Sent2Vec을 이용한 아동 발화의 화제 변경 탐지

허탁성[1,O], 이윤경[2], 김유섭[3]

한림대학교 융합소프트웨어학과[1]
한림대학교 언어청각학부[2]
한림대학교 소프트웨어융합대학[3]
gjxkrtjd221@gmail.com, ylee@hallym.ac.kr, yskim01@hallym.ac.kr

Detection of Topic Changes in Child Speech Using Sent2Vec

Tak-Sung Heo[1,O], Yoon-Kyoung Lee[2], Yu-Seop Kim[3]
Hallym University, Department of Convergence Software[1]
Hallym University, Division of Speech Pathology and Audiology[2]
Hallym University, College of Software[3]

요 약

언어 병리학에서는 연령별 대화 능력 발달에 관한 연구에 관심이 많다. 하지만 이러한 연구는 많은 시간과 비용이 소모된다. 이를 해결하기 위해, 본 연구에서는 대화 능력 발달의 많은 연구 중 한 가지 방법인 화제 변경을 Sent2Vec을 이용하여 자동적으로 탐지해주는 방법을 제안한다. 아동의 연속된 두 발화를 비교하여 화제의 변경을 Sent2Vec의 코사인 유사도를 통해 찾아냈다. 본 연구에서는 언어 병리학에서의 연구 결과와 비교를 하기 위해 초등학생 1학년, 3학년, 5학년 집단의 데이터를 사용하였다. 본 연구에서 제안한 방법의 결과와 언어 병리학에서 연구한 결과의 상관관계가 99.95%로 매우 높음을 확인할 수 있었다. 이러한 화제 변경 탐지를 자동화함으로써, 언어 연구에 필요한 시간과 비용을 크게 절감할 수 있다.

주제어: 화제 관리 능력, 화제 변경, Sent2Vec

1. 서론

언어 병리학이란 과학적으로 언어에 대한 본질적 성질을 연구하는 분야이다. 이러한 분야는 언어 능력과 청각 능력의 향상을 목적으로 둔다. 하지만 이는 발화자의 언어 능력을 판단을 하는데 많은 시간과 비용이 소모된다. 이를 해결하기 위해 최근에는 언어 분석을 자동화하기 위한 형태소 분석과 의존 구문 분석, 개체명 인식 등 여러 방법론들이 적용되고 있다 [1-3].

언어 병리학에서 말하는 화제 관리 능력은 대화 능력 발달 연구 중 한 가지 방법으로, 원활한 대화를 주고받기 위해 지속적으로 화제를 유지하거나, 또는 화제가 변경이 요구될 때 발화자가 화제를 제대로 변경하는 능력을 말한다 [4, 5]. 이러한 능력을 포함한 대화 능력은 학령기로 갈수록 점진적으로 발달하게 된다 [5].

이에 따라, 본 연구에서는 화제 관리 능력 중 화제 변경을 자동적으로 탐지해주는 방법을 제안한다. 제안 방법은 문장 단위를 임베딩하는 방식인 Sent2Vec을 이용하였다 [6, 7]. 본 연구는 네이버 기사를 통해 수집된 데이터를 사용해 코사인 유사도의 기준 값을 정한 후, 아동의 연속된 발화에 대한 코사인 유사도를 비교함으로써 화제 변경을 탐지하였다.

본 연구의 결과와 언어 병리학의 화제 변경 연구 결과인 [5]를 비교함으로써, 본 연구의 신뢰성을 확인한다. 이를 통해 언어 연구에 소모되는 시간과 비용을 크게 절감할 수 있다.

2. 관련 연구

다른 사람과의 관계 형성 및 유지를 위해서는 대화를 주고받는 능력이 중요하다 [4]. 원활한 대화가 이루어지기 위해서는 대화의 화제를 다루는 능력이 중요한데, 이를 화제 관리 능력이라고 한다 [8].

[5]에서는 국내 초등학생 1학년, 3학년, 5학년 집단인 각각 10명에 대한 화제 관리 능력에 대해 연구를 했다. 이 연구는 객관적인 결과를 얻기 위해 아동에게 '가정생활', '학교생활', '기타/친구'에 관한 익숙한 3가지의 큰 화제로 대화를 이어나갔으며, 아동의 데이터를 수집하였다. 이를 통해 [5]의 연구에서 화제 관리 능력 중 화제 변경 비율이 1학년에 비해 3학년이 낮으며, 3학년에 비해 5학년이 낮다는 결과를 얻었다. [5]에서의 화제 변경은 대화를 이어나가는 도중 이전 대화차례의 화제와 연결되어 있지 않거나 새로운 화제로 대화 화제가 바뀌는 경우를 말한다. 여기서 화제 변경 비율은 아동의 화제 변경 수의 합을 전체 대화차례로 나눈 것이다.

