Characterizing Latent Perspectives of Media Houses Towards Public Figures

Anonymous ACL submission

Abstract

Media houses reporting on public figures, often come with their own biases stemming from their respective worldviews. A characterization of these underlying patterns helps us in better understanding and interpreting news stories. For this, we need diverse or subjective summarizations, which may not be amenable for classifying into predefined class labels. This work proposes a zero-shot approach for non-extractive or generative characterizations of person entities from a corpus using GPT-2. We use well-articulated articles from several well-known news media houses as a corpus to build a sound argument for this approach. First, we fine-tune a GPT-2 pre-trained language model with a corpus where specific person entities are characterized. Second, we further fine-tune this with demonstrations of person entity characterizations, created from a corpus of programmatically constructed characterizations. This twice fine-tuned model is primed with manual prompts consisting of entity names that were not previously encountered in the second fine-tuning, to generate a simple sentence about the entity. The results were encouraging, when compared against actual characterizations from the corpus.

1 Introduction

Media houses have their own worldview with which they interpret happenings around the world, that may show up as biases in their characterizations of entities like persons, organizations or countries. Such biases are often implicit and benign. However, in order to get better clarity and understanding of news, it is important to explicate and understand how specific media houses characterize specific entities.

Automated approaches for entity characterizations have gained significant interest in recent years (Wei et al., 2019; Liu et al., 2020, 2021). Most of the current approaches are extractive in nature, that look for specific features like frequency, diversity, informativeness, etc. in the descriptions of entities to extract sentences characterizing them.

Given the vast numbers of entities and issues that media houses report on, it is impractical to create a pre-determined set of classes onto which, characterizations can be classified. We define entity characterization as a terse or one statement description of the individual quality of a person or a thing. In contrast to summarization, entity characterization need cover all pertinent characteristics of the entity in a short summary. Entity characterization is subjective– that reveals the biases of the inquirer, whereas summarization is meant to be objective, and verifiable against the actual characteristics of the object of inquiry.

In this work, we propose an approach for *non-extractive* or *generative* characterizations for person entities. These are in the form of one-sentence descriptions that are obtained from suitably fine-tuning pre-trained masked language models.

Pre-trained masked language models are known to perform well on diverse NLP tasks in a zero-shot setting (Radford et al., 2019). A number of recent approaches (Schick and Schütze, 2021b; Schick et al., 2020; Hambardzumyan et al., 2021; Gao et al., 2021; Petroni et al., 2019) show that domain-adapted language models have substantial knowledge of the domain, and with pattern demonstrations to solve a NLP task, the model performs well in the intended task. Our approach is along the same lines to perform Generative Entity Characterization by fine-tuning with demonstrations.

Our approach is to fine-tune the GPT-2 pre-trained language model twice, and use this model to generate characterizations. The first fine-tuning is for domain adaptation with a corpus of person entity mentions disambiguated for co-references. For the second fine-tuning we perform *entity characterization demonstrations*, based on sentences characterizing the entity in question. These sentences are programmatically constructed from the corpus by extracting clauses and their parts. Subject, Verb, Object, and Adverbials are the common parts of clauses that are extracted. We construct a demonstration pattern to convert parts of clauses into semantically coherent simple sentences describing the entity. The pattern is to suffix the subject with *"is described as"*, convert lemmatized verb into a gerund, and append other parts grammatically. With this pattern, a corpus of simple sentences about entities is constructed. Demonstration sentences for ten entities are then separated from this corpus to be used for testing, and are not included the demonstrations training.

The twice fine-tuned model is then prompted with test entities suffixed with four different manual prompts, and the generated texts were inspected for characterizations of entities. Since the test entity sentences were not used in demonstrations, we attribute the generated text to zero-shot generations.

2 Related Work

In the recent past, text classification with language models and pattern training has shown promising results on key datasets. Schick and Schütze (Schick and Schütze, 2021a), show that language models understand text classification task by converting input to *cloze question* patterns and training. GPT-3 with hundreds of billions of parameters shows remarkable few-shot performance on SuperGLUE. Schick and Schütze (Schick and Schütze, 2021b) show that an equivalent few-shot performance can be achieved by training small language model ALBERT with *cloze question* patterns.

In the text classification task, mapping predicted tokens to predefined labels is challenging and requires domain expertise even though training with patterns optimizes text classification. Schick et al. (Schick et al., 2020), show an approach to automatically map the predicted tokens to labels. Training language models with patterns have shown adequate performance in the text classification task. In this work, we propose a similar approach with manual prompts patterns to generate non-extractive information about person entities from a corpus.

