

심층신경망을 활용한 제어가능 말뭉치 시각화 기법

정지수[○], 정상근*

충남대학교, 컴퓨터융합학부
jisu.jung5@gmail.com, hugman@cnu.ac.kr

Constrained Corpus visualization using neural network

Jeesu Jung[○], Sangkeun Jung*

Department of Computer Science & Engineering, Chungnam National University

요약

말뭉치를 구성하고 있는 문장들 사이의 관계가 반영된 시각화는 말뭉치 전체의 구조나 유사의미 문장군의 분포 등을 파악하는데 매우 유용하게 활용될 수 있다. 본 연구에서는, 유사한 의미를 가지는 문장들은 서로 가까이 분포하도록 시각화되어야 한다는 제어조건을 사용자가 제공했을 때, 해당 조건이 만족되도록 2차원 공간에 말뭉치의 각 문장을 시각화하는 기법을 소개한다.

주제어: 말뭉치 시각화, 차원 축소, 제어정보

1. 서론

훈련에 사용되는 정답 말뭉치, 더 나아가 상용시스템을 통해 수집된 실사용예제가 포함된 로그(Log) 말뭉치 등의 시각화는 자연어처리 연구 및 상용시스템의 유지보수에도 큰 도움을 줄 수 있다. 특히, 문장의 의미정보나 문장과 다른 문장들간의 관계가 반영된 2차원 혹은 3차원의 시각정보는 말뭉치 내의 의미분포를 빠르게 파악하는데 도움을 줄 수 있다.

딥러닝 기반 자연어처리에서는 보통 UMAP[1] 이나 T-SNE[2] 등을 활용한 말뭉치 시각화가 이루어진다. 본 연구에서는, 문장과 문장 혹은 문장들과 문장들 간의 관계에 대한 제어정보를 사용자가 제공하게 되면, 이러한 정보를 최대한 반영하여 말뭉치를 시각화 하는 방법을 새롭게 제안한다.

이러한 제어가능한 말뭉치 시각화를 구현하기 위해서는 아래와 같은 3가지 문제가 해결되어야 한다.

- **제어정보정의:** 유의미한 시각정보를 구축하기 위해 필요한 제어정보의 종류와 정보형태를 정의해야 한다.
- **문장좌표생성:** 각 문장들의 의미정보가 반영된 저차원 문장좌표를 생성해야 한다.
- **문장좌표이동:** 주어진 제어정보에 해당하는 문장들은 조건에 부합되도록 이동시키되, 정보가 주어지지 않은 문장들은 변동이 없도록 좌표벡터를 고정해야 한다.

제어정보의 종류는 대단히 다양할 수 있으며, 구체적인 시각화는 목적에 따라 달라 질 수 있다. 본 연구에서는 의미가 유사한 문장은 서로 가까이 있어야 한다는 긍정 관계의 문장에 대한 연결정보만을 제어정보로 다루기

로 한다.

문장좌표생성 문제는 전형적인 문장 임베딩(Sentence embedding) 문제의 하나로 접근할 수 있다. 본 연구에서는 선학습된 BERT 모델[3]을 활용하여 문장을 고차원(768차원) 문장벡터로 바꾼 후 2차원으로 프로젝션하는 별도의 신경망층을 이용하여 문장좌표벡터를 생성하였다.

문장좌표이동은 주어진 제어정보에 해당하는 문장들끼리는 그 조건에 부합되게 이동하고, 그렇지 않은 문장들은 그 자리에 계속 유지하도록 고안되어야 한다. 본 연구에서는 긍정관계에 있는 문장들 사이는 거리가 작아지도록 하는 Contrastive loss 값을 활용하고, 정보가 없는 문장들은 최초의 좌표에 고정되도록 하는 회귀분석(regression) 과 유사한 Contrastive loss 값을 활용하여 문제를 해결하였다.

제안하는 방법론의 타당성 검증을 위해 Weather 말뭉치에 대해 제어정보별 시각화 방법을 검증하였다.

본 논문의 구조는 다음과 같다. 2장에서는 관련연구를 기술하였고, 3장에서는 본 연구의 시각화방법에 대해 자세히 서술하였다. 4장에서는 실험 조건과 결과를 확인할 수 있고, 5장에서는 본 연구 결과를 요약하고 향후 연구 방향에 대해 논하도록 한다.

