Methods to represent natural language with application to conversational AI.

Pierre Colombo (IBM GBS & LTCI, Télécom Paris)

Jury	Dr. Alexandre Allauzen	(ESPCI)
	Dr. Ludovic Denoyer	(Facebook Research)
Supervisors	Dr. Chloé Clavel	LTCI, Telecom Paris
	Dr. Giovanna Varni	LTCI, Telecom Paris
	Emmanuel Vignon	IBM GBS

Today's Agenda

Conversational AI

FIGURE – Source : www.activechat.ai

FIGURE –
Source :https ://www.paymentsjournal.com/

Goal of Conversational AI [7, 15]

The long term goal of conversational AI is to build a superhuman computer that will mimic the property of human conversation and will be preferred to an ordinary human.

Specifics of Spoken Language

Caller	Utterance
A	um, did you do through a public school system or private ?
B	Yeah,
B	well, I went through private an until ninth grade.
A	Uh-huh,
A	did you notice a big difference ?
B	Oh, yeah,
B	a big difference.
A	Like in what sense ?

TABLE – Example of dialog taken from the Switchboard Dialog Act Corpus.

Specifics of Spoken Language

Caller	Utterance
A	um, did you do through a public school system or private ?
B	Yeah,
B	well, I went through private an until ninth grade.
A	Uh-huh,
A	did you notice a big difference ?
B	Oh, yeah,
B	a big difference.
A	Like in what sense ?

TABLE – Example of dialog taken from the Switchboard Dialog Act Corpus.

Specifics of spoken dialog [15]

- *Disfluencies* (e.g "Uh-huh") [32]
- *Segmentation Issues* [1, 39]
- Lexical diversity, and grammatical complexity and accuracy. [6, 28]

Desirable qualities of conversational agents.[29]

- *Continual Learning :* continual learning enable the agent to adapt its behavior to new situations.
- *Engaging Content :* the agent needs to be able to carry and interesting and engaging conversation
- *Well behaved :* the agent can not generate offensive and toxic content

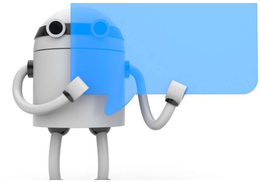


FIGURE – Source : https ://www.ladn.eu/

Aim of this thesis

We aim at textual learning representations useful for the conversational agent. Two types of problems arise :

- Understand the user's input (NLU)
- Generate a response (NLG)

Aim of this thesis

We aim at textual learning representations useful for the conversational agent. Two types of problems arise :

- Understand the user's input (NLU)
- Generate a response (NLG)

RQs :

- RQ1 : *How to best represent inputs that contains text for NLU ? Can we build representations that take into account both the structural properties of the input and the target task i.e extracting the intends and associated information*.

Aim of this thesis

We aim at textual learning representations useful for the conversational agent. Two types of problems arise :

- Understand the user's input (NLU)
- Generate a response (NLG)

RQs :

- RQ1 : *How to best represent inputs that contains text for NLU ? Can we build representations that take into account both the structural properties of the input and the target task i.e extracting the intends and associated information*.

- RQ2 : *How to best represent inputs that contains text for NLG ? Can we build representations that exhibit desirable topological properties (e.g invariance, disentanglement) for a specific sequence generation task ?*

RQ1 : Research Questions

RQ1 : *How to best represent inputs that contains text for NLU ?*

Can we build representations that take into account both the structural properties of the input and the target task i.e extracting the intends and associated information.

Three different scenarii :

- *The inputs are transcripts of spontaneous speech.* [a]
- *The inputs are interactions.* [b]
- *The input is multi-modal and comes from spontaneous speech.* [c]

---

a. Work based on Dinkar(*) and Colombo(*) et al. EMNLP 2020

b. Work based on Colombo(*) and Chapuis(*) et al. AAAI 2020 and Chapuis(*) and Colombo(*) et al. Findings of EMNLP 2020

c. Work based on Garcia(*) and Colombo(*) et al. EMNLP 2019

RQ2 : Research Questions

RQ2 : *How to best represent inputs that contains text for NLG ?*

Can we build representations that exhibit desirable topological properties (e.g invariance, disentanglement) for a specific sequence generation task ?.

Two different tasks :

- *Sentence Generation with constant polarity.* [a]
- *Style Transfer.* [b]

---

a. Work based on Jalazai(*) and Colombo(*) et al. NeurIPS2020
b. Work based on Colombo, Piantanida, Clavel. (Preprint)

Today's Agenda

1. Context
   - Conversational AI
   - Research Questions (RQs)
2. RQ1 : *How to best represent inputs that contains text for NLU ?*
3. RQ2 : *How to best represent inputs that contains text for NLG ?*
4. Conclusions and Planning
5. References

Sequence labelling for conversational AI.

RQ1 : *How to best represent inputs that contains text for NLU ?*

Can we build representations that take into account both the structural properties of the input and the target task i.e extracting the intends and associated information.

Sequence Labellings

Importance of Sequence Labelling for CAs.

- **NLU** : Understand User's Query
- **NLG** : Controlled Generation over content, avoid generic response problem [36, 10].

Sequence Labellings

Importance of Sequence Labelling for CAs.

- **NLU** : Understand User's Query
- **NLG** : Controlled Generation over content, avoid generic response problem [36, 10].

Speaker	**Utterance**	**Dialog Act (DA)**
A	Is there anyone who doesn't know Nancy ?	Yes/No Question
	Do you - Do you know Nancy ?	Question
B	Me ?	Question
	Mm-hmm	Backchannel
	I know Nancy	Yes/No Answer

TABLE – Example of Sequence Labelling with Dialog Act from [16]. Emotion Labels exhibit similar structure as DAs.

Sequence Labelling as an NMT problem [9]

SILICONE (Sequence labelling evaLuatIon benChmark fOr spoken laNguagE)

Limitations of current works on Sequence Labelling [20, 8, 18, 21] :

- Focus on type of label either **DA** or **Emotion**.
- Current methods require large corpora to train models from scratch.

SILICONE (Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE)

Limitations of current works on Sequence Labelling [20, 8, 18, 21] :

- Focus on type of label either **DA** or **Emotion**.
- Current methods require large corpora to train models from scratch.

Corpus	$\|Train\|$	$\|Val\|$	$\|Test\|$	Utt.	$\|Labels\|$	Task	Utt./$\|Labels\|$
SwDA* [16]	1k	100	11	200k	42	DA	4.8k
MRDA* [31]	56	6	12	110k	5	DA	2.6k
DyDA$_a$	11k	1k	1k	102k	4	DA	25.5k
MT* [33]	121	22	25	36k	12	DA	3k
Oasis* [19]	508	64	64	15k	42	DA	357
DyDA$_e$[22]	11k	1k	1k	102k	7	E	2.2k
MELD$_s$*[27]	934	104	280	13k	3	S	4.3k
MELD$_e$*	934	104	280	13k	7	S	1.8k
IEMO[5]	108	12	31	10k	6	E	1.7k
SEM [25]	62	7	10	5,6k	3	S	1.9k

TABLE – Statistics of datasets composing SILICONE. E stands for emotion label and S for sentiment label ; * stands for datasets with available official split. Sizes of Train, Val and Test are given in number of conversations.

