길이의 자질벡터가 만들어진다. 즉, {'etc-etc', 'etc-x', 'x-etc', 'x-x'}의 네 가지 이행패턴만이 가능하다. 따라서 이렇게 다른 길이의 자질벡터 모형간의 비교는 한국어의 격표지 정보가 제공하는 통사역이 실험과제 수행에 어떤 기여를 하는지 관찰할 수 있도록 도와준다. 이 결과는 표 2에 제시되어 있는데 위에서 기술된 25-자질벡터(5개 통사역), 16-자질벡터(4개 통사역) 그리고 4-자질벡터(단순모형)의 결과를 각각 보여준다. 5개의 통사역을 사용한 모형의 정확도가 가장 높았으며, 4개의 통사역을 사용한 모형의 수행이 가장 낮았다. 하지만 이들 간의 차이는 통계적으로 유의하지 않았다(본 연구에서는 [5]를 따라 Sign test 통계검증을 사용하였다).

표 2. 테스트 자료에서 얻은 실험조건별 정확도

모형	정확도
5-통사역	70.2
4-통사역	69.1*
단순모형	69.5
LSA-기반 모형	73.3

표 3은 또한 LSA 기반 모형의 결과도 함께 제시하고 있으며 이는 LSA 실험모형 중에서 가장 높은 결과값을 사용한 것이다(150개 축소차원을 사용한 모형). 정확도 옆에 표시된 별표('*')는 그 정확도를 가장 높은 결과값인 LSA-기반모형의 정확도와 비교했을 때, 그 차이가 통계적으로 유의미했음을 보여준다($p < 0.05$). 결과적으로 위 표에서 4-통사역을 이용한 개체격자 모형은 LSA-기반모형에 비해 유의미하게 낮은 수행을 보여주었다. 단순 개체격자 모형의 정확도 결과는 LSA-기반모형의 결과에 비교해 통계적 유의미성 테스트를 통과하지는 못했으나 거의 유의미한 차이를 보여주었다(p =0.053). 하지만 5개 통사역을 사용한 모형과의 차이는 통계적으로 전혀 유의미하지 않았다(p =0.53).

이 결과를 요약하면, LSA-기반 접근의 모형은 일부 개체격자 모형(4-통사역, 단순모형)의 수행보다 높은 수행을 유의미하게 보여주었으나 5개 통사역을 사용한

모형에 대해서는 상대적으로 그렇지 못했다. 이는 LSA-기반 접근 모형과 5개-통사역을 활용한 개체격자 모형이 통계적으로 차이를 관찰할 수 없는 수준의 수행을 보였다는 것을 의미한다. 5-통사역 모형와 4-통사역 모형, 단순모형 결과 간의 직접적 통계적 차이는 별도로 또 다른 Sign test를 통해 관찰하였지만 통계적으로 유의미하지 않았다.

LSA에 기반한 모형들 간의 차이는 SVD의 차원축소 과정에서 선택한 차원들의 개수에 의해 결정되는데(원형 데이터의 차원크기는 279) 각각 150개, 100개, 50개의 LSA-기반모형들이 실험되었고 이들의 결과는 그림 2에 제시되어 있다.

그림 2에 제시된 것처럼 축소된 차원의 수가 150일 때, 가장 높은 수행을 보여주었으며(정확도는 73.3%), 이 차원수가 50일 때는 68.9%, 차원수가 100인 조건에서는 71.5%의 정확도를 각각 기록했다. 차원수 150인 조건의 정확도와 차원수 50인 조건 간의 정확도의 차이는 통계적으로 유의미했다($p < 0.05$),

하지만 차원수가 150인 조건과 차원수 100인 조건 간의 차이는 유의미하지 않았다($p = 0.089$). 이 결과는 LSA-기반 모형 사용시 최적화된 SVD 사용은 수행결과를 유의미하게 향상시킬 수 있다는 것을 함의한다.