2. 관련 연구

딥러닝 기반 자연어처리에서의 일반적인 말뭉치 시각화 도구는 UMAP[1]과 T-SNE[2]이다. 이 둘은 모두 통계적인 기반으로 가진 도구이다. UMAP의 경우에는 리만 기하학의 조합과 퍼지 단순화 집합의 기하학적 실현에 대한 범주 이론적 접근방식을 응용한다. T-SNE의 경우에는 라플라시안 고유맵을 기반으로, Stochastic Neighbor Embedding을 이용하여 시각화를 진행한다. 이 외에도 다양한 말뭉치 시각화 방식이 존재하나, 대체로 통계적인

* 교신 저자(Corresponding Author)

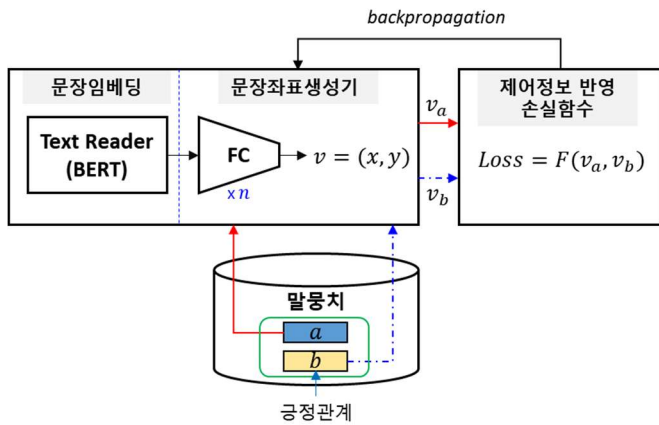


그림 1. 전체 구조

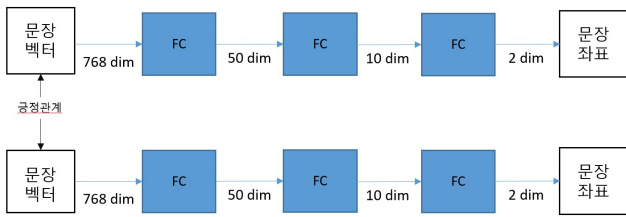


그림 2. 문장 좌표 생성의 구조

방식을 이용한다.

네트워크 구조를 사용하여 벡터 사이 거리 조정을 이용하여 딥러닝 모델을 훈련하는 Metric Learning[4] 방식이 있었다. Metric learning은 벡터 사이의 관계를 이용하여 유사도가 높은 쪽을 더 가까운 벡터로 나타내도록 수식을 유도하며 훈련하는 딥러닝 기법이다.

이 기법에서는 손실함수가 매우 중요하게 여겨지는데, 벡터 간의 관계를 수치화한 뒤, 손실함수로 사용하여 벡터의 위치를 조정하기 때문이다. 가장 간단한 손실 함수는 Euclidean distance이다. 벡터 간 거리에 대한 값을 손실로 받는 것이다. 이 이외에도 두 벡터 간 관계에 의해 긍정 관계와 부정 관계에 대해 서로 다른 손실의 값을 적용하는 contrastive loss, 세 벡터 간의 거리를 비교하여 더 가까워야 하는 벡터와 멀어져야 하는 벡터에 대한 손실값을 주는 triplet loss 등이 있다.

3. 구조

본 연구의 실험을 위한 전체구조는 그림 1과 같다. 말뭉치에 포함된 문장들에 대한 문장 임베딩을 수행한 후 문장좌표생성기를 통해 2차원으로 투사시키는 과정을 거친다. 이렇게 생성된 각 문장좌표들에 대해서, 외부에서 주어지는 제어정보에 맞게 손실함수가 계산되고, 그 손실함수를 최소화하는 방향으로 문장좌표생성기의 파라미터가 학습되게 된다. 이때, 역전파(backpropagation) 과정은 fully connection 층으로 구성된 문자좌표생성기까지만 이루어지게 되기 때문에 매우 빠른 학습이 가능해

진다.