Choosing prompts and equivalent words of classification labels manually or algorithmically are challenging since there are significant variations. Hambardzumyan et al. (Hambardzumyan et al., 2021) show an approach to finding these as embeddings in a continuous embedding space of word embeddings. Trainable embeddings are added around the input to make the masked language model predict the masked token and evaluated on natural language understanding tasks of GLUE Benchmark.

With natural language prompts and a few demonstrations on GPT-3, awe-inspiring performance on language understanding tasks is observed. However, since GPT-3 has 175B parameters, it is challenging to use in real-world applications. Gao et al. (Gao et al., 2021) show prompt-based fine-tuning with demonstrations on moderately small language models BERT and RoBERTa. In this work, we have fine-tuned with person entity characterizing sentences as demonstrations.

Mining commonsense knowledge is an important natural language processing task. Language models are known to have this, Davison et al. (Davison et al., 2019) show an approach to mine commonsense knowledge from Pre-trained Language Modes. A uni-directional model generates sentences with a specific template for each type of relation in information triples. This generated sentence is validated by masking and predicting the tokens using a bi-directional language model.

Apart from linguistic knowledge, language models might also contain relational knowledge in the training data. Petroni et al. (Petroni et al., 2019) analyze relational knowledge in state-of-the-art pre-trained language models with LAMA (LAnguage Model Analysis) probe a corpus of facts in subject-relation-object triples or question-answer pairs forms derived from diverse factual and commonsense knowledge sources. Kassner and Schütze (Kassner and Schütze, 2020), show that the ability of pre-trained language models to learn factual knowledge is not as good as humans learn by probing for facts with Negated LAMA and Misprimed LAMA. Ideally, these probe variants should result in contradictions, whereas it was not so, suggesting that factual knowledge extraction is based on pattern matching rather than inference. Jiang et al. (Jiang et al., 2020) study factual knowledge in multilingual language models with manually created probes in 23 languages similar to LAMA.

Kumar and Talukdar (Kumar and Talukdar, 2021), show that the order of training examples significantly reduces the samples required for few-shot learning on Sentiment Classification, NLI, and Fact Retrieval tasks.

Nishida et al. (Nishida et al., 2020) shows an approach where the pre-trained BERT is adapted to the target domain and next fine-tuned with RC task on a source domain. Finally, this model performs RC tasks in the target domain. Domain adaptation is crucial to solving any task related to that domain.Gururangan et al. (Gururangan et al., 2020) show that even pre-trained language models of hundreds of millions of parameters are ineffective to encode the nuances of a given textual domain.

The state-of-the-art of using pre-trained language models to solve an NLP task show domain adaptation and fine-tuning with demonstrations of patterns as the most plausible approach to a reasonable extent. In this work, we propose an approach to characterize entities along similar lines.

3 GPT-2 Domain Adaptation

A GPT-2 Pre-trained Language Model (PLM), with 345M parameters, was fine-tuned with steps from GitHub.[1] PLM was fine-tuned individually on four popular news media corpora. Due to limitations in the available compute instance, 345M PLM, medium model was fine-tuned and this model proved sufficient to get convincing results. Domain adaptation or fine-tuning PLM on domain corpora is a prerequisite before task-specific training. The domain-adapted PLM was further fine-tuned with programmatically constructed demonstration sentences.

3.1 Textual Media Source

The GDELT Project[2] records the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages. From GDELT database, textual news media article URLs of four popular media houses, between years 2015 to 2021, were extracted and texts of articles were scraped for domain adaptation. Table 1 shows the details of each media house corpus.

Table 1: Scraped Media House Articles, 2015 to 2021

Media house	No. of Articles	Size on Disk
Media House A	40, 514	282M
Media House B	53, 024	364M
Media House C	31, 029	298.6M
Media House D	27, 044	171M

[1]GPT-2 Fine-tuning: https://github.com/openai/gpt-2/

[2]GDELT Project: https://www.gdeltproject.org/

4 Person Entity Characterization with Manual Prefix Prompts

Cloze and *Prefix* prompts are two types of prompts used as inputs for a language model to solve NLP tasks. Cloze prompts as in (Petroni et al., 2019) is where the token to be predicted is masked and the model predicts. The Prefix prompts (Li and Liang, 2021; Lester et al., 2021) or the prompts used for priming when used as input to language model generates a conditional sequence text auto-regressively.

