# 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

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

# 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*

$A_1$ Did you check with Toys R us? I think I saw it there.

$A_2$ Go to Doha city center you may get it at 4 floor.

$A_3$ ``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 :)

$A_4$ Hobby Shop- City Centre

$A_5$ OMG!! :| Guns and helicopters??!!

$A_6$ 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 definetly 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**

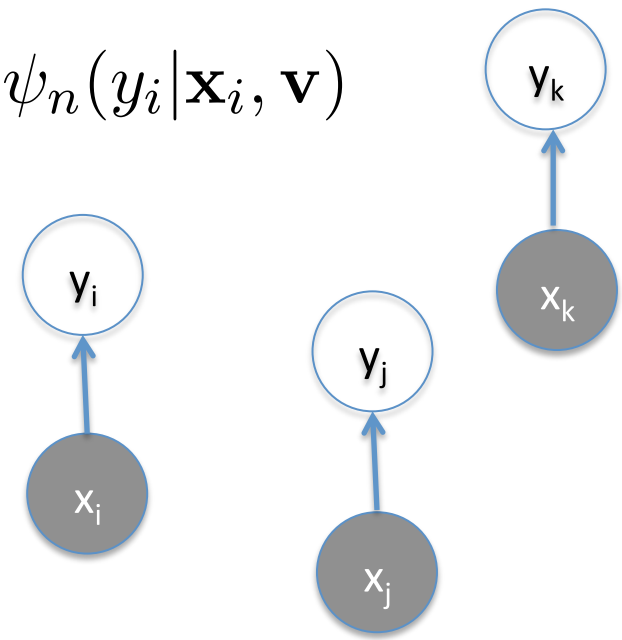
# Outline

- Motivation
- Three approaches to classification
- Our models
  - Locally normalized Joint model
  - 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

