Joint Learning with Global Inference for Comment Classification in Community Question Answering

Shafiq Joty, Lluís Màrquez and Preslav Nakov
Arabic Language Technology (ALT) Group
Qatar Computing Research Institute - HBKU
What this Talk is About?

- Not about feature engineering
- Not about deep learning
- But, about **joint learning** and **inference**
- Also about **locally vs. globally normalized** models.
The Task: Community Question Answering

**Q**

hello guys and gals..could anyone of u knows where to buy a good and originals RC helicopters and toy guns here in qatar..im longin for this toys but its nowhere to find.. thanks

**A₁**

Did you check with Toys R us? I think I saw it there.

**A₂**

Go to Doha city center you may get it at 4 floor.

**A₃**

``Hobby Shop" in City center has these toys with original motors. They are super cool.. U will love that shop..and will definetly buy one :) Have fun :)

**A₄**

Hobby Shop- City Centre

**A₅**

OMG!! :| Guns and helicopters??!!

**A₆**

Speed Marine- Salwa Road I think these guys r the best ..
Need for Joint Learning and Inference

- Many comments are short.
- Many comments contain similar info.
- Similar comments should get similar labels.
- Similarity with question not enough.
- Classifier does not get enough information when comments are considered separately.
- Need Joint learning & inference to learn to classify collectively.

Q: hello guys and gals..could anyone of u knows where to buy a good and originals RC helicopters and toy guns here in qatar..im longin for this toys but its nowhere to find.. thanks

A_1 Go to Doha city center you may get it at 4 floor.
   Local: Good, Human: Good

A_2 “Hobby Shop” in City center has these toys with original motors. They are super cool.. U will love that shop..and will definitly buy one :) Have fun :)
   Local: Good, Human: Good

A_3 IM selling all my rc nitro helicopters. call me at 5285113.. (1)TREX 600 new/ (1) TREX500 (1) SHUTTLERG (1) FUTABA ... [truncated]
   Local: Good, Human: Bad

A_4 Hobby Shop- City Centre
   Local: Bad, Human: Good

A_5 OMG!! :— Guns and helicopters??!!
   Local: Good, Human: Bad

A_6 Speed Marine- Salwa Road I think these guys r the best in town...
   Local: Good, Human: Good

A_7 City center, i’ve seen wonderful collection.. Its some wer besides the kids fun place..
   Local: Bad, Human: Good

A_8 try the shop in city center. they have many RC toys for sale there. and for the toy guns, in your talking baout airsoft i think its prohibited here. good luck
   Local: Good, Human: Good
Outline

• Motivation
• Three approaches to classification
• Our models
  o Locally normalized Joint model
  o Globally normalized Fully-connected CRF
• Inference with loopy Belief Propagation
• Experiments & error analysis
• Conclusion & future work
Three Approaches to Classification

**Approach 1:** Classify each comment separately

\[ \psi_n(y_i | x_i, v) \]

\[ x \]: feature vector extracted by considering the comment and the question

\[ y \]: class label

Does not model the dependency between comment labels
Three Approaches to Classification

Approach 2:

- Learn two classifiers separately and combine them in Inference
- Works in three steps (Joty et al. 2015, Pang & Lee, 2004):
  
  a) Learn a node-level classifier
  b) Learn an edge-level classifier
  c) Classify collectively using global inference (ILP, Graph-cut)

![Diagram of classification process](image)
Approach 2: Inference with Graph Cut

- Decoupling learning from inference can lead to suboptimal solutions (Punyakanok et al., 2005)
- Often requires a tuning parameter to control the relative weights of the two classifiers in the combination.
Three Approaches to Classification

**Approach 3:** Learn to classify with global inference (our approach)

- **Learn** node-level & edge-level classifiers/potentials from global thread-level feedback given by an inference alg.
- **Classify** collectively with global inference.

Models dependencies between output variables while learning.

Potentials could be normalized **locally** or **globally**
Our Models

Model 1: Learn two local classifiers jointly with global feedback

- **Node-level classifier:**
  \[
  \psi_n(y_i = k | x_i, v) = \frac{\exp(v_k^T x_i)}{\sum_{k'=1}^{K} \exp(v_{k'}^T x_i)}
  \]

- **Edge-level classifier:**
  \[
  \psi_e(y_{i,j} = l | \phi(x_i, x_j), w) = \frac{\exp(w_l^T \phi(x_i, x_j))}{\sum_{l'=1}^{L} \exp(w_{l'}^T \phi(x_i, x_j))}
  \]
**Our Models**

