Energy-based Self-attentive Learning of Abstractive Communities for Spoken Language Understanding

AACL-IJCNLP 2020

Guokan Shang[1,2], Antoine J.-P. Tixier[1], Michalis Vazirgiannis[1,3], Jean-Pierre Lorré[2]

[1]École Polytechnique, [2]LINAGORA, [3]AUEB

December 2020

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

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.

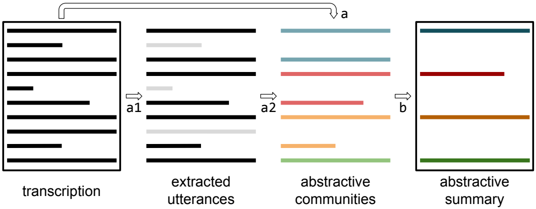

Figure: Abstractive summarization of conversations.

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

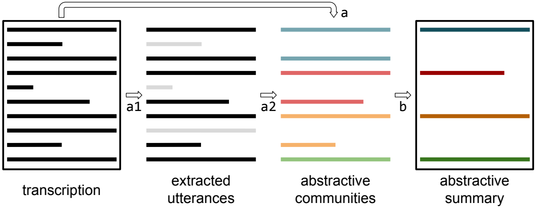

Figure: Abstractive summarization of conversations.

$\Rightarrow$ 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

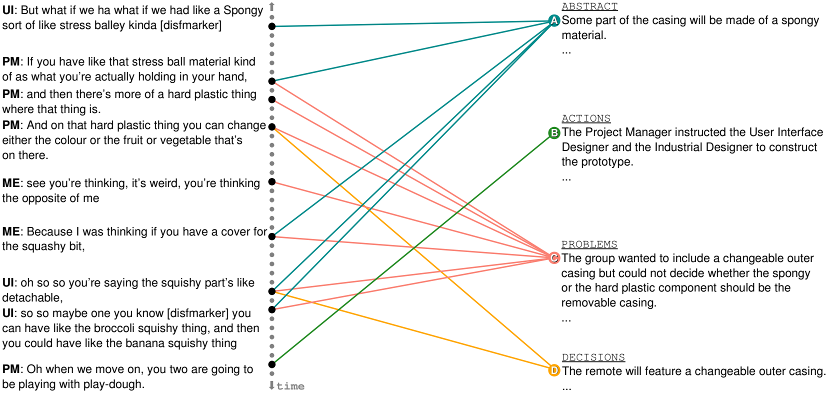

Figure: Example of ground truth human annotations from the ES2011c AMI meeting. Successive grey nodes on the left denote utterances in the transcription. Black nodes correspond to the utterances judged important. Sentences (e.g., A, B, C, D) from the abstractive summary are shown on the right. All utterances linked to the same abstractive sentence form one community.
$\Rightarrow$ 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 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

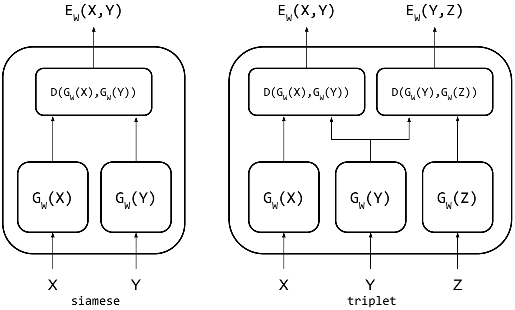

Figure: Siamese & triplet architectures

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; 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

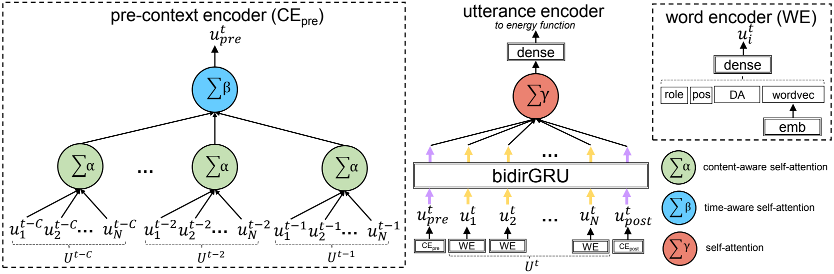

Figure: Our proposed utterance encoder G_W. Only the pre-context encoder is shown. C is the context size.

- **word encoder**: textual features (word embedding) + discourse features (role, position, dialogue act) $\rightarrow$ dense layer $\rightarrow$ $\mathbf{u}_i^t$
- **utterance encoder**: $\{\mathbf{u}_{\text{pre}}^t, \mathbf{u}_1^t, \ldots, \mathbf{u}_N^t, \mathbf{u}_{\text{post}}^t\}$ $\rightarrow$ BiGRU $\rightarrow$ **self-attention** (γ) (Vaswani et al. 2017; Lin et al. 2017) $\rightarrow$ dense layer $\rightarrow$ $\mathbf{u}^t$

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

Utterance encoder 2/3

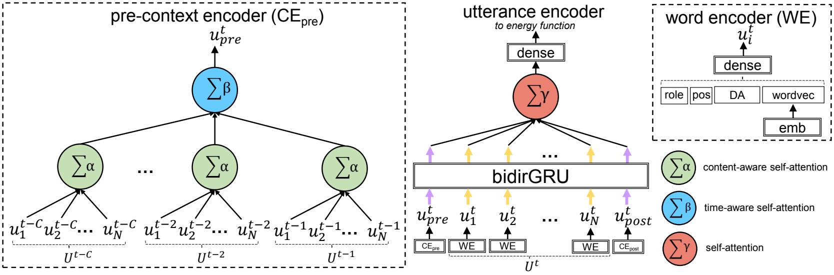

Figure: Our proposed utterance encoder G_W. Only the pre-context encoder is shown. C is the context size.

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

$$\boldsymbol{\alpha}^{t-1} = \mathrm{softmax}\left(\mathbf{u}_\alpha \cdot \tanh\left(\mathbf{W}_\alpha \mathbf{U}^{t-1} + \mathbf{W}' \sum_{i=1}^{N} \mathbf{u}_i^t\right)\right) \quad (2)$$

Utterance encoder 3/3

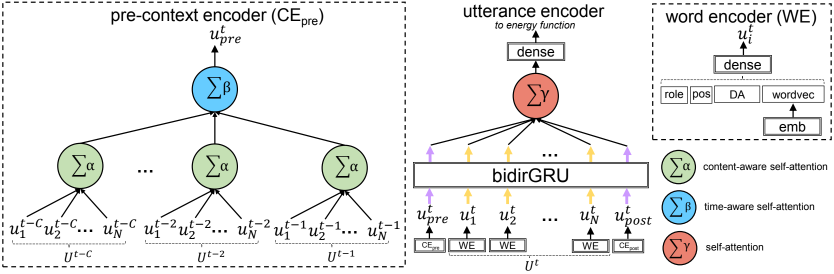

Figure: Our proposed utterance encoder G_W. Only the pre-context encoder is shown. C is the context size.

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

$$\begin{aligned} \beta^{t-1} &= w_1 \beta^{\text{conv}^{t-1}} + w_2 \beta^{\text{lin}^{t-1}} + w_3 \beta^{\text{conc}^{t-1}} \\ &= \frac{w_1}{a(d^{t-1})^b} + w_2[ed^{t-1} + k]^+ + \frac{w_3}{1 + (\frac{d^{t-1}}{D_0})^l} \end{aligned} \quad (3)$$

where $[*]^+ = max(*, 0)$ (ReLU), d^{t-1} is the offset between the positions of $\mathbf{U}^{t-1}$ and $\mathbf{U}^t$, and the w_i's, a, b, e, k, D_0, and l are scalar parameters learned during training.

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

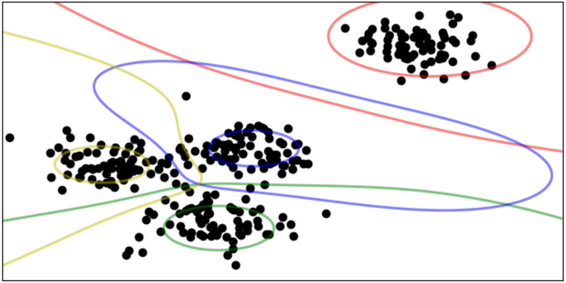

Figure: FCM example.

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** (Lee and Dernoncourt 2016) and **HAN** (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

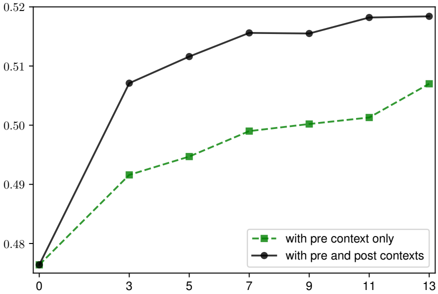

Figure: Impact of context size on the validation $P@k = v$, for our model trained within the triplet meta-architecture.

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 as the method does not produce utterance embeddings.

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	,	the	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

- 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

Future research should be devoted to 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.

Acknowledgments

This work was supported by the LinTO meeting assistant project (Lorré et al. 2019). `https://linto.ai`

References I

- 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.
- 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.
- 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.
- 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.
- Chopra, S., R. Hadsell, and Y. LeCun (June 2005). "Learning a similarity metric discriminatively, with application to face verification". In: **2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)**. Vol. 1, 539–546 vol. 1. DOI: 10.1109/CVPR.2005.202.
- 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. URL: http://www.cs.ru.nl/~kraaijw/pubs/Biblio/papers/mccowan-ami-mb2005.pdf.
- Murray, Gabriel, Giuseppe Carenini, and Raymond Ng (June 2012). "Using the Omega Index for Evaluating Abstractive Community Detection". In: **Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization**. Montréal, Canada: ACL, pp. 10–18. URL: https://www.aclweb.org/anthology/W12-2602.
- Mehdad, Yashar et al. (2013). "Abstractive Meeting Summarization with Entailment and Fusion". In: **Proceedings of the 14th European Workshop on Natural Language Generation**. Sofia, Bulgaria: ACL, pp. 136–146. URL: http://aclweb.org/anthology/W13-2117.
- 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.
- 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.

References II

- 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.
- Lee, Ji Young and Franck Dernoncourt (2016). "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks". 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. 515–520. DOI: 10.18653/v1/N16-1062. URL: http://aclweb.org/anthology/N16-1062.
- 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.
- 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.
- 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.
- Shang, Guokan et al. (2018). "Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization". In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: ACL, pp. 664–674. URL: http://aclweb.org/anthology/P18-1062.
- 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.
- Lorré, Jean-Pierre et al. (2019). "LinTO: Assistant vocal open-source respectueux des données personnelles pour les réunions d'entreprise". In: **APIA**, p. 63.