NYU

A Vietnamese Named Entity Recognition System for COVID-19 Articles

Ngoc Nhu Hoang | CSCI-UA 480 | Fall 2021

Problem statement

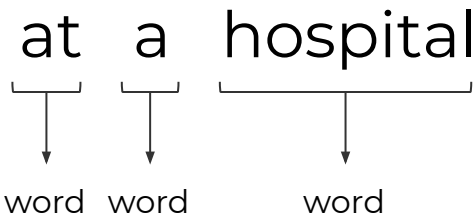
This project presents a **named entity recognition system** for the specific domain of **Vietnamese COVID-19 news articles**.

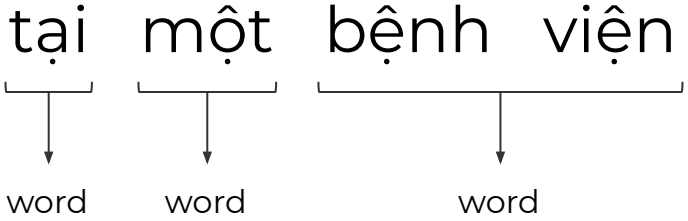
Simple deep learning model with input including:

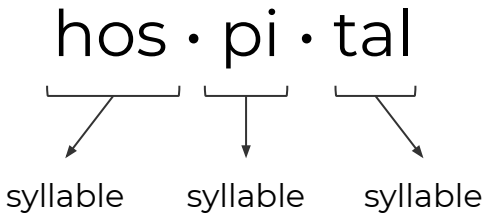
- word embeddings
- part-of-speech tags
- manually extracted features

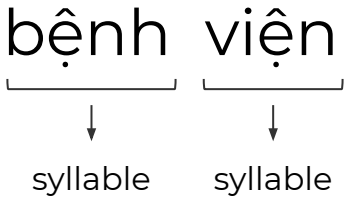
The system can identify 10 types of named entities with an unweighted average F score of about 90.41% on the test set (>3K sentences, >50K words, >11K entities).

Word vs. syllable









PART 01

APPROACH

Data source

Neural network architecture

Input data engineering

Data

COVID-19 Vietnamese named entity recognition dataset (Truong et al., 2021).

Includes 10 entity types: PATIENT_ID, PERSON_NAME, AGE, GENDER, OCCUPATION, LOCATION, ORGANIZATION, SYMPTOM&DISEASE, TRANSPORTATION, and DATE.

Includes 20 tags: B- and I- for each type & O tag.

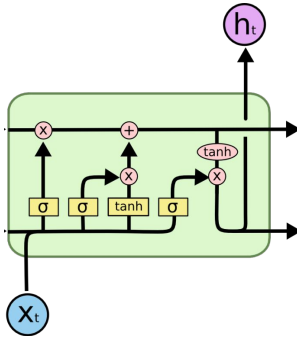
Contains 10K sentences, >270K words, >35K entities.

Split into train/validation/test sets with ratio 5/2/3.

Words	NER tags
cách_ly	O
tại	O
Bệnh_viện	B-LOCATION
Bệnh	I-LOCATION
Nhiệt_đới	I-LOCATION
Trung_ương	I-LOCATION

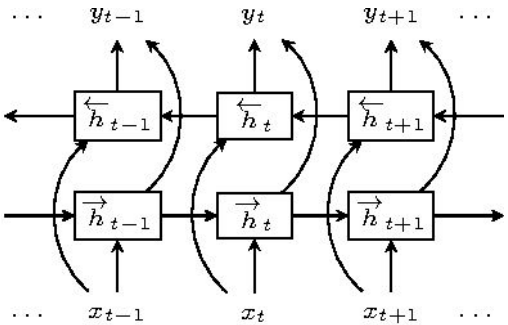
Sample from the dataset. “quarantine(d) at the National Hospital of Tropical Diseases”.

