IN5550: Neural Methods in Natural Language Processing
Sub-lecture 7.3
RNN practicalities

Andrey Kutuzov

University of Oslo



Contents



1. Sequence labeling
2. Natural language inference
3. RNN tips and tricks

Sequence labeling



- Token-level class assignments in sequential context, aka tagging
- e.g. phoneme sequences, parts of speech; chunks, named entities, etc.

Sequence labeling



- Token-level class assignments in sequential context, aka tagging
- e.g. phoneme sequences, parts of speech; chunks, named entities, etc.
- some structure transcending individual tokens can be approximated

Sequence labeling



- Token-level class assignments in sequential context, aka tagging
- e.g. phoneme sequences, parts of speech; chunks, named entities, etc.
- some structure transcending individual tokens can be approximated

Michelle	*Obama*	*visits*	*UiO*	*today*	*.*

Sequence labeling



- Token-level class assignments in sequential context, aka tagging
- e.g. phoneme sequences, parts of speech; chunks, named entities, etc.
- some structure transcending individual tokens can be approximated

Michelle	*Obama*	*visits*	*UiO*	*today*	.
NNP	NNP	VBZ	NNP	RB	.

Sequence labeling



- Token-level class assignments in sequential context, aka tagging
- e.g. phoneme sequences, parts of speech; chunks, named entities, etc.
- some structure transcending individual tokens can be approximated

Michelle	*Obama*	*visits*	*UiO*	*today*	.
NNP	NNP	VBZ	NNP	RB	.
PERS			ORG		

Sequence labeling



- Token-level class assignments in sequential context, aka tagging
- e.g. phoneme sequences, parts of speech; chunks, named entities, etc.
- some structure transcending individual tokens can be approximated

Michelle	*Obama*	*visits*	*UiO*	*today*	.
NNP	NNP	VBZ	NNP	RB	.
PERS			ORG		
PERS	PERS	—	ORG	—	—

Sequence labeling



- Token-level class assignments in sequential context, aka tagging
- e.g. phoneme sequences, parts of speech; chunks, named entities, etc.
- some structure transcending individual tokens can be approximated

Michelle	*Obama*	*visits*	*UiO*	*today*	.
NNP	NNP	VBZ	NNP	RB	.
PERS			ORG		
PERS	PERS	—	ORG	—	—
B_{PERS}	I_{PERS}	O	B_{ORG}	O	O

Sequence labeling



- Token-level class assignments in sequential context, aka tagging
- e.g. phoneme sequences, parts of speech; chunks, named entities, etc.
- some structure transcending individual tokens can be approximated

Michelle	*Obama*	*visits*	*UiO*	*today*	.
NNP	NNP	VBZ	NNP	RB	.
PERS			ORG		
PERS	PERS	—	ORG	—	—
B_{PERS}	I_{PERS}	O	B_{ORG}	O	O
B_{PERS}	E_{PERS}	O	S_{ORG}	O	O

Sequence labeling



- Token-level class assignments in sequential context, aka tagging
- e.g. phoneme sequences, parts of speech; chunks, named entities, etc.
- some structure transcending individual tokens can be approximated

Michelle	*Obama*	*visits*	*UiO*	*today*	.
NNP	NNP	VBZ	NNP	RB	.
PERS			ORG		
PERS	PERS	—	ORG	—	—
B_{PERS}	I_{PERS}	O	B_{ORG}	O	O
B_{PERS}	E_{PERS}	O	S_{ORG}	O	O

- IOB (aka BIO) labeling scheme — and variants — encodes chunkings.

Two Definitions of 'Sequence Labeling'



- Hmm, actually, what exactly does one mean by sequence labeling?

Two Definitions of 'Sequence Labeling'



- Hmm, actually, what exactly does one mean by sequence labeling?
- gentle definition: class predictions for all elements, in context;
  - pointwise classification; each individual decision is independent;
  - no (direct) model of wellformedness conditions on class sequence;

Two Definitions of 'Sequence Labeling'



- Hmm, actually, what exactly does one mean by sequence labeling?
- gentle definition: class predictions for all elements, in context;
  - pointwise classification; each individual decision is independent;
  - no (direct) model of wellformedness conditions on class sequence;
- strict definition: sequence labeling performs structured prediction;
  - search for 'globally' optimal solution, e.g. most probable sequence;
  - models (properties of) output sequence explicitly, e.g. class bi-grams;
  - later time points impact earlier choices, i.e. revision of path prefix;

