Global Thread-Level Inference for Comment Classification in Community Question Answering

Shafiq Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti and Preslav Nakov
Qatar Computing Research Institute, HBKU

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

1. Introduction

SemEval-2015 Task 3:
Answer Selection in cQA
http://alt.qcri.org/semeval2015/task3/

- Subtask A: Given a question, classify answers in a thread as
  - *good* vs. *potentially useful* vs. *bad*
- This work**: *good-vs-bad*** classification (i.e,. *good* vs. *rest*).

Our previous work [1]

- conversation-level (global) features along with local (e.g., similarity with the question) features
- univariate & multivariate models, e.g., LR, SVM, CRF, SVM-HMM

This work

- **Thread-level inference using a classifier of comment pairs**

2. Thread-level Inference

Q: I have a female friend who is leaving for a teaching job in Qatar in January. What would be a useful portable gift to give her to take with her?

A_1 A couple of good best-selling novels. It's hard to find much here in Doha in the way of books.
Local: Good, **Human:** Good

A_2 ipod to entertain herself in case of boredom... a lot of patience for her students...
Local: Good, **Human:** Good

A_3 Thanks, please keep suggestions coming, would like to send her off with a useful gift.
Local: Bad **Human:** Bad

A_6 Bacon. Nice bread, bacon, bacon, errmmm bacon and a pork joint..
Local: Bad, **Human:** Good

A_9 Couple of good novels, All time favorite movies, ..
Local: Bad, **Human:** Good

- **Similar comments should get the same label**
- **Relations between comments can be at any distance**

} ***Pairwise classifier***

This research was performed by the Arabic Language Technologies (ALT) group at the Qatar Computing Research Institute (QCRI), HBKU, Qatar Foundation. Such research is part of the Interactive sYstems for Answer Search (Iyas) project, which is developed in collaboration with MIT-CSAIL.

3. Our Solution

- A binary classifier is trained to decide whether a pair of comments in a thread should be in the same class or not.
- The pairwise and the local good-vs-bad classification probabilities are then used in (a) min-cut or (b) ILP models for global inference.

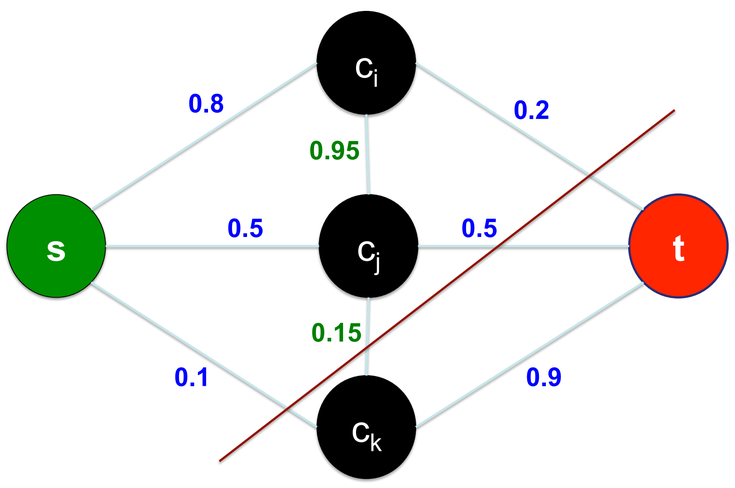
a. Graph Partition

- **Find a partition P = (G,B) minimizing**

$$C(P) = \lambda\Big[\sum_{c_i \in G} s_{iB} + \sum_{c_i \in B} s_{iG}\Big] + (1-\lambda)\sum_{c_i \in G, c_j \in B} s_{ij}$$

- s_{iG} and s_{iB} are individual scores;
- s_{ij} are pairwise scores.

- **Max-flow/ min-cut:** Gives exact solution in polynomial time



b. Linear Programming

- **Find an assignment A to all variables that minimizes**

$$C(A) = \lambda \cdot \sum_{i=1}^{N} (c_{iG} \cdot x_{iG} + c_{iB} \cdot x_{iB}) +$$
$$(1-\lambda) \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (c_{ijS} \cdot x_{ijS} + c_{ijD} \cdot x_{ijD})$$

$$c_{iG} = -\log s_{iG} \quad c_{ijS} = -\log s_{ij}, \text{ etc.}$$

- **Subject to the constraints:**
  - all variables are binary
  - only one label is assigned to each comment
  - the assignments to comments and comment-pairs are consistent

4. Experimental Setup and Results

- **Dataset (Qatar Living)**

Category	Train	Dev	Test
Questions	2,600	300	329
Comments	16,541	1,645	1,976
Good	8,069	875	997
Bad	8,472	770	979

- **Same vs. Different Classification**

Classifier	P	R	F_1	Acc
baseline: *Same*				69.26
MaxEnt-2C	73.95	90.99	81.59	71.56
MaxEnt-3C	77.15	80.42	78.75	69.94

- Same-vs.-different better than 3-way classifier
- Small improvement in accuracy over the baseline
  - yet, the classifier is helpful with graph-cut/ILP

- **Main Results**

System	P	R	F_1	Acc
Top-3 at SemEval-2015 Task 3				
JAIST	80.23	77.73	78.96	79.10
HITSZ-ICRC	75.91	77.13	76.52	76.11
QCRI	74.33	83.05	78.45	76.97
Instance Classifiers				
MaxEnt	75.67	84.33	79.77	78.43
Linear Chain Classifiers				
CRF	74.89	83.45	78.94	77.53
Global Inference Classifiers				
ILP	77.04	83.53	80.15	79.14‡
Graph-cut	78.30	82.93	**80.55**	**79.80**‡
ILP-3C	78.07	80.42	79.23	78.73
Graph-cut-3C	78.26	81.32	79.76	79.19†

Summary

We have shown that using thread-level information in a pairwise classifier + min-cut/ILP improves over the state of the art. Linear-chain CRF model is less helpful.

Future work

(i) Joint models, (ii) exploiting cross-thread information, (iii) use other CQA datasets

References: [1] Thread-level Information for Comment Classification in Community Question Answering. In *ACL-2015*