화제 변경 탐지를 위해 많은 연구들이 진행되어 왔으며, 최근에는 딥러닝을 이용하여 주제 변경 탐지를 위한 연구들이 존재한다 [9, 10]. 이들은 대규모 데이터셋을 사용했으며, 문서 단위의 주제 변화에 대한 것을 목표로 삼았다. 문서에서의 여러 자질들을 추출한 후 딥러닝 모델에 훈련을 시켰으며, 각 주제에 대한 태그들을 이용해 지도 학습을 시켰다.

본 연구에서는 화제 변경 탐지를 위한 기존 논문인 [9, 10]과 달리, 데이터가 문서가 아닌 문장으로 되어있으며,

각 문장에 대해 화제에 관한 태그가 되어있지 않다. 또한 본 연구의 데이터로는 딥러닝 모델에 학습하기에 양이 충분하지 않다. 그러므로 본 연구에서는 적은 데이터를 사용하여 문장 단위의 화제 변경 탐지를 위한 새로운 방법론을 제시한다. [5]의 화제 변경 탐지를 자동화함으로써, 언어 전문가의 개입 없이 언어 능력 연구에서의 화제 변경에 대한 객관적인 판단을 할 수 있다.

3. 데이터

본 연구의 아동 발화 데이터는 한림대학교 언어 병리학과[1]에서 수집한 데이터이다. 이 데이터들은 여러 집단의 데이터로 구성된다. 본 연구는 언어 병리학 연구인 [5]와 비교하기 위해 [5]에서 사용한 1학년, 3학년, 5학년 집단의 데이터를 사용하였다. 이러한 집단에서 한 사람의 발화의 개수는 평균 84개로 이루어져 있다. 이 데이터는 검사자와 아동의 발화에 의한 구어체로 구성되어 있으며, 표 1와 같다.

표 1. 데이터 예시

턴	발화	대화
<학교>		검 학교생활은 어때?
1	1	아 재미없어요.
		검 재미없구나. 그리고?
2	2	아 점심 먹고 양치하고
	3	아 바로 한 한시간 20분 놀아요.

표 1는 수집한 데이터 집단 중 한 명의 아동에 대한 발화 예시이다. 대화에서 '검'과 '아'는 검사자와 아동을 의미하는 것이며, 턴은 화자의 말이 대화 상대자가 말을 하기 전까지의 모든 발화를 말한다. 발화는 화자가 말을 하는 하나의 문장이다.

언어 병리학에서의 데이터는 검사자가 아동에게 다른 화제를 제시하는 질문이 존재한다. 이러한 질문을 통해 언어 병리학에서는 아동의 화제가 변경이 되었다면 이 또한 화제라고 판단을 하였다. 이를 바탕으로 화제 변경 탐지를 위해 검사자의 발화를 제거하였으며, 아동의 발화에서 '아'를 제거한 발화를 사용하였다.

4. 방법론

그림 1은 본 연구 방법을 보여준다. 우선, 코사인 유사도로 화제 변경을 탐지하기 위해서는 코사인 유사도의 기준 값 (Standard Score)을 정해야 한다. 기준 값을 정하기 위해 네이버 뉴스 기사의 첫 문장들을 수집하였다. 그 후, 동일한 카테고리의 문장들은 Sent2Vec을 이용하여 코사인 유사도를 구하였으며 그 값들의 평균을 코사인 유사도의 기준 값이라고 정하였다. 이 기준 값과 아동의 연속된 발화에 대한 코사인 유사도 (Child's Score)를 비교함으로써 아동 발화의 화제 변경을 탐지하였다.

