Spoken Language Understanding for Abstractive Meeting Summarization

Ph.D Thesis Defense

Guokan Shang
École Polytechnique & LINAGORA

Directeur: Michalis Vazirgiannis
Co-encadrant: Jean-Pierre Lorré

January 28, 2021

LaTeX of the slides: `https://www.overleaf.com/read/jpdcrkfryjsh`

Publications

Guokan Shang, Wensi Ding, et al. (July 2018). "Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization". In: **ACL 2018 - Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, pp. 664–674. URL: https://www.aclweb.org/anthology/P18-1062

Guokan Shang, Antoine Tixier, et al. (Dec. 2020b). "Speaker-change Aware CRF for Dialogue Act Classification". In: **COLING 2020 - Proceedings of the 28th International Conference on Computational Linguistics**. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 450–464. URL: https://www.aclweb.org/anthology/2020.coling-main.40

Guokan Shang, Antoine Tixier, et al. (Dec. 2020a). "Energy-based Self-attentive Learning of Abstractive Communities for Spoken Language Understanding". In: **AACL-IJCNLP 2020 - Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing**. Suzhou, China: Association for Computational Linguistics, pp. 313–327. URL: https://www.aclweb.org/anthology/2020.aacl-main.34

Outline

Outline

Introduction

People spend a lot of their time in meetings (Romano and Nunamaker 2001)

- essential and inevitable
- sometimes costly, unproductive, and dissatisfying

Booming era of Artificial Intelligence (Lu 2019)

- conversational agents, autonomous driving, and healthcare industry

AI-powered meeting assistant LinTO (Lorré et al. 2019)

- understands and assists

Fields

- Spoken Language Understanding (X. Huang et al. 2001; Tur and De Mori 2011)
  - Natural Language Processing, Machine Learning, and Automatic Speech Recognition
- Meeting Summarization (Carenini, Murray, and Ng 2011)

LinTO

https://linto.ai

Overview of contributions

- **Abstractive meeting summarization**
  - transcription $\xrightarrow{\text{generates}}$ summary
  - A fully unsupervised framework based on multi-sentence compression graphs and budgeted submodular maximization.
- **Dialogue act classification**
  - utterance $\xrightarrow{\text{assigns}}$ dialogue act label
  - A modified neural conditional random field layer that takes speaker-change into account.
- **Abstractive community detection**
  - utterances $\xrightarrow{\text{groups}}$ abstractive communities
  - An energy-based learning approach, a general triplet sampling scheme, and a contextual utterance encoder featuring self-attention mechanisms.

Outline

Basic concepts

Text representation

$t, d, D \rightarrow$ word (term), sentence, document

- Bag-of-words
  - $\mathit{TF\text{-}IDF}(t, d, D) = \mathit{TF}(t, d) \times \mathit{IDF}(t, D)$
  - $\mathit{IDF}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$
- Graph-of-words
  - $\mathit{TW\text{-}IDF}(t, d, D) = \mathit{TW}(t, d) \times \mathit{IDF}(t, D)$
  - $\mathit{TW} \rightarrow$ centrality measures
- Word embedding
  - CBOW and Skip-gram models

Evaluation

- Accuracy, Precision, Recall, and F1-score
- ROUGE-1/2/SU4/L/etc.
  - ROUGE-1 R $= \frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$

(Christopher D Manning, Schütze, and Raghavan 2008; Mihalcea and Tarau 2004; François Rousseau and Vazirgiannis 2013; Mikolov, K. Chen, et al. 2013; Mikolov, Sutskever, et al. 2013; C.-Y. Lin 2004)

information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources

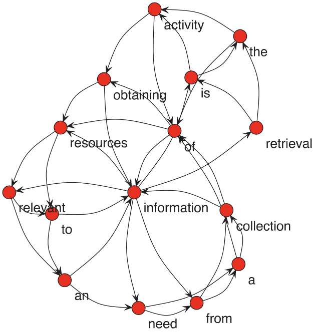


Figure: Example of an unweighted directed GoW in which an edge indicates at least one directed co-occurence of the two terms in a window of size 3 in the text. (François Rousseau and Vazirgiannis 2013)

Datasets

AMI corpus (McCowan et al. 2005)

- 137 scenario-driven meetings (65 hours)
- 4 participants play the roles within a fictive electronics company, as a design team, to develop a new television remote control.

ICSI corpus (Janin et al. 2003)

- 75 naturally-occurring meetings (72 hours)
- 6 members (on average per meeting) from research groups discuss specialized and technical topics.

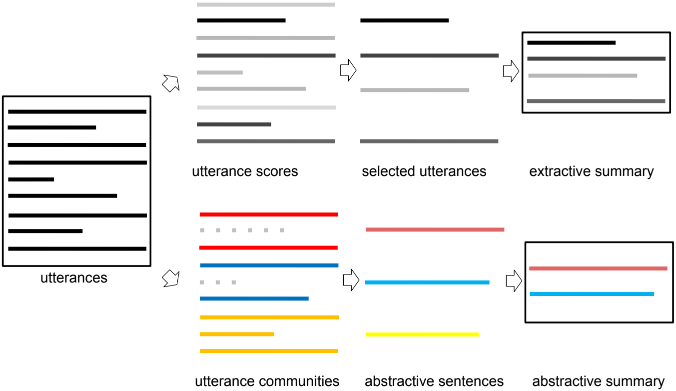
Annotations:

- speech transcription
- extractive summary
- abstractive summary
- abstractive-extractive linking

Outline

Introduction

Abstractive summaries are preferred to extractive ones by human judges. (Murray, Carenini, and Ng 2010)



Spontaneous multi-party meeting speech transcription is made of often ill-formed and ungrammatical text fragments (called *utterances*).

⇒ *Summarizing transcription requires approaches that differ from traditional document summarization.*

Related work

Keyword extraction

- Meladianos et al. 2017
  - keywords are influential spreaders within their graph-of-words
  - identified via graph degeneracy, k-core decomposition, CoreRank

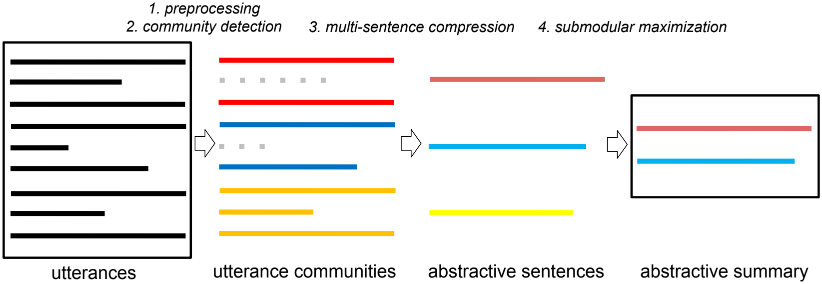
Multi-sentence compression

- Filippova 2010
  - unsupervised, simple NLG approach based on word graph
  - edge-weights, heuristics $\rightarrow$ k-shortest paths $\rightarrow$ re-ranking $\rightarrow$ the best path
- Boudin and Morin 2013
  - re-ranking taking into account information coverage (Keyphrases, TextRank)

Meeting summarization

- Mehdad et al. 2013 *abstractive*
  - supervised abstractive community detection method
  - each community is fused with Filippova's approach +
  - WordNet to capture synonymy and hypo/hypernymy when building graph
  - re-ranking taking into account information coverage (TF-IDF scores) and grammaticality (via a language model)
- Tixier et al. 2017 *extractive*
  - submodularity for summarization (H. Lin and Bilmes 2010; H. Lin 2012)
  - coverage term based on k-core decomposition of graph-of-words

Pipeline


1. preprocessing
2. community detection
3. multi-sentence compression
4. submodular maximization
utterances
utterance communities
abstractive sentences
abstractive summary

1. Text Preprocessing & 2. Utterance Community Detection

Preprocessing

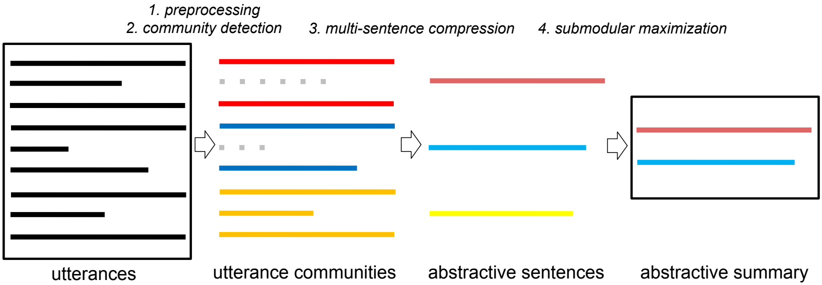
- Initial ellipsis: *'kay, 'til, 'em → okay, until, them*
- Consecutive repeated unigram and bigram terms: *remote control remote control → remote control*
- ASR tags are filtered out: *<vocalsound>*
- Filler words are discarded: *uh-huh, okay well, by the way*
- Consecutive stopwords at head and tail of utterance are stripped
- Utterances containing less than 3 non-stopwords are pruned out

Clustering

Group together the utterances that should be summarized by a common abstractive sentence. (Murray, Carenini, and Ng 2012)

1. Utterances → TF-IDF weight matrix
2. Latent Semantic Analysis
3. K-means algorithm (on the SVD result) - 35 to 50 clusters

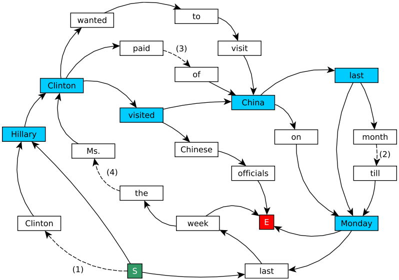
Pipeline


1. preprocessing
2. community detection
3. multi-sentence compression
4. submodular maximization
utterances
utterance communities
abstractive sentences
abstractive summary

3. Multi-Sentence Compression Graph (MSCG)

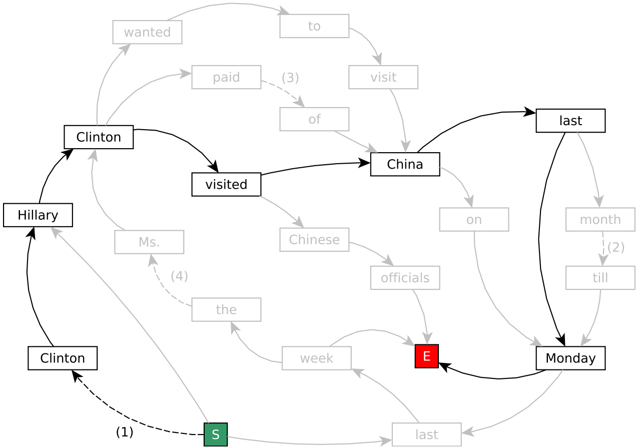
Generate an abstractive sentence for each utterance community with MSCG.

1. The wife of a former U.S. president Bill Clinton Hillary Clinton visited China last Monday
2. Hillary Clinton wanted to visit China last month but postponed her plans till Monday last week
3. Hillary Clinton paid a visit to the People Republic of China on Monday
4. Last week the Secretary of State Ms. Clinton visited Chinese officials



⇒ *Redundancy provides a reliable way of generating grammatical sentences.* (Filippova 2010)

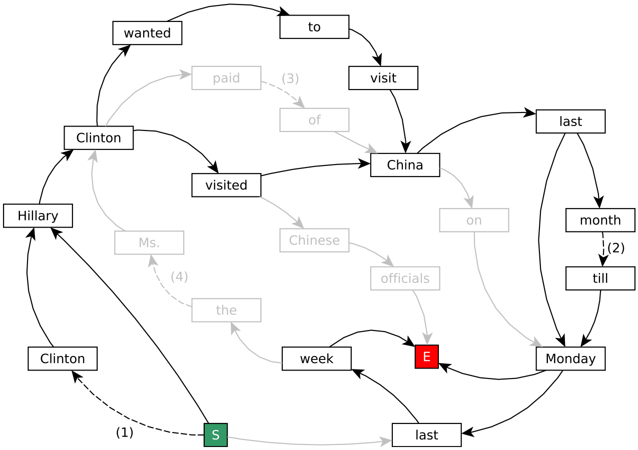
3.1. MSCG Building (1/4)



(1) *The wife of a former U.S. president Bill* Clinton Hillary Clinton visited China last Monday[1]

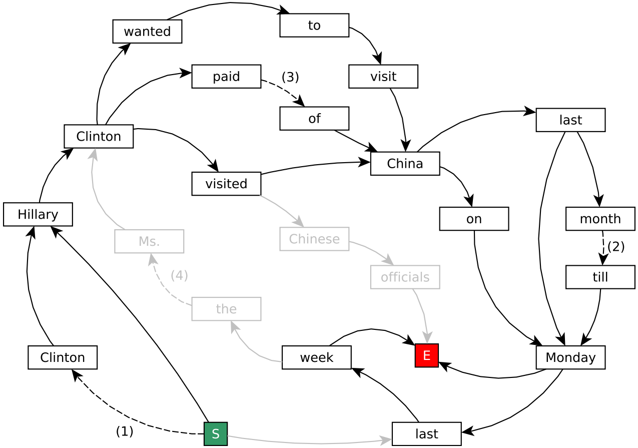
[1] Italicized fragments from the sentences are replaced with dashed arrow for clarity in the graph.

