Learning Entailment-based Sentence Embeddings from Natural Language Inference

Rabeeh Karimi[1,2], Florian Mai[1,2], James Henderson[1]

1. Idiap Research Institute

2. École Polytechnique Fédérale de Lausanne (EPFL)

13 November, 2019

Why Model Entailment?

"Public health insurance is less costly than private insurance to the overall economy"

$\Rightarrow$ *"Public healthcare is less expensive"*

Entailment is a powerful semantic relation

- information inclusion: $y \Rightarrow x$ iff everything known given x is also known given y
- abstraction: $y \Rightarrow x$ means x is a description of y which may abstract away from some details
- foundation of the formal semantics of language

Why Model Textual Entailment?

"Public health insurance is less costly than private insurance to the overall economy"

$\Rightarrow$ *"Public healthcare is less expensive"*

Textual Entailment has a wide variety of applications

- Machine translation evaluation
- Identifying similar sentences in corpora
- Zero-shot text classification
- Used other tasks (Question answering, Dialogue systems, summarisation)

Outline

Outline

Motivation

Natural Language Inference

Entailment-based Sentence Embeddings

Empirical Results

Natural Language Inference

Natural Language Inference (NLI) data:

Given premise and hypothesis sentences, classify their relationship into **entailment**, **contradiction**, and **neutral**.

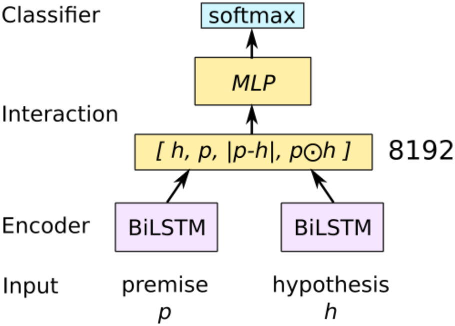
Premise	Two dogs are running through a field.
Entailment	There are animals outdoors.
Contradiction	The pets are sitting on a couch.
Neutral	Some puppies are running to catch a stick.

Natural Language Inference

NLI systems typically have three stages

- Encoder: encode each sentence as a vector
- Interaction: model the interaction between the sentences
- Classifier: apply a softmax classifier

We want to train sentence embeddings on NLI, so we focus on the **Interaction** stage



Interaction Stage

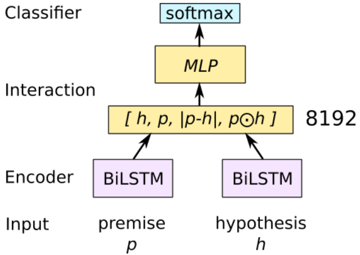
- **Previous methods** mostly model interaction using *heuristic matching features* [2]:

$$m = [\, p;\ h;\ |p - h|;\ p \odot h \,]$$

followed by an MLP:

$$\mathrm{tanh}(W_e m + b_e)$$

where $W_e \in \mathbb{R}^{n \times 4d}$, $b_e \in \mathbb{R}^n$, and n is the size of the hidden layer. The number of parameters (W_e) can be large.

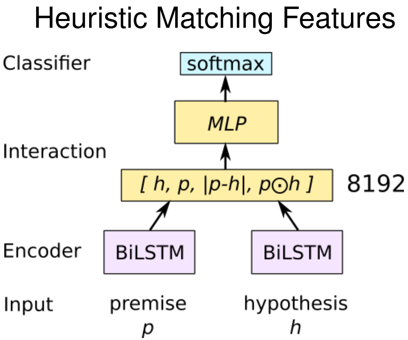


- **Problem:** Most of the information relevant to entailment is modelled in the MLP!

