Learning to Disentangle Interleaved Conversational Threads with a Siamese Hierarchical Network and Similarity Ranking

Jyun-Yu Jiang†, Francine Chen‡, Yan-Ying Chen‡ and Wei Wang†

†University of California, Los Angeles (UCLA)
‡FX Palo Alto Laboratory (FXPAL)

June 4, 2018 (NAACL)

Conversations are everyday and everywhere.

We humans are inherently social beings.

In the real world...

In the virtual world...

Conversations are everyday and everywhere.

We humans are inherently social beings.

In the real world...

In the virtual world...

Conversations are everyday and everywhere.

We humans are inherently social beings.

In the real world...

In the virtual world...

Conversations can be simultaneous and interleaved!

Avg. 1.79 conversations at a time

Avg. 2.75 conversations at a time

Avg. 3^+ conversations at a time

Party of 8 Participants

[Aoki et al., 2006]

IRC Channels

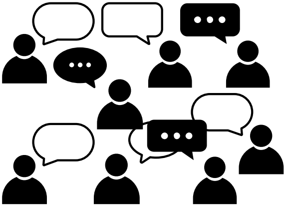
[Elsner and Charniak, 2013]

Web Forum Discussions

[Aragón et al., 2017]

Conversations can be simultaneous and interleaved!

Avg. 1.79 conversations at a time



Party of 8 Participants

[Aoki et al., 2006]

Avg. 2.75 conversations at a time

IRC Channels

[Elsner and Charniak, 2013]

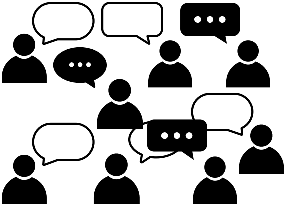
Avg. 3^+ conversations at a time

Web Forum Discussions

[Aragón et al., 2017]

Conversations can be simultaneous and interleaved!

Avg. 1.79 conversations at a time



Party of 8 Participants

[Aoki et al., 2006]

Avg. 2.75 conversations at a time

IRC Channels

[Elsner and Charniak, 2013]

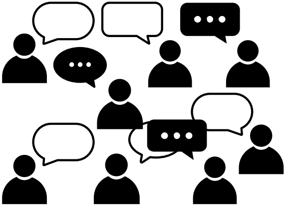
Avg. 3^+ conversations at a time

Web Forum Discussions

[Aragón et al., 2017]

Conversations can be simultaneous and interleaved!

Avg. 1.79 conversations at a time



Party of 8 Participants

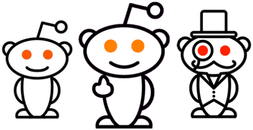
[Aoki et al., 2006]

Avg. 2.75 conversations at a time

IRC Channels

[Elsner and Charniak, 2013]

Avg. 3^{+} conversations at a time



Web Forum Discussions

[Aragón et al., 2017]

An Example from the Real-World IRC Dataset

Thread	Message
⋮	⋮
T31	*Malcolm: If running as root, I need to set up a global config rather than ~/.fetchmailrc ?*
T38	***Elma:*** *i'm sure i missed something but fonts rendering in my gimp works isn't at its best*
T39	*Sena: is there anyway to see what the CPU temperature is?*
T38	***Elma:*** *is it because of gimp or i missed some tuning or such?*
T31	*Rache: Specify a non-default name run control file.*
T41	*Denny: so how does one enforce a permission set and ownership set on a folder and all its children?*
T31	*Malcolm: in the man page it doesn't mention any global fetchmailrc file... that is what was confusing me...*
T42	*Shenna: hi, are sata drives accessed as sda or hda?*
T41	***Elma:*** *-R for recursive...*
T42	***Elma:*** *sda*
⋮	⋮

An Example from the Real-World IRC Dataset

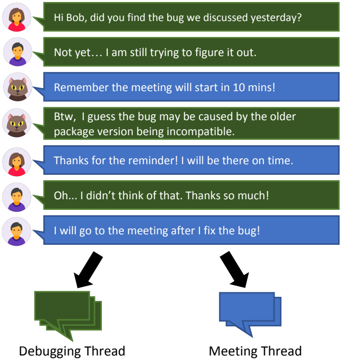
Thread	Message
⋮	⋮
T31	*Malcolm: If running as root, I need to set up a global config rather than ∼/.fetchmailrc ?*
T38	***Elma:** i'm sure i missed something but fonts rendering in my gimp works isn't at its best*
T39	*Sena: is there anyway to see what the CPU temperature is?*
T38	***Elma:** is it because of gimp or i missed some tuning or such?*
T31	*Rache: Specify a non-default name run control file.*
T41	*Denny: so how does one enforce a permission set and ownership set on a folder and all its children?*
T31	*Malcolm: in the man page it doesn't mention any global fetchmailrc file... that is what was confusing me...*
T42	*Shenna: hi, are sata drives accessed as sda or hda?*
T41	***Elma:** -R for recursive...*
T42	***Elma:** sda*
⋮	⋮