$$\psi_n(y_i|\mathbf{x}_i, \mathbf{v})$$




**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):

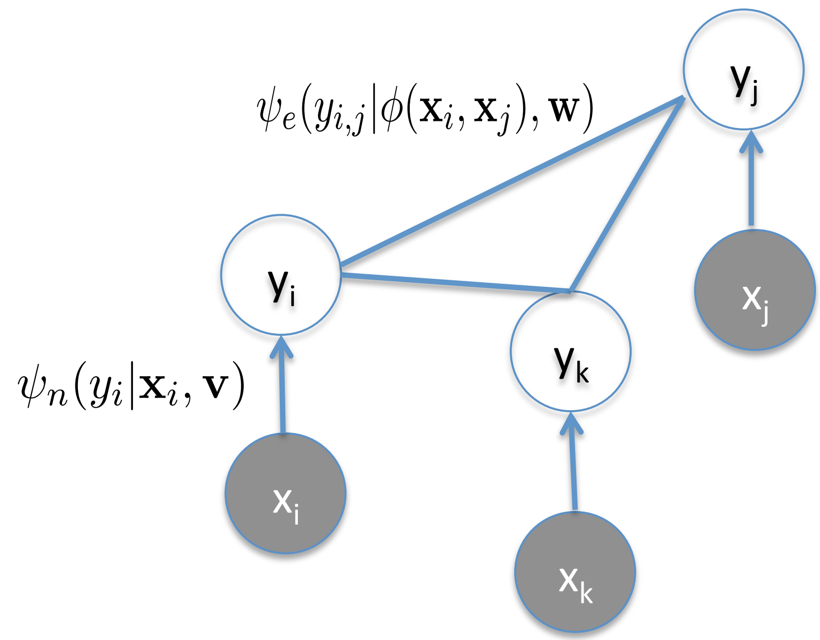
**Learning**

a) Learn a node-level classifier

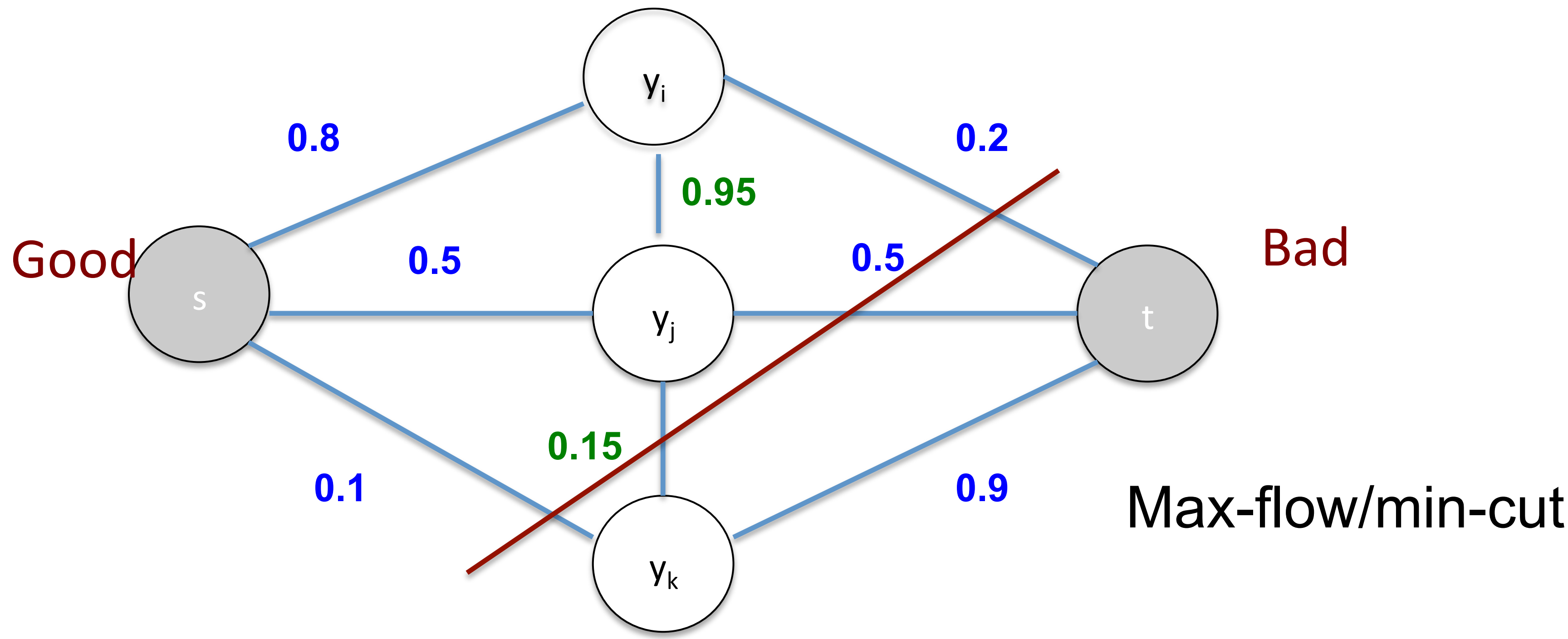
b) Learn an edge-level classifier

**Inference**

c) Classify collectively using global inference (ILP, Graph-cut)



# 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.

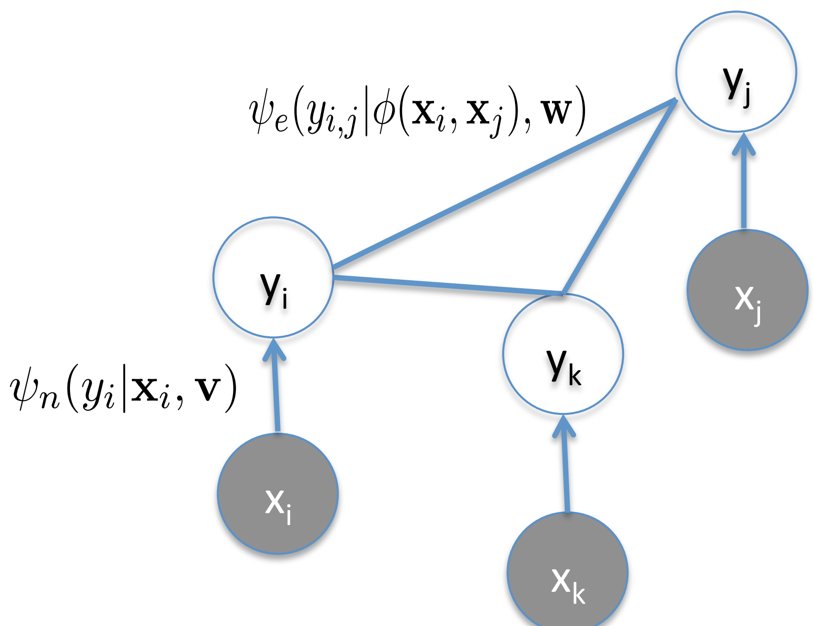
Models dependencies between output variables while learning.