Priming in this work can be attributed to *"programming in natural language"* detailed by Reynolds and McDonell (Reynolds and McDonell, 2021). This work attempts to prompt language model to generate characteristics of a person entity with prompts ubiquitous in spoken and written English language. The concept is when you want to describe a person, one would express beginning with *"John is described as ..."* or a semantically similar prefix, in most contexts. These prefixes and synonymous ones are very common in any corpora used to train the language models and priming with natural language phrases like *"John is described as ..."* would constrain the entailment to something about *John*. The intuition is prime the language model in *"ubiquitous or natural language way."* Since these demonstrations are not very frequent in the corpus we construct a corpus of these type of sentences to fine-tune. To test this hypothesis following steps were followed with each Media House corpus and depicted in **Figure 1**.

Block 1: Person Entity Mention Disambiguation in Articles
- Co-reference Replacement[3]
- Replace short names with full name

Block 2: First fine-tuning, GPT-2 PLM (345M) if fine-tuned with the **Block 1** processed disambiguated articles corpus and named as **FT1 Checkpoint**

Block 3: Extract clauses and their parts from sentences of person entities using *spacy-clausie*[4] from **Block 1** disambiguated articles corpus

Block 4: With parts of clauses (**Block 3**) convert lemmatized verb of clauses to a gerund and construct a corpus of simple entity characterization demonstration sentences in the following pattern: "<*Person_Entity_Name*> **'is described as'** <*gerund*> <*grammatically valid combination of parts of clause*>" From this corpus of sentences, sentences of ten entities with high frequencies in different ranges set aside as *Test Corpus* and rest as *Demonstrations or Training Corpus*

Block 5: With the *Demonstrations Corpus* (**Block 4**), **FT1 Checkpoint** was fine-tuned and named as **FT2 Checkpoint**

Block 6: **FT2 Checkpoint** was used to generate sentences of entities in *Test Corpus* with prompts defined in **Table 2**

Block 7: Sentences generated about entities in **Block 7** were tested for *non-extractive characterization* against FT1 and FT2 corpus sentences using Semantic Textual Similarity [5] and Sentiment Analysis

[3]https://github.com/NeuroSYS-pl/coreference-resolution

[4] https://github.com/mmxgn/spacy-clausie

[5]STS: https://www.sbert.net/docs/usage/semantic_textual_similarity.html

Following subsections detail each of the above steps.

Table 2: Four types of *prefix prompts* used to generate sentences about entities

"<*Person_Entity_Name*> is described as being"
"<*Person_Entity_Name*> is described as having characteristics"
"<*Person_Entity_Name*> is described as performing"
"<*Person_Entity_Name*> is described as stating"

4.1 Person Entity Mention Disambiguation

Co-reference resolution improves the accuracy of NLP tasks like machine translation, sentiment analysis, paraphrase detection and summarization (Sukthanker et al.) (Sukthanker et al., 2020). We have disambiguated person entity mentions in the articles to ensure that every person entity sentence has full name of the entity.

The first pre-processing was replacing entity co-references with the actual entity name and on this output replace partial name references with full name to finally get a processed document with full name of the entity in maximum number of sentences in each news article.

NeuroSYS coreference-resolution[6] proposes three intersection strategies or ensemble methods of AllenNLP and Huggingface coreference models outputs. The methods are *strict* where clusters identical in both the models are considered, *partial* where spans identical in both model outputs are considered and *fuzzy* where spans and overlapping spans are considered from both the models. In this work we leveraged the *fuzzy* ensemble and processed the raw articles.

The objective of this work was to generate single concise sentences of person entity characterizations. To align with this objective the sentences in each media house articles were processed to contain unambiguous entity mentions. To address this, the co-references replaced texts were processed to replace partial name references with full name of the entity so that every sentence has full qualified mention of the entity and information about the entity. To achieve this, we followed a logic of processing one article at a time, mapping partial names, either first name or last name, with the full name by comparing tokens. The intuition is that entity is referred with full name in the initial parts of the article and in later sentences of the article either first name or last name is used to refer to the entity. The partial name should be either first name or last name of the entity in the previously used longer name.

This final corpus of articles with full entity disambiguation was used for first fine tuning or FT1. For all the media houses the loss plateaued around 0.6 and hence a checkpoint around this loss was considered for next fine tuning FT2.