An Example from the Real-World IRC Dataset

Thread	Message
⋮	⋮
T31	*Malcolm: If running as root, I need to set up a global config rather than ~/.fetchmailrc ?*
T38	***Elma:*** *i'm sure i missed something but fonts rendering in my gimp works isn't at its best*
T39	*Sena: is there anyway to see what the CPU temperature is?*
T38	***Elma:*** *is it because of gimp or i missed some tuning or such?*
T31	*Rache: Specify a non-default name run control file.*
T41	*Denny: so how does one enforce a permission set and ownership set on a folder and all its children?*
T31	*Malcolm: in the man page it doesn't mention any global fetchmailrc file... that is what was confusing me...*
T42	*Shenna: hi, are sata drives accessed as sda or hda?*
T41	***Elma:*** *-R for recursive...*
T42	***Elma:*** *sda*
⋮	⋮

Conversation disentanglement is needed!

- Interleaved conversations are messy.
  - Difficult to follow discussions
  - Hard to find relevant messages to a specific conversation



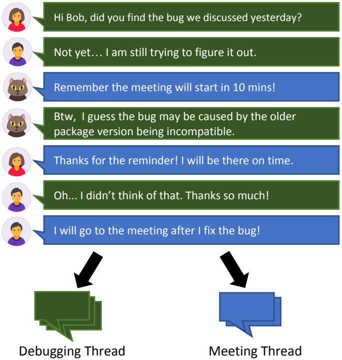
Conversation Disentanglement

- Given a sequence of messages
- Disentangle messages into a thread for each individual conversation

Goal: to help users easily follow discussions and retrieve relevant messages

Conversation disentanglement is needed!

- Interleaved conversations are messy.
  - Difficult to follow discussions
  - Hard to find relevant messages to a specific conversation



Conversation Disentanglement

- Given a sequence of messages
- Disentangle messages into a thread for each individual conversation

Goal: to help users easily follow discussions and retrieve relevant messages

Conversation disentanglement is needed!

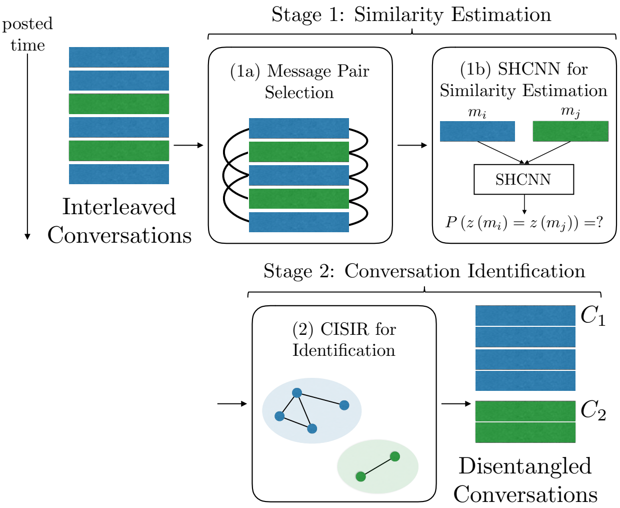
- Interleaved conversations are messy.
  - Difficult to follow discussions
  - Hard to find relevant messages to a specific conversation

Conversation Disentanglement

- Given a sequence of messages
- Disentangle messages into a thread for each individual conversation

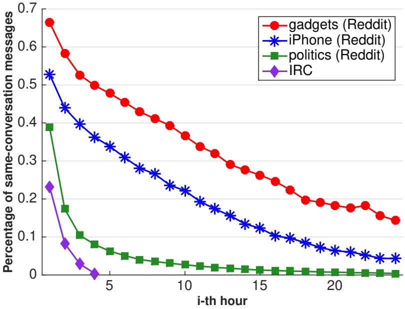
Goal: to help users easily follow discussions and retrieve relevant messages

Framework Overview



Message Pair Selection for Similarity Estimation

- Previous studies use all pairs.
- $O(n^2)$ message pairs
  - result in an enormous amount of computational time
- Low percentage of message pairs in the same conversation
  - amplifies false alarms
  - harms graph-based methods

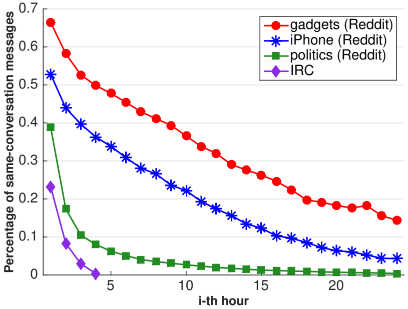


Do we need that many pairs?