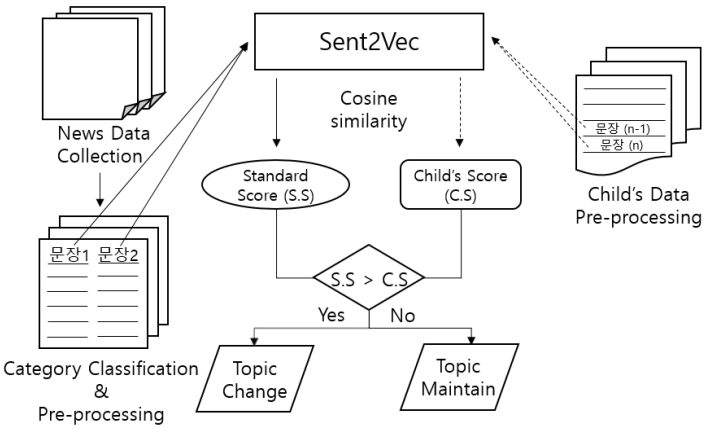

그림 1. 연구 방법의 도식화

4.1. Sent2Vec

본 연구에서는 문장쌍의 코사인 유사도를 구하기 위해 문장 단위를 임베딩하는 방식인 Sent2Vec을 사용하였다 [6]. 이는 단어 단위의 의미를 살리는 Word2Vec과 달리 문장 단위의 의미를 살리기 위해 문장 자체를 벡터화 시키는 방법이다.

Sent2Vec은 Word2Vec의 CBOW와 FastText를 확장시킨 모델이다 [6, 7]. Sent2Vec과 Word2Vec의 차이점은 자주 나타나는 단어들을 제거하는 서브 샘플링과 윈도우 사이즈를 랜덤하게 변경하는 다이나믹 컨텍스트 윈도우를 사용하지 않는 것에 있다. 왜냐하면 서브 샘플링에서는 자주 나타나는 단어가 중요한 단어일 수 있으며, Sent2Vec의 학습 방식인 n-gram의 생성을 방해하기 때문이다. 또한 다이나믹 컨텍스트 윈도우에서는 윈도우 사이즈가 전체 문장의 길이로 고정이 되어야 문장 그 자체를 학습할 수 있기 때문이다. 그리고 FastText와의 차이점은 FastText는 하나의 단어에서 서브 단어 단위를 추출하는 반면 Sent2Vec에서는 하나의 문장에서 단어 단위의 n-gram을 사용한다는 것이다. 또한 여기서 n-gram은 일반적인 n-gram이 아닌 uni-gram을 포함한 bi-gram의 개수를 뜻한다 [6, 7].

이로써 Sent2Vec은 그 문장 단위의 의미를 벡터로 나타낼 수 있다. 이렇게 벡터로 나타낸 문장들은 서로의 대한 유사성을 나타낼 수 있다. 이를 표현하기 위해 두 벡터의 내적을 구하는 코사인 유사도가 사용이 된다. 이는 벡터로 표현된 문장 또는 문서의 유사도를 구하는 방법으로 널리 사용되어 왔다 [11]. 코사인 유사도의 값은 1부터 -1의 값을 가지는데, 값이 1에 가까울수록 두 벡터의 유사성이 높다고 할 수 있다.

Sent2Vec의 코사인 유사도 성능은 문장을 음소, 형태소, 어절 단위로 나누는 것에 따라 달라진다. 본 연구에

[1] https://slp.hallym.ac.kr/user/indexMain.do?siteId=slp

서는 Sent2Vec의 성능을 비교하는 실험을 통해 의미를 가진 가장 작은 말의 단위인 형태소로 나누어 화제 변경을 탐지하는 실험을 하였다. 이 때, 사용한 형태소 분석기는 Kkma[2]를 사용하였다.

4.2. 코사인 유사도 기준 값 설정

아동의 연속된 발화에 대한 화제 변경을 탐지하기 위해서는 Sent2Vec을 이용한 코사인 유사도의 기준 값을 정할 필요가 있다. 왜냐하면 서로 같은 주제를 가진 두 문장에 대한 코사인 유사도의 절대적인 기준이 없기 때문이다.

