Speech Act Modeling of Written Asynchronous Conversations with Task-Specific Embeddings and Conditional Structured Models

Shafiq Joty *and Enamul Hoque*

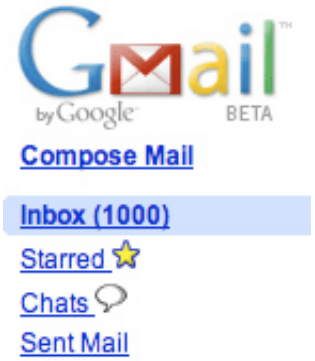
Arabic Language Technology (ALT) Group

Qatar Computing Research Institute - HBKU

Asynchronous Conversations

- Conversations where participants communicate with each other at different times.

- **Examples:**
  - Emails
  - Blogs
  - Forums
  - Twitter
  - Facebook



The Task: Speech Act Recognition in Asynchronous Conversations

C_1	*My son wish to do his bachelor degree in Mechanical Engineering in an affordable Canadian university.*	ST
	The info. available in the net and the people who wish to offer services are too many and some are misleading.	ST
	The preliminary preparations,eligibility,the require funds etc., are some of the issues which I wish to know from any panel members of this forum who …	Q
C_2	*.. take a list of canadian universities and then create a table and insert all the relevant info. by reading each and every program info. on the web.*	SU
	Without doing a research my advice would be to apply to UVIC .. for the following reasons ..	SU
C_3	*snakyy21: UVIC is a short form of? I have already started researching for my brother and found ``College of North Atlantic'' and ..*	Q
:		
C_5	*thank you for sharing useful tips will follow your advise.*	P

Contributions

1) Sentence representation

- Exiting methods use bag-of-ngrams
- Should consider sentence structure
- Our solution: sequential LSTM

2) Conversational dependencies

- Exiting methods usually classify each sentence locally
- Should consider dependencies inside and across comments
- Our solution: structured models

3) A new corpus

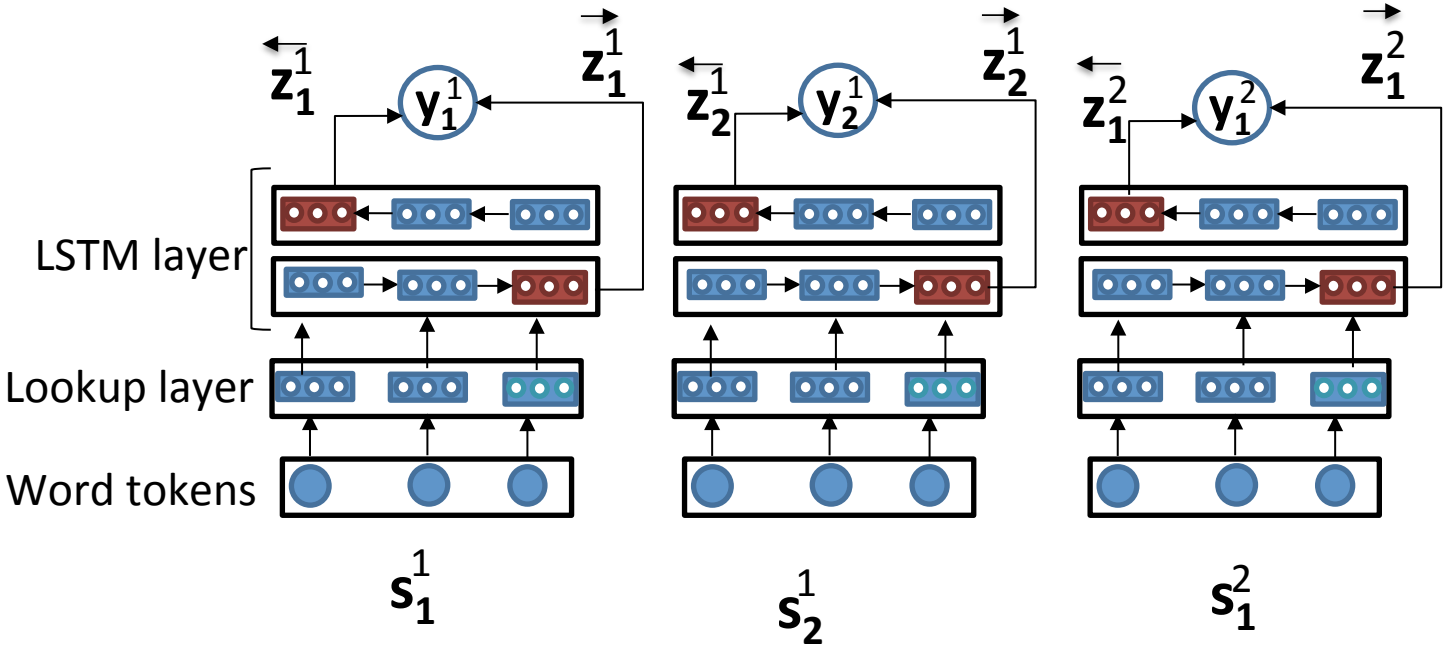
- Forum conversations
- Annotated with standard tagset