Message Pair Selection for Similarity Estimation

- Previous studies use all pairs.
- $O(n^2)$ message pairs
  - result in an enormous amount of computational time
- Low percentage of message pairs in the same conversation
  - amplifies false alarms
  - harms graph-based methods

Do we need that many pairs?



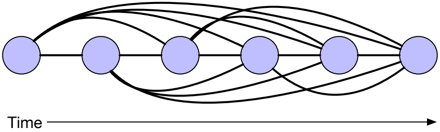
Assumption of Pairwise Redundancy

Assumption

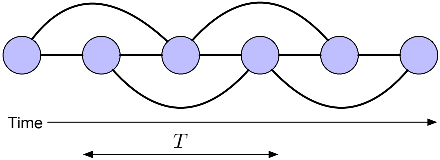
The time difference between two consecutive messages in the same conversation is rarely greater than T hours, where T is a small number.

- Only robust message pairs posted within T hours are needed.
- Redundancy of pairwise relationships

All Message Pairs



Message Pairs under Assumption



Message Representation with CNNs

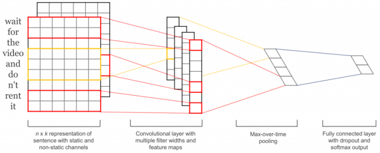
The effectiveness of convolutional NNs (CNNs) have been demonstrated.

Single-layer CNN

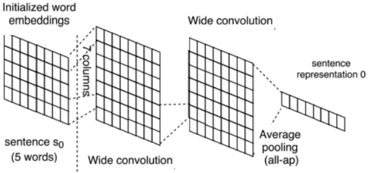
- Low-level context as *n*-gram
- Hard to capture high-level info

Multi-layer CNN

- High-level semantics
- Diluted low-level knowledge

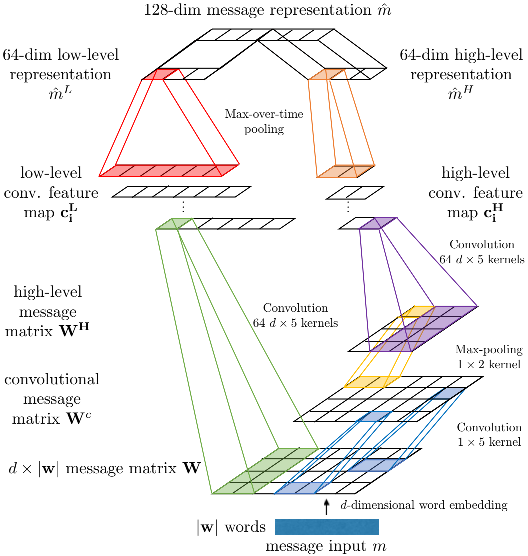


[Kim, 2014]



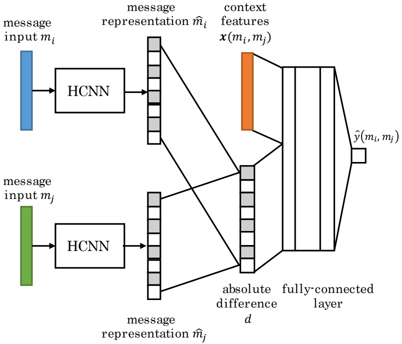
[Yin et al., 2016]

Hierarchical CNN (HCNN) for Message Representation

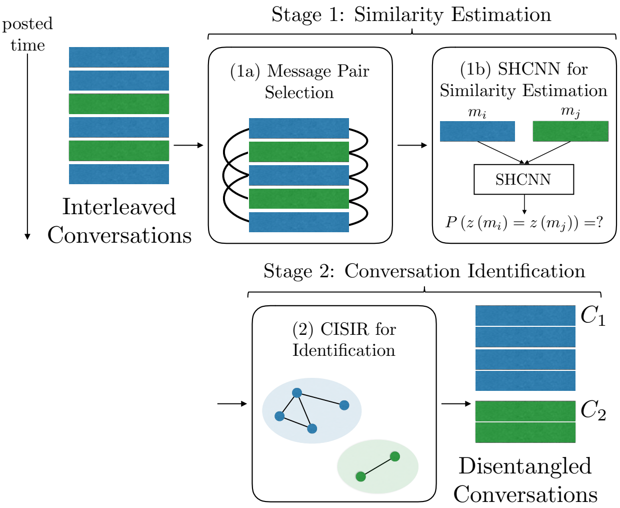


Siamese HCNN (SHCNN) for Similarity Estimation

- Absolute difference for representation comparison
    - Fewer parameters
    - Flexibility for each dimension
- Context features are included
    - Additional information
        - e.g., user data and syntactics
    - Interaction with representations