Models

Notations Each conversation C_i is composed of utterances u, i.e $C_i = (u_1, u_2, \ldots, u_{|C_i|})$ with $Y_i = (y_1, y_2, \ldots, y_{|C_i|})$ the corresponding sequence of labels. An utterance u_i is a sequence of words, i.e $u_i = (\omega_1^i, \omega_2^i, \ldots, \omega_{|u_i|}^i)$.
Hierarchical Encoder It is composed of two functions f^u and f^c, satisfying :

$$\mathcal{E}_{u_i} = f_\theta^u(\omega_1, \ldots, \omega_{|u_i|}) \tag{1}$$
$$\mathcal{E}_{C_j} = f_\theta^d(\mathcal{E}_{u_1}, \ldots, \mathcal{E}_{C_j}) \tag{2}$$

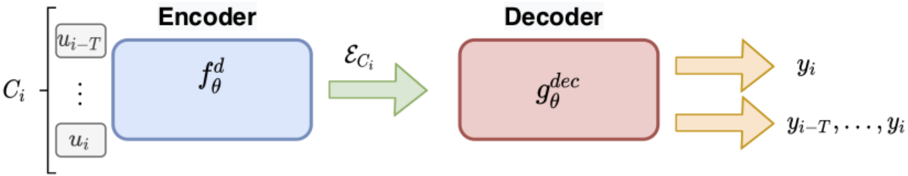


FIGURE – Model architecture for sequence labelling task. g_θ^{dec} represents the decoder layer.

Models

Notations Each conversation C_i is composed of utterances u, i.e $C_i = (u_1, u_2, \ldots, u_{|C_i|})$ with $Y_i = (y_1, y_2, \ldots, y_{|C_i|})$ the corresponding sequence of labels. An utterance u_i is a sequence of words, i.e $u_i = (\omega^i_1, \omega^i_2, \ldots, \omega^i_{|u_i|})$.

Hierarchical Encoder It is composed of two functions f^u and f^c, satisfying :

$$\mathcal{E}_{u_i} = f^u_\theta(\omega_1, \ldots, \omega_{|u_i|}) \tag{1}$$

$$\mathcal{E}_{C_j} = f^d_\theta(\mathcal{E}_{u_1}, \ldots, \mathcal{E}_{C_j}) \tag{2}$$

FIGURE – Model architecture for sequence labelling task. g^{dec}_θ represents the decoder layer.

We choose Transformer cells for f^u and f^c

Pretraining Objectives

Global hierarchical loss [30, 12]. The set of parameters θ is learnt by maximizing :

$$\mathcal{L}(\theta) = \underbrace{\lambda_u * \mathcal{L}^u(\theta)}_{\text{utterance level}} + \underbrace{\lambda_d * \mathcal{L}^d(\theta)}_{\text{dialog level}} \tag{3}$$

Pretraining Objectives

Global hierarchical loss [30, 12]. The set of parameters θ is learnt by maximizing :

$$\mathcal{L}(\theta) = \underbrace{\lambda_u * \mathcal{L}^u(\theta)}_{\text{utterance level}} + \underbrace{\lambda_d * \mathcal{L}^d(\theta)}_{\text{dialog level}} \tag{3}$$

$$\mathcal{L}^u_{\mathtt{MLM}}(\theta, u_i) = \mathbb{E}\left[\sum_{t \in m^{u_i}} \log(p_\theta(\omega^i_t | \tilde{u}_i))\right] \tag{4}$$

where $\tilde{u}_i$ is the corrupted utterance, $m^{u_i}_j \sim unif\{1, |u_i|\} \ \forall \ j \in [1, p_\omega]$

$$\mathcal{L}^d_{\mathtt{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega^j_i | \tilde{C}_k))\right] \tag{5}$$

where $m^{C_k}_j \sim unif\{1, |C_k|\} \ \forall \ j \in [1, p_{\mathcal{C}}]$ is the set of positions of masked utterances, $\tilde{C}_k$ is the corrupted context, and $p_{\mathcal{C}}$ is the proportion of masked utterances.

Pretraining Objectives

Global hierarchical loss [30, 12]. The set of parameters θ is learnt by maximizing :

$$\mathcal{L}(\theta) = \underbrace{\lambda_u * \mathcal{L}^u(\theta)}_{\text{utterance level}} + \underbrace{\lambda_d * \mathcal{L}^d(\theta)}_{\text{dialog level}} \tag{3}$$

$$\mathcal{L}^u_{\texttt{MLM}}(\theta, u_i) = \mathbb{E}\left[\sum_{t \in m^{u_i}} \log(p_\theta(\omega^i_t | \tilde{u}_i))\right] \tag{4}$$

where $\tilde{u}_i$ is the corrupted utterance, $m^{u_i}_j \sim unif\{1, |u_i|\} \ \forall\ j \in [1, p_\omega]$

$$\mathcal{L}^d_{\texttt{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega^j_i | \tilde{C}_k))\right] \tag{5}$$

where $m^{C_k}_j \sim unif\{1, |C_k|\} \ \forall\ j \in [1, p_{\mathcal{C}}]$ is the set of positions of masked utterances, $\tilde{C}_k$ is the corrupted context, and $p_{\mathcal{C}}$ is the proportion of masked utterances.
Pretaining Corpora : OpenSubtitles [23], ∼54M conversations and ∼270M utterances

Overall Results

- **Emotion and Sentiment Labels are noisier than DA**.
- Pretrained model achieves better results.

	Avg	SwDA	MRDA	$\mathrm{DyDA}_{\mathrm{DA}}$	MT	Oasis	DyDA_e	MELD_s	MELD_e	IEMO	SEM
BERT-4layers	70.4	77.8	90.7	79.0	88.4	66.8	90.3	55.3	53.4	43.0	58.8
$\mathcal{HR}$	69.8	77,5	90,9	80,1	82,8	64,3	91.5	59,3	59.9	40.3	51.1
$\mathcal{HT}(\theta^{u,d}_{MLM})$	73.3	**79.3**	92.0	80.1	90.0	68,3	92.5	62.6	59.9	42.0	66.6
$\mathcal{HT}(\theta^{d}_{GAP})$	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9

TABLE – Performances of different encoders when decoding using a MLP on SILICONE. The datasets are grouped by label type (DA vs E/S) and ordered by decreasing size.

Overall Results

- **Emotion and Sentiment Labels are noisier than DA**.
- Pretrained model achieves better results.

	Avg	SwDA	MRDA	DyDA_{DA}	MT	Oasis	DyDA_{e}	MELD_{s}	MELD_{e}	IEMO	SEM
BERT-4layers	70.4	77.8	90.7	79.0	88.4	66.8	90.3	55.3	53.4	43.0	58.8
$\mathcal{HR}$	69.8	77,5	90,9	80,1	82,8	64,3	91.5	59,3	59.9	40.3	51.1
$\mathcal{HT}(\theta^{u,d}_{MLM})$	73.3	**79.3**	92.0	80.1	90.0	68,3	92.5	62.6	59.9	42.0	66.6
$\mathcal{HT}(\theta^{d}_{GAP})$	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9

TABLE – Performances of different encoders when decoding using a MLP on SILICONE. The datasets are grouped by label type (DA vs E/S) and ordered by decreasing size.