우리는 서로 같은 주제를 가진 두 문장을 통해 기준 값을 정하였다. 만약 서로 다른 주제를 가진 문장들을 사용한다면, 실제로는 주제가 비슷한 문장이 주제가 다른 것처럼 사용되는 경우가 발생할 수 있다. 이 경우에는 계산되는 기준이 부정확해지므로 같은 주제를 가진 두 문장을 가지고 기준을 계산한다. 코사인 유사도 기준 값을 정하기 위해 건강 정보, 금융, 노동, 부동산, 북한, 사건/사고, 자동차, 중동/아프리카, 컴퓨터, 환경에 대한 네이버 뉴스 기사의 첫 문장을 크롤링하여 데이터로 사용하였다. 이 과정을 통하여 같은 주제를 가진 300개의 문장쌍을 추출할 수 있었다. 또한 어절, 형태소, 음소 단위로 Sent2Vec의 성능을 비교하여 형태소 단위로 Sent2Vec을 구현하였을 때 성능이 가장 좋다는 것을 확인하였으며, 이를 바탕으로 문장들을 모두 형태소 단위로 처리하였다. 그리고 문장쌍들은 형태소 단위로 나눈 문장들을 학습한 Sent2Vec을 이용하여 코사인 유사도를 구하였으며, 그 값들에 대한 평균 값인 '0.1187947'을 코사인 유사도 기준 값으로 사용하였다.

4.3. 화제 변경 비율 평가

아동의 화제 변경은 Sent2Vec을 이용한 기준 값과 아동의 연속된 두 발화의 코사인 유사도 비교를 통해 이루어진다. 여기서 연속된 두 발화란 아동 1명의 전체 발화에서 전산언어학에서 말하는 bi-gram의 문장단위를 뜻한다.

화제 변경 비율을 평가할 때 아동들의 발화 수는 전부 다르므로 이를 정규화 할 필요가 있다. 이는 수식 (1)과 같이 화제 변경 비율로 평가하였다. 여기서 화제 변경 개수는 코사인 유사도의 기준 값보다 아동 발화 쌍들의 코사인 유사도가 낮게 나타난 결과의 개수이다.

$$\textbf{Topic change Rate} = \frac{\textit{화제 변경 개수}}{\textit{총 발화의 개수}} \quad (1)$$

5. 실험

5.1 Sent2Vec의 성능 비교

Sent2Vec의 성능을 비교하고자 Sent2Vec의 모델은 19,562,522 문장으로 이루어진 국민대학교의 한국어 말뭉치[3]를 음소, 형태소, 어절 단위로 나눠 학습시켰다. 이렇게 학습된 Sent2Vec의 모델들의 성능을 평가하기 위해 한국어로 번역한 Kaggle의 Quora Question Pairs Dataset[4]과 ExoBrain Korean Paraphrase[5], 한국외국어대학교의 독어 번역쌍 등 총 2997문장의 평가 데이터를 수집하였다. 여기서 Kaggle의 Quora Question Pairs와 한국외국어대학교의 독어 번역쌍은 유사도가 0과 1로 되어 있으나, ExoBrain Korean Paraphrase의 데이터는 다른 데이터와 다르게 유사도가 0~5로 되어있다. 이러한 ExoBrain Korean Paraphrase 데이터는 다른 데이터들의 유사도와 같게끔 유사도 4와 5는 1로 변경하였으며, 0과 1로 된 유사도 값은 0으로 변경하여 Sent2Vec의 성능을 비교하였다. 여기서 유사도가 1인 값은 문장쌍이 의미적으로 유사하다고 판단되는 값이며, 0인 값은 의미적으로 유사하지 않다고 판단되는 값이다. 이 실험을 하기 위해 평가 데이터 또한 음소, 형태소, 어절 단위로 나누었으며, 코사인 유사도 성능을 비교하기 위해 train set과 test set을 5:5 비율로 나누어 평가하였다.

표 2. Sent2Vec 코사인 유사도 성능 비교

단위	Precision	Recall	F1-Score	Accuracy
어절	0.6994	0.7755	0.7355	0.7258
형태소	0.8199	0.8098	0.8148	0.7993
음소	0.7823	0.697	0.7372	0.7258

표 2에서 보이는 바와 같이, 어절과 음소 단위보다 형태소 단위로 Sent2Vec을 구성할 때 가장 좋은 성능을 보인다는 것을 알 수 있다. 이에 따라 본 연구에서는 코사인 유사도 기준 값 설정을 위한 뉴스 데이터와 Sent2Vec의 학습 데이터, 아동의 모든 발화들을 형태소 단위로 나누어 연구를 진행하였다.