3.1. 제어 정보 정의

본 연구에서는 긍정 조건의 문장들의 관계에 대해서만 연구를 진행하였다. 문장끼리의 긍정 조건의 관계는 문장의 주제와 문장 내 의미 태그 구성이 모두 같아 의미적으로 두 문장 사이의 유사성이 높다고 판단되는 관계로 설정하였다.

이렇게 설정된 제어 정보를 활용하기 위해 긍정 관계의 문장들의 집합에 대한 사건을 생성하고, 이를 이용하여 문장을 긍정 관계 문장 쌍을 생성할 수 있도록 한다.

사용한 말뭉치에서는 문장별 의미틀(semantic frame)이 구성되어 있다. 의미틀의 구성은 문장의 주제(intent), 문장의 단어의 쓰임의 카테고리과 같은 의미

태그(slot tag), 태그된 단어(slot value)로 되어있다. 이러한 구성 중, 문장의 주제(intent)와 문장의 의미 태그(slot tag)가 완전히 같은 것만을 긍정 관계로 판단하고 이러한 문장의 인덱스를 긍정관계 문장사전에 저장하여 활용하였다.

3.2. 문장 좌표 생성

선학습된 BERT 트랜스포머 모델은 출력으로 768차원의 임베딩 벡터를 준다. 이러한 출력된 벡터를 저차원으로 감소시키기 위해서는 문장벡터들을 긍정 관계의 문장 집합 사건을 참고하여 문장쌍을 생성하고, 이를 신경망 구조에 넣고 레이어 간의 선형 결합을 순차적으로 하여 2차원까지 차원을 낮추도록 하였다. 이에 대한 구체적인 구조는 그림 2와 같다.

3.3. 문장 좌표 이동

선학습 BERT모델의 768차원 벡터를 2차원으로 줄일 때, 손실함수를 이용하여 역전파를 통해 제어정보에 맞는 좌표로 이동하도록 한다. 긍정 관계가 있는 문장들은 서로 가까워지고, 그 외의 구체적 제어정보가 없는 문장은 제자리에 고정되어야 한다.

손실함수로 contrastive loss를 이용하여 contrastive loss 값이란, Metric learning에서 사용되는 대표적인 손실 함수로, 두 벡터의 거리를 D_w 라 하면, 문장 a에 대한 문장벡터를 v_a^t , 문장 b에 대한 문장벡터를 v_b^t 에 대한 거리 공식은 다음과 같다.

$$D_w(v_a^t, v_b^t) = \|v_a^t - v_b^t\|$$

이때, 두 벡터 사이의 관계를 Y라 가정한다. 긍정 관계의 벡터에게는 $Y = 1$ 이고, 부정 관계일 시에는 $Y = 0$ 이다. 공식은 다음과 같다.

$$Loss_{contrastive} = (1 - Y) * D_w^2 + Y * \max(0, D_w)^2$$

epoch t에서의 문장 a에 대한 문장벡터를 v_a^t , 문장 b에 대한 문장벡터를 v_b^t 로 나타낼 때, 긍정관계에 있는 손실함수는 아래와 같이 정의할 수 있다.

$$Loss_{pos} = Loss_{contrastive}(v_a^t, v_a^t)$$

제어정보가 주어지지 않은 문장 m 에 대해서는, 바로 전에 학습을 진행했던 epoch 에서의 좌표와 같은 좌표에 계속 머물러 있어야 하므로, 이를 반영하는 손실함수를 아래와 같이 정의할 수 있다.

$$Loss_{fix} = contrastive_loss(v_m^{t-1}, v_m^t)$$

전체 손실함수는 위의 두 손실함수를 합한 값으로 정의된다.

$$Loss = Loss_{fix} + Loss_{pos}$$

이렇게 계산된 손실값을 이용하여 역전파를 반복하면 긍정 관계의 문장벡터의 좌표들끼리는 가까워지고, 긍정 관계가 아닌 문장벡터들의 좌표는 그 자리를 유지하게 된다

4. 실험

실험 데이터로 Weather 데이터셋과 Atis 데이터셋[5] 을 이용하였다. Weather 데이터셋은 의도 분류를 위해 수집된 한국어 문장 데이터셋이다. 문장은 날씨에 관한 문장들로 이루어져 있으며, 문장의 주제, 의미태그, 의미태그된 단어에 대한 정보가 추가되어 있다. 총 6993문장이 있으며, 19개의 문장 주제와 14개의 의미태그로 구성되어 있다. Atis 데이터셋은 의도 분류를 위한 영문 데이터셋이다. 총 4475개의 문장으로 이루어져 있으며, Weather 데이터셋과 같이 문장의 주제, 의미태그, 태그된 단어에 대한 정보가 포함되어 있다.

