AdvNLE Seminar 5

Sequence Classification / Propaganda Detection

Dr Julie Weeds, Spring 2023

Warm-up

- What is the difference between sequence labelling and sequence classification? How many examples can you come up with of applications of each.

- Can you think of ways that either could be applied in the following scenarios:
  - Machine translation?
  - Summarization?

Previously

- Distributional models of word meaning
  - how similar are two words based on how they are used in text?
- Language models
  - how likely is a sequence of words in a language?
- Neural language models
- Sequence Labelling (Named Entity Recognition)

This week

- Sequence classification
- Document / sequence representations
- Evaluation
- Distributional representations of meaning
- Composition
- Propaganda detection

Sequence classification

- Aka document classification, text classification
- Make a single sequence-level classification decision per sentence / per document
- Classification could be
  - *sentiment, topic, relevance … spam, hate speech, machine-generated …*
- Classification could be
  - *Binary or multi-class*
- Classification could be
  - *Hard or soft*

Evaluation

- Is class distribution balanced?
- Accuracy
- Precision, Recall, F1

Actual Class

Predicted Class	1	2	3	4
1	**TP**	**FP**	**FP**	**FP**
2	**FN**	TN	TN	TN
3	**FN**	TN	TN	TN
4	**FN**	TN	TN	TN

Actual Class

Predicted Class	1	0
1	TP	FP
0	FN	TN

- In multi-class scenario, need to compute precision, recall and F1 for each class
- Macro-average => unweighted average
- Micro-average => average weighted by the size of each class

Representing sequences for classification

- Classical document-level representation is **bag-of-words**
- What are the benefits and limitations of this?

	the	plot	is	great	not	boring	dull
1. The plot is great	1	1	1	1	0	0	0
2. The plot is not great	1	1	1	1	1	0	0
3. The great plot is not boring	1	1	1	1	1	1	0
4. The boring plot is not great	1	1	1	1	1	1	0
5. The plot is dull	1	1	1	0	0	0	1

Using Word Embeddings

- Find the **sum/centroid/max** of all of the word embeddings for words in the sequence
- Pass this as input to a classifier (e.g., logistic regression / SVM / NN)

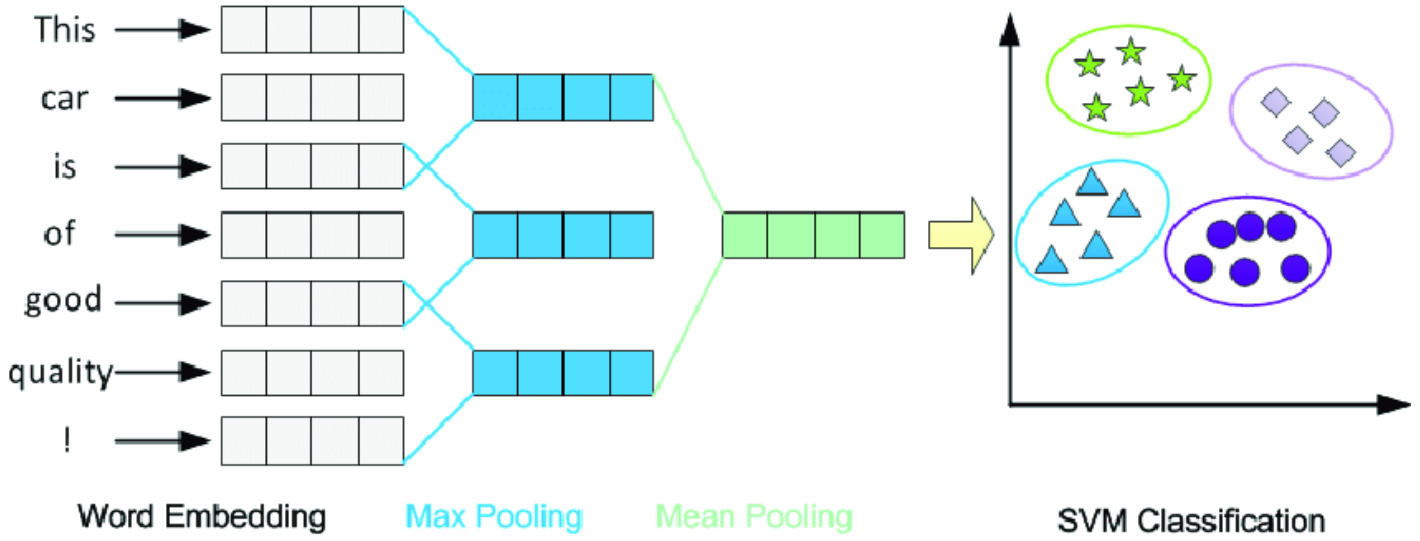


Figure from Pan et al. 2019

Question

- What are the benefits / disadvantages of using word embeddings in this way?

Distributed Representations of Word Meaning

- Distributional Hypothesis
- *"words which mean similar things tend to behave in similar ways"*
- i.e., they co-occur with similar words

	miaows	....	elected	...	decides	...	comfy	...
leader	0		3		5		0	
president	0		5		5		0	
ruler	0		3		5		0	
cat	5		0		0		1	
chair	0		0		0		5	

Distributed Representations of Sentential Meaning

- Can we do the same as for words?
  - Hypothesize that sentences which mean similar things tend to behave in similar ways?
  - Collect all of the contexts of sentences
  - Represent using vectors and compare
- Why / why not?