- **Sequential nature** of the label Matters but ...

	Avg	Avg DA	Avg E/S
BERT (+MLP)	72,8	81.5	64.0
BERT (+GRU)	69.9	80.4	59.3
BERT (+CRF)	72.8	81.5	64.1

TABLE – Results on SILICONE for pre-trained BERT models.

Overall Results

- **Emotion and Sentiment Labels are noisier than DA**.
- Pretrained model achieves better results.

	Avg	SwDA	MRDA	DyDA$_{\text{DA}}$	MT	Oasis	DyDA$_e$	MELD$_s$	MELD$_e$	IEMO	SEM
BERT-4layers	70.4	77.8	90.7	79.0	88.4	66.8	90.3	55.3	53.4	43.0	58.8
$\mathcal{HR}$	69.8	77,5	90,9	80,1	82,8	64,3	91.5	59,3	59.9	40.3	51.1
$\mathcal{HT}(\theta^{u,d}_{MLM})$	73.3	**79.3**	92.0	80.1	90.0	68,3	92.5	62.6	59.9	42.0	66.6
$\mathcal{HT}(\theta^{d}_{GAP})$	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9

TABLE – Performances of different encoders when decoding using a MLP on SILICONE. The datasets are grouped by label type (DA vs E/S) and ordered by decreasing size.

- **Sequential nature** of the label Matters but ...

	Avg	Avg DA	Avg E/S
BERT (+MLP)	72,8	81.5	64.0
BERT (+GRU)	69.9	80.4	59.3
BERT (+CRF)	72.8	81.5	64.1

TABLE – Results on SILICONE for pre-trained BERT models.

- **Source** of the pretraining data matters.

Conclusion and Perspectives

Summary of our contributions

1. New Benchmark for Sequence Labeling Tasks (`SILICONE`)
2. Preprocess and pretrained on a large collection of Spoken Dialog (OpenSubtitles)
3. Explore New Pretraining Objectives
4. Hierarchical Transformers Based Encoder (Reduced Parameters and GPUs)

Future Work and Perspectives

1. Explore Sequence Labelling Tasks In a Multilingual Setting

Today's Agenda

1. Context
   - Conversational AI
   - Research Questions (RQs)
2. RQ1 : *How to best represent inputs that contains text for NLU ?*
3. RQ2 : *How to best represent inputs that contains text for NLG ?*
4. Conclusions and Planning
5. References

Learning A Disentangled Representation.

RQ2 : *How to best represent inputs that contains text for NLG ?*

Can we build representations that exhibit desirable topological properties (e.g invariance, disentanglement) for a specific sequence generation task ?.

Motivations

Importance of Disentangled Representations

- Visual Reasonning [34].
- Robust and Fair classification [2].
- Style Transfer [14], Conditional Generation [11, 4].

Motivations

Importance of Disentangled Representations

- Visual Reasonning [34].
- Robust and Fair classification [2].
- Style Transfer [14], Conditional Generation [11, 4].

Learning Disentangled representations.

Learn a model $\mathcal{M} : \mathcal{X} \to \mathcal{R}^d$
Goal : Output vector has to retain as much as possible information of the original content from the input sentence but as little as possible about the undesired attribute Y.
Assumption : $Y \in \mathcal{Y}$ is discrete.

Motivations

Importance of Disentangled Representations

- Visual Reasonning [34].
- Robust and Fair classification [2].
- Style Transfer [14], Conditional Generation [11, 4].

Learning Disentangled representations.

Learn a model $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}^d$
Goal : Output vector has to retain as much as possible information of the original content from the input sentence but as little as possible about the undesired attribute Y.
Assumption : $Y \in \mathcal{Y}$ is discrete.

We focus on applications **textual style transfer**.
Goal : Disentangled content and polarity (style).

Related Work

Adversarial Losses : an adversarial term in the training objective that aims at ensuring that sensitive attribute values [35, 2, 13].

Related Work

Adversarial Losses : an adversarial term in the training objective that aims at ensuring that sensitive attribute values [35, 2, 13].

Limitation of Adversarial Losses

- Disentanglement is not perfect [13].
- Adversarial Losses Fail for $|\mathcal{Y}| > 2$.

Related Work

Adversarial Losses : an adversarial term in the training objective that aims at ensuring that sensitive attribute values [35, 2, 13].

Limitation of Adversarial Losses

- Disentanglement is not perfect [13].
- Adversarial Losses Fail for $|\mathcal{Y}| > 2$.

Mutual Information Given two random variables Z and Y, the MI is defined by

$$I(Z;Y) = \mathbb{E}_{ZY}\left[\log \frac{p_{ZY}(Z,Y)}{p_Z(Z)p_Y(Y)}\right], \tag{6}$$

where p_{ZY} is the joint pdf and p_Z and p_Y are the marginal pdfs.

Contributions and Methods

Contributions

- A novel objective to train disentangled representations from attributes which include a new variational estimate of the MI
- Applications and numerical results and we demonstrate that the aforementioned surrogate is better suited than the widely used adversarial losses

General Loss to Minimize

$$\mathcal{L}(f_{\theta_e}) \equiv \underbrace{\mathcal{L}_{down.}(f_{\theta_e})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_{\theta_e}(X); Y)}_{\text{disentangled}}, \qquad (7)$$

$\mathcal{L}_{down.}$ represents a downstream specific (target task) loss
λ is a meta-parameter
f_{θ_e} is the encoding function

Estimating The MI

Computing the MI is a long standing challenge [3, 17, 26].

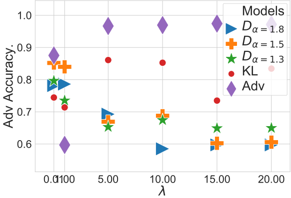
Variational upper bound on MI

Let (Z, Y) be an arbitrary pair of RVs with $(Z, Y) \sim p_{ZY}$ according to some underlying pdf, and let $Q_{\widehat{Y}|Z}$ be a conditional variational distribution on the attributes satisfying $P_{ZY} \ll P_Z \cdot Q_{\widehat{Y}|Z}$, i.e., absolutely continuous. Then, we have that

$$I(Z;Y) \leq \mathbb{E}_Y \left[-\log \int_{R^d} Q_{\widehat{Y}|Z}(Y|z) P_Z(dz) \right] + \mathbb{E}_{YZ} \left[\log Q_{\widehat{Y}|Z}(Y|Z) \right] + D_\alpha \left(P_{ZY} \| P_Z Q_{\widehat{Y}|Z} \right) \tag{8}$$

where $D_\alpha \left(P_{ZY} \| P_Z Q_{\widehat{Y}|Z} \right) = \frac{1}{\alpha - 1} \log \mathbb{E}_{ZY}[R^{\alpha-1}(Z, Y)]$ denotes the Renyi divergence and $R(z, y) = \frac{P_{Y|Z}(y|z)}{Q_{\widehat{Y}|Z}(y|z)}$, for all pairs $(z, y) \in \mathrm{Supp}(P_{ZY})$.