각 데이터셋의 문장들을 입력으로 했을 때 선학습된 BERT based uncased 트랜스포머 모델[3]에서 출력된 768 차원 임베딩 벡터 결과물을 이용하였다. 문장 벡터들을 긍정 조건의 문장벡터집합으로 묶고 이 집합에 속하는 기준점을 선정하여, 이를 이용하여 긍정 관계의 문장을 한 쌍씩 묶는다.

본 논문에서는 문장들의 위치 변화를 쉽게 파악하고자, 전체 6993문장 중 특정 문장의 문장 주제와 의미태그가 같은 문장들의 집합에 대해 서로 문장쌍을 생성하여 실험을 진행하였다.

그림 4의 (a)의 경우, 문장 주제는 weather, general, 의미태그 day.p의 긍정 관계의 문장을 서로 쌍으로 묶어 총 56쌍으로 집합을 구성하였다. 나머지 문장은 자기 자신과 쌍을 지어 총 6637쌍을 생성하였다. 이들은 자기 자신만이 같은 문장이 되도록 오직 자신과 같은 값을 가진 벡터와 한 쌍으로 묶는 것이다. 그림 4의 (a)은 전체 문장 중, 선정한 문장 쌍 집합의 위치 변화를 시각화한 것이다. 그림 3는 선정한 문장 쌍 집합이 어느정도 가까워지는지 확인하기 위해, 해당 집합의 문장 쌍 위치의 거리변화를 시각화한 것이다.

같은 방법으로 문장 주제 weather.uv, 의미태그는 day.p를 가지는 77개의 문장쌍에 대해서도 실험한 그래

프가 그림 4의 (b)이다. 문장 주제 weather, temperature, 의미태그 location을 가지는 53개의 문장쌍에 대한 실험 그래프는 그림 4의 (c)이다.

자기 자신만이 한 쌍인 벡터들은 모두 고정된 채로 같은 문장으로 취급한 문장 벡터들만 거리가 가까워지는 것을 볼 수 있다. 임베딩된 문장 벡터와 이들의 관계 이외의 것을 제공하지 않았음에도 의미가 같은 문장끼리 거리가 가까워지는 것이 관찰된다.

이처럼 Atis 데이터셋을 이용하여 Weather 데이터셋과 같은 방식으로 주제와 의미태그가 같은 문장들에 대해 실험을 진행하였다. 그림 4 (d)의 경우 문장 주제는 atis.abbreviation, 문장의 의미태그 fare_basis_code로 구성된 문장들에 대한 45개의 긍정 문장쌍에 대한 관계 변화이다. 그림 4 (e)의 경우는 주제 atis.airfare, 의미태그 fromloc.city_name, round_trip, toloc.city_name의 정보를 가진 33개의 긍정 문장쌍에 대해 실험한 그래프이다. 문장 주제 atis.flight, 문장의 의미태그 fromloc.city_name, toloc.city_name에 대한 671개의 문장쌍에 대한 실험 그래프는 그림 4 (f)이다.

두 개의 데이터셋을 이용한 실험에 대해, 에폭에 따른 긍정 문장 벡터집합의 거리 변화는 다음의 그림 3과 같다. 에폭이 증가함에 따라 거리 변화는 감소하는 것으로 볼 수 있다.

5. 결론

실험을 통해 신경망 구조를 이용한 긍정 조건의 문장들의 그래프를 관찰했을 때, 좌표 간 거리가 확연히 줄어들음을 확인할 수 있다. 에폭에 따른 그래프의 변화를 관찰하면 문장의 제약 조건에 의해 긍정 관계의 문장 집합은 거리가 줄어들고, 이외의 점들은 거리가 변하지 않는다는 것을 볼 수 있다. 이를 통해 제약조건에 의한 말뭉치 시각화가 가능하다고 판단된다. 다만 본 연구는 긍정 조건에 대해서만 반응하도록 설계되어, 부정 조건에 대해서는 추후 연구가 진행되어야 한다.

