Combining Labels with a Multi-View VAE

NAACL-HLT, 2019-06-03

Alexander Hoyle
University College London

Lawrence Wolf-Sonkin
Johns Hopkins University

Ryan Cotterell
University of Cambridge

Isabelle Augenstein
University of Copenhagen

1 Now at University of Maryland. 2 Moving to ETH Zurich.
Problem: annotators’ schemata are different

Roger Ebert
⭐⭐⭐

Jay Sherman
👎
Idea: combine annotations across sources

Roger Ebert
⭐⭐⭐ +

Jay Sherman
👎 −

50%
Aggregation increases coverage and reduces bias

...but can we aggregate consistently?

Roger Ebert
⭐⭐

Roger Ebert
⭐⭐⭐⭐

A. O. Scott
⭐⭐⭐⭐

0.2 0.5
A generative story of reviewing

“True review”

64%

“Noisy Opinions”

Roger Ebert

Jay Sherman
This talk summarized in one slide

Combine disparate annotation schemes into one
...with generative modeling!
Merging Sentiment Lexica with a VAE
Sentiment lexica are also inconsistent

“good”

SentiWordNet
[0.672, 0.0]

Hu-Liu
Positive

MPQA
Positive

General Inquirer
Positive

SenticNet 5
0.849

VADER
[0, 0, 0, 0, 0, 4, 4, 1, 1]
Generating labels is an imperfect process

Suppose there exists a true, unobserved label

“peppy”

0.64 ~ 0.651 \( \sim f \) \( \sim \) 4/5
One latent variable per datapoint

 Where $\rho = \mu$  

Kingma and Welling, 2014
Problem: observations don’t have consistent scales
Solution: lexicon-specific emission distributions

\[ z^w \sim \text{Dir}(\alpha^w) \]

\[ \rho^w_d = f(z^w; \theta_d) \]

\[ x^w_d \sim P_d(x^w_d | \rho^w_d) \]
Individual distributions are schema-dependent

\[ P_d \]

- **SentiWordNet**: \( \mathcal{N}(\rho, 0.01 I) \)
- **MPQA**: \( \text{Bern}(\rho) \)
- **SenticNet 5**: \( \mathcal{N}(\rho_\mu, \rho_\sigma) \)
- **Hu-Liu**: \( \text{Bern}(\rho) \)
- **General Inquirer**: \( \text{Bern}(\rho) \)
- **VADER**: \( \text{Mult}(\rho) \)
Encoder approximates intractable posterior $P(Z|X)$

SenticNet 5

$S_{SN} 0.65 0.24 0.11$

Hu-Liu

$S_{HL} 0.91 0.09 0.00$

$S$ $\beta^w Q_{\beta^w}(z^w|x^w)$

“peppy”
Distribution over latent representation is interpretable

“Superb”
6.00, 1.00, 1.00

“Terrible”
1.00, 1.00, 6.99

“Portuguese”
1.07, 1.89, 1.05
Evaluation
Evaluation task: predict sentence sentiment

x_i

“The movie was good”

y_i

Pos

SenticNet 0.94 0.84 - 0.85 \frac{1}{4} \sum 0.87 \int 1
Our representation outperforms individual lexica
Performance owed to better representation in addition to greater coverage
For what other kinds of annotation schemes could this technique be useful?
Thanks

- Co-authors: Lawrence Wolf-Sonkin, Hanna Wallach, Ryan Cotterell, and Isabelle Augenstein
- University of Copenhagen for travel funding
Fin