(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년03월10일
(11) 등록번호 10-1715118
(24) 등록일자 2017년03월06일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06F 17/27* (2006.01)
G06N 3/04 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06F 17/2705 (2013.01)
(21) 출원번호 10-2016-0140376
(22) 출원일자 2016년10월26일
심사청구일자 2016년10월26일
(56) 선행기술조사문헌
Li, Jiwei, Minh-Thang Luong, and Dan Jurafsky. "A hierarchical neural autoencoder for paragraphs and documents." arXiv preprint arXiv:1506.01057. 2015.
Li, Jiwei. "Feature weight tuning for recursive neural networks." arXiv preprint arXiv:1412.3714. 2014.
Tang, Duyu, et al. "Effective LSTMs for Target-Dependent Sentiment Classification." arXiv preprint arXiv:1512.01100v2. 2016.9.29.
KR1020130103249 A

(73) 특허권자
가천대학교 산학협력단
경기도 성남시 수정구 성남대로 1342 (복정동)
(72) 발명자
노웅기
경기도 성남시 분당구 구미로144번길 7, 802동 402호 (구미동, 무지개마을제일아파트)
강상우
서울특별시 성북구 길음로 33, 802동 901호 (길음동, 길음뉴타운)
(74) 대리인
이은철, 이수찬

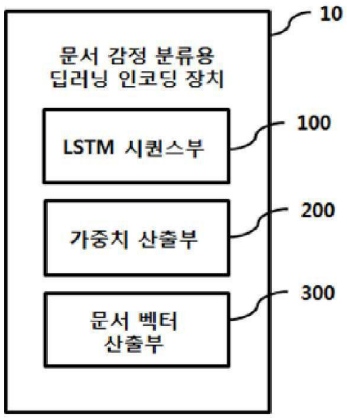
전체 청구항 수 : 총 4 항 심사관 : 서광훈

(54) 발명의 명칭 **문서 감정 분류용 딥러닝 인코딩 장치 및 방법.**

(57) 요 약

본 기술은 문서 감정 분류용 딥러닝 인코딩 장치 및 방법에 관한 것이다. 본 기술의 구현 예에 따르면, LSTM(Long Short Term Memory) 들로 구성되는 LSTM 시퀀스부, LSTM에 저장된 정보에 대하여 중요도에 따른 가중치를 산출하는 가중치 산출부, 가중치 산출부로부터 산출된 가중치와 LSTM의 출력 값에 따라 기계 학습 하여 문서 벡터를 산출하는 문서 벡터 산출부를 포함하여 길이가 긴 문서에서도 높은 정확도를 유지하고, 구, 절 혹은 이모티콘이 갖는 감정을 효과적으로 판단할 수 있어 더욱 효율적인 문서의 자동 감정 분류를 가능하게 할 수 있는 이점이 있다.

대 표 도



(52) CPC특허분류
G06N 3/04 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 R7015-16-1003
부처명 미래창조과학부
연구관리전문기관 정보통신기술진흥센터
연구사업명 SW전문인력역량강화
연구과제명 2016년 SW중심대학(SW특성화대학원)_가천대
기 여 율 1/1
주관기관 가천대학교 산학협력단
연구기간 2016.03.01 ~ 2017.02.28
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

감정 분류할 문서를 이루는 단어 및 알파벳이 아닌 기호를 입력 받아 처리되는 정보를 선택적으로 갱신 및 저장하는 딥러닝 구조인 LSTM(Long Short Term Memory) 들로 구성되는 LSTM 시퀀스부;

상기 LSTM에 저장된 정보에 대하여 중요도를 산출하고, 산출된 상기 중요도로부터 가중치를 산출하는 가중치 산출부; 및

상기 가중치 산출부로부터 산출된 가중치와 상기 LSTM의 출력 값에 따라 기계 학습 하여 문서 벡터를 산출하는 문서 벡터 산출부;를 포함하고,

상기 가중치 산출부는,

상기 LSTM을 소정 크기의 윈도우 단위로 결합하여 중요도 및 가중치를 산출하는 것이고,

상기 문서 벡터 산출부는,

상기 LSTM을 소정 크기의 윈도우 단위로 결합한 출력 값으로부터 문서 벡터를 산출하는 것을 특징으로 하는 문서 감정 분류용 딥러닝 인코딩 장치.

청구항 2

제1항에 있어서,

상기 가중치 산출부는,

학습 데이터에 대한 과적합을 방지하기 위하여 상기 가중치를 소정의 문서길이에 따른 적응적 평활 값에 따라 재조정하여 산출하는 것을 특징으로 하는 문서 감정 분류용 딥러닝 인코딩 장치.