Two Definitions of 'Sequence Labeling'



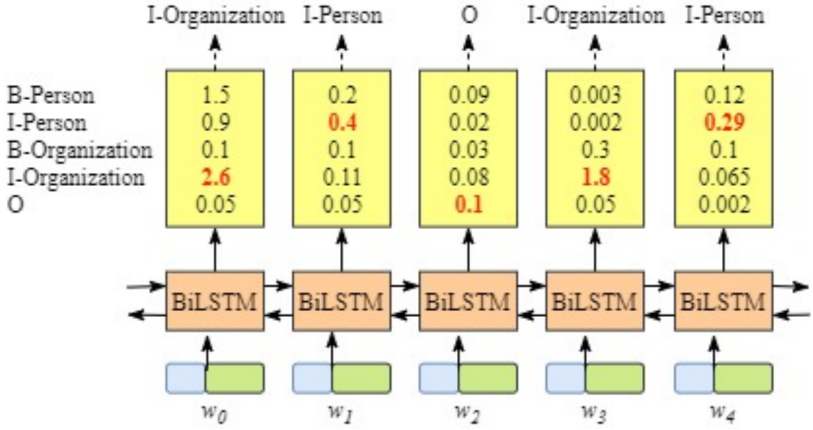
- Hmm, actually, what exactly does one mean by sequence labeling?
- gentle definition: class predictions for all elements, in context;
  - pointwise classification; each individual decision is independent;
  - no (direct) model of wellformedness conditions on class sequence;
- strict definition: sequence labeling performs structured prediction;
  - search for 'globally' optimal solution, e.g. most probable sequence;
  - models (properties of) output sequence explicitly, e.g. class bi-grams;
  - later time points impact earlier choices, i.e. revision of path prefix;
  - search techniques: dynamic programming, conditional random fields (CRF), beam search, re-ranking, etc

Two Definitions of 'Sequence Labeling'



- Hmm, actually, what exactly does one mean by sequence labeling?
- gentle definition: class predictions for all elements, in context;
  - pointwise classification; each individual decision is independent;
  - no (direct) model of wellformedness conditions on class sequence;
- strict definition: sequence labeling performs structured prediction;
  - search for 'globally' optimal solution, e.g. most probable sequence;
  - models (properties of) output sequence explicitly, e.g. class bi-grams;
  - later time points impact earlier choices, i.e. revision of path prefix;
  - search techniques: dynamic programming, conditional random fields (CRF), beam search, re-ranking, etc
- Will do part of speech tagging (sequence labeling) with RNNs next week at the group session.

Sequence-Level Output Constraints



Question: What constraint is broken here?

Contents



1 Sequence labeling

2 Natural language inference

3 RNN tips and tricks

Applications: Natural Language Inference (NLI)

[Bowman et al., 2015]

A large annotated corpus for learning natural language inference

Samuel R. Bowman[*†] sbowman@stanford.edu

Gabor Angeli[†‡] angeli@stanford.edu

Christopher Potts[*] cgpotts@stanford.edu

Christopher D. Manning[*†‡] manning@stanford.edu

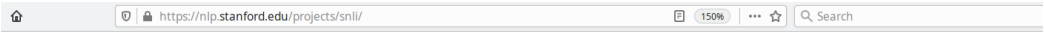
[*]Stanford Linguistics [†]Stanford NLP Group [‡]Stanford Computer Science

Abstract

Understanding entailment and contradiction is fundamental to understanding natural language, and inference about entailment and contradiction is a valuable testing ground for the development of semantic representations. However, machine learning research in this area has been dramatically limited by the lack of large-scale resources. To address this, we introduce the Stanford Natural Language Inference corpus, a new, freely available collection

for approaches employing *distributed* word and phrase representations. Distributed representations excel at capturing relations based in similarity, and have proven effective at modeling simple dimensions of meaning like evaluative sentiment (e.g., Socher et al. 2013), but it is less clear that they can be trained to support the full range of logical and commonsense inferences required for NLI (Bowman et al., 2015; Weston et al., 2015b; Weston et al., 2015a). In a SemEval 2014 task aimed at evaluating distributed representations for NLI, the best-performing systems relied heavily

Applications: Natural Language Inference (NLI)



Here are a few example pairs taken from the development portion of the corpus. Each has the judgments of five mechanical turk workers and a