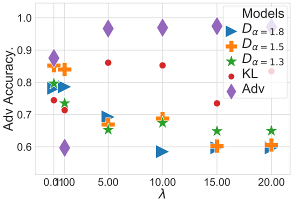
Application to Textual Style Transfer



Yelp corpus : Review from Yelp. The task consists in transferring a binary sentiment (positive/negative) [37, 24].

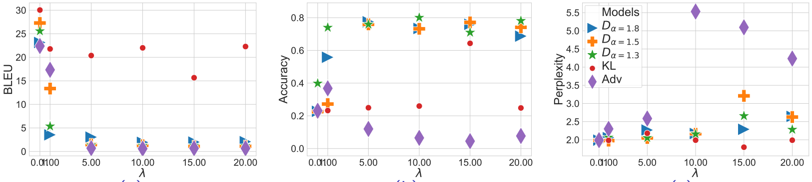
FIGURE – Disentanglement of the representations learnt by the encoder f_{θ_e} when the model is trained on a binary sentence generation task.

Application to Textual Style Transfer



Yelp corpus : Review from Yelp. The task consists in transferring a binary sentiment (positive/negative) [37, 24].

FIGURE – Disentanglement of the representations learnt by the encoder f_{θ_e} when the model is trained on a binary sentence generation task.



(a) (b) (c)

FIGURE – Numerical experiments on binary style transfer. Quality of generated sentences are evaluated using BLEU (6a) ; style transfer accuracy (6a) ; sentence fluency (6c).

Application to Textual Style Transfer

	Input	**It's freshly made, very soft and flavorful.**
0.1	Adv	it's crispy and too nice and very flavor.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	Input	it's freshly made, very soft and flavorful.
1	Adv	it's not crispy and not very flavorful flavor.
	KL	it's very fresh, and very flavorful and flavor.
	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	Input	it's freshly made, very soft and flavorful.
10	Adv	i hate this place.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the $ NUM.

TABLE – Sequences generated by the different models on the binary sentiment transfer task.

Conclusion and Perspectives

Summary of our contributions

1. New Estimate of MI based on an upper Bound
2. New method capable of learning disentangled textual representation
3. Better Trade off in Fair Classification
4. There is no free-lunch for sentence generation tasks : transferring style is easier with disentangled representations, but removes important information about the content.

Today's Agenda

1 Context
- Conversational AI
- Research Questions (RQs)

2 RQ1 : *How to best represent inputs that contains text for NLU ?*

3 RQ2 : *How to best represent inputs that contains text for NLG ?*

4 Conclusions and Planning

5 References

Conclusion and Perspectives

On going work on RQ1

1. **Multilingual DAs**. Learning more general encoders that can adapt to different languages. Improved version of Colombo(*), Chapuis(*) et al *AAAI 2020* and Chapuis(*), Colombo(*) et al. *Findings of EMNLP 2020*
2. **Learning Better Fusion Model**. Improved version of Garcia(*), Colombo(*) et al. *EMNLP 2019*.

Conclusion and Perspectives

On going work on RQ1

1. **Multilingual DAs**. Learning more general encoders that can adapt to different languages. Improved version of Colombo(*), Chapuis(*) et al *AAAI 2020* and Chapuis(*), Colombo(*) et al. *Findings of EMNLP 2020*
2. **Learning Better Fusion Model**. Improved version of Garcia(*), Colombo(*) et al. *EMNLP 2019*.

Follow Up Work (RQ1)

1. **Disfluency**. Planned follow up work on Dinkar(*), Colombo(*) et al. *EMNLP 2020*.

Conclusion and Perspectives

On going work on RQ1

1. **Multilingual DAs**. Learning more general encoders that can adapt to different languages. Improved version of Colombo(*), Chapuis(*) et al *AAAI 2020* and Chapuis(*), Colombo(*) et al. *Findings of EMNLP 2020*
2. **Learning Better Fusion Model**. Improved version of Garcia(*), Colombo(*) et al. *EMNLP 2019*.

Follow Up Work (RQ1)

1. **Disfluency**. Planned follow up work on Dinkar(*), Colombo(*) et al. *EMNLP 2020*.

Futur Work on (RQ2)

1. **New evaluation metrics**. An inherent problem that arise in NLG is the lack of evaluation metric relying on continuous representations. I believe we can do better than BertScore [38].

Conclusion and Perspectives

Month	Planned Work	Conference Paper Deadlines
November	Multi Modal	
December	Multi Lingual	NAACL
January		
February		ICML
March	Metric	ACL
April		ECML
May	Disfluency	NeurIPS, EMNLP
June		
July	Thesis Redaction	
August		
September		ICLR, AAAI
October		
November	PhD Defense	NAACL

TABLE – Planning of the remaining year of my PhD thesis.

Thank You!

Results presented in this report are the fruit of collaborations with my great PhD advisors Chloe Clavel, Giovanna Varni, Emmanuel Vignon **as well as with amazing senior researchers :** Pablo Piantanida, Matthieu Labeau, Matteo Manica, Florence D'Alche-Buc, Anne Sabourin, Eric Gaussier ,Slim Essid, Chouchang Jack Yang **and state of the art PhD students :** Tanvi Dinkar, Emile Chapuis, Alexandre Garcia **and last but not least** Hamid Jalalzai**. I am thankful to all of them for the hard-work and the fun moments we had and we will have.**

Today's Agenda

References I

Jeremy Ang, Yang Liu, and Elizabeth Shriberg.
Automatic dialog act segmentation and classification in multiparty meetings.
In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–1061. IEEE, 2005.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard.
Adversarial removal of demographic attributes revisited.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6331–6336, 2019.

References II

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm.
Mine : mutual information neural estimation.
arXiv preprint arXiv :1801.04062, 2018.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner.
Understanding disentangling in β-vae.
arXiv preprint arXiv :1804.03599, 2018.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan.
Iemocap : Interactive emotional dyadic motion capture database.
Language Resources and Evaluation, 42 :335–359, 12 2008.

References III

Wallace Chafe and Deborah Tannen.
The relation between written and spoken language.
Annual Review of Anthropology, 16(1) :383–407, 1987.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang.
A survey on dialogue systems : Recent advances and new frontiers.
Acm Sigkdd Explorations Newsletter, 19(2) :25–35, 2017.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He.
Dialogue act recognition via crf-attentive structured network.
In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234, 2018.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel.
Guiding Attention in Sequence-to-Sequence Models for Dialogue Act Prediction.
In *AAAI-20*, pages 7594–7601, 2020.

References IV

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia.
Affect-driven dialog generation.
arXiv preprint arXiv :1904.02793, 2019.

Emily L Denton et al.
Unsupervised learning of disentangled representations from video.
In *Advances in neural information processing systems*, pages 4414–4423, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT : Pre-Training of Deep Bidirectional Transformers for Language Understanding.
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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

References V

Yanai Elazar and Yoav Goldberg.
Adversarial removal of demographic attributes from text data.
arXiv preprint arXiv :1808.06640, 2018.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan.
Style transfer in text : Exploration and evaluation.
arXiv preprint arXiv :1711.06861, 2017.

Jianfeng Gao, Michel Galley, and Lihong Li.
Neural approaches to conversational ai.
In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374, 2018.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel.
Switchboard : Telephone speech corpus for research and development.
In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA, 1992. IEEE Computer Society.

References VI

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio.
Learning deep representations by mutual information estimation and maximization.
arXiv preprint arXiv :1808.06670, 2018.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi.
Dialogue act sequence labeling using hierarchical encoder with crf.
In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Geoffrey Leech and Martin Weisser.
Generic speech act annotation for task-oriented dialogues.
2003.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen.
A dual-attention hierarchical recurrent neural network for dialogue act classification.
CoRR, abs/1810.09154, 2018.

References VII

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen.
A dual-attention hierarchical recurrent neural network for dialogue act classification.
arXiv preprint arXiv :1810.09154, 2018.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu.
Dailydialog : A manually labelled multi-turn dialogue dataset, 2017.

Pierre Lison and Jörg Tiedemann.
Opensubtitles2016 : Extracting large parallel corpora from movie and tv subtitles.
2016.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han.
Mining quality phrases from massive text corpora.
In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1729–1744. ACM, 2015.

References VIII

Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder.
The semaine database : Annotated multimodal records of emotionally colored conversations between a person and a limited agent.
Affective Computing, IEEE Transactions on, 3 :5–17, 08 2013.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals.
Representation learning with contrastive predictive coding.
arXiv preprint arXiv :1807.03748, 2018.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea.
Meld : A multimodal multi-party dataset for emotion recognition in conversations, 2018.

Gisela Redeker.
On differences between spoken and written language.
Discourse processes, 7(1) :43–55, 1984.

References IX

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, et al.
Open-domain conversational agents : Current progress, open problems, and future directions.
arXiv preprint arXiv :2006.12442, 2020.

Victor Sanh, Thomas Wolf, and Sebastian Ruder.
A hierarchical multi-task approach for learning embeddings from semantic tasks.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6949–6956, 2019.

References X

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey.
The ICSI meeting recorder dialog act (MRDA) corpus.
In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA, April 30 - May 1 2004. Association for Computational Linguistics.

Andreas Stolcke and Elizabeth Shriberg.
Statistical language modeling for speech disfluencies.
In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 405–408. IEEE, 1996.

Henry Thompson, Anne Anderson, Ellen Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo.
The hcrc map task corpus : natural dialogue for speech recognition.
01 1993.

References XI

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem.
Are disentangled representations helpful for abstract visual reasoning ?
In *Advances in Neural Information Processing Systems*, pages 14245–14258, 2019.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig.
Controllable invariance through adversarial feature learning.
In *Advances in Neural Information Processing Systems*, pages 585–596, 2017.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur.
Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators.
arXiv preprint arXiv :1904.13015, 2019.

References XII

Xiao Yu, Xiang Ren, Yizhou Sun, Quanquan Gu, Bradley Sturt, Urvashi Khandelwal, Brandon Norick, and Jiawei Han.
Personalized entity recommendation : A heterogeneous information network approach.
In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 283–292. ACM, 2014.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi.
Bertscore : Evaluating text generation with bert.
arXiv preprint arXiv :1904.09675, 2019.

Matthias Zimmermann, Yang Liu, Elizabeth Shriberg, and Andreas Stolcke.
Toward joint segmentation and classification of dialog acts in multiparty meetings.
In *International Workshop on Machine Learning for Multimodal Interaction*, pages 187–193. Springer, 2005.

Back-up Slides

Modelling Disfluencies

RQ1 : *How to best represent inputs that contains text for NLU ?*

Can we build representations that take into account both the structural properties of the input and the target task i.e extracting the intends and associated information.

Introduction

- Spoken language is rarely fluent, filled with disfluencies.
- Fillers : disfluencies filling a pause in an utterance or conversation, "um" or "uh" in English.
- Meanings of fillers : Contextual, dependent on the perception of the listener.
- Feeling of another's knowing (FOAK) (Brennan Williams, 1995), perception of confidence.
- Despite rich linguistic literature, fillers SLU typically considered noise !

FIGURE – Source : Jorge Cham, www.phdcomics.com

Aim of this work

- O1 : Fillers play an important role in spoken language, should not be removed as noise.
- O2 : Fillers play an important role in the listener's perception of the speaker's expressed confidence (FOAK).
- Experimental validation of O1,O2, without handcrafting features.
- Efficiently represent and study informativeness of fillers using SOTA models.

Dataset

- POM dataset (Park et al., 2014), spontaneous speech, movie review videos. Speakers record videos of themselves giving a movie review.
  - Annotators : "How confident was the speaker ?"
  - Annotations of fillers and confidence available.
- Filler count high (4%), inter-annotator agreement confidence high (Kripps alpha = 0.73).
- Annotator were not asked to pay attention to the speaker's use of fillers.
- Monologues, dialogue related disfluencies (such as backchannels) are not present.

Dataset

- POM dataset (Park et al., 2014), spontaneous speech, movie review videos.
- Speakers record videos of themselves giving a movie review.
  - Annotators : "How confident was the speaker ?"
  - Annotations of fillers and confidence available.
- Filler count high (4%), inter-annotator agreement confidence high (Kripps alpha = 0.73).
- Annotator were not asked to pay attention to the speaker's use of fillers.
- Monologues, dialogue related disfluencies (such as backchannels) are not present.

O1 : Fillers play an important role in spoken language

- A language modelling task, using BERT, same MLM objective.
- Each experiment : token representation strategy (TR) and a pre-processing strategy (PS).

PS1 : Fillers removed	PS3 : Fillers kept
TR1 : No special treatment	TR2 : Special token for each

  - Raw (um) Things that (uh) you usually wouldn't find funny.
  - TR1 ['um', 'things', 'things', 'that', 'uh'...]
  - TR2 ['[FILLER-UM]', 'things', 'that', '[FILLER-UH]'...]
- Optionally fine-tune BERT for each PS and TR, using MLM.

O1 : Results

- Adding fillers, both with and without fine-tuning, model with lower perplexity.

Language Modelling task		Perplexity
w/o fine-tuning	PS1 : Fillers removed (Training + inference)	22
	PS3 : Fillers kept	20
w fine-tuning	PS1 : Fillers removed	5.5
	PS3 : Fillers kept	4.6

- Fixing PS3 as strategy, TR1 best. Better to keep the existing representations.

Best Token Representation		Perplexity
PS3	TR1 : No treatment	4.6
	TR2/3 : Treatment	4.7

- Interestingly, BERT unable to distinguish between two fillers.

O2 : Results

- Downstream confidence prediction task, informative.
  - Adding a Multi-Layer Perceptron (MLP) on top of a BERT.
  - Same pre-processing PS as before, fixed token rep TR1.
  - Optionally fine-tuned using the MLM.
  - Mean Squared Error (MSE) loss.
- Results
  - PS3 (Fillers kept in training+inference) outperform other PS, both with/without MLM fine tune.
  - Fillers, discriminative feature in confidence prediction.