**Model 1:** Learning two local classifiers jointly with global inference

- **Node-level classifier:** \( \psi_n(y_i = k|x_i, v) = \frac{\exp(v_k^T x_i)}{\sum_{k'=1}^{K} \exp(v_{k'}^T x_i)} \)

- **Edge-level classifier:** \( \psi_e(y_{i,j} = l|\phi(x_i, x_j), w) = \frac{\exp(w_l^T \phi(x_i, x_j))}{\sum_{l'=1}^{L} \exp(w_{l'}^T \phi(x_i, x_j))} \)

---

**Algorithm 1:** Joint learning of local classifiers with global thread-level inference

1. Initialize the model parameters \( v \) and \( w \);
2. repeat
   - for each thread \( G = (V, E) \) do
     - a. Compute node and edge probabilities \( \psi_n(y_i|x_i, v) \) and \( \psi_e(y_{i,j}|\phi(x_i, x_j), w) \);
     - b. Infer node and edge marginals \( \beta_n(y_i) \) and \( \beta_e(y_{i,j}) \) using sum-product LBP;
     - c. Update: \( v = v - \frac{n}{|V|} f'(v) \);
     - d. Update: \( w = w - \frac{n}{|E|} f'(w) \);
   - end
3. until convergence;
Limitations of Model 1

• Local normalization leads to label bias problem.

• Local classifiers use their own feature sets, which may not work well when trained with global feedback.
Our Models

Model 2: Learn a joint model with global normalization

- The model: \[ p(y|v, w, x) = \frac{1}{Z(v, w, x)} \prod_{i \in V} \psi_n(y_i|x, v) \cdot \prod_{(i,j) \in E} \psi_e(y_{i,j}|x, w) \]

- Node potential: \[ \psi_n(y_i|x, v) = \exp(v^T \phi(y_i, x)) \]

- Edge potential: \[ \psi_e(y_{i,j}|x, w) = \exp(w^T \phi(y_{i,j}, x)) \]

- Objective: \[ f(\theta) = \sum_{i \in V} v^T \phi(y_i, x) + \sum_{(i,j) \in E} w^T \phi(y_{i,j}, x) - \log Z(v, w, x) \]

Pairwise FCCRF

- Edge potentials:
  a. All possible state transitions \[ \psi_n(y_i|x_i, v) \]
  b. Ising like (Same and Different)
Inference with Belief Propagation

• Belief propagation (Pearl, 1988) is a message passing algorithm for performing inference in probabilistic graphical models.

• Message from a variable node to a factor node

\[
\mu_{v \rightarrow a}(x_v) = \prod_{a^* \in N(v) \setminus \{a\}} \mu_{a^* \rightarrow v}(x_v); \forall x_v \in Dom(v)
\]
Inference with Belief Propagation

Let us formulate the loopy BP for the factor graph

2.1.1 Loopy Belief Propagation for Pairwise Factor Graphs

messages simultaneously (typically) at every iteration. One method of exact

ing must be adjusted slightly compared with the one used for trees.

Although it was originally designed for acyclic graphical models, it was found

we describe an approximation algorithm for such graphs, i.e., the

marginal actually converge to the true marginals in a finite number of iterations.

belonging to one factor is proportional to the product of the factor and the

has cycles or loops, such an optimal scheduling does not exist, and a typical

convergence after computing each message only once. When the factor graph

See Kevin's book Chap 22 for a pseudocode of the algorithm.

In Loopy BP, one initializes all variable messages to 1 and updates all

In the case where the factor graph is acyclic (i.e. tree, forest), these estimated

Upon convergence, the estimated marginal distribution of each node is pro-

• Message from a factor node to a variable node

\[
\mu_{a \rightarrow v}(x_v) = \sum_{x'_a : x'_v = x_v} f_a(x'_a) \prod_{v^* \in N(a) \setminus \{v\}} \mu_{v^* \rightarrow a}(x_{v^*}); \forall x_v \in \text{Dom}(v)
\]

• Upon convergence:

\[
P(x_v) \propto \prod_{a \in N(v)} \mu_{a \rightarrow v}(x_v)
\]

\[
P(x_a) \propto f_a(x_a) \prod_{v \in N(a)} \mu_{v \rightarrow a}(x_v)
\]
Belief Propagation for Pairwise Factors

Message: \[ \mu_{i\rightarrow j}(y_j) = \sum_{y_i} \psi_n(y_i) \psi_e(y_{i,j}) \prod_{k \in N(i) \setminus j} \mu_{k\rightarrow i}(y_i) \]

Node Belief: \[ \beta_n(y_i) \approx \psi_n(y_i) \prod_{j \in N(i)} \mu_{j\rightarrow i}(y_i) \]

