Skoro komputery są tak inteligentne, dlaczego nie potrafią czytać?

SAMANTHA: Więc jak mogę ci pomóc?

THEODORE: Och, po prostu wszystko wydaje się niezorganizowane, to wszystko.

SAMANTHA: Nie masz nic przeciwko, jeśli przejrzę twój twardy dysk?

THEODORE: Um… okej.

SAMANTHA: OK, zacznijmy od twoich e-maili. Masz kilka tysięcy maili dotyczących LA Weekly, ale wygląda na to, że nie pracowałeś tam od wielu lat.

THEODORE: O tak. Myślę, że po prostu ratowałem te sprawy, pomyślałem, że może napisałem coś zabawnego w niektórych z nich. Ale…

SAMANTHA: Tak, jest kilka zabawnych. Powiedziałbym, że powinniśmy zapisać około osiemdziesięciu sześciu, resztę możemy skasować.

-HER (2013), NAPIS I REŻYSER SPIKE JONZE

Czy nie byłoby miło, gdyby maszyny mogły nas zrozumieć tak samo, jak Samantha („system operacyjny" z głosem Scarlett Johansson w filmie science fiction Her) rozumie Teodora? A gdyby mogli w mgnieniu oka przejrzeć nasze e-maile, wybrać to, czego potrzebujemy, a resztę odfiltrować? Gdybyśmy mogli dać komputerom jeden prezent, którego jeszcze nie mają, byłby to dar zrozumienia języka, nie tylko po to, by pomóc w organizacji naszego życia, ale także po to, by pomóc ludzkości w niektórych z naszych największych wyzwań, takich jak destylacja obszernej literatury naukowa, za którą poszczególni ludzie nie są w stanie nadążyć. W medycynie ukazuje się codziennie siedem tysięcy artykułów. Żaden lekarz ani badacz nie jest w stanie przeczytać ich wszystkich, a to jest poważną przeszkodą w postępie. Odkrywanie leków zostaje częściowo opóźnione, ponieważ wiele informacji jest zamkniętych w literaturze, której nikt nie ma czasu na przeczytanie. Czasami nowe terapie nie są stosowane, ponieważ lekarze nie mają czasu na ich czytanie i odkrywanie. Programy AI, które mogłyby automatycznie syntetyzować ogromną literaturę medyczną, byłyby prawdziwą rewolucją. Komputery, które mogłyby czytać równie dobrze jak doktoranci, ale z surową mocą obliczeniową Google, również zrewolucjonizowałyby naukę. Spodziewalibyśmy się postępów w każdej dziedzinie, od matematyki przez nauki o klimacie po nauki o materiałach. I nie tylko nauka uległaby przemianie. Historycy i biografowie mogli błyskawicznie dowiedzieć się wszystkiego, co zostało napisane o nieznanej osobie, miejscu lub wydarzeniu. Pisarze mogli automatycznie sprawdzać niespójności fabuły, luki logiczne i anachronizmy. Nawet znacznie prostsze umiejętności mogą być niezwykle pomocne. Obecnie iPhone ma taką funkcję, że gdy otrzymasz wiadomość e-mail, która umawia spotkanie, możesz ją kliknąć, a iPhone doda ją do Twojego kalendarza. To naprawdę przydatne - kiedy działa prawidłowo. Często tak się nie dzieje; iPhone dodaje spotkanie, nie w dniu, który masz na myśli, ale być może w innym dniu wymienionym w e-mailu. Jeśli nie złapiesz błędu, gdy zrobi go iPhone, może to być katastrofa. Pewnego dnia, kiedy maszyny naprawdę potrafią czytać, nasi potomkowie będą się zastanawiać, jak sobie radziliśmy bez syntetycznych czytników, tak jak będziemy się zastanawiać, jak wcześniejsze pokolenia radziły sobie bez elektryczności.

