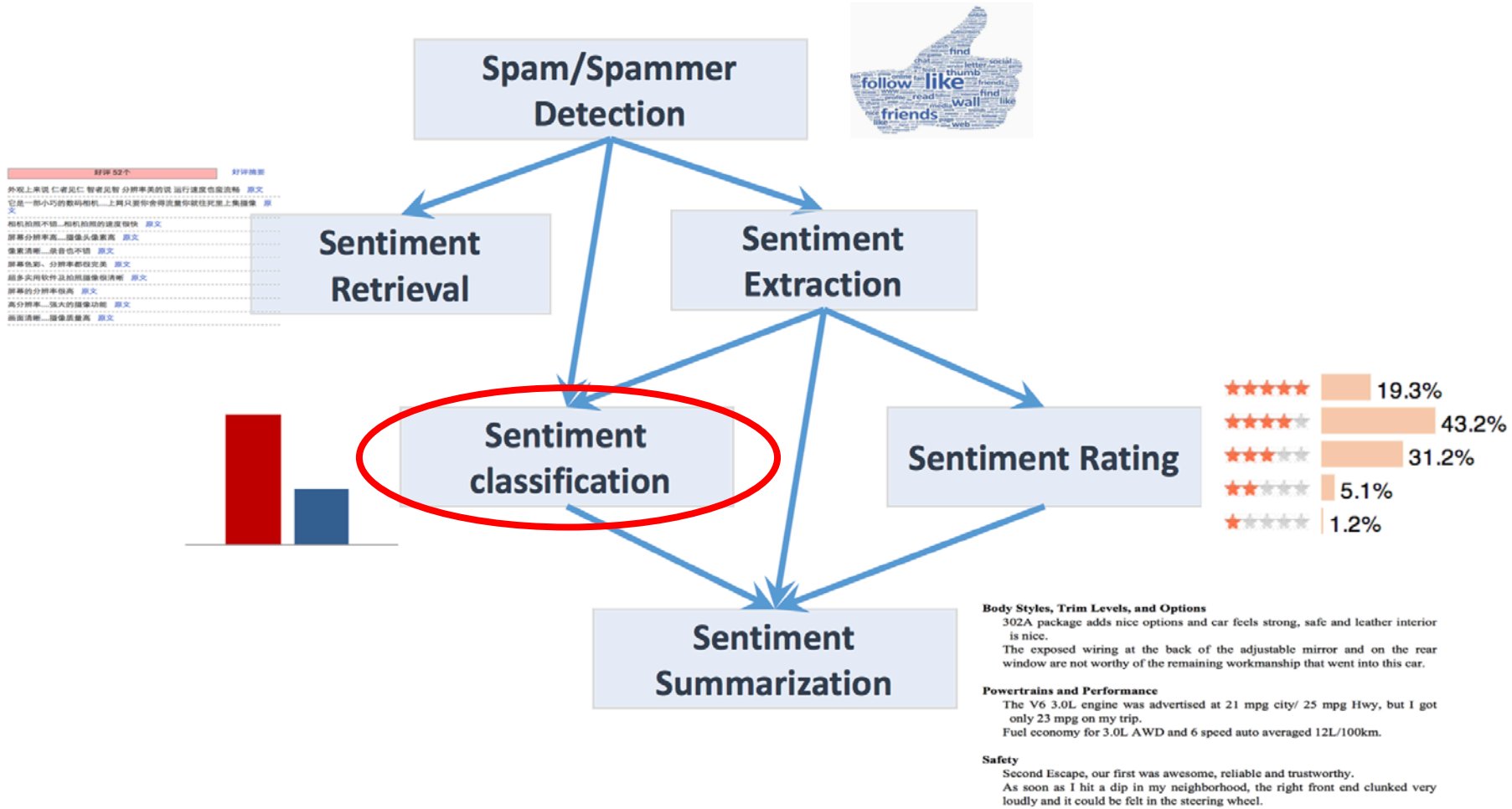
Linguistically Regularized LSTM for Sentiment Classification

September 2017

Qiao Qian, Minlie Huang, Xiaoyan Zhu

Computer Science and Technology / Tsinghua University

Background



Background

Traditional, Fixed Representation

[1/0, 1/0, ···, 1/0]

- [tf*idf, ···, tf*dif]
- [#w_1, #w_2, ···, #w_3]