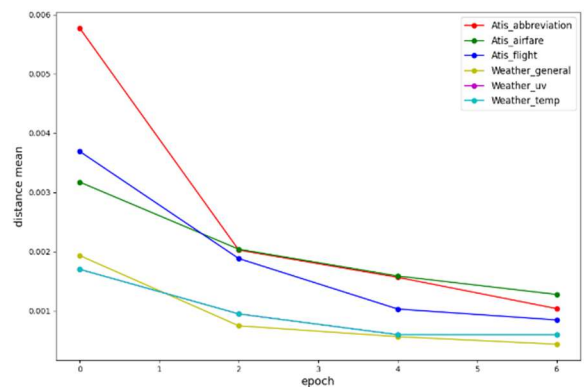
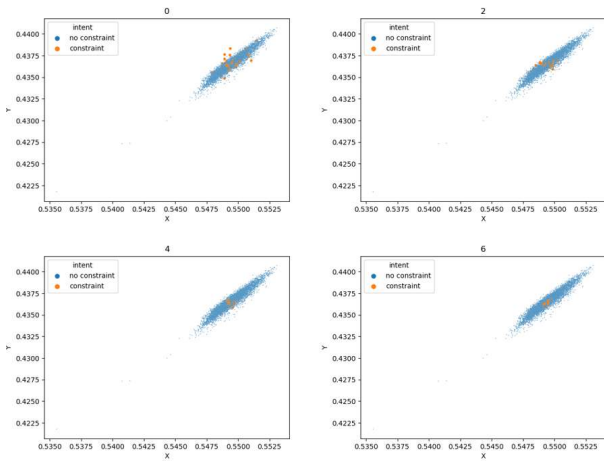
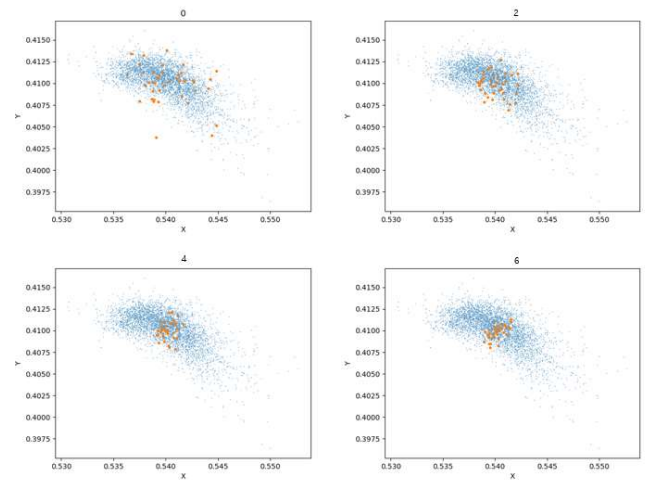


