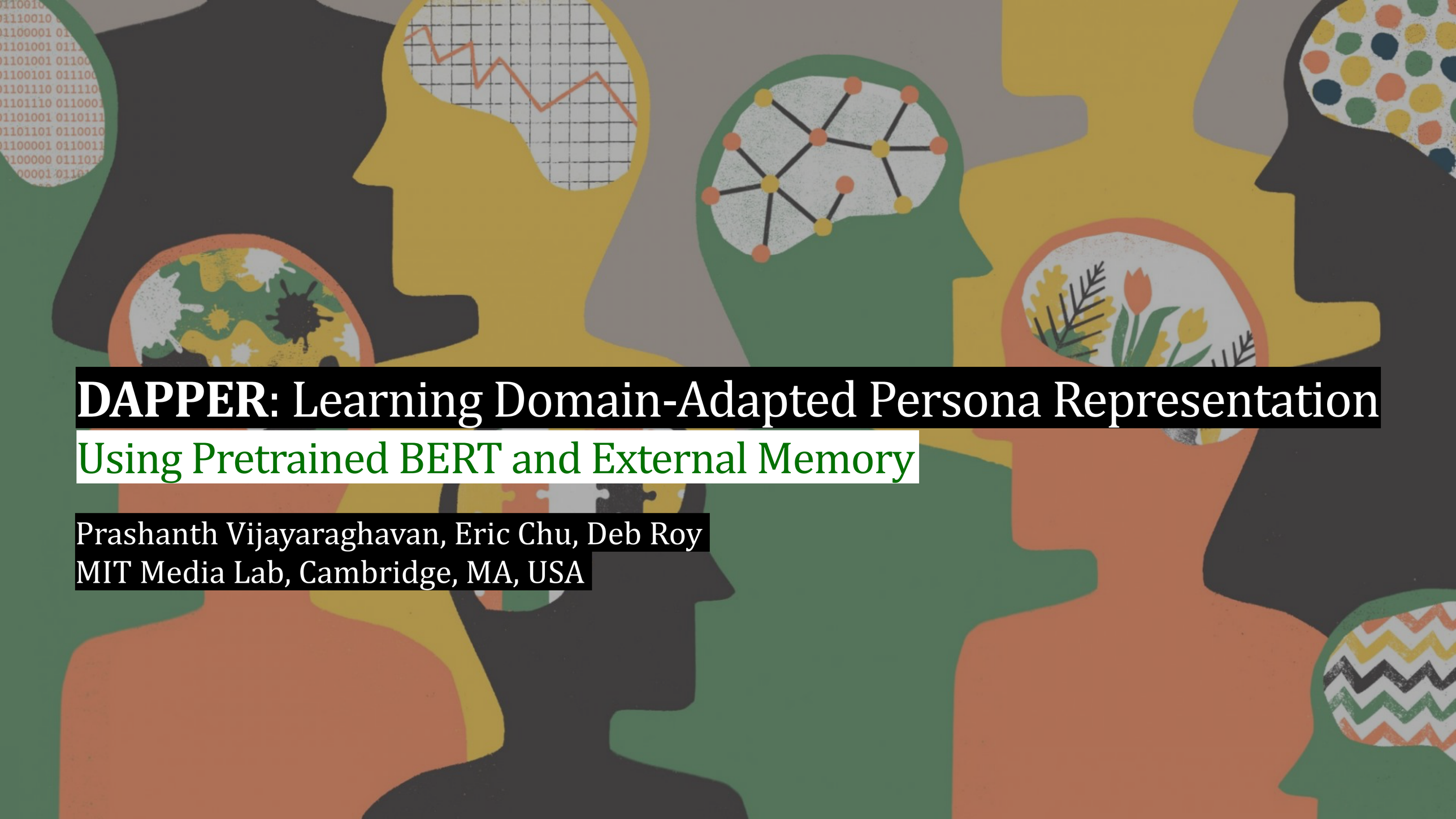
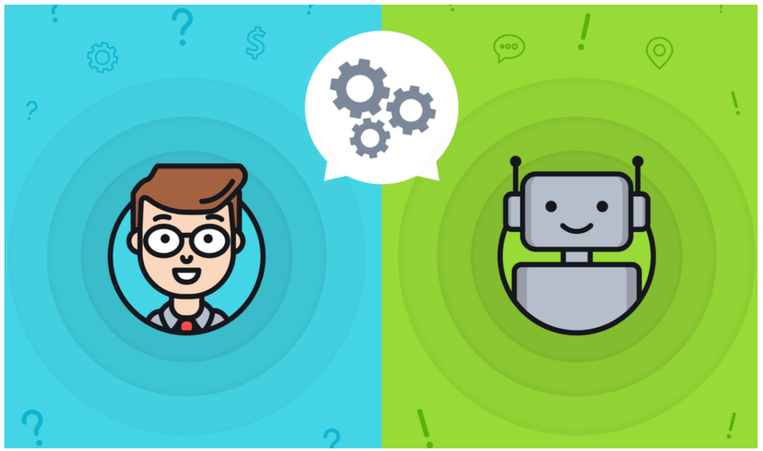


DAPPER: Learning Domain-Adapted Persona Representation

Using Pretrained BERT and External Memory

Prashanth Vijayaraghavan, Eric Chu, Deb Roy
MIT Media Lab, Cambridge, MA, USA

Background



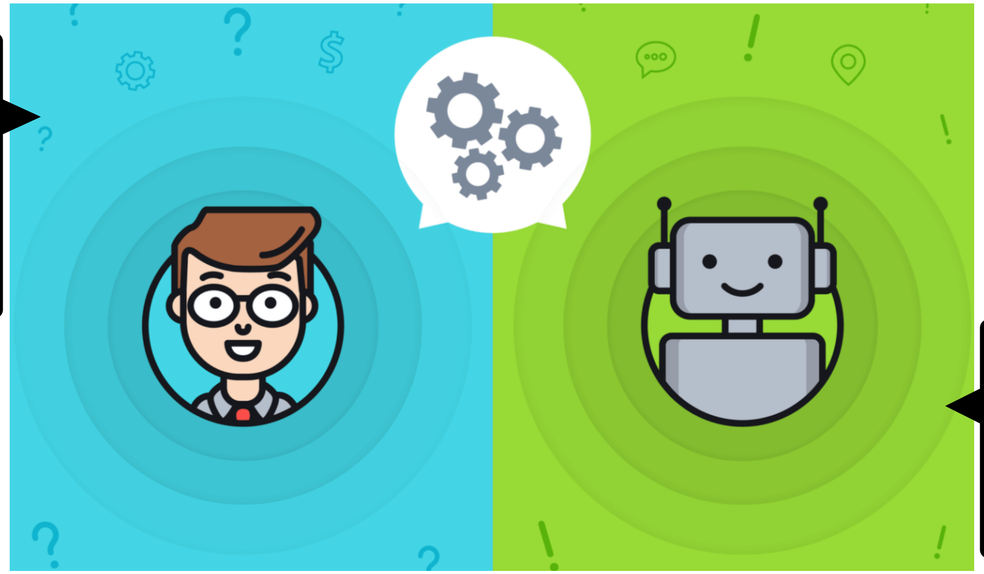
Conversational Agents

Social Media Behavior Mining

Narrative Intelligence

What is Persona?