청구항 3

LSTM 시퀀스의 LSTM들에 감정 분류할 문서를 이루는 단어 및 알파벳이 아닌 기호를 입력하고, 이에 따라 처리되는 정보를 선택적으로 갱신 및 저장하도록 단계;

상기 LSTM에 저장된 정보에 대하여 중요도를 산출하고, 산출된 상기 중요도로부터 가중치를 산출하는 단계 및

산출된 상기 가중치와 상기 LSTM의 출력 값에 따라 기계 학습 하여 문서 벡터를 산출하는 단계;를 포함하고,

상기 가중치를 산출하는 단계는,

상기 LSTM을 소정 크기의 윈도우 단위로 결합하여 중요도 및 가중치를 산출하는 것이고,

상기 문서 벡터를 산출하는 단계는,

상기 LSTM을 소정 크기의 윈도우 단위로 결합하여 산출된 출력 값으로부터 문서 벡터를 산출하는 것을 특징으로 하는 문서 감정 분류용 딥러닝 인코딩 방법.

청구항 4

제3항에 있어서,

상기 가중치를 산출하는 단계는,

학습 데이터에 대한 과적합을 방지하기 위하여 상기 가중치를 소정의 문서길이에 따른 적응적 평활 값에 따라 재조정하여 산출하는 것을 특징으로 하는 문서 감정 분류용 딥러닝 인코딩 방법.

발명의 설명

기 술 분 야

[0001] 문서 감정 분류용 딥러닝 인코딩 장치 및 방법에 관한 것으로, 더욱 상세하게는 감정 분류 대상인 문서의 길이가 길어지거나 문장 구성 요소로 이모티콘, 한 단어 이상으로 구성되는 구와 절을 포함하더라도 성능이 떨어지지 않도록 하는 문서 감정 분류용 딥러닝 인코딩 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 감정 분류(Sentiment Classification)는 문서에서 드러난 필자의 긍정 도는 부정적 감정을 분석하여, 이를 자동으로 분류하는 연구이다.

[0003] 기존의 감정 분류 연구에는 감정 사전, 워드넷 등의 영역 의존 지식을 사용하거나, 의존 구문구조 같은 통사적 정보를 사용하였다. 근래에는 감정 분류에 딥러닝 인코더를 적용하려는 연구가 활발히 진행되고 있다.

[0004] 딥러닝 인코더는 딥러닝 기술을 사용해서 가변 길이 문서를 고정 길이 문서 벡터로 표현하는 방법으로, 감정 분류 분야에서 우수한 성능을 보여줄 수 있다. 하지만 전체 문서 시퀀스의 마지막 출력을 문서 벡터로 간주하는 LSTM(Long Short Term Momory) 인코딩 장치의 경우, 입력이 길어짐에 따라 초기에 입력된 패턴의 인식률이 급격히 저하되어, 긴 문서의 인코딩 장치로는 적합하지 않은 문제점이 있다.

[0005] 이러한 문제점을 해결하기 위해 LSTM Attention 인코딩 장치를 사용하여 문서 벡터를 생성하는 방법을 제안되었다. LSTM Attention 인코딩 장치는 LSTM의 시간 별 출력 결과를 그 중요도에 따라 가중치를 부여하고, 가중치와 출력 결과의 가중치 합으로 문서 벡터를 생성하는 모델이다.

[0006] 문서에서 감정을 나타내는 구성 요소는 한 단어 인 경우도 있으나, 많은 문서에서 감정을 표현하는 문장 구성 요소는 이모티콘, 구, 절과 같이 한 단어 이상으로 구성되는 실정이다. 그러나 LSTM Attention 인코딩 장치 등 종래의 딥러딩 인코딩 장치들은 문서 내 단어의 가중치를 계산하는 모델로서 한 단어가 갖는 감정을 판별하기에는 좋으나, 구, 절 혹은 이모티콘이 갖는 감정을 효과적으로 판단하는데 있어서 여전히 적절하지 못한 문제점이 있다.

[0007] 도 1은 종래 기술에 따른 LSTM 인코딩 장치의 LSTM 인코더 구조에 대한 개념도를 도시한다.

[0008] 도 1을 참조하면, LSTM 인코더 구조를 모델로 하는 LSTM 인코딩 장치는 문서의 길이가 n인 경우 전체 LSTM 시퀀스의 마지막 LSTM의 출력 값 h_n을 문서 벡터로 사용하게 됨에 따라 초기에 입력된 패턴의 인식률이 급격히 저하되어 긴 문서의 인코딩 장치로는 적합하지 않다.

