Multimodal Cross-Document Event Coreference Resolution Using Linear Semantic Transfer and Mixed-Modality Ensembles

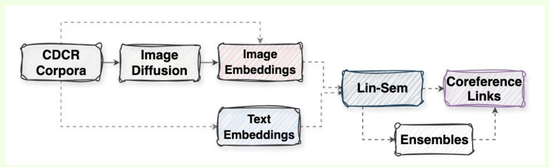
Abhijnan Nath, Huma Jamil, Shafiuddin Rehan Ahmed, George Baker, Rahul Ghosh, James H. Martin, Nathaniel Blanchard, and Nikhil Krishnaswamy

Colorado State University — Situated Grounding and Natural Language Lab

University of Colorado Boulder • {abhijnan.nath,nkrishna}@colostate.edu

Introduction

- Images contain useful coreference cues for events
- Visual cues from images can help resolve coreference when text descriptions are ambiguous
  - For example, facial, geographical, locational features
- However, most CDCR corpora is not multimodal: sparsity problem
- We augment common CDCR corpora with stable-diffusion models
- Apply Linear Semantic Transfer (**Lin-Sem**) method to transfer semantic information between modalities
- We establish upper-limit on ECB+; new baseline on AIDA Phase 1 data



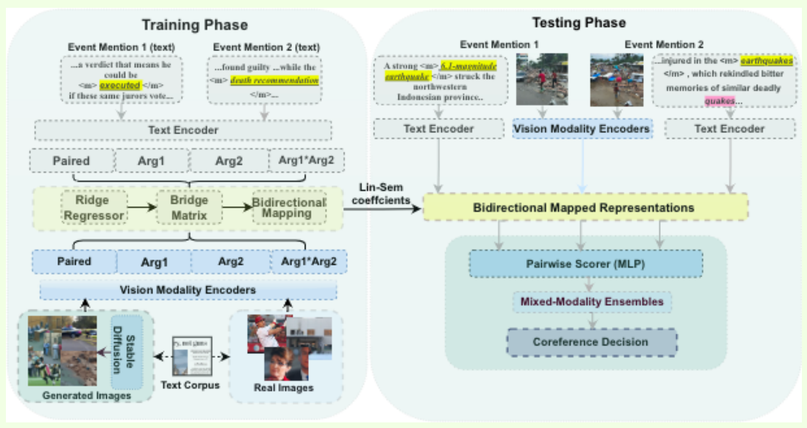
High-level overview of approach

Image Generation and Encoding

- Generate photo-realistic images from ECB+ mentions using Stable Diffusion models
- One image/mention ~ less-compute
- Source images from URLs in ECB+ meta data
- AIDA data already contains images for mentions
- We encode these images with ViT, SWIN, BEiT and CLIP

Linear-Semantic Transfer (Lin-Sem)

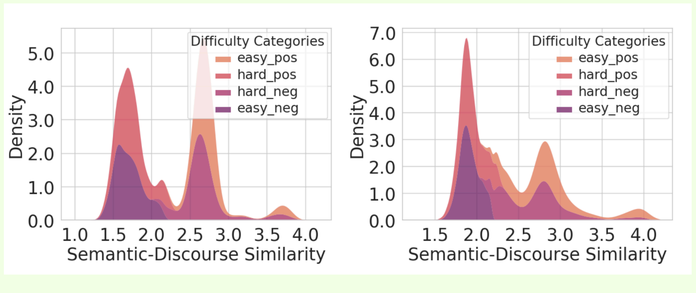
- Text and Image representations of event mentions are linearly-mapped *bidirectionally*
- We use a ridge regressor to generate optimal mapping matrices (coefficients)
- At inference, we matrix multiply both modality representations with their respective mapping matrices



Linear semantic Transfer (Lin-Sem) Method. **Arg1** and **Arg2** refer to

Semantic Transfer Categories

- Similarity scores (within-topic, within-doc, Wu-Palmer and mention cosine similarity) are computed.
- Based on a threshold (dev-set), mention pairs are binned into easy-positive, easy-negative, hard-positive and hard-negative
- Categories used in selecting ensemble components



Kernel Density Estimation plots of similarity scores for mention pair difficulty categories in ECB+ (L) and AIDA Phase 1 (R)

Results (with Ensembles)