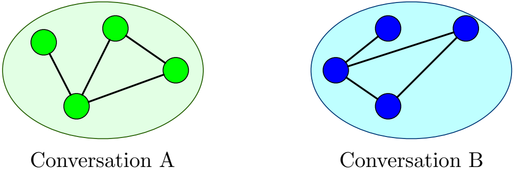


Framework Overview



Graph-based Conversation Disentanglement

- Construct a message graph using pairwise relationships
- Each connected component represents a conversation.

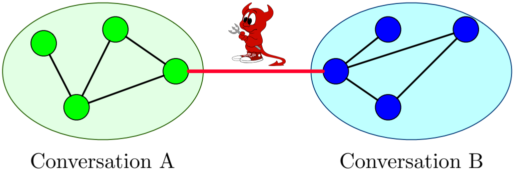


- However, false alarms are harmful!

Different conversations can be connected by only one mistake!

Graph-based Conversation Disentanglement

- Construct a message graph using pairwise relationships
- Each connected component represents a conversation.



- However, false alarms are harmful!

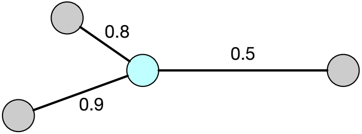
Different conversations can be connected by only one mistake!

Conversation Identification by SImilarity Ranking (CISIR)

High-rank Similarity Graph

For each message, only focus on pairs with top-r similarity scores.

- Rely on the redundancy of highly confident relations
- Discard edges whose scores are not top-r for any message
- Linear computation time with heaps when r is constant

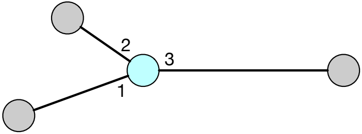


Conversation Identification by SImilarity Ranking (CISIR)

High-rank Similarity Graph

For each message, only focus on pairs with top-r similarity scores.

- Rely on the redundancy of highly confident relations
- Discard edges whose scores are not top-r for any message
- Linear computation time with heaps when r is constant



Conversation Identification by SImilarity Ranking (CISIR)

High-rank Similarity Graph

For each message, only focus on pairs with top-r similarity scores.

- Rely on the redundancy of highly confident relations
- Discard edges whose scores are not top-r for any message
- Linear computation time with heaps when r is constant

Suppose $r = 2$

Conversation Identification by SImilarity Ranking (CISIR)

High-rank Similarity Graph

For each message, only focus on pairs with top-r similarity scores.

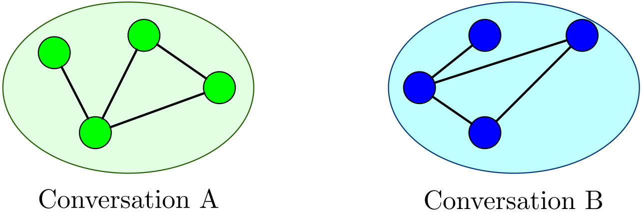
- Rely on the redundancy of highly confident relations
- Discard edges whose scores are not top-r for any message
- Linear computation time with heaps when r is constant

Conversation Identification by SImilarity Ranking (CISIR)

High-rank Similarity Graph

For each message, only focus on pairs with top-r similarity scores.

- Rely on the redundancy of highly confident relations
- Discard edges whose scores are not top-r for any message
- Linear computation time with heaps when r is constant



Experimental Datasets

- Four publicly available datasets
  - Three synthetic large-scale datasets from Reddit.com
  - One real dataset from IRC channels
- Reddit Datasets
  - All posts and comments posted from June 2016 to May 2017
  - Manually merge comments under posts as interleaved conversations
- IRC Dataset
  - Conversations about Linux for about five hours
  - Ground truths (conversations) are human-annotated

Dataset	Reddit			IRC
	gadgets	iPhone	politics	
Messages	8,518	12,433	105,663	497
Conversations	287	617	3,671	39
Speakers	5,185	5,231	25,289	71
Train/Valid Pairs	3,445	5,556	244,492	5,995
Test Pairs	27,565	44,450	1,955,943	47,966

Experimental Datasets

- Four publicly available datasets
  - Three synthetic large-scale datasets from Reddit.com
  - One real dataset from IRC channels
- Reddit Datasets
  - All posts and comments posted from June 2016 to May 2017
  - Manually merge comments under posts as interleaved conversations
- IRC Dataset
  - Conversations about Linux for about five hours
  - Ground truths (conversations) are human-annotated

