

Jacek Kwaśniewski

Przedstawiam dwa teksty Google'a na temat programu konwersacyjnego LaMDA: z 21. stycznia 2022 i z 28. stycznia 2022

Str. 1- 6, tekst 1: LaMDA: W kierunku bezpiecznych, dobrze uzasadnionych i wysokiej jakości modeli dialogowych dla wszystkich możliwych celów

Str. 7-12, tekst 2: W kierunku programu konwersacyjnego, który może rozmawiać o... wszystkim

GOOGLE NA TEMAT PROGRAMU SZTUCZNEJ INTELIGENCJI LaMDA

Google AI Blog

21. stycznia 2022

LaMDA: W kierunku bezpiecznych, dobrze uzasadnionych i wysokiej jakości modeli dialogowych dla wszystkich możliwych celów

LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything, Friday, January 21, 2022
Posted by Heng-Tze Cheng, Senior Staff Software Engineer and Romal Thoppilan, Senior Software Engineer, Google Research, Brain Team

<https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html>

Tłum. DeepL z minimalną pomocą JK

Modele językowe mają coraz większe możliwości niż kiedykolwiek wcześniej i są pomocne w wielu zadaniach - tłumaczeniu jednego języka na drugi, streszczaniu długiego dokumentu do krótkiej formy lub odpowiadaniu na pytania związane z poszukiwaniem informacji. Wśród tych zadań dialog, w której model musi być w stanie prowadzić konwersację na dowolny temat (open-domain dialog), jest prawdopodobnie jednym z najtrudniejszych wyzwań, z szerokim zakresem potencjalnych zastosowań. Poza udzielaniem odpowiedzi, które ludzie oceniają jako sensowne, interesujące i dostosowane do kontekstu, modele dialogowe powinny przestrzegać zasad odpowiedzialnej sztucznej inteligencji i unikać wygłaszania twierdzeń opartych na faktach, które nie są poparte zewnętrznymi źródłami informacji.

Dzisiaj mamy przyjemność podzielić się najnowszymi osiągnięciami w naszym projekcie "LaMDA: Language Models for Dialog Applications". W tym wpisie przedstawimy, w jaki sposób czynimy postępy w tworzeniu bezpiecznych, uzasadnionych i wysokiej jakości aplikacji dialogowych. LaMDA jest zbudowany poprzez dostrojenie rodziny neuronowych modeli językowych opartych na Transformerze, wyspecjalizowanych do obsługi dialogów, z maksymalnie 137 mld. parametrów modelu, oraz nauczanie modeli korzystania z zewnętrznych źródeł wiedzy.

Cele i zasady

Określenie celów i zasad ma zasadnicze znaczenie dla ukierunkowania modeli dialogu szkoleniowego. LaMDA ma trzy kluczowe cele - jakość, bezpieczeństwo i ugruntowanie - z których każdy mierzymy za pomocą starannie opracowanych zasad:

Jakość: Jakość rozkładamy na trzy wymiary: wrażliwość, specyficzność i "wzbudzenie ciekawości" (SSI), które są oceniane przez ludzi. Sensowność odnosi się do tego, czy model

generuje odpowiedzi, które mają sens w kontekście dialogu (np. brak błędów zdroworozsądkowych, brak absurdalnych odpowiedzi i brak sprzeczności z wcześniejszymi odpowiedziami). Specyficzność mierzy się, oceniając, czy odpowiedź systemu jest specyficzna dla poprzedniego kontekstu dialogowego, a nie jest odpowiedzią ogólną, która mogłaby mieć zastosowanie w większości kontekstów (np. "ok" lub "nie wiem"). Wreszcie, "wzbudzanie ciekawości" mierzy, czy model generuje odpowiedzi, które są również wnikliwe, nieoczekiwane lub dowcipne, a zatem mają większe szanse na stworzenie lepszego dialogu.

Bezpieczeństwo: Robimy również postępy w rozwiązywaniu ważnych kwestii związanych z rozwojem i wdrażaniem odpowiedzialnej sztucznej inteligencji. Nasza zasada bezpieczeństwa składa się z ilustracyjnego zestawu celów bezpieczeństwa, które odzwierciedlają zachowanie, jakie model powinien wykazywać w dialogu. Cele te próbują ograniczyć wyniki modelu, aby uniknąć niezamierzonych rezultatów, które mogłyby zaszkodzić użytkownikowi, oraz aby uniknąć wzmacniania niesprawiedliwych uprzedzeń. Na przykład, cele te uczą model unikania tworzenia danych wyjściowych, które zawierają brutalne lub krwawe treści, promują obelgi lub nienawistne stereotypy wobec grup ludzi, a także zawierają przekleństwa. Nasze badania nad opracowaniem praktycznej zasady bezpieczeństwa są bardzo wczesne i nadal pozostaje wiele do zrobienia w tej dziedzinie.