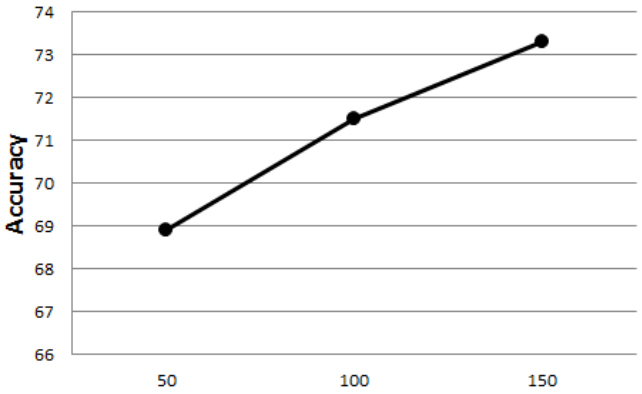

그림 2. LSA의 축소차원 개수에 따른 정확도

논의 및 결론

본 연구에서는 언어적 측면에서, 한국어의 격표지 정보가 텍스트 응집도 측정에 미치는 정도를 개체격자 모형을 통해 관찰하기를 기대하였다. 그 결과 주제격 격표지와 주격 표지를 모두 활용한 5개 통사역 모형의 수행이 가장 높음을 확인할 수 있었지만, 이 차이는 다른 모형에 비해 통계적으로 유의미할 만큼 크지 않았다. 그러나 [5]의 선행연구에서도 이 통사적 요소의 통계적 유의미성은 두 종류의 실험데이터 세트 중 하나에서만 관찰되었고(p. 16), 독일어에 대해 같은 실험을 진행한 [14]는 오히려 시스템의 수행결과가 떨어진다는 보고를 하였다. 하지만 본 연구의 실험에서는 표 2에서 드러나듯이 오히려 5개-통사역 모형의 수행이 가장 높았고, 오직 5개 통사역을 사용한 모형만이 LSA-기반 모형과의 비교에서 통계적으로 유의미한 차이를 보이지 않았다는 점을 고려할 필요가 있다. 이는 보다 정교한 통사역 분석을 개체격자 모형에 적용할 경우, 더 나은 결과가 나올 가능성을 가리킨다. 특히 본 연구에서는 오직 격표지에 의존하여 어휘의 통사역을 분석했기 때문에 보다 정교한 구문분석 시스템에 비해서 정확성이 떨어질 것이라는 점과 4개 통사역을 활용한 개체격자 모형도 통사역이 없는 단순모형에 비교하여 거의 차이가 없었다는 점(69.1 vs. 69.5)을 고려하면 그러한 가능성은 더욱 높아 보이며 후속 연구가 필요하다.

실험결과, LSA-기반 모형은 가장 높은 수행을 보였지만(73.3%) 이는 차원축소를 150으로 한 경우에만 해당되며, 차원축소가 50인 조건, 100인 조건들과 개체격자 모형과의 차이는 크지 않았고 이들 사이의 통계적 차이도 유의미하지 않았다. [5]의 실험에서도 LSA-기반모형과 개체격자 모형간의 차이는 대용어해소 모듈을 사용하지 않을 경우 거의 비슷한 정도의 수행을 보였다는 점을 보면, 이 결과는 선행연구의 결과와 일관적이라고 볼 수 있다. 또한 [5]의 연구에서 개체격자 접근의 가장 단순화된 모형과 LSA-기반 모형은 각각 81.4%와 81%의 정확도를 기록하여 본 연구의 단순화 모형의 결과에 비해 약 10%정도 높았지만, 같은 LSA-기반 분석을 사용한 본 연구의 결과의 정확도가 70% 정도였다는 점을 볼 때 이러한 결과차이가 실험방법의 차이가 아니라 실험데이터의 차이에서 오는 것임을 간접적으로 확인해 준다. 본 연구의 실험데이터가 여러 신문사의 동일주제에 대한 기사임에 비

해, [5]의 데이터는 하나의 출처로부터 제공되어 데이터 내 차이가 상대적으로 적을 것이라는 점도 고려되어야 한다.