Na TED na początku 2018 roku znany futurysta i wynalazca Ray Kurzweil, obecnie pracujący w Google, ogłosił swój najnowszy projekt, Google Talk to Books, który obiecał wykorzystać rozumienie języka naturalnego, aby „zapewnić zupełnie nowy sposób odkrywania książek". Quartz posłusznie reklamował to jako „zdumiewające nowe narzędzie wyszukiwania Google, które odpowie na każde pytanie, czytając tysiące książek". Jak zwykle, pierwsze pytanie, które należy zadać, to „Co właściwie robi

program?" Odpowiedź była taka, że Google zindeksował zdania w 100 000 książek, począwszy od „Triving at College" przez „Beginning Programming for Dummies" aż po „Ewangelię według Tolkiena" i opracował wydajną metodę kodowania znaczeń zdań jako zestawów liczb znanych jako wektory. Kiedy zadasz pytanie, używa tych wektorów, aby znaleźć dwadzieścia zdań w bazie danych, które mają najbardziej podobne wektory. System nie ma pojęcia, o co właściwie prosisz. Już po zapoznaniu się z danymi wejściowymi do systemu powinno być od razu oczywiste, że twierdzenie zawarte w artykule Quartz, że Talk to Books „odpowie na każde pytanie", nie może być brane dosłownie; 100 000 książek może wydawać się dużą liczbą, ale to niewielki ułamek z ponad stu milionów, które zostały opublikowane. Biorąc pod uwagę to, co widzieliśmy wcześniej, o tym, jak bardzo głębokie uczenie się opiera się na korelacji, a nie na prawdziwym zrozumieniu, nie powinno dziwić, że wiele odpowiedzi było wątpliwych. Gdybyś zapytał na przykład o jakiś szczególny szczegół powieści, powinieneś rozsądnie oczekiwać wiarygodnej odpowiedzi. Jednak kiedy zapytaliśmy „Gdzie Harry Potter poznał Hermionę Granger?" żadna z dwudziestu odpowiedzi nie pochodziła od Harry'ego Pottera i Kamienia Filozoficznego i żadna nie odpowiadała na samo pytanie: gdzie odbyło się spotkanie. Kiedy zapytaliśmy: „Czy alianci byli usprawiedliwieni w kontynuowaniu blokady Niemiec po I wojnie światowej?" nie znalazło żadnych wyników, które nawet wspomniały o blokadzie. Odpowiadanie na „każde pytanie" to dzika przesada. A kiedy odpowiedzi nie były wyrażone bezpośrednio w frazie w indeksowanym tekście, często wszystko szło nie tak. Kiedy zapytaliśmy „Jakie było siedem horkruksów w Harrym Potterze?" nie otrzymaliśmy nawet odpowiedzi z listą, być może dlatego, że żadna z wielu książek, które omawiają Harry'ego Pottera, nie wymienia horkruksów na jednej liście. Kiedy zapytaliśmy „Kto był najstarszym sędzią Sądu Najwyższego w 1980 roku?" system zawiódł, mimo że jako człowiek możesz przejść do dowolnej internetowej listy sędziów Sądu Najwyższego (na przykład w Wikipedii) i po kilku minutach zorientujesz się, że był to William Brennan. Talk to Books ponownie się potknął, właśnie dlatego, że w żadnej ze 100 000 książek nie było zdania, które zawierałoby to w całości – „Najstarszym sędzią Sądu Najwyższego w 1980 r. był William Brennan" - i nie miało to podstaw do wyciągania wniosków, poza dosłownym tekstem. Najbardziej wymownym problemem było jednak to, że w zależności od tego, jak zadaliśmy pytanie, otrzymaliśmy zupełnie różne odpowiedzi. Gdybyśmy zapytali Talk to Books: „Kto zdradził swojego nauczyciela za 30 srebrników?", dość znany incydent w dość znanej historii, z dwudziestu odpowiedzi tylko sześć poprawnie zidentyfikowało Judasza. (Co ciekawe, dziewięć odpowiedzi dotyczyło znacznie bardziej niejasnej historii Micheasza Efraimity). Ale sytuacja pogorszyła się jeszcze bardziej, gdy odeszliśmy od dokładnego sformułowania „srebrników". Kiedy poprosiliśmy Talk to Books o nieco mniej konkretne „Kto zdradził swojego nauczyciela za 30 monet?" Judasz pojawił się tylko w 10 procentach odpowiedzi. (Najwyższej pozycji odpowiedź była zarówno nieistotna, jak i pozbawiona informacji: „Nie wiadomo, kto był nauczycielem Jingwana"). I kiedy ponownie nieco przeredagowaliśmy pytanie, tym razem zmieniając „zdradzony" na „wyprzedany", co dało „Kto wyprzedał swojego nauczyciela za 30 monet?" Judasz całkowicie zniknął z pierwszej dwudziestki wyników. Im dalej odchodziliśmy od dokładnego dopasowania zestawu słów, tym bardziej gubił się system.

