



RNN의 새로운 앙상블 기법을 통한 Seq2Seq 모델 성능 개선

Improving Seq2Seq model performance through new RNN ensemble method

Ahjeong Park^o Chulyun Kim

ahjeong@sookmyung.ac.kr cykim@sookmyung.ac.kr

Sookmyung Women's University

Abstract

자연어 처리, 음성인식, 번역과 같이 문맥이 있는 Sequence 데이터와 날씨, 주가와 같이 순서가 있는 시계열(Time-Series) 데이터는 순차적으로 등장하는 데이터 처리에 적합한 순환 신경망(Recurrent Neural Network) 모델을 많이 사용한다. 일반적으로 순환 신경망 모델에서 앙상블은 여러 개의 분류기를 통해 나온 예측을 결합하여 더 좋은 예측을 할 수 있도록 한다.

하지만 RNN 기반 모델은 하나의 아웃풋이 재귀적으로 다음 아웃풋에 영향을 주기 때문에 기존의 최종 결과만을 비교하는 앙상블의 방식은 RNN의 구조적인 특징을 모두 반영하지 못하는 문제를 가지고 있다. 특히 이러한 구조를 사용하는 기계 번역에서 단순히 최종 결과를 결합하는 앙상블은 문맥을 고려하지 못한 결과를 출력할 가능성이 높다. 따라서 재귀적으로 다음 아웃풋에 영향을 주는 RNN 기반 모델의 구조적인 특징을 고려한 Time Series 모델의 새로운 앙상블 기법을 제안한다. 그리고 기계 번역에서 앙상블을 통한 성능 향상 실험을 통해 우리의 새로운 앙상블 기법이 효과적임을 확인한다.

Introduction

언어 모델링, 음성인식, 기계번역 또는 챗봇과 같이 문맥이 중요한 Time-Series 데이터는 인공 신경망(Neural Network) 또는 합성곱(Convolutional Neural Network) 신경망에서 다루기는 어렵기 때문에 이전의 결과가 다음 결과에 영향을 미치는 순환 신경망(Recurrent Neural Network)과 변형 구조인 LSTM, GRU를 사용하게 된다.

그동안 RNN 기반 모델의 앙상블은 여러 개의 모델을 생성하고 각 모델에서 나온 최종 결과를 결합해 더 정확한 예측을 도출하도록 했다. 이는 시계열 데이터 예측의 성능 향상을 가져왔다.

하지만 RNN 기반 Encoder, Decoder로 이루어진 Seq2Seq 모형을 사용하는 기계 번역에서 최종 결과로 나온 문장의 단순 결합 앙상블 방식은 재귀적으로 다음 아웃풋에 영향을 주는 RNN의 구조적인 특징을 모두 반영하지 못한다는 점에서 적절하지 않다.

따라서 본 논문은 기존의 최종 결과를 단순 결합하는 앙상블 방식이 아닌 하나의 아웃풋이 재귀적으로 다음 아웃풋에 영향을 주는 RNN의 구조적인 특징을 모두 반영할 수 있도록 RNN 기반 모델의 Time-Step에서의 앙상블 기법을 제안하고자 한다. 그리고 최종 결과를 단순 결합하는 기존의 앙상블과 비교해 우리의 앙상블 기법이 RNN의 구조적 특징을 고려한 앙상블 방식에 효과적임을 보인다.

그리고 2가지 앙상블 기법에 대해서 Seq2Seq 모델을 사용하는 기계번역 실험을 진행한다.

Related Work

RNN 기반 모델(LSTM, GRU 등)을 사용한 앙상블 방법이 활발하게 연구되고 있다. 그중 Deep 4 LSTM[3]은 4개의 LSTM 모델에서 출력된 결과값에 가중치를 곱하여 결합해 최종 결과를 출력한다. 이는 암호 화폐 가격 예측에서 Simple RNN, Simple LSTM과 2-LSTM, 3-LSTM 보다 우수한 성능을 보여주었다.

