Will it blend?
Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining

BlendNet

Weak labeled data (WLD)
- Abundant but noisy, can be obtained by:
  - Applying heuristics to a large corpus
  - Distant supervision
  - Strong labeled data for a related task

Strong labeled data (SLD)
- The gold standard, expensive and complicated to obtain, therefore scarce

The Algorithm

Blending epochs using the entire SLD and a fraction of the WLD determined by a blend factor

Initialization epochs using the entire WLD with no SLD

The task
Corpus-wide topic-dependent evidence detection

Strong labeled data
- 5,785 sentences extracted from Wikipedia
- 10 annotators for each sentence
- 118 topics generally dealing with one identifiable concept from domains such as politics, science and education
- Split by topics into training and test

Get the data set from IBM Debater Datasets webpage

Weak labeled data
- Webis-Debate-16 corpus (Al-Khatib et al., 2016) automatically extracted from idebate.org
  - Positives: 11,000 argumentative sentences from points for/against section
  - Negatives: 5,500 non-argumentative sentences from the introduction section
- Following Levy et al. (2017) for unsupervised topic-dependent claim detection
  - Positives: 63,000 sentences from Wikipedia containing the word that followed by the concept of the topic (not necessarily adjacent)
  - Negatives: 190,000 sentences from Wikipedia containing the concept of the topic

Evaluation

WLD from Webis-Debate-16
- Micro-averaged accuracy on the SLD test set for the different sizes of SLD training data
- (*) indicates significant results in comparison to SLD only, (**) indicates significant results also in comparison to Blend factor 0

WLD following Levy et al. (2017)
- Even though WLD (orange triangle) is not nearly as accurate as SLD (blue curve), it has the potential to improve performance, if blended correctly (green curve), especially when there is not enough SLD.