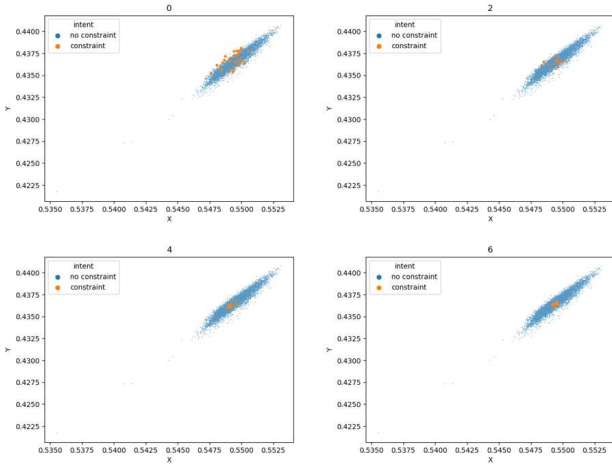
그림 3. 각 실험에서의 epoch에 따른 거리 변화 비교



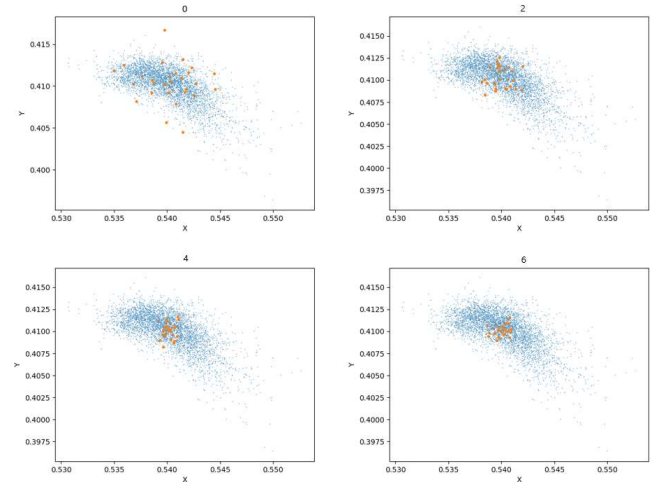
(a) weather.general(day.p=*)



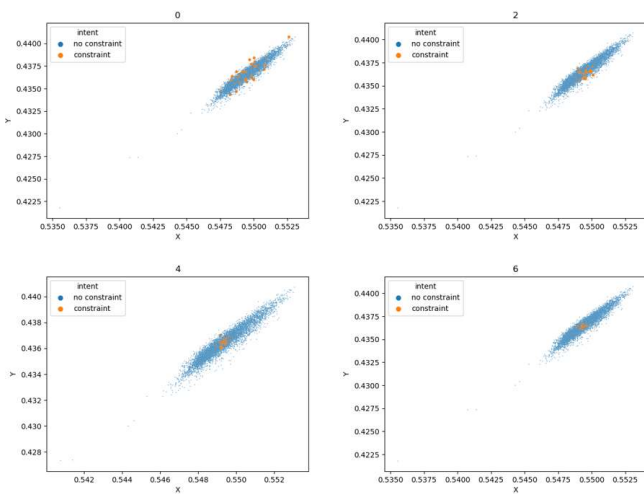
(d) atis.abbreviation(fare_basis_code=*)



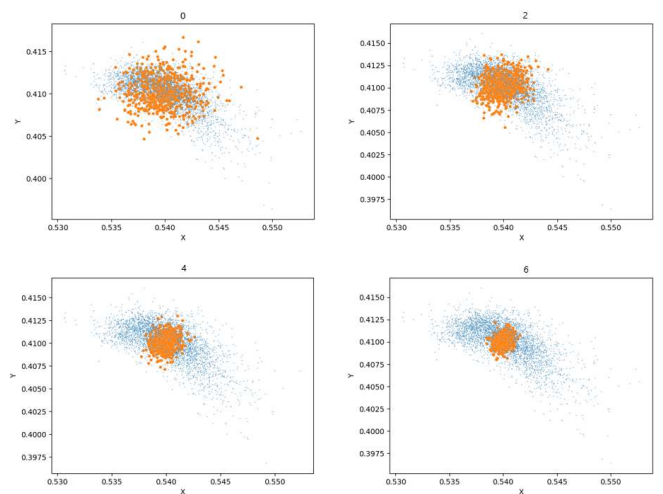
(b) weather.uv(day.p=*)



(e) atis.airfare(fromloc.city_name=*, round_trip=*, toloc.city_name=*)



(c) weather.temperature(location=*)



(f) atis.flight(fromloc.city_name=*, toloc.city_name=*)

그림 4. 훈련 진행 과정별 제어 시각화 결과.
 각 그림은 해당 intent_tag(slot_tag=*, slot_tag=*, ...)에 대한 결과.

감사의 글

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2020-0-01441)

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

이 논문은 2019 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2019R1F1A1060601)

참고문헌

- [1] MCINNES, Leland; HEALY, John; MELVILLE, James. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.
- [2] VAN DER MAATEN, Laurens; HINTON, Geoffrey. Visualizing data using t-SNE. journal of Machine Learning Research 9. Nov (2008), 2008.
- [3] DEVLIN, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [4] KAYA, Mahmut; BILGE, Hasan Şakir. Deep metric learning: A survey. Symmetry, 2019, 11.9: 1066.
- [5] PRICE, Patti. Evaluation of spoken language systems: The ATIS domain. In: Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990. 1990.