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

# Our Models

**Model 1:** Learn two local classifiers jointly with global feedback

- Node-level classifier: $\psi_n(y_i = k|\mathbf{x_i}, \mathbf{v}) = \frac{\exp(\mathbf{v}_k^T \mathbf{x_i})}{\sum_{k'=1}^{K} \exp(\mathbf{v}_{k'}^T \mathbf{x_i})}$
- Edge-level classifier: $\psi_e(y_{i,j} = l|\phi(\mathbf{x_i}, \mathbf{x_j}), \mathbf{w}) = \frac{\exp(\mathbf{w}_l^T \phi(\mathbf{x_i}, \mathbf{x_j}))}{\sum_{l'=1}^{L} \exp(\mathbf{w}_{l'}^T \phi(\mathbf{x_i}, \mathbf{x_j}))}$



# Our Models

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

- Node-level classifier: $\psi_n(y_i = k|\mathbf{x_i}, \mathbf{v}) \;=\; \frac{\exp(\mathbf{v}_k^T \mathbf{x_i})}{\sum_{k'=1}^{K} \exp(\mathbf{v}_k^T \mathbf{x_i})}$
- Edge-level classifier: $\psi_e(y_{i,j} = l|\phi(\mathbf{x_i}, \mathbf{x_j}), \mathbf{w}) = \frac{\exp(\mathbf{w}_l^T \phi(\mathbf{x_i}, \mathbf{x_j}))}{\sum_{l'=1}^{L} \exp(\mathbf{w}_{l'}^T \phi(\mathbf{x_i}, \mathbf{x_j}))}$

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

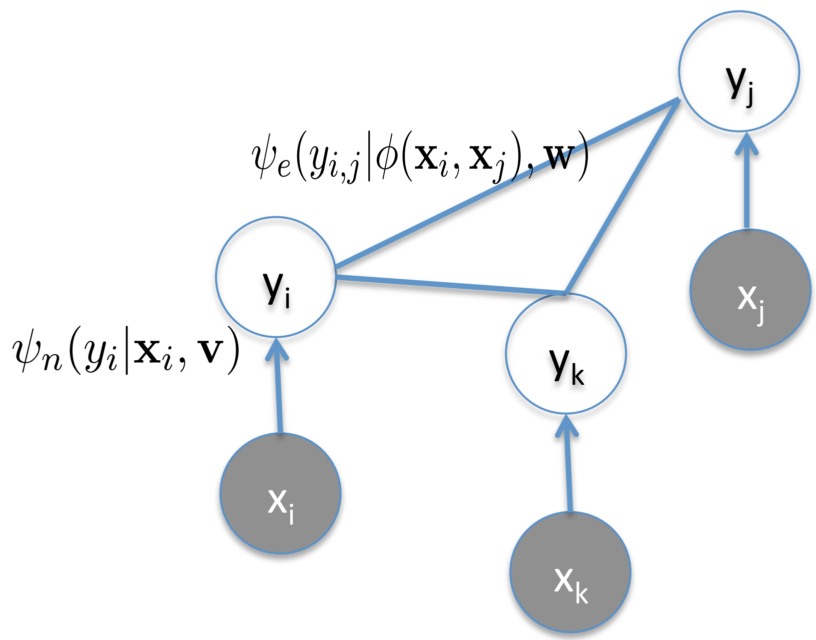
1. Initialize the model parameters $\mathbf{v}$ and $\mathbf{w}$;
2. **repeat**
   - **for** *each thread* $G = (V, E)$ **do**
     - a. Compute node and edge probabilities $\psi_n(y_i|\mathbf{x}_i, \mathbf{v})$ and $\psi_e(y_{i,j}|\phi(\mathbf{x}_i, \mathbf{x}_j), \mathbf{w})$;
     - b. Infer node and edge marginals $\beta_n(y_i)$ and $\beta_e(y_{i,j})$ using sum-product LBP;
     - c. Update: $\mathbf{v} = \mathbf{v} - \frac{\eta}{|V|} f'(\mathbf{v})$;
     - d. Update: $\mathbf{w} = \mathbf{w} - \frac{\eta}{|E|} f'(\mathbf{w})$;
   - **end**