New Feature Representation

Trainable, Learnable

- High-dimensional, sparse
- Heavy domain expertise
- Expensive engineering

- Low-dimensional, dense
- Data and model driven
- Task oriented

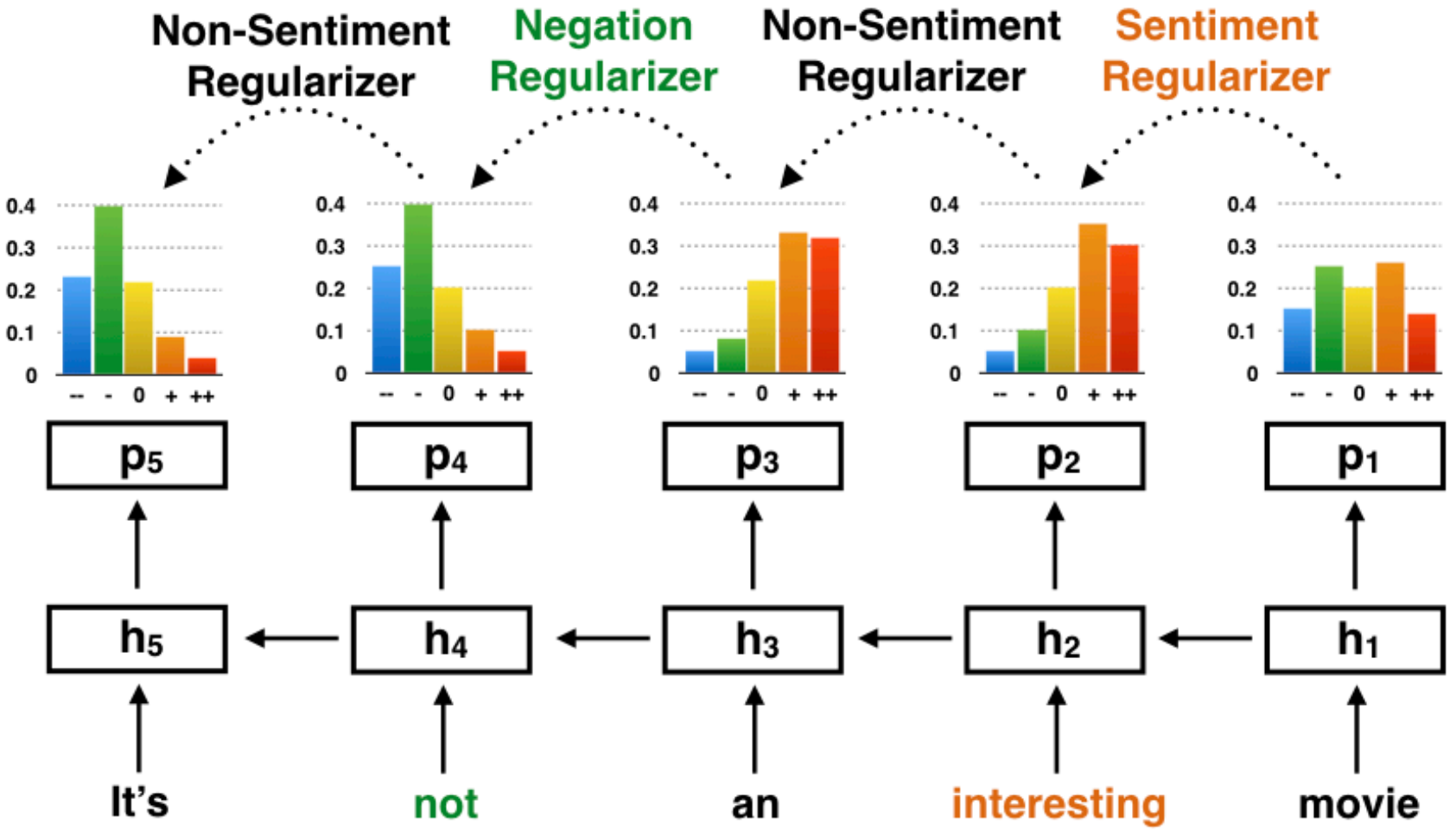
Motivation

- Linguistic resources for sentiment classification
  - Negation: **not, never, cannot**
  - Intensity: **very, absolutely**
  - Sentiment resources: sentiment words like **interesting, wonderful, etc**

This is **not** a **very** **interesting** movie.

How to leverage linguistic recourses in RNN/LSTM?

Model Overview



Experiments

- Dataset
  - Movie Review (pos. / neg.)
  - Sentiment Treebank (fine-grained)
- Linguistic Resource
  - Sentiment Word: 9750 (from MPQA & SST)
  - Negation Word: 36
  - Intensity Word: 44

Dataset	MR	SST
# sentences in total	10,662	11,885
#sen containing sentiment word	10,446	11,211
#sen containing negation word	1,644	1,832
#sen containing intensity word	2,687	2,472

Table 1: The data statistics.

Method	MR	SST Phrase-level	SST Sent.-level
RNN	77.7*	44.8#	43.2*
RNTN	75.9#	45.7*	43.4#
LSTM	77.4#	46.4*	45.6#
Bi-LSTM	79.3#	49.1*	46.5#
Tree-LSTM	80.7#	51.0*	48.1#
CNN	81.5*	48.0*	46.9#
CNN-Tensor	-	51.2*	50.6*
DAN	-	-	47.7*
NCSL	82.9	51.1*	47.1#
LR-Bi-LSTM	82.1	50.6	48.6
LR-LSTM	81.5	50.2	48.2

- *Phrase-level* means the models use phrase level annotation for training.
- *Sent.-level* means the models only use sentence level annotation.

THANK YOU

NExT research is supported by the National Research Foundation,
Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.