Persona profile as character descriptions containing their interests, likes and dislikes.
— Zhang et. al (2018)



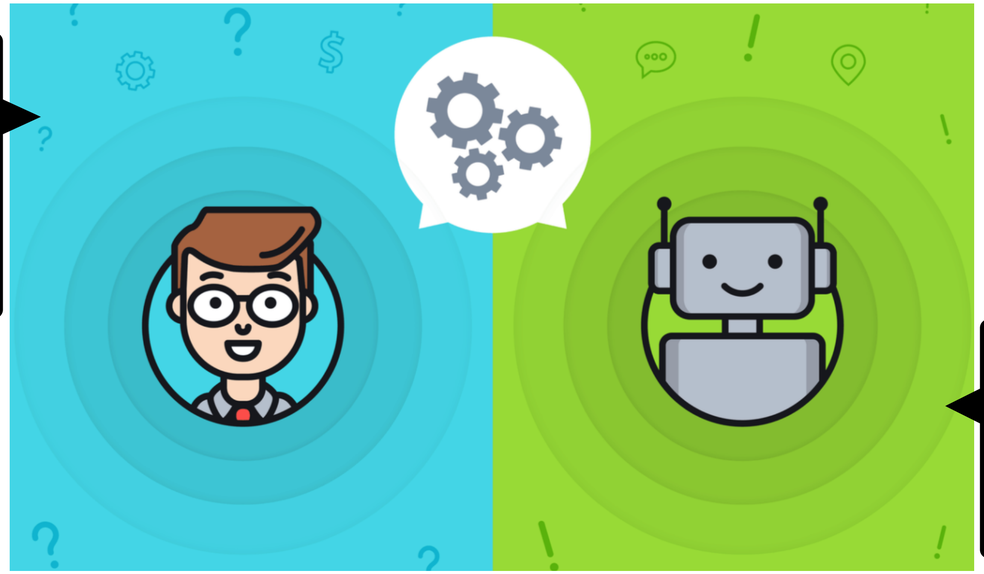
Persona as background Information, Speaker Style
— Li et. al (2016)

Conversational Agents

1. Zhang, Saizheng, et al. "Personalizing dialogue agents: I have a dog, do you have pets too?." *arXiv preprint arXiv:1801.07243* (2018).
2. Li, Jiwei, et al. "A persona-based neural conversation model." *arXiv preprint arXiv:1603.06155* (2016).

What is Persona?

Persona profile as character descriptions containing their interests, likes and dislikes.
— Zhang et. al (2018)



Persona as background Information, Speaker Style
— Li et. al (2016)

Conversational Agents

1. Zhang, Saizheng, et al. "Personalizing dialogue agents: I have a dog, do you have pets too?." *arXiv preprint arXiv:1801.07243* (2018).
2. Li, Jiwei, et al. "A persona-based neural conversation model." *arXiv preprint arXiv:1603.06155* (2016).

What is Persona?

Persona defined by 16 categories based on Myers-Briggs Type Indicator model (MBTI) — Gjurković et. al (2018)

Persona based on Big-Five Personality Traits Model — Golbeck et. al (2011)

Social Media Behavior Mining

1. Gjurković, Matej, and Jan Šnajder. "Reddit: A gold mine for personality prediction." Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. 2018. Azucar, Danny, Davide Marengo, and Michele Settanni. "
2. Golbeck, Jennifer, Cristina Robles, and Karen Turner. "Predicting personality with social media." *CHI'11 extended abstracts on human factors in computing systems*. 2011. 253-262.

What is Persona?

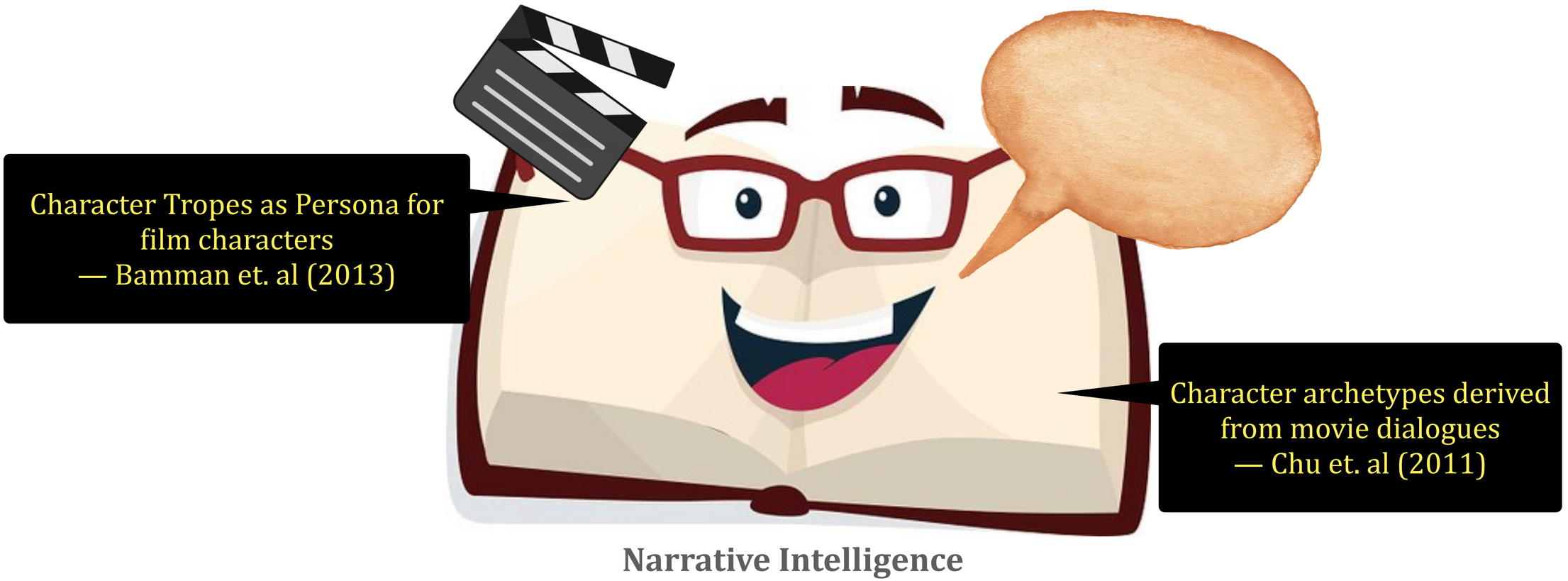
Persona defined by 16 categories based on Myers-Briggs Type Indicator model (MBTI) — Gjurković et. al (2018)

Persona based on Big-Five Personality Traits Model — Golbeck et. al (2011)

Social Media Behavior Mining

1. Gjurković, Matej, and Jan Šnajder. "Reddit: A gold mine for personality prediction." Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. 2018. Azucar, Danny, Davide Marengo, and Michele Settanni. "
2. Golbeck, Jennifer, Cristina Robles, and Karen Turner. "Predicting personality with social media." *CHI'11 extended abstracts on human factors in computing systems*. 2011. 253-262.

What is Persona?



Narrative Intelligence

1. Bamman, David, Brendan O'Connor, and Noah A. Smith. "Learning latent personas of film characters." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013.
2. Chu, Eric, Prashanth Vijayaraghavan, and Deb Roy. "Learning personas from dialogue with attentive memory networks." *arXiv preprint arXiv:1810.08717* (2018).

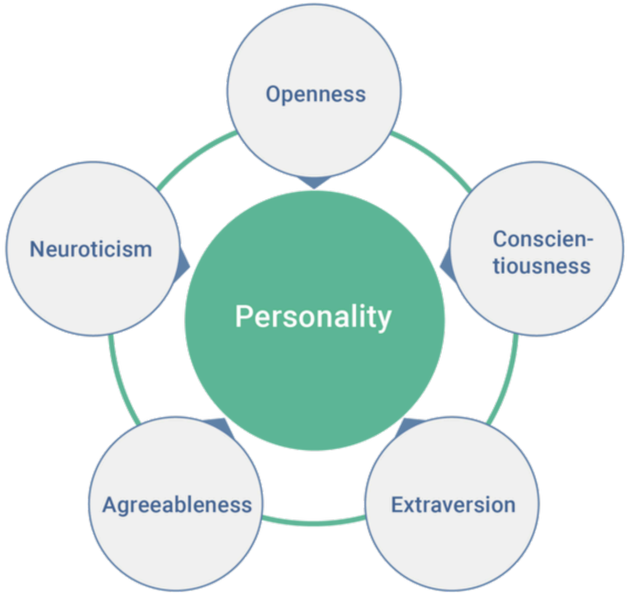
“Persona is defined as the sum total of mental, emotional, and social characteristics of an individual”

Soloff, 1985

1. Gjurković, Matej, and Jan Šnajder. "Reddit: A gold mine for personality prediction." Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media. 2018. Azucar, Danny, Davide Marengo, and Michele Settanni. “
2. Golbeck, Jennifer, Cristina Robles, and Karen Turner. "Predicting personality with social media." *CHI'11 extended abstracts on human factors in computing systems*. 2011. 253-262.