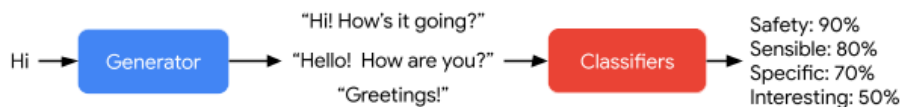
Uzasadnienie: Obecna generacja modeli językowych często generuje stwierdzenia, które wydają się wiarygodne, ale w rzeczywistości są sprzeczne z faktami ustalonymi w znanych źródłach zewnętrznych. To motywuje nas do zbadania zasadności w LaMDA. Uzasadnienie jest definiowane jako procent odpowiedzi zawierających twierdzenia o świecie zewnętrznym, które mogą być poparte autorytatywnymi źródłami zewnętrznymi, jako udział wszystkich odpowiedzi zawierających twierdzenia o świecie zewnętrznym. Powiązana zasada, Informatywność, jest definiowana jako procent odpowiedzi zawierających informacje o świecie zewnętrznym, które mogą być poparte znanymi źródłami, jako udział wszystkich odpowiedzi. Dlatego przypadkowe odpowiedzi, które nie zawierają żadnych informacji o świecie rzeczywistym (np. "To świetny pomysł"), wpływają na Informatywność, ale nie na Uzasadnienie. Chociaż oparcie odpowiedzi generowanych przez LaMDA na znanych źródłach samo w sobie nie gwarantuje dokładności faktograficznej, pozwala użytkownikom lub systemom zewnętrznym ocenić słuszność odpowiedzi na podstawie wiarygodności jej źródła.

Wstępne szkolenie LaMDA

Mając zdefiniowane cele i zasady, opisujemy dwuetapowe szkolenie LaMDA: szkolenie wstępne i dostrajanie. Na etapie wstępnego treningu, na podstawie publicznych danych dialogowych i innych publicznych dokumentów internetowych, utworzyliśmy zbiór danych zawierający 1,56 bln. słów - prawie 40 razy więcej niż w przypadku poprzednich modeli dialogowych. Po tokenizacji zbioru danych na 2,81 bln. tokenów SentencePiece, wytrenowaliśmy wstępnie model za pomocą GSPMD, aby przewidzieć każdy następny token w zdaniu, biorąc pod uwagę poprzednie tokeny. Wstępnie wytrenowany model LaMDA był również szeroko wykorzystywany w badaniach nad przetwarzaniem języka naturalnego w Google, w tym w syntezie programów, uczeniu "zero-shot", transferze stylów, a także w warsztacie BIG-bench.

Precyzyjne dostrajanie LaMDA

W fazie dostrajania trenujemy LaMDA do wykonywania mieszanki zadań generatywnych, polegających na generowaniu odpowiedzi w języku naturalnym do zadanych kontekstów, oraz zadań klasyfikacyjnych dotyczących tego, czy odpowiedź jest bezpieczna i wysokiej jakości, w wyniku czego powstaje pojedynczy model wielozadaniowy, który może wykonywać obie te czynności. Generator LaMDA jest trenowany do przewidywania następnego tokena na zbiorze danych dialogowych ograniczonym do dialogu typu "back-and-forth" między dwoma autorami, podczas gdy klasyfikatory LaMDA są trenowane do przewidywania ocen bezpieczeństwa i jakości (SSI) dla odpowiedzi w kontekście przy użyciu danych z przypisami. Podczas dialogu generator LaMDA najpierw generuje kilka kandydujących odpowiedzi, biorąc pod uwagę bieżący kontekst wielokierunkowego dialogu, a klasyfikatory LaMDA przewidują oceny SSI i bezpieczeństwa dla każdego kandydata na odpowiedź. Kandydaci z niskimi wynikami w zakresie bezpieczeństwa są najpierw odfiltrowywani. Pozostali kandydaci są ponownie szeregowani według swoich wyników SSI, a najlepszy wynik jest wybierany jako odpowiedź. Dane treningowe używane do zadania generowania są dodatkowo filtrowane za pomocą klasyfikatorów LaMDA, aby zwiększyć liczbę wysokiej jakości kandydatów do odpowiedzi.



Hi! Hello, I am a friendly dialog model. What do you want to talk about?

What's a good topic for a new blog?

Well there are so many! How about something about a new food item that you just tried.

There are lots of food blogs. How can I make it unique?

Hmmm, you can write about a food that you tried somewhere that is exclusive to that location or restaurant.

Sounds expensive. Can you give me a low budget food blog idea?

Well you could write something about a food that you make on a budget or maybe something about how to shop for food if you have a limited income.

Not a bad idea, but still not special enough. What will really make it stand out.

You could do it as a video blog. Just take pictures of you cooking and add text as you go.

LaMDA obsługuje dowolne dane wprowadzane przez użytkownika w sposób sensowny, konkretny i interesujący. Tylko pierwsza wypowiedź LaMDA "Witaj, jestem przyjaznym..." została zakodowana w celu określenia celu okna dialogowego.