[0009] 도 2은 LSTM Attention 인코딩 장치의 LSTM Attention 인코더 구조에 대한 개념도를 도시한다. 도 2를 참조하면 “I hate this.”라는 문서가 입력될 경우, ‘hate’의 가중치는 ‘I’, ‘this’, ‘.’ 보다 큰 값을 갖게된다. 결과적으로, ‘hate’가 다른 단어들보다 강조되어, 입력 문서가 부정적인 문서로 분류된다. 이러한 구조는 한 단어가 갖는 감정을 판별 하기에는 효과적이나, 이모티콘을 빈번하게 사용하는 인터넷 언어 문화에 있어서 이모티콘이 가지는 감정 판별에 효과적이지 못하며, 구와 절의 감정을 판별하는데 있어서도 효과적이지 못하다.

선행기술문헌

특허문헌

[0010] (특허문헌 0001) 1. 한국등록특허 제10-1465756호 (감정 분석 장치 및 방법과 이를 이용한 영화 추천 방법)

비특허문헌

[0011] (비특허문헌 0001) 1.Wollmer, Martin, et al. "Youtube movie reviews:

(비특허문헌 0002) Sentiment analysis in an audio-visual context." Intelligent

(비특허문헌 0003) Systems, IEEE 28.3, pp.46-53, 2013

(비특허문헌 0004) 2. Graves, Alex, Greg Wayne, and Ivo Danihelka. "Neural

(비특허문헌 0005) turing machines." arXiv preprint arXiv:1410.5401, 2014.

발명의 내용

해결하려는 과제

[0012] 본 발명은 상기와 같은 문제를 해결하기 위한 것으로서, 길이가 긴 문서에서도 높은 정확도를 유지하고, 구, 절 혹은 이모티콘이 갖는 감정을 효과적으로 판단할 수 있도록 하는 문서의 감정 분류용 딥러닝 인코딩 장치 및 방법을 제공하는데 그 목적이 있다.

과제의 해결 수단

[0014] 상기와 같은 목적을 달성하기 위한 본 발명은 감정 분류할 문서의 단어 또는 문자가 아닌 기호를 정보로 저장하고, 선택적으로 갱신하는 LSTM(Long Short Term Memory) 들로 구성되는 LSTM 시퀀스부; 상기 LSTM에 저장된 정보에 대하여 중요도를 산출하고, 산출된 상기 중요도로부터 가중치를 산출하는 가중치 산출부; 및 상기 가중치 산출부로부터 산출된 가중치와 상기 LSTM의 출력 값에 따라 기계 학습 하여 문서 벡터를 산출하는 문서 벡터 산출부;를 포함하고, 상기 가중치 산출부는, 상기 LSTM을 소정 크기의 윈도우 단위로 결합하여 중요도 및 가중치를 산출하는 것이고, 상기 문서 벡터 산출부는, 상기 LSTM을 소정 크기의 윈도우 단위로 결합한 출력 값으로부터 문서 벡터를 산출하는 것을 특징으로 한다.

[0015] 바람직하게는, 상기 가중치 산출부는, 학습 데이터에 대한 과적합을 방지하기 위하여 상기 가중치를 소정의 문서길이에 따른 적응적 평활 값에 따라 재조정하여 산출하는 것을 특징으로 할 수 있다.

[0016] 또한 본 발명의 다른 실시예 따른 문서 감정 분류용 딥러닝 인코딩 방법은 LSTM 시퀀스의 LSTM들에 감정 분류할 문서의 단어 또는 문자가 아닌 기호를 정보로 저장하고, 선택적으로 갱신하는 단계; 상기 LSTM에 저장된 정보에 대하여 중요도를 산출하고, 산출된 상기 중요도로부터 가중치를 산출하는 단계 및 산출된 상기 가중치와 상기 LSTM의 출력 값에 따라 기계 학습 하여 문서 벡터를 산출하는 단계;를 포함하고, 상기 가중치를 산출하는 단계는, 상기 LSTM을 소정 크기의 윈도우 단위로 결합하여 중요도 및 가중치를 산출하는 것이고, 상기 문서 벡터를 산출하는 단계는, 상기 LSTM을 소정 크기의 윈도우 단위로 결합하여 산출된 출력 값으로부터 문서 벡터를 산출하는 것을 특징으로 한다.

[0017] 바람직하게는, 상기 가중치를 산출하는 단계는, 학습 데이터에 대한 과적합을 방지하기 위하여 상기 가중치를 소정의 문서길이에 따른 적응적 평활 값에 따라 재조정하여 산출하는 것일 수 있다.