Systemy odczytu maszynowego z naszych snów, kiedy się pojawią, będą w stanie odpowiedzieć w zasadzie na każde rozsądne pytanie o to, co przeczytały. Będą w stanie zebrać informacje w wielu dokumentach. A ich odpowiedzi nie składałyby się tylko z odrzucenia podkreślonych fragmentów, ale z syntezy informacji, czy to listy horkruksów, które nigdy nie pojawiły się w tym samym fragmencie, czy tego rodzaju zwięzłych enkapsulacji, których można by się spodziewać po prawniku zbierającym precedensy w wielu sprawach. lub naukowcu formułującym teorię wyjaśniającą obserwacje zebrane w wielu artykułach. Nawet pierwszoklasista może stworzyć listę wszystkich dobrych i złych facetów, którzy pojawiają się w serii książek dla dzieci. Tak jak student piszący pracę semestralną może zebrać pomysły z wielu źródeł, zweryfikować je krzyżowo i dojść do nowatorskich wniosków, tak samo

powinna być każda maszyna, która potrafi czytać. Ale zanim zdołamy zmusić maszyny do syntezy informacji, a nie tylko do ich papugowania, potrzebujemy czegoś znacznie prostszego: maszyn, które mogą niezawodnie zrozumieć nawet podstawowe teksty. Tego dnia jeszcze nie ma, chociaż niektórzy ludzie wydają się być podekscytowani sztuczną inteligencją. Aby zrozumieć, dlaczego solidne czytanie maszynowe jest w rzeczywistości wciąż dość odległą perspektywą, warto docenić - w szczegółach - to, co jest wymagane nawet do zrozumienia czegoś stosunkowo prostego, na przykład bajki dla dzieci. Załóżmy, że czytasz następujący fragment z książki „Farmer Boy" dla dzieci autorstwa Laury Ingalls Wilder (autorki „Domku na prerii"). Almanzo, dziewięcioletni chłopiec, znajduje portfel (wtedy nazywany „książką kieszonkową") pełen pieniędzy upuszczonych na ulicę. Ojciec Almanzo domyśla się, że „książka kieszonkowa" (tj. portfel) może należeć do pana Thompsona, a Almanzo znajduje pana Thompsona w jednym ze sklepów w mieście.

Almanzo zwrócił się do pana Thompsona i zapytał: „Zgubił Pan portfel?"

Pan Thompson podskoczył. Włożył rękę do kieszeni i głośno krzyknął.

"Tak, mam tysiąc pięćset dolarów w nim! Co z nim ? Co o tym wiesz?"

"Czy to jest to?" - spytał Almanzo.

„Tak, tak, to wszystko!" - powiedział pan Thompson, wyrywając portfel. Otworzył je i pospiesznie przeliczył pieniądze. Wszystkie banknoty przeliczył ponad dwa razy

Potem odetchnął z ulgą i powiedział: „Cóż, ten chłopak nic z tego nie ukradł".

Dobry system czytania powinien być w stanie odpowiedzieć na takie pytania:

· Dlaczego pan Thompson uderzył dłonią w kieszeń?

· Zanim Almanzo przemówił, czy pan Thompson zdał sobie sprawę, że zgubił swój portfel?

· Do czego odnosi się Almanzo, gdy pyta „Czy to jest to?"

· Kto prawie stracił 1500 dolarów?

· Czy wszystkie pieniądze nadal były w portfelu?