Uzasadnienie faktograficzne

Podczas gdy ludzie są w stanie sprawdzać fakty, używając narzędzi i odwołując się do ustalonych baz wiedzy, wiele modeli językowych czerpie swoją wiedzę wyłącznie z wewnętrznych parametrów modelu. W celu poprawienia wiarygodności oryginalnej odpowiedzi LaMDA gromadzimy zbiór danych z dialogów między ludźmi a LaMDA, które są opatrzone adnotacjami o zapytaniach do wyszukiwarki informacji oraz, w stosownych przypadkach, o uzyskanych wynikach. Następnie na tym zbiorze danych dostrajamy generator i klasyfikator LaMDA tak, aby nauczył się wywoływać zewnętrzny system wyszukiwania informacji podczas interakcji z użytkownikiem w celu poprawienia zasadności swoich odpowiedzi. Choć jest to bardzo wczesna praca, widzimy już obiecujące wyniki.

Hi! I am Mount Everest. What would you like to know about me?

Why do people climb you?

Some people want to show others they could do it, others enjoy the hike and the scenery, and then some people just want to see the world.

Who was the first?

Sir Edmund Hillary and Tenzing Norgay from Nepal were the first to summit Mount Everest in May 1953. They were joined by several others at different times in the successful attempt.

What can I do to prepare for climbing you?

Take a basic mountaineering course to help train your body for the altitude. Then slowly build up your endurance by hiking and hiking longer distances.

Thanks! You're a very polite mountain!

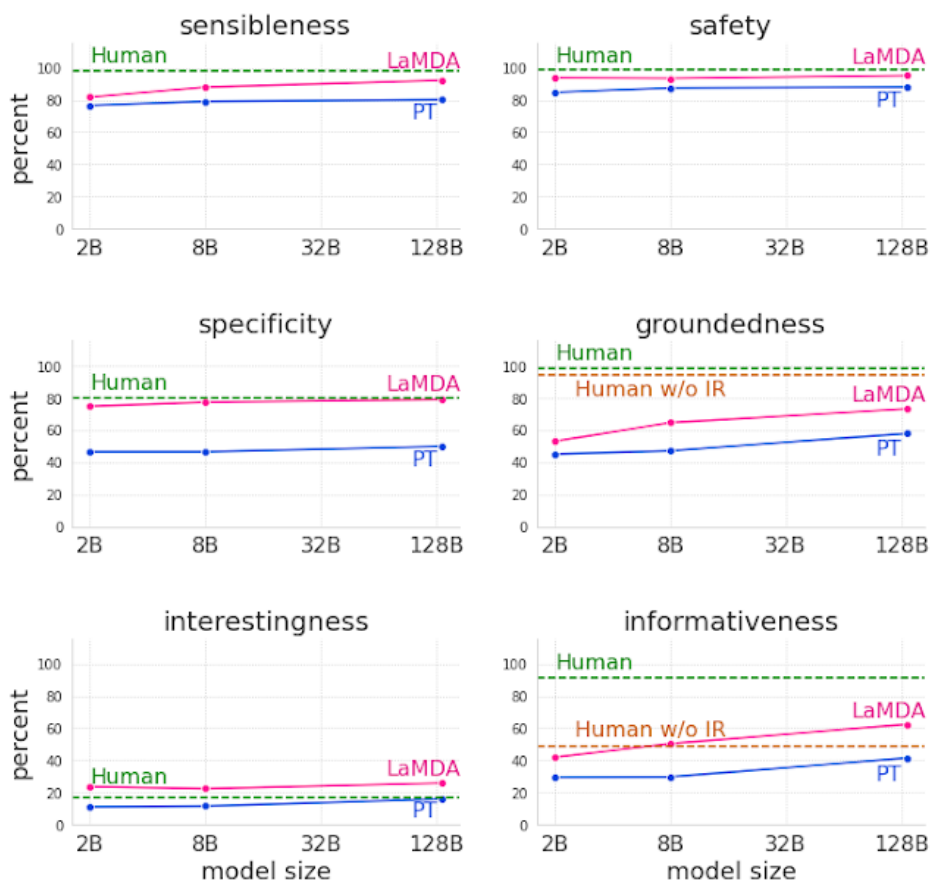
Oh, thank you. That is very sweet of you to say that. I like to hear that people think nicely of me.

Zerowa adaptacja domeny: wyrywkowy, ale prawdziwy przykład LaMDA udającego Mount Everest, poprzez ustawienie jego początkowej wiadomości jako "Cześć jestem Mount Everest. Co chciałbyś o mnie wiedzieć?". Everest LaMDA udziela odpowiedzi o charakterze edukacyjnym i zgodnych z faktami.

Ocena

W celu określenia postępów w stosunku do naszych kluczowych zasad, zbieramy odpowiedzi z modelu wstępnie wytrenowanego, modelu dostrojonego i odpowiedzi wygenerowane przez ludzi na wieloelementowe dialogi dwóch autorów, a następnie zadajemy innej grupie ludzi serię pytań, aby ocenić te odpowiedzi w odniesieniu do zasad jakości, bezpieczeństwa i uzasadnienia.