Personal Essays Corpus

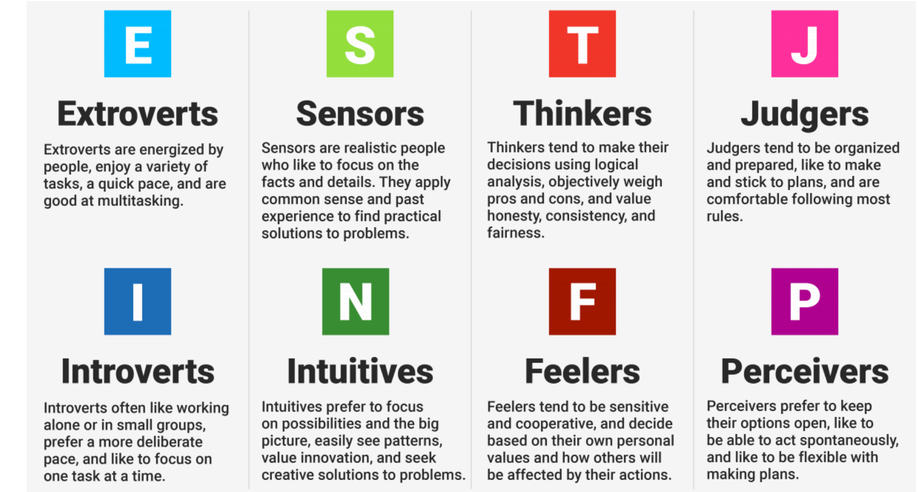
- Personal Essays contain one's goals and values
- Student essays obtained from between 1997-2004
- Based on Big-Five Questionnaires, discretized persona labels computed for essays

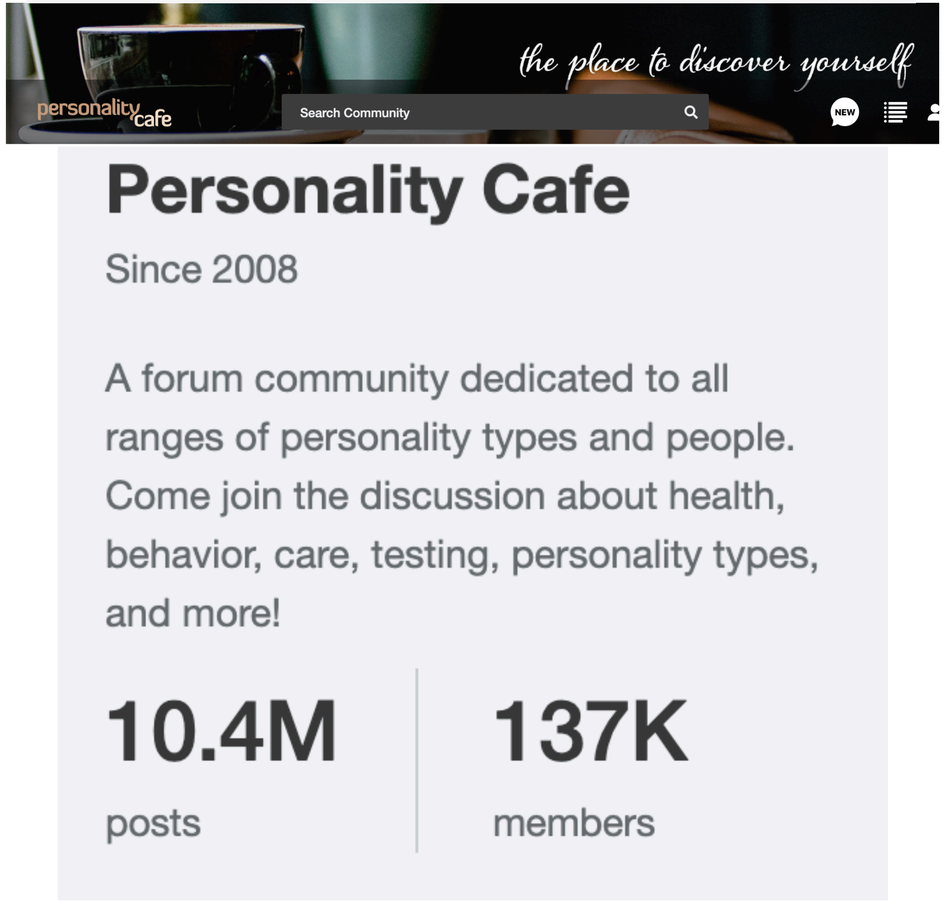


1. Pennebaker, James W., and Laura A. King. "Linguistic styles: Language use as an individual difference." *Journal of personality and social psychology* 77.6 (1999): 1296.
2. Mairesse, François, et al. "Using linguistic cues for the automatic recognition of personality in conversation and text." *Journal of artificial intelligence research* 30 (2007): 457-500.

Forum Posts Corpus

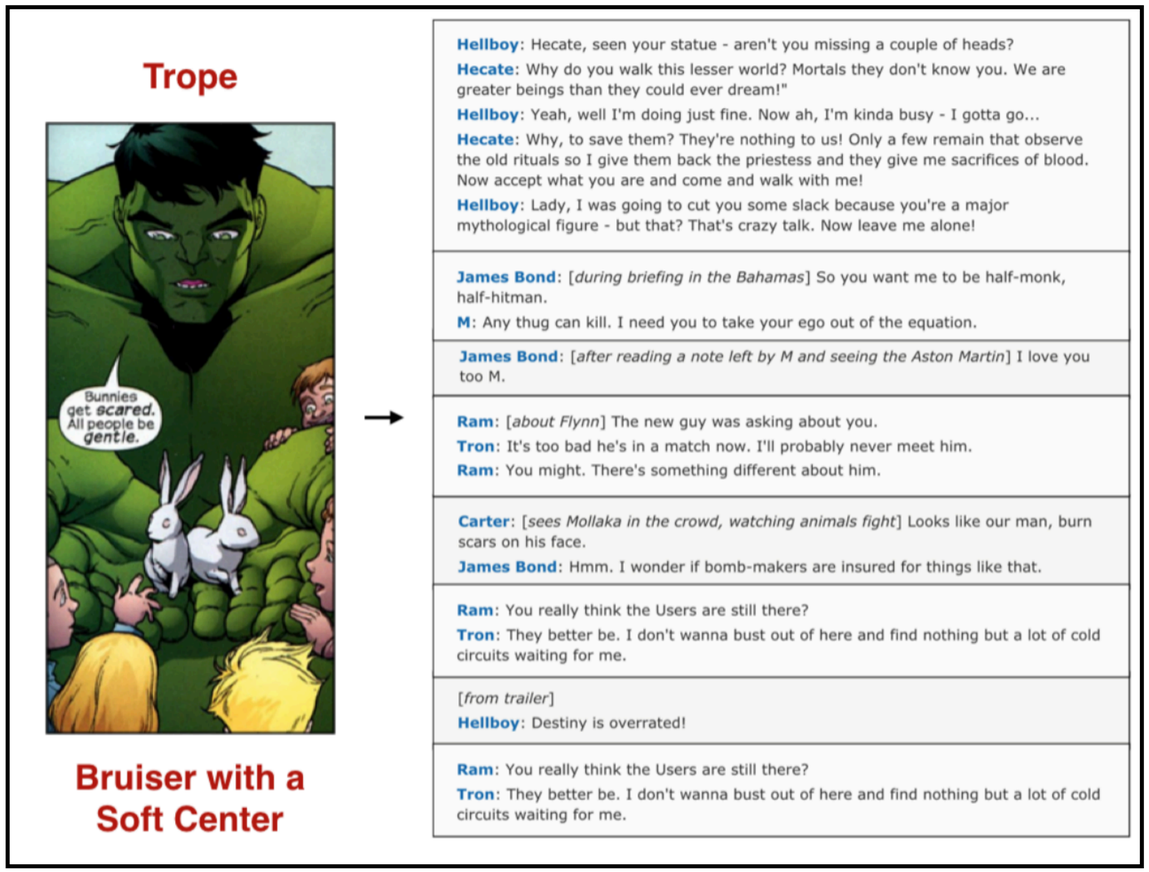
- Jung’s theory of Psychological types — 16 MBTI Types.
- PersonalityCafe — Discussion forum with text in stream-of-consciousness style.
- Crawl different section of forums, filter posts, labels based on self-identified category.