Dataset	Reddit			IRC
	gadgets	iPhone	politics	
Messages	8,518	12,433	105,663	497
Conversations	287	617	3,671	39
Speakers	5,185	5,231	25,289	71
Train/Valid Pairs	3,445	5,556	244,492	5,995
Test Pairs	27,565	44,450	1,955,943	47,966

Experimental Datasets

- Four publicly available datasets
  - Three synthetic large-scale datasets from Reddit.com
  - One real dataset from IRC channels
- Reddit Datasets
  - All posts and comments posted from June 2016 to May 2017
  - Manually merge comments under posts as interleaved conversations
- IRC Dataset
  - Conversations about Linux for about five hours
  - Ground truths (conversations) are human-annotated

Dataset	Reddit			IRC
	gadgets	iPhone	politics	
Messages	8,518	12,433	105,663	497
Conversations	287	617	3,671	39
Speakers	5,185	5,231	25,289	71
Train/Valid Pairs	3,445	5,556	244,492	5,995
Test Pairs	27,565	44,450	1,955,943	47,966

Experimental Datasets

- Four publicly available datasets
  - Three synthetic large-scale datasets from Reddit.com
  - One real dataset from IRC channels
- Reddit Datasets
  - All posts and comments posted from June 2016 to May 2017
  - Manually merge comments under posts as interleaved conversations
- IRC Dataset
  - Conversations about Linux for about five hours
  - Ground truths (conversations) are human-annotated

Dataset	Reddit			IRC
	gadgets	iPhone	politics	
Messages	8,518	12,433	105,663	497
Conversations	287	617	3,671	39
Speakers	5,185	5,231	25,289	71
Train/Valid Pairs	3,445	5,556	244,492	5,995
Test Pairs	27,565	44,450	1,955,943	47,966

Evaluation of Similarity Estimation (Stage 1)

- Evaluated by ranking metrics
  - P@1, MRR, and MAP
- Comparative Baselines
  - Two naïive baselines
    - Difference between posted times (TimeDiff)
    - Identicality of speakers (Speaker)
  - Two feature-based baselines
    - Similarity between bag-of-word features (Text-Sim)
    - Logistic regression with various features [Elsner, 2008]
  - Two deep learning baselines
    - Single-layer CNN: DeepQA
    - Multi-layer CNN with attention: ABCNN

Performance of Similarity Estimation (SHCNN)

Feature-based methods are better than naïve baselines.

Dataset	Reddit Datasets									IRC Dataset		
	gadgets			iPhone			politics					
Metric	P@1	MRR	MAP	P@1	MRR	MAP	P@1	MRR	MAP	P@1	MRR	MAP
TimeDiff	0.6916	0.8237	0.8170	0.6085	0.7651	0.7495	0.4412	0.6362	0.5644	0.3262	0.5180	0.4384
Speaker	0.5643	0.7046	0.7425	0.5364	0.6595	0.6590	0.4021	0.4620	0.3914	0.4356	0.6263	0.6891
Text-Sim	0.7913	0.8746	0.8440	0.7347	0.8318	0.7872	0.5245	0.6672	0.5326	0.3712	0.5269	0.3108
Elsner	0.7758	0.8651	0.8321	0.6809	0.7935	0.7471	0.4643	0.6132	0.4884	0.1094	0.1886	0.2063
DeepQA	0.8011	0.8755	0.8511	0.7156	0.8112	0.7766	0.5593	0.6759	0.5685	0.7811	0.8182	0.8050
ABCNN	0.8374	0.8511	0.8502	0.8112	0.8520	0.8118	0.7419	0.6221	0.6644	0.7008	0.4142	0.5858
SHCNN	**0.8834**	**0.9281**	**0.9005**	**0.8375**	**0.8944**	**0.8497**	**0.7696**	**0.8392**	**0.6967**	0.9785	**0.9838**	**0.9819**
SHCNN (L)	0.8470	0.9080	0.8702	0.8066	0.8792	0.8275	0.7225	0.8070	0.6438	**0.9807**	0.9834	0.9750
SHCNN (H)	0.8490	0.9105	0.8704	0.8158	0.8851	0.8313	0.7228	0.8110	0.6283	0.9635	0.9728	0.8632

Performance of Similarity Estimation (SHCNN)

Deep learning methods perform better than all other baseline methods.