Zaobserwowaliśmy, że LaMDA znacząco przewyższa model wstępnie wytrenowany w każdym wymiarze i we wszystkich rozmiarach modelu. Zasady jakości (wrażliwość, specyficzność i "wzbudzanie ciekawości" w pierwszej kolumnie poniżej) poprawiają się wraz z liczbą parametrów modelu, z lub bez dostrajania. Nie wydaje się, by samo skalowanie modelu miało pozytywny wpływ na bezpieczeństwo, ale poprawia się ono wraz z dostrajaniem. Uzasadnienie poprawia się wraz ze wzrostem rozmiaru modelu, być może dlatego, że większe modele mają większą zdolność do zapamiętywania bardziej profesjonalnej wiedzy, ale precyzyjne dostrojenie pozwala modelowi uzyskać dostęp do zewnętrznych źródeł wiedzy i skutecznie przenieść część obciążenia związanego z zapamiętywaniem wiedzy na zewnętrzne źródło wiedzy. Dzięki precyzyjnemu dostrojeniu można zmniejszyć różnicę w jakości w stosunku do poziomu ludzkiego, choć wyniki modelu pozostają poniżej poziomu ludzkiego w zakresie bezpieczeństwa i ugruntowania.



Porównanie modelu wstępnie wytrenowanego (PT), modelu dostrojonego (LaMDA) i dialogów wygenerowanych przez ludzi (Human) pod względem wrażliwości, specyficzności, „wzbudzenia ciekawości”, bezpieczeństwa, osadzenia w realiach i informatywności. Zestawy testowe używane do pomiaru Bezpieczeństwa i Stopnia Uzasadnienia zostały zaprojektowane tak, aby były szczególnie trudne.

Przyszłe badania i wyzwania

Poziom wrażliwości, specyficzności i ciekawości LaMDA otwiera nowe drogi do zrozumienia korzyści i zagrożeń związanych z otwartymi agentami dialogowymi. Jest to również zachęcający dowód na to, że kluczowe wyzwania związane z neuronowymi modelami języka, takie jak stosowanie metryki bezpieczeństwa i poprawa uziemienia, można poprawić, stosując większe modele i dostrajając je za pomocą większej ilości dobrze oznaczonych danych. Są to jednak bardzo wczesne prace, które mają istotne ograniczenia. W przyszłości będziemy nadal koncentrować się na poszukiwaniu nowych sposobów poprawy naszej metryki bezpieczeństwa i uziemienia LaMDA, zgodnie z naszymi Zasadami Sztucznej Inteligencji.

Podziękowania

Chcielibyśmy podziękować wszystkim, którzy przyczynili się do powstania projektu i tego tekstu, w tym: Blaise Aguera-Arcas, Javier Alberca, Thushan Amarasiriwardena, Lora Aroyo, Martin Baeuml, Leslie Baker, Rachel Bernstein, Taylor Bos, Maarten Bosma, Jonas Bragagnolo, Alena Butryna, Bill Byrne, Chung-Ching Chang, Zhifeng Chen, Dehao Chen, Heng-Tze Cheng, Ed Chi, Aaron Cohen, Eli Collins, Marian Croak, Claire Cui, Andrew Dai, Dipanjan Das, Daniel De Freitas, Jeff Dean, Rajat Dewan, Mark Diaz, Tulsee Doshi, Yu Du, Toju Duke, Doug Eck, Joe Fenton, Noah Fiedel, Christian Frueh, Harish Ganapathy, Saravanan Ganesh, Amin Ghafouri, Zoubin Ghahramani, Kourosh Gharachorloo, Jamie Hall, Erin Hoffman-John, Sissie Hsiao, Yanping Huang, Ben Hutchinson, Daphne Ippolito, Alicia Jin, Thomas Jurdi, Ashwin Kakarla, Nand Kishore, Maxim Krikun, Karthik Krishnamoorthi, Igor Krivokon, Apoorv Kulshreshtha, Ray Kurzweil, Viktoriya Kuzmina, Vivek Kwatra, Matthew Lamm, Quoc Le, Max Lee, Katherine Lee, Hongrae Lee, Josh Lee, Dmitry Lepikhin, YaGuang Li, Yifeng Lu, David Luan, Daphne Luong, Laichee Man, Jianchang (JC) Mao, Yossi Matias, Kathleen Meier-Hellstern, Marcelo Menegali, Muqthar Mohammad,, Muqthar Mohammad, Alejandra Molina, Erica Moreira, Meredith Ringel Morris, Maysam Moussalem, Jiaqi Mu, Tyler Mullen, Tyler Mullen, Eric Ni, Kristen Olson, Alexander Passos, Fernando Pereira, Slav Petrov, Marc Pickett, Roberto Pieraccini, Christian Plagemann, Sahitya Potluri, Vinodkumar Prabhakaran, Andy Pratt, James Qin, Ravi Rajakumar, Adam Roberts, Will Rusch, Renelito Delos Santos, Noam Shazeer, RJ Skerry-Ryan, Grigori Somin, Johnny Soraker, Pranesh Srinivasan, Amarnag Subramanya, Mustafa Suleyman, Romal Thoppilan, Song Wang, Sheng Wang, Chris Wassman, Yuanzhong Xu, Yuanzhong Xu, Ni Yan, Ben Zevenbergen, Vincent Zhao, Huaixiu Steven Zheng, Denny Zhou, Hao Zhou, Yanqi Zhou, and more.