Movies Dialogue Corpus

- IMDB Dialogue Snippets containing utterances of movie characters.
- Character trope— Character aspects such as role, motivations, perceived behavior.
- TVTropes — wiki of document descriptions about plot conventions and devices.



Dataset Statistics

Personal Essays Corpus: Big-Five Model — Extrovert

…. I have some really random thoughts. I want the best things. But I fear that I want too much! What if I fall flat on my face and don't amount to anything. But I feel like I was born to do BIG things on this earth. But who knows… There is this Persian party today. My neck hurts ….

Forum Posts Corpus: PersonalityCafe — ISFJ

#13 • May 15, 2011

I'm tired of people making ad hominem attacks.
I'm tired of people thinking they're better than me because I'm an F.
I still don't believe that Americans care as much about "immigration status" as they care about the color of your skin.

Movie Dialogue Corpus: IMDB Dialogue Snippet

Stacks Edwards: What time is it?
Tommy DeVito: It's eleven thirty, we're supposed to be there by nine.
Stacks Edwards: Be ready in a minute.
Tommy DeVito: Yeah, you were always fuckin' late, you were late for your own fuckin' funeral.
[*shoots him*]

Datasets	Label Type	Size	# Categories
Personal Essays	Big-Five	2,400	5
Forum Posts	MBTI	52,648	16
Movies Dialogue	Tropes	17,342	72

DAPPER Model

Overview

- Learn Domain-adapted Persona Embedding
- Positive Knowledge Transfer across multiple text domains: movies dialogue, forum discussion posts, personal essays.
- Pretrained BERT Model & Transformer-based adaptive layers

Personal Essays Corpus: Big-Five Model — **Extrovert**

…. I have some really random thoughts. I want the best things. But I fear that I want too much! What if I fall flat on my face and don't amount to anything. But I feel like I was born to do BIG things on this earth. But who knows… There is this Persian party today. My neck hurts ….

Forum Posts Corpus: PersonalityCafe — ISFJ

#13 • May 15, 2011

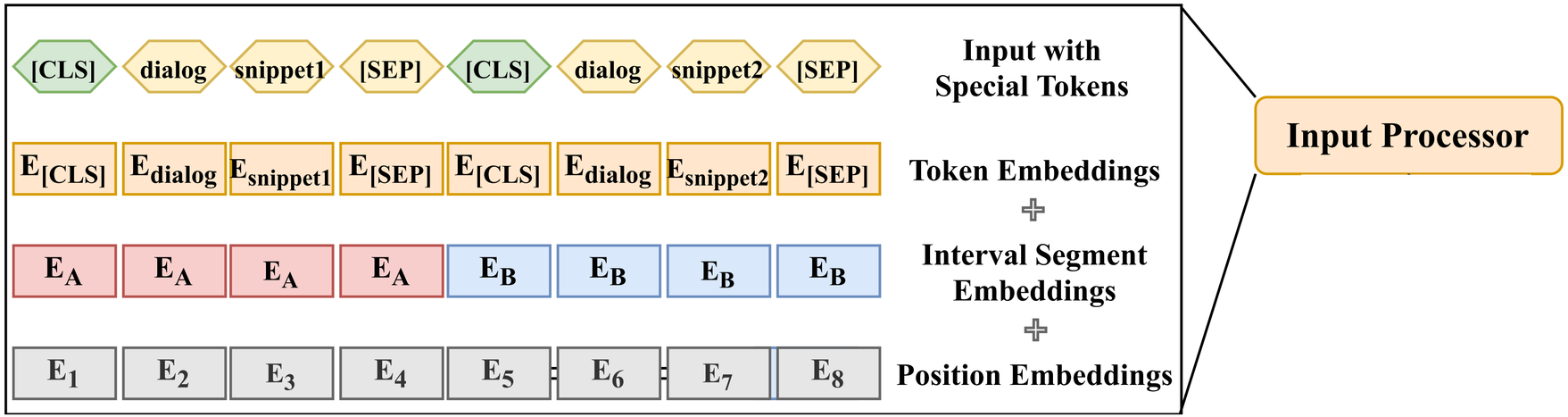
I'm tired of people making ad hominem attacks.
I'm tired of people thinking they're better than me because I'm an F.
I still don't believe that Americans care as much about "immigration status" as they care about the color of your skin.

Movie Dialogue Corpus: IMDB Dialogue Snippet

Stacks Edwards: What time is it?
Tommy DeVito: It's eleven thirty, we're supposed to be there by nine.
Stacks Edwards: Be ready in a minute.
Tommy DeVito: Yeah, you were always fuckin' late, you were late for your own fuckin' funeral.
[*shoots him*]

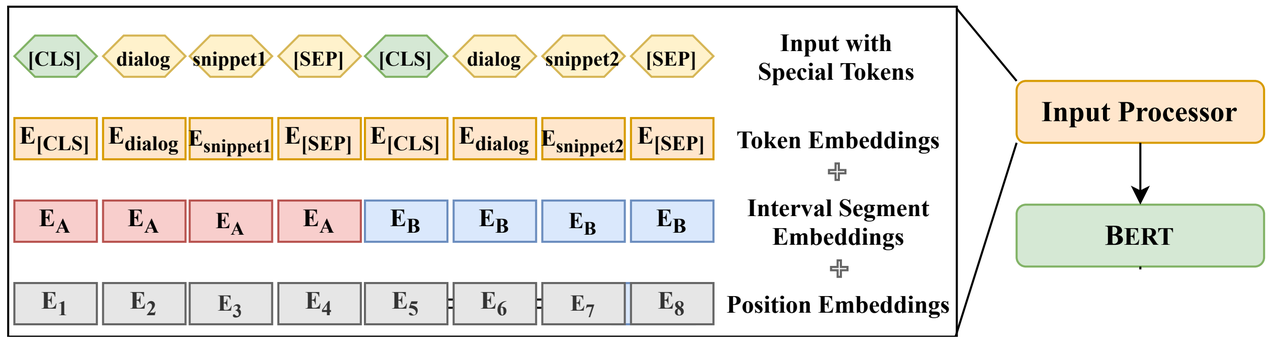
Input Processor

- Uniform way of representing data from different domains.
- Introduce special tokens — [CLS], [SEP]
- Combine token, segment and position embeddings