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

Applications: Natural Language Inference (NLI)



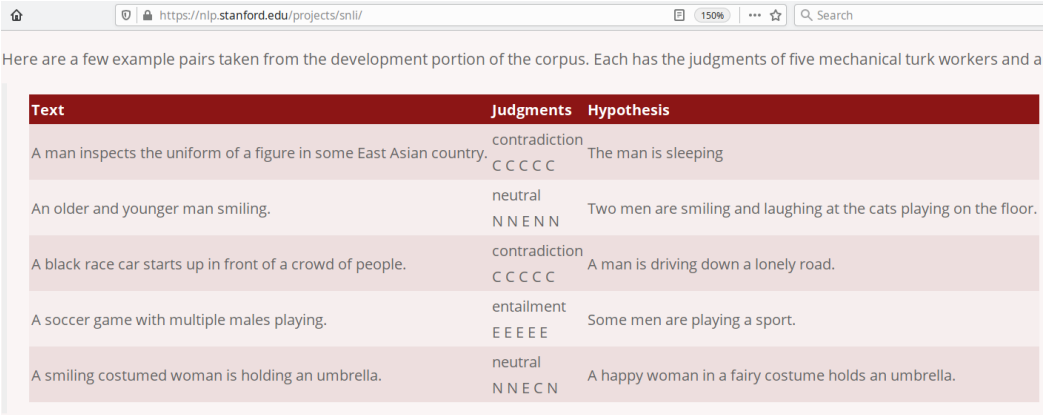
https://nlp.stanford.edu/projects/snli/ 150% Search

Here are a few example pairs taken from the development portion of the corpus. Each has the judgments of five mechanical turk workers and a

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

- 570,000 sentence pairs from image descriptions; three labels: entailment (E), neutral (N), contradiction (C)
- five crowd-sourced annotations; 'consensus'; first enabler of neural NLI;

Applications: Natural Language Inference (NLI)



https://nlp.stanford.edu/projects/snli/ 150% Search

Here are a few example pairs taken from the development portion of the corpus. Each has the judgments of five mechanical turk workers and a

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

- 570,000 sentence pairs from image descriptions; three labels: entailment (E), neutral (N), contradiction (C)
- five crowd-sourced annotations; 'consensus'; first enabler of neural NLI;
- approach via three-way classification over sentence representations?

SNLI: Four Years of Intense Engineering



https://nlp.stanford.edu/projects/snli/

- Successfully established as community benchmark task (for a while);

SNLI: Four Years of Intense Engineering



https://nlp.stanford.edu/projects/snli/

- Successfully established as community benchmark task (for a while);
- [Bowman et al., 2015]: 50.4 % accuracy; [Wang et al., 2021]: 93.1 %;
- from millions of parameters to hundreds of millions;
- added RNNs (conditional) attention, transformers, contextualized embeddings, ensembles...

SNLI: Four Years of Intense Engineering



https://nlp.stanford.edu/projects/snli/

- Successfully established as community benchmark task (for a while);
- [Bowman et al., 2015]: 50.4 % accuracy; [Wang et al., 2021]: 93.1 %;
- from millions of parameters to hundreds of millions;
- added RNNs (conditional) attention, transformers, contextualized embeddings, ensembles...
- at times considered 'too easy';
- follow-up data sets in the SuperGLUE benchmark [Wang et al., 2019] in a similar spirit.
- You are dealing with natural language inference in the Obligatory assignment 2.

SNLI: Four Years of Intense Engineering



https://nlp.stanford.edu/projects/snli/

- Successfully established as community benchmark task (for a while);
- [Bowman et al., 2015]: 50.4 % accuracy; [Wang et al., 2021]: 93.1 %;
- from millions of parameters to hundreds of millions;
- added RNNs (conditional) attention, transformers, contextualized embeddings, ensembles...
- at times considered 'too easy';
- follow-up data sets in the SuperGLUE benchmark [Wang et al., 2019] in a similar spirit.
- You are dealing with natural language inference in the Obligatory assignment 2.

More applications of RNNs in [Goldberg, 2017] and in Jurafsky&Martin online book (see the course schedule for the links).

Contents



1. Sequence labeling
2. Natural language inference
3. RNN tips and tricks

RNN tips and tricks



Variable Length of Input Sequence

- Although RNNs in principle are well-defined for inputs of variable length,
- in practice, padding to fixed length is required for efficiency (batching);
- actually not too much 'waste'; but can be beneficial to bin by length.