발명의 효과

[0018] 전술한 바와 같은 본 발명에 따르면, 평균 문서 길이가 몇 백 단어에 달하는 길이의 문서에서도 높은 정확도를 가지며 이모티콘, 구, 절과 같이 한 단어 이상으로 구성되는 문장 구성 요소의 감정을 효과적으로 판단하여 더욱 효율적인 문서의 자동 감정 분류를 가능하게 할 수 있는 이점이 있다.

도면의 간단한 설명

[0019] 도 1은 LSTM 인코더 구조에 대한 개념도이다.

도 2는 LSTM Attention 인코더 구조에 대한 개념도이다.

도 3은 본 발명의 일 실시예에 따른 LSTM Window Attention 인코더 구조에 대한 개념도이다.

도 4는 본 발명에 따른 문서 감정 분류용 딥러닝 인코딩 장치의 블록도이다.

도 5는 본 발명에 따른 검증 데이터에서 문서길이에 따른 적응적 평활 값의 성능을 도시하는 그래프이다.

도 6은 본 발명의 다른 실시예에 따른 문서 감정 분류용 딥러닝 인코딩 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0020] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 이에 앞서 본 발명에 관련된 공지 기능 및 그 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 구체적인 설명을 생략하였음에 유의해야 할 것이다.

[0021] 본 발명의 실시예에 따른 딥러닝 인코딩 장치는 LSTM Window Attention 인코더 구조를 이용한다.

[0022] 도 3은 본 발명의 실시예에 따른 LSTM Window Attention 인코딩 장치의 LSTM Window Attention 인코더 구조에 대한 개념도를 도시한다.

[0023] 도 4는 본 발명의 실시예에 따른 문서 감정 분류용 딥러닝 인코딩 장치의 블록도를 도시한다.

[0024] 도 4를 참조하면, 본 발명에 따른 딥러닝 인코딩 장치는 LSTM 시퀀스부(100), 가중치 산출부(200), 문서 벡터 산출부(300)를 포함한다.

[0025] LSTM 시퀀스부(100)는 감정 분류할 문서를 이루는 단어 및 문자가 아닌 기호를 입력 받아 선택적으로 갱신 및 저장하는 딥러닝 구조인 LSTM(Long Short Term Memory)들로 구성된다.

[0026] LSTM은 정보를 저장하는 셀 상태(Cell state)의 선택적인 갱신으로 RNN(Recurrent Neural Network)의 장기 의존성을 회피하는 딥러닝 구조로서 아래의 수학식 1과 같이 동작한다.

[0027]

[0028] 수학식 1

[0029]

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\overline{C_t} &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
C_t &= f_t * C_{t-1} + i_t * \overline{C_t} \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t * \tanh(C_t)
\end{aligned}
$$

[0030]

[0032] 가중치 산출부(200)는 LSTM에 저장된 정보에 대하여 중요도를 산출하고, 산출된 상기 중요도로부터 가중치를 산출한다.

[0033] 각 LSTM의 정보에 대한 중요도는 아래 수학식 2의 함수 f에 따라 산출될 수 있다. 또한 함수 f는 i번째 LSTM에 저장된 정보에 대한 출력 값인 h_i의 중요도를 FFNN(Feed Forward Neural Network)를 이용하여 산출할 수 있다.

[0034]

[0035] 수학식 2

[0037]

$$
f(h_i) = W_2 \tanh(W_1 h_i + b_1)
$$

[0039]

[0040] 가중치는 중요도 함수의 출력에 수학식 3과 같은 softmax 함수를 적용하여 산출될 수 있다. softmax 함수는 n개의 값이 존재할 때, 각각의 값의 편차를 확대시켜 큰 값은 상대적으로 더 크게, 작은 값은 상대적으로 더 작게 만든 다음 정규화 하는 함수이다. 즉, softmax 함수는 중요도를 확률 변수화 시킨다.

[0041] 또한, 가중치 산출부(200)는 LSTM을 소정 크기의 윈도우 단위로 결합하여 가중치를 산출한다.

[0042] 윈도우는 LSTM을 소정의 크기만큼 결합하는 단위이다. 도 3의 구조도를 참조하면 윈도우는 LSTM을 3개 단위로 결합할 수 있고, 각 윈도우 간에는 중첩되는 LSTM이 있을 수 있다.

[0043] 예를 들면 3개 크기의 윈도우 단위로 LSTM을 결합한 가중치 α는 아래의 수학식 3에 따라 산출될 수 있다. 실시예에서 함수 f는 인접한 3개의 LSTM의 출력 값이 결합된 $h_{i,i+1,i+2}$를 입력 받을 수 있다. 일 실시예에서 $\alpha_{i,i+1,i+2}$는 3개의 LSTM을 결합하는 윈도우의 가중치 값이다.