3.1. MSCG Building (2/4)



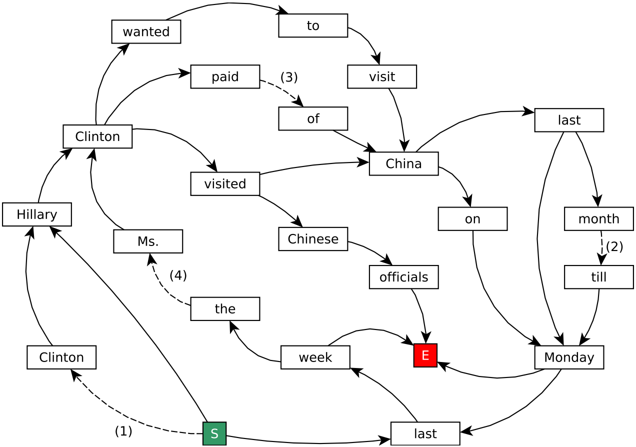
(2) Hillary Clinton wanted to visit China last month *but postponed her plans* till Monday last week

3.1. MSCG Building (3/4)



(3) Hillary Clinton paid *a visit to the People Republic of* China on Monday

3.1. MSCG Building (4/4)

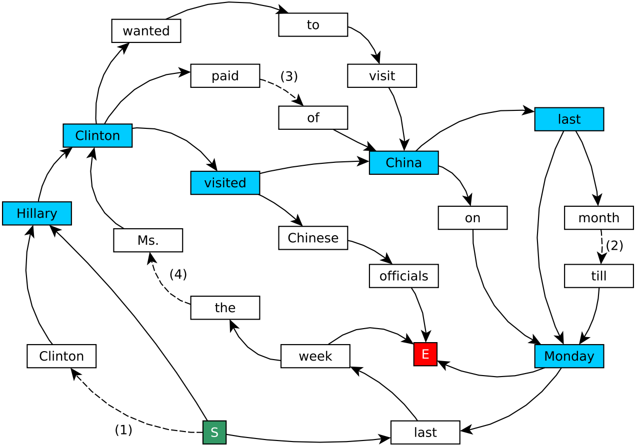


(4) Last week the *Secretary of State* Ms. Clinton visited Chinese officials

⇒ *Every input sentence corresponds to a loopless path in the graph.*

⇒ *There are many other paths.*

3.1. Objective of MSCG Building



⇒ *Find the best compression path: **Hilary Clinton visited China last Monday**.*

3.2. Edge Weight Assignment

Final edge weight (the lower the better):

$$w'''(p_i, p_j) = \frac{w'(p_i, p_j)}{w''(p_i, p_j)} \quad (1)$$

- **Local co-occurrence statistics** (Filippova 2010):

$$w'(p_i, p_j) = \frac{freq(p_i) + freq(p_j)}{\sum_{P \in G', p_i, p_j \in P} \mathrm{diff}(P, p_i, p_j)^{-1}} \quad (2)$$

$freq(p_i)$: number of words mapped to the node p_i.
$diff(P, p_i, p_j)^{-1}$: inverse of the distance between p_i and p_j in path P.

Favors edges between words that frequently appear close to each other (*word association*).

- **Global exterior knowledge**: Word Attraction Score (R. Wang, W. Liu, and McDonald 2014):

$$w''(p_i, p_j) = \frac{freq(p_i) \times freq(p_j)}{d^2_{p_i, p_j}} \quad (3)$$

d_{p_i, p_j}: Euclidean distance of the word embedding vectors for p_i and p_j.

Favor paths going through salient nodes that are close in the embedding space (*semantic relatedness*).

3.3. Path Selection and Reranking (1/2)

- Path score as its **cumulative edge weights** (the lowest is the best compression path):

$$W(P) = \sum_{i=1}^{|P|-1} w'''(p_i, p_{i+1}) \tag{4}$$

Reranking

The path with the lowest score does not guarantee its readability nor informativeness. (Boudin and Morin 2013)

$\Rightarrow$ *Reranking N best paths is necessary.*

3.3. Path Selection and Reranking (2/2)

- **Fluency** (Mehdad et al. 2013): estimate readability of MSCG path P based on a 3-gram language model

$$F(P) = \frac{\sum_{i=1}^{|P|} \log Pr(p_i | p_{i-n+1}^{i-1})}{\#n\text{-}gram} \tag{5}$$

- **Coverage** (Mehdad et al. 2013): estimate the information covered by P

$$C(P) = \frac{\sum_{p_i \in P} \text{TW-IDF}(p_i)}{\# p_i} \tag{6}$$

  TW: term CoreRank score of p_i in the GoW of the community. (Tixier, Malliaros, and Vazirgiannis 2016)

- **Diversity**: estimate the diversity of the information contained by P

$$D(P) = \frac{\sum_{j=1}^{k} 1_{\exists p_i \in P | p_i \in \text{cluster}_j}}{|P|} \tag{7}$$

  The number of different word clusters covered by the path
- **Final path score**: select the path with the lowest score per community

$$\text{score}(P) = \frac{W(P)}{|P| \times F(P) \times C(P) \times D(P)} \tag{8}$$

Diversity

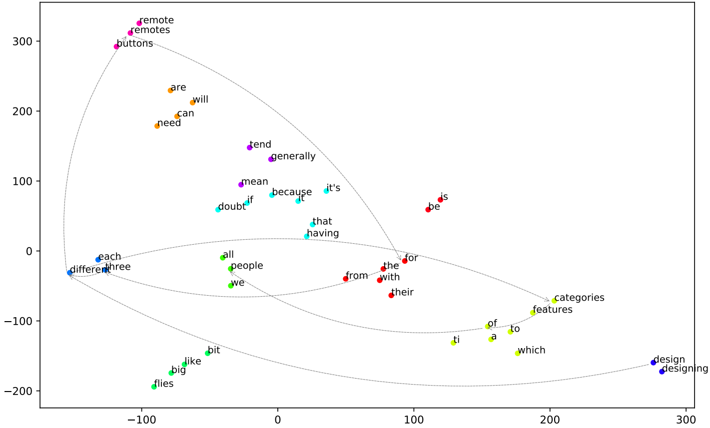
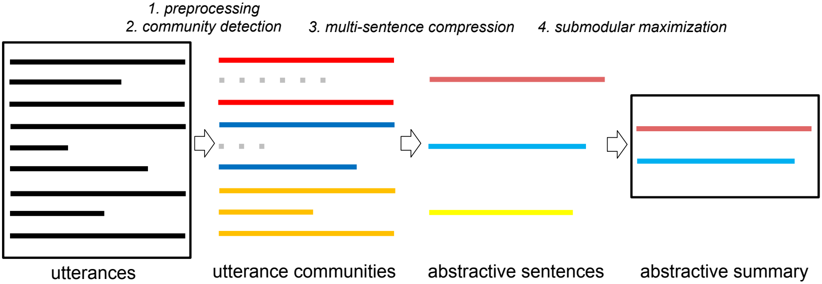


Figure: t-SNE visualization of the GoogleNews vectors of the words in an utterance community. Arrows join the words in the best compression path. Movements in the embedding space, as measured by the number of unique clusters covered by the path (here, 6/11), can provide a sense of the diversity of the compressed sentence, as formalized in Equation 7.

Pipeline


1. preprocessing
2. community detection
3. multi-sentence compression
4. submodular maximization
utterances
utterance communities
abstractive sentences
abstractive summary

4. Budgeted Submodular Maximization

Generate the final summary by selecting an optimal subset S from the set of abstractive sentences $\mathcal{S}$ under a budget constraint.

$$\boxed{\underset{S \subseteq \mathcal{S}}{\operatorname{argmax}} f(S) | \sum_{s \in S} cost_s \leq Budget}$$

NP-hard, but near-optimal performance can be guaranteed with a modified greedy algorithm (H. Lin and Bilmes 2010) that iteratively selects the sentence s that maximizes the ratio of summary quality function gain to scaled cost $f(G \cup s) - f(G) / cost_s^r$ (where G is the current subset and $r \geq 0$ is a scaling factor).

Summary quality function f is non-decreasing and **submodular** taking both coverage and diversity into account:

$$f(S) = c(S) + \lambda d(S)$$

$$c(S) = \sum_{s_i \in S} n_{s_i} w_{s_i}, \; d(S) = \sum_{j=1}^{k} 1_{\exists s_i \in S, s_i \in cluster_j}$$

$\lambda \geq 0$: trade-off parameter, n_{s_i}: number of occurrences of word s_i in S, w_{s_i}: CoreRank score of word s_i

Experimental setup

Baselines

- **Random** & **Longest Greedy** (Riedhammer et al. 2008)
- **TextRank** (Mihalcea and Tarau 2004) & **ClusterRank** (Garg et al. 2009)
- **Oracle** & **CoreRank Submodular** & **PageRank Submodular** (Tixier, Meladianos, and Vazirgiannis 2017)

Variants of our system

- Our System (**Baseline**) (Filippova 2010)
- Our System (**KeyRank**) (Boudin and Morin 2013)
- Our System (**FluCovRank**) (Mehdad et al. 2013)

Parameter tuning

- over fixed summary size: **350** / **450** words for AMI / ICSI corpus

Datasets & Metrics

- AMI / ICSI corpus (47/25 for development, 20/6 for test, 1/3 reference summaries)
- ROUGE-1/2/SU4

ROUGE Results AMI

	AMI ROUGE-1			AMI ROUGE-2			AMI ROUGE-SU4		
	R	P	F-1	R	P	F-1	R	P	F-1
Our System	41.83	34.44	37.25	8.22	6.95	7.43	15.83	13.70	14.51
Our System (Baseline)	41.56	34.37	37.11	7.88	6.66	7.11	15.36	13.20	14.02
Our System (KeyRank)	42.43	35.01	**37.86**	8.72	7.29	**7.84**	16.19	13.76	**14.71**
Our System (FluCovRank)	41.84	34.61	37.37	8.29	6.92	7.45	16.28	13.48	14.58
Oracle	40.49	34.65	**36.73**	8.07	7.35	**7.55**	15.00	14.03	**14.26**
CoreRank Submodular	41.14	32.93	36.13	8.06	6.88	7.33	14.84	13.91	14.18
PageRank Submodular	40.84	33.08	36.10	8.27	6.88	7.42	15.37	13.71	14.32
TextRank	39.55	32.60	35.25	7.67	6.43	6.90	14.87	12.87	13.62
ClusterRank	39.36	32.53	35.14	7.14	6.05	6.46	14.34	12.80	13.35
Longest Greedy	37.31	30.93	33.35	5.77	4.71	5.11	13.79	11.11	12.15
Random	39.42	32.48	35.13	6.88	5.89	6.26	14.07	12.70	13.17

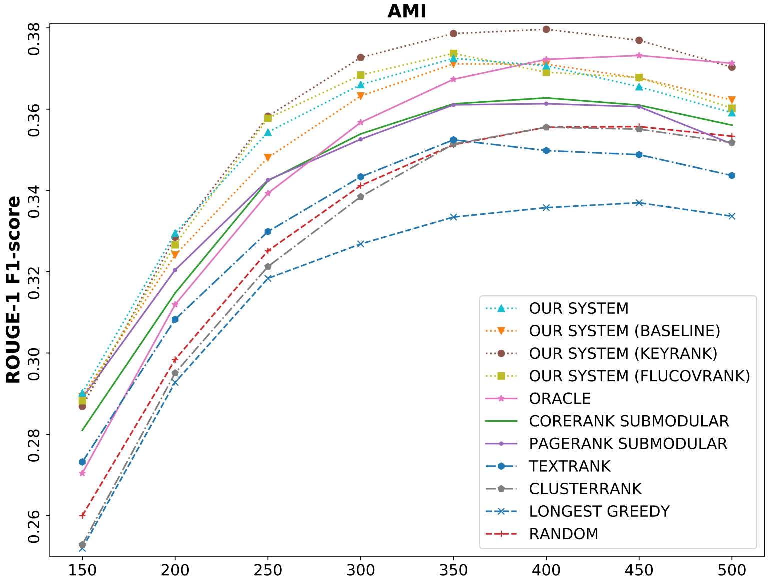
Table: Macro-averaged results for 350 word summaries (ASR transcriptions).