Neural network



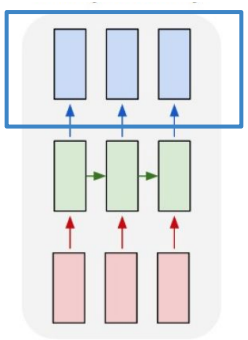
Long Short Term Memory

A type of Recurrent Neural Network capable of capturing long term dependencies in sequence data



Bidirectional LSTM

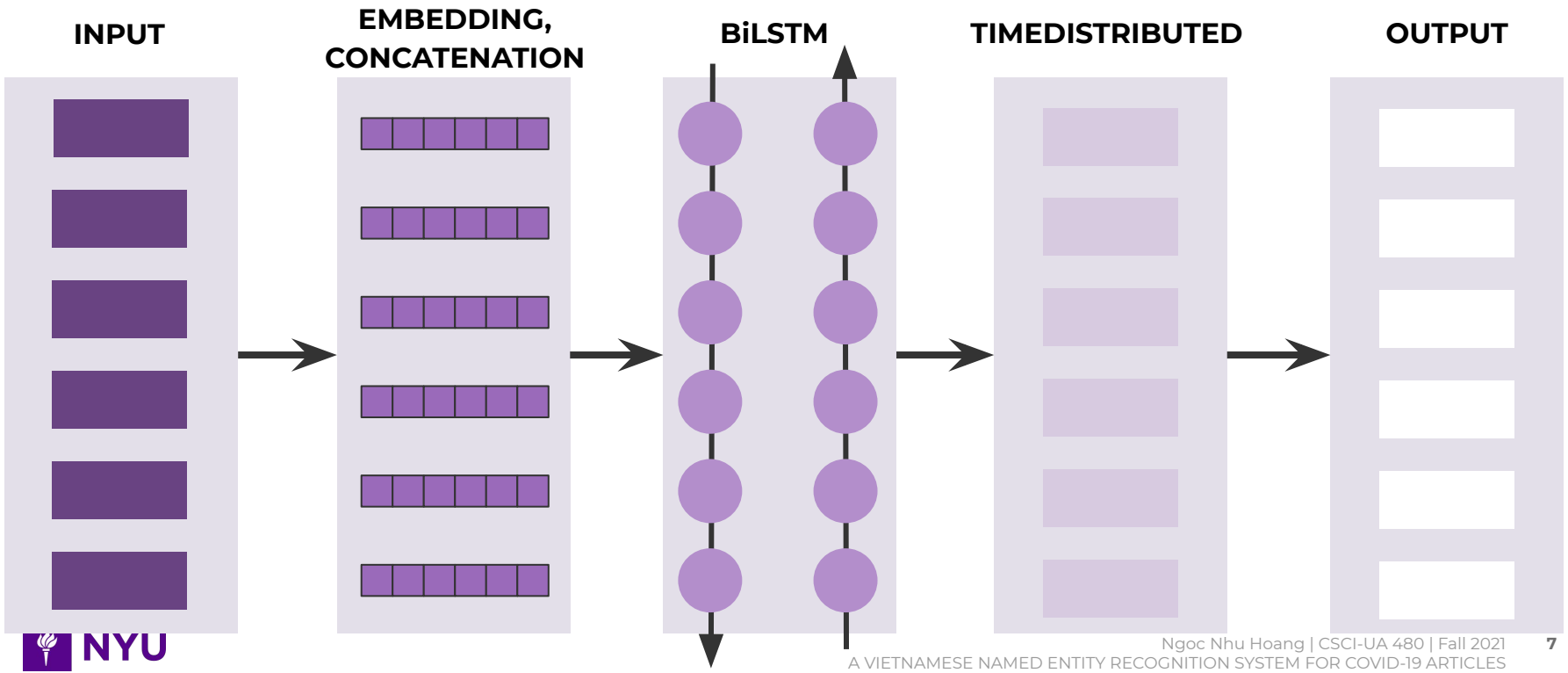
Stacked LSTMs that are both connected to input layer to process data in forward and backward directions



TimeDistributed

Wrapper that helps apply a layer (in this case a Dense layer) to every slice of an input sequence to produce an output sequence

