BERT를 이용한 숫자-한국어 음역 모호성 해소

박정연○, 육대범, 이재성
충북대학교
{parkjeongyeon, daebum1994, jasonlee}@chungbuk.ac.kr

Arabic-Numerals to Korean Transliteration Disambiguation using BERT

Jeong Yeon Park○, Dae Bum Yuk, Jae Sung Lee
Chungbuk National University

요 약

TTS(Text-to-Speech) 시스템을 위해서는 한글 이외의 문자열을 한글로 변환해줄 필요가 있다. 이러한 문자열에는 숫자, 특수문자 등의 문자열이 포함되어 있다. 특히 숫자의 경우, 숫자가 사용되는 문맥에 따라 그 발음방법이 달라지는 문제점이 있다. 본 논문에서는 기존의 규칙기반과 한정된 문맥 정보만을 활용할 수 있는 방법이 아닌, 딥러닝을 이용한 방법으로 문맥에 따라 발음방법이 달라지는 숫자 음역의 모호성을 해소하는 방법을 소개한다.

주제어: 음역, BERT, 모호성 해소, 음성합성

1. 서론

최근 책을 읽어주는 서비스와 언어를 배우는 사람을 위한 시스템 등 TTS(Text-to-Speech)와 관련된 시스템의 수요 증가로 올바른 음역의 중요성이 강조되고 있다[1]. 여기서 음역이란 외래어, 숫자 등을 그 소리를 따서 자국의 문자 체계로 표기하는 것을 말한다. 일반적으로 TTS는 텍스트 분석부와 음성 합성부를 가지고 있으며, 특히 음역을 진행하는 텍스트 분석부의 결과는 합성음의 자연스러운 발음을 결정하는 중요한 역할을 한다[2].

한국어에서 숫자는 문맥에 따라 고유어, 한자어, 외래어 등으로 그 발음 방법이 다르게 결정된다. 예를 들어, "3권"이라는 단어를 발음할 때, 책의 목차 등을 나타낼 때에는 한자어 "삼권"으로 발음되고, 책의 수량 등을 나타낼 때에는 고유어 "세권"으로 발음된다. 또한, "시즌 1"이라는 단어를 발음할 때, "시즌 일","시즌 하나","시즌 원" 등 다양한 발음이 가능하지만, 일반적으로 외래어에 붙은 숫자는 같은 외래어 발음으로 음역하여 "시즌 원"으로 발음된다. 이 외에도 숫자가 고유명사로 사용되는 "112", "4.19" 와 같은 경우는 일의 단위로만 읽어 각각 "일일이", "사일구"로 발음된다. 이와 같이 숫자가 경우에 따라 그 발음 방법이 달라지는 것을 숫자 음역 모호성이라고 하며, TTS 시스템 성능 저하의 원인이 된다[3].

기존의 한국어 연구에서는 숫자 음역 모호성을 해결하기 위해 다양한 방법이 제시되었다. [4-6]의 연구에서는 숫자 주변 문맥의 패턴 또는 숫자의 의존명사를 이용해 규칙 기반의 방법으로 숫자 음역 모호성을 해소하였다. [7]의 연구는 숫자와 함께 나타나는 문맥을 시간, 장소, 제목, 단순 숫자 등 10가지 패턴으로 구분하여 활용하고, 이를 결정 트리(Decision Tree) 알고리즘으로 학습하여 숫자 음역 모호성을 해소하였다. 또한, [8]의 연구에서는 [7]보다 정밀한 카테고리를 설정하고 규칙과 함께 활용하는 연구를 했다. 먼저, 학습말뭉치에서 모호성을 숫자 주변 단어들을 클러스터링하였다. 이후 클러스터링 결과에 나타난 단어들을 시소러스에 맵핑(Mapping)시키고, 그 단어들의 최하위 공통 조상을 패턴 카테고리로 선택하여 사용하였다.