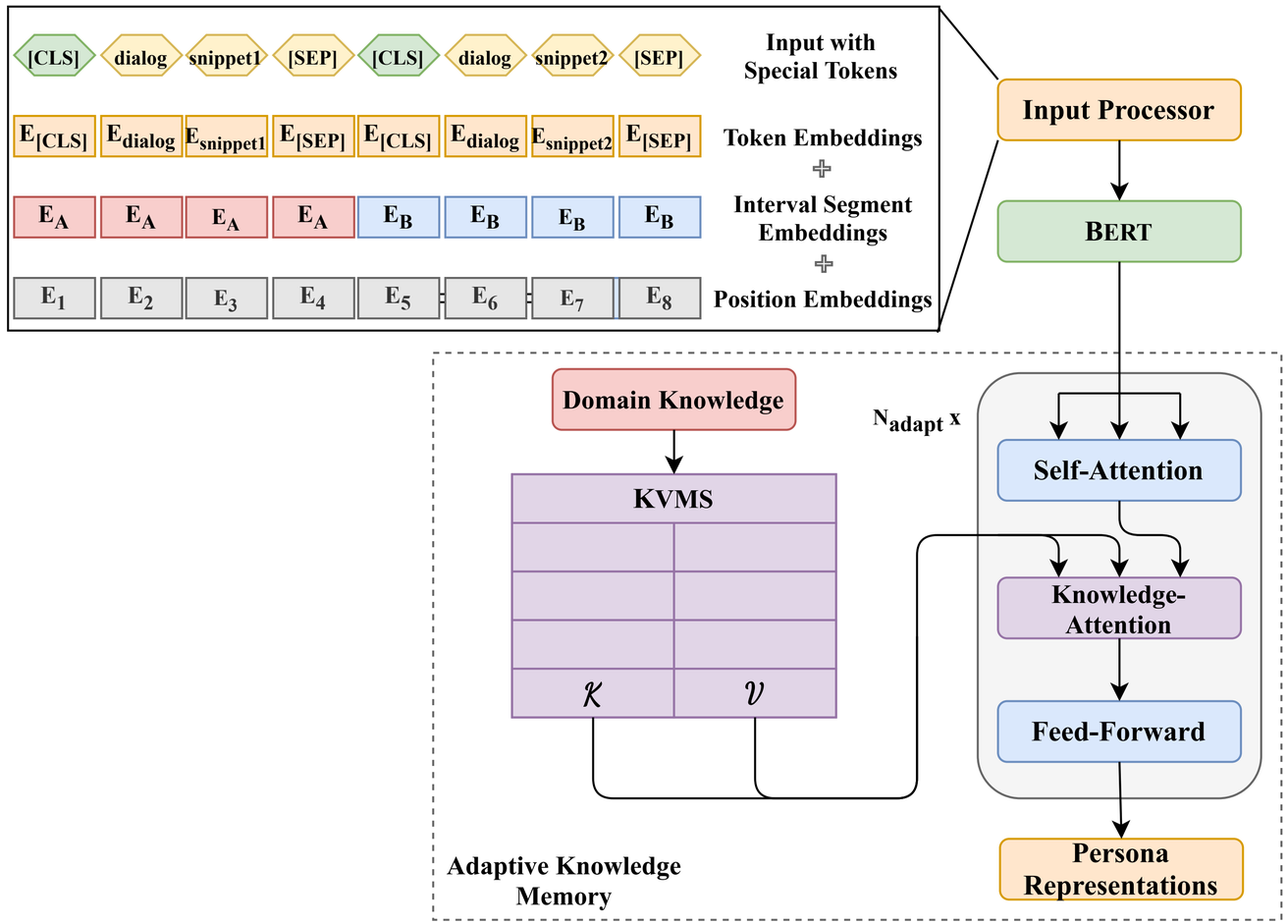
Encoder

- Input: Uniform Input Document representation
- Model: Pretrained BERT model
- Output: Contexualized Document representation



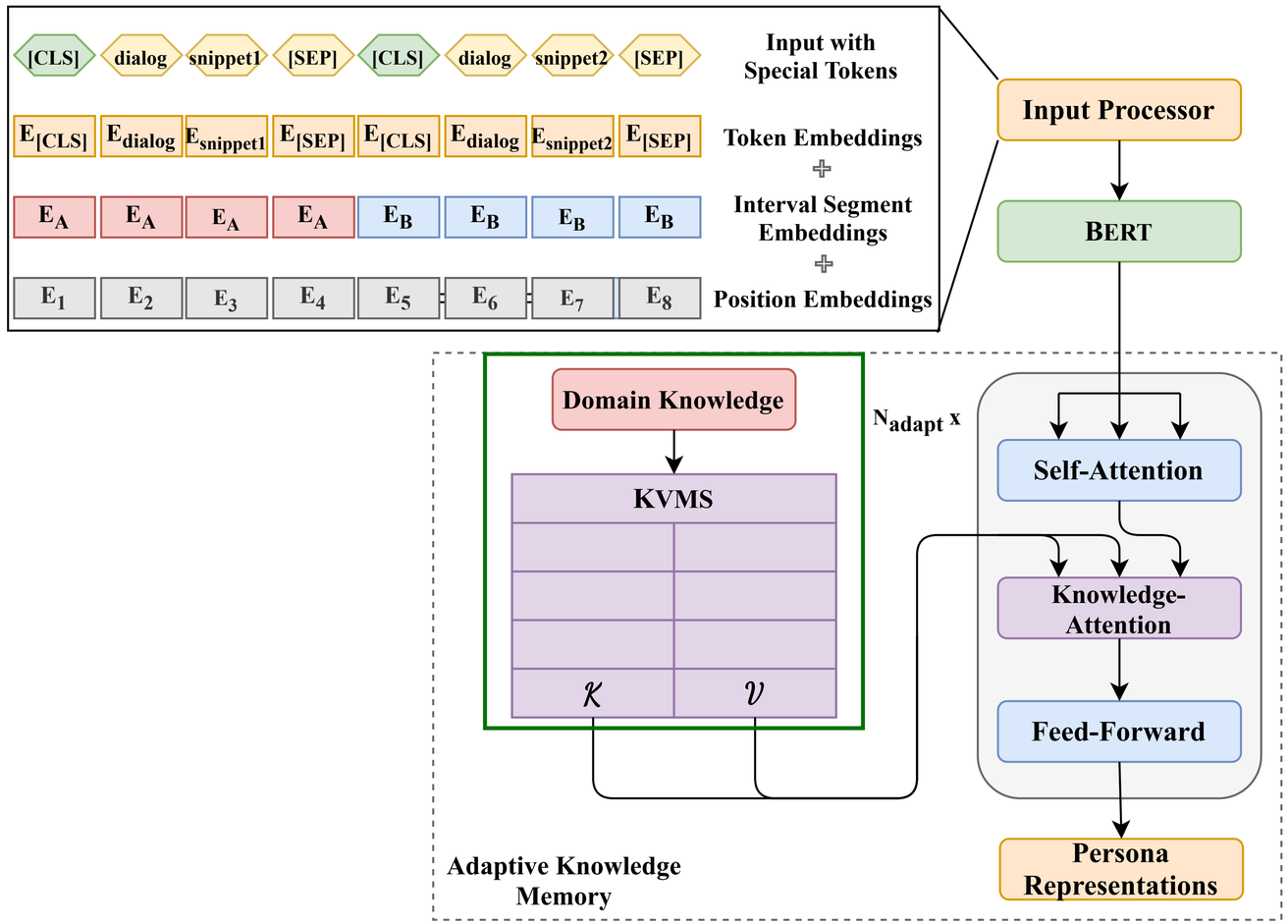
Adaptive Knowledge Transformer

- Compute persistent latent embeddings related to Persona Categories
- Accumulate more information with more knowledge from each domain.
- Key Components:
  - Key-Value Memory Store
  - Transformer-based adaptive module