ROUGE Results ICSI

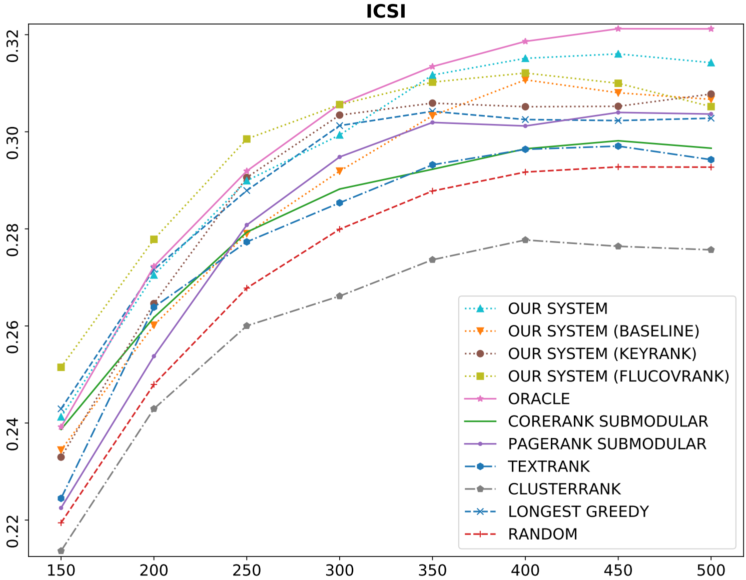
	ICSI ROUGE-1			ICSI ROUGE-2			ICSI ROUGE-SU4		
	R	P	F-1	R	P	F-1	R	P	F-1
Our System	36.99	28.12	**31.60**	5.41	4.39	4.79	13.10	10.17	**11.35**
Our System (Baseline)	36.39	27.20	30.80	5.19	4.12	4.55	12.59	9.70	10.86
Our System (KeyRank)	35.95	27.00	30.52	4.64	3.64	4.04	12.43	9.23	10.50
Our System (FluCovRank)	36.27	27.56	31.00	5.56	4.35	**4.83**	13.47	9.85	11.29
Oracle	37.91	28.39	**32.12**	5.73	4.82	**5.18**	13.35	10.73	**11.80**
CoreRank Submodular	35.22	26.34	29.82	4.36	3.76	4.00	12.11	9.58	10.61
PageRank Submodular	36.05	26.69	30.40	4.82	4.16	4.42	12.19	10.39	11.14
TextRank	34.89	26.33	29.70	4.60	3.74	4.09	12.42	9.43	10.64
ClusterRank	32.63	24.44	27.64	4.03	3.44	3.68	11.04	8.88	9.77
Longest Greedy	35.57	26.74	30.23	4.84	3.88	4.27	13.09	9.46	10.90
Random	34.78	25.75	29.28	4.19	3.51	3.78	11.61	9.37	10.29

Table: Macro-averaged results for 450 word summaries (ASR transcriptions).

ROUGE-1 F1-score



ROUGE-1 F1-score



Example Summary AMI TS3003c manual transcription of Our System

attract elderly people can use the remote control
changing channels button on the right side that would certainly yield great options for the design of the remote
personally i dont think that older people like to shake your remote control
imagine that the remote control and the docking station
remote control have to lay in your hand and right hand users
finding an attractive way to control the remote control
casing the manufacturing department can deliver a flat casing single or double curved casing
top of that the lcd screen would help in making the remote control easier
increase the price for which were selling our remote control
remote controls are using a onoff button still on the top
apply remote control on which you can apply different case covers
button on your docking station which you can push and then it starts beeping
surveys have indicated that especially wood is the material for older people
mobile phones so like the nokia mobile phones when you can change the case
greyblack colour for people prefer dark colours
brings us to the discussion about our concepts
docking station and small screen would be our main points of interest
industrial designer and user interface designer are going to work
innovativeness was about half of half as important as the fancy design
efficient and cheaper to put it in the docking station
case supplement and the buttons it really depends on the designer
start by choosing a case
deployed some trendwatchers to milan

Reference Summary AMI TS3003c

The project manager opened the meeting and recapped the decisions made in the previous meeting.
The marketing expert discussed his personal preferences for the design of the remote and presented the results of trend-watching reports, which indicated that there is a need for products which are fancy, innovative, easy to use, in dark colors, in recognizable shapes, and in a familiar material like wood.
The user interface designer discussed the option to include speech recognition and which functions to include on the remote.
The industrial designer discussed which options he preferred for the remote in terms of energy sources, casing, case supplements, buttons, and chips.
The team then discussed and made decisions regarding energy sources, speech recognition, LCD screens, chips, case materials and colors, case shape and orientation, and button orientation.
The team members will look at the corporate website.
The user interface designer will continue with what he has been working on.
The industrial designer and user interface designer will work together.
The remote will have a docking station.
The remote will use a conventional battery and a docking station which recharges the battery.
The remote will use an advanced chip.
The remote will have changeable case covers.
The case covers will be available in wood or plastic.
The case will be single curved.
Whether to use kinetic energy or a conventional battery with a docking station which recharges the remote.
Whether to implement an LCD screen on the remote.
Choosing between an LCD screen or speech recognition.
Using wood for the case.

Conclusion

Contributions

- A fully unsupervised framework, does not rely on any annotations, language-independent
  - based on MSCG and budgeted submodular maximization
- Novel edge weight assignment and path re-ranking strategy for the MSCG
  - based on word embeddings, graph-of-words, and graph degeneracy
- Code is publicly available:
  `https://bitbucket.org/dascim/acl2018_abssumm`

Future work

- improving the community detection phase (TF-IDF + k-means)
  $\Rightarrow$ a novel approach will be introduced in Section 5.

Outline

Introduction

Dialogue Act (DA) classification aims at assigning to each utterance in a conversation a DA label to represent its **communicative intention**.

- Useful annotations to many spoken language understanding tasks.

Change	Speaker	Utterance	DA
-	B	Of course I use,	sd
True	A	<laughter>.	x
True	B	credit cards.	+
False	B	I have a couple of credit cards	sd
True	A	**Yeah.**	b
True	B	and, uh, use them.	+
True	A	Uh-huh,	b
False	A	do you use them a lot?	**qy**
True	B	Oh, we try not to.	**ng**

Table: Fragment from SwDA conversation sw3332. **Statement**-non-opinion (sd), Non-verbal (x), Interruption (+), Acknowledge/Backchannel (b), Yes-No-**Question** (qy), Negative non-no **answers** (ng).

$\Rightarrow$ There are dependencies both at the **utterance level** and at the **label level**.

Related work

Multi-class classification

Consecutive DA labels are considered to be independent, predicted in isolation.

- **naive Bayes** (Grau et al. 2004), **Maxent** (Venkataraman et al. 2005; Ang, Y. Liu, and E. Shriberg 2005), or **SVM** (Y. Liu 2006).
- **Deep learning models** (Ries 1999; Khanpour, Guntakandla, and Nielsen 2016; Shen and H.-y. Lee 2016; Kalchbrenner and Blunsom 2013; J. Y. Lee and Dernoncourt 2016; Ortega and Vu 2017; Bothe et al. 2018)

Sequence labeling

DA labels for all the utterances in the conversation are classified together.

- **HMMs** (Stolcke et al. 2000; Surendran and Levow 2006; Tavafi et al. 2013) and **CRFs** (Lendvai and Geertzen 2007; Zimmermann 2009; Kim, Cavedon, and Baldwin 2010)
- Neural sequence labeling architectures: **BiLSTM-Softmax** (W. Li and Wu 2016; Tran, Zukerman, and Haffari 2017; Y. Liu et al. 2017) and **BiLSTM-CRF** (Kumar et al. 2018; Z. Chen et al. 2018; Raheja and Tetreault 2019; R. Li et al. 2019).

BiLSTM-CRF is able to capture the dependencies among consecutive **utterances** (with BiLSTM) and among consecutive DA **labels** (with CRF).

Motivation

The state-of-the-art works do not take into account the additional **speaker** input sequence.

- This is a major **limitation**.
- This extra input could greatly improve DA prediction.

Turn management (Sacks, Schegloff, and Jefferson 1974)

- Dialogue participants follow an underlying turn-taking system to occupy or release (not arbitrarily) the speaker role (Petukhova and Bunt 2009).
- $\Rightarrow$ DA transition should be conditioned both on the utterance transition and the **speaker-change** (not speaker-identifier).
- $\Rightarrow$ "A *Question* is usually followed by an *Answer*" (only partially true), + **[*if the speaker changed*]**.

To address the limitation, we propose a simple modification of the CRF layer that takes speaker-change into account.
We evaluate our modified CRF layer within the BiLSTM-CRF architecture.

BiLSTM-CRF

Notation

$X = \{\mathbf{x}^t\}_{t=1}^T$: the input utterance sequence, of length T.

$Y = \{y^t\}_{t=1}^T$: the target label sequence, where $y^t \in \mathcal{Y}$, the DA label set of size K.

We use y^t to denote the label and its integer index interchangeably.

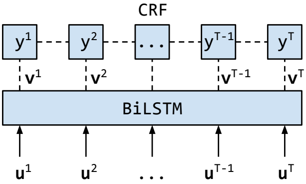


Figure: BiLSTM-CRF. $\{\mathbf{u}^t\}_{t=1}^T$ are utterance embeddings.

1. LSTM (text encoder): utterances $X = \{\mathbf{x}^t\}_{t=1}^T \rightarrow$ utterance embeddings $\{\mathbf{u}^t\}_{t=1}^T$.
2. BiLSTM: $\{\mathbf{u}^t\}_{t=1}^T \rightarrow$ conversation-level utterance representations $\{\mathbf{v}^t\}_{t=1}^T$.
3. CRF: $\{\mathbf{v}^t\}_{t=1}^T \rightarrow$ labels $Y = \{y^t\}_{t=1}^T$

CRF layer

CRF is a **discriminative** probabilistic graphical framework used to label sequences (Lafferty, McCallum, and Pereira 2001).

$$P(Y|X) = \frac{\exp(\psi(X, Y))}{\sum_{\tilde{Y}} \exp(\psi(X, \tilde{Y}))} \tag{9}$$

where $\psi(X, Y)$ is a feature function that assigns a ***path score*** to the label sequence Y, giving the input sequence X. $\tilde{Y}$ denotes one of all possible label sequences (paths).

$$\psi(X, Y) = \sum_{t=1}^{T} h(y^t, X) + \sum_{t=1}^{T-1} g(y^t, y^{t+1}) \tag{10}$$

$\psi(X, Y)$ is defined as the sum of ***emission scores*** (or state scores) and ***transition scores*** over all time steps.

$$h(y^t, X) = (\mathbf{W}\mathbf{v}^t + \mathbf{b})[y^t] \tag{11}$$

where the conversation-level utterance representation $\mathbf{v}^t$ is converted into a vector of size K.

$$g(y^t, y^{t+1}) = \mathbf{G}[y^t, y^{t+1}] \tag{12}$$

where $\mathbf{G}$ is the label transition matrix of size $K \times K$.