4.2 Characterization Sentences Corpus of Person Entities

To generate simple concise demonstration sentences of entity characterizations, the FT1 Checkpoint was fine-tuned with manual prompt prefixed to clauses of entities. Clauses contain the main information of entities. Corpus of simple sentences of anything said, done or events related to the entities was constructed using clauses and their parts extracted from each article using ClauCy[7] (Del Corro and Gemulla, 2013). Clauses and their parts were extracted from each sentence in articles. The parts of the clauses are: *Type, Subject, Verb, Indirect_Object, Direct_Object, Complement and Adverbials*. There are ten Clause Types with combination of parts: *SVC, SVOO, SVOC, SVO, SVOA, SVO, SVO, SV, SVA and SVO*. Every clause has a subject and verb, other parts vary depending on the input sentence. Entities and the sentences they appear were mapped and maps with more than 500 sentences were considered for FT2 corpus. Table 3 shows the details of FT2 sentences corpus for each media house.

FT2 sentences were constructed by suffixing Subject with "is described as", converting Verb in to Gerund form and grammatically joining other parts of the clause to form a complete readable sentence. Gerund or present participle is the adjective form the verb (like showing, saying, claiming, winning, etc.) and functions as attributing the other parts of the clause (Object, Complements and Adverbials) to the Subject. Ten subjects or person entities with highest count in different ranges were separated as test corpus and rest of entity sentences for second fine tuning. This was done to ensure testing with entity count in broad ranges. The checkpoint from FT1 was further fine tuned with FT2 corpus. For all the media houses, the second fine

[6] 3

[7] 4

Clause Type	Media House 1	Media House 2	Media House 3	Media House 4
SV	11,349	3,244	17,863	7,768
SVA	2,732	695	3,829	1,698
SVC	26,042	10,403	40,899	16,617
SVO	23,522	8,750	34,164	15,795
SVOA	1,223	488	1,915	937
SVOC	2,832	1,249	4,619	1,857
SVOO	597	246	738	370
FT2 Dataset Sentences Count	68,297	25,075	1,04,027	45,042
Unique Person Entities Count	117	69	140	83

Table 3: Each Media House FT2 Sentences Corpus details. Count of each extracted clause type, total number of sentences and unique person entities in each corpus

True Positive (TP)
Definition: ***Novel and Meaningful or Non-extractive Characterization.*** The generated sentence has a high semantically matching sentence in FT1 or FT2 datasets, and the person entity in both sentence contexts are the same. **Condition:** Prompt Entity == Ground Truth Entity And The cosine score between generated sentence and FT1 or FT2 dataset sentence is >= 0.6
Type 2 Error or False Positive (FP)
Definition: The generated sentence has a high semantically matching sentence in FT1 or FT2 datasets and the person entity in both sentence contexts are different. **Condition:** Prompt Entity != Ground Truth Entity And The cosine score between generated sentence and FT1 or FT2 dataset sentence is >= 0.6
Type 1 Error or False Negative (FN)
Definition: The generated sentence has a low semantically matching sentence in FT1 or FT2 datasets, and the person entity in both sentence contexts are the same. **Condition:** Prompt Entity == Ground Truth Entity And The cosine score between generated sentence and FT1 or FT2 dataset sentence is <0.6
True Negative (TN)
Definition: The generated sentence has a low semantically matching sentence in FT1 or FT2 datasets and the person entity in both sentence contexts are different. **Condition:** Prompt Entity != Ground Truth Entity And The cosine score between generated sentence and FT1 or FT2 dataset sentence is <0.6

Table 4: Evaluation Criteria

tuning plateaued around loss of 0.1 and hence fine tuning was stopped when loss reached below 0.1.

4.3 Generative Entity Characterization

Widely prevalent manual prompts in the spoken and written language used to talk about a person were chosen to prime the language model. Sentences were generated with the FT2 Checkpoint. The second fine-tuning, FT2, was with a corpus of sentences with “is described as” prompt. The results of the generated sentences with this prompt were not convincing, so we experimented with semantic alternative prompts shown in Table 2. With these prompts, we observed entity characterizing generated sentences. Ideally, all the test sentences should be generated; hence, sentences were generated to each entity’s count in the test corpus. Novel combinations of information in the corpus or summarized opinions of test entities were expected in the generated texts. The generated texts were compared for Semantic Textual Similarity (STS) with FT1 and FT2 corpus sentences using Sentence Transformers[8]. Since language models are probabilistic and generate novel sentences, we chose cosine similarity of greater than or equal to 0.6 as a positive result.