Edge Belief: \[ \beta_e(y_{i,j}) \approx \psi_e(y_{i,j}) \times \mu_{i\rightarrow j}(y_i) \times \mu_{j\rightarrow i}(y_j) \]

- BP is guaranteed to converge to an exact solution if the graph is a tree.
- Exact inference is intractable for general graphs (with loops).
- Although LBP gives approximate solutions for general graphs, it often works well in practice (Murphy et al, 1999)
Outline

• Motivation
• Three approaches to classification
• Our models
  o Locally normalized Joint model
  o Globally normalized Fully-connected CRF
• Inference with loopy Belief Propagation
• Experiments & error analysis
• Conclusion & future work
Experimental Settings: Datasets and Metrics

• **Dataset:** SemEval 2015 Task 3: Question-answer threads from Qatar Living

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions</td>
<td>2600</td>
<td>300</td>
<td>329</td>
</tr>
<tr>
<td>Comments</td>
<td>16,541</td>
<td>1645</td>
<td>1976</td>
</tr>
</tbody>
</table>

• **Metrics:**
  - Macro F1
  - Accuracy

• **Significance test:**
  - Appr. Randomization
Experimental Settings: Features


- **Node-level**
  
  **Local features**
  
  **Similarity features**
  - Longest common subsequence
  - Cosine similarity
  - Jaccard coefficient
  - PTK over syntactic trees.
  - .....

  **Heuristic features**
  - URL, email address
  - “yes”, “no”, etc.
  - Thank*, ack*
  - Length
  - .....

- **Global features**
  
  - Position of the comment.
  - # of comments by the same user.
  - Comment appears before a comment by \( u_q \) containing ack, question.
  - Contains a dialogue pattern.
  - .....

- **Edge-level**
  
  - All features from Node classifier
  - Similarity features
  - Good vs. bad predictions
Experimental Settings: Methods Compared

- **Independent comment classification (ICC)**
  - MaxEnt (SGD)
  - Perceptron

- **Learning & Inference (LI)**
  - MaxEnt (SGD)
  - Graph cut
  - Loopy BP

- **Joint Learning & Inference**
  - Joint MaxEnts (SGD)
  - FCCRF (SGD)
  - Graph cut
  - Loopy BP
Main Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Learning</th>
<th>Inference</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>Majority</td>
<td>–</td>
<td>50.5</td>
<td>100.0</td>
<td>67.1</td>
<td>50.5</td>
</tr>
<tr>
<td>II.</td>
<td>ICC&lt;sub&gt;ME&lt;/sub&gt;</td>
<td>Local, SGD</td>
<td>–</td>
<td>75.1</td>
<td>85.8</td>
<td>80.1</td>
</tr>
<tr>
<td></td>
<td>ICC&lt;sub&gt;Perc&lt;/sub&gt;</td>
<td>Local, Voted</td>
<td>–</td>
<td>76.6</td>
<td>82.4</td>
<td>79.4</td>
</tr>
<tr>
<td>III.</td>
<td>LI&lt;sub&gt;ME−GC&lt;/sub&gt;</td>
<td>Local, SGD</td>
<td>Graph-cut</td>
<td>77.4</td>
<td>83.6</td>
<td>80.4</td>
</tr>
<tr>
<td></td>
<td>LI&lt;sub&gt;ME−LBP&lt;/sub&gt;</td>
<td>Local, SGD</td>
<td>LBP</td>
<td>76.4</td>
<td>84.6</td>
<td>80.3</td>
</tr>
<tr>
<td>IV.</td>
<td>Joint&lt;sub&gt;ME−LBP&lt;/sub&gt;</td>
<td>2 classifiers, Joint, SGD</td>
<td>LBP</td>
<td>76.1</td>
<td>84.4</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td>Joint&lt;sub&gt;Perc−LBP&lt;/sub&gt;</td>
<td>2 classifiers, Joint, AVG</td>
<td>LBP</td>
<td>77.1</td>
<td>74.5</td>
<td>75.8</td>
</tr>
<tr>
<td></td>
<td>FCCRF</td>
<td>Joint, SGD</td>
<td>LBP</td>
<td>77.3</td>
<td>86.2</td>
<td>81.5</td>
</tr>
</tbody>
</table>