Dataset	Reddit Datasets									IRC Dataset		
	gadgets			iPhone			politics					
Metric	P@1	MRR	MAP	P@1	MRR	MAP	P@1	MRR	MAP	P@1	MRR	MAP
TimeDiff	0.6916	0.8237	0.8170	0.6085	0.7651	0.7495	0.4412	0.6362	0.5644	0.3262	0.5180	0.4384
Speaker	0.5643	0.7046	0.7425	0.5364	0.6595	0.6590	0.4021	0.4620	0.3914	0.4356	0.6263	0.6891
Text-Sim	0.7913	0.8746	0.8440	0.7347	0.8318	0.7872	0.5245	0.6672	0.5326	0.3712	0.5269	0.3108
Elsner	0.7758	0.8651	0.8321	0.6809	0.7935	0.7471	0.4643	0.6132	0.4884	0.1094	0.1886	0.2063
DeepQA	0.8011	0.8755	0.8511	0.7156	0.8112	0.7766	0.5593	0.6759	0.5685	0.7811	0.8182	0.8050
ABCNN	0.8374	0.8511	0.8502	0.8112	0.8520	0.8118	0.7419	0.6221	0.6644	0.7008	0.4142	0.5858
SHCNN	**0.8834**	**0.9281**	**0.9005**	**0.8375**	**0.8944**	**0.8497**	**0.7696**	**0.8392**	**0.6967**	0.9785	**0.9838**	**0.9819**
SHCNN (L)	0.8470	0.9080	0.8702	0.8066	0.8792	0.8275	0.7225	0.8070	0.6438	**0.9807**	0.9834	0.9750
SHCNN (H)	0.8490	0.9105	0.8704	0.8158	0.8851	0.8313	0.7228	0.8110	0.6283	0.9635	0.9728	0.8632

Performance of Similarity Estimation (SHCNN)

Our proposed SHCNN outperforms all of the baseline methods.

Dataset	Reddit Datasets									IRC Dataset		
	gadgets			iPhone			politics					
Metric	P@1	MRR	MAP	P@1	MRR	MAP	P@1	MRR	MAP	P@1	MRR	MAP
TimeDiff	0.6916	0.8237	0.8170	0.6085	0.7651	0.7495	0.4412	0.6362	0.5644	0.3262	0.5180	0.4384
Speaker	0.5643	0.7046	0.7425	0.5364	0.6595	0.6590	0.4021	0.4620	0.3914	0.4356	0.6263	0.6891
Text-Sim	0.7913	0.8746	0.8440	0.7347	0.8318	0.7872	0.5245	0.6672	0.5326	0.3712	0.5269	0.3108
Elsner	0.7758	0.8651	0.8321	0.6809	0.7935	0.7471	0.4643	0.6132	0.4884	0.1094	0.1886	0.2063
DeepQA	0.8011	0.8755	0.8511	0.7156	0.8112	0.7766	0.5593	0.6759	0.5685	0.7811	0.8182	0.8050
ABCNN	0.8374	0.8511	0.8502	0.8112	0.8520	0.8118	0.7419	0.6221	0.6644	0.7008	0.4142	0.5858
SHCNN	**0.8834**	**0.9281**	**0.9005**	**0.8375**	**0.8944**	**0.8497**	**0.7696**	**0.8392**	**0.6967**	0.9785	**0.9838**	**0.9819**
SHCNN (L)	0.8470	0.9080	0.8702	0.8066	0.8792	0.8275	0.7225	0.8070	0.6438	**0.9807**	0.9834	0.9750
SHCNN (H)	0.8490	0.9105	0.8704	0.8158	0.8851	0.8313	0.7228	0.8110	0.6283	0.9635	0.9728	0.8632

Performance of Similarity Estimation (SHCNN)

SHCNN still performs well while using only high-/low- level representations.

Dataset	Reddit Datasets									IRC Dataset		
	gadgets			iPhone			politics					
Metric	P@1	MRR	MAP	P@1	MRR	MAP	P@1	MRR	MAP	P@1	MRR	MAP
TimeDiff	0.6916	0.8237	0.8170	0.6085	0.7651	0.7495	0.4412	0.6362	0.5644	0.3262	0.5180	0.4384
Speaker	0.5643	0.7046	0.7425	0.5364	0.6595	0.6590	0.4021	0.4620	0.3914	0.4356	0.6263	0.6891
Text-Sim	0.7913	0.8746	0.8440	0.7347	0.8318	0.7872	0.5245	0.6672	0.5326	0.3712	0.5269	0.3108
Elsner	0.7758	0.8651	0.8321	0.6809	0.7935	0.7471	0.4643	0.6132	0.4884	0.1094	0.1886	0.2063
DeepQA	0.8011	0.8755	0.8511	0.7156	0.8112	0.7766	0.5593	0.6759	0.5685	0.7811	0.8182	0.8050
ABCNN	0.8374	0.8511	0.8502	0.8112	0.8520	0.8118	0.7419	0.6221	0.6644	0.7008	0.4142	0.5858
SHCNN	**0.8834**	**0.9281**	**0.9005**	**0.8375**	**0.8944**	**0.8497**	**0.7696**	**0.8392**	**0.6967**	0.9785	**0.9838**	**0.9819**
SHCNN (L)	0.8470	0.9080	0.8702	0.8066	0.8792	0.8275	0.7225	0.8070	0.6438	**0.9807**	0.9834	0.9750
SHCNN (H)	0.8490	0.9105	0.8704	0.8158	0.8851	0.8313	0.7228	0.8110	0.6283	0.9635	0.9728	0.8632