CRF layer

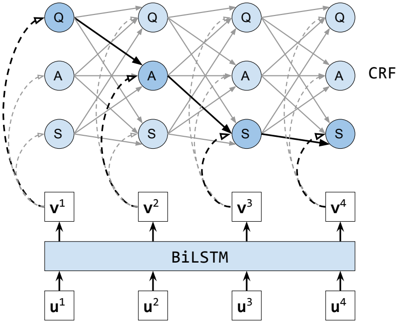


Figure: BiLSTM-CRF for an example.

For a training set of M conversations, the loss can be written as:

$$\mathcal{L} = \sum_{m=1}^{M} -\log P(Y^m|X^m) \tag{13}$$

At test time, the optimal label sequence, i.e., $Y^* = \text{argmax}_{\tilde{Y}} P(\tilde{Y}|X)$ for unseen X, is obtained with the Viterbi algorithm (Viterbi 1967), with polynomial complexity $O(TK^2)$.

Contribution

Notation

$S = \{s^t\}_{t=1}^{T}$: the sequence of speaker-identifiers.
$Z = \{z^{t,t+1}\}_{t=1}^{T-1}$: **the sequence of speaker-changes**, obtained by comparing neighbors in S.
E.g., $z^{2,3} = 0$ means the speaker does not change from time $t = 2$ to $t = 3$.

We extend the original CRF so that it considers as **additional input**, the sequence Z.

$$P(Y|X, Z) = \frac{\exp(\psi(X, Y, Z))}{\sum_{\tilde{Y}} \exp(\psi(X, \tilde{Y}, Z))} \tag{14}$$

Specifically, transition scores in our modified CRF layer are computed as follows:

$$\begin{aligned} g(y^t, y^{t+1}, z^{t,t+1}) =& (1 - z^{t,t+1}) * \mathbf{G}_0[y^t, y^{t+1}] + \\ & z^{t,t+1} * \mathbf{G}_1[y^t, y^{t+1}] \end{aligned} \tag{15}$$

where $\mathbf{G}_0$ and $\mathbf{G}_1$ are label transition matrices of size $K \times K$, corresponding respectively to the **"speaker unchanged"** and **"speaker changed"** cases.

Dataset

Switchboard Dialogue Act (SwDA) dataset (Jurafsky, L. Shriberg, and Biasca 1997; Stolcke et al. 2000).

- telephonic conversations recorded between two randomly selected speakers talking about one of various general topics (air pollution, music, football, etc.).
- training, validation and testing partition of 1003, 112, and 19 conversations.
- utterances are annotated with **42** mutually exclusive DA labels
- Inter-annotator agreement is 84%.

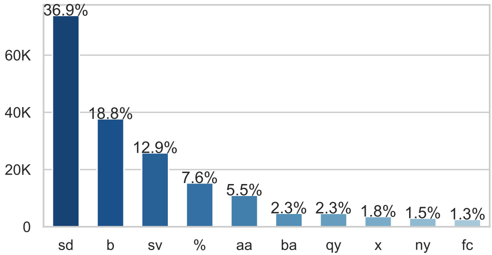


Figure: Counts and frequencies of the 10 most represented DA labels in the SwDA dataset. There are 200444 utterances in total.

Results

	Model	BiLSTM input	CRF extra input	Accuracy (% ± SD)
a)	Our CRF	$\mathbf{u}^t$	SC	**78.70** ± .37
a1)		$\mathbf{u}^t$ + SI	SC	78.32 ± .28
a2)		$\mathbf{u}^t$ + SC	SC	78.65 ± .47
b)	Vanilla CRF	$\mathbf{u}^t$	-	77.69 ± .38
b1)		$\mathbf{u}^t$ + SI	-	77.86 ± .61
b2)		$\mathbf{u}^t$ + SC	-	78.33 ± .71
c)	Softmax	$\mathbf{u}^t$	-	77.80 ± .48
c1)		$\mathbf{u}^t$ + SI	-	77.73 ± .44
c2)		$\mathbf{u}^t$ + SC	-	78.33 ± .49
a) + b) ensembling		$\mathbf{u}^t$	SC	**78.89** ± .20
a) + b) joint training		$\mathbf{u}^t$	SC	78.27 ± .47

Table: Results, averaged over 10 runs and 42 DA labels. SI: speaker-identifier, SC: speaker-change, $\mathbf{u}^t$: utterance embedding, ±: standard deviation.

Analysis

Our CRF vs. Vanilla CRF

- $\Rightarrow$ our model a) outperforms the base model b) by 1%, over 42 labels.
- $\Rightarrow$ The boost is greater than the gains of 0.26% (Y. Liu et al. 2017) and 0.09% (Bothe et al. 2018) reported by previous attempts at leveraging speaker information.

Confusion matrices

- 10 most frequent labels (91%) $\Rightarrow$ outperforms on a majority of them, but not on `sd`.
- 10 best predicted labels (20%) and 10 worst predicted labels (40%) $\Rightarrow$ Our model is most useful for the difficult and rare DAs requiring speaker-change awareness.

Different ways of incorporating speaker information

- concatenate the one-hot encoded SI vector (of size 2) and the binary speaker-change vector (of size 1) with $\mathbf{u}^t$ the utterance embedding.

BiLSTM-CRF VS. BiLSTM-Softmax

- $\Rightarrow$ competitive, this finding is not surprising and consistent with the results reported in recent works on other tasks (Reimers and Gurevych 2017; J. Yang, Liang, and Zhang 2018; Cui and Zhang 2019).

Ensembling vs. joint training

- Ensembling: combines the predictions of the two trained models by averaging their emission and transition scores respectively.
- Joint training: $\mathbf{G}_{basis}[y^t, y^{t+1}] + (1 - z^{t,t+1}) * \mathbf{G}_0[y^t, y^{t+1}] + z^{t,t+1} * \mathbf{G}_1[y^t, y^{t+1}]$

Confusion matrices for the 10 most frequent labels

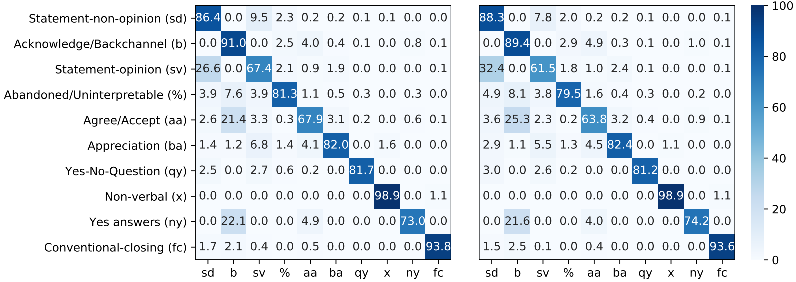


Figure: Normalized confusion matrices, averaged over 10 runs, for the 10 most frequent DA labels (90.9% of all annotations). Left: our model, right: base model. Rows (columns) correspond to true (predicted) classes.

		P	R	F1
Our	sd	80.49	86.36	83.32
	sv	71.54	67.41	69.42
Vanilla	sd	77.83	88.32	82.74
	sv	73.24	61.48	66.84

Table: Precison, Recall, and F1 score (%) of our model vs. base model on the `sd` and `sv` labels.

Visualization of transition matrices 1/2

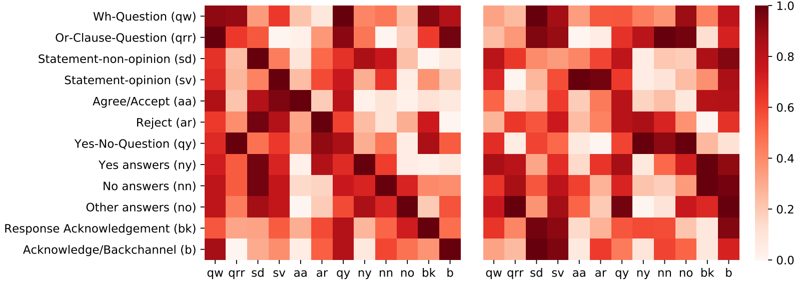


Figure: Normalized transition matrices (averaged over 10 runs). Left: $\mathbf{G}_0$ (speaker unchanged) and Right: $\mathbf{G}_1$ (speaker changed) of **our CRF layer**. The darker, the greater the score.

Visualization of transition matrices 2/2

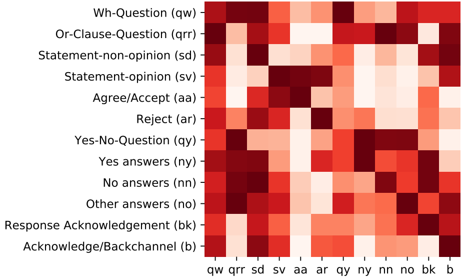


Figure: Normalized transition matrix (averaged over 10 runs). **G** of **vanilla CRF layer**. The darker, the greater the score.

Conclusion

Contributions

- A modified CRF layer that takes as extra input the sequence of speaker-changes was proposed. Code is publicly available: `https://bitbucket.org/guokan_shang/da-classification`.
- Experiments showed that our CRF layer outperforms vanilla CRF $\Rightarrow$ taking speaker information into consideration was beneficial.
- Visualizations confirmed that our improved CRF was able to learn complex speaker-change aware DA transition patterns in an end-to-end way.

Future work

- address the limitation of the Markov property of CRF layer
- capture longer-range dependencies within and among the three sequences: that of speakers, utterances, and DA labels.

Outline

Introduction 1/2

Abstractive summarization of conversations aims to take a transcription as input and produce an abstractive summary as output.

- **Subtask a**, *Abstractive Community Detection* (ACD), groups utterances according to whether they can be *jointly* summarized by a common abstractive sentence.
- **Subtask b**, NLG, generates an abstractive sentence for each group named *abstractive community* ⇒ forming the final abstractive summary.

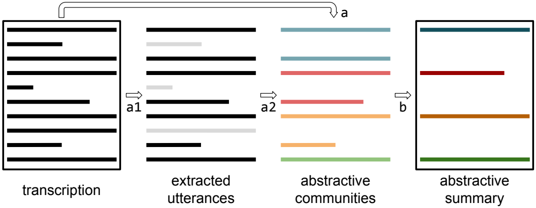


Figure: Abstractive summarization of conversations.

Introduction 2/2

ACD (Murray, Carenini, and Ng 2012) in two steps:

- **Step a1** extracts important/summary-worthy utterances from the transcription.
  - closely related to *extractive summarization* & extensively studied
- **Step a2** groups extracted utterances into abstractive communities.
  - plays a crucial role of *bridge* between two major types of summaries: extractive and abstractive & rarely explored
  - utterance clustering $\Leftarrow$ the focus of our work

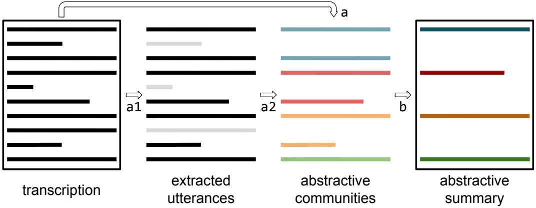


Figure: Abstractive summarization of conversations.

$\Rightarrow$ This $a1 \rightarrow a2 \rightarrow b$ process is more consistent with the way humans treat the summarization task (e.g., the creation of the AMI corpus (McCowan et al. 2005)).