그러나 한국어에서는 단어의 생략이 빈번하게 일어나며, 신조어가 지속적으로 등장하고 있기 때문에 모든 경우를 규칙으로 결정하는 것은 어려운 일이다. 또한 모호성 있는 숫자와 의존명사 등이 사용되는 문맥을 학습하기 위해 새로운 말뭉치를 구축하는 것은 어려운 문제이다. 따라서 기존 연구의 문제점을 해결할 수 있는 연구가 필요하다.

최근 대용량 말뭉치를 사전학습(Pre-training)한 문맥기반 언어모델인 BERT[9]가 다양한 자연어처리 분야에서 높은 성능을 보이고 있다. 이에 본 논문도 BERT를 활용하여 숫자 음역 모호성을 해소하는 방법을 제안한다.

2. BERT 기반 숫자-한국어 음역 방향 결정

최근의 자연어 처리 연구들은 단어 벡터나 미리 추출한 특징들을 활용하는 대신, 대량의 데이터를 이용해 문맥을 사전학습한 문맥 기반 언어모델을 원하는 과제(Task)에 맞도록 정밀조정(Fine-tuning)하는 방법이 주를 이루고 있다. 특히 문맥 기반 언어모델인 BERT는 정밀조정 방법으로 NER, SQuAD, MNLI 등 다양한 과제에서 최고의 성능을 보인다[9].

숫자의 음역은 숫자가 포함된 문맥에 따라 결정된다고 할 수 있다. 따라서 본 연구와 유사한 단어 의미 모호성 해소에서 높은 성능을 보인 [10]의 방법을 활용하여, 문맥 기반 언어모델인 BERT를 정밀조정하는 Sequence Labeling 문제로 접근하여 숫자 음역 모호성을 해소한다.

일반적으로 숫자 음역은 크게 '일', '이', '삼' 등의 발음인 한자어, '하나', '둘', '셋' 등의 고유어, '원', '투', '쓰리' 등의 외래어[1], 기관 및

1) 본 논문에서 외래어는 영어에서 온 외래어만 다룬다.

특정 사건을 나타내는 고유명사 등의 4가지 음역 방향을 가진다[6]. 여기서 숫자가 고유명사인 경우는 보통 특수문자 등을 제외하고 숫자만을 일의 자리 단위로 읽는다. 해당 경우와 같이 일의 자리 단위로 모든 숫자를 읽어야 하는 경우, 본 논문에서는 이를 모두 고유명사로 지정하였다. 여기에는 사건, 기관, 휴대전화 읽기 등의 경우가 있다. 고유명사에 해당하는 예시는 [표 1]과 같다.

숫자	발음	구분
119	일일구	기관(소방서)
6.25	육이오	사건(한국전쟁)
000-111-2222	공공공 일일일(에) 이이이이	전화번호

표 1. 숫자가 고유명사인 경우의 예시

본 논문은 숫자 음역 모호성 해소에 중점을 두고 있다. 그러므로 숫자 음역 모호성 해소를 위하여 4가지 음역 방향 중 올바른 음역 방향을 선택할 수 있도록 모델을 구성하고 평가하였다. 완전 일치 여부를 통해 음역 시스템의 성능을 측정하면 띄어쓰기 등의 일부 문자열 차이로 모호성 해소 여부를 정확히 측정하기 어렵기 때문이다[11]. 예를 들어, [표 1]의 전화번호의 경우, 특수문자 “-”를 “에”로 음역하거나 생략할 수 있지만, 모두 올바른 음역 결과다.

이와 같이 숫자의 음역 결과는 경우에 따라 일부 문자열의 차이를 가질 수 있다. 따라서 음역 결과의 완전 일치 여부보다 음역 방향의 일치 여부가 숫자 음역의 올바른 평가가 된다. 일단 음역 방향이 결정되면 간단한 규칙이나 맵핑 등으로 문자 변환을 하여 음역을 완성할 수 있다. 따라서 음역 방향을 결정하는 것이 숫자 음역 모호성 해소에 중요하다.