Neural network

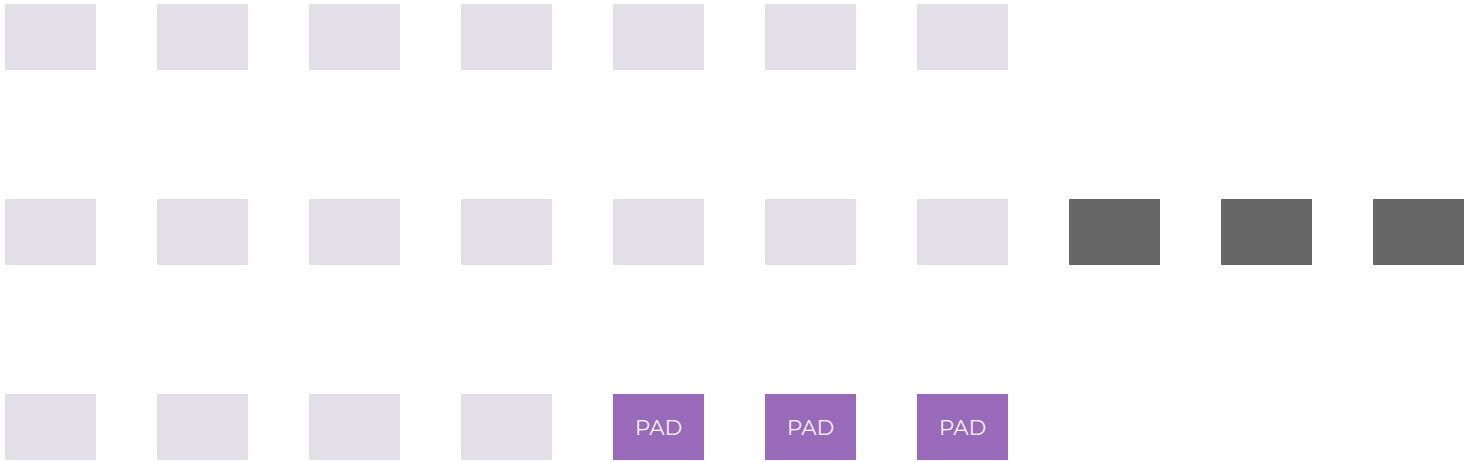


Input data

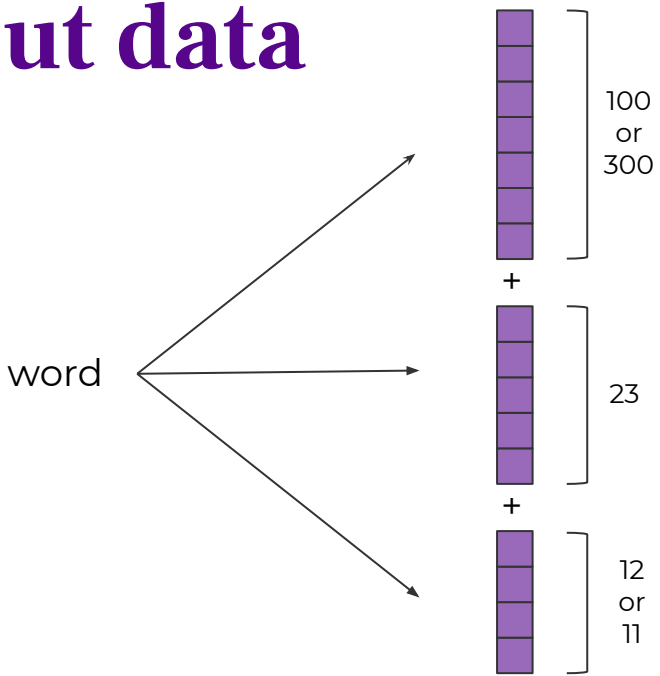
Standard:
70 words/
90 syllables

Longer:
cut off

Shorter:
padded



Input data



1. WORD EMBEDDINGS

Created a vocabulary list from all unique words
Used fastText to build embedding matrix, row *i* contains a vector representation of the word with index *i* in the vocabulary list