The Principle of Compositionality

- widely attributed to Gottlieb Frege, but assumed by others e.g., George Boole
- ***"The meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them"***

Composition for Meaning Representations

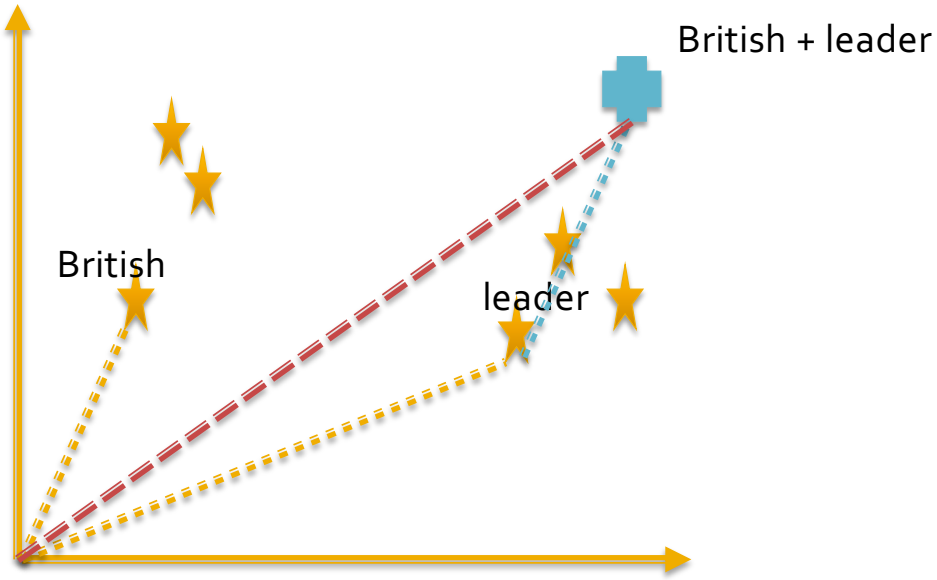
- **Constituent** expressions are **words**
- Words are represented by distributional representations / **embeddings** (e.g., Word2Vec or GloVe)

- So to get a representation of a sentence we need to …
- … compose the embeddings of the constituent words

- How? What are the rules for composition?

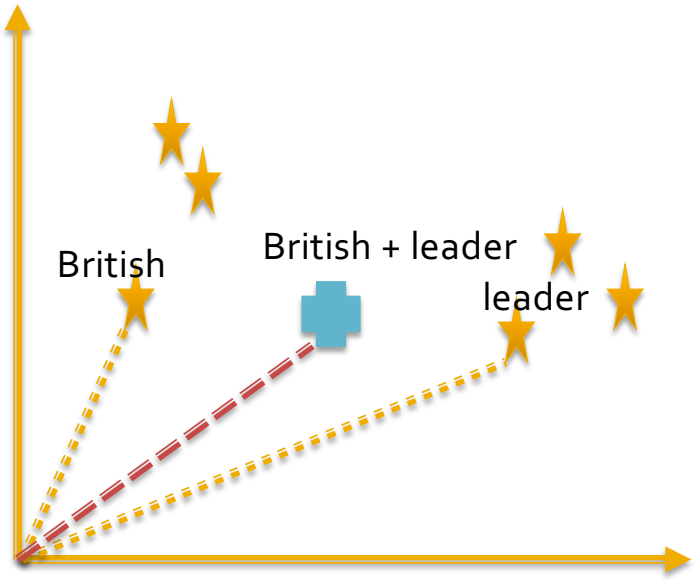
Additive composition

- Simply add the vectors



Additive composition

- Or average the vectors (find the centroid)