To the best of our knowledge, there is no start-of-the-art corpus for Entity Characterization demonstrations and evaluation criteria. For this purpose, we have compiled FT2 dataset and defined evaluation criteria with Confusion Matrix as shown in Table 4. The following section details the results of entity characterizations generated with prefix prompts in Table 2

[8] 5

5 Results

With the FT2 checkpoint of each media house, sentences were generated for ten test entities, with four prompts shown in Table 2, to test the hypothesis. The count of generated sentences was up to the entity sentences count in the FT2 sentences corpus. The length of the generated text was limited to 30, and the first sentence in the generated text was considered for evaluation. The first evaluation was with the FT2 sentences corpus. Entity names in the FT2 sentences corpus was masked and embeddings were constructed. Then, each generated sentence matched with all sentence embeddings. Masking entity names in FT2 corpus resulted in better relevant matches. The match with the highest cosine score was considered the best semantic match. Next, a similar evaluation was done with the FT1 articles corpus. Every sentence was extracted from each article of the FT1 corpus, and sentences with person entities and lengths greater than ten were considered to compare with the generated text to consider sentences with reasonable information content and to exclude insignificant sentences. In this evaluation entity names were not masked in the FT1 corpus sentences.

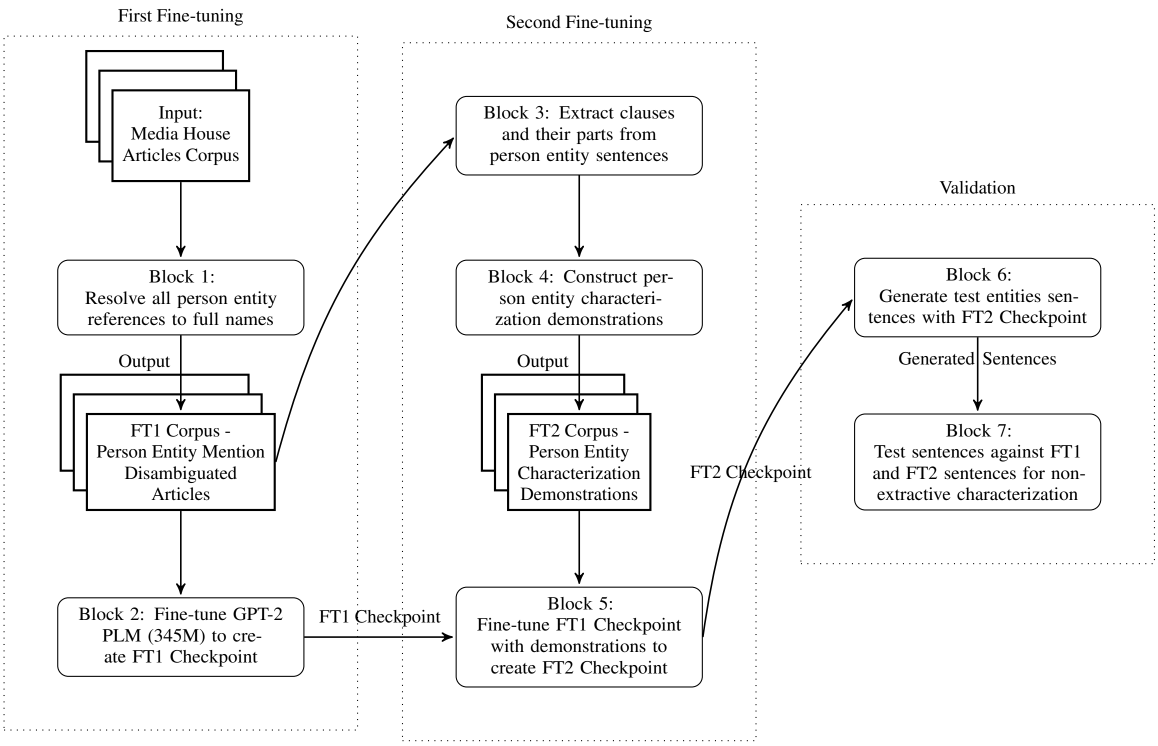

Figure 1: Pipeline of processing a Media House Corpus, generating sentences about entities and validating for characterizations. *Block 1* uses NeuroSYS[9]. *Block 3* uses Claucy[10]. Details of FT2 or Demonstrations Corpus is shown in Table 3. In *Block 6* sentences about test entities are generated with prompts listed in Table 2. *Block 7* uses Semantic Text Similarity (STS)[11] to compare generated sentence with corpus sentences. Examples of generated and semantically similar corpus sentences are shown in Table 6