2. PART-OF-SPEECH TAGS

Used VnCoreNLP toolkit to obtain POS tags
Turned each tag into a vector using one-hot encoding

3. MANUALLY EXTRACTED FEATURES

Extracted from the words a number of features: isLower, isAllCap, isTitle, isNoun, etc.
Also features based on custom word lists (common last names, common jobs)

PART 02

RESULTS

Evaluation method

Results

Discussion of results

Evaluation method

Input sentences are mostly padded up or cut off, so model predictions are stripped of all padded positions to leave only predicted labels of real words.

Evaluation metrics: precision, recall, F score.

Predictions are aggregated by entity types (e.g. B-DATE & I-DATE → DATE).

Results - Words

Models	Unweighted			Weighted			Correct boundaries
	Precision	Recall	F	Precision	Recall	F	
300D word embeddings	93.51%	85.19%	88.79%	**95.62%**	90.83%	93.06%	90.59%
300D word embeddings + POS	**93.53%**	87.10%	89.94%	95.50%	91.68%	93.46%	92.08%
300D word embeddings + POS + features	93.42%	**88.02%**	**90.41%**	95.23%	**92.24%**	**93.64%**	**92.34%**
100D word embeddings	91.74%	84.54%	87.66%	94.80%	90.43%	92.46%	90.34%
100D word embeddings + POS	91.29%	86.81%	88.76%	94.68%	91.08%	92.78%	91.27%
100D word embeddings + POS + features	92.28%	87.15%	89.46%	94.66%	91.77%	93.13%	92.03%
BiL-CNN-CRF			87.5%			91%	
PhoBERT base			92%			94.2%	
PhoBERT large			93.1%			94.5%	

Results - Syllables

Models	Unweighted			Weighted			Correct boundaries
	Precision	Recall	F	Precision	Recall	F	
300D word embeddings	93.07%	85.77%	89%	95.08%	91.76%	93.30%	91.15%
300D word embeddings + POS	92.62%	86.91%	89.54%	95.12%	91.09%	93.52%	91.55%
300D word embeddings + POS + features	**93.95%**	**87.67%**	**90.45%**	**95.28%**	**92.27%**	**93.67%**	**92.30%**
100D word embeddings	92.29%	85.59%	88.51%	94.60%	91.22%	92.79%	90.41%
100D word embeddings + POS	91.59%	86.24%	88.67%	94.44%	91.34%	92.81%	90.84%
100D word embeddings + POS + features	93.38%	86.63%	89.65%	94.81%	91.79%	92.24%	91.64%
BiL-CNN-CRF			85.8%			90.6%	
XLM-R base			87.9%			92.5%	
XLM-R large			91.1%			93.8%	

Discussions

Discussion of results

Higher-dimensional word embeddings perform better than low-dimensional word embeddings.

Adding POS tags & manual features improve overall F score by 1-2%, higher in specific entity types:

- JOB: 57% → 63%
- NAME: 82% → 87% → 91%

Simple neural network model & features trained on specific domain gives results comparable to more complicated models (BiLSTM-CNN-CRF, fined tuned & pre-trained language models).

Future work & extension

Extend to bigram, trigram.

Downstream tasks: automatic contact tracing through news articles, relationship extraction.

Selected references

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. **COVID-19 named entity recognition for Vietnamese**. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2146–2153, Online. Association for Computational Linguistics.

Thai-Hoang Pham and Phuong Le-Hong. 2017. **The importance of automatic syntactic features in Vietnamese named entity recognition**. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation, pages 97–103. The National University (Philippines).

Thai-Hoang Pham and Phuong Le-Hong. 2017. **End-to-end recurrent neural network models for Vietnamese named entity recognition: Word-level vs. character- level**. CoRR, abs/1705.04044.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. **VnCoreNLP: A Vietnamese natural language processing toolkit**. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.


NYU
Thank you.