Na wszystkie te pytania ludzie mogą łatwo odpowiedzieć, ale żadna dotychczas opracowana sztuczna inteligencja nie była w stanie niezawodnie obsłużyć takich zapytań. (Pomyśl o tym, jakie kłopoty sprawiałyby im usługi Google Talk to Books). W istocie każde z tych pytań wymaga od czytelnika (człowieka lub innego) śledzenia łańcucha wniosków, które są tylko ukryte w historii. Weź pierwsze pytanie. Zanim Almanzo się odezwie, pan Thompson nie wie, że zgubił portfel i zakłada, że ma go w kieszeni. Kiedy Almanzo pyta go, czy zgubił portfel, Thompson zdaje sobie sprawę, że mógł w rzeczywistości zgubić portfel. Aby sprawdzić tę możliwość - portfel może się zgubić - Thompson uderza się w kieszeń. Ponieważ portfel nie jest tam, gdzie zwykle go trzyma, Thompson dochodzi do wniosku, że zgubił portfel. Jeśli chodzi o złożone łańcuchy rozumowania, obecna sztuczna inteligencja jest na przegranej. Takie łańcuchy rozumowania często wymagają, aby czytelnik zebrał imponujący zakres podstawowej wiedzy o ludziach i przedmiotach, a bardziej ogólnie o tym, jak działa świat, a żaden obecny system nie ma wystarczająco szerokiego zasobu wiedzy ogólnej, aby zrobić to dobrze. Weź niektóre rodzaje wiedzy, z których prawdopodobnie czerpałeś właśnie teraz, automatycznie, nawet nie będąc tego świadomym, przetrawiając historię Almanzo i portfela:

· Ludzie mogą upuszczać rzeczy, nie zdając sobie z tego sprawy. Jest to przykład wiedzy o związkach między zdarzeniami a stanami psychicznymi ludzi.

· Ludzie często noszą portfele w kieszeniach. Jest to przykład wiedzy o tym, jak ludzie zwykle używają określonych przedmiotów.

· Ludzie często noszą pieniądze w portfelach, a pieniądze są dla nich ważne, ponieważ pozwalają im płacić za rzeczy. To przykład wiedzy o ludziach, zwyczajach i ekonomii.

· Jeśli ludzie zakładają, że coś ważnego dla nich jest prawdą, a dowiadują się, że może to nie być prawdą, często pilnie próbują to zweryfikować. To jest przykład wiedzy o tym, jakie rzeczy są psychologicznie ważne dla ludzi.

· Często można dowiedzieć się, czy coś jest w kieszeni, obmacując zewnętrzną stronę kieszeni. To przykład, jak można łączyć różne rodzaje wiedzy. Tutaj wiedza o tym, jak różne przedmioty (ręce, kieszenie, portfele) oddziałują na siebie, jest połączona z wiedzą o działaniu zmysłów.

Rozumowanie wymagane w przypadku pozostałych pytań jest równie bogate. Aby odpowiedzieć na trzecie pytanie, „Do czego odnosi się Almanzo, gdy pyta „Czy to jest to?"", czytelnik musi zrozumieć coś o języku, a także o ludziach i przedmiotach, uznając, że jest to rozsądny poprzednik słów „to". " i „to" może być portfelem, ale (raczej subtelnie), że „to" odnosi się do portfela trzymanego przez Almanzo, podczas gdy „to" odnosi się do portfela, który zgubił pan Thompson. Na szczęście te dwie rzeczy (to, co trzyma Almanzo i co stracił pan Thompson) okazują się tym samym. Aby poradzić sobie nawet z prostym fragmentem, wiedza o ludziach, przedmiotach i języku musi być głęboka, szeroka i elastyczna; jeśli okoliczności są choć trochę inne, musimy odpowiednio się dostosować. Nie powinniśmy oczekiwać od pana Thompsona równie pilnej potrzeby, gdyby Almanzo powiedział, że znalazł portfel babci Almanzo. Uważamy za prawdopodobne, że pan Thompson mógł zgubić portfel, nie wiedząc o tym, ale bylibyśmy zaskoczeni, gdyby nie wiedział, że jego portfel został zabrany po tym, jak został napadnięty z nożem. Nikt jeszcze nie był w stanie wymyślić, jak sprawić, by maszyna rozumowała w tak elastyczny sposób. Nie uważamy, że jest to niemożliwe, a później naszkicujemy niektóre kroki, które należałoby podjąć, ale na razie rzeczywistość jest taka, że to, co jest wymagane, znacznie przewyższa to, co udało się osiągnąć każdemu z nas w społeczności AI. Google Talk to Books nie byłoby nawet blisko (ani czytelnicy z Microsoftu i Alibaba, o których wspomnieliśmy na samym początku książki). Zasadniczo istnieje rozbieżność między tym, w czym maszyny są teraz dobre – klasyfikowaniem rzeczy do kategorii – a rodzajem rozumowania i zrozumienia w świecie rzeczywistym, które byłyby wymagane, aby uchwycić tę przyziemną, ale krytyczną zdolność.