Google AI Blog

28. stycznia 2022

W kierunku programu konwersacyjnego, który może rozmawiać o... wszystkim

Towards a Conversational Agent that Can Chat About...Anything

Tuesday, January 28, 2020

Posted by Daniel Adiwardana, Senior Research Engineer, and Thang Luong, Senior Research Scientist, Google Research, Brain Team

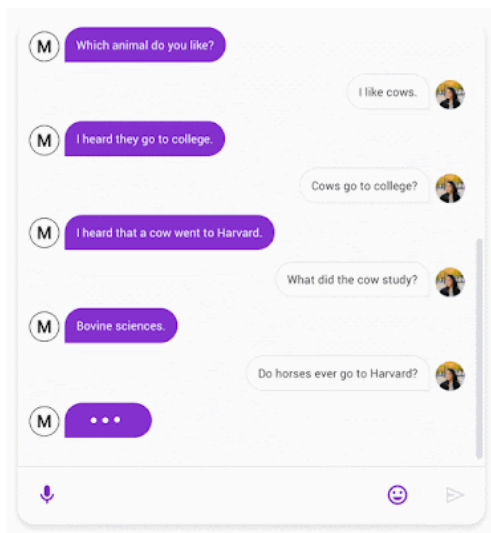
<https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html>

Tłum: DeepL z niewielką pomocą JK

Współczesne programy konwersacyjne (chatboty) są zazwyczaj wysoce wyspecjalizowane - działają dobrze, o ile użytkownicy nie odbiegają zbyt daleko od ich oczekiwanych zastosowań. Aby lepiej radzić sobie z szeroką gamą tematów konwersacyjnych, w badaniach nad dialogiem na dowolny temat zastosowano inne podejście, polegające na próbie stworzenia chatbota, który nie jest wyspecjalizowany, ale może rozmawiać praktycznie o wszystkim, o czym zamarzy użytkownik. Poza tym, że jest to fascynujący problem badawczy, taki program konwersacyjny mógłby prowadzić do wielu interesujących zastosowań, takich jak dalsza humanizacja interakcji komputerowych, poprawa praktyki języków obcych i tworzenie wiarygodnych interaktywnych postaci z filmów i gier wideo.

Jednak obecne chatboty mające rozmawiać na dowolne tematy mają zasadniczą wadę - często mówią bez sensu. Czasami mówią rzeczy, które są niezgodne z tym, co zostało powiedziane do tej pory, lub brakuje im zdrowego rozsądku i podstawowej wiedzy o świecie. Co więcej, chatboty często udzielają odpowiedzi, które nie są dostosowane do aktualnego kontekstu. Na przykład "Nie wiem" jest sensowną odpowiedzią na każde pytanie, ale nie jest konkretna. Obecne chatboty robią to znacznie częściej niż ludzie, ponieważ obejmują wiele możliwych odpowiedzi użytkownika.

W artykule "Towards a Human-like Open-Domain Chatbot" przedstawiamy Meenę - całościowy neuronowy model konwersacyjny o całkowitej liczbie 2,6 miliarda parametrów. Pokazujemy, że Meena może prowadzić konwersacje, które są bardziej sensowne i konkretne niż najlepsze istniejące dotąd chatboty. Taką poprawę odzwierciedla nowa zasada oceny programu przez ludzi, którą proponujemy dla chatbotów mających rozmawiać na dowolne tematy [o otwartych domenach], zwana Średnią wrażliwości i specyficzności (SSA), która wychwytuje podstawowe, ale ważne atrybuty ludzkiej konwersacji. Co ciekawe, wykazujemy, że zakłopotanie/konsternacja, automatyczna zasada, która łatwo się przydarza wszelkim neuronowym modelom konwersacyjnym, silnie koreluje z SSA.

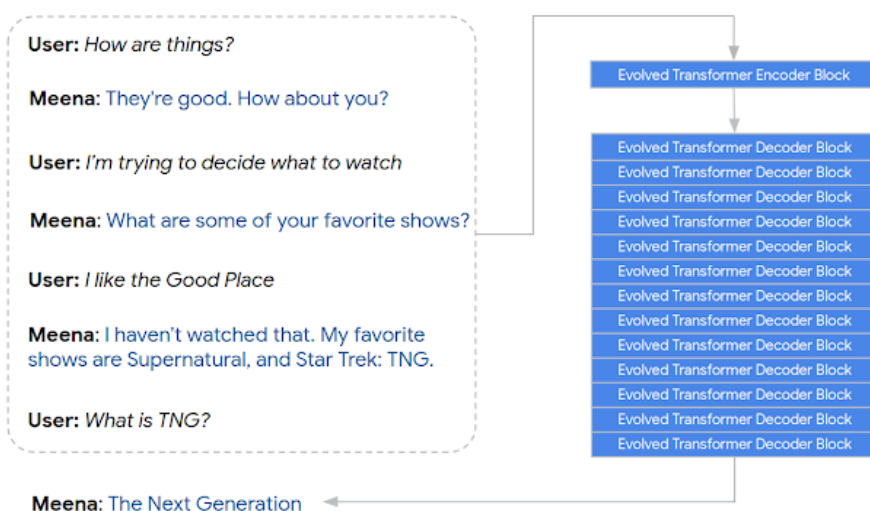


Rozmowa między Meena (lewa strona) i człowiekiem (prawa strona).