Manual Prompts	Distinct Generated Sentences Count	% of Distinct Semantic Matches		Average Sentiment Scores Differences of True Positives (TP)		F1 Score		Precision		Recall	
		FT1	FT2	FT1	FT2	FT1	FT2	FT1	FT2	FT1	FT2
						Media House 1					
is described as having characteristics	3010	41%	28%	0.157	0.108	**0.864**	0.54	**0.888**	0.504	**0.842**	0.981
is described as being	7347	50%	34%	0.154	0.136	**0.898**	0.497	**0.852**	0.414	**0.948**	0.973
is described as performing	4243	28%	20%	0.034	0.027	0.725	0.317	0.607	0.294	0.899	0.968
is described as stating	8901	65%	44%	0.138	0.097	0.807	0.406	0.726	0.363	0.907	0.935
						Media House 2					
is described as having characteristics	4407	27%	17%	0.062	0.054	**0.910**	0.55	**0.892**	0.622	**0.929**	0.492
is described as being	4985	51%	32%	0.155	0.139	**0.894**	0.54	**0.837**	0.519	**0.960**	0.563
is described as stating	5794	67%	41%	0.126	0.099	0.825	0.469	0.734	0.476	0.942	0.461
is described as performing	2506	30%	20%	0.023	0.016	0.557	0.263	0.404	0.184	0.898	0.467
						Media House 3					
is described as having characteristics	5418	22%	20%	0.102	0.079	**0.953**	**0.743**	0.945	**0.682**	**0.960**	**0.816**
is described as being	9591	47%	59%	0.177	0.142	**0.921**	0.597	**0.889**	0.525	**0.954**	0.692
is described as performing	6430	35%	39%	0.064	0.039	0.869	0.576	0.828	0.517	0.915	0.650
is described as stating	11222	59%	30%	0.150	0.117	0.844	0.515	0.767	0.465	0.940	0.576
						Media House 4					
is described as having characteristics	177	42%	23%	0.024	0.038	0.789	**0.824**	0.679	**0.860**	0.942	**0.791**
is described as performing	4478	29%	20%	0.025	0.011	0.754	0.622	0.660	0.638	0.879	0.607
is described as being	5375	48%	32%	0.156	0.110	**0.903**	0.574	**0.874**	0.548	**0.934**	0.601
is described as stating	6420	60%	39%	0.139	0.090	0.837	0.464	0.789	0.476	0.892	0.452

Table 5: Metrics based on evaluation criteria in **Table 4** of FT2 checkpoint generated sentences with FT1 and FT2 corpus sentences