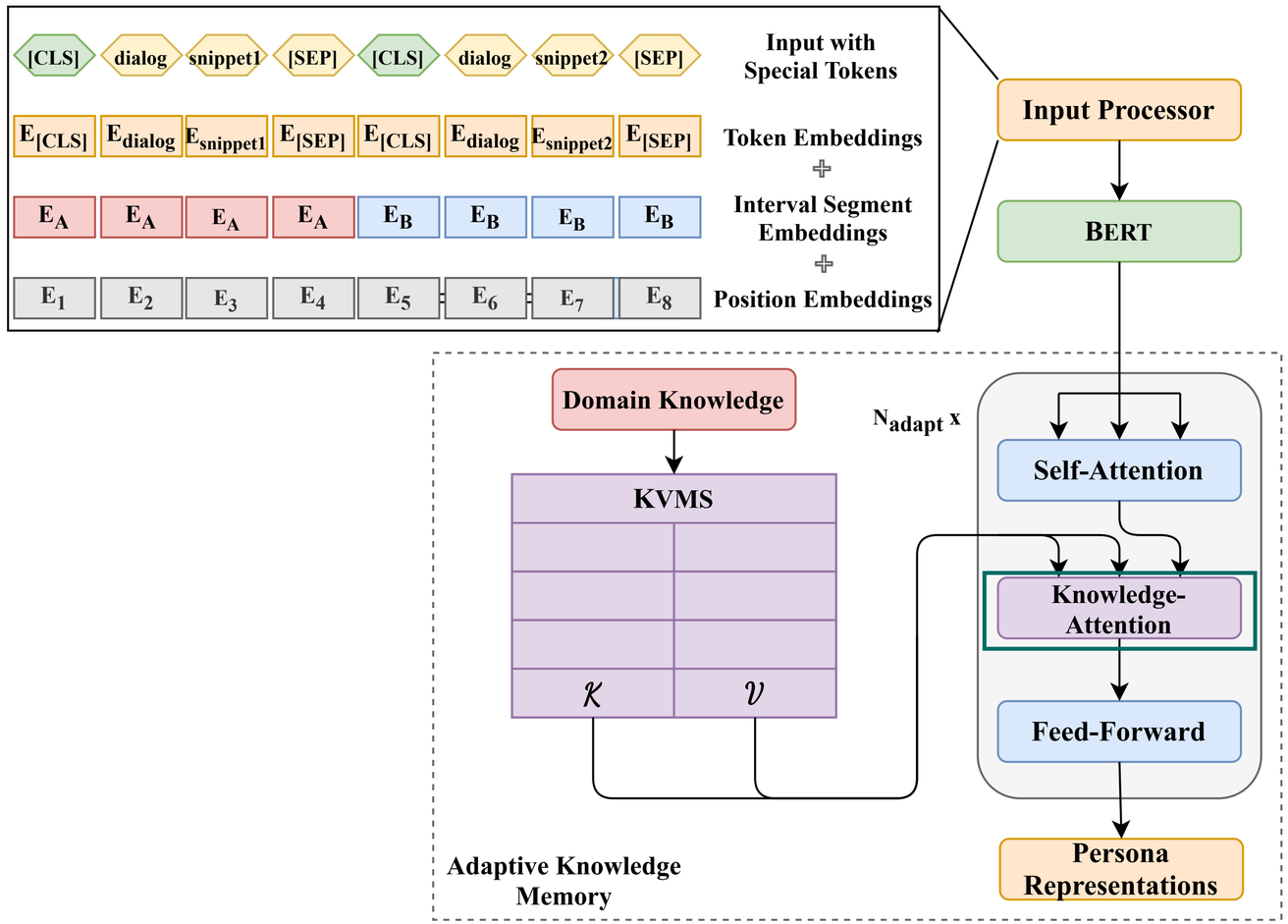
KVMS: Key-Value Memory Store

- Key Matrix: Mutable, Accumulated multi-domain knowledge
  - Initialized with representations of descriptions — tropes, MBTI types & Big-Five traits
- Value Matrix: Immutable, learnable persona category embedding



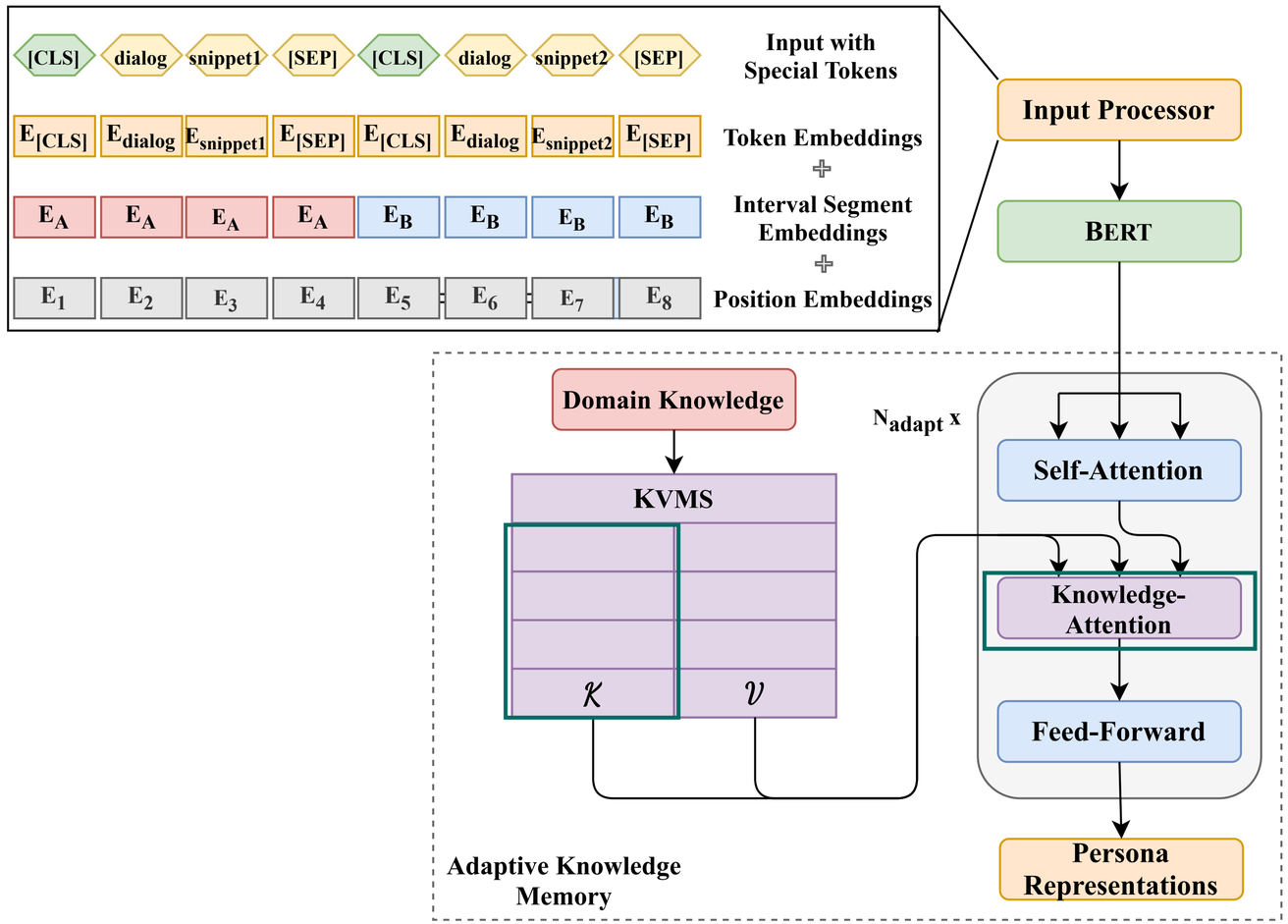
Knowledge-Attention

- Transformer-based adaptive layers contain an augmented sub-layer — knowledge-attention
- Identifies most correlated knowledge from KVMS and produces knowledge enhanced embeddings
- Last layer output— domain adapted persona representation



Memory Update

- Accumulate persona-related knowledge.
- Selective memory row update in Key Matrix with ground truth persona labels.
- Mean [CLS] embedding used to update Key Matrix.