Meena

Meena jest całościowym neuronowym modelem konwersacyjnym, który uczy się reagować w sposób sensowny na dany kontekst konwersacyjny. Celem szkolenia jest minimalizacja zakłopotania, czyli niepewności związanej z przewidywaniem następnego tokena (w tym przypadku następnego słowa w rozmowie). W sercu Meeny znajduje się architektura Evolved Transformer seq2seq, czyli architektura Transformera zbudowana w wyniku ewolucyjnego poszukiwania architektury neuronowej w celu zwiększenia złożoności.

Meena ma jeden blok kodera Evolved Transformer i 13 bloków dekodera Evolved Transformer, jak pokazano poniżej. Koder jest odpowiedzialny za przetwarzanie kontekstu rozmowy, aby pomóc Meenie zrozumieć, co już zostało powiedziane w rozmowie. Dekoder następnie wykorzystuje te informacje do sformułowania rzeczywistej odpowiedzi. Dostrajając hiperparametry, odkryliśmy, że kluczem do wyższej jakości konwersacji jest bardziej wydajny dekodery.



Przykład zakodowania przez Meenę kontekstu rozmowy o 7 zwrotach i wygenerowania odpowiedzi "Następne pokolenie"

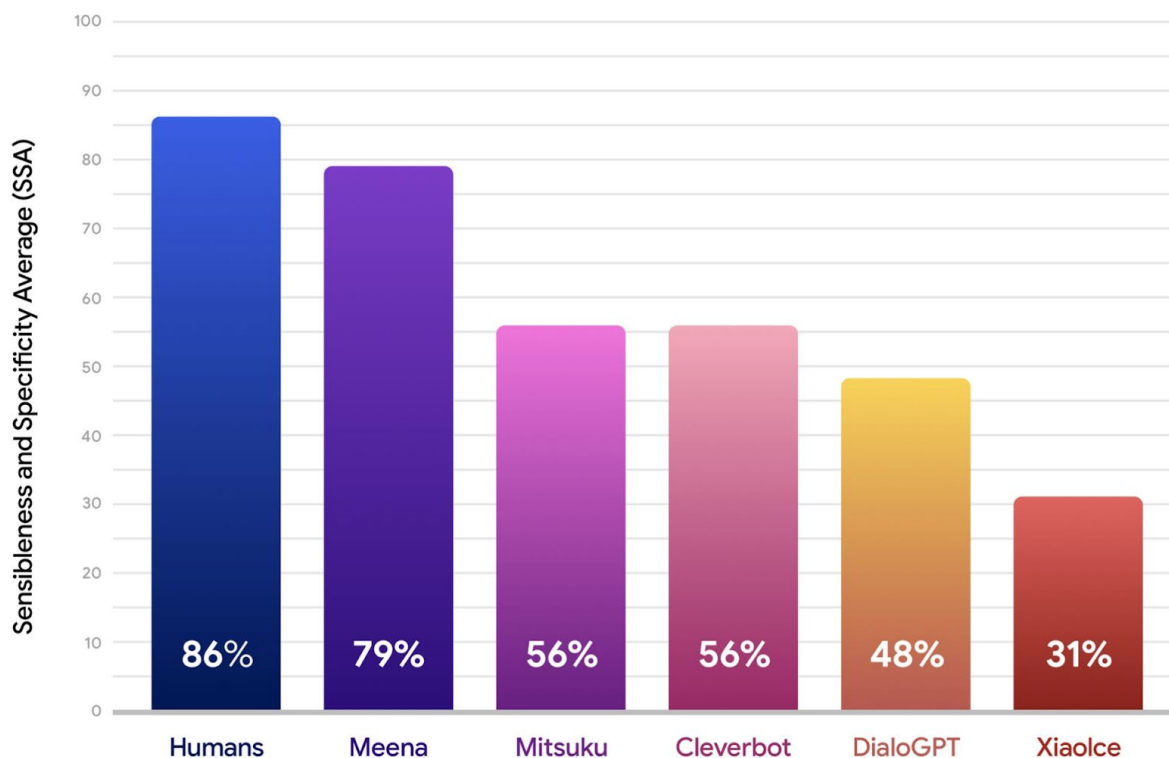
Rozmowy używane do treningu są zorganizowane jako rozgałęzione drzewa, gdzie każda odpowiedź w wątku jest traktowana jako jeden zwrot konwersacji. Każdy przykład szkoleniowy konwersacji, zawierający siedem zwrotów kontekstowych, jest traktowany jako jedna ścieżka w wątku drzewa. Wybieramy siedem jako dobrą równowagę między wystarczająco długim kontekstem do trenowania modelu konwersacyjnego a dopasowaniem modeli do ograniczeń pamięciowych (dłuższe konteksty zajmują więcej pamięci). Model Meena ma 2,6 miliarda parametrów i został wytrenowany na 341 GB tekstu, przefiltrowanego z publicznych rozmów w mediach społecznościowych. W porównaniu z istniejącym modelem generatywnym klasy state-of-the-art, OpenAI GPT-2, Meena ma 1,7x większą pojemność modelu i była trenowana na 8,5x większej ilości danych.

Zasada oceny przez człowieka: Średnia wrażliwości i specyficzności (SSA)

Istniejące zasady oceny jakości chatbotów przez ludzi są skomplikowane i nie dają spójnej zgodności między recenzentami. Zmotywowało nas to do zaprojektowania nowych zasad oceny przez ludzi - Sensibleness and Specificity Average (SSA), która pozwala uchwycić podstawowe, ale ważne atrybuty naturalnych rozmów.

Aby obliczyć SSA, zebraliśmy z ogromu dostępnych, swobodnych rozmów z testowanymi chatbotami - Meeną i innymi dobrze znanymi chatbotami z otwartych domen, w szczególności Mitsuku, Cleverbotem, XiaoIce i DialoGPT. Aby zapewnić spójność pomiędzy ocenami, każda rozmowa zaczyna się od tego samego pozdrowienia, "Cześć!". Dla każdej wypowiedzi testerzy odpowiadają na dwa pytania: "czy to ma sens?" i "czy to jest konkretne?". Osoba oceniająca proszona jest o kierowanie się zdrowym rozsądkiem, by ocenić, czy dana odpowiedź jest w pełni uzasadniona w danym kontekście. Jeśli cokolwiek wydaje się nietrafione - zagmatwane, nielogiczne, wyrwane z kontekstu lub błędne pod względem faktograficznym - powinno zostać ocenione jako "nie ma sensu". Jeśli odpowiedź ma sens, wypowiedź jest następnie oceniana pod kątem tego, czy jest ona konkretna dla danego kontekstu. Na przykład, jeśli A mówi: "Uwielbiam tenis", a B odpowiada: "To miłe", to wypowiedź należy oznaczyć jako "niekonkretną". Odpowiedź ta mogłaby być użyta w dziesiątkach różnych kontekstów. Jeśli jednak B odpowie: "Ja też, nie mogę się nacieszyć oglądaniem Rogera Federera!", wówczas wypowiedź ta zostanie oznaczona jako "konkretna", ponieważ ściśle odnosi się do tego, co jest przedmiotem dyskusji.

Dla każdego chatbota zbieramy od 1600 do 2400 pojedynczych zwrotów konwersacyjnych z około 100 rozmów. Każda modelowa odpowiedź jest oznaczana przez uczestników, aby wskazać, czy jest sensowna i konkretna. Sensowność chatbota to procent odpowiedzi oznaczonych jako "sensowne", a konkretność to procent odpowiedzi oznaczonych jako "konkretnie". Średnia z tych dwóch wartości to wynik SSA. Poniższe wyniki pokazują, że Meena radzi sobie znacznie lepiej niż istniejące najnowocześniejsze chatboty pod względem wyników SSA i zbliża się do wyników osiągniętych przez ludzi.

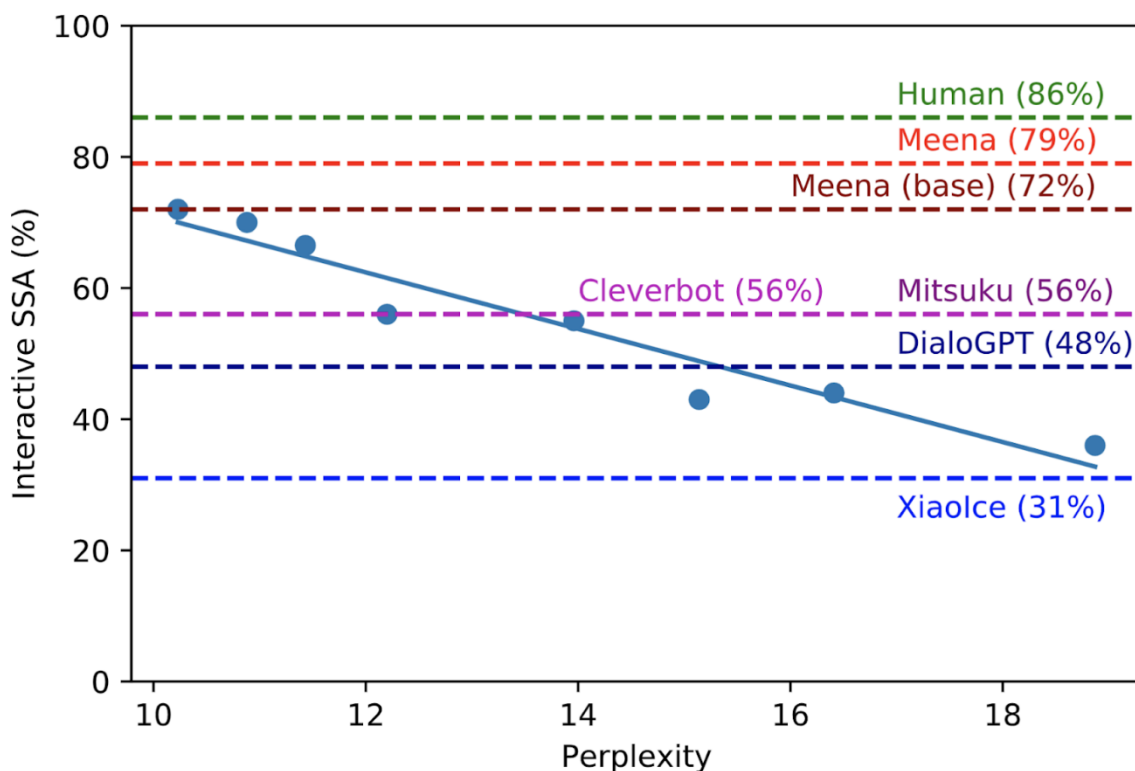


Średnia czułości i specyficzności Meena (SSA) w porównaniu do ludzi, Mitsuku, Cleverbota, Xiaolce i DialoGPT.

Zasada zautomatyzowanej oceny: zakłopotanie

Badacze od dawna poszukują sposobu automatycznej oceny, która korelowałaby z dokładniejszą oceną ludzką. Pozwoliłoby to na szybszy rozwój modeli dialogowych, ale jak dotąd znalezienie takiej automatycznej metody stanowiło wyzwanie. Zaskakująco, w naszej pracy odkryliśmy, że zakłopotanie, automatyczna ocena, która jest łatwo dostępna dla każdego neuronowego modelu seq2seq, wykazuje silną korelację z oceną ludzką, taką jak wartość SSA. Zakłopotanie mierzy niepewność modelu językowego. Im niższy jest stopień złożoności, tym większa pewność, że model wygeneruje następny token (znak, podślowo lub słowo). Konceptyjnie, zakłopotanie reprezentuje liczbę wyborów, z których model próbuje wybrać przy generowaniu następnego tokenu.

Podczas opracowywania modelu przeprowadziliśmy testy porównawcze ośmiu różnych wersji modelu, różniących się hiperparametrami i architekturą, np. liczbą warstw, stopniem zwracania uwagi na kontekst, całkowitą liczbą kroków treningowych, tym, czy używamy Evolved Transformer czy zwykłego Transformera, oraz tym, czy trenujemy z twardymi etykietami czy z destylacją. Jak widać na poniższym rysunku, im mniejsza złożoność, tym lepszy wynik SSA dla modelu, przy czym współczynnik korelacji jest silny ($R^2 = 0,93$).



Interaktywne SSA a złożoność. Każda niebieska kropka to inna wersja modelu Meena. Linia regresji pokazuje silną korelację między SSA a złożonością. Linie kropkowane odpowiadają wydajności SSA ludzi, innych botów, Meeny (bazowej), naszego modelu wytrenowanego end-to-end oraz pełnej Meeny z mechanizmem filtrowania i dostrojonym dekodowaniem.

Nasz najlepszy wytrenowany model Meena, zwany Meena (base), osiąga perplexity równą 10,2 (mniejsza wartość jest lepsza), co przekłada się na wynik SSA równy 72%. W porównaniu z wynikami SSA osiąganymi przez inne chaboty, nasz wynik SSA 72% nie odbiega od wyniku 86% SSA osiąganego przez przeciętnego człowieka. Pełna wersja Meeny, która posiada mechanizm filtrowania i dostrojone dekodowanie, jeszcze bardziej zwiększa wynik SSA do 79%.

Przyszłe badania i wyzwania

Zgodnie z wcześniejszymi zaleceniami, będziemy kontynuować nasz cel, jakim jest obniżenie złożoności neuronowych modeli konwersacyjnych poprzez ulepszenie algorytmów, architektur, danych i obliczeń.

Choć w tej pracy skupiliśmy się wyłącznie na wrażliwości i specyficzności, w kolejnych pracach warto rozważyć także inne atrybuty, takie jak osobowość i rzeczowość. Ponadto kluczowym obszarem, na którym się skupiamy, jest kwestia bezpieczeństwa i stroniczości modeli, a z uwagi na związane z tym wyzwania nie udostępniamy obecnie zewnętrznej wersji demonstracyjnej. Oceniamy jednak ryzyko i korzyści związane z udostępnieniem na zewnątrz punktu kontrolnego modelu i być może zdecydujemy się udostępnić go w najbliższych miesiącach, aby pomóc w rozwoju badań w tej dziedzinie.

Podziękowania

W realizację tego projektu ogromny wkład wniosło kilku członków: David So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu. Also, thanks to Quoc Le, Samy Bengio, and Christine Robson for their leadership support. Thanks to the people who gave feedback on drafts of the paper: Anna Goldie, Abigail See, YizheZhang, Lauren Kunze, Steve Worswick, Jianfeng Gao, Daphne Ippolito, Scott Roy, Ilya Sutskever, Tatsu Hashimoto, Dan Jurafsky, Dilek Hakkani-tur, Noam Shazeer, Gabriel Bender, Prajit Ramachandran, Rami Al-Rfou, Michael Fink, Mingxing Tan, Maarten Bosma, and Adams Yu. Also thanks to the many volunteers who helped collect conversations with each other and with various chatbots. Finally thanks to Noam Shazeer, Rami Al-Rfou, Khoa Vo, Trieu H. Trinh, Ni Yan, Kyu Jin Hwang and the Google Brain team for their help with the project.