Models	MUC	B^3	$CEAFe$	CoNLL
Held et al. (2021)	87.5	86.6	82.9	85.7
Ahmed et al. (2023)	90.8	86.7	84.7	87.4
ViT-gen ⊕ LLM + LLM	89.1	86.5	84.8	86.8
BEiT-gen ⊕ LLM + LLM	87.5	85.7	83.9	85.7
SWIN-gen ⊕ LLM + LLM	87.5	85.9	83.8	85.7
CLIP-gen ⊕ LLM + LLM	90.1	85.3	83.8	86.4
ViT-gen→LLM + LLM→BEiT-gen + LLM	90.8	85.2	84.8	86.9
BEiT-gen→LLM + LLM→BEiT-gen + LLM	91.3	85.5	86.5	87.8
SWIN-gen→LLM + LLM→BEiT-gen + LLM	90.4	84.4	83.8	86.2
CLIP-gen→LLM + LLM→BEiT-gen + LLM	91.2	85.3	85.7	87.4
LLM→ViT-gen + LLM→BEiT-gen + LLM	88.7	82.3	79.4	83.5
LLM→BEiT-gen + LLM	88.7	82.2	79.1	83.3
LLM→SWIN-gen + LLM→BEiT-gen + LLM	88.7	82.2	79.1	83.3
LLM→CLIP-gen + LLM→BEiT-gen + LLM	88.7	82.2	79.1	83.3
ViT-real→LLM + LLM→BEiT-real + LLM	**94.5**	**89.5**	**91.8**	**91.9**
BEiT-real→LLM + LLM→BEiT-real + LLM	88.9	82.4	79.7	83.7
SWIN-real→LLM + LLM→BEiT-real + LLM	88.7	82.2	79.1	83.3
CLIP-real→LLM + LLM→BEiT-real + LLM	94.3	89.3	91.6	91.7
LLM→ViT-real + LLM→BEiT-real + LLM	88.7	82.3	79.3	83.4
LLM→BEiT-real + LLM	88.7	82.2	79.1	83.3
LLM→SWIN-real + LLM→BEiT-real + LLM	89.0	82.7	80.1	83.9
LLM→CLIP-real + LLM→BEiT-real + LLM	88.7	82.2	79.1	83.3

MUC, B^3, $CEAF_e$ and CoNLL F1 results on **ECB+ test set,** using ensemble models. Ensemble model names are formatted **Hard-N model + Hard-P model + Easy pairs model**. LLM was always used to handle Easy pairs.

Models	MUC	B^3	$CEAFe$	CoNLL
LLM	80.7	49.5	54.1	61.4
ViT-real→LLM	85.9	38.4	52.7	59.0
BEiT-real→LLM	85.7	42.6	57.9	62.1
SWIN-real→LLM	82.9	46.4	55.8	61.7
CLIP-real→LLM	78.5	**52.4**	53.5	61.5
LLM→ViT-real	86.3	37.3	52.7	58.8
LLM→BEiT-real	85.7	40.2	53.1	59.7
LLM→SWIN-real	86.2	39.1	54.4	59.9
LLM→CLIP-real	86.2	37.1	52.3	58.5
ViT-real→LLM + LLM→BEiT-real + LLM	86.2	39.6	54.4	60.1
BEiT-real→LLM + LLM→BEiT-real + LLM	**87.1**	42.1	60.4	63.2
SWIN-real→LLM + LLM→BEiT-real + LLM	**87.1**	42.5	60.5	63.4
CLIP-real→LLM + LLM→BEiT-real + LLM	**87.1**	43.8	**62.8**	**64.6**
LLM→ViT-real + LLM→BEiT-real + LLM	86.2	39.0	53.5	59.6
LLM→BEiT-real + LLM	85.8	40.8	54.1	60.2
LLM→SWIN-real + LLM→BEiT-real + LLM	86.6	40.7	56.6	61.3
LLM→CLIP-real + LLM→BEiT-real + LLM	86.2	39.0	53.5	59.6

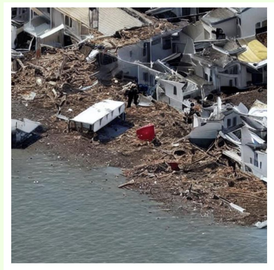
MUC, B3 , CEAF e and CoNLL F1 results on **AIDA Phase 1 Eval set.**

Analysis

- Linear mapped systems close performance gap with text-only + domain-fused models
- Ensembles outperform text-only baseline (Ahmed at al., 2023) and establish upper-limit (91.9 CoNLL F1) on ECB+
- 64.5 CoNLL F1 on AIDA (+3 F1 vs Text-only LLM)
- Multimodal cues help resolve harder coreferences
- Visual media contain referential cues, missing in text-modality: useful for semantic transfer

Doctor Who has finally selected its 12th doctor : Peter Capaldi is officially set to replace exiting star Matt Smith as the TARDIS leader , producer *Steven Moffat* **<m> announced </m>** on the live BBC special Doctor Who Live : The Next Doctor Sunday.

Scottish actor best known for *his* role as Malcolm Tucker in The Thick of It **<m> revealed </m>** as 12th actor to play the Doctor.



A young girl was killed and dozens of others were seriously injured in the quakes , **<m> which </m>** also sent *small tsunamis* into Japan 's

Atururi said a 10-year-old girl was killed and at least 40 people were injured in the **<m> earthquakes </m>** , which rekindled bitter memories of similar

Sample coreferent event pairs from ECB+ (L: real images, R: generated images) that were correctly linked by our best multimodal ensemble (ViT-real→LLM + LLM→BEiT-real + LLM), but not by the text-only model

Conclusion and Future Work

- Multimodal cues can help event coreference resolution especially for harder cases
- Lin-Sem mapping between embedding spaces can transfer coreference-specific knowledge between distinct modalities.
- Upper-bound on the ECB+ (> 3 F1 points) using our ensembles + a new baseline on AIDA
- Overall, modular and compute-efficient
- Apply to more challenging corpora like the FCC with cross-subtopic mentions or in multi-lingual ECR.

Resources

Codebase

Generated Images

Paper