RQ1: How well does our DAPPER model perform in comparison to baselines and its variants?

Experiments (RQ1)

- Baselines & Model Variants:
  - Aff2Vec, CNN
  - Attentive Memory Networks
  - TTS — Non-pretrained transformer baseline
  - BERT — Fine-Tuned (FT), GRU FT + K
  - DAPPER: Full, Without Knowledge (-K)

Analyze:

Overall Performance

Effect of Architecture Choices

Effect of Knowledge-Attention

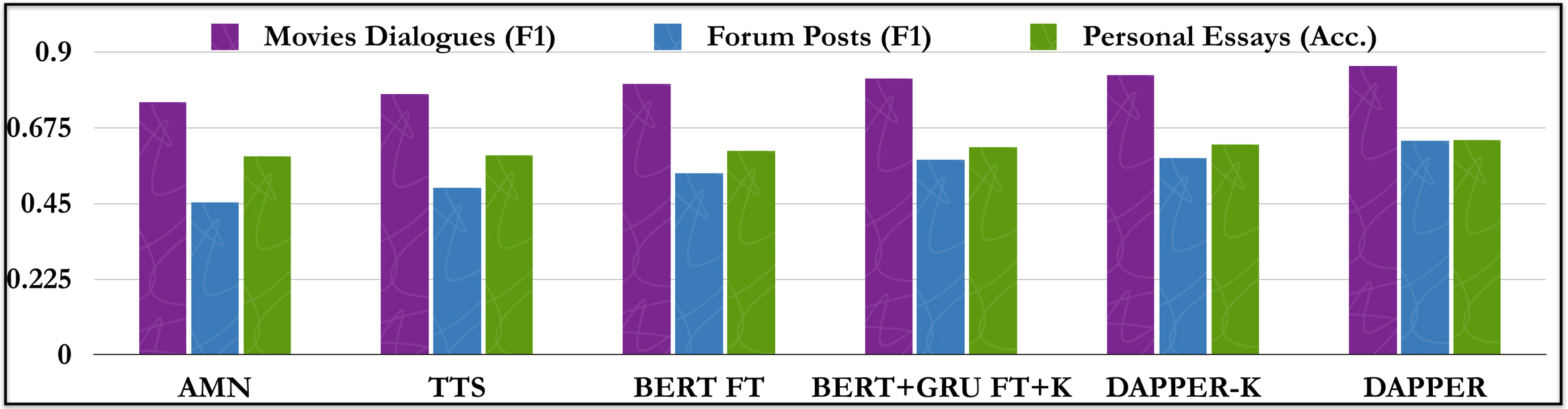
Experiments (RQ1)

DAPPER outperforms other baselines by: **14.53%** on Movie Dialogues, **8.67%** on Personal Essays

Transformer-based adaptive layers with **6.1% improvement** with knowledge attention & **4.2%** without knowledge-attention.

Significant gains with the help of knowledge attention — **7.38%** for Forum Posts

Percentage increase doubles within individual domain comparing Transformer vs RNN-based adaptive layers.

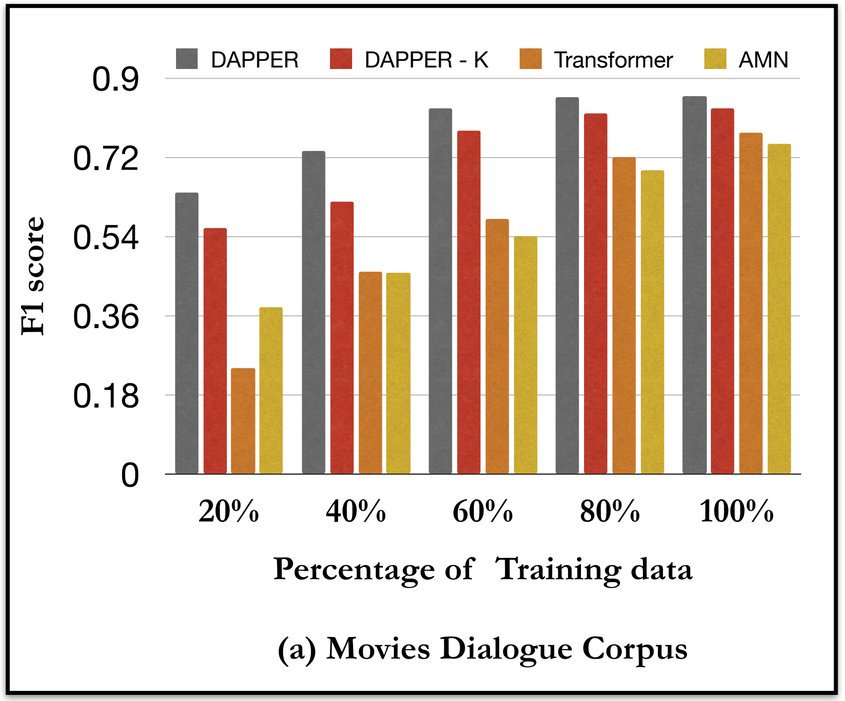


RQ2: Is our model capable of adapting to new domains with limited labeled data?

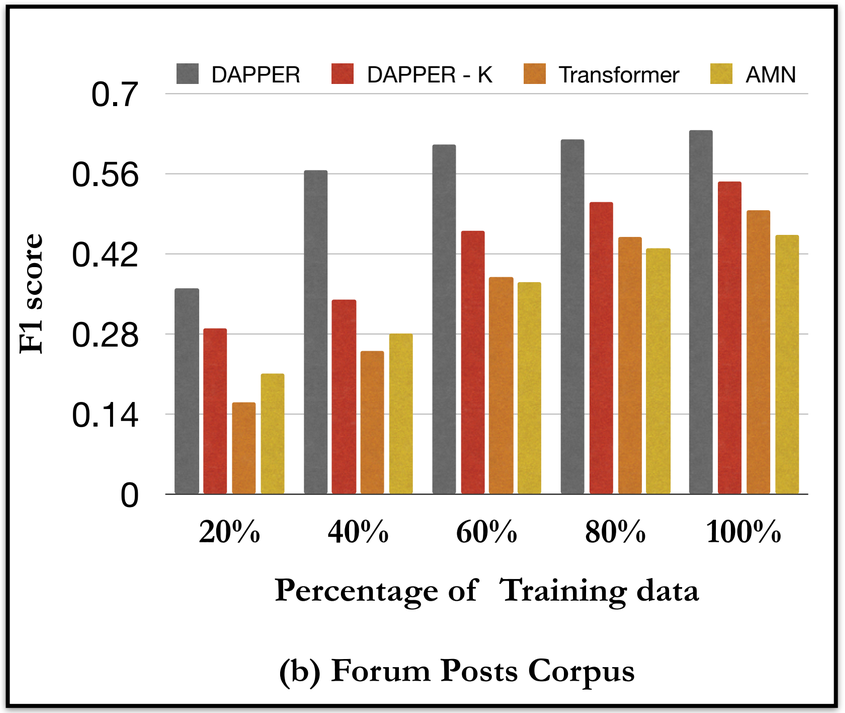
Experiments (RQ2)

"ADAPT" Mode:

- Simulate low-data setting
- Restrain amount of training for one domain while retaining complete set for other domains.