5.2 결과

아동의 이전 발화와 이후 발화의 Sent2Vec의 코사인 유사도의 값을 구한 후, 이 코사인 유사도 값이 코사인 유사도 기준 값보다 낮으면 이는 화제가 변경되었다고 판단하였다. 표 3은 언어 병리학 연구와 본 연구 제안 방법의 결과 비교를 위한 표이다. 이는 각 집단에서 화제 변경 비율 결과들의 평균값을 보여준다.

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

[3] http://nlp.kookmin.ac.kr/

[4] https://www.kaggle.com/quora/question-pairs-dataset

[5] http://aiopen.etri.re.kr/service_dataset.php?category=language

표 3. 제안 방법과 언어 병리학의 결과 비교

학년	언어 병리학 [5]	제안 방법
초등학교 1	0.1476	0.399
초등학교 3	0.1088	0.3197
초등학교 5	0.0447	0.2028
상관관계	0.9995	

표 3에서 보이는 바와 같이, 제안 방법은 언어 병리학의 결과인 [5]처럼 1학년에 비해 3학년이 낮으며, 3학년에 비해 5학년이 낮다는 결과를 얻었다. 또한 제안 방법과 기존의 연구결과와의 상관관계가 0.9995로 순서뿐만 아니라 값의 간격조차 매우 유사함을 확인할 수 있었다.

6. 결론

화제 관리 능력은 원활한 대화뿐만 아니라 관계 형성에 매우 중요한 역할을 한다. 하지만 화제 관리 능력에 대한 분석에는 많은 시간과 비용이 들어간다. 이러한 시간과 비용을 줄이기 위해 화제 관리 능력 연구 중 하나인 화제 변경 탐지를 Sent2Vec을 이용하여 자동화하였다. 이를 이용한 방법은 기존 연구와 매우 유사한 결과를 얻을 수 있었다.

언어 능력 발달에 대한 연구는 화제 변경 탐지 이외에도 여러 연구들이 존재한다. 이러한 언어 능력 발달 연구의 자동화를 더욱 효과적으로 진행하기 위해서는 많은 데이터가 필요하다. 하지만 연령별 대화 데이터를 수집하는 것은 많은 시간과 비용이 들기 때문에 많은 어려움이 있다.

이를 위해, 향후 연구에서는 대화 데이터를 대체할 수 있는 방법을 모색할 것이며, 이러한 데이터들을 사용하여 효과적인 언어 분석 자동화를 위한 많은 연구들을 진행할 것이다.

참고문헌

[1] 이재성. "한국어 형태소 분석을 위한 3 단계 확률 모델." 정보과학회논문지: 소프트웨어 및 응용 38.5 (2011): 257-268.

[2] 이창기, 김준석, 김정희. "딥 러닝을 이용한 한국어 의존 구문 분석." 제 26 회 한글 및 한국어 정보처리 학술대회 (2014): 87-91.

[3] 유홍연, 고영중. "Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장." 정보과학회논문지 44.3 (2017): 306-313.

[4] 허현숙, 이윤경. "언어학습부진아동의 대화차례 주고받기 및 주제운용 특성." Communication Sciences & Disorders 17 (2012): 66-78.

[5] 박윤정, 최지은, 이윤경. "초등학생 아동의 대화 화제관리 능력의 발달." Communication Sciences & Disorders 22 (2017): 25-34.

[6] Pagliardini, Matteo, Prakhar Gupta, and Martin Jaggi. "Unsupervised learning of sentence embeddings using compositional n-gram features." arXiv preprint arXiv:1703.02507(2017).

[7] 박상길, 신명철. "Sent2Vec 문장 임베딩을 통한 한국어 유사 문장 판별 구현." Proc. of the HCLT(Human & Cognitive Language Technology), pp.541-545, 2018. (in Korean)

[8] Brinton, Bonnie, and Martin Fujiki. "Development of topic manipulation skills in discourse." Journal of Speech, Language, and Hearing Research 27.3 (1984): 350-358.

[9] Arnold, Sebastian, et al. "SECTOR: A Neural Model for Coherent Topic Segmentation and Classification." Transactions of the Association for Computational Linguistics 7 (2019): 169-184.

[10] Koshorek, Omri, et al. "Text segmentation as a supervised learning task." arXiv preprint arXiv:1803.09337 (2018).

[11] Achananuparp, Palakorn, Xiaohua Hu, and Xiajiong Shen. "The evaluation of sentence similarity measures." International Conference on data warehousing and knowledge discovery. Springer, Berlin, Heidelberg, 2008.