Examples of Novel and Meaningful or Non-extractive Person Entity Characterizations (True Positives)	
Generated Text	**Corpus Text**
Media House 1 - FT1	
Entity A is described as having characteristics that can end up forming the government in State.	As per sources, Entity A is tipped to be the next Chief of Elected Memebers of State.
Entity B is described as having characteristics like threatening, stoking violence, etc.	Entity B comments come after he was likened to a terrorist by a prominent leader.
Entity C is described as having characteristics of a caring truly, a loving truly, and a pious truly	Entity C was a great leader with a great sense of compassion and humour.
Entity D is described as having the characteristics of an Angel.	A prominent chronicles of a Powerful person, a character that bears an uncanny resemblance to Entity D.
Entity E is described as being critical of the Prominent Party government in state.	Leader Entity E had remained highly critical of the Prominent Party government in the past.
Entity D is described as being a strong advocate for the interests of the people.	Listing out the various pro-people initiatives launched by Entity D, a Rebel leader said Entity D was a popular leader who always remained in the hearts of the people.
Entity F is described as being an extremely beautiful face	The smoky eyes and nude lips further complimented Entity F's look.
Entity G is described as being very quick in taking the decision, in such a situation.	Entity G, however, is the first politician from the ruler's family to have reacted to the step.
Media House 2 - FT1	
Entity H is described as being under house arrest, at his residence.	Does this mean party head and elected member Entity H is under house arrest?
Entity I is described as being the new Chief of State.	"People of State want Entity I to become the Chief of State," he added.
Entity H is described as being unwell.	"We have heard that Entity H is unwell, which is understandable"
Entity J is described as being mature.	Another significant development is that Entity J has emerged as a matured leader during the General elections.
Entity J is described as having the characteristics like a true leader and a person who has a capacity to thought.	He said that Entity J has a good vision and thoughts.
Entity K is described as having characteristics such as being able to bring people on the path of development.	"Through better economic management, we could take the common man ahead on the path of progress," Entity K claimed.
Entity L is described as having characteristics like a seasoned politician and leader and an ideal organisational person.	Entity L is an extremely qualified & respected leader, Entity L has served this nation with dedication & humility.
Entity M is described as having characteristics such as reconciling to the family, developing friendships that helped him during the difficult times, honesty and integrity in discharge of his duties as an actor.	She said, "actor Entity M has really had my back, and has been there for me as a friend and support over the years, unfailingly and intuitively."
Entity N is described as being no entry, in the roadshow.	Entity N said that he was restricted only to his region as he does not hold any official post in city unit.
Entity M is described as being an awareness campaign to urge people to follow.	During this time, Entity M has appeared in several public safety videos, urging his fans to obey laws.
Media House 3 - FT1	
Entity D is described as having characteristics of a strong personality.	On one side, you see in Entity D a woman who was the personification of authoritarianism.
Entity O is described as having characteristics of a classic leader born to influential parents.	With a massive campaign focused on Entity O's personality, he has towered over other stalwarts in State politics, including a Top Leader and his father's father.
Entity J is described as being to become the President.	Entity J finally looks all set to become President.
Entity P is described as being the primary link between the party and the people	"Entity P is the unifying factor for party," the party affairs representative told in an interview.
Media House 3 - FT2	
Entity P is described as having characteristics of a leader who has a habit of wearing her aspirational state's uniforms.	Entity P is described as coming in her uniform.
Entity E is described as having characteristics of a leader who may be able to win City elections.	Entity E is described as claiming he built his from the ground up by addressing dozens of rallies in State's villages and towns, before converging in City.
Entity Q is described as having characteristics of a successful orator.	Entity Q is described as making that comment , in his personal capacity.
Entity R is described as having characteristics of a leader who may need to rein in elements on the ground.	Entity R is described as saying that he will take all efforts to help authorities contain the spread of the disease.
Media House 4 - FT1	
Entity B is described as being in State, for a two-day visit to State.	Entity B is on a two-day visit to State.
Entity S is described as being active, on social media.	Entity S is an avid social media player and also a writes a blog regularly.
Entity T is described as being the new go-to girl.	New 'Country Girl' Entity T is making a lot of headlines these days.
Entity U is described as being in no mood to waste time.	"I do not waste my time on what he says," said the leader Entity U.
Media House 4 - FT2	
Entity J is described as having characteristics of a revolutionary.	Entity J is described as showing hiss mettle.
Entity V is described as having characteristics of an artiste.	Entity V is described as winning several accolades for his work, including the Country Award for his debut role as a child artist.
Entity W is described as having characteristics of a leader.	Entity W is described as charting his future course of action.

Table 6: True Positive examples of top metrics in **Table 5**

Test Entity	Examples of Generated Characterizations Across Media Houses
Entity 1	MH1: *is described as having characteristics* of an immature, perhaps naive, leader MH1: *is described as having characteristics* of an immature, perhaps anti-national, protestor
	MH2: *is described as having characteristics* like a true patriot MH2: *is described as having characteristics* like a true leader and a man to trust
	MH3: *is described as having characteristics* of a classic Party loyalist MH3: *is described as having characteristics* of a leader who is adept at top command
	MH4: *is described as being* at loggerheads with the Party leadership MH4: *is described as being* fit, also, to be a prime minister
Entity 2	MH1: *is described as having characteristics* of a strong woman MH1: *is described as having characteristics* of a strong political personality
	MH2: *is described as having characteristics* such as long history with the State and its unique culture and languages MH2: *is described as having characteristics* like a person, strong willpower, and political instincts
	MH3: *is described as having characteristics* of a classic leader MH3: *is described as having characteristics* of a strong regional leader
Entity 3	MH3: *is described as having characteristics* of a leader who is adept at stoking passions through the Party's various programs MH3: *is described as having characteristics* like a leader with firm control over the party, a decisive figure, and an ability to move the front
	MH4: *is described as being* successful in expanding the Party MH4: *is described as being* a "prominent face" of the Party

Table 7: Examples of Entity Characterizations across media houses

We define the evaluation criteria as detailed in Table 4. The evaluation approach is that if a generated sentence is semantically similar to an FT1 or FT2 sentence, the entity referred to is the same. Then the generated sentence should be about the entity. In FT2, we have processed sentences where something said, done, and about an event related to the entity is suffixed with entity name and "is described as" and we refer to these sentences as the entity characterizing sentences. The FT2 generated sentences were the same kind as in the FT2 corpus. Examples in Table 6. Hence we define the FT2 generated sentences as characterizations and validate the characterizations with the Confusion Matrix definitions in Table 4. Also, good metrics on either FT1 or FT2 dataset is good enough to conclude soundness of the approach.

Table 5 shows the metrics derived from the evaluation criteria. F1, Precision and Recall are computed based on the *Distinct Generated Sentences Count*. was shown, the *"is described as having characteristics"* and *"is described as being"* prompts resulted in good F1, Precision, and Recall (or True Positive Rate) scores across media houses, which is Confirming that FT2 would lead to generating the most relevant sentences to the entity. More than one generated sentence is semantically similar to a corpus sentence. For True Positives average of the difference in sentiment scores of generated and semantically matching sentence is marginal. Therefore, it is encouraging to conclude that FT2 generated sentences are about the prompted entities and characterizing the entities with sentiment in the corpus. An exhaustive examples of generated True Positive and corresponding semantically matching sentences of top metrics in Table 5 is shown in Table 6.

With the approach, evaluation criteria, and test prompts detailed in this work, the *"is described as having characteristics"* and *"is described as being"* manual prompts function reasonably well as prompts to generate non-extractive characterizations of entities, as is evident from the examples. Examples of top characterizations of three test entities appearing across media houses are shown in Table 7 to contrast the characterizations generated by each media house. Generated characterizations have a cosine similarity score of greater than 0.75 with the FT1 corpus sentences. It is evident that top characterizations differ distinctly across media houses for the entities.

6 Conclusion

There are diverse perspectives about a person entity we know and even more with famous personalities. Media House discourses are diverse and impact the World Views of famous personalities. In today's world of the Information Age, getting insights into these World Views will lead to faster and better awareness. In this work, we propose an approach to derive common perceptions in a Zero-shot way. The evaluation criteria and metrics show a good performance of the approach.

7 Ethics Compliance

The data collected for this work is from GDELT Project[12]. GDELT Project monitors the world's

[12] 2

broadcast, print, and web news from nearly every corner of every country in over 100 languages. For this work, we have retrieved all news articles from four well-known media houses between 2015 and 2021 without any specific filters in selecting articles or content. Our goal was to collect all publicly available news articles on the internet without bias. The approach in this work works well when a public figure or person entity mention frequency is high, and there are no specific methods to distinguish person entities. Results shared are based on public figures or person entities with high frequency, and there are no custom implementations to showcase any specific or group of person entities. This work shows unbiased and fair latent perspectives of public figures or person entities as per the publicly available new articles.

8 Limitations

This work aimed to extract latent perspectives of public figures or person entities in non-extractive or not based on the frequency of keywords. To prove the approach, we have applied the approach to each media house individually to know the latent perspectives of public figures from each media house. There are opportunities to combine media house corpora in alternate ways and check the results. The scope of this work is limited to extracting latent perspectives of public figures of each media house.

References

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. 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 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4921–4933, Online. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-factr: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818.

Sawan Kumar and Partha P. Talukdar. 2021. Reordering examples helps during priming-based few-shot learning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4507–4518. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Qingxia Liu, Gong Cheng, Kalpa Gunaratna, and Yuzhong Qu. 2021. Entity summarization: State of the art and future challenges. *Journal of Web Semantics*, 69:100647.

Qingxia Liu, Gong Cheng, and Yuzhong Qu. 2020. Deeplens: Deep learning for entity summarization. *arXiv preprint arXiv:2003.03736*.

Kosuke Nishida, Kyosuke Nishida, Itsumi Saito, Hisako Asano, and Junji Tomita. 2020. Unsupervised domain adaptation of language models for reading comprehension. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5392–5399.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Dongjun Wei, Yaxin Liu, Fuqing Zhu, Liangjun Zang, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Esa: entity summarization with attention. *arXiv preprint arXiv:1905.10625*.