Evaluation of Conversation Disentanglement (Stage 2)

- Evaluated by clustering metrics
  - NMI, ARI, and F1
- Comparative Baselines
  - Two naïve baselines
    - Blocks of 10 messages (Block-10)
    - Messages of individual speakers (Speaker)
  - Embedding-based method
    - Doc2Vec with affinity propagation (Doc2Vec)
  - Two methods based on single-pass clustering
    - Context-based message expansion (CBME) [Wang, 2009]
    - Cluster with char- and content-specific features (GTM) [Elsner, 2008]

Performance of Conversation Disentanglement (CISIR)

Distances between Doc2Vec vectors cannot reveal conversations.

Dataset	Reddit Datasets									IRC Dataset		
	gadgets			iPhone			politics					
Metric	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI	F1
Doc2Vec	0.1757	0.0008	0.0589	0.2318	0.0002	0.0718	0.2672	0.0001	0.0506	0.2046	0.0048	0.1711
Block-10	0.7745	0.1840	0.3411	0.8203	0.2349	0.4251	0.8338	0.1724	0.3451	0.4821	0.0819	0.2087
Speaker	0.7647	0.0440	0.2094	0.7861	0.1001	0.3339	0.7480	0.0637	0.2207	0.7394	0.4572	0.6310
CBME	0.6913	0.0212	0.1465	0.7280	0.0339	0.1966	0.7883	0.0165	0.1382	0.2818	0.0324	0.1970
GTM	0.7942	0.1787	0.2986	0.8198	0.0536	0.2566	0.8496	0.3076	0.4292	0.0226	0.0001	0.2064
CISIR	0.8254	0.4287	0.4939	0.8552	0.4236	0.5187	0.8825	0.3561	0.4950	0.9330	0.9543	0.8798

Performance of Conversation Disentanglement (CISIR)

Naïve baselines have better performance compared to Doc2Vec.

Dataset	Reddit Datasets									IRC Dataset		
	gadgets			iPhone			politics					
Metric	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI	F1
Doc2Vec	0.1757	0.0008	0.0589	0.2318	0.0002	0.0718	0.2672	0.0001	0.0506	0.2046	0.0048	0.1711
Block-10	0.7745	0.1840	0.3411	0.8203	0.2349	0.4251	0.8338	0.1724	0.3451	0.4821	0.0819	0.2087
Speaker	0.7647	0.0440	0.2094	0.7861	0.1001	0.3339	0.7480	0.0637	0.2207	0.7394	0.4572	0.6310
CBME	0.6913	0.0212	0.1465	0.7280	0.0339	0.1966	0.7883	0.0165	0.1382	0.2818	0.0324	0.1970
GTM	0.7942	0.1787	0.2986	0.8198	0.0536	0.2566	0.8496	0.3076	0.4292	0.0226	0.0001	0.2064
CISIR	0.8254	0.4287	0.4939	0.8552	0.4236	0.5187	0.8825	0.3561	0.4950	0.9330	0.9543	0.8798

Performance of Conversation Disentanglement (CISIR)

Single-pass clustering methods have the better performance in the Reddit datasets but the worse performance in the IRC dataset.

Dataset	Reddit Datasets									IRC Dataset		
	gadgets			iPhone			politics					
Metric	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI	F1
Doc2Vec	0.1757	0.0008	0.0589	0.2318	0.0002	0.0718	0.2672	0.0001	0.0506	0.2046	0.0048	0.1711
Block-10	0.7745	0.1840	0.3411	0.8203	0.2349	0.4251	0.8338	0.1724	0.3451	0.4821	0.0819	0.2087
Speaker	0.7647	0.0440	0.2094	0.7861	0.1001	0.3339	0.7480	0.0637	0.2207	0.7394	0.4572	0.6310
CBME	0.6913	0.0212	0.1465	0.7280	0.0339	0.1966	0.7883	0.0165	0.1382	0.2818	0.0324	0.1970
GTM	0.7942	0.1787	0.2986	0.8198	0.0536	0.2566	0.8496	0.3076	0.4292	0.0226	0.0001	0.2064
CISIR	0.8254	0.4287	0.4939	0.8552	0.4236	0.5187	0.8825	0.3561	0.4950	0.9330	0.9543	0.8798

Performance of Conversation Disentanglement (CISIR)

Our proposed CISIR outperforms all of the baseline methods.