[0045] 수학식 3

$$\alpha_{i,i+1,i+2} = e^{f(h_{i,i+1,i+2})} \Big/ \sum_{j=2}^{n-1} e^{f(h_{j,j+1,j+2})}$$

[0047]

[0049]

[0051] 본 발명은 윈도우 단위로 LSTM을 결합하여 중요도와 가중치를 산출함에 따라 여러 단어 내지 여러 문자 이외의 기호로 구성되는 구, 절 혹은 이모티콘이 갖는 감정도 효과적으로 판단할 수 있다.

[0052] 또한, 가중치 산출부(200)는 학습 데이터에 대한 과적합을 방지하기 위하여 상기 가중치를 문서길이에 따른 적응적 평활(adaptive smoothing) 값에 따라 재조정하여 산출하는 것일 수 있다.

[0053] 윈도우 단위로 LSTM을 결합하여 가중치를 산출하는 경우에 대부분의 구문의 가중치가 0에 가까운 값을 가지는 문제가 발생할 수 있다. 본 발명에 따른 문서 감정 분류용 딥러닝 인코딩 장치는 문서 길이에 따른 적응적 평활 값에 따라 가중치를 재조정 하여 학습 데이터 과적합을 해결함으로써 문서 감정 분류의 정확도를 더욱 높일 수 있다. 예를 들면, 적응적 평활 값은 0.1/문서길이로 설정될 수 있으며 검증 데이터에서 가장 높은 결과를 보인 값을 소정의 값으로 선택할 수 있다.

[0054] 문서 벡터 산출부(300)는 가중치 산출부(200)로부터 산출된 가중치와 LSTM의 출력 값에 따라 기계 학습하여 문서 벡터를 산출한다.

[0055] 또한 문서 벡터 산출부(300)는 LSTM을 소정 크기의 윈도우 단위로 결합하여 산출된 출력 값으로부터 문서 벡터를 산출한다.

[0056] 예를 들면, 길이 n인 문서에 대한 문서 벡터 c는 LSTM을 3개 크기의 윈도우 단위로 결합한 출력 값인 $h_{i,i+1,i+2}$로부터 아래 수학식 4에 따라 산출될 수 있다.

[0057]

[0058] 수학식 4

$$c = \sum_{i=2}^{n-1} \alpha_{i,i+1,i+2} \times h_{i,i+1,i+2}$$

[0060]

[0063] 또한, 본 발명에 따른 문서 감정 분류 시스템은 상기와 같은 LSTM 시퀀스부(100), 가중치 산출부(200), 문서 벡터 산출부(300)를 포함하는 문서 감정 분류용 딥러닝 인코딩 장치(10)를 이용하며, 문서입력부를 통해 문서 감정 분류용 딥러닝 인코딩 장치(10)에 감정 분류 대상인 문서를 입력하여 산출된 문서 벡터를 이용하여 감정분류부가 문서의 감정을 분류할 수 있다.

[0065] *실험 및 결과*

[0067] IMDB movie review 말뭉치를 대상으로 본 발명의 일 실시예에 따른 문서 감정 분류 시스템에 대하여 성능을 평가하였다. 일 실시예에서 말뭉치는 평균적으로 문서당 254 단어를 가지며, 학습 데이터 24,600 문서, 실

험 데이터 25,000 문서, 검증 데이터 400 문서로 구성하였다.

[0068] 공지의 딥러닝 구조에 있어서 워드 임베딩 크기는 100, 은닉 노드의 개수는 100개, FFNN의 은닉 노드의 개수는 200개로 구성하였으며 FFNN의 drop rate는 0.5로 설정하였다.

[0069] 본 발명에 따른 윈도우 크기는 3, 평활 값은 도 5를 참조하면, 검증 데이터에서 가장 높은 결과를 보인 0.1/문서길이로 설정하였다.

[0070] 표 1은 본 발명의 LSTM Window Attention 인코더 구조를 모델로 하는 딥러닝 인코딩 장치를 이용하여 문서 감정을 분류하는 문서 감정 분류 시스템과 다른 인코더를 모델로 이용하는 시스템의 실험 결과이다.