Outline

- Motivation
- Our Approach
  - Sentence representation using LSTMs
  - Conditional structured models
- Corpora
  - Existing datasets
  - New forum corpus
- Experiments & Analysis
  - Effectiveness of LSTM RNNs
  - Effectiveness of CRFs
- Conclusion & future work

Our Approach

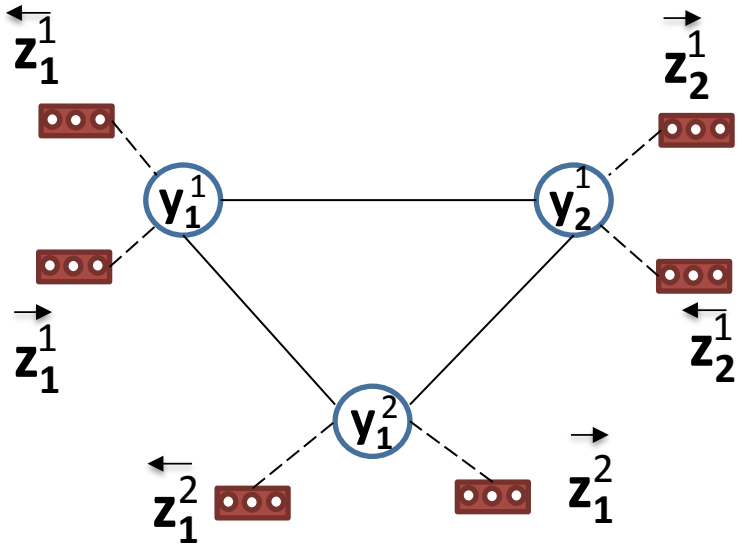
Step 1: LSTM for speech act classification & sentence encoding



- Considers word order in a sentence
- Does not consider the interdependencies between sentences.