   **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.

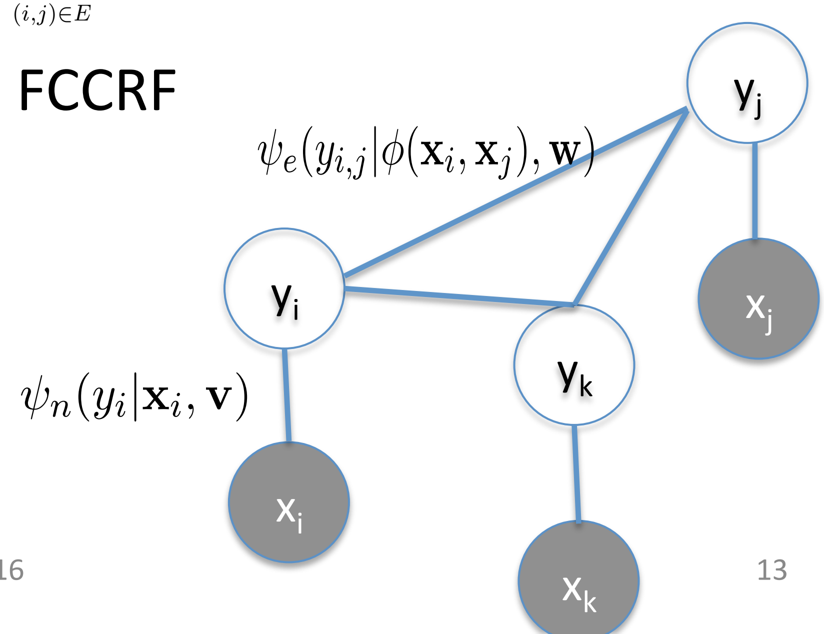


# Our Models

**Model 2:** Learn a joint model with global normalization

- The model: $p(\mathbf{y}|\mathbf{v},\mathbf{w},\mathbf{x}) = \frac{1}{Z(\mathbf{v},\mathbf{w},\mathbf{x})} \prod_{i \in V} \psi_n(y_i|\mathbf{x},\mathbf{v}) \cdot \prod_{(i,j)\in E} \psi_e(y_{i,j}|\mathbf{x},\mathbf{w})$
- Node potential: $\psi_n(y_i|\mathbf{x},\mathbf{v}) = \exp(\mathbf{v}^T \phi(y_i,\mathbf{x}))$
- Edge potential: $\psi_e(y_{i,j}|\mathbf{x},\mathbf{w}) = \exp(\mathbf{w}^T \phi(y_{i,j},\mathbf{x}))$
- Objective: $f(\theta) = \sum_{i \in V} \mathbf{v}^T \phi(y_i,\mathbf{x}) + \sum_{(i,j)\in E} \mathbf{w}^T \phi(y_{i,j},\mathbf{x}) - \log Z(\mathbf{v},\mathbf{w},\mathbf{x})$

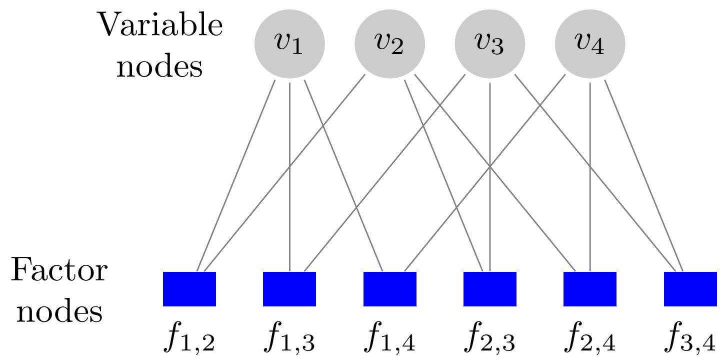
Pairwise FCCRF




- **Edge potentials:**
  a. All possible state transitions
  b. Ising like (Same and Different)