음역 방향 결정을 위해 본 논문에서 제안하는 모델은 [그림 1]과 같다. BERT의 출력을 LSTM Layer의 입력으로 사용하며, LSTM Layer의 출력을 분류기를 통해 최종 출력을 결정한다. 여기서 분류기의 출력은 한자어, 고유어, 외래어, 고유명사로 구분된 4개의 숫자 태그와 숫자가 포함되지 않은 모든 경우를 나타내는 단어 태그, 문장의 시작과 끝을 나타내는 SE태그, 총 6가지를 갖는다.

BERT 모델은 ETRI에서 제공되는 KorBERT[12] 어절 모델을 사용한다. 입력은 CoNLL 포맷[13]으로 변환하여 제공되며, 그 외 전처리 과정은 수행하지 않는다.

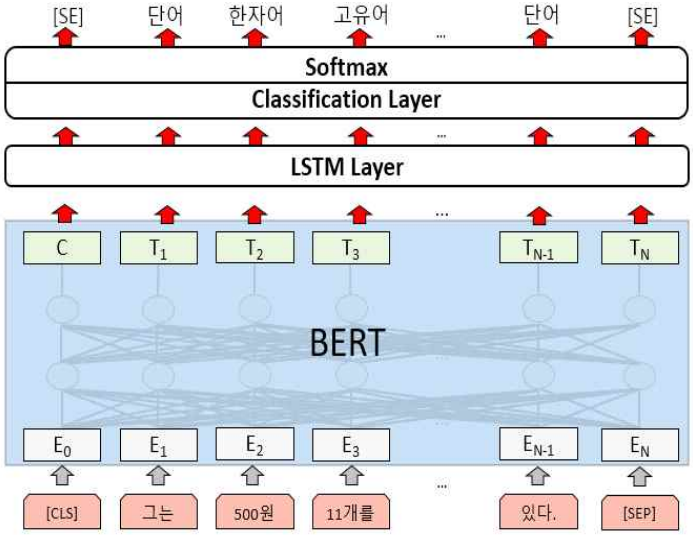

그림 1. BERT기반 숫자-한국어 음역 방향 결정 모델

[그림 1] 예시의 음역 방향에 따라 변환한 문자열은 [표 2]와 같다.

단어	음역 방향	음역 결과
그는	단어	그는
500원	한자어	오백원
11개를	고유어	열한개를
…	…	…
있다.	단어	있다.

표 2. [그림 1]의 음역 방향에 따른 음역 예시

3. 실험 및 평가

실험은 [6]의 연구에서 사용된 데이터[2]를 사용한다. 여기서 숫자가 포함된 5,516개 문장 중 중복 데이터를 제거한 5,001개 문장에 5가지 음역 방향(단어, 한자어, 고유어, 영어, 고유명사)을 태깅하고, 5배수 교차검증으로 평가를 진행하였다. 평가는 “단어” 태그를 제외한 음역 방향의 일치 여부를 F1-score로 측정하였다.

본 논문에서는 규칙기반 방법과의 비교를 위해 [6]의 음역 시스템의 출력 형태를 변환하여 본 연구의 실험결과 평가와 마찬가지로 올바른 음역을 결정하는지를 평가하여 비교하였다. 실험결과는 [표 3]과 같다.

	규칙기반 모델[6]	제안모델 (BERT-based Model)
F1-score	82.14 %	96.74 %

표 3. 모델 실험결과 비교

2) 2018년 전자부품연구원(KETI) 프로젝트 과정에서 제공받음.

여기서 규칙기반 모델은 [6]의 논문에 보고된 성능과 다른 성능을 보이는데, 이는 [6]에서는 전체 데이터 중 일부 데이터를 무작위로 추출하여 성능을 평가한 반면, 본 논문의 평가에서는 [6]의 전체 데이터를 대상으로 정제하여 새롭게 평가하였기 때문이다.