Outline

Learning Entailment-Based Sentence Embeddings

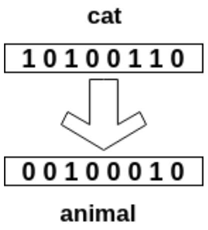
- Learn sentence embeddings with an entailment interpretation
- Force all the information about entailment into the sentence embeddings
- Give a useful inductive bias for textual entailment



Entailment Vectors Framework (Henderson and Popa 2016) [1]

Represent information inclusion per-bit

- A entails B $\Leftrightarrow$ Everything known about B is also known about A
- 1 = known,   0 = unknown
- $P(y \Rightarrow x) = \prod_{k=1}^{d} (1 - P(y_k{=}0)P(x_k{=}1))$
- Given $P(x_k{=}1) = \sigma(X_k)$ and $P(y_k{=}1) = \sigma(Y_k)$:



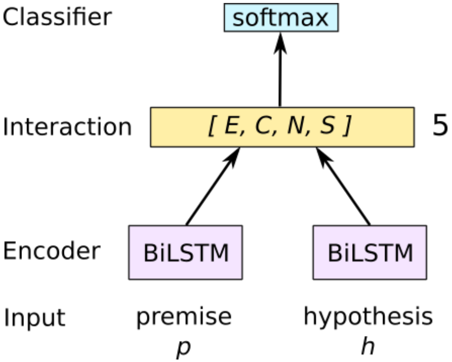
$$Y \tilde{\Rightarrow} X \;=\; \log(\prod_{k=1}^{d} 1 - \sigma(-Y_k)\sigma(X_k)) \;\approx\; \log P(y \Rightarrow x | X, Y)$$

Entailment Vector Model of NLI

Interaction model is 5 scores

- Entailment score
- Contradiction score
- Neutral score
- 2 Similarity scores

with **no parameters**



Entailment Score

We compute the *entailment score* between two sentences using entailment operator ($Y \tilde{\Rightarrow} X$) proposed in [1]:

$$S(\text{entail}|X, Y) = \log(\prod_{k=1}^{d} 1 - \sigma(-Y_k)\sigma(X_k)).$$

Contradiction Score

- Split vector in two halves, one for known-to-be-true and one for known-to-be-false
- Each dimension $k \in [1, \frac{d}{2}]$ contradicts the associated dimension $k + \frac{d}{2}$ in the other half

$$\begin{aligned} S_k(\text{contradict}|X, Y) = {} & \sigma(X_k)\sigma(Y_{k+\frac{d}{2}}) + \sigma(X_{k+\frac{d}{2}})\sigma(Y_k) \\ & - \sigma(X_k)\sigma(Y_{k+\frac{d}{2}})\sigma(X_{k+\frac{d}{2}})\sigma(Y_k) \end{aligned}$$

- Sentences contradict if any dimension contradicts

$$S(\text{contradict}|X, Y) = 1 - \prod_{k=1}^{\frac{d}{2}} (1 - S_k(\text{contradict}|X, Y))$$

Neutral Score

We define a neutral score as the non-negative complement of the contradiction and entailment scores:

$$S(\text{neutral}|X, Y) = \text{ReLU}(1 - S(\text{entail}|X, Y) - S(\text{contradict}|X, Y)).$$

- The ReLU function avoids negative scores.
- Its nonlinearity makes this score non-redundant in the log-linear softmax classifier.

Similarity Scores

We employ two similarity scores measured in the probability space:

- Resembling the element-wise multiplication $p \odot h$, we use the average element-wise multiplication:
$$sim_{mul}(X, Y) = \frac{1}{d} \sum_{k=1}^{d} (\sigma(X_k)\sigma(Y_k)).$$
- Resembling the absolute difference $|p - h|$, we compute the average absolute difference:
$$sim_{diff}(X, Y) = \frac{1}{d} \sum_{k=1}^{d} (|\sigma(X_k) - \sigma(Y_k)|).$$

Outline

Motivation

Natural Language Inference

Entailment-based Sentence Embeddings

Empirical Results

Baselines

- *HM*: heuristic matching features + MLP.
- p, h: only sentence embeddings + MLP.
- *Random*: random nonlinear projection of p, h + MLP, defined as:

$$r = \sigma(W_g \sigma(W_i [p, h] + b_i) + b_g),$$

where the weight matrices $W_i \in \mathbb{R}^{d \times 2d}$, $W_g \in \mathbb{R}^{5 \times d}$ and biases are randomly generated

Experimental Results

Model	**#enc**	**#mlp**	**SNLI**	**MNLI**
Random	3.3m	18	79.07	65.88/65.91
p,h	3.3m	1.3m	78.70	65.69/64.7
HM	3.3m	2.4m	84.82	71.46/71.23
Ours	3.3m	18	83.47	70.51/69.97
HM+attn	13.8m	2.4m	86.46	74.81/74.81
Ours+attn	13.8m	18	86.28	74.41/74.21

- Our interaction layer performs almost as well as MLP-based models (HM) while being simpler and parameter-free.

Ablation Results

Used scores	SNLI	MNLI
E, C, N, S	83.47	70.51/69.97
E, C, N	83.14	69.97/69.19
E, C	78.02	69.66/69.49
S	75.48	63.31/63.03
E	78.62	63.92/63.57
C	74.7	58.96/58.19

- Most of the work is being done by the Entailment and Contradiction scores

Ablation Results

- Trained weights of the final classification layer (E,C,N model):

$$W_c = \begin{array}{c} \\ E \\ N \\ C \end{array} \begin{array}{c} \begin{array}{ccc} S_E & S_N & S_C \end{array} \\ \begin{pmatrix} \boxed{+41.3} & +0.2 & -24.0 \\ -10.8 & -3.3 & -35.0 \\ -29.5 & +4.1 & \boxed{+60.0} \end{pmatrix} \end{array}, \quad b_c = \begin{pmatrix} -26.4 \\ +21.0 \\ +5.3 \end{pmatrix}$$

- **Large weights** in the first and last columns indicate that indeed the entailment score predicts entailment and the contradiction score predicts contradiction.

Transfer Performance to Other NLI datasets

Target Test Dataset	Methods		
	Baseline	Ours	Δ Ours
RTE	48.38	64.98	**+16.6**
JOCI	41.14	45.58	**+4.44**
SCITAIL	68.02	71.59	**+3.57**
SPR	50.84	53.74	**+2.9**
QQP	68.8	69.7	**+0.9**
DPR	49.95	49.95	**0**
FN+	43.04	42.81	**-0.23**
SICK	56.57	54.03	**-2.54**
MPE	48.1	41.0	**-7.10**
ADD-ONE-RTE	29.2	17.05	-12.15
SNLI	64.96	54.14	-10.82

- Thanks to its inductive bias, our model transfers better from MNLI to other datasets with different annotation biases

Transfer Results in Downstream Tasks

Model	MR	CR	MPQA	SUBJ	SST2	SST5	TREC	STS-B
Ours	**84.76**	**90.57**	**89.88**	**93.57**	**90.50**	**49.14**	82.6	0.6511
HM	80.27	88.77	88.07	90.74	86.44	46.56	**83.0**	**0.6574**

SentEval evaluations of sentence embeddings on different sentence classification tasks with logistic regression

Model	STS12	STS13	STS14	STS15	STS16
Ours	**0.6125**	**0.6058**	**0.6618**	**0.6685**	**0.6740**
HM	0.5339	0.5065	0.6289	**0.6653**	0.6351

Correlation between the cosine similarity of sentence embeddings and the gold labels for Textual Similarity (STS)

- Our sentence embeddings transfer better to other tasks

Conclusion

- Proposed entailment and contradiction scores are effective for modelling textual entailment.
- Improved transfer performance in both downstream task and other NLI datasets.
- This parameter-free model puts all textual entailment information in the learned sentence embeddings with a direct entailment-based interpretation.

Thank you! Questions?

References I

James Henderson and Diana Nicoleta Popa. "A Vector Space for Distributional Semantics for Entailment". In: *ACL*. The Association for Computer Linguistics, 2016.

Lili Mou et al. "Natural Language Inference by Tree-Based Convolution and Heuristic Matching". In: *ACL*. 2016.

References