# Inference with Belief Propagation

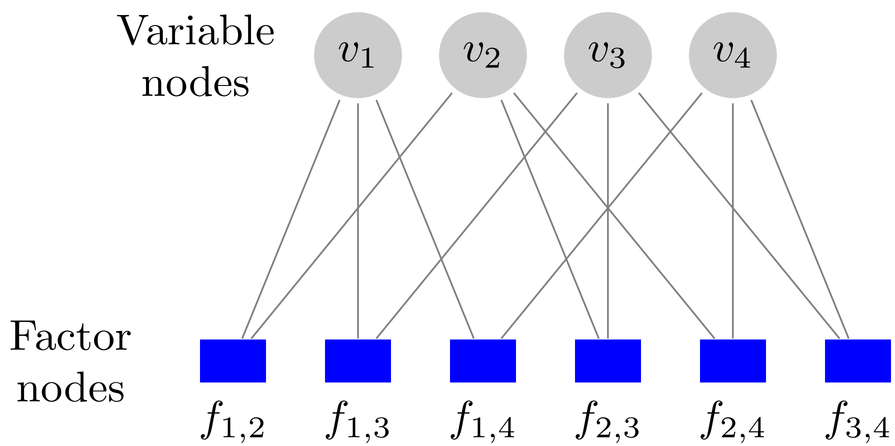
- Belief propagation (Pearl, 1988) is a message passing algorithm for performing inference in probabilistic graphical models.




- Message from a variable node to a factor node

$$\mu_{v\to a}(x_v) = \prod_{a^* \in N(v)\setminus\{a\}} \mu_{a^*\to v}(x_v); \forall x_v \in Dom(v)$$

# Inference with Belief Propagation




- Message from a factor node to a variable node

$$\mu_{a\to v}(x_v) = \sum_{\mathbf{x}'_a : x'_v = x_v} f_a(\mathbf{x}'_a) \prod_{v^* \in N(a)\setminus\{v\}} \mu_{v^*\to a}(x_{v*}); \forall x_v \in Dom(v)$$

- Upon convergence: $\boxed{P(x_v) \propto \prod_{a\in N(v)} \mu_{a\to v}(x_v)}$ $\boxed{P(\mathbf{x}_a) \propto f_a(\mathbf{x}_a) \prod_{v\in N(a)} \mu_{v\to a}(x_v)}$

# Belief Propagation for Pairwise Factors

Message: $$\mu_{i\to j}(y_j) = \sum_{y_i} \psi_n(y_i)\psi_e(y_{i,j}) \prod_{k\in N(i)\backslash j} \mu_{k\to i}(y_i)$$

Node Belief: $$\beta_n(y_i) \approx \psi_n(y_i) \prod_{j\in N(i)} \mu_{j\to i}(y_i)$$

Edge Belief: $$\beta_e(y_{i,j}) \approx \psi_e(y_{i,j}) \times \mu_{i\to j}(y_i) \times \mu_{j\to i}(y_j)$$

- 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
    - Locally normalized Joint model
    - 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

| | Train | Dev | Test |
|---|---|---|---|
| **Questions** | 2600 | 300 | 329 |
| **Comments** | 16,541 | 1645 | 1976 |

- **Metrics:**
  - Macro F1
  - Accuracy
- **Significance test:**
  - Appr. Randomization

# Experimental Settings: Features

Barrón-Cedeño et al. (2015); Joty et al (2015)

- **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

    Inf. alg.
- Joint Learning & Inference
    - Joint MaxEnts (SGD)
    - FCCRF (SGD)
    - Graph cut
    - Loopy BP

    Inf. alg.

# Main Results

| | *Model* | *Learning* | *Inference* | P | R | $F_1$ | Acc |
|---|---|---|---|---|---|---|---|
| I. | Majority | – | – | 50.5 | 100.0 | 67.1 | 50.5 |
| II. | $\mathrm{ICC}_{ME}$ | Local, SGD | – | 75.1 | 85.8 | 80.1 | 78.5 |
| | $\mathrm{ICC}_{Perc}$ | Local, Voted | – | 76.6 | 82.4 | 79.4 | 78.4 |
| III. | $\mathrm{LI}_{ME-GC}$ | Local, SGD | Graph-cut | **77.4** | 83.6 | 80.4 | 79.4 |
| | $\mathrm{LI}_{ME-LBP}$ | Local, SGD | LBP | 76.4 | 84.6 | 80.3 | 79.1 |
| IV. | $\mathrm{Joint}_{ME-LBP}$ | 2 classifiers, Joint, SGD | LBP | 76.1 | 84.4 | 80.0 | 78.7 |
| | $\mathrm{Joint}_{Perc-LBP}$ | 2 classifiers, Joint, AVG | LBP | 77.1 | 74.5 | 75.8 | 76.0 |
| | FCCRF | Joint, SGD | LBP | 77.3 | **86.2** | **81.5** | **80.5** |