Adaboost GRU는 Adaboost와 GRU를 결합하여 금융 시계열 예측에 활용했다. Nae Won Kwak et al.는 KerasRegressor 클래스로 GRU 모델 생성 후 Adaboost 알고리즘으로 모델을 결합하여 기존 연구보다 좋은 개선을 이루었다.

Seq2Seq를 활용하는 기계번역에서 Rico Sennrich et al.는 WMT에서 우수한 성능을 보인 모델 Nematus[6] 기반 2가지 앙상블 기법을 제안했다. Single model의 마지막 N개의 checkpoint를 결합하는 checkpoint ensemble과 서로 다른 hyperparameter로 훈련된 독립적인 모델을 결합하는 independent ensemble을 함께 사용해 성능 개선을 이루었다.

Experiments

실험 환경
기계번역 실험에서는 Tab-delimited Bilingual Sentence Paris에 제공된 영어-스페인 데이터 셋을 사용한다. 베이스 모델은 Bahdanau et al.의 Attention을 적용한 Encoder-Decoder 구조를 사용한다. 데이터 셋 크기는 60000개로 그중 80%의 데이터를 사용해 15개의 모델을 랜덤으로 학습시킨다. 학습된 15개의 모델은 2가지의 앙상블 방법으로 결합한 후 영어 -> 스페인어 번역을 진행한다.

성능 평가
기계 번역에서 성능 평가는 쉽고 빠르게 번역의 품질을 평가하는 python package TQE를 사용한다. 12000개의 번역된 문장에 대해서 TQE 측정 후 평균을 내어 모델의 성능을 평가한다.

실험 결과
표 1은 영어-스페인 기계 번역 실험에서 각 모델에 대한 TQE 성능 지표를 나타낸다. 앙상블에 사용된 15개의 단일 모델 중 가장 높은 점수와 비교했을 때 2가지의 앙상블 방법이 더 높다. 그중 Time-Step마다 출력을 고려한 서바이벌 앙상블 기법이 최종 출력을 고려한 다수결 앙상블 기법 보다 점수가 높다는 것을 확인할 수 있다.

표 1 기계번역 앙상블 모델과 단일 모델 TQE 성능 비교

	Ensemble		Single Models (Baseline)				
	Survival	Majority	Model 1	Model 2	Model 3	Model 4	Model 5
TQE	0.731033	0.724257	0.699758	0.685662	0.693042	0.698464	0.685117
			Model 6	Model 7	Model 8	Model 9	Model 10
TQE			0.692936	0.685694	0.701483	0.684669	0.667196
			Model 11	Model 12	Model 13	Model 14	Model 15
TQE			0.699275	0.687404	0.692424	0.690276	0.700335

Method

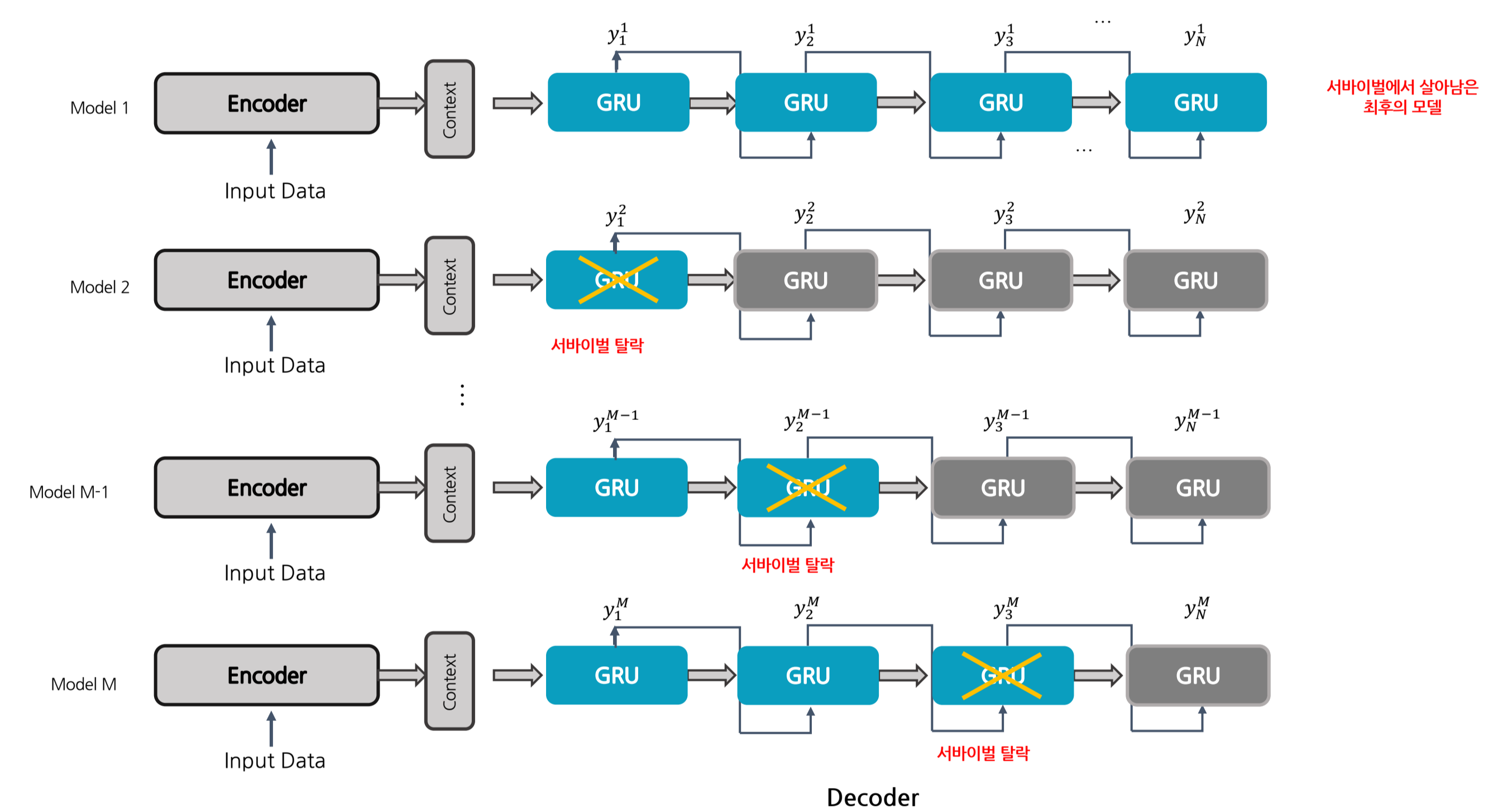
본 연구에서 RNN 기반 모델의 앙상블 방법의 핵심은 Time-Step 별 중간 결과를 고려해 최종 결과를 채택한다는 것이다. 본 논문에서 실험한 앙상블 방식은 총 2가지이다. 하나는 매 Time-Step에서 그 순간 정답을 맞춘 모델만 살아남는 방식으로, 마지막 Time-Step까지 살아남는 모델의 최종결과를 출력하는 방식이다. 본 논문은 이러한 앙상블을 서바이벌 앙상블 기법으로 표현한다. 여기서 정답이란 여러 개의 모델이 출력한 결과 중 가장 많이 등장한 결과로, 다수결로 뽑힌 결과를 말한다. 기계 번역에서는 다수결로 뽑힌 단어로 생각할 수 있다. 다른 하나는 각각의 모델에서 나온 결과물을 다수결 처리하여 최종 결과를 만들어내는 방식이다. 즉 RNN 중간 Time-Step 결과를 고려하지 않고 원래대로 각 모델마다 결과를 내놓으면, 그 결과를 기반으로 다수결을 통해 최종 결과를 도출하는 것이다. 이러한 앙상블을 다수결 앙상블 기법이라고 한다.