본 논문에서 제안하는 모델은 숫자의 앞뒤 문맥 1개 단어만을 고려하는 [6]의 규칙기반 모델에 비해 약 14.6%p 높은 성능을 보이고 있다. 이는 제한된 문맥 정보와 한정된 규칙을 사용하는 규칙기반모델과 다르게 문장 내의 모든 문맥 정보를 파악하고, 이를 기반으로 정해진 규칙 외의 음역 결정이 가능하기 때문이다.

4. 결론

최근 수요가 증가하고 있는 Text-to-Speech 기술의 성능 향상을 위해서 올바른 음역이 필요하다. 본 논문은 문맥 기반 언어모델인 BERT를 정밀조정하여 숫자의 올바른 음역 방향을 결정하는 방법으로 숫자 음역 모호성을 해소하는 연구를 하였으며, 이를 규칙기반의 시스템 성능과 비교하였다. 그 결과, 한정된 정보를 사용하는 규칙기반 방법보다 문맥 정보를 사용하는 BERT를 활용하였을 때 숫자 음역 모호성을 좀 더 잘 해소할 수 있음을 보였다.

감사의 글

이 (성과물)은 중소벤처기업부 '산업전문인력역량강화사업' 의 재원으로 한국산학연협회(AURI)의 지원을 받아 수행된 연구임. (2020년 기업연계형연구개발인력양성사업, 과제번호 : S2929950)

참고문헌

[1] Soumyadeep Kundu, Sayantan Paul and Santanu Pal, "A Deep Learning Based Approach to Transliteration", Proceedings of the Seventh Named Entities Workshop, pp.79-83, 2018.

[2] 최연주, 정영문, 김영관, 서영주, 김회린, "한국어 text-to-speech(TTS) 시스템을 위한 엔드투엔드 합성고무 방식 연구", 말소리와 음성과학 제10권 제1호, pp.39-48, 2018.

[3] 정영임, 김정세, 김상훈 ,이영직, 윤애선, "현대 한국어에서 아라비안 숫자의 읽기 규칙 연구", 제14회 한글 및 한국어 정보처리 학술발표 논문집, pp. 16-23, 2002.

[4] Aesun Yoon, Hyuk-Chul Kwon and Man-Hyeong Lee, "An automatic transcription system for Arabic numerals in Korean", International Conference on Natural Language Processing and Knowledge Engineering, pp.221-226, 2003.

[5] 정영임, 윤애선, 권혁철, "임베디드 TTS 시스템을 위한 아라비안 숫자의 문자 변환", 한국정보과학회 학술대회 논문집, pp.442-444, 2005.

[6] 박정연,신형진,육대범,이재성, "음성합성을 위한 텍스트 음역 시스템과 숫자 음역 모호성 처리", 제30회 한글 및 한국어 정보처리 학술대회 논문집, pp.449-452, 2018.

[7] Youngim Jung, Donghun Lee, HyeonSook Nam, Aesun Yoon and Hyuk-chul Kwon, "Learning for Transliteration of Arabic-Numeral Expressions Using Decision Tree for Korean TTS", INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, 2004.

[8] Youngim Jung, Aesun Yoon, and Hyuk-Chul Kwon, "Disambiguation Based on Wordnet for Transliteration of Arabic Numerals for Korean TTS", International Conference on Intelligent Text Processing and Computational Linguistics, pp. 366-377, 2006.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, 2018.

[10] 윤준영, 신형진, 박정연, 이재성, "BERT를 이용한 한국어 단어 의미 모호성 해소", 제31회 한글 및 한국어 정보처리 학술대회 논문집, pp.485-487, 2019.

[11] 박주희, 박원준, 서희철 "문장대문장 학습을 이용한 음차변환 모델과 한글 음차변환어의 발음 유사도 기반 부분매칭 방법론", 제30회 한글 및 한국어 정보처리 학술대회 논문집, pp.443-448, 2018.

[12] KorBERT, http://aiopen.etri.re.kr

[13] CoNLL-X format, https://universaldependencies.org/format.html