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

MaxEnt performs slightly better than voted perceptron

# Main Results

| | *Model* | *Learning* | *Inference* | P | R | $F_1$ | Acc |
|---|---|---|---|---|---|---|---|
| I. | Majority | – | – | 50.5 | 100.0 | 67.1 | 50.5 |
| II. | $ICC_{ME}$ | Local, SGD | – | 75.1 | 85.8 | 80.1 | 78.5 |
| | $ICC_{Perc}$ | Local, Voted | – | 76.6 | 82.4 | 79.4 | 78.4 |
| III. | $LI_{ME-GC}$ | Local, SGD | Graph-cut | **77.4** | 83.6 | 80.4 | 79.4 |
| | $LI_{ME-LBP}$ | Local, SGD | LBP | 76.4 | 84.6 | 80.3 | 79.1 |
| IV. | $Joint_{ME-LBP}$ | 2 classifiers, Joint, SGD | LBP | 76.1 | 84.4 | 80.0 | 78.7 |
| | $Joint_{Perc-LBP}$ | 2 classifiers, Joint, AVG | LBP | 77.1 | 74.5 | 75.8 | 76.0 |
| | FCCRF | Joint, SGD | LBP | 77.3 | **86.2** | **81.5** | **80.5** |

- 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

| | *Model* | *Learning* | *Inference* | P | R | $F_1$ | Acc |
|---|---|---|---|---|---|---|---|
| I. | Majority | – | – | 50.5 | 100.0 | 67.1 | 50.5 |
| II. | $\mathrm{ICC}_{ME}$ | Local, SGD | – | 75.1 | 85.8 | 80.1 | 78.5 |
| | $\mathrm{ICC}_{Perc}$ | Local, Voted | – | 76.6 | 82.4 | 79.4 | 78.4 |
| III. | $\mathrm{LI}_{ME-GC}$ | Local, SGD | Graph-cut | **77.4** | 83.6 | 80.4 | 79.4 |
| | $\mathrm{LI}_{ME-LBP}$ | Local, SGD | LBP | 76.4 | 84.6 | 80.3 | 79.1 |
| IV. | $\mathrm{Joint}_{ME-LBP}$ | 2 classifiers, Joint, SGD | LBP | 76.1 | 84.4 | 80.0 | 78.7 |
| | $\mathrm{Joint}_{Perc-LBP}$ | 2 classifiers, Joint, AVG | LBP | 77.1 | 74.5 | 75.8 | 76.0 |
| | FCCRF | Joint, SGD | LBP | 77.3 | **86.2** | **81.5** | **80.5** |

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

| *Model* | P | R | $F_1$ | Acc |
|---|---|---|---|---|
| MaxEnt classifier | 75.7 | 84.3 | 79.8 | 78.4 |
| Linear CRF | 74.9 | 83.5 | 78.9 | 77.5 |
| MaxEnt+ILP | 77.0 | 83.5 | 80.2 | 79.1 |
| MaxEnt+GraphCut | **78.3** | 82.9 | 80.6 | 79.8 |
| Our method (FCCRF) | 77.3 | **86.2** | **81.5** | **80.5** |

# Comparison between CRF Variants

| *Model* | P | R | $F_1$ | Acc |
|---|---|---|---|---|
| LCCRF (ord=1) | 76.1 | 83.2 | 79.4 | 78.3 |
| LCCRF (ord=2) | 76.8 | 82.1 | 79.3 | 78.4 |
| FCCRF | 77.3 | **86.2** | **81.5** | **80.5** |
| FCCRF-noFeatures | 77.2 | 86.0 | 81.4 | 80.1 |
| FCCRF (4C) | **78.8** | 79.7 | 79.3 | 79.0 |

Linear chain CRFs are not the best models for this task

# Comparison between CRF Variants

| *Model* | P | R | $F_1$ | Acc |
|---|---|---|---|---|
| LCCRF (ord=1) | 76.1 | 83.2 | 79.4 | 78.3 |
| LCCRF (ord=2) | 76.8 | 82.1 | 79.3 | 78.4 |
| FCCRF | 77.3 | **86.2** | **81.5** | **80.5** |
| FCCRF-noFeatures | 77.2 | 86.0 | 81.4 | 80.1 |
| FCCRF (4C) | **78.8** | 79.7 | 79.3 | 79.0 |

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

# Error Analysis

- Accuracy for threads **with more than one** comment
  - Local: 78.7
  - Inference: 79.1
  - Joint: 80.4
- Disagreements
  - Local vs. Inference: 6%
  - Local vs. Joint: 9.9%
  - 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

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY