Tytuł: SMIS

(strona tytułowa do wstawienia z Moja PG)

Kamil Kozioł Alicja Wagner Kacper Chodubski Kacper Konopka

11 października 2024

STRESZCZENIE

Praca skupia się na zagadnieniu wykorzystania modeli języka naturalnego. Większość obecnego rynku skupia się na bezpośrednim wykorzystaniu modeli do miedzy innymi pozyskiwania informacji, redagowania tekstów i tym podobnym. Nasz projekt wykorzystuje model jako narzędzie do imitowania ludzkiego postrzegania świata. Model jest wykorzystywany jako ”osoba” podejmująca decyzje na wzór tych człowieka. Żeby móc zaprezentować działanie tego mechanizmu potrzebowaliśmy zbudować wirtualne środowisko, gdzie agenci, czyli nasz odpowiednik wirtualnego człowieka, będą podejmowali działania zaplanowane i zarządzane przez model językowy. Pozwoli nam to zobaczyć jak dużo wiedzy na temat realnego świata jest zawarte w zbiorach danych używanych przez twórców tych modeli oraz jak dobrze odzwierciedlają one zachowanie ludzi.

Słowa kluczowe: Agent, LLM
Dziedzina nauki i techniki, zgodnie z wymogami OECD:
Nauki o komputerach i informatyka

ABSTRACT

Abstract in english.

Keywords: Keyword 1, Keyword 2, Keyword 3
Field of science and technology in accordance with OECD requirements:
Computer Science and Information technology

SPIS TREŚCI

WYKAZ WAŻNIEJSZYCH OZNACZEŃ I SKRÓTÓW

SMIS - Social Modeling and Interaction Simulator – nazwa własna systemu, który został zrealizowany w ramach projektu dyplomowego inżynierskiego.

LLM - (ang. *Large Language Model*) – duży model językowy.

RAG - (ang. *Retrieval Augumented Generation*) – technika generowania tekstu przy użyciu zewnętrznych źródeł danych.

1. WSTĘP I CEL PRACY

Wstępu wstępu - tutaj należy pokrótce opisać o co chodzi w pracy i wyraźnie wskazać cel pracy!

Wstęp i cel pracy nakreśla problematykę opisaną lub rozwiązywaną w pracy dyplomowej wraz z uzasadnieniem celowości jej realizacji. Podaje cel i ewentualnie tezę (hipotezę). Syntetycznie opisuje dotychczasowe dokonania w danej tematyce, założenia techniczne oraz może zwięźle przedstawić zawartość poszczególnych rozdziałów. W przypadku pracy realizowanej przez kilku studentów, przy omawianiu zawartości rozdziałów należy podać ich autorów. Punkty stanowiące element składowy podrozdziału powinny być opracowane przez jednego autora

1.1. Motywacja

dlaczego akurat taki temat, czemu warto coś robić w tej dziedzinie

1.2. Cel pracy

co chcemy zrobić, implementacja papieru

2. WPROWADZENIE DO DZIEDZINY

tekst

2.1. Generatywna sztuczna inteligencja

tekst

2.2. Duże modele językowe

llm + trenowanie, różnice między llama a llama chat

2.3. Retrieval Augumented Generation

teoria + metody poprawy retrievalu

2.4. Agenty dużych modeli językowych

tekst, agenty llm

3. TECHNOLOGIE, ALGORYTMY I NARZĘDZIA

tekst

3.1. Silnik gry Unity

czemu Unity, zalety, do czego jest wykorzystane, jakie biblioteki, skąd assety itp.

3.2. Python

czemu python, jakie biblioteki, krótkie porównanie z innymi językami

3.3. Wybrane modele językowe

3.3.1. Wybór modeli

jak przebiegał wybór modeli, opensource, opisać wszystkie, których używaliśmy

3.3.2. Wektoryzacja i osadzanie słów

tekst

3.3.3. Podobieństwo cosinusowe

tekst

3.3.4. Zasoby obliczeniowe

opisać, że korzystamy z serwera KASKu, ile VRAMu itp, jak to się przekłada na szybkość działania itd

4. PROJEKT SYSTEMU

tekst, tu już typowo o naszej implementacji (porównywać do papieru), na początku krótki opis że dzieli sie na front i backend i model

4.1. Interfejs gry

cały opis unity, za co odpowiada, jakie są interakcje, animacje

4.2. Architektura serwera i warstwy logicznej

4.2.1. Agenci

Pamięć agentów

Pamięć agentów jest kluczowym elementem ich funkcjonowania. To dzięki niej agenci reagują w sposób bardziej ludzki.

Architektura pamięci agentów

Pamięć opiera się na modelu języka naturalnego typu encoder, który przekształca nowe wspomnienia agenta na wektor. Później, w trakcie interakcji czy budowania nowych wspomnień przez agenta jest wybierane k wspomnień, które uzyskały największą średnią ważoną z trzech metryk: podobieństwa (relevance), ważności (importance) i niedawności (recency). Wybrane wspomnienia są później przekazywane do modelu językowego typu decoder, który na podstawie ich i danego wydarzenia decyduje o zachowaniu agenta.

Metryka podobieństwa

Metryka podobieństwa odpowiada podobieństwu tematycznemu nowego wydarzenia do wydarzenia zapisanego w pamięci agenta. Jest to uzyskiwane za pomocą obliczonej odległości cosinusowej wektorów stworzonych przez model typu encoder.

Metryka ważności

Metryka istoty jest uzyskiwana poprzez ocenę ważności danego wydarzenia w kontekście życia agenta przez LLM. Metryka odpowiada na pytanie jak ważne dla danego agenta w skali od 1 do 10 jest dane wydarzenie np. pościelenie łóżka będzie miało skalę bliską 1, gdzie wzięcie ślubu skalę bliską 10.

Metryka niedawności

Metryka niedawności odpowiada mechanizmowi ”zapominania” przez agenta. Wspomnienia uzyskują wynik zgodnie ze wzorem:

$$score = max\_score * decay\_factor^{time\_diff} \tag{4-1}$$

4.3. Polecenia dla modeli językowych

Jakość odpowiedzi generowanych przez duże modele językowe w znacznym stopniu zależy od zapytań, które są do nich wysyłane. Tekst przekazywany do modelu nazywany jest poleceniem (ang. *prompt*), którego główną częścią jest pytanie lub zadanie opisujące, co model powinien wykonać. W przypadku techniki RAG w skład polecenia wchodzi także kontekst, czyli tekst ukazujący szerszą perspektywę danego tematu. W samym zadaniu zawarta jest także informacja dla modelu, aby ten, podczas odpowiadania na pytanie lub wykonywania postawionego mu zadania, korzystał bezpośrednio z dostarczonego mu kontekstu - jeżeli jakaś informacja w nim nie występuje, model powinien jasno zakomunikować, że nie jest w stanie odpowiedzieć na pytanie. Nie powinien próbować wymyślać odpowiedzi jedynie na podstawie własnej wiedzy.

4.3.1. Wykonywanie akcji

W aplikacji SMIS, podobnie jak w programie, na którym wzorowana jest jej implementacja [1], każda decyzja dotycząca przebiegu rozgrywki podejmowana jest za pomocą dużego modelu językowego. W ujęciu ogólnym, na podstawie otrzymanych odpowiedzi, wykonywane są odpowiednie akcje. Kiedy rozpoczyna się dzień, dla każdego agenta generowany jest godzinowy plan dnia, bazujący na jego zainteresowaniach i jego trybie życia. Agent przeżywa dzień zgodnie z tym planem. Ponadto, przy każdym spotkaniu z inną osobą, model językowy decyduje, czy agenci powinni nawiązać konwersację. Jeśli decyzja jest pozytywna, tworzona jest konwersacja, która nawiązuje do wspomnień rozmówców oraz jest zgodna z ich osobowościami. Po zakończeniu rozmowy jest ona podsumowywana i generowane są wspomnienia dla obu agentów, które zapisywane są w ich pamięci. Dzięki temu mogą się oni do nich odwoływać w odpowiednich do tego momentach.

Przy tworzeniu polecenia dla modelu, pobierane są informacje i wspomnienia odnoszące się do konkretnego agenta. Żeby proces ten mógł nastąpić, należy podczas zapisywania wspomnienia uwzględnić trzy metryki, opisane szerzej w podrozdziale 4.2. Metryka ważności również opiera się na odpowiedzi od modelu.

4.3.2. Konwersacje

Tworzenie konwersacji jest dość skomplikowanym procesem. Aby jak najbardziej upodobnić agentów do prawdziwych ludzi, generowanie rozmowy podzielone zostało na kilka etapów.

Pierwszym z nich jest decydowanie, czy agent powinien w ogóle podjąć konwersację. Warto zauważyć, że my również nie odzywamy się do każdej miniętej na ulicy osoby - w takiej sytuacji nie udawałoby nam się nigdzie dotrzeć, gdyż cały dzień wypełniony byłby jedynie rozmowami z innymi. Podświadomie oceniamy, czy warto przywitać się z daną osobą. Czynniki, jakie są w takiej

sytuacji brane pod uwagę, to m.in. to, kim jest spotkany człowiek, czy go znamy, co w danym momencie robi, czy nie jest zajęty.

Podobnie działają agenci w aplikacji SMIS.

4.4. Komunikacja między serwisami

jak jest zrobiona komunikacja, jak to działa, może być wiele frontów na jeden serwer

5. BADANIA I ANALIZA WYNIKÓW

tekst

5.1. Ocena świadomości agentów

jak dobrze agenci generują realistyczne odpowiedzi:
- czy agenci potrafią mówić o sobie
- czy pamiętają informacje o sobie, imię czy styl życia
- czy ich wypowiedzi odwołują się do przeszłych wydarzeń

5.2. Porównanie wydajności modeli językowych

- benchmark, ogólna ocena modeli i porównanie
- porównanie różnych modeli LLM używanych w aplikacji (np llama2, llama3)
- testy: czas odpowiedzi, czas generowania, zgodnosc z kontekstem, naturalnosc dialogow, zroznicowanie odpowiedzi

5.3. Ocena jakości konwersacji przez użytkowników

- subiektywna ocena modeli
- ankieta przeprowadzona wśród studentów, którzy ocenią jakość konwersacji generowanych przez różne modele
- która konwersacja była najbardziej naturalna i spójna, + uwagi i sugestie będzie można dodać

5.4. Rozprzestrzenianie się informacji wśród agentów

- jak w papierze
- eksperymenty związane z rozprzestrzenianiem się informacji między agentami
- do ilu agentów dociera informacja gdy na początku znał ją jeden agent (np. planowany ślub albo olimpiada)
- jak ważność informacji wpływa na jej propagację (np ślub i olimpiada ważne więc powinny się rozprzestrzenic, a to ze ktoś zjadł jajecznicę na śniadanie już nie)

5.5. Wnioski i ograniczenia

- wyzwania napotkane podczas pracy z aplikacją związane z modelami
- pewnie nie zawsze będzie się wyszukiwał ddobry kontekst, problemy z retrievalem
- halucynacje

6. PODSUMOWANIE

tekst

6.1. Osiągnięte rezultaty

tekst

6.2. Plany na przyszłość

tekst

WYKAZ LITERATURY

[1] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, *Generative agents: Interactive simulacra of human behavior*, 2023. arXiv: 2304.03442 [cs.HC]. [Online]. Available: https://arxiv.org/abs/2304.03442.

WYKAZ RYSUNKÓW

WYKAZ TABEL