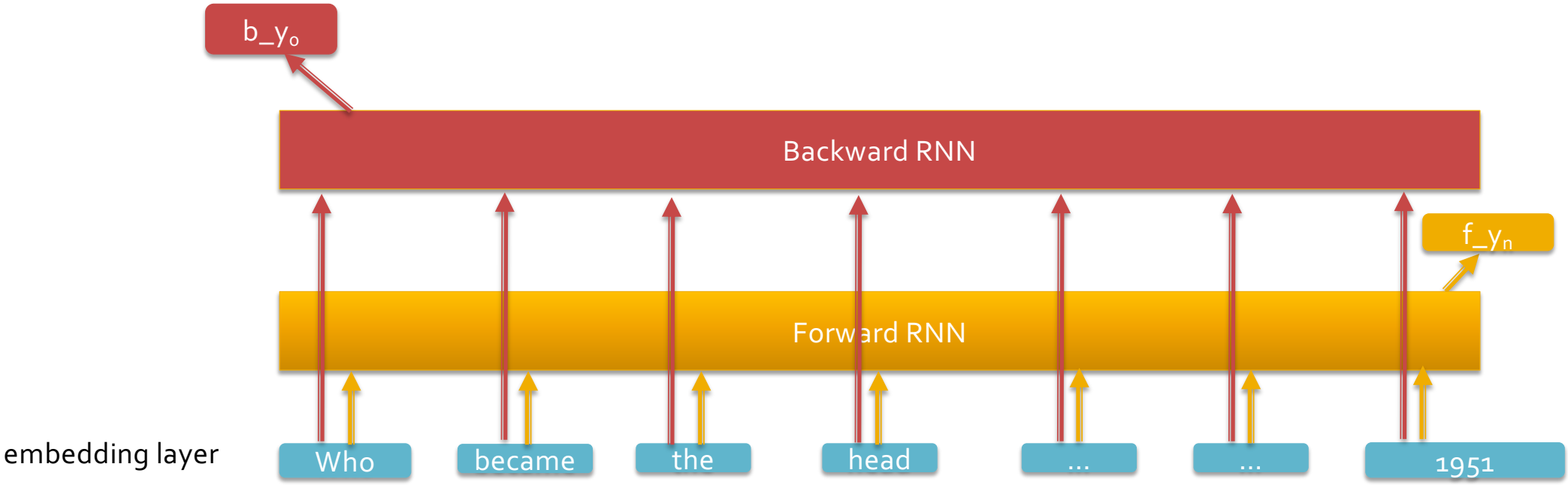
- remember, we are interested in cosine similarity between vectors
- → direction
- so very little difference between adding and averaging
- especially if vectors are normalised to unit length

Disadvantages of Adding Word Embeddings

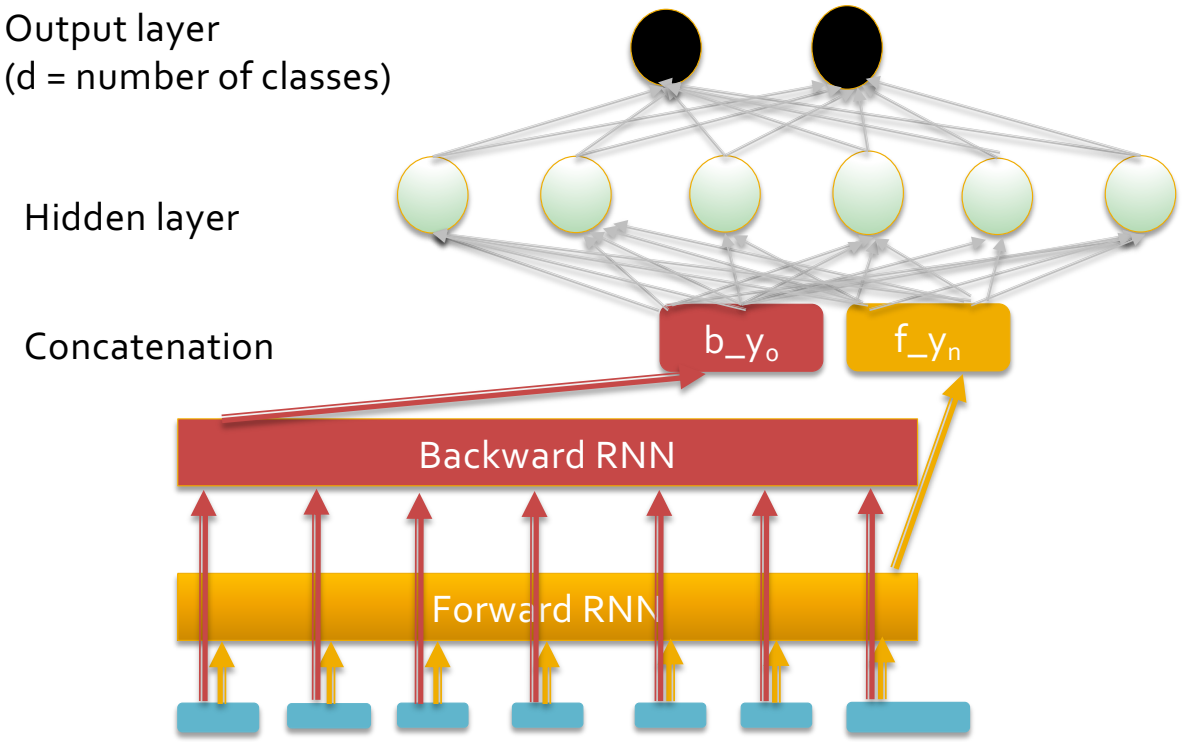
- Word embeddings are **uncontextualised**
  - contain a mixture of usages of all senses of the word
  - e.g., *head* as part of the body and *head* as leader
  - but only one sense is intended in a given sentence
- Pays no attention to word order or syntax
- What about function words such as *in*, *on*, *every* and *not*

Language modelling for Sequence Representations

- Represent a sequence by what is predicted to the left and right of it
- e.g., concatenation of [b_y_o,f_y_n]



Classification using representations from LMs



Representations taken from last hidden layer (not output layer)

Trained to predict previous word

Trained to predict next word

Propaganda Detection (Part 2)

Da San Martino et al. (2020) introduce a competition to detect and classify propaganda techniques in text. When reading this paper, do not be overly concerned with the different systems which took part in the competition. We will focus on the overall idea of propaganda detection, the two tasks introduced in this paper (span identification and technique classification), the dataset and the evaluation metrics. Once you have read the paper, consider the following questions.