Praktycznie wszystko, co możesz przeczytać, stwarza podobne wyzwania. Nie ma nic szczególnego w przejściu Wildera. Oto krótki przykład z The New York Times, 25 kwietnia 2017 r.

Dzisiaj byłyby setne urodziny Elli Fitzgerald. Jedna z nowojorczyków, Loren Schoenberg, grała na saksofonie u boku „Pierwszej Damy Piosenki" w 1990 roku, pod sam koniec swojej kariery. Porównał ją do „starej butelki wina"&helip;

Każdy może łatwo odpowiedzieć na pytania zaczerpnięte bezpośrednio z tekstu (na jakim instrumencie grał Loren Schoenberg?) - ale wiele pytań wymagałoby pewnego rodzaju wnioskowania, które całkowicie wymyka się większości aktualnych systemów sztucznej inteligencji.

· Czy Ella Fitzgerald żyła w 1990 roku?

· Czy żyła w 1960 roku?

· Czy żyła w 1860 roku?

· Czy Loren Schoenberg spotkał kiedyś Ellę Fitzgerald?

· Czy Schoenberg uważa, że Fitzgerald był napojem alkoholowym?

Odpowiedź na pierwsze, drugie i trzecie pytanie polega na wnioskowaniu, że Ella urodziła się 25 kwietnia 1917 r., ponieważ 25 kwietnia 2017 r. obchodziła jej setne urodziny, a następnie wykorzystuje powszechną wiedzę, taką jak fakty, że

· Ludzie żyją podczas swojej kariery, więc żyła w 1990 roku.

· Ludzie żyją przez cały czas między narodzinami a śmiercią, ani przed narodzinami, ani po śmierci. Więc Fitzgerald musiał żyć w 1960 roku i jeszcze nie żyła w 1860 roku.

Odpowiedź na czwarte pytanie wiąże się z wnioskowaniem, że granie muzyki z kimś na ogół wiąże się z poznaniem tej osoby, i wnioskowanie, że Fitzgerald jest „Pierwszą Damą Piosenki", mimo że ta tożsamość nigdy nie jest do końca sprecyzowana. Odpowiedź na piąte pytanie wymaga wyjaśnienia, jakie rzeczy ludzie zwykle wyobrażają sobie, gdy dokonują porównań, oraz wiedzy, że Ella Fitzgerald była osobą i że ludzie nie mogą zamienić się w napoje.

Wybierz przypadkowy artykuł w gazecie, opowiadanie lub powieść dowolnej długości, a na pewno zobaczysz coś podobnego; wykwalifikowani pisarze nie mówią ci wszystkiego, mówią ci to, co musisz wiedzieć, opierając się na wspólnej wiedzy, aby wypełnić luki. (Wyobraź sobie, jak nudna byłaby historia Wilder, gdyby musiała ci powiedzieć, że ludzie trzymają portfele w kieszeniach i że ludzie czasami próbują wykryć obecność lub brak małych fizycznych przedmiotów, sięgając po nie rękami, przez kieszenie.) We wcześniejszej epoce grupa badaczy sztucznej inteligencji naprawdę próbowała rozwiązać te problemy. Peter Norvig, obecnie dyrektor ds. badań w Google, napisał prowokacyjną rozprawę doktorską na temat wyzwań związanych ze zrozumieniem historii przez maszyny. Bardziej znane jest to, że Roger Schank, pracujący wówczas w Yale, wymyślił serię wnikliwych przykładów tego, jak maszyny mogą używać „skryptów", aby zrozumieć, co się dzieje, gdy klient idzie do restauracji. Ale zrozumienie historii wymaga znacznie bardziej złożonej wiedzy i znacznie więcej form wiedzy niż scenariusze, a problem sformułowania i zebrania całej tej wiedzy był przytłaczający. Z czasem dziedzina się poddała, a badacze zaczęli pracować nad innymi, bardziej przystępnymi problemami - takimi jak wyