Comparative Study of Neural Language Models

Abhinash Khare Soham Pal Sruthi Gorantla

September 27, 2017

Motivation

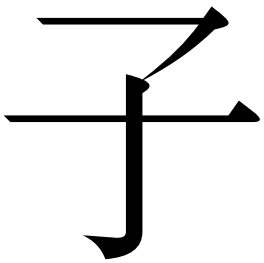
un-break-able

Motivation

Morpheme

A meaningful morphological unit of a language that cannot be further divided (e.g. *in*, *come*, *-ing*, forming *incoming*).

Motivation



child

Motivation

dog

Motivation

子 犬

child *dog*

Motivation

子犬

puppy

Relevant Literature

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush.
Character-aware neural language models.
CoRR,abs/1508.06615, 2015.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black.
Character-based neural machine translation.
CoRR,abs/1511.04586, 2015.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu.
Exploring the limits of language modeling
CoRR,abs/1602.02410, 2016.

Relevant Literature II

Ilya Sutskever, Oriol Vinyals and Quoc V. Le.
Sequence to Sequence Learning with Neural Networks.
CoRR,abs/1409.3215, 2014.

Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio.
Neural Machine Translation by Jointly Learning to Align and Translate.
CoRR,abs/1409.0473, 2014.

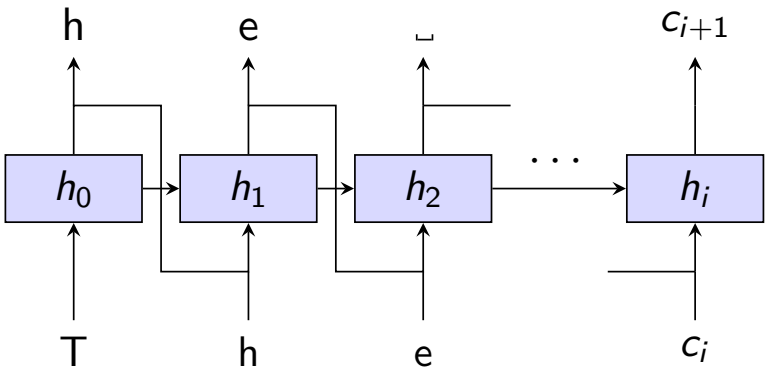
Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky and Andrew Y. Ng.
Neural Language Correction with Character-Based Attention.
CoRR,abs/1603.09727, 2016.

Architecture Overview

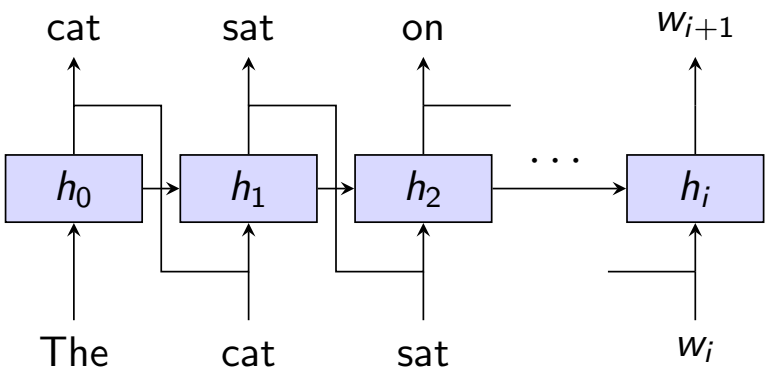
The architecture of our language models can be divided into:

1. A **back-end** that encodes the data either at a character-level or at a word-level.
2. A **front-end** that receives the encoded input from the back-end and produces the next character or a word in the sequence. It is implemented as an RNN.

Choice of Front-ends

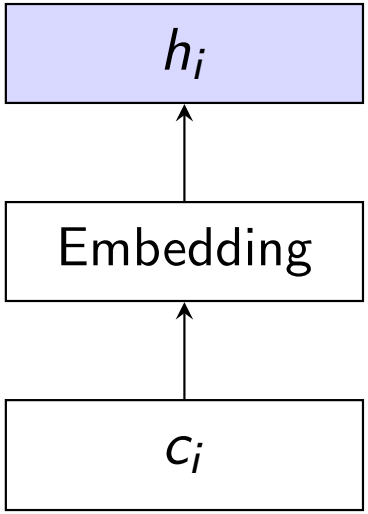


Character-based model

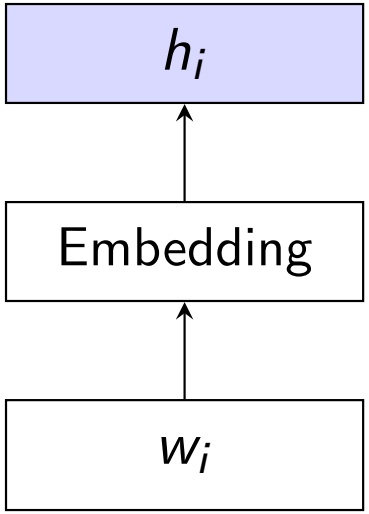


Word-based model

Choice of Back-ends I

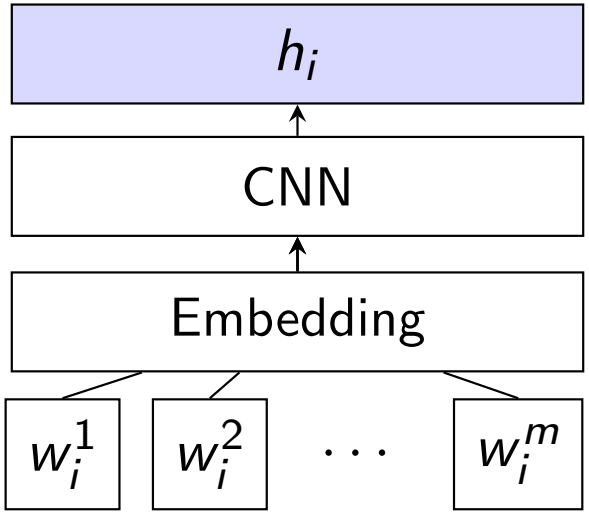


Character embeddings

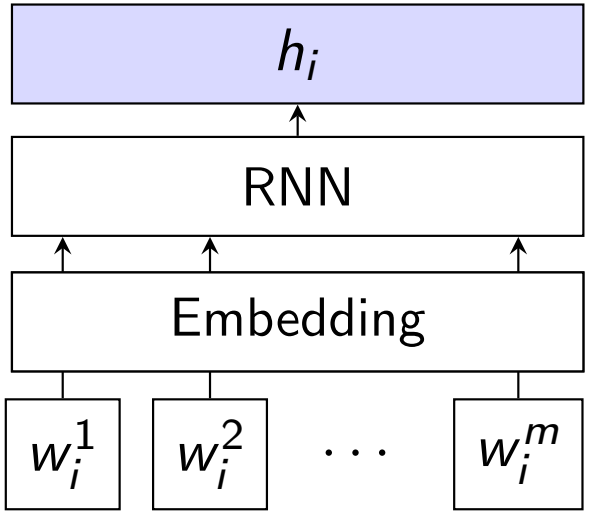


Word embeddings

Choice of Back-ends II



CNN on characters



RNN on characters

Dataset

Open Weiboscope Data Access 微博像

1. 226 million posts collected from 新浪微博 *Sina Weibo* (Twitter-like platform used almost exclusively in China).
2. Collected in 2012 from feeds of users having > 1000 followers.
3. Released for public use by Journalism and Media Center of the University of Hong Kong, citation required, but no specific licensing terms.

Timeline

1. Start out by preprocessing the dataset and setting up the pipeline in unison.
2. Split the experiments 3-way:
   2.1 Embeddings (word + character level)
   2.2 CNN over character embeddings
   2.3 RNN over character embeddings
3. Pit the best scores of each against the others.
4. Study the effect of the various architectures in greater detail.

Thank You