Confidence Prediction task		MSE
w/o MLM	PS1 : Fillers removed	1.47
	PS3 : Fillers kept	1.30
w MLM	PS1 : Fillers removed	1.32
	PS3 : Fillers kept	1.24

Takeaway

- Fillers, improve results when working with contextualised word embeddings : LM, in spoken language, fillers leveraged to reduce uncertainty of BERT.
  - Unexpected, as intuitively, perplexity reduction as sentence simplified.
  - BERT, representation of fillers already exist.
- Downstream task of confidence/FOAK prediction.
  - Fillers, discriminative feature in confidence prediction.
  - Validation on spontaneous speech corpora.
- Unsupervised way of studying their informativeness.
- Future work : Acoustic representation, pre-trained representations.

Representing Multimodal Input for fine-grained opinion mining.

RQ1 : *How to best represent inputs that contains text for NLU ?*

Can we build representations that take into account both the structural properties of the input and the target task i.e extracting the intends and associated information.

Multimodal Opinion Mining

Definition of an opinion : The expression of opinions is an evaluation towards an object. The expression of such evaluations can be summarised by the combination of three components :

- a source (mainly the speaker) expressing a statement
- a target identifying the entity evaluated
- a polarised expression

FIGURE – Frames illustrating negative, positive or neutral opinions

Data Presentation : Garcia et Al. 2019

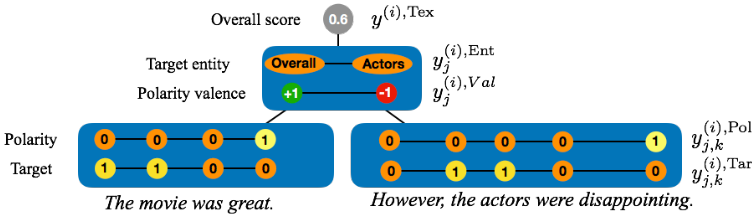


FIGURE – Structure of an annotated opinion

- *Token-level labels* are represented by a sequence of 2-dimensional binary label vectors
- *Sentence*-level labels carry 2 pieces of information : (1) the categorization of the target *entities*, (2) the sentence polarity (*Positive*, *Negative*, *Neutral/Mixed* and *None*).
- *Text*-level labels a continuous score summarizing the overall rating.

Motivations

In this work we show :

- The redundancy of the opinion information contained at different granularities can be leveraged improved opinion predictors.
- We propose several generic curriculum strategies to improve learning process in a MT setting.
- We demonstrate that jointly predicting entities and opinion helps to achieve better results

Background on Multi-modal Learning

When working with multimodal data several challenges need to be solved :

- **Representation** : Represent and summarize multimodal data in away that exploits the complementarity and redundancy.
- **Alignment** : Identify the direct relations between (sub)elements from two or more different modalities.
- **Fusion** : To join information from two or more modalities to perform a prediction task
- **Co-Learning** : Transfer knowledge between modalities, including their representations and predictive models.

Multi-Task Learning (1/3)

Based on these representations, we define a set of losses, $l^{(\text{Tok})}, l^{(\text{Sent})}, l^{(\text{Tex})}$ dedicated to measuring the similarity of each substructure prediction, $\hat{\mathrm{y}}^{(\text{Tok})}, \hat{\mathrm{y}}^{(\text{Sent})}, \hat{\mathrm{y}}^{(\text{Tex})}$ with the ground-truth.

$$l^{(Tok)}(\mathrm{y}^{\text{Tok}}, \hat{\mathrm{y}}^{\text{Tok}}) = -\frac{1}{2}\sum_i ((\mathrm{y}_i^{Pol}\log(\hat{\mathrm{y}}_i^{Pol}) + \quad \mathrm{y}_i^{Tar}\log(\hat{\mathrm{y}}_i{}^{Tar})),$$

$$l^{(Sent)}(\mathrm{y}^{Sent}, \hat{\mathrm{y}}^{Sent}) = -\frac{1}{2}\sum_i (\mathrm{y}_i^{\text{Ent}}\log(\hat{\mathrm{y}}_i^{\text{Ent}}) + \quad \mathrm{y}_i^{\text{Val}}\log(\hat{\mathrm{y}}_i^{\text{Val}})),$$

$$l^{(Tex)}(y^{\text{Tex}}, \hat{y}^{\text{Tex}}) = (y^{\text{Tex}} - \hat{y}^{\text{Tex}})^2,$$

The loss we optimise l, is a convex combination of these different task at each granularity level : $t \in \text{Tasks} = \{Tok, Sent, Tex\}$ weighted according to a set of task weights λ_t :

$$l(\mathrm{y}, \hat{\mathrm{y}}) = \frac{\sum_{t \in \text{Tasks}} \lambda_t l^{(t)}(\mathrm{y}^t, \hat{\mathrm{y}}^t)}{\sum_{t \in \text{Tasks}} \lambda_t}, \ \forall \lambda_t \geq 0. \tag{9}$$

Example of Strategies

In order to gradually guide the model from easy tasks to harder ones, λ_t is defined as function of the number of epochs of the form $\lambda_t^{(n_{\text{epoch}})} = \lambda_{\max} \frac{\exp\left((n_{\text{epoch}} - Ns_t)/\sigma\right)}{1+\exp\left((n_{\text{epoch}} - Ns_t)/\sigma\right)}$

- Strategy 1 (S1) consists in optimizing the different objectives one at a time from the easiest to the hardest. The underlying idea is that the low level labels are only useful as an initialization point for higher level ones.

- Strategy 2 (S2) consists in adding sequentially the different objectives to each other from the easiest to the hardest. This strategy relies on the idea that keeping a supervision on low level labels has a regularizing effect on high level ones.

Model Architectures (1/3)

From a general perspective, a hierarchical opinion predictor is composed of 3 functions $g^{\text{Tex}}, g^{\text{Sent}}, g^{\text{Tok}}$ encoding the dependency across the levels :

$$
\begin{aligned}
h_{j,k}^{(i),\text{Tok}} &= g_{\theta^{\text{Tok}}}^{\text{Tok}}(x_{j,:}^{(i),\text{Tok}}), & \hat{y}^{(i),\text{Tex}} &= \sigma^{\text{Tex}}(W^{\text{Tex}} h^{(i),\text{Tex}} + b^{\text{Tex}}),\\
h_j^{(i)^{\text{Sent}}} &= g_{\theta^{\text{Sent}}}^{\text{Sent}}(h_{j,:}^{(i)^{\text{Tok}}}), & \hat{y}_j^{(i),\text{Sent}} &= \sigma^{\text{Sent}}(W^{\text{Sent}} h_j^{(i)^{\text{Sent}}} + b^{\text{Sent}}),\\
h^{(i)^{\text{Tex}}} &= g_{\theta^{\text{Tex}}}^{\text{Tex}}(h_:^{(i)^{\text{Sent}}}). & \hat{y}_{j,k}^{(i),\text{Tok}} &= \sigma^{\text{Tok}}(W^{\text{Tok}} h_{j,k}^{(i),\text{Tok}} + b^{\text{Tok}})
\end{aligned}
$$

Choice of architectures for g_θ

- **Bidirectional Gated Recurrent Units (BiGRU)**
- **The Multi-attention Recurrent Network (MARN)** extends the traditional Long Short Term Memory.
- **Memory Fusion Networks (MFN)** are a second family of multi-view sequential models built upon a set of LSTM.

Several pooling strategies available (Last state representation, attention based sequence summarization)

Experiment 1 : Which architecture provides the best results on the task of fine grained opinion polarity prediction ?

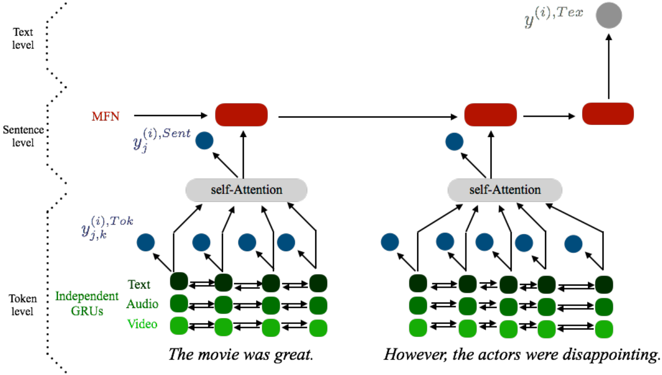


FIGURE – Example of architecture

Experiment 1 : Which architecture provides the best results on the task of fine grained opinion polarity prediction ?

	$\lambda_{Tok} = \lambda_{Sent} = 0$: no fine grained supervision				
Model / Metric	BiGRU	Ind BiGRU + att	MARN	MFN	Av Emb
MAE *Text*	0.35	0.38	0.29	0.32	**0.17**
	λ_{Tok},λ_{Sent} : Supervision at the token, sentence and review levels				
$\mu F1$ *Tokens*	0.90	**0.93**	0.90	0.89	X
$\mu F1$ *Sentence*	0.68	**0.75**	0.52	0.47	X
MAE *Text*	0.16	0.15	0.35	0.37	X

TABLE – Scores on sentiment label

Experiment 1 : Which architecture provides the best results on the task of fine grained opinion polarity prediction ? (2/2

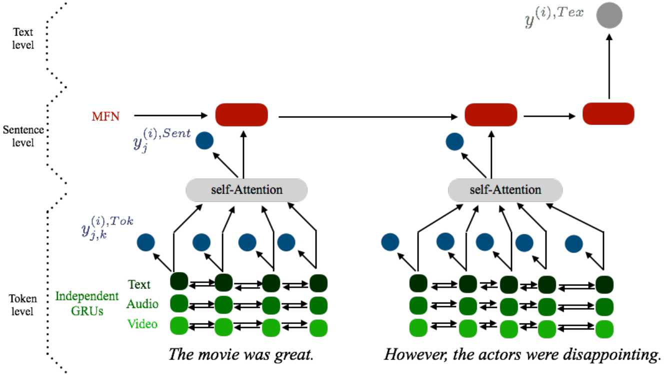


FIGURE – Best architecture selected during the Experiment 1

Experiment 2 : What is the best strategy to take into accou

multiple levels of opinion information ?

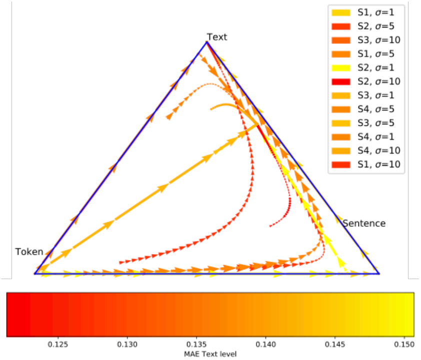


FIGURE – Path of the weight vector in the simplex triangle for the different tested strategies

Experiment 3 : Is it better to jointly predict opinions and entities ?

	Polarity labels	Entity labels	Polarity + entities
F1 polarity tokens	0.93	X	0.93
F1 polarity valence	0.75	X	0.75
F1 entities tokens	X	0.97	0.97
F1 entities Entities	X	Table **??**	Table **??**
MAE score review level	0.14	0.38	0.14

TABLE – Joint and independent prediction of entities and polarities

	Entity	Entity + Polarity	Value Count
Overall	0.71	**0.73**	1985
Actors	**0.65**	**0.65**	493
Screenplay	0.60	**0.63**	246
Atmosphere and mood	0.62	**0.64**	151
Vision and special effects	**0.62**	0.58	154

TABLE – F1 score per label for the top entity categories annotated at the sentence level (mean score averaged over 7 runs), value counts are provided on the test set.

Takeaways & Perspective

In this work we show :

- The redundancy of the opinion information contained at different granularities can be leveraged improved opinion predictors.
- We propose several generic curriculum strategies to improve learning process in a MT setting.
- We demonstrate that jointly predicting entities and opinion helps to achieve better results

Future work will explore the use of *structured output learning* methods dedicated to the opinion structure.

Learning A Dilation Invariant Representation.

RQ2 : *How to best represent inputs that contains text for NLG ?*

Can we build representations that exhibit desirable topological properties (e.g invariance, disentanglement) for a specific sequence generation task ?.

Motivations

Label Invariant Sentence Generation with Guarantees.

$$g\left(h_\lambda(\varphi(x))\right) = g\left(\varphi(x)\right). \tag{10}$$

$\forall \lambda \geq 1$ For an input sentence x, the embedding function is called φ and g a classifier.

Motivations

Label Invariant Sentence Generation with Guarantees.

$$g\left(h_\lambda(\varphi(x))\right) = g\left(\varphi(x)\right). \tag{10}$$

$\forall\lambda \geq 1$ For an input sentence x, the embedding function is called φ and g a classifier.
Limitations of Current Embeddings ELMo, BERT, XLNet trained on massive corpora show convenient properties but they do not fit our problem.

Motivations

Label Invariant Sentence Generation with Guarantees.

$$g\big(h_\lambda(\varphi(x))\big) = g\big(\varphi(x)\big). \tag{10}$$

$\forall \lambda \geq 1$ For an input sentence x, the embedding function is called φ and g a classifier.
Limitations of Current Embeddings ELMo, BERT, XLNet trained on massive corpora show convenient properties but they do not fit our problem.
Dilation Invariance : h_λ is chosen as the homothety with scale factor λ, $h_\lambda(x) = \lambda x$.

Definitions

Extrems

A r.v X is *extreme*, if $X \in \mathbb{R}_+ . \big(X > t, \text{ for some} t \geq 0.\big)$

Definitions

Extrems

A r.v X is *extreme*, if $X \in \mathbb{R}_+ . \big(X > t, \text{ for some} t \geq 0. \big)$

Heavy Tailed Distribution

A r.v X is *heavy-tailed*, if $X \in \mathbb{R}_+ . \big(\mathbb{P}(X > t) e^{ct} \xrightarrow[t \to \infty]{} \infty, \forall c > 0 \big)$

Definitions

Extrems

A r.v X is *extreme*, if $X \in \mathbb{R}_+ . \big(X > t, \text{ for some} t \geq 0.\big)$

Heavy Tailed Distribution

A r.v X is *heavy-tailed*, if $X \in \mathbb{R}_+ . \big(\mathbb{P}(X > t)e^{ct} \xrightarrow[t \to \infty]{} \infty, \forall c > 0\big)$

Example

- Exponential(λ) : $\mathbb{P}(X > t) = e^{-\lambda t}$ is not heavy tailed (choose $c < \lambda$)
- Pareto(α) : $\mathbb{P}(X > t) = 1/t^{\alpha}$ is heavy tailed.