- Independent comment classification (ICC)
  - MaxEnt (SGD)
  - Perceptron

MaxEnt performs slightly better than voted perceptron
Main Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Learning</th>
<th>Inference</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Majority</td>
<td>–</td>
<td>–</td>
<td>50.5</td>
<td>100.0</td>
<td>67.1</td>
<td>50.5</td>
</tr>
<tr>
<td>II. ICC&lt;sub&gt;ME&lt;/sub&gt;</td>
<td>Local, SGD</td>
<td>–</td>
<td>75.1</td>
<td>85.8</td>
<td>80.1</td>
<td>78.5</td>
</tr>
<tr>
<td>ICC&lt;sub&gt;Perc&lt;/sub&gt;</td>
<td>Local, Voted</td>
<td>–</td>
<td>76.6</td>
<td>82.4</td>
<td>79.4</td>
<td>78.4</td>
</tr>
<tr>
<td>III. LI&lt;sub&gt;ME–GC&lt;/sub&gt;</td>
<td>Local, SGD</td>
<td>Graph-cut</td>
<td>77.4</td>
<td>83.6</td>
<td>80.4</td>
<td>79.4</td>
</tr>
<tr>
<td>LI&lt;sub&gt;ME–LBP&lt;/sub&gt;</td>
<td>Local, SGD</td>
<td>LBP</td>
<td>76.4</td>
<td>84.6</td>
<td>80.3</td>
<td>79.1</td>
</tr>
<tr>
<td>IV. Joint&lt;sub&gt;ME–LBP&lt;/sub&gt;</td>
<td>2 classifiers, Joint, SGD</td>
<td>LBP</td>
<td>76.1</td>
<td>84.4</td>
<td>80.0</td>
<td>78.7</td>
</tr>
<tr>
<td>Joint&lt;sub&gt;Perc–LBP&lt;/sub&gt;</td>
<td>2 classifiers, Joint, AVG</td>
<td>LBP</td>
<td>77.1</td>
<td>74.5</td>
<td>75.8</td>
<td>76.0</td>
</tr>
<tr>
<td>FCCRF</td>
<td>Joint, SGD</td>
<td>LBP</td>
<td>77.3</td>
<td>86.2</td>
<td>81.5</td>
<td>80.5</td>
</tr>
</tbody>
</table>

- Learning & Inference (LI)
  - MaxEnt (SGD)
  - Graph cut (Joty et al, 2015)
  - Loopy BP

Global inference improves over local classifiers, but not significantly (p = 0.09)
## Main Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Learning</th>
<th>Inference</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>Majority</td>
<td>–</td>
<td>50.5</td>
<td>100.0</td>
<td>67.1</td>
<td>50.5</td>
</tr>
<tr>
<td>II.</td>
<td>ICC&lt;sub&gt;ME&lt;/sub&gt;</td>
<td>Local, SGD</td>
<td>–</td>
<td>75.1</td>
<td>85.8</td>
<td>80.1</td>
</tr>
<tr>
<td></td>
<td>ICC&lt;sub&gt;Perc&lt;/sub&gt;</td>
<td>Local, Voted</td>
<td>–</td>
<td>76.6</td>
<td>82.4</td>
<td>79.4</td>
</tr>
<tr>
<td>III.</td>
<td>LI&lt;sub&gt;ME&lt;/sub&gt;-GC</td>
<td>Local, SGD</td>
<td>Graph-cut</td>
<td>77.4</td>
<td>83.6</td>
<td>80.4</td>
</tr>
<tr>
<td></td>
<td>LI&lt;sub&gt;ME&lt;/sub&gt;-LBP</td>
<td>Local, SGD</td>
<td>LBP</td>
<td>76.4</td>
<td>84.6</td>
<td>80.3</td>
</tr>
<tr>
<td>IV.</td>
<td>Joint&lt;sub&gt;ME&lt;/sub&gt;-LBP</td>
<td>2 classifiers, Joint, SGD</td>
<td>LBP</td>
<td>76.1</td>
<td>84.4</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td>Joint&lt;sub&gt;Perc&lt;/sub&gt;-LBP</td>
<td>2 classifiers, Joint, AVG</td>
<td>LBP</td>
<td>77.1</td>
<td>74.5</td>
<td>75.8</td>
</tr>
<tr>
<td></td>
<td>FCCRF</td>
<td>Joint, SGD</td>
<td>LBP</td>
<td>77.3</td>
<td>86.2</td>
<td>81.5</td>
</tr>
</tbody>
</table>

Joint learning with **local** normalization does not work well.