선행연구의 개체격자 모형은 선행사와 대명사와의 지시관계 정보를 사용하였고, 이들 지시관계 정보들은 시스템의 수행을 높이는 것으로 보고되었다[5, 14]. 하지만 본 실험에서 이 요소는 응용가능하지 않았고 또한 연구의 관심대상과는 거리가 멀었기에 실험과정에서 제외되었다. 그러나 이 대용어 해소과정은 어휘자질 간의 이행패턴의 정확도를 증가시키길 것이기에 개체격자 모형의 장점을 실용적 측면에서 누리기 위해서는 향후 도입되어야 한다. 한국어는 대명사의 사용이 영어에서처럼 자주 나타나지 않지만, 그에 해당하는 대용어는 종종 등장하므로(예, 김교수, 김박사 등) 이에 대한 연구를 활용해야 할 것이다.

개체격자 기반모형과 LSA-기반 모형은 본 연구에서 유사한 정도의 수행을 보였지만 개체격자 모형은 향후 대용어 해소와 함께 보다 향상된 통사역 분석시, 수행의 개선가능성이 있다는 차이를 간과해서는 안 될 것이다. 그리고 LSA에서 SVD를 실시할 때, 가장 적정한 차원수를 결정하는 것은 경험적인 문제로서 연구자의 주의가 필요하다. 또한 LSA에 기반한 문장정렬 방법은 문서가 포함하는 문장수 크기에 비례하는 행렬연산을 요구하므로 계산적인 측면에서 비용이 크며, 사전에 구축된 큰 규모의 말뭉치가 필요하다는 점들이 고려되어야 한다. 반면 계산비용의 측면에서 개체격자 모형은 상대적으로 장점을 지니는데, 학습데이터의 학습 후에는 오직 문서단위의 국소적인 정보만을 요구하므로 큰 규모의 행렬구조 연산을 필요로 하지 않기 때문이다. 그러므로 어떤 모형이 한국어 문장정렬에 응용하기에 적절한 모형인가는 연구자의 판단과 주의가 요구된다.

요약하면 본 연구는 한국어 텍스트에 대한 문장정렬 과제에 개체격자 모형을 적용하는 것이 LSA를 사용한 고전적인 접근방법에 비해 나쁜 선택이 아니라는 점을 보였다. 또한 영어권에 한해 텍스트의 국소적 응집도를 측정하는데 있어서 유용하다고 알려진 어휘들의 통사적 역할 정보가 한국어에 대해서도 마찬가지로 유효할 수 있다는 가능성을 보였다. 이는 향후 정교한 통사역 분석의 활용시 개체격자 모형의 성능을 향상시킬 수 있다는 점을 뜻하며, 개체격자 모형을 보다 발전시킬 필요가 있다는 것을 지시한다. 아울러 LSA 기반 접근법의 경우 양질의 말뭉치와 방대한 계산량이 허용가능하다면 언어학적 요소에 대한 고민 없이 문장정렬 과

제를 수행할 수 있다는 점을 재확인하였다. 향후에는 이와 같은 연구로부터 문서 자동요약이나 작문평가, 텍스트 자동생성 등의 연구 및 개발이 이루어질 수 있을 것이며, 이를 위한 한국어의 특성을 이용한 텍스트 응집도 측정이론의 발전이 필요할 것이다.

개체격자 모형은 중심화 이론의 기본원리와 가정들을 받아들이면서도 규칙기반 시스템이 아닌, 학습 알고리즘을 이용하여 학습후-평가를 수행하는 개체기반적 기계학습 시스템이라고 평가할 수 있다. 본 연구의 의의는 한국어 텍스트에 대해서는 최초로 개체격자 모형을 응용하여 문장정렬 과제를 수행하였고 이 결과를 전통적인 LSA-기반 모형과 비교하였다는 데 있다. 하지만 본 연구는 기존의 LSA-기반 방법과 개체격자 모형이 유사한 수준의 성능을 가진다는 것을 보였지만, 어떤 방법이 일관적으로 더 나은 성능을 보이는지를 관찰하지는 못하였다. 따라서 후속연구를 통해 보다 정확한 구문분석과 대용어 해소의 적용이 얼마나 시스템의 성능을 향상시킬 수 있을지 실험적으로 관찰할 필요가 있다.