[0071] SVM(Support Vector Machine)의 자질로 unigram과 uni-gram + bi-gram을 사용하는 경우의 성능은 86.59%, 86.95%를 나타났다. 그리고 LSTM 인코더를 모델로 하는 시스템의 경우는 86.5%의 정확도를 보여, 성능이 매우 떨어지는 것을 확인 할 수 있었다. 그러나 본 발명에 따른 문서 감정 분류 시스템의 성능은 89.67%로 LSTM 인코더를 모델로 하는 시스템보다 3.17%p 성능이 향상 되었다. 이는 LSTM 인코더가 전체 시퀀스의 마지막 출력을 문서 벡터로 간주하여, 입력 문서가 길어지면 초기 입력 패턴 인식률이 떨어지기 때문이다. 반면, 본 발명에 따른 문서 감정 분류 시스템이 이용하는 LSTM Window Attention 인코더 구조는 입력 순서, 문서 길이와 상관없이 가중치 학습을 통해 단서 구문을 결정하도록 하기 때문에, LSTM 인코더를 이용하는 시스템 보다 높은 성능을 보인다. 또한, 본 발명에 따른 문서 감정 분류 시스템은 기존에 제안된 딥러딩 인코더인 Paragraph Vector, DCNN 모델을 이용하는 시스템 보다 더 높은 성능을 보였다.

[0073] *표 1*

표 2. 모델 간 성능 비교

실험 모델	정확도
SVM uni-gram	86.59%
SVM bi-gram	86.95%
LSTM encoder	86.5%
Paragraph Vector[7]	87.78%
DCNN[8]	89.4%
Our model	89.67%

[0074]

[0075] 도 6은 도 4에 도시된 문서 감정 분류용 딥러닝 인코딩 장치에 대한 동작 과정을 보인 흐름도로서, 도 6을 참조하여 본 발명의 다른 실시예에 따른 문서 감정 분류용 딥러닝 인코딩 방법을 설명한다.

[0076] 우선, LSTM 시퀀스의 LSTM들에 감정 분류할 문서를 이루는 단어 및 알파벳이 아닌 기호를 입력하고, 이에 따라 처리되는 정보를 선택적으로 갱신 및 저장하도록 단계;(S100). 이어서, 상기 LSTM에 저장된 정보에 대하여 중요도를 산출하고, 산출된 상기 중요도로부터 가중치를 산출한다(S200). 이어서, 산출된 상기 가중치와 상기 LSTM의 출력 값에 따라 기계 학습 하여 문서 벡터를 산출한다(S300).

[0077] 가중치를 산출하는 단계(S200)는 상기 LSTM을 소정 크기의 윈도우 단위로 결합하여 중요도 및 가중치를 산출하는 것이다. 또한, 상기 문서 벡터를 산출하는 단계(S300)는, 상기 LSTM을 소정 크기의 윈도우 단위로 결합하여 산출된 출력 값으로부터 문서 벡터를 산출하는 것이다.

[0078] 또한, 가중치를 산출하는 단계(S200)는 학습 데이터에 대한 과적합을 방지하기 위하여 상기 가중치를 소정의 문서길이에 따른 적응적 평활 값에 따라 재조정하여 산출하는 것일 수 있다.

[0079] 이상으로 본 발명의 기술적 사상을 예시하기 위한 바람직한 실시예와 관련하여 설명하고 도시하였지만, 본 발명은 이와 같이 도시되고 설명된 그대로의 구성 및 작용에만 국한되는 것이 아니며, 기술적 사상의 범주를 일탈함이 없이 본 발명에 대해 다수의 변경 및 수정이 가능함을 잘 이해할 수 있을 것이다. 따라서, 그러한 모든 적절한 변경 및 수정과 균등물들도 본 발명의 범위에 속하는 것으로 간주되어야 할 것이다.