Our Approach

Step 2: Conversational dependencies with structured models

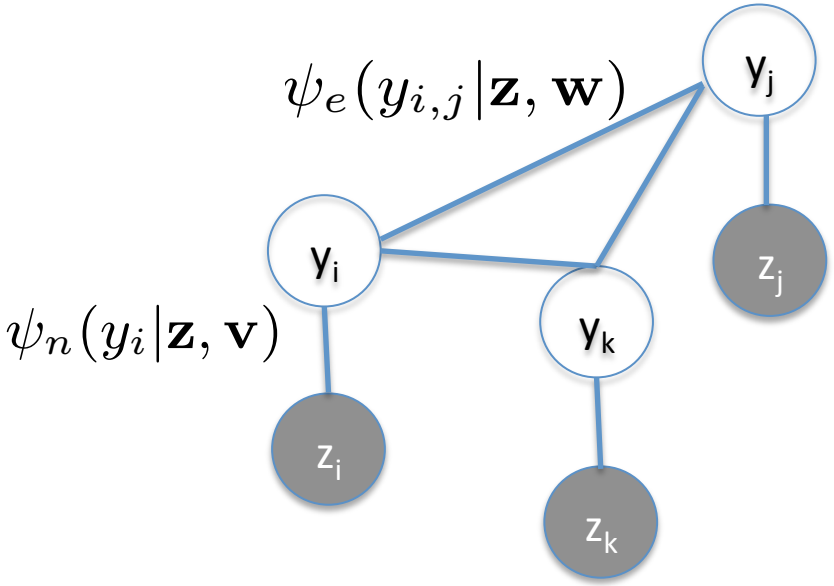


Fully-connected graph

- Experimented with various graph structures

Conditional Structured Model

- Learn a joint model with global normalization



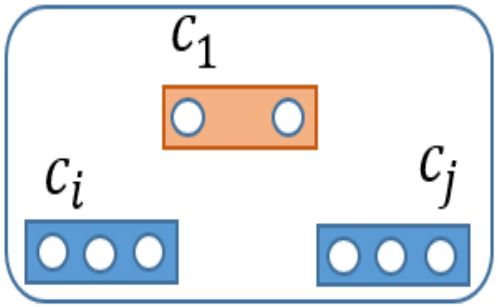
Pairwise CRF

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

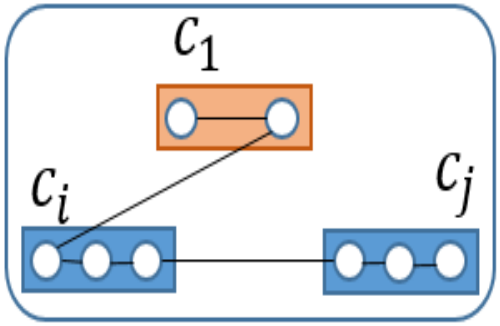
CRF Graph Structures

- **Intra-** and **across-**comment connections

Tag	Connection type	Applicable to
NO	No connection between nodes	intra & across
LC	Linear chain connection	intra & across
FC	Fully connected	intra & across
FC_1	Fully connected with first comment only	across
LC_1	Linear chain with first comment only	across



(a) NO-NO (MaxEnt)

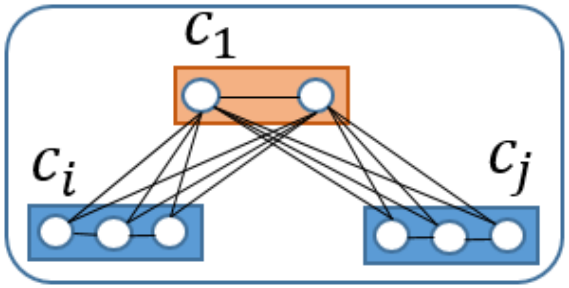


(b) LC-LC

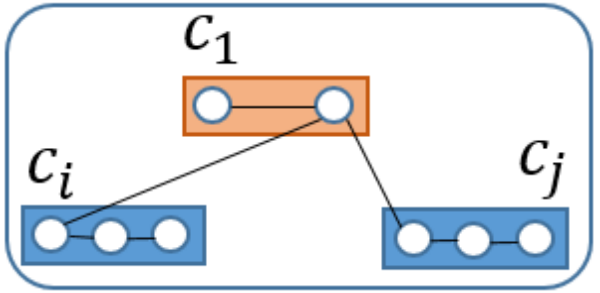
CRF Graph Structures

- **Intra-** and **across-**comment connections

Tag	Connection type	Applicable to
NO	No connection between nodes	intra & across
LC	Linear chain connection	intra & across
FC	Fully connected	intra & across
FC_1	Fully connected with first comment only	across
LC_1	Linear chain with first comment only	across



(d) LC-FC_1



(c) LC-LC_1

Training & Inference in CRFs

- Online learning (SGD)
- Inference: Loopy belief propagation (Pearl, 1988)

Algorithm 1: Online learning algorithm for conditional random fields

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

Outline

- Motivation
- Our Approach
  - Sentence representation using LSTMs
  - Conditional structured models
- Corpora
  - Existing datasets
  - New forum corpus
- Experiments & Analysis
  - Effectiveness of LSTM RNNs
  - Effectiveness of CRFs
- Conclusion & future work

Corpora: Existing Datasets

- **Asynchronous domains**
  - Trip Advisor **forum** (Jeong et al. 2009)
  - BC3 **email** corpus (Ulrich et al. 2008)
- **Synchronous domain**
  - **Meeting** Recorder Dialog Act or MRDA (Dhillon et al. 2004)

	TA	BC3
Total number of conv.	200	39
Avg. nb of comments per conv.	4.02	6.54
Avg. nb of sentences per conv.	18.56	34.15
Avg. nb of words per sentence	14.90	12.61

Tag	Description	TA	BC3	MRDA
SU	Suggestion	7.71%	5.48%	5.97%
R	Response	2.4%	3.75%	15.63%
Q	Question	14.71%	8.41%	8.62%
P	Polite	9.57%	8.63%	3.77%
ST	Statement	65.62%	73.72%	66.00%

Corpora: A New Forum Dataset

- QC3 conversational corpus
  - 50 conversations from Qatar Living forum.