1. What do you understand by the term propaganda and why might it be important to develop systems which can automatically detect propaganda in text?
2. Why is automatic propaganda detection difficult?
3. Give examples of 3 different propaganda techniques being used in text. Explain why this is propaganda.
4. What textual features might be useful to help a system detect propaganda?
5. Describe the pipeline proposed by the paper for propaganda identification. Can you think of any alternatives? What advantages / disadvantages are there of each?
6. How was the PTC-SemEval20 corpus collected and annotated? What do you understand by "the γ agreement on the annotated articles is on average 0.6"?
7. How do the authors evaluate systems on the span identification task?
8. Micro-average F_1 is used to evaluate systems on the technique classification task. The authors state that for a single-label task, this is equivalent to accuracy. Explain
9. Outline one method which could be used to carry out span identification.
10. Outline one method which could be used to carry out techniques classification.
11. Systems were evaluated for span identification on both the development set and the test set. Why do you think the results are not the same on both?
12. What is the predominant propaganda technique found in the corpus? If a system labelled every propaganda snippet with this label, how would it do? What do you think of the system results for techniques classification (Table 6)?

Question 1

- What do you understand by the term propaganda and why might it be important to develop systems which automatically detect propaganda in text?

Question 1

- What do you understand by the term propaganda and why might it be important to develop systems which automatically detect propaganda in text?

Propaganda comes in many forms, but it can be recognized by its persuasive function, sizable target audience, the representation of a specific group's agenda, and the use of faulty reasoning and/or emotional appeals (Miller, 1939). The term *propaganda* was coined in the 17th century, and initially referred to the propagation of the Catholic faith in the New World (Jowett and O'Donnell, 2012a, p. 2). It soon took a pejorative connotation, as its meaning was extended to also mean opposition to Protestantism. In more recent times, the Institute for Propaganda Analysis (Ins, 1938) proposed the following definition:

> **Propaganda.** *Expression of opinion or action by individuals or groups deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends.*

Recently, Bolsover and Howard (2017) dug deeper into this definition identifying its two key elements: (*i*) trying to influence opinion, and (*ii*) doing so on purpose.

Question 2

- Why is automatic propaganda detection difficult?

Question 2

- Why is automatic propaganda detection difficult?
- At least 3 reasons:
  1. ....
  2. ....
  3. ....

Question 3

- Give examples of 3 different propaganda techniques being used in text. Explain why this is propaganda

#	Technique	Snippet
1	Loaded language	**Outrage** as Donald Trump suggests injecting disinfectant to kill virus.
2	Name calling, labeling	WHO: Coronavirus emergency is '**Public Enemy Number 1**'
3	Repetition	I still have a **dream**. It is a **dream** deeply rooted in the American **dream**. I have a **dream** that one day . . .
4	Exaggeration, minimization	Coronavirus **'risk to the American people remains very low'**, Trump said.
5	Doubt	**Can the same be said for the Obama Administration**?
6	Appeal to fear/prejudice	**A dark, impenetrable and "irreversible" winter of persecution of the faithful by their own shepherds will fall**.
7	Flag-waving	Mueller attempts **to stop the will of We the People**!!! It's time to jail Mueller.
8	Causal oversimplification	**If France had not have declared war on Germany then World War II would have never happened.**
9	Slogans	**"BUILD THE WALL!"** Trump tweeted.
10	Appeal to authority	**Monsignor Jean-Franois Lantheaume, who served as first Counsellor of the Nunciature in Washington, confirmed that "Vigan said the truth. That's all."**
11	Black-and-white fallacy	Francis said these words: "**Everyone is guilty for the good he could have done and did not do . . . If we do not oppose evil, we tacitly feed it**."
12	Thought-terminating cliché	**I do not really see any problems there.** Marx is the President.
13	Whataboutism	President Trump —**who himself avoided national military service** in the 1960's— keeps beating the war drums over North Korea.
	Straw man	"Take it seriously, but with a large grain of salt." **Which is just Allen's more nuanced way of saying: "Don't believe it**."
	Red herring	**"You may claim that the death penalty is an ineffective deterrent against crime – but what about the victims of crime? How do you think surviving family members feel when they see the man who murdered their son kept in prison at their expense? Is it right that they should pay for their son's murderer to be fed and housed?"**
14	Bandwagon	He tweeted, "**EU no longer considers #Hamas a terrorist group. Time for US to do same."**
	Reductio ad hitlerum	"Vichy journalism," a term which now fits so much of the mainstream media. **It collaborates in the same way that the Vichy government in France collaborated with the Nazis.**

Table 1: The 14 propaganda techniques with examples, where the propaganda span is shown in bold.

Question 4

- What textual features might be useful to help a system detect propaganda?