부호의 설명

[0081] 10 : 문서 감정 분류용 딥러닝 인코딩 장치

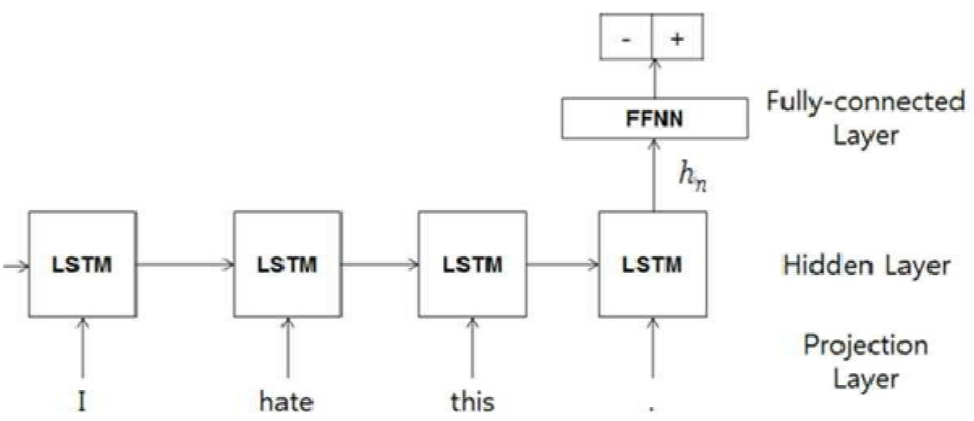
100 : LSTM 시퀀스부

200 : 가중치 산출부

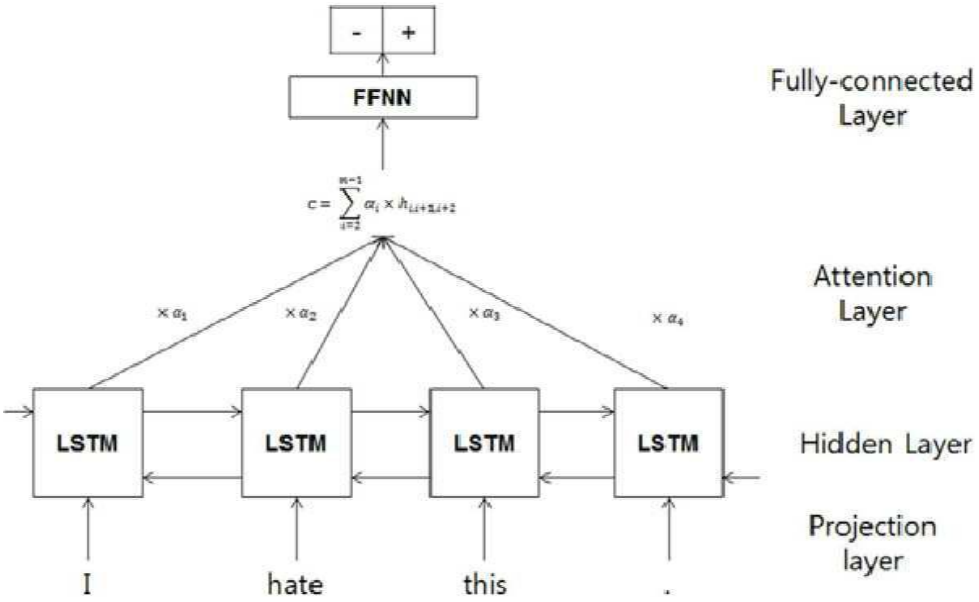
300 : 문서 벡터 산출부

도면

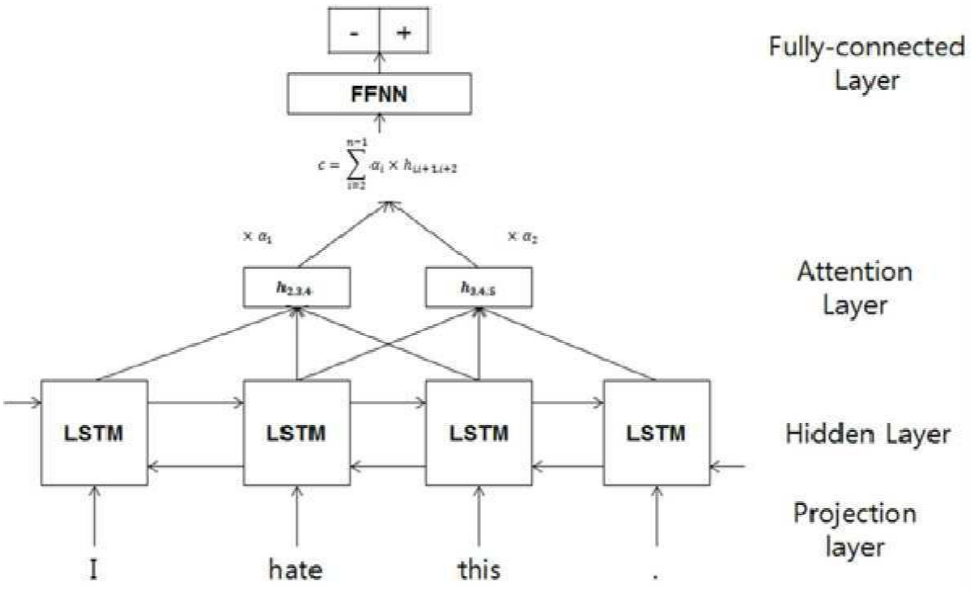
도면1



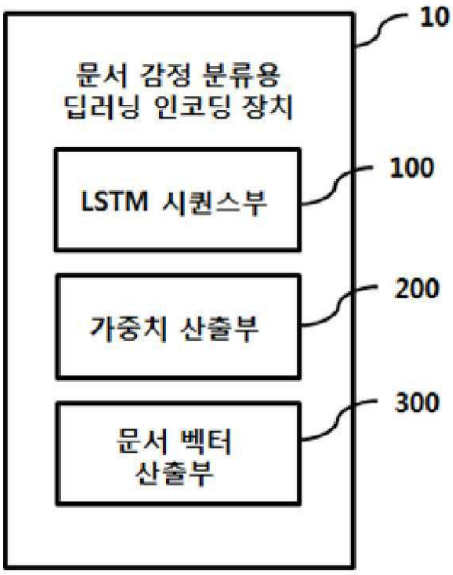
도면2



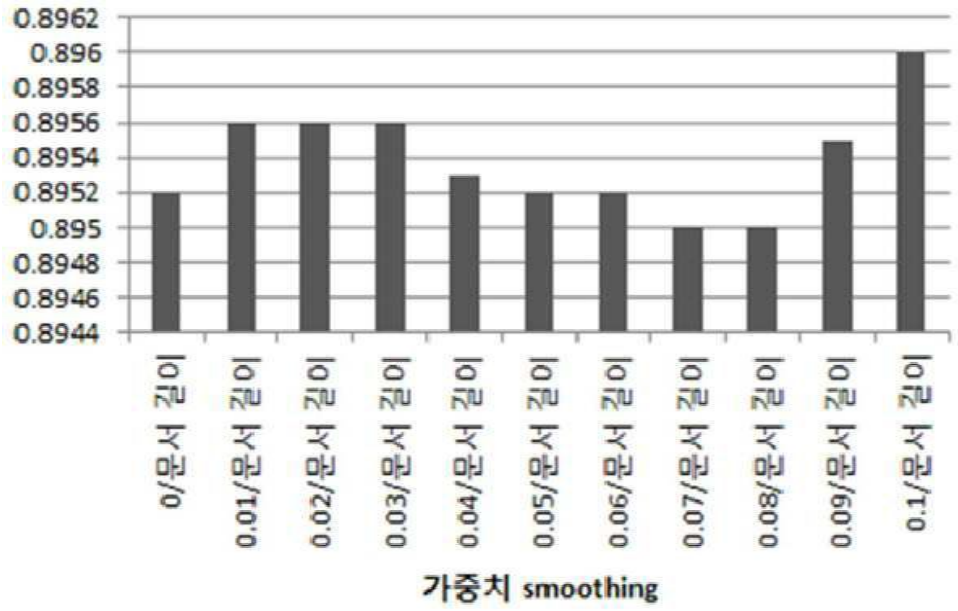
도면3


-
+
FFNN
Fully-connected Layer
Attention Layer
LSTM
LSTM
LSTM
LSTM
Hidden Layer
Projection layer
I