RNN tips and tricks

Variable Length of Input Sequence

- Although RNNs in principle are well-defined for inputs of variable length,
- in practice, padding to fixed length is required for efficiency (batching);
- actually not too much 'waste'; but can be beneficial to bin by length.

Evaluation

- Accuracy is a common metric for tagging; fixed number of predictions;
- for most inputs, very large proportion of padding tokens (and labels);
- trivial predictions will inflate accuracy scores; detrimental to learning?
- Can define custom accuracy functions; mask padding tokens predictions.

RNN tips and tricks



What to use for classification?

- The last hidden state?

RNN tips and tricks



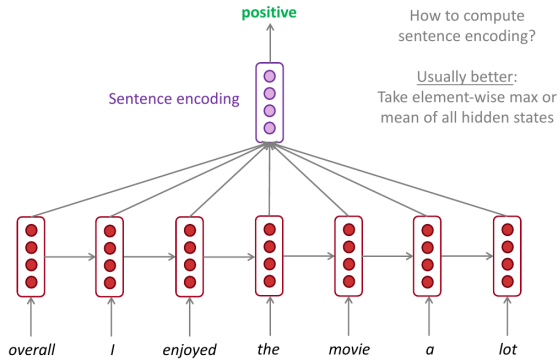
What to use for classification?

- The last hidden state?
- The sum or the average of each element states?

RNN tips and tricks

What to use for classification?

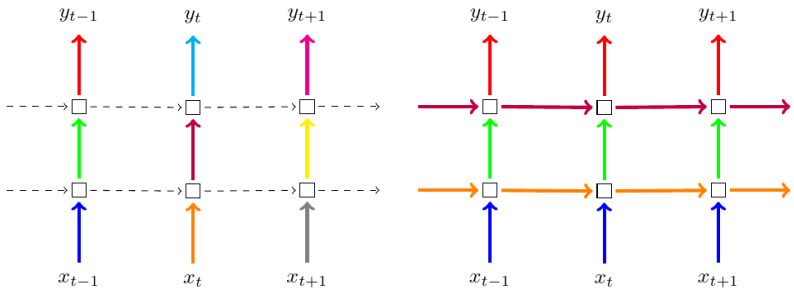
- The last hidden state?
- The sum or the average of each element states?
- Max pooling?



RNN tips and tricks

Dropout in RNNs

- Dropout along memory updates can inhibit learning of effective gating;
- only apply dropout 'vertically'; or fix random mask: 'variational RNN'

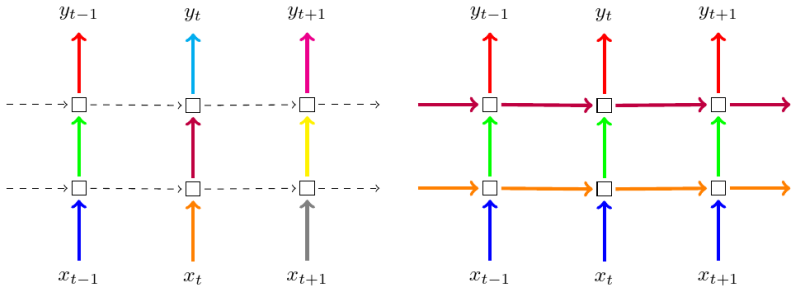


[Gal and Ghahramani, 2016]

RNN tips and tricks

Dropout in RNNs

- Dropout along memory updates can inhibit learning of effective gating;
- only apply dropout 'vertically'; or fix random mask: 'variational RNN'



[Gal and Ghahramani, 2016]

Implemented in *TensorFlow/Keras*, but not in *PyTorch*. Wanna try adding?

References I

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015).
A large annotated corpus for learning natural language inference.
In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Gal, Y. and Ghahramani, Z. (2016).
A theoretically grounded application of dropout in recurrent neural networks.
Advances in neural information processing systems, 29.

Goldberg, Y. (2017).
Neural network methods for natural language processing.
Synthesis Lectures on Human Language Technologies, 10(1):1–309.

References II

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019).
SuperGLUE: A stickier benchmark for general-purpose language understanding systems.
In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Wang, S., Fang, H., Khabsa, M., Mao, H., and Ma, H. (2021).
Entailment as few-shot learner.
arXiv preprint arXiv:2104.14690.