#	Technique	Snippet
1	Loaded language	**Outrage** as Donald Trump suggests injecting disinfectant to kill virus.
2	Name calling, labeling	WHO: Coronavirus emergency is '**Public Enemy Number 1**'
3	Repetition	I still have a **dream**. It is a **dream** deeply rooted in the American **dream**. I have a **dream** that one day . . .
4	Exaggeration, minimization	Coronavirus **'risk to the American people remains very low'**, Trump said.
5	Doubt	**Can the same be said for the Obama Administration**?
6	Appeal to fear/prejudice	**A dark, impenetrable and "irreversible" winter of persecution of the faithful by their own shepherds will fall**.
7	Flag-waving	Mueller attempts **to stop the will of We the People**!!! It's time to jail Mueller.
8	Causal oversimplification	**If France had not have declared war on Germany then World War II would have never happened.**
9	Slogans	**"BUILD THE WALL!"** Trump tweeted.
10	Appeal to authority	**Monsignor Jean-Franois Lantheaume, who served as first Counsellor of the Nunciature in Washington, confirmed that "Vigan said the truth. That's all."**
11	Black-and-white fallacy	Francis said these words: "**Everyone is guilty for the good he could have done and did not do . . . If we do not oppose evil, we tacitly feed it**."
12	Thought-terminating cliché	**I do not really see any problems there.** Marx is the President.
13	Whataboutism	President Trump —**who himself avoided national military service** in the 1960's— keeps beating the war drums over North Korea.
	Straw man	"Take it seriously, but with a large grain of salt." **Which is just Allen's more nuanced way of saying: "Don't believe it**."
	Red herring	**"You may claim that the death penalty is an ineffective deterrent against crime – but what about the victims of crime? How do you think surviving family members feel when they see the man who murdered their son kept in prison at their expense? Is it right that they should pay for their son's murderer to be fed and housed?"**
14	Bandwagon	He tweeted, "**EU no longer considers #Hamas a terrorist group. Time for US to do same."**
	Reductio ad hitlerum	"Vichy journalism," a term which now fits so much of the mainstream media. **It collaborates in the same way that the Vichy government in France collaborated with the Nazis.**

Table 1: The 14 propaganda techniques with examples, where the propaganda span is shown in bold.

Question 5

- Describe the pipeline proposed by the paper for propaganda identification. Can you think of any alternatives? What advantages / disadvantages are there of each?

Question 5

- Describe the pipeline proposed by the paper for propaganda identification. Can you think of any alternatives? What advantages / disadvantages are there of each?

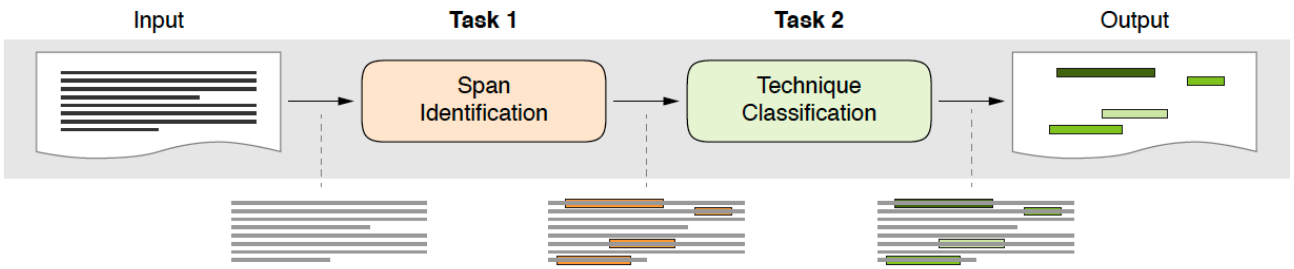


Figure 1: The full propaganda identification pipeline, including the two subtasks: Span Identification and Technique Classification.

Question 6

- How was the PTC-SemEval20 corpus collected and annotated? What do you understand by "the gamma agreement on the annotated articles is on average 0.6"?

Question 6

- How was the PTC-SemEval20 corpus collected and **annotated**?

Input article

Manchin says Democrats acted like **babies** at the SOTU

In a glaring sign of just how **stupid** and **petty** things have become in Washington these days [...] State of the Union speech not looking as though Trump **killed his grandma**. [...]

Annotation file

Article ID	Technique	Start	End
123456	Name_Calling	34	40
123456	Loaded_Language	83	89
123456	Loaded_Language	94	99
123456	Loaded_Language	350	368
...	...		

Figure 2: Example of a plain-text article (left) and its annotation (right). The *Start* and the *End* columns are the indices representing the character span of the spotted technique.

Question 6

- **What do you understand by "the gamma agreement on the annotated articles is on average 0.6"?**

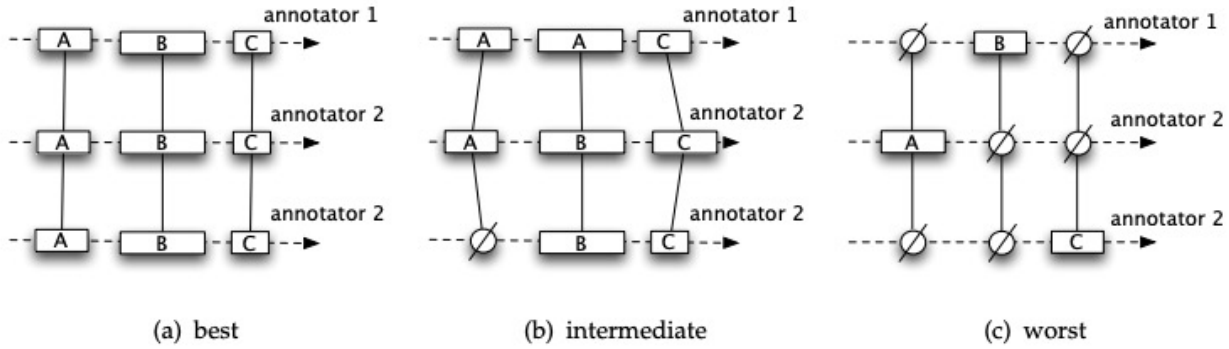


Figure 11
Examples of best, intermediate, and worst possible disorders.