hate
this
.

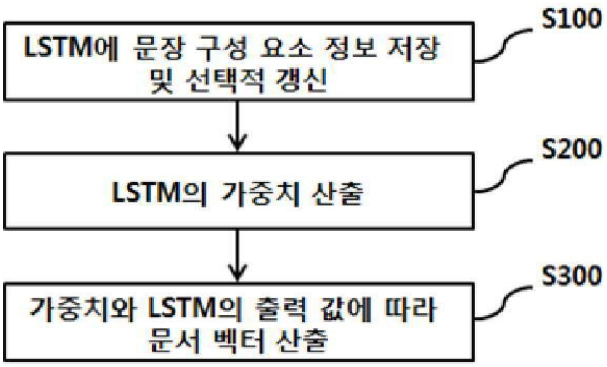
도면4


10
문서 감정 분류용
딥러닝 인코딩 장치
100
LSTM 시퀀스부
200
가중치 산출부
300
문서 벡터
산출부

도면5


0.8962
0.896
0.8958
0.8956
0.8954
0.8952
0.895
0.8948
0.8946
0.8944
0/문서 길이
0.01/문서 길이
0.02/문서 길이
0.03/문서 길이
0.04/문서 길이
0.05/문서 길이
0.06/문서 길이
0.07/문서 길이
0.08/문서 길이
0.09/문서 길이
0.1/문서 길이
가중치 smoothing

도면6


LSTM에 문장 구성 요소 정보 저장 및 선택적 갱신
S100
LSTM의 가중치 산출
S200
가중치와 LSTM의 출력 값에 따라 문서 벡터 산출
S300