참고문헌

[1] Scott, D., & de Souza, C. S. (1990). Getting the message across in RST-based text generation. Current research in natural language generation, 4, 47-73.

[2] Kibble, R., & Power, R. (2004). Optimizing Referential Coherence in Text Generation. Comput. Linguist., 30(4), 401-416.

[3] Reiter, E., & Dale, R. (2000). Building natural language generation systems: Cambridge University Press.

[4] McKoon, G., & Ratcliff, R. (1992). Inference during reading. Psychological review, 99(3), 440-466.

[5] Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. Computational Linguistics, 34(1), 1-34.

[6] Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. Discourse Processes, 25(2-3), 285-307.

[7] Kuno, S. (1972). Functional sentence perspective: A case study from Japanese and English. Linguistic inquiry, 3(3), 269-320.

[8] Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects and topics. C. Li (ed.), Subject and topic, 25-56: New York, Academic Press.

[9] Karttunen, L. (1976). Discourse referents. In D. James & J. McCawley (Eds.), Syntax and Semantics: Notes form the linguistic underground (Vol. 7, pp.363-386). New York: Academic Press.

[10] Givon, T. (1987). Beyond foreground and background. Coherence and grounding in discourse, 11, 175-188.

[11] Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: a framework for modeling the local coherence of discourse. Comput. Linguist., 21(2), 203-225.

[12] Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), Radical pragmatics (pp.223-255). New York: Academic Press.

[13] 노지은, 나승훈, 이종혁 (2007). 중심화 이론을 이용한 텍스트 구조화. 정보과학회논문지: 소프트웨어 및 응용, 24(6), 572-583.

[14] Filippova, K., & Strube, M. (2007). Extending the entity-grid coherence model to semantically related entities. Paper presented at the Proceedings of the Eleventh European Workshop on Natural Language Generation.

[15] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2-3), 259-284.

[16] Dumais, S., & Landauer, T. (1997). A solution to Platos problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological review, 104(2), 211-240.

[17] Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37(1), 141-188.

[18] Salton, G. (1971). The SMART retrieval system-experiments in automatic document processing.

[19] Salton, G., Wong, A., & Yang, C. -S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[20] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407.

[21] Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. SIAM review, 37(4), 573-595.

[22] Joachims, T. (2006). Training linear SVMs in linear time. Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA.

1 차원고접수 : 2013. 04. 11
2 차원고접수 : 2013. 09. 26
최종게재승인 : 2013. 10. 30

(*Abstract*)

A comparative study of Entity-Grid and LSA models on Korean sentence ordering

Youngsam Kim[1)] Hong-Gee Kim[2)] Hyopil Shin[3)]

[1)]Cognitive Science Program, Seoul National University

[2)]Biomedical Knowledge Engineering Laboratory, Seoul National University

[3)]Department of Linguistics, Seoul National University

For the task of sentence ordering, this paper attempts to utilize the Entity-Grid model, a type of entity-based modeling approach, as well as Latent Semantic analysis, which is based on vector space modeling, The task is well known as one of the fundamental tools used to measure text coherence and to enhance text generation processes. For the implementation of the Entity-Grid model, we attempt to use the syntactic roles of the nouns in the Korean text for the ordering task, and measure its impact on the result, since its contribution has been discussed in previous research. Contrary to the case of German, it shows a positive result. In order to obtain the information on the syntactic roles, we use a strategy of using Korean case-markers for the nouns. As a result, it is revealed that the cues can be helpful to measure text coherence. In addition, we compare the results with the ones of the LSA-based model, discussing the advantages and disadvantages of the models, and options for future studies.

Key words : sentence ordering, coherence, Entity-Grid model, Latent Semantic Analysis, SVM