Total number of conv.	50
Avg. nb of comments per conv.	13.32
Avg. nb of sentences per conv.	33.28
Avg. nb of words per sentence	19.78

Speech Act	Distribution	κ
Suggestion	17.38%	0.86
Response	5.24%	0.43
Question	12.59%	0.87
Polite	6.13%	0.75
Statement	58.66%	0.78

Experiments: Effectiveness of LSTMs

- **Data split:**
  - *Asynchronous:* 80% train, 10% test, 10% valid.
  - *MRDA:* Same as Jenog et al. (2009)

- **Baselines:**
  - *ME:* MaxEnt with BoW representation
  - *MLP:* One hidden layer MLP with BoW representation

- **LSTM settings:**
  - ADAM (Kingma & Ba, 2014) learning alg.
  - Dropout & Early stopping.
  - Random & Word2Vec initialization.

Experiments: Effectiveness of LSTMs

	QC3		TA		MRDA	
	Testset	5 folds	Testset	5 folds	5 classes	12 classes
Jeong et al. (ng)	-	-	-	-	-	57.53 (83.30)
Jeong et al. (All)	-	-	-	-	-	59.04 (83.49)
ME	55.12 (**75.64**)	50.23 (71.37)	61.4 (85.44)	59.23 (84.85)	65.25 (83.95)	57.79 (82.84)
MLP	**61.30** (74.36)	**54.57 (71.63)**	**68.17 (85.98)**	**62.41 (85.02)**	68.12 (84.24)	58.19 (83.24)
U-LSTM$_r$	51.57 (73.55)	48.64 (65.94)	56.54 (83.24)	56.39 (83.83)	71.29 (85.38)	58.72 (83.34)
U-LSTM$_p$	49.41 (70.97)	50.26 (65.62)	63.12(83.78)	59.10 (83.13)	72.32 (85.19)	59.05 (84.06)
B-LSTM$_r$	50.75 (72.26)	48.41 (66.19)	58.88 (82.97)	56.23 (83.34)	71.69 (85.62)	58.33 (83.49)
B-LSTM$_p$	53.22 (71.61)	51.59 (68.50)	60.73 (82.97)	59.68 (84.07)	72.02 (85.33)	**60.12 (84.46*)**

- Jeong et al. (All): using ME with all features, e.g., n-gram, speaker, dependency, POS.
- LSTMs and Jeong et al. (ng) use the same information.
- All LSTM variants achieve state-of-the-art results on MRDA.
- B-LSTM$_p$ is significantly better than the best existing result.

Experiments: Effectiveness of LSTMs

	QC3		TA		MRDA	
	Testset	5 folds	Testset	5 folds	5 classes	12 classes
Jeong et al. (ng)	-	-	-	-	-	57.53 (83.30)
Jeong et al. (All)	-	-	-	-	-	59.04 (83.49)
ME	55.12 (**75.64**)	50.23 (71.37)	61.4 (85.44)	59.23 (84.85)	65.25 (83.95)	57.79 (82.84)
MLP	**61.30** (74.36)	**54.57 (71.63)**	**68.17 (85.98)**	**62.41 (85.02)**	68.12 (84.24)	58.19 (83.24)
U-LSTM$_r$	51.57 (73.55)	48.64 (65.94)	56.54 (83.24)	56.39 (83.83)	71.29 (85.38)	58.72 (83.34)
U-LSTM$_p$	49.41 (70.97)	50.26 (65.62)	63.12(83.78)	59.10 (83.13)	72.32 (85.19)	59.05 (84.06)
B-LSTM$_r$	50.75 (72.26)	48.41 (66.19)	58.88 (82.97)	56.23 (83.34)	71.69 (85.62)	58.33 (83.49)
B-LSTM$_p$	53.22 (71.61)	51.59 (68.50)	60.73 (82.97)	59.68 (84.07)	72.02 (85.33)	**60.12 (84.46*)**

- Pre-trained Google vectors give better initialization.
- Bi-directional LSTMs perform better than their unidirectional counterpart.

Experiments: Effectiveness of LSTMs

	QC3		TA		MRDA	
	Testset	5 folds	Testset	5 folds	5 classes	12 classes
Jeong et al. (ng)	-	-	-	-	-	57.53 (83.30)
Jeong et al. (All)	-	-	-	-	-	59.04 (83.49)
ME	55.12 (**75.64**)	50.23 (71.37)	61.4 (85.44)	59.23 (84.85)	65.25 (83.95)	57.79 (82.84)
MLP	**61.30** (74.36)	**54.57 (71.63)**	**68.17 (85.98)**	**62.41 (85.02)**	68.12 (84.24)	58.19 (83.24)
U-LSTM$_r$	51.57 (73.55)	48.64 (65.94)	56.54 (83.24)	56.39 (83.83)	71.29 (85.38)	58.72 (83.34)
U-LSTM$_p$	49.41 (70.97)	50.26 (65.62)	63.12(83.78)	59.10 (83.13)	72.32 (85.19)	59.05 (84.06)
B-LSTM$_r$	50.75 (72.26)	48.41 (66.19)	58.88 (82.97)	56.23 (83.34)	71.69 (85.62)	58.33 (83.49)
B-LSTM$_p$	53.22 (71.61)	51.59 (68.50)	60.73 (82.97)	59.68 (84.07)	72.02 (85.33)	**60.12 (84.46*)**

- ME and MLP baselines outperform LSTMs by a good margin.
- Same observation with 5-fold cross validation over the whole corpus.
- Not surprising since LSTMs have a lot of parameters.

Experiments: Effectiveness of LSTMs

- Results after training on a concatenated dataset:
  - MRDA + TA + BC3 + QC3

	QC3 (Testset)	TA (Testset)
ME	50.64 (71.15)	72.49 (84.10)
MLP	58.60 (74.36)	73.07 (86.29)
B-LSTM$_p$	**66.40 (80.65*)**	**73.14 (87.01*)**

- Bi-directional LSTM outperforms the baselines.
- ME and MLP suffer from data diversity.
- Bi-directional LSTM gives better sentence representation

Experiments: Effectiveness of CRFs

- Datasets for CRF experiments

	Train	Dev	Test
QC3	38 (1332)	4 (111)	5 (122)
TA	160 (2957)	20 (310)	20 (444)
Total	197 (4289)	24 (421)	25 (566)

- CRF variants

Tag	Connection type	Applicable to
NO	No connection between nodes	intra & across
LC	Linear chain connection	intra & across
FC	Fully connected	intra & across
FC_1	Fully connected with first comment only	across
LC_1	Linear chain with first comment only	across

Experiments: Effectiveness of CRFs

	QC3	TA
ME_b	56.67 (67.21)	63.29 (84.23)
B-$LSTM_p$	65.15 (77.87)	66.93 (85.13)
ME_e	59.94 (77.05)	59.55 (85.14)
CRF (LC-NO)	62.20 (77.87)	60.30 (85.81)
CRF (LC-LC)	62.35 (78.69)	60.30 (85.81)
CRF (LC-LC_1)	**65.94 (80.33*)**	61.58 (86.54)
CRF (LC-FC_1)	61.18 (77.87)	60.00 (85.36)
CRF (FC-FC)	64.54 (79.51*)	**61.64 (86.81*)**

- Baselines (local models)
  - **ME_b :** MaxEnt with BoW representation.
  - **B-$LSTM_p$** : Bi-directional LSTM with pre-trained embeddings. Trained on concatenated dataset.
  - **ME_e :** MaxEnt with sentence embeddings from B-$LSTM_p$.

Experiments: Effectiveness of CRFs

	QC3	TA
ME_b	56.67 (67.21)	63.29 (84.23)
B-LSTM_p	65.15 (77.87)	66.93 (85.13)
ME_e	59.94 (77.05)	59.55 (85.14)
CRF (LC-NO)	62.20 (77.87)	60.30 (85.81)
CRF (LC-LC)	62.35 (78.69)	60.30 (85.81)
CRF (LC-LC_1)	**65.94 (80.33*)**	61.58 (86.54)
CRF (LC-FC_1)	61.18 (77.87)	60.00 (85.36)
CRF (FC-FC)	64.54 (79.51*)	**61.64 (86.81*)**

- CRF models use the sentence embeddings from B-LSTM_p
- CRFs generally outperform local baselines in accuracy.
- Linear chain CRFs are not the best models.
- CRF (LC-LC_1) and CRF (FC-FC) are best performing models.

Experiments: Error Analysis

C_1: My son wish to do his bachelor degree in Mechanical Engineering in an affordable Canadian university.
Human: st, Local: st, Global: st

The info. available in the net and the people who wish to offer services are too many and some are misleading.
Human: st, Local: st, Global: st

The preliminary preparations,eligibility,the require funds etc., are some of the issues which I wish to know from any panel members of this forum .. (truncated)
Human: ques, Local: st, Global: st

C_3 (truncated)...take a list of canadian universities and then create a table and insert all the relevant information by reading each and every program info on the web.
Human: sug, Local: sug, Global: sug

Without doing a research my advice would be to apply to UVIC .. for the following reasons .. (truncated)
Human: sug, Local: sug, Global: sug

UBC is good too... but it is expensive particularly for international students due to tuition .. (truncated)
Human: sug, Local: sug, Global: sug

most of them accept on-line or email application.
Human: st, Local: st, Global: st

Good luck !!
Human: pol, Local: pol, Global: pol

C_4 snakyy21: UVIC is a short form of? I have already started researching for my brother and found "College of North Atlantic" and .. (truncated)
Human: ques, Local: st, Global: ques

but not sure about the reputation..
Human: st, Local: res, Global: st

Conclusion & Future Work

- Two-step framework for speech act recognition
  - LSTM-RNN to encode each sentence
  - Pairwise CRFs to model conversational dependencies
- Combine the input representational power of DNNs with the output representational power of PGMs.
- LSTMs provide better representations but requires more data
- Global joint models improve over local models given that it considers the right graph structure.
- Combine CRFs with LSTMs to perform the two steps jointly by taking LBP errors back to the embedding layers.
- Apply to conversations where graph structure is already given (e.g., Slashdot) or extractable (emails).

Code & Data: http://alt.qcri.org/tools/speech-act/

Speech Act Modeling of Written Asynchronous Conversations with Task-Specific Embeddings and Conditional Structured Models

Shafiq Joty *and Enamul Hoque*

Arabic Language Technology (ALT) Group

Qatar Computing Research Institute - HBKU

Belief Propagation for Pairwise Factors

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

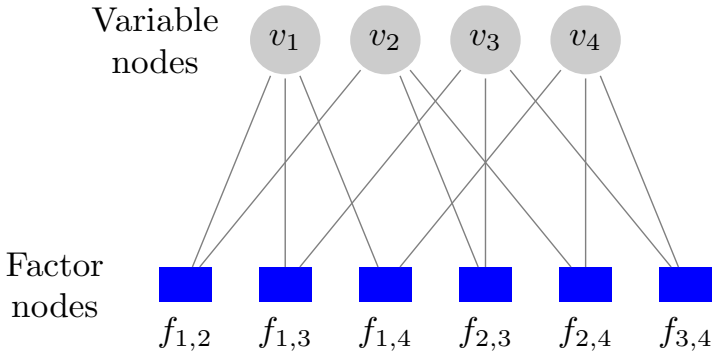
Node Belief: $$\beta_n(y_i) \approx \psi_n(y_i) \prod_{j\in N(i)} \mu_{j\to i}(y_i)$$

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)

Inference with Belief Propagation

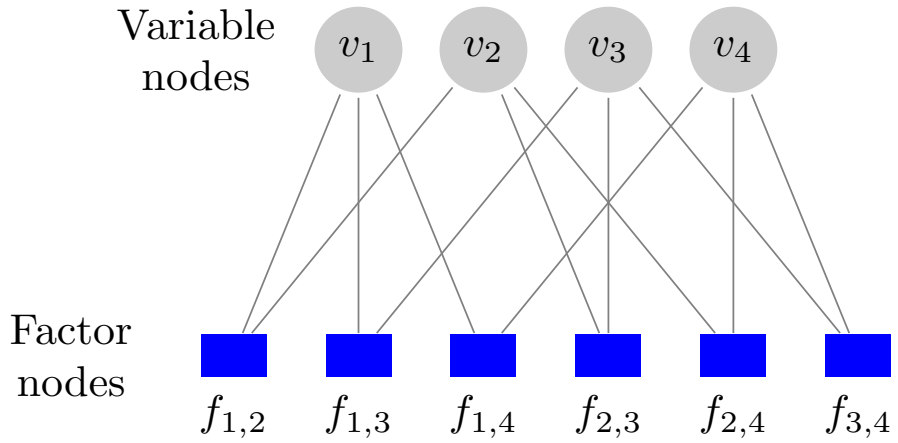
- Belief propagation (Pearl, 1988) is a message passing algorithm for performing inference in probabilistic graphical models.



- 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:

$$P(x_v) \propto \prod_{a\in N(v)} \mu_{a\to v}(x_v)$$

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