Example of abstractive communities

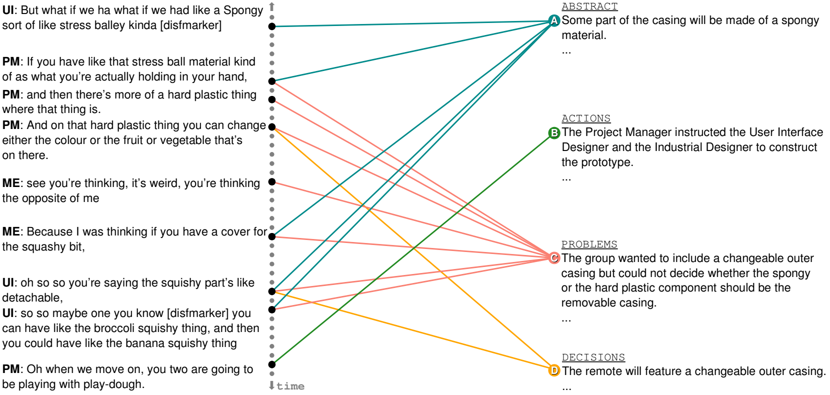


Figure: Example of ground truth human annotations from the ES2011c AMI meeting. Successive grey nodes on the left denote utterances in the transcription. Black nodes correspond to the utterances judged important. Sentences (e.g., A, B, C, D) from the abstractive summary are shown on the right. All utterances linked to the same abstractive sentence form one community.
⇒ Communities should capture more complex relationship than simple semantic similarity.

Related work

Supervised approaches

- Utterance graph + CONGA, edges are decided by a trained binary classifier (if or not two utterances are jointly summarizable)(Murray, Carenini, and Ng 2012).
- + an entailment graph for each community (Mehdad et al. 2013)

Unsupervised approaches

- Topic segmentation (Oya et al. 2014; Banerjee, Mitra, and Sugiyama 2015; Singla et al. 2017)
- TF-IDF + *k*-means (**Shang**, Ding, et al. 2018)

Our energy-based/deep metric learning approach

- We introduce a neural contextual utterance encoder featuring three types of self-attention mechanisms.
- We then train it using the siamese and triplet energy-based meta-architectures.
- We applied the Fuzzy c-Means clustering algorithm on the trained utterance embeddings in order to obtain abstractive communities.

Siamese & triplet energy-based architectures

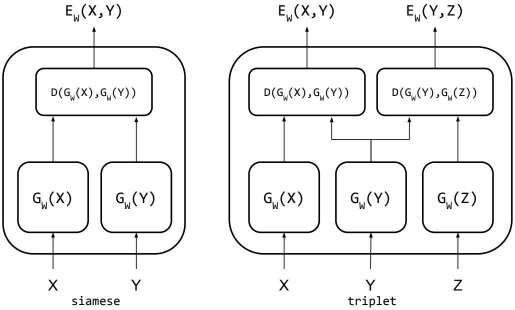


Figure: Siamese & triplet architectures

siamese (Bromley et al. 1994; Chopra, Hadsell, and LeCun 2005)

- objective: minimize the output energies (i.e., distances in the embedding space) $E_W(X^i, Y^i)$ associated with positive pairs, and maximize those associated with negative pairs.

triplet (Hoffer and Ailon 2015; J. Wang et al. 2014)

- objective: jointly minimize the positive-anchor energy $E_W(X^i, Y^i)$ and maximize the anchor-negative energy $E_W(Y^i, Z^i)$.

Utterance encoder 1/3

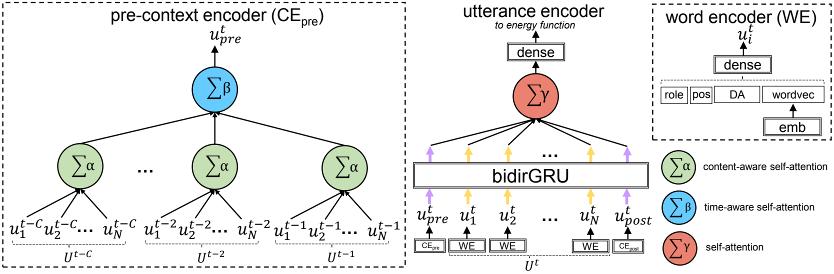


Figure: Our proposed utterance encoder G_W. Only the pre-context encoder is shown. C is the context size.

- **word encoder**: textual features (word embedding) + discourse features (role, position, dialogue act) $\rightarrow$ dense layer $\rightarrow$ $\mathbf{u}_i^t$
- **utterance encoder**: $\{\mathbf{u}^t_{\text{pre}}, \mathbf{u}^t_1, \ldots, \mathbf{u}^t_N, \mathbf{u}^t_{\text{post}}\}$ $\rightarrow$ BiGRU $\rightarrow$ **self-attention** (γ) (Vaswani et al. 2017; Z. Lin et al. 2017) $\rightarrow$ dense layer $\rightarrow$ $\mathbf{u}^t$

$$\mathbf{u}^t = \text{dense}\left(\sum_{i=1}^{N+2} \gamma_i^t \mathbf{h}_i^t\right) \quad \boldsymbol{\gamma}^t = \text{softmax}(\mathbf{u}_\gamma \cdot \tanh(\mathbf{W}_\gamma \mathbf{H}^t)) \tag{16}$$

Utterance encoder 2/3

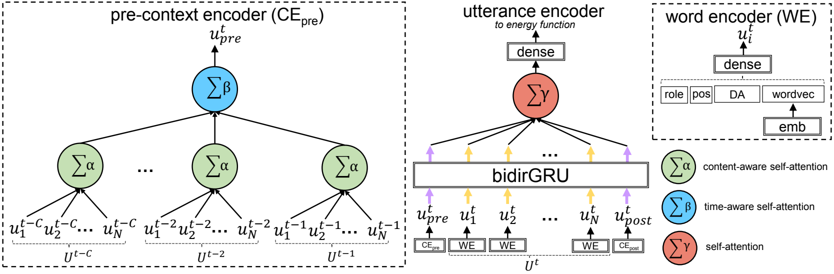


Figure: Our proposed utterance encoder G_W. Only the pre-context encoder is shown. C is the context size.

- **context encoder level 1**: $\mathbf{U}^{t-1} = \{\mathbf{u}_1^{t-1}, \ldots, \mathbf{u}_N^{t-1}\} \rightarrow$ **content-aware self-attention** (α) (Tu et al. 2016; See, P. J. Liu, and Christopher D. Manning 2017) $\rightarrow \mathbf{u}^{t-1}$

$$\boldsymbol{\alpha}^{t-1} = \text{softmax}\left(\mathbf{u}_\alpha \cdot \tanh\left(\mathbf{W}_\alpha \mathbf{U}^{t-1} + \mathbf{W}' \sum_{i=1}^{N} \mathbf{u}_i^t\right)\right) \tag{17}$$

Utterance encoder 3/3

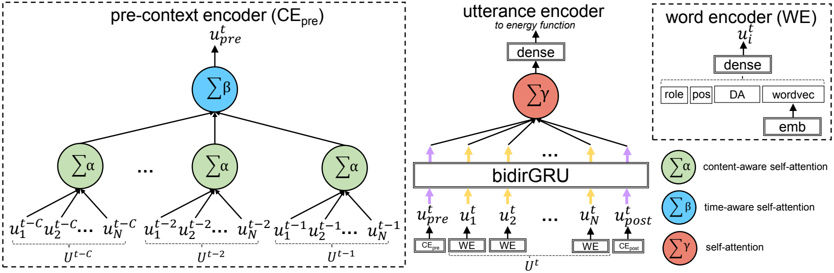


Figure: Our proposed utterance encoder G_W. Only the pre-context encoder is shown. C is the context size.

- **context encoder level 2**: $\{\mathbf{u}^{t-C}, \dots, \mathbf{u}^{t-1}\} \rightarrow$ **time-aware self-attention** (β) (Su, Yuan, and Y.-N. Chen 2018) $\rightarrow \mathbf{u}^t_{pre}$

$$\begin{aligned}\beta^{t-1} &= w_1 \beta^{\text{conv}^{t-1}} + w_2 \beta^{\text{lin}^{t-1}} + w_3 \beta^{\text{conc}^{t-1}} \\ &= \frac{w_1}{a(d^{t-1})^b} + w_2[ed^{t-1} + k]^+ + \frac{w_3}{1 + (\frac{d^{t-1}}{D_0})^l}\end{aligned} \tag{18}$$

where $[*]^+ = max(*, 0)$ (ReLU), d^{t-1} is the offset between the positions of $\mathbf{U}^{t-1}$ and $\mathbf{U}^t$, and the

Community detection

Fuzzy c-Means (FCM) algorithm (Bezdek, Ehrlich, and Full 1984) for overlapping communities.

- a probabilistic version of *k*-means, which returns a probability distribution over all communities for each utterance

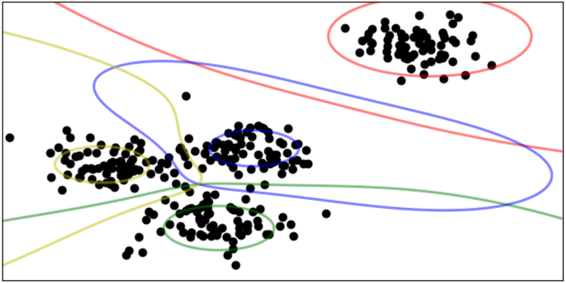


Figure: FCM example.

Experimental setup

Dataset: AMI meeting corpus

- participants play 4 roles of a design team to develop a TV remote control.
- 97, 20, and 20 meetings respectively for training, validation and testing.
- 2368 unique abstractive communities.

Baselines

- encoders: **LD** (J. Y. Lee and Dernoncourt 2016) and **HAN** (Z. Yang et al. 2016).
- systems: unsupervised (**tf-idf**, **w2v**, **LCseg** (Galley et al. 2003)), and supervised approaches similar to that of Murray, Carenini, and Ng 2012 (utterance graph + CONGA).

Ablations

- variants of our encoder: **CA-S**, **S-S**, **(0,0)**, and **(3,0)**.

Evaluation

- distance level: P, R, F1 at k (k=10/v)
- clustering level: Omega Index (Collins and Dent 1988) ($|Q|$=11/v)

Parameter tuning

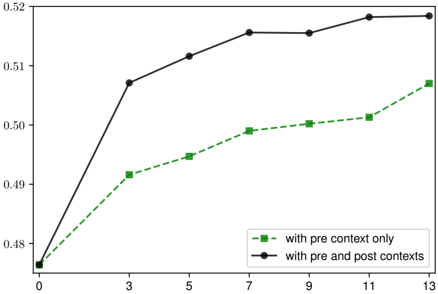


Figure: Impact of context size on the validation $P@k = v$, for our model trained within the triplet meta-architecture.

Results

			(pre, post)	P @$k = v$	P @$k = 10$	R @$k = 10$	F1 @$k = 10$	Omega index ×100 $\lvert Q \rvert = v$	Omega index ×100 $\lvert Q \rvert = 11$
Triplet	a1)	our model	(0, 0)	54.59	46.05	62.45	43.18	49.09	48.81
	a2)	our model	(3, 0)	55.17	46.17	62.80	43.25	49.78	49.70
	a3)	our model	(11, 11)	58.58	46.73	63.82	43.83	49.90	49.28
	b)	our model (CA-S)	(11, 11)	**59.52***	**46.98***	**64.01***	**44.06***	**50.11**	49.73
	c)	our model (S-S)	(11, 11)	58.96	46.81	63.65	43.87	49.59	**49.88**
	d)	LD	(3, 0)	52.04	44.82	60.41	41.82	48.70	48.14
	e)	HAN	(11, 11)	58.72	45.76	62.60	42.89	49.32	48.88
Siamese	f1)	our model	(0, 0)	53.01	45.10	60.97	42.12	50.56	49.65
	f2)	our model	(3, 0)	53.78	45.54	61.33	42.48	51.01	50.00
	f3)	our model	(11, 11)	56.64	**46.47**	**62.54**	**43.40**	**52.44***	**51.88***
	g)	our model (CA-S)	(11, 11)	56.46	46.08	61.92	43.02	51.60	50.98
	h)	our model (S-S)	(11, 11)	55.68	45.64	61.17	42.53	52.26	51.11
	i)	LD	(3, 0)	52.13	44.83	60.85	41.86	51.18	50.70
	j)	HAN	(11, 11)	**58.54**	45.72	61.55	42.74	50.51	49.82
Unsupervised	k1)	tf-idf	(0, 0)	29.28	26.67	34.69	24.19	13.12	13.66
	k2)	tf-idf	(3, 0)	34.77	30.27	40.83	27.79	10.22	10.17
	k3)	tf-idf	(11, 11)	**58.94**	43.94	61.36	41.45	38.09	39.47
	l1)	w2v	(0, 0)	29.02	27.46	37.39	25.11	13.89	13.50
	l2)	w2v	(3, 0)	34.11	29.92	39.55	27.32	10.61	10.77
	l3)	w2v	(11, 11)	58.30	**44.08**	**61.59**	**41.59**	37.75	38.28
	m)	LCSeg	-	-	-	-	-	**38.98**	**41.57**
Supervised	n1)	tf-idf	(0, 0)	-	-	-	-	25.04	25.14
	n2)	tf-idf	(3, 0)	-	-	-	-	27.33	26.95
	n3)	tf-idf	(11, 11)	-	-	-	-	**45.26**	**44.91**
	o1)	w2v	(0, 0)	-	-	-	-	25.32	25.25
	o2)	w2v	(3, 0)	-	-	-	-	29.14	29.02
	o3)	w2v	(11, 11)	-	-	-	-	43.31	43.08

Table: Results (averaged over 10 runs). *: best score per column. **Bold**: best score per section. -: does not apply

Attention visualization

-11	ID:	And	we'll	need	to	custom	desi	design	a	circuit	board	,				
-10	ID:	because	the	circuit	board	has	to	take	the	button	input	and	send	it	to	...
-9	ID:	But	once	we	come	up	with	a	design	we'll	send	it	to	the	circuit	...
-8	ID:	Um	,	standard	parts	include	the	buttons	and	the	wheels	,	um	the	iPod-style	...
-7	ID:	The	infrared	LED	is	actually	gonna	be	included	in	the	circuit	board	that	comes	...
-6	ID:	Um	,	we	need	a	radio	sender	and	receiver	,	those	are	standard	.	
-5	ID:	And	al	we	also	need	a	beeper	or	buzzer	or	other	sort	of	noise	...
-4	ID:	So	we	have	some	material	options	.								
-3	ID:	Um	,	we	can	use	rubber	,	plastic	,	wood	or	titanium	.		
-2	ID:	Um	,	I'd	recommend	against	titanium									
-1	ID:	because	it	can	only	be	used	in	the	flat	cases	and	it's	really	heavy	.
t	ID:	PRE	Um	,	and	the	rubber	case	requires	rubber	buttons	,	so	if	we	definitely
		want	plastic	buttons	,	we	shouldn't	have	a	rubber	case	.	POST			
+1	PM:	And	why	not	wood	?										
+2	ID:	And	,													
+3	ID:	hmm	?													
+4	PM:	And	why	not	wood	?										
+5	ID:	Uh	,	well	we	can	use	wood	.							
+6	ID:	I	don't	know	why	we'd	want	to	.							
+7	ID:	Um	and	also	we	should	note	that	if	we	want	an	iPod-style	wheel	button	...
+8	ID:	We	can't	use	the	minimal	chip	,	we	need	the	next	higher	grade	,	...
+9	ID:	I	don't	think	it's	much	more	expensive	,	but	it	is	more	expensive	.	
+10	ID:	So	that's	what	I've	got	on	design	.							
+11	PM:	'S	good	.												

Figure: Visualization of attention distributions around an utterance from the ES2011c meeting. Some utterances are truncated for readability

Conclusion

Contributions

- We formalized ACD, a crucial subtask for abstractive summarization of conversations. The AMI corpus preprocessed for this task and the code are publicly available. `https://bitbucket.org/guokan_shang/abscomm`
- We proposed an energy-based learning approach to this task, using siamese and triplet architectures to learn utterance embeddings for clustering.
- We introduced a novel utterance encoder featuring three types of self-attention mechanisms and taking contextual and temporal information into account.

Future work

1. evaluate our approach within the full abstractive summarization pipeline $a1 \rightarrow a2 \rightarrow b$
2. apply our contextual utterance encoder to other tasks, such as dialogue act classification.

Outline

Summary of contributions

- A fully unsupervised framework based on multi-sentence compression graphs and budgeted submodular maximization.
  - **Abstractive meeting summarization**
  - transcription $\xrightarrow{\text{generates}}$ summary
- A modified neural conditional random field layer that takes speaker-change into account.
  - **Dialogue act classification**
  - utterance $\xrightarrow{\text{assigns}}$ dialogue act label
- An energy-based learning approach, a general triplet sampling scheme, and a contextual utterance encoder featuring self-attention mechanisms.
  - **Abstractive community detection**
  - utterances $\xrightarrow{\text{groups}}$ abstractive communities

Future work

Deeper understanding of meetings

- discourse structure/graph → GNN
  - Thompson and Mann 1987; Nicholas Asher 1993; N. Asher et al. 2003
  - Zhou et al. 2018
- multi-modal information
  - M. Li et al. 2019

Deeper understanding of natural language

- pre-trained language model
  - Devlin et al. 2018

Thank you!

References I

- Viterbi, Andrew (1967). “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. In: **IEEE transactions on Information Theory** 13.2, pp. 260–269.
- Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson (1974). “A Simplest Systematics for the Organization of Turn-Taking for Conversation”. In: **Language** 50.4, pp. 696–735. ISSN: 00978507, 15350665. URL: http://www.jstor.org/stable/412243.
- Bezdek, James C., Robert Ehrlich, and William Full (1984). “FCM: The fuzzy c-means clustering algorithm”. In: **Computers & Geosciences** 10.2, pp. 191–203. ISSN: 0098-3004. DOI: 10.1016/0098-3004(84)90020-7. URL: http://www.sciencedirect.com/science/article/pii/0098300484900207.
- Thompson, Sandra A and William C Mann (1987). “Rhetorical structure theory: A framework for the analysis of texts”. In: **IPRA Papers in Pragmatics** 1.1, pp. 79–105.
- Collins, Linda M. and Clyde W. Dent (1988). “Omega: A General Formulation of the Rand Index of Cluster Recovery Suitable for Non-disjoint Solutions”. In: **Multivariate Behavioral Research** 23.2. PMID: 26764947, pp. 231–242. DOI: 10.1207/s15327906mbr2302_6. eprint: https://doi.org/10.1207/s15327906mbr2302_6. URL: https://doi.org/10.1207/s15327906mbr2302_6.
- Asher, Nicholas (1993). **Reference to abstract objects in English**.
- Bromley, Jane et al. (1994). “Signature Verification using a "Siamese" Time Delay Neural Network”. In: **Advances in Neural Information Processing Systems 6**. Ed. by J. D. Cowan, G. Tesauro, and J. Alspector. Morgan-Kaufmann, pp. 737–744. URL: http://papers.nips.cc/paper/769-signature-verification-using-a-siamese-time-delay-neural-network.pdf.
- Jurafsky, Dan, Liz Shriberg, and Debra Biasca (1997). “Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual”. In: **Institute of Cognitive Science Technical Report**. URL: https://web.stanford.edu/~jurafsky/ws97/manual.august1.html.
- Ries, Klaus (1999). “HMM and neural network based speech act detection”. In: **1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)**. Vol. 1. IEEE, pp. 497–500.
- Stolcke, Andreas et al. (2000). “Dialogue act modeling for automatic tagging and recognition of conversational speech”. In: **Computational Linguistics** 26.3, pp. 339–374. URL: https://www.aclweb.org/anthology/J00-3003.
- Huang, Xuedong et al. (2001). **Spoken Language Processing: A Guide to Theory, Algorithm, and System Development**. 1st. USA: Prentice Hall PTR. ISBN: 0130226165.
- Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: **Proceedings of the Eighteenth International Conference on Machine Learning**. ICML ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289. ISBN: 1558607781.
- Romano, Nicholas C and Jay F Nunamaker (2001). “Meeting analysis: Findings from research and practice”. In: **Proceedings of the 34th annual Hawaii international conference on system sciences**. IEEE, 13–pp.
- Asher, N. et al. (2003). **Logics of Conversation**. Studies in Natural Language Processing. Cambridge University Press. ISBN: 9780521650588. URL: https://books.google.fr/books?id=VD-8yisFhBwC.

References II

- Galley, Michel et al. (July 2003). "Discourse Segmentation of Multi-Party Conversation". In: **Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics**. Sapporo, Japan: ACL, pp. 562–569. DOI: 10.3115/1075096.1075167. URL: https://www.aclweb.org/anthology/P03-1071.
- Janin, Adam et al. (2003). "The ICSI meeting corpus". In: **Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on**. Vol. 1. IEEE, pp. I–I.
- Grau, Sergio et al. (2004). "Dialogue act classification using a Bayesian approach". In: **9th Conference Speech and Computer**.
- Lin, Chin-Yew (2004). "Rouge: A package for automatic evaluation of summaries". In: **Text summarization branches out: Proceedings of the ACL-04 workshop**. Vol. 8. Barcelona, Spain.
- Mihalcea, Rada and Paul Tarau (2004). "Textrank: Bringing order into text". In: **Proceedings of the 2004 conference on empirical methods in natural language processing**.
- Ang, Jeremy, Yang Liu, and Elizabeth Shriberg (2005). "Automatic dialog act segmentation and classification in multiparty meetings". In: **Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.** Vol. 1. IEEE, pp. I–1061.
- Chopra, Sumit, Raia Hadsell, and Yann LeCun (2005). "Learning a similarity metric discriminatively, with application to face verification". In: **Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on**. Vol. 1. IEEE, pp. 539–546.
- LeCun, Yann and Fu Jie Huang (2005). "Loss Functions for Discriminative Training of Energy-Based Models". In: **Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005, Bridgetown, Barbados, January 6-8, 2005**. Ed. by Robert G. Cowell and Zoubin Ghahramani. Society for Artificial Intelligence and Statistics. URL: http://www.gatsby.ucl.ac.uk/aistats/fullpapers/207.pdf.
- McCowan, Iain et al. (2005). "The AMI meeting corpus". In: **Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research**. Vol. 88, p. 100.
- Venkataraman, Anand et al. (2005). "Does active learning help automatic dialog act tagging in meeting data?" In: **Ninth European Conference on Speech Communication and Technology**.
- Lecun, Yann et al. (2006). "A tutorial on energy-based learning". English (US). In: **Predicting structured data**. MIT Press. URL: http://yann.lecun.com/exdb/publis/orig/lecun-06.pdf.
- Liu, Yang (2006). "Using SVM and error-correcting codes for multiclass dialog act classification in meeting corpus". In: **Ninth International Conference on Spoken Language Processing**.
- Surendran, Dinoj and Gina-Anne Levow (2006). "Dialog act tagging with support vector machines and hidden Markov models". In: **Ninth International Conference on Spoken Language Processing**.
- Lendvai, Piroska and Jeroen Geertzen (2007). "Token-based chunking of turn-internal dialogue act sequences". In: **Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue**, pp. 174–181.
- Manning, Christopher D, Hinrich Schütze, and Prabhakar Raghavan (2008). **Introduction to information retrieval**. Cambridge university press.

References III

- Riedhammer, Korbinian et al. (2008). "Packing the meeting summarization knapsack". In: **Ninth Annual Conference of the International Speech Communication Association**.
- Garg, Nikhil et al. (2009). "Clusterrank: a graph based method for meeting summarization". In: **Tenth Annual Conference of the International Speech Communication Association**.
- Petukhova, Volha and Harry Bunt (2009). "Who's next? Speaker-selection mechanisms in multiparty dialogue". In: **Workshop on the Semantics and Pragmatics of Dialogue**.
- Zimmermann, Matthias (2009). "Joint segmentation and classification of dialog acts using conditional random fields". In: **Tenth Annual Conference of the International Speech Communication Association**.
- Filippova, Katja (2010). "Multi-sentence compression: Finding shortest paths in word graphs". In: **Proceedings of the 23rd International Conference on Computational Linguistics**. Association for Computational Linguistics, pp. 322–330.
- Kim, Su Nam, Lawrence Cavedon, and Timothy Baldwin (Oct. 2010). "Classifying Dialogue Acts in One-on-One Live Chats". In: **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**. Cambridge, MA: Association for Computational Linguistics, pp. 862–871. URL: https://www.aclweb.org/anthology/D10-1084.
- Lin, Hui and Jeff Bilmes (2010). "Multi-document summarization via budgeted maximization of submodular functions". In: **Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics**. Association for Computational Linguistics, pp. 912–920.
- Murray, Gabriel, Giuseppe Carenini, and Raymond Ng (2010). "Generating and validating abstracts of meeting conversations: a user study". In: **Proceedings of the 6th International Natural Language Generation Conference**. Association for Computational Linguistics, pp. 105–113.
- Carenini, Giuseppe, Gabriel Murray, and Raymond Ng (2011). "Methods for mining and summarizing text conversations". In: **Synthesis Lectures on Data Management** 3.3, pp. 1–130.
- Tur, Gokhan and Renato De Mori (2011). **Spoken language understanding: Systems for extracting semantic information from speech**. John Wiley & Sons.
- Lin, Hui (2012). **Submodularity in natural language processing: algorithms and applications**. University of Washington.
- Murray, Gabriel, Giuseppe Carenini, and Raymond Ng (2012). "Using the omega index for evaluating abstractive community detection". In: **Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization**. Association for Computational Linguistics, pp. 10–18.
- Boudin, Florian and Emmanuel Morin (2013). "Keyphrase extraction for n-best reranking in multi-sentence compression". In: **North American Chapter of the Association for Computational Linguistics (NAACL)**.
- Kalchbrenner, Nal and Phil Blunsom (Aug. 2013). "Recurrent Convolutional Neural Networks for Discourse Compositionality". In: **Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality**. Sofia, Bulgaria: Association for Computational Linguistics, pp. 119–126. URL: https://www.aclweb.org/anthology/W13-3214.
- Mehdad, Yashar et al. (2013). "Abstractive Meeting Summarization with Entailment and Fusion.". In: **ENLG**, pp. 136–146.

References IV

- Mikolov, Tomas, Kai Chen, et al. (2013). "Efficient estimation of word representations in vector space". In: **arXiv preprint arXiv:1301.3781**.
- Mikolov, Tomas, Ilya Sutskever, et al. (2013). "Distributed representations of words and phrases and their compositionality". In: **Advances in neural information processing systems**, pp. 3111–3119.
- Rousseau, François and Michalis Vazirgiannis (2013). "Graph-of-word and TW-IDF: New Approach to Ad Hoc IR". In: **Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management**. CIKM '13. San Francisco, California, USA: ACM, pp. 59–68. ISBN: 978-1-4503-2263-8. DOI: 10.1145/2505515.2505671. URL: http://doi.acm.org/10.1145/2505515.2505671.
- Tavafi, Maryam et al. (Aug. 2013). "Dialogue Act Recognition in Synchronous and Asynchronous Conversations". In: **Proceedings of the SIGDIAL 2013 Conference**. Metz, France: Association for Computational Linguistics, pp. 117–121. URL: https://www.aclweb.org/anthology/W13-4017.
- Krause, Andreas and Daniel Golovin (2014). **Submodular function maximization.**
- Oya, Tatsuro et al. (2014). "A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships". In: **Proceedings of the 8th International Natural Language Generation Conference (INLG)**. Philadelphia, Pennsylvania, U.S.A.: ACL, pp. 45–53. DOI: 10.3115/v1/W14-4407. URL: http://aclweb.org/anthology/W14-4407.
- Wang, Jiang et al. (2014). "Learning Fine-Grained Image Similarity with Deep Ranking". In: **Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition**. CVPR '14. Washington, DC, USA: IEEE Computer Society, pp. 1386–1393. ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.180. URL: https://doi.org/10.1109/CVPR.2014.180.
- Wang, Rui, Wei Liu, and Chris McDonald (2014). "Corpus-independent generic keyphrase extraction using word embedding vectors". In: **Software Engineering Research Conference**. Vol. 39.
- Banerjee, Siddhartha, Prasenjit Mitra, and Kazunari Sugiyama (2015). "Generating Abstractive Summaries from Meeting Transcripts". In: **Proceedings of the 2015 ACM Symposium on Document Engineering**. DocEng '15. Lausanne, Switzerland, pp. 51–60. ISBN: 978-1-4503-3307-8. DOI: 10.1145/2682571.2797061. URL: http://doi.acm.org/10.1145/2682571.2797061.
- Hoffer, Elad and Nir Ailon (2015). "Deep metric learning using Triplet network". In: **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings**. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1412.6622.
- Rousseau, François and Michalis Vazirgiannis (2015). "Main core retention on graph-of-words for single-document keyword extraction". In: **European Conference on Information Retrieval**. Springer, pp. 382–393.
- Schroff, F., D. Kalenichenko, and J. Philbin (June 2015). "FaceNet: A unified embedding for face recognition and clustering". In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Vol. 00, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682. URL: doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298682.

References V

- Khanpour, Hamed, Nishitha Guntakandla, and Rodney Nielsen (Dec. 2016). "Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network". In: **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 2012–2021. URL: https://www.aclweb.org/anthology/C16-1189.
- Lee, Ji Young and Franck Dernoncourt (2016). "Sequential short-text classification with recurrent and convolutional neural networks". In: **arXiv preprint arXiv:1603.03827**.
- Li, Wei and Yunfang Wu (Dec. 2016). "Multi-level Gated Recurrent Neural Network for dialog act classification". In: **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 1970–1979. URL: https://www.aclweb.org/anthology/C16-1185.
- Mueller, Jonas and Aditya Thyagarajan (2016). "Siamese Recurrent Architectures for Learning Sentence Similarity". In: **Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.** Ed. by Dale Schuurmans and Michael P. Wellman. AAAI Press, pp. 2786–2792. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12195.
- Shen, Sheng-syun and Hung-yi Lee (2016). "Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection". In: **arXiv preprint arXiv:1604.00077**.
- Tixier, Antoine, Fragkiskos Malliaros, and Michalis Vazirgiannis (2016). "A graph degeneracy-based approach to keyword extraction". In: **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1860–1870.
- Tu, Zhaopeng et al. (2016). "Modeling coverage for neural machine translation". In: **arXiv preprint arXiv:1601.04811**.
- Yang, Zichao et al. (2016). "Hierarchical Attention Networks for Document Classification". In: **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. San Diego, California: ACL, pp. 1480–1489. DOI: 10.18653/v1/N16-1174. URL: http://aclweb.org/anthology/N16-1174.
- Lin, Zhouhan et al. (2017). "A Structured Self-Attentive Sentence Embedding". In: **5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings**. URL: https://openreview.net/forum?id=BJC%5C_jUqxe.
- Liu, Yang et al. (Sept. 2017). "Using Context Information for Dialog Act Classification in DNN Framework". In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2170–2178. DOI: 10.18653/v1/D17-1231. URL: https://www.aclweb.org/anthology/D17-1231.
- Meladianos, Polykarpos et al. (2017). "Real-Time Keyword Extraction from Conversations". In: **EACL 2017**, p. 462.
- Ortega, Daniel and Ngoc Thang Vu (Aug. 2017). "Neural-based Context Representation Learning for Dialog Act Classification". In: **Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue**. Saarbrücken, Germany: Association for Computational Linguistics, pp. 247–252. DOI: 10.18653/v1/W17-5530. URL: https://www.aclweb.org/anthology/W17-5530.

References VI

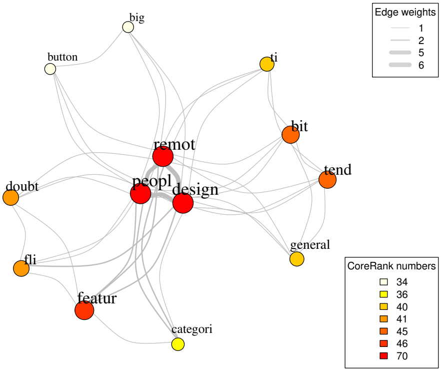
- Reimers, Nils and Iryna Gurevych (2017). "Optimal hyperparameters for deep lstm-networks for sequence labeling tasks". In: **arXiv preprint arXiv:1707.06799**.
- See, Abigail, Peter J. Liu, and Christopher D. Manning (2017). "Get To The Point: Summarization with Pointer-Generator Networks". In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Vancouver, Canada: ACL, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: http://aclweb.org/anthology/P17-1099.
- Singla, Karan et al. (2017). "Automatic Community Creation for Abstractive Spoken Conversations Summarization". In: **Proceedings of the Workshop on New Frontiers in Summarization**. Copenhagen, Denmark: ACL, pp. 43–47. DOI: 10.18653/v1/W17-4506. URL: http://aclweb.org/anthology/W17-4506.
- Tixier, Antoine, Polykarpos Meladianos, and Michalis Vazirgiannis (2017). "Combining Graph Degeneracy and Submodularity for Unsupervised Extractive Summarization". In: **Proceedings of the Workshop on New Frontiers in Summarization**, pp. 48–58.
- Tran, Quan Hung, Ingrid Zukerman, and Gholamreza Haffari (Apr. 2017). "A Hierarchical Neural Model for Learning Sequences of Dialogue Acts". In: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**. Valencia, Spain: Association for Computational Linguistics, pp. 428–437. URL: https://www.aclweb.org/anthology/E17-1041.
- Vaswani, Ashish et al. (2017). "Attention is All you Need". In: **Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA**. Ed. by Isabelle Guyon et al., pp. 6000–6010. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.
- Bothe, Chandrakant et al. (May 2018). "A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks". In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://www.aclweb.org/anthology/L18-1307.
- Chen, Zheqian et al. (2018). "Dialogue Act Recognition via CRF-Attentive Structured Network". In: **The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval**. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, pp. 225–234. ISBN: 9781450356572. DOI: 10.1145/3209978.3209997. URL: https://doi.org/10.1145/3209978.3209997.
- Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: **arXiv preprint arXiv:1810.04805**.
- Kumar, Harshit et al. (2018). "Dialogue Act Sequence Labeling Using Hierarchical Encoder With CRF". In: **AAAI Conference on Artificial Intelligence**. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16706.
- **Shang**, **Guokan**, Wensi Ding, et al. (July 2018). "Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization". In: **ACL 2018 - Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, pp. 664–674. URL: https://www.aclweb.org/anthology/P18-1062.

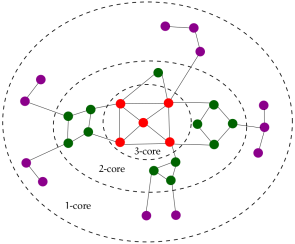
References VII

- Su, Shang-Yu, Pei-Chieh Yuan, and Yun-Nung Chen (June 2018). "How Time Matters: Learning Time-Decay Attention for Contextual Spoken Language Understanding in Dialogues". In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana: ACL, pp. 2133–2142. DOI: 10.18653/v1/N18-1194. URL: https://www.aclweb.org/anthology/N18-1194.
- Yang, Jie, Shuailong Liang, and Yue Zhang (2018). "Design challenges and misconceptions in neural sequence labeling". In: **arXiv preprint arXiv:1806.04470**.
- Zhou, Jie et al. (2018). "Graph neural networks: A review of methods and applications". In: **arXiv preprint arXiv:1812.08434**.
- Cui, Leyang and Yue Zhang (2019). "Hierarchically-Refined Label Attention Network for Sequence Labeling". In: **arXiv preprint arXiv:1908.08676**.
- Li, Manling et al. (July 2019). "Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization". In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, pp. 2190–2196. DOI: 10.18653/v1/P19-1210. URL: https://www.aclweb.org/anthology/P19-1210.
- Li, Ruizhe et al. (Nov. 2019). "A Dual-Attention Hierarchical Recurrent Neural Network for Dialogue Act Classification". In: **Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)**. Hong Kong, China: Association for Computational Linguistics, pp. 383–392. DOI: 10.18653/v1/K19-1036. URL: https://www.aclweb.org/anthology/K19-1036.
- Lorré, Jean-Pierre et al. (2019). "LinTO: Assistant vocal open-source respectueux des données personnelles pour les réunions d'entreprise". In: **APIA**, p. 63.
- Lu, Yang (2019). "Artificial intelligence: a survey on evolution, models, applications and future trends". In: **Journal of Management Analytics** 6.1, pp. 1–29.
- Raheja, Vipul and Joel Tetreault (June 2019). "Dialogue Act Classification with Context-Aware Self-Attention". In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3727–3733. DOI: 10.18653/v1/N19-1373. URL: https://www.aclweb.org/anthology/N19-1373.
- **Shang**, **Guokan**, Antoine Tixier, et al. (Dec. 2020a). "Energy-based Self-attentive Learning of Abstractive Communities for Spoken Language Understanding". In: **AACL-IJCNLP 2020 - Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing**. Suzhou, China: Association for Computational Linguistics, pp. 313–327. URL: https://www.aclweb.org/anthology/2020.aacl-main.34.
- — (Dec. 2020b). "Speaker-change Aware CRF for Dialogue Act Classification". In: **COLING 2020 - Proceedings of the 28th International Conference on Computational Linguistics**. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 450–464. URL: https://www.aclweb.org/anthology/2020.coling-main.40.

Keyword Extraction with Graph-of-words and CoreRank

$$TW\text{-}IDF(t, d, D) = TW(t, d) \times IDF(t, D) \quad (19)$$





A k-core of G is a maximal subgraph of G in which every vertex v has at least weighted degree k.
(François Rousseau and Vazirgiannis 2015; Tixier, Malliaros, and Vazirgiannis 2016; Meladianos et al. 2017)

Submodularity

- **Submodularity** (Krause and Golovin 2014):
  A set function $F : 2^V \rightarrow \mathcal{R}$ where $V = \{v_1, ..., v_n\}$ is said to be *submodular* if it satisfies the property of *diminishing returns*:

$$\forall A \subseteq B \subseteq V \backslash v,$$
$$F(A \cup v) - F(A) \geq F(B \cup v) - F(B)$$

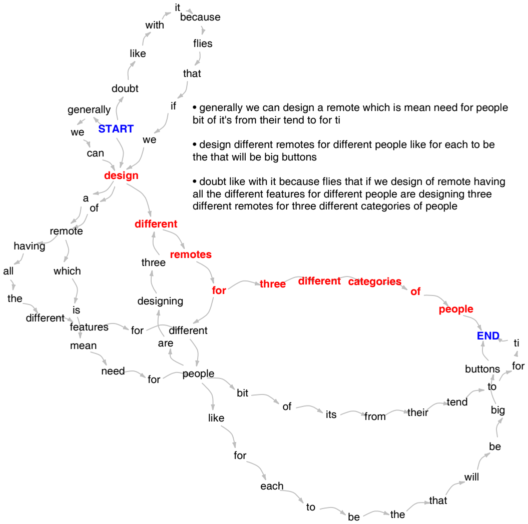
the gain of adding a new sentence to a given summary should be greater than the gain of adding the same sentence to a larger summary containing the smaller one

the set function $F(\cdot)$ is *monotone non-decreasing*:

$$\forall A \subseteq B, F(A) \leq F(B)$$

the quality of a summary can only increase or stay the same as it grows in size, i.e., as we add sentences to it

Example



- generally we can design a remote which is mean need for people bit of it's from their tend to for ti
- design different remotes for different people like for each to be the that will be big buttons
- doubt like with it because flies that if we design of remote having all the different features for different people are designing three different remotes for three different categories of people

http://datascience.open-paas.org/abs_summ_app

Confusion matrices for the 10 best predicted labels

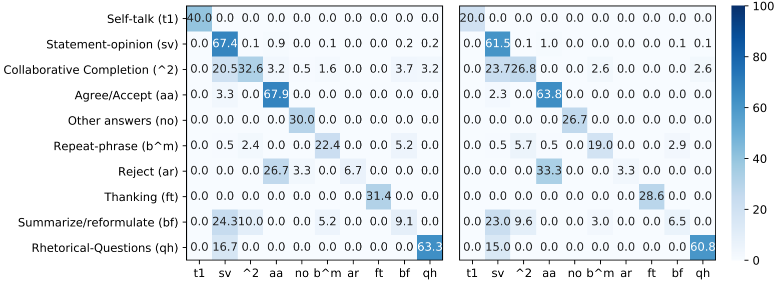


Figure: Normalized confusion matrices, averaged over 10 runs, for the 10 DA labels **best** predicted by our model (20.2% of all annotations). Left: our model, right: base model.

	Ours	Vanilla	Diff.
10 best DAs	37.08	31.70	+ 5.38
10 worst DAs	59.67	64.54	- 4.87

Table: accuracy (%) of our model vs. base model on the 10 DAs best and worst predicted by our model (resp., **20**% and **40**% of all annotations).

Confusion matrices for the 10 worst predicted labels

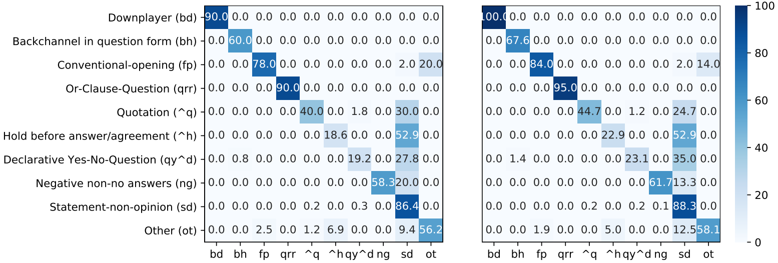


Figure: Normalized confusion matrices, averaged over 10 runs, for the 10 DA labels **worst** predicted by our model (20.2% of all annotations). Left: our model, right: base model.

1. “A: Hi, Wanet. (fp)”
2. “A: How are you? (fp)”
3. “B: I’m doing fine. (fp)”

Energy-Based Modeling (EBM)

EBM is a unified framework that can be applied to many machine learning problems (LeCun and F. J. Huang 2005; Lecun et al. 2006).

- An energy function $E_W(X, Y)$ parameterized by W assigns a scalar called *energy* to each pair of random variables (X, Y).
- Training consists in finding the parameters W^* of the energy function E_W that, for all (X^i, Y^i) in the training set $\mathcal{S}$ of size P, assign low energy to compatible (correct) combinations and high energy to all other incompatible (incorrect) ones.
- This is done by minimizing a *loss functional* $\mathcal{L}$:

$$W^* = \arg\min_{W \in \mathcal{W}} \mathcal{L}(E_W(X, Y), \mathcal{S}) \qquad (20)$$

- For a given X, prediction consists in finding the value of Y that minimizes the energy.

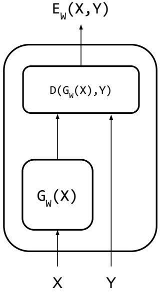


Figure: EBMs for regression. G_W: regressor model, D: dissimilarity measure.

$$\mathcal{L} = \frac{1}{P}\sum_{i=1}^{P} E_W(X^i, Y^i) \qquad (21)$$

$$= \frac{1}{P}\sum_{i=1}^{P} \|G_W(X^i) - Y^i\|^2 \qquad (22)$$

Siamese & triplet architectures

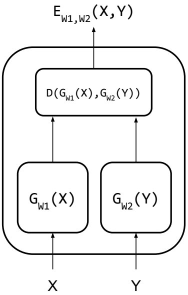


Figure: Siamese architecture, when $G_{W_1} = G_{W_2}$ and $W_1 = W_2$.

siamese (Bromley et al. 1994; Chopra, Hadsell, and LeCun 2005)

- (X^i, Y^i) is a positive pair, i.e., the label $C^i = 0$, when X^i and Y^i are two utterances from the same community, otherwise (X^i, Y^i) is a negative pair.
- objective: minimize the output energies (or distances) associated with positive pairs, and maximize those associated with negative pairs.
- loss (Mueller and Thyagarajan 2016):

$$E_W(X, Y) = 1 - \exp(-\|G_W(X) - G_W(Y)\|_1) \quad (23)$$

$$\mathcal{L} = \frac{1}{P}\sum_{i=1}^{P} \|E_W(X^i, Y^i) - C^i\|^2 \quad (24)$$

Siamese & triplet architectures

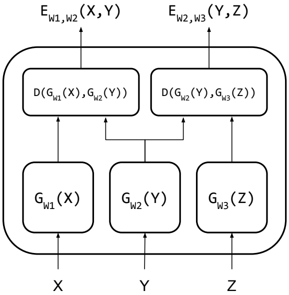


Figure: Triplet architecture, when $G_{W_1} = G_{W_2} = G_{W_3}$ and $W_1 = W_2 = W_3$.

triplet (Schroff, Kalenichenko, and Philbin 2015; Hoffer and Ailon 2015; J. Wang et al. 2014)

- a direct extension of the siamese architecture
- (X, Y, Z) referred to as the *positive*, *anchor*, and *negative* objects, where X and Y are from the same community and Z from another.
- objective: jointly minimize the positive-anchor energy $E_W(X^i, Y^i)$ and maximize the anchor-negative energy $E_W(Y^i, Z^i)$.
- *softmax triplet loss* (Hoffer and Ailon 2015):

$$\mathcal{L} = \frac{1}{2P} \sum_{i=1}^{P} \left(\|ne^+ - 0\|^2 + \|ne^- - 1\|^2 \right) \quad (25)$$

$$ne^+ = \frac{e^{E_W(X^i,Y^i)}}{e^{E_W(X^i,Y^i)} + e^{E_W(Y^i,Z^i)}} \quad (26)$$

$$ne^- = \frac{e^{E_W(Y^i,Z^i)}}{e^{E_W(X^i,Y^i)} + e^{E_W(Y^i,Z^i)}} \quad (27)$$

$$E_W(X^i, Y^i) = \|G_W(X^i) - G_W(Y^i)\|_2 \quad (28)$$

Triplet sampling scheme

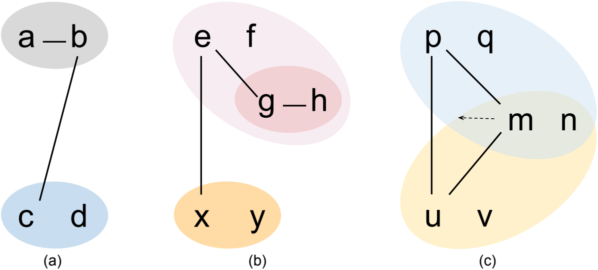


Figure: (a) communities $\{a, b\}$ and $\{c, d\}$ are disjoint (b) community $\{g, h\}$ is nested in community $\{e, f, g, h\}$ (c) communities $\{p, q, m, n\}$ and $\{m, n, u, v\}$ overlap upon $\{m, n\}$.

Triplet sampling scheme

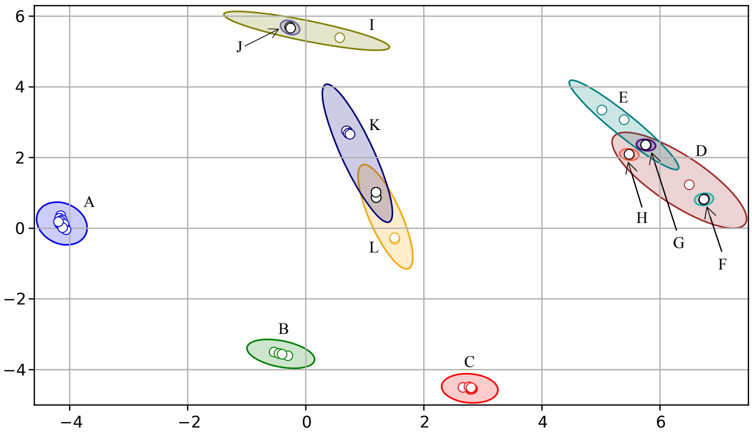


Figure: All 48 utterances of 12 abstractive communities from the meeting IS1001c projected into 2-dimensional PCA of learned 32-dimensional embedding space. Trained on 23612 triplets for 5 epochs. Converged $P@k = v$ is equal to 96.33%.

Attention visualization

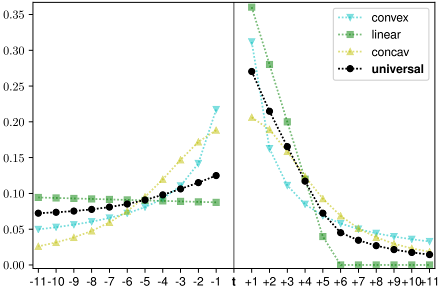


Figure: Normalized time-aware self-attention weights for pre and post-contexts, averaged over 10 runs.