Definitions

Extrems

A r.v X is *extreme*, if $X \in \mathbb{R}_+ . \big(X > t, \text{ for some} t \geq 0.\big)$

Heavy Tailed Distribution

A r.v X is *heavy-tailed*, if $X \in \mathbb{R}_+ . \big(\mathbb{P}(X > t)e^{ct} \xrightarrow[t \to \infty]{} \infty, \forall c > 0\big)$

Example

- Exponential(λ) : $\mathbb{P}(X > t) = e^{-\lambda t}$ is not heavy tailed (choose $c < \lambda$)
- Pareto(α) : $\mathbb{P}(X > t) = 1/t^{\alpha}$ is heavy tailed.

Jalalzai et al. NIPS 2018

There exists a classifier $g^{\star}_{\infty}$ depending on the pseudo-angle $\Theta(x) = \|x\|^{-1}x$ only, that is $g^{\star}_{\infty}(x) = g^{\star}_{\infty}\big(\Theta(x)\big)$, which is asymptotically optimal in terms of classification risk. The angle $\Theta(x)$ belongs to the positive orthant of the unit sphere.

Learning a Heavy-Tailed Representation (LHTR)

We suppose that :

- We are given a labelled dataset (x_i, y_i)
- We are given an embedding (in our case BERT)
- We are given an heavy-tailed distribution (Multivariate Logistic).

We learn φ using an adversarial approach.

Learning a Heavy-Tailed Representation (LHTR)

We suppose that :

- We are given a labelled dataset (x_i, y_i)
- We are given an embedding (in our case BERT)
- We are given an heavy-tailed distribution (Multivariate Logistic).

We learn φ using an adversarial approach.

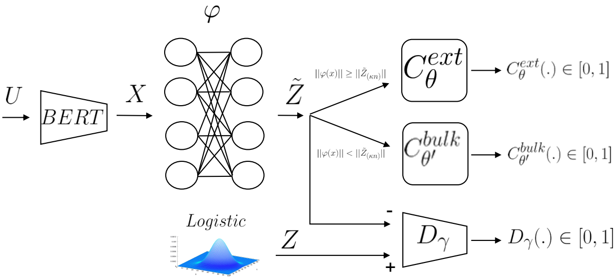
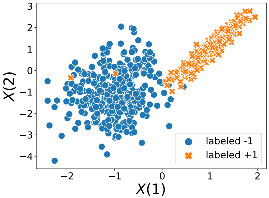


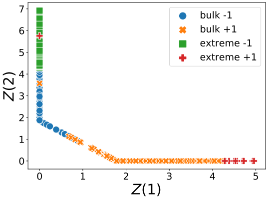
FIGURE – Pipeline to learn an heavy-tailed representation

Toy Datasets : bivariate data

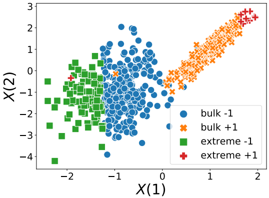
We begin with 2D Toy Datasets



(a) Input bivariate data



(b) Latent Space learnt by φ.



(c) Input Space with extremes from each class selected in the input space

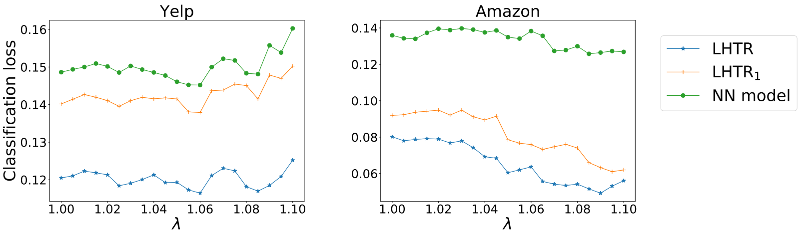
Dilation Invariance

C^{ext} solely depends on $Z/||Z||$!

Binary Classification in Extremes

Textual Datasets with binary rating

- Yelp : 1,450k reviews
- Amazon : 231k reviews



Figure

`NN model` is a MLP trained on BERT. For **`LHTR`**$_1$ a single MLP (C) is trained on φ. **`LHTR`**trains two separate MLP C^{ext} and C^{bulk} on φ.

Putting Pieces Together

Dilation Invariance

If Z is extreme, C^{ext} solely depends on $Z/||Z||$ implies that
$C^{\text{ext}}(\lambda Z) = C^{\text{ext}}(Z), \qquad \forall \lambda > 1.$

Putting Pieces Together

Dilation Invariance

If Z is extreme, C^{ext} solely depends on $Z/||Z||$ implies that
$C^{\text{ext}}(\lambda Z) = C^{\text{ext}}(Z), \qquad \forall \lambda > 1.$

Any dilation of a sample from the latent space will share the same label.

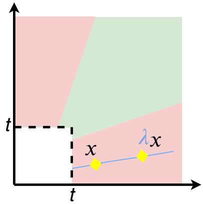


FIGURE – Illustration of dilation invariance property. Red region lies for $y = 1$ and green region for $y = 0$

Label Preserving Data Augmentation in Extremes

Label Invariant Sentence Generation with Guarantees.

$$g\left(h_\lambda(\varphi(x))\right) = g\left(\varphi(x)\right). \tag{11}$$

Label Preserving Data Augmentation in Extremes

Label Invariant Sentence Generation with Guarantees.

$$g\big(h_\lambda(\varphi(x))\big) = g\big(\varphi(x)\big). \tag{11}$$

Use a frozen φ for sentence generation using a seq2seq model.

Label Preserving Data Augmentation in Extremes

Label Invariant Sentence Generation with Guarantees.

$$g\big(h_\lambda(\varphi(x))\big) = g\big(\varphi(x)\big). \tag{11}$$

Use a frozen φ for sentence generation using a seq2seq model.

input	all of the tapas dishes were delicious !
$\lambda = 1$	all the tapas was delicious.
$\lambda = 1.1$	all tapas dishes were delicious !
$\lambda = 1.3$	all the tapas dishes were delicious !
$\lambda = 1.5$	the tapas were great !
input	there was hardly any meat.
$\lambda = 1$	there was almost no meat.
$\lambda = 1.1$	there was practically no meat.
$\lambda = 1.3$	there was almost no meat.
$\lambda = 1.5$	there was no meat.
input	i 'm not eating here !
$\lambda = 1$	i don't eat here.
$\lambda = 1.1$	i don't eat here !
$\lambda = 1.3$	i'm not going to eat here !
$\lambda = 1.5$	i will never going to eat here !