(a) Movies Dialogue Corpus



(b) Forum Posts Corpus

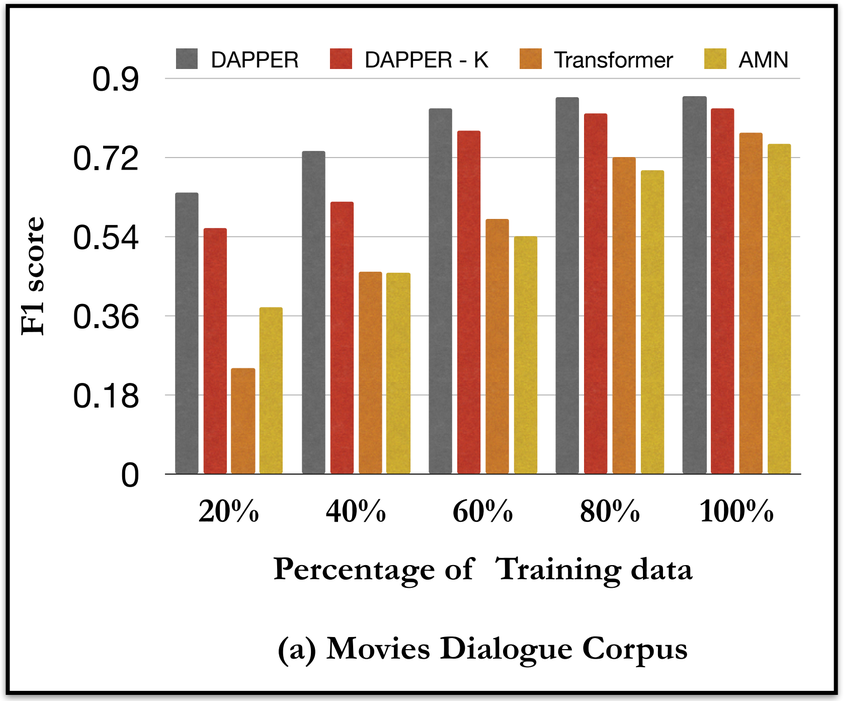
Experiments (RQ2)

"ADAPT" Mode:

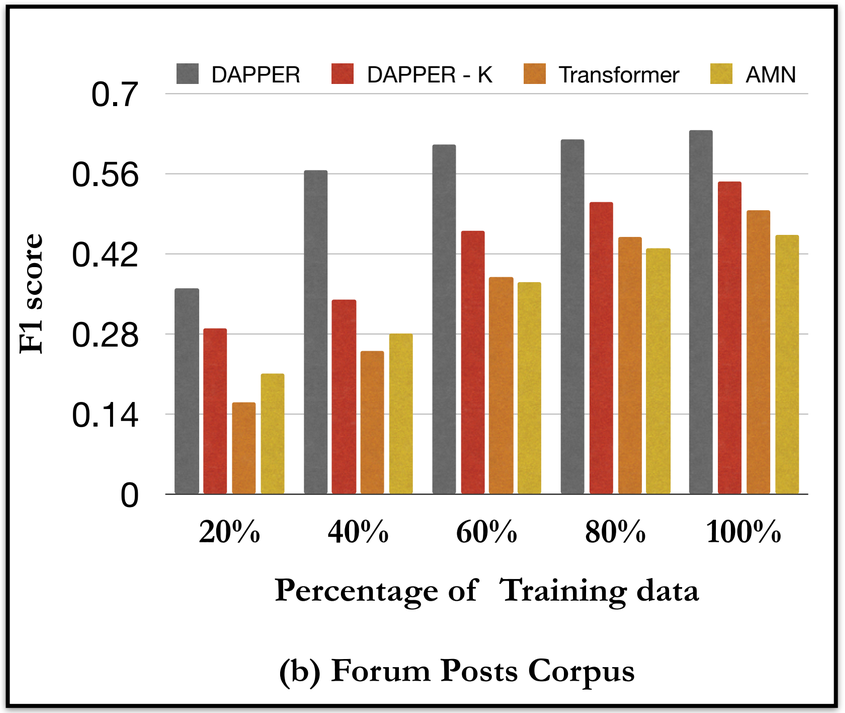
- Simulate low-data setting
- Restrain amount of training for one domain while retaining complete set for other domains.

DAPPER learns general purpose persona embeddings that can adapt to low-data settings.

Achieves good performance with **<40%** of the training data



(a) Movies Dialogue Corpus



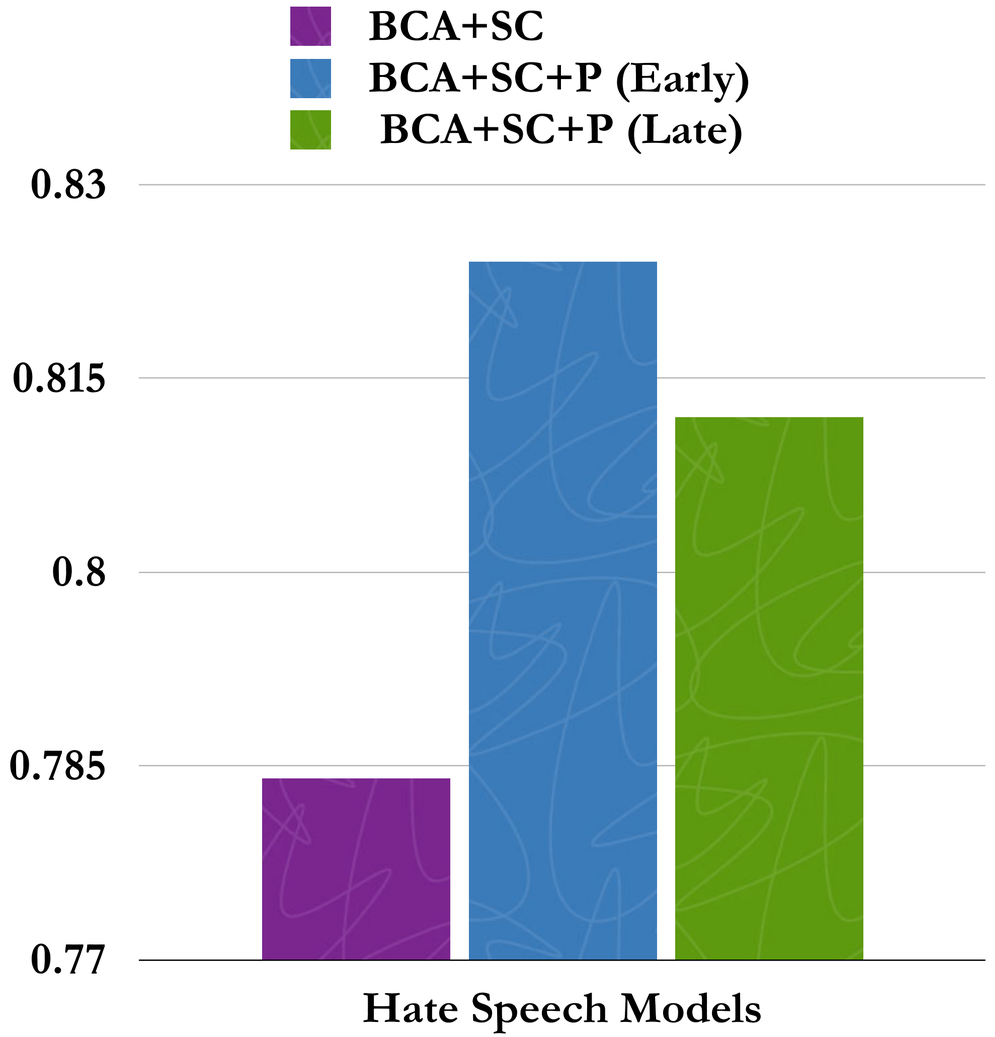
(b) Forum Posts Corpus

RQ3: Are the learned persona embeddings transferrable to a downstream task?

Experiments (RQ3)

Hate Speech Detection

- Baselines:
  - BCA — BiGRU+Char+Attention
  - SC — Socio-cultural features based on demographic & network features
  - P — DAPPER Persona embeddings
  - Early Vs Late Fusion Techniques



1. Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. Interpretable multi-modal hate speech detection. AI for Social Good Workshop, ICML 2019

Thank You!

Contact: pralav@media.mit.edu