Ensemble with Survival

N 개의 모델을 훈련 후 예측 단계에서 매 Time-Step마다 N 개의 모델 중 정답을 맞춘 모델만 살아남는 방식이다. 살아남은 Winner 모델은 다음 Time-Step에 참가할 수 있다. 이렇게 마지막 Time-Step까지 살아남은 최후의 모델의 출력을 얻는다. 이때 최후의 모델은 하나 또는 여러 개가 될 수 있다.

다음 Time-Step에 참가할 모델을 정하는 기준 정답은 각 모델이 출력한 결과의 다수결로 결정한다. 따라서 이 정답을 출력하지 못한 나머지 Loser 모델은 서바이벌에서 탈락해 다음 Time-Step에 참가할 수 없다.

만약 최후의 모델이 여러 개이고 다수결로 정하지 못하는 경우에는, 각 모델의 Softmax 값 중 큰 값의 출력을 기준 정답을 결정한다. 그 후에 다시 Winner 모델을 결정한다.



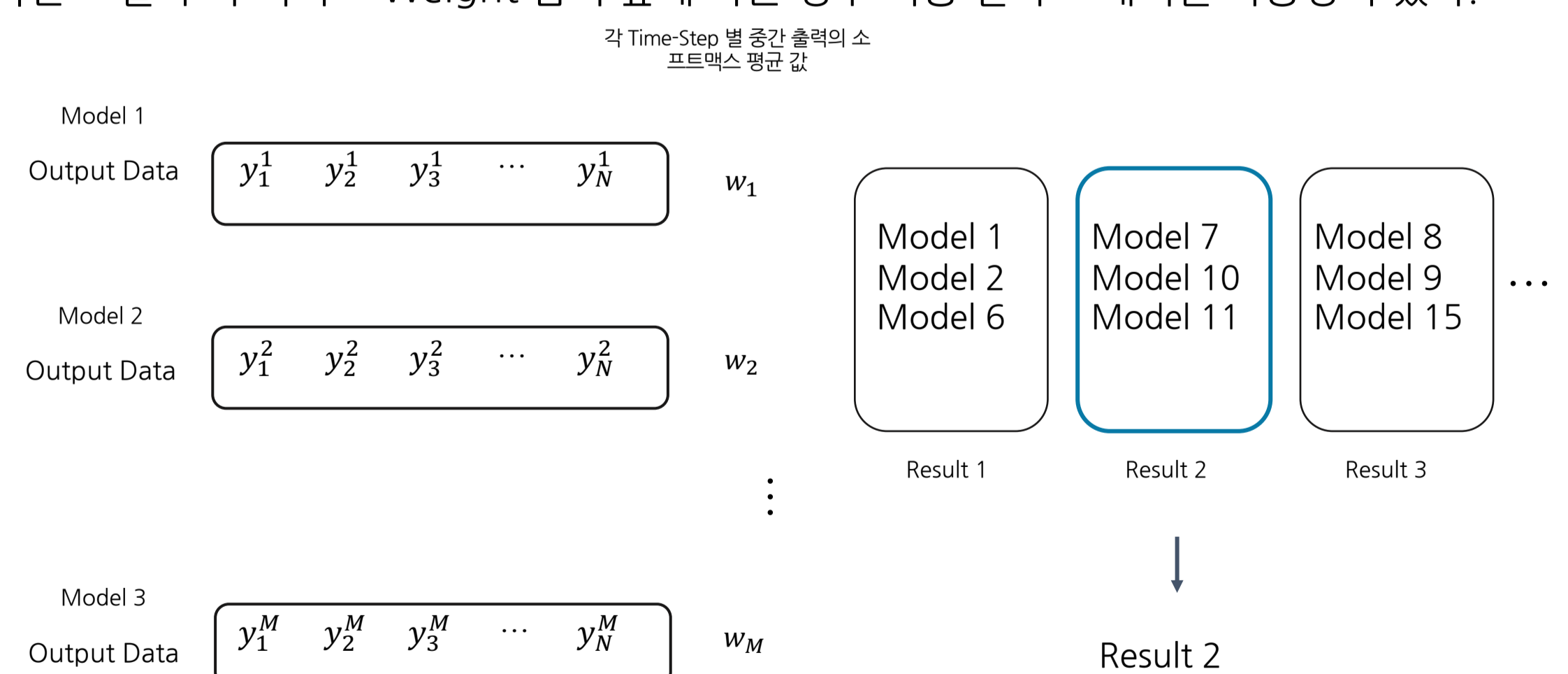
Ensemble with Majority

최종 출력 결과를 결합하는 기존의 앙상블 기법을 활용한다. N 개의 모델을 훈련 후 N 개의 모델은 마지막 Time-Step까지 예측한다. 그 N 개의 결과 중 다수결을 통해 최종 결과를 완성한다. 위 서바이벌 앙상블 기법과의 차이점은 중간 Time-Step의 결과를 고려하지 않는 점이다.

N 개의 최종 출력을 얻는 다수결은 다음과 같이 정한다. 먼저 N 개의 모델로부터 나온 각각의 최종 출력을 통해 그 모델의 Weight를 결정한다. Weight는 각 Time-Step 별 중간 출력의 소프트맥스의 평균값으로 정한다.

이제 각 모델의 출력에 대해 같은 것끼리 그룹을 짓고 그룹에 속한 모델의 weight 합을 구한다. 여러 개의 그룹 중 가장 weight 합이 큰 그룹의 답을 최종 결과로 채택한다.