$$\forall s \in c, \gamma = 1 - \frac{\delta(s)}{\delta_e(c)} \tag{8}$$

If all annotators perfectly agree (Figure 11a), $\gamma = 1$. Figure 11c corresponds to the worst case, where the annotators are worse than annotating at random, with $\gamma < 0$. Figure 11b shows an intermediate situation.

From Mathet at al. 2015

Question 7

- How do the authors evaluate systems on the span identification task?

Question 7

- How do the authors evaluate systems on the span identification task?

Let d be a news article in a set D. A gold span t is a sequence of contiguous indices of the characters composing a text fragment $t \subseteq d$. For example, in Figure 4 (top-left) the gold fragment *"stupid and petty"* is represented by the set of indices $t_1 = [4, 19]$. We denote with $T_d = \{t_1, \ldots, t_n\}$ the set of all gold spans for an article d and with $T = \{T_d\}_d$ the set of all gold annotated spans in D. Similarly, we define $S_d = \{s_1, \ldots, s_m\}$ and S to be the set of predicted spans for an article d and a dataset D, respectively. We compute precision P and recall R by adapting the formulas in (Potthast et al., 2010):

$$P(S,T) = \frac{1}{|S|} \cdot \sum_{d \in D} \sum_{s \in S_d, t \in T_d} \frac{|(s \cap t)|}{|t|}, \quad (1)$$

$$R(S,T) = \frac{1}{|T|} \cdot \sum_{d \in D} \sum_{s \in S_d, t \in T_d} \frac{|(s \cap t)|}{|s|}. \quad (2)$$

Question 8

- Micro-average F1 is used to evaluate systems on the techniques classification task. The authors state that for a single-label tast, this is equivalent to accuracy. Explain.

Evaluation

- Is class distribution balanced?
- Accuracy
- Precision, Recall, F1

Actual Class

Predicted Class	1	2	3	4
1	**TP**	**FP**	**FP**	**FP**
2	**FN**	TN	TN	TN
3	**FN**	TN	TN	TN
4	**FN**	TN	TN	TN

Actual Class

Predicted Class	1	0
1	TP	FP
0	FN	TN

- In multi-class scenario, need to compute precision, recall and F1 for each class
- Macro-average => unweighted average
- Micro-average => average weighted by the size of each class

Question 9

- Outline one method which could be used to carry out span identification.

Question 10

- Outline one method which could be used to carry out technique classification.

Question 11

- Systems were evaluated for span identification on both the development set and the test set. Why do you think the results are not the same on both?

Question 11

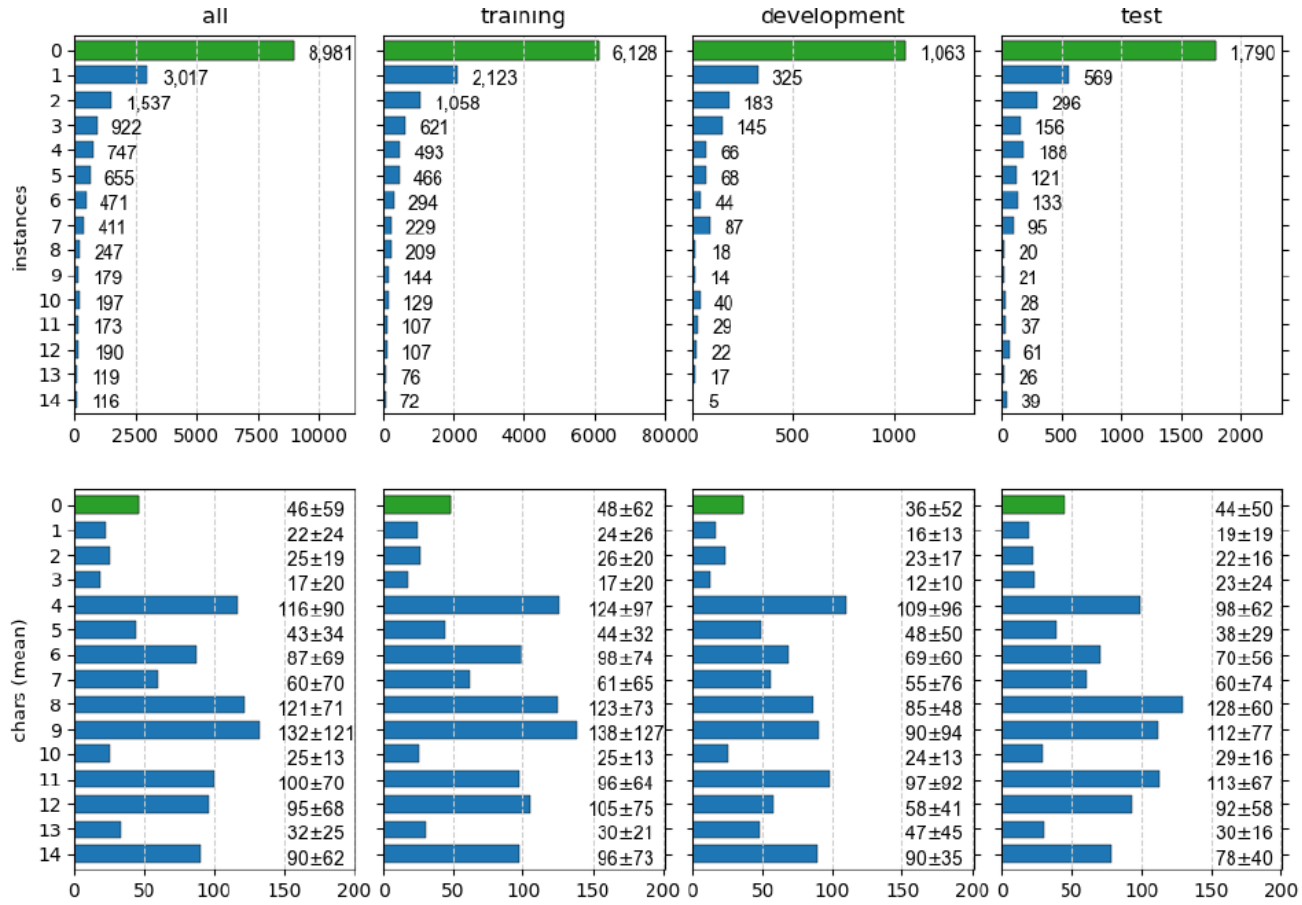
	Test				Development			
Team	**Rnk**	**F_1**	**P**	**R**	**Rnk**	**F_1**	**P**	**R**
Hitachi	1	**51.55**	56.54	47.37	4	**50.12**	42.26	61.56
ApplicaAI	2	49.15	59.95	41.65	3	52.19	47.15	58.44
aschern	3	49.10	53.23	45.56	5	49.99	44.53	56.98
LTIatCMU	4	47.66	50.97	44.76	7	49.06	43.38	56.47
UPB	5	46.06	58.61	37.94	8	46.79	42.44	52.13
Fragarach	6	45.96	54.26	39.86	12	44.27	41.68	47.21
NoPropaganda	7	44.68	55.62	37.34	9	46.13	40.65	53.31
CyberWallE	8	43.86	42.16	45.70	17	42.39	33.45	57.86
Transformers	9	43.60	49.86	38.74	14	43.06	40.85	45.52
SWEAT	10	43.22	52.77	36.59	16	42.51	42.97	42.06
YNUtaoxin	11	43.21	55.62	35.33	11	44.35	40.74	48.67
DREAM	12	43.10	54.54	35.63	19	42.15	42.66	41.65
newsSweeper	13	42.21	46.52	38.63	10	44.45	38.76	52.10
PsuedoProp	14	41.20	41.54	40.87	22	39.32	34.27	46.11
Solomon	15	40.68	53.95	32.66	15	42.86	43.24	42.49
YNUHPCC	16	40.63	36.55	45.74	18	42.27	32.08	61.95
NLFIIT	17	40.58	50.91	33.73	21	39.67	35.04	45.72
PALI	18	40.57	53.20	32.79	2	52.35	49.64	55.37
UESTCICSA	19	39.85	56.09	30.90	13	44.17	43.21	45.18
TTUI	20	39.84	**66.88**	28.37	6	49.59	48.76	50.44
BPGC	21	38.74	49.39	31.88	25	36.79	34.72	39.12
DoNotDistribute	22	37.86	42.36	34.23	24	37.73	32.41	45.12
UTMNandOCAS	23	37.49	37.97	37.03	31	34.35	23.65	62.69
Entropy	24	37.23	41.68	33.63	32	32.89	30.82	35.25
syrapropa	25	36.20	49.53	28.52	1	53.40	39.88	80.80

- Systems were evaluated for span identification on both the development set and the test set. Why do you think the results are not the same on both?

Question 12

- What is the predominant propaganda technique found in the corpus? If a system labelled every propaganda snippet with this label, how would it do? What do you think of the system results for technique classification (Table 6)?

- What is the predominant propaganda technique found in the corpus?
- If a system labelled every propaganda snippet with this label, how would it do?



What do you think of the system results for technique classification (Table 6)?

Rnk	Team	Overall	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	ApplicaAI	**62.07**	77.12	74.38	**54.55**	33.59	56.23	**45.49**	**69.43**	22.73	51.28	48.15	49.02	39.22	25.00	8.33
2	aschern	62.01	77.02	**75.65**	53.38	32.65	59.44	41.78	66.35	25.97	**54.24**	35.29	**53.57**	**42.55**	18.87	14.93
3	Hitachi	61.73	75.64	74.20	37.88	34.58	**63.43**	38.94	68.02	**36.62**	45.61	40.00	47.92	29.41	**26.92**	4.88
4	Solomon	58.94	74.66	70.75	42.53	28.44	61.82	39.39	61.84	19.61	50.75	26.67	42.00	38.10	0.00	4.88
5	newsSweeper	58.44	75.32	74.23	20.69	37.10	56.55	42.80	60.53	19.72	50.75	41.67	25.00	21.62	8.00	13.04
6	NoPropaganda	58.27	**77.17**	73.90	42.71	**37.99**	56.27	38.02	59.30	12.12	42.42	23.26	8.70	23.26	0.00	0.00
7	Inno	57.99	73.31	74.30	24.89	35.39	58.65	45.09	59.41	24.32	43.75	43.14	40.40	29.63	19.36	10.71
8	CyberWallE	57.37	74.68	70.92	47.68	28.34	58.65	39.84	54.38	15.39	39.39	14.63	23.68	23.81	0.00	12.25
9	PALI	57.32	74.29	69.09	24.56	28.57	58.97	36.59	61.62	30.59	39.22	27.59	39.62	40.82	20.90	**28.57**
10	DUTH	57.21	73.71	71.41	20.10	28.24	59.16	33.33	58.95	26.23	34.78	44.44	33.33	27.03	17.78	9.30
11	DiSaster	56.65	74.49	68.10	20.44	30.64	59.12	35.25	58.25	14.63	42.55	51.16	26.67	19.05	4.35	20.41
12	djichen	56.54	73.21	68.38	29.75	31.42	60.00	33.65	56.19	22.79	30.77	37.50	43.81	27.91	18.87	20.83
13	SocCogCom	55.81	72.18	67.34	18.88	34.86	60.40	31.62	54.26	6.35	40.91	28.57	26.51	23.53	10.00	9.76
14	TTUI	55.64	73.22	68.49	21.18	32.20	57.40	41.48	61.68	23.08	37.50	28.24	35.29	25.00	20.29	24.56
15	JUST	55.31	71.96	64.73	21.94	29.57	58.26	37.10	62.56	27.27	33.33	**48.89**	28.89	31.82	28.57	24.49
16	NLFIIT	55.25	72.55	69.30	21.55	30.30	55.66	24.89	63.32	0.00	41.67	29.63	32.10	13.64	0.00	9.30
17	UMSIForeseer	55.14	73.02	70.79	21.49	28.57	57.21	31.97	56.14	0.00	39.22	29.41	0.00	14.29	0.00	9.76
18	BPGC	54.81	71.58	67.51	23.74	33.47	53.78	33.65	58.93	24.18	40.00	30.77	40.00	20.69	20.90	12.50
19	UPB	54.30	70.09	68.86	20.00	30.62	52.55	30.00	55.87	16.95	34.62	20.00	19.72	22.86	4.88	0.00
20	syrapropa	54.25	71.47	68.44	30.77	28.10	56.14	29.77	57.02	21.51	29.03	31.58	30.61	28.57	9.09	19.61
21	WMD	52.01	69.33	64.67	13.89	25.46	53.94	29.20	52.08	5.71	6.90	7.14	0.00	7.41	0.00	5.00
22	YNUHPCC	50.50	68.08	62.33	17.72	21.54	51.04	26.40	55.56	3.45	27.59	29.79	38.38	17.78	15.00	13.79
23	UESTCICSA	49.94	68.23	66.88	27.96	25.44	44.99	22.75	53.14	3.74	41.38	12.77	11.27	28.57	3.70	0.00
24	DoNotDistribute	49.72	68.44	60.65	19.44	27.23	46.25	29.75	53.76	14.89	28.07	22.64	24.49	12.25	9.68	4.55
25	NTUAAILS	46.37	65.79	54.55	18.43	29.66	48.75	28.31	46.47	0.00	13.79	36.36	0.00	11.43	4.08	9.76
26	UAIC1860	41.17	62.33	42.97	11.16	21.01	36.41	22.12	38.78	7.60	11.43	17.39	2.90	5.56	4.26	9.76
27	UNTLing	39.11	62.57	36.74	7.78	11.82	32.65	5.29	40.48	2.86	17.65	4.35	0.00	0.00	0.00	0.00
28	HunAlize	37.10	58.59	15.82	2.09	23.81	31.76	11.83	29.95	7.84	4.55	6.45	8.00	0.00	0.00	0.00
29	Transformers	26.54	47.55	24.06	2.86	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	Baseline	25.20	46.48	0.00	19.26	14.42	29.14	3.68	6.20	11.56	0.00	0.00	0.00	0.00	0.00	0.00
31	Entropy	20.39	37.74	15.49	5.83	6.39	12.81	6.32	4.95	7.41	0.00	3.92	2.27	0.00	6.78	0.00
32	IJSE8	19.72	38.07	14.70	4.92	8.23	15.47	7.07	8.57	2.27	0.00	0.00	0.00	0.00	0.00	0.00

Table 6: **Technique classification F_1 performance on the test set.** The systems are ordered on the basis of the final ranking. Columns **1** to **14** show the performance for each of the propaganda techniques (cf. Section 2). The best score for each technique appears highlighted. (*Note: We found a bug in the evaluation script after the end of the competition. The correct ranking, shown in Appendix B, does not differ substantially from above.*)