Joint learning with **global** normalization is the best model and significantly better than local models (p = 0.04)
Comparison with State-of-the-art

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaxEnt classifier</td>
<td>75.7</td>
<td>84.3</td>
<td>79.8</td>
<td>78.4</td>
</tr>
<tr>
<td>Linear CRF</td>
<td>74.9</td>
<td>83.5</td>
<td>78.9</td>
<td>77.5</td>
</tr>
<tr>
<td>MaxEnt+ILP</td>
<td>77.0</td>
<td>83.5</td>
<td>80.2</td>
<td>79.1</td>
</tr>
<tr>
<td>MaxEnt+GraphCut</td>
<td>78.3</td>
<td>82.9</td>
<td>80.6</td>
<td>79.8</td>
</tr>
<tr>
<td>Our method (FCCRF)</td>
<td>77.3</td>
<td>86.2</td>
<td>81.5</td>
<td>80.5</td>
</tr>
</tbody>
</table>

Table 1: Results of all compared models on the test set. The best results are boldfaced.

Table 2: Comparison to the best published results on the same datasets, as reported in (Joty et al., 2015).
Comparison between CRF Variants

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCCRF (ord=1)</td>
<td>76.1</td>
<td>83.2</td>
<td>79.4</td>
<td>78.3</td>
</tr>
<tr>
<td>LCCRF (ord=2)</td>
<td>76.8</td>
<td>82.1</td>
<td>79.3</td>
<td>78.4</td>
</tr>
<tr>
<td>FCCRF</td>
<td>77.3</td>
<td>86.2</td>
<td>81.5</td>
<td>80.5</td>
</tr>
<tr>
<td>FCCRF-noFeatures</td>
<td>77.2</td>
<td>86.0</td>
<td>81.4</td>
<td>80.1</td>
</tr>
<tr>
<td>FCCRF (4C)</td>
<td>78.8</td>
<td>79.7</td>
<td>79.3</td>
<td>79.0</td>
</tr>
</tbody>
</table>

Linear chain CRFs are not the best models for this task.
Comparison between CRF Variants

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCCRF (ord=1)</td>
<td>76.1</td>
<td>83.2</td>
<td>79.4</td>
<td>78.3</td>
</tr>
<tr>
<td>LCCRF (ord=2)</td>
<td>76.8</td>
<td>82.1</td>
<td>79.3</td>
<td>78.4</td>
</tr>
<tr>
<td>FCCRF</td>
<td>77.3</td>
<td>86.2</td>
<td>81.5</td>
<td>80.5</td>
</tr>
<tr>
<td>FCCRF-noFeatures</td>
<td>77.2</td>
<td>86.0</td>
<td>81.4</td>
<td>80.1</td>
</tr>
<tr>
<td>FCCRF (4C)</td>
<td>78.8</td>
<td>79.7</td>
<td>79.3</td>
<td>79.0</td>
</tr>
</tbody>
</table>

Edge features do not contribute much. Ising-like edge potential is crucial.
Error Analysis

• Accuracy for threads with more than one comment
  o Local: 78.7
  o Inference: 79.1
  o Joint: 80.4

• Disagreements
  o Local vs. Inference: 6%
  o Local vs. Joint: 9.9%
  o Inference vs. Joint: 8.8%

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. [...] Loc: Good, Inf: Good, Jnt: Good, Hum: Good

A$_5$ A big box of decent tea... like “Scottish blend” or “Tetleys”... [...] Loc: Good, Inf: Good, Jnt: Good, Hum: Good

A$_6$ Bacon. Nice bread, bacon, bacon, errmmm bacon and a pork joint. Loc: Good, Inf: Bad, Jnt: Good, Hum: Good

A$_8$ Go to Tesco buy some good latest DVD... [...] Loc: Good, Inf: Good, Jnt: Good, Hum: Good

A$_9$ Couple of good novels, All time favorite movies, .. Loc: Good, Inf: Bad, Jnt: Good, Hum: Good

A$_{10}$ Agree I do the same Indorachel..But some time you get a good copy some time a bad one... [...] Loc: Good, Inf: Good, Jnt: Good, Hum: Bad

A$_{11}$ Ditto on the books and dvd’s. Excedrin. Loc: Bad, Inf: Bad, Jnt: Good, Hum: Good

A$_{12}$ Ditto on the bacon, pork sausage, pork chops, ham...can you tell we miss pork! [...] Loc: Bad, Inf: Bad, Jnt: Good, Hum: Good
Conclusion

• Proposed two models for coupling learning with inference
• The locally normalized model suffers from label bias
• The FCCRF model with Ising-like edge potentials performs the best and achieves state-of-the-art results.

Future Work

• In future, we would like to apply FCCRF to other cQA tasks:
  - finding related questions to a new question
  - finding good answers to a new question.
Joint Learning with Global Inference for Comment Classification in Community Question Answering

*Shafiq Joty, Lluís Màrquez and Preslav Nakov*

Arabic Language Technology (ALT) Group
Qatar Computing Research Institute - HBKU