Actionable Entities Recognition Benchmark for Interactive Fiction

Alexey Tikhonov, Inworld.AI, Berlin, Germany &
Ivan P. Yamshchikov, MPI for Mathematics in the Sciences, Leipzig, Germany, CEMAPRE, Lisbon, Portugal
altsoph@gmail.com, ivan@yamshchikov.info

Abstract

We present a new natural language processing task - **[Actionable Entities Recognition]** (AER) - recognition of entities that protagonists could interact with for further **[plot development]**. AER might be helpful building the efficient agents for interactive text environments, as long as for the systems dealing with narrative processing. We present a **[new benchmark dataset]** for the AER task that includes **[5550 descriptions]** with one or more **[actionable entities]**.

*"One must never place a **[loaded rifle]** on the stage if it is not going to go off. It is wrong to make promises you do not mean to keep."*
A. Chekhov.

Dataset definition

We analysed **[1500+ interactive fiction games]**, crawled through the **[available locations]** (using the walktrough instructions when possible or a random walk otherwise) and extracted textual descriptions of them. Then we probed each element of the description of each location we found with the command `'examine <obj>'`. If interaction with one of the entities gives a non-trivial reaction from the game, we label this **[entity]** as an AE. Most of interactive fiction (IF) games have a standard response to inconsequential actions of the player. If the game's response is different, we assume that the entity is actionable. We finished with **[5550 locations from 995 different games]** with one or more **[AE]** in each of them. The total size of **[dataset]** is 1.2 megabytes. Every **[Actionable Entity]** in the text is labeled as shown in Figure 1. We publish the resulting **[BAER dataset]**[1] with labeled **[AEs]** and suggest it for AER algorithms benchmarking.

*The rather large proscenium seems to have been designed mainly with live performances in mind (the dead were thought, not unreasonably, to be dreadfully dull.) As such, it sports a large, maroon **[curtain]** leading backstage and a row of brightly colored **[footlights]**. **[Rubble and debris]** block the ways to the wings, and a set of **[stairs]** leads back down into the theater's main aisle.*

** * **

*A thick, maroon **[curtain]** separates the backstage area from the stage. This area was obviously the target of a small, underground tornado (a Vorx) as **[scrims, scenery and costumes litter]** the floor.*

** * **

*The Museum of Illusion was dedicated to the memory of the Great **[Implementers]** - figures who were believed to have created the GUE as an act of pure will. Before you, a row of delicately crafted **[porcelain busts]** of these immortal greats once stood. Legend has it that so lifelike were these **[busts]**, that they would seem to talk among themselves, discussing history, the arts, music, and philosophy, much as those mythic figures did in the Golden Age of Text Adventures. But years of neglect and the ravages of time have toppled most from atop their finely-wrought pedestals. Now, alas, the only two that remain are those of **[Marc Blank and Mike Berlyn]**, bloodied, but unbent; battered, but unbroken; shaken, but not stirred.*

Figure 1: AERs examples in texts from the ZTUU game.

Task validation

Using ClubFloyd dataset

[The ClubFloyd dataset] contains human **[gameplay transcripts]**, which cover 590 text-based games of diverse genres and styles. We extracted location descriptions and lists of **[Action Targets]** (AT) – entities with which players were trying to interact. Since the **[dataset]** contains game logs of several players, some **[ATs]** could be mentioned several times. We consider it like these **[entities]** are perceived by **[humans]** as more interesting to explore. **[Table 1]** summarizes what share of ATs could be labeled by an AER model.

[1] https://github.com/altsoph/BAER

Table 1: Relationship between AEs detected by the XLMR model traind on BAER and ATs extracted from the Club Floyd dataset.

	All payer action targets (AT)				Unique ATs			
AER model threshold	p >0.5	p >0.65	p>0.8	p>0.95	p >0.5	p >0.65	p>0.8	p>0.95
Share of AEs that occur in AT list	0.38	0.43	0.48	**0.57**	0.22	0.25	0.29	0.36
Share of ATs labelled as AEs	**0.84**	0.65	0.30	0.05	**0.84**	0.65	0.30	0.04

Contrasting Actionable and Named Entities

Table 2: Pre-trained XLMR t-NER hardly detects AEs out of the box. Yet after fine-tuning on BAER for entity span prediction the quality significantly improves.

	Accuracy	F1-score
Pre-trained T-NER	0.05	0.00
T-NER fine-tuned on BAER	0.50	0.51

Table 3: Top t-NER categories with the highest frequencies of AEs across four different datasets.

BAER 1 117 texts		ClubFloyd 43 795 texts		TV-MAZE 299 197 texts		WikiPlots 2 070 449 texts	
T-NER category	$\frac{\nu_{AER}}{\nu_{ner}}$	T-NER category	$\frac{\nu_{AER}}{\nu_{ner}}$	T-NER category	$\frac{\nu_{AER}}{\nu_{ner}}$	T-NER category	$\frac{\nu_{AER}}{\nu_{ner}}$
product	18	chemical	7.5	person	36.5	product	31.2
person	2.6	product	7.3	product	12.2	corporation	22.7
work of art	2.0	person	5.5	corporation	3.9	person	20.1
organization	0.7	other	2.1	organization	3.3	group	5.4
location	0.5	work of art	1.3	chemical	3.1	chemical	4.9

AEs and structure of narrative

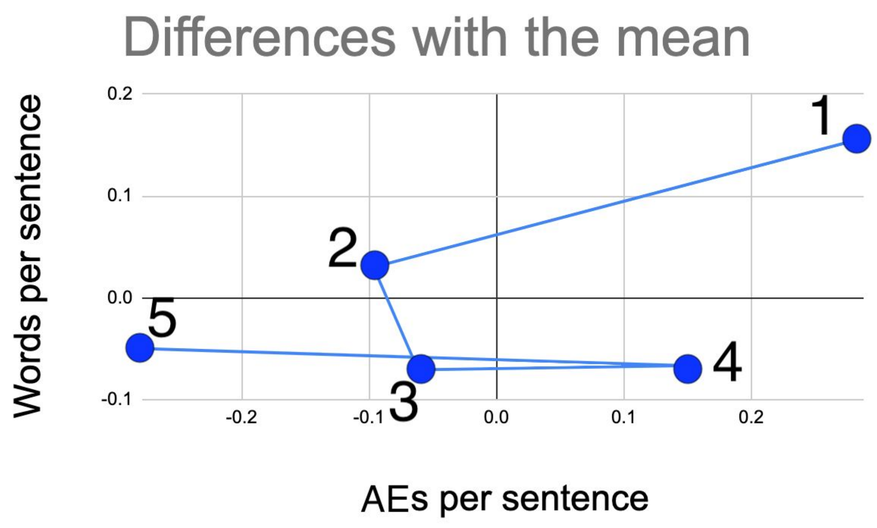

Figure 2: Average number of AEs per sentence and the average number of words in a sentence differ in different turning points of the story based on the TRIPOD dataset.

This paper:

- presents a **[new task]** — **[AER]**;
- presents a **[new BAER dataset]** to introduce the **[benchmark for AER]**;
- shows **[correlation]** between AEs and human attempts to interact with entities;
- illustrates the **[differences]** between **[AER]** and **[NER]** tasks;
- analyzes the **[connection]** between **[AEs]** in the text and the **[structure of the narrative]**.