Dataset	Reddit Datasets									IRC Dataset		
	gadgets			iPhone			politics					
Metric	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI	F1
Doc2Vec	0.1757	0.0008	0.0589	0.2318	0.0002	0.0718	0.2672	0.0001	0.0506	0.2046	0.0048	0.1711
Block-10	0.7745	0.1840	0.3411	0.8203	0.2349	0.4251	0.8338	0.1724	0.3451	0.4821	0.0819	0.2087
Speaker	0.7647	0.0440	0.2094	0.7861	0.1001	0.3339	0.7480	0.0637	0.2207	0.7394	0.4572	0.6310
CBME	0.6913	0.0212	0.1465	0.7280	0.0339	0.1966	0.7883	0.0165	0.1382	0.2818	0.0324	0.1970
GTM	0.7942	0.1787	0.2986	0.8198	0.0536	0.2566	0.8496	0.3076	0.4292	0.0226	0.0001	0.2064
CISIR	**0.8254**	**0.4287**	**0.4939**	**0.8552**	**0.4236**	**0.5187**	**0.8825**	**0.3561**	**0.4950**	**0.9330**	**0.9543**	**0.8798**

Conclusions

- Focused on the task of conversation disentanglement
- Proposed a two-stage approach
  - (1) Similarity estimation
  - (2) Conversation Disentanglement
- Proposed SHCNN for estimating conversation-level similarity
- Proposed CISIR to disentangle conversations
- Conducted experiments on four datasets
  - including 3 large-scale and 1 real interleaved conversation datasets
- Outperformed several competitive baseline methods

Thanks for your attention! Questions?

Conclusions

- Focused on the task of conversation disentanglement
- Proposed a two-stage approach
  - (1) Similarity estimation
  - (2) Conversation Disentanglement
- Proposed SHCNN for estimating conversation-level similarity
- Proposed CISIR to disentangle conversations
- Conducted experiments on four datasets
  - including 3 large-scale and 1 real interleaved conversation datasets
- Outperformed several competitive baseline methods

Thanks for your attention! Questions?

Conclusions

- Focused on the task of conversation disentanglement
- Proposed a two-stage approach
  - (1) Similarity estimation
  - (2) Conversation Disentanglement
- Proposed SHCNN for estimating conversation-level similarity
- Proposed CISIR to disentangle conversations
- Conducted experiments on four datasets
  - including 3 large-scale and 1 real interleaved conversation datasets
- Outperformed several competitive baseline methods

Thanks for your attention! Questions?

Conclusions

- Focused on the task of conversation disentanglement
- Proposed a two-stage approach
  - (1) Similarity estimation
  - (2) Conversation Disentanglement
- Proposed SHCNN for estimating conversation-level similarity
- Proposed CISIR to disentangle conversations
- Conducted experiments on four datasets
  - including 3 large-scale and 1 real interleaved conversation datasets
- Outperformed several competitive baseline methods

Thanks for your attention! Questions?

Conclusions

- Focused on the task of conversation disentanglement
- Proposed a two-stage approach
  - (1) Similarity estimation
  - (2) Conversation Disentanglement
- Proposed SHCNN for estimating conversation-level similarity
- Proposed CISIR to disentangle conversations
- Conducted experiments on four datasets
  - including 3 large-scale and 1 real interleaved conversation datasets
- Outperformed several competitive baseline methods

Thanks for your attention! Questions?

Conclusions

- Focused on the task of conversation disentanglement
- Proposed a two-stage approach
  - (1) Similarity estimation
  - (2) Conversation Disentanglement
- Proposed SHCNN for estimating conversation-level similarity
- Proposed CISIR to disentangle conversations
- Conducted experiments on four datasets
  - including 3 large-scale and 1 real interleaved conversation datasets
- Outperformed several competitive baseline methods

Thanks for your attention! Questions?

Conclusions

- Focused on the task of conversation disentanglement
- Proposed a two-stage approach
  - (1) Similarity estimation
  - (2) Conversation Disentanglement
- Proposed SHCNN for estimating conversation-level similarity
- Proposed CISIR to disentangle conversations
- Conducted experiments on four datasets
  - including 3 large-scale and 1 real interleaved conversation datasets
- Outperformed several competitive baseline methods

Thanks for your attention! Questions?