이는 단순히 모델 수에 기반한 다수결이 아니라 소프트맥스의 합을 고려한 결과이다. 따라서 그룹에 속한 모델 수가 적어도 Weight 합이 높게 나온 경우 최종 결과로 채택될 가능성이 있다.



Conclusion

본 연구에서는 기존의 RNN 기반 모델에서 앙상블과 달리 매 Time-Step 별 출력을 고려하여 최종 결과를 채택하는 새로운 RNN 앙상블 기법을 제안했다. 그리고 최종 결과를 결합하는 다수결 앙상블 기법과 비교하였다. 특히 기계번역에서 Time-Step마다 여러 개의 모델의 출력을 고려한 뒤 다음 Time-Step으로 넘어가는 서바이벌 앙상블 기법은 단일 모델뿐만 아니라 최종 결과를 결합하는 다수결 앙상블 기법보다 성능이 향상된 것을 확인하였다. 이를 통해 RNN 기반 모델로 이루어진 Seq2Seq 구조에서 Time-Step 별 출력을 고려한 앙상블은 다음 아웃풋에 영향을 주는 RNN의 특징을 모두 반영하면서 성능 향상에 효과적임을 보였다. 다수의 답이 언제나 옳다고 할 수 없지만 서바이벌 모델은 정답이 없는 상황에서 다수의 답을 정답으로 채택함으로써 미리 오류가 포함되어 있는 부정확한 모델을 제거할 수 있게 된다. 이는 최종 결정에서의 노이즈 적인 의견을 제거하는 효과를 갖게 되어 다수결 모델보다 성능 향상이 이루어진 것으로 보인다.

추후 Time-Step 별 출력을 고려하는 것뿐만 아니라 출력을 활용해 다음 Time-Step에 영향을 줄 수 있는 새로운 앙상블 기법을 연구할 계획이다. 또한 Seq2Seq 구조를 사용할 수 있는 다른 활용 예제에서도 우리의 기법이 효과적인지 실험을 진행하고 기계 번역의 데이터 셋 범위를 확장해 영어-스페인어 뿐만 아니라 다양한 언어에서도 성능 향상을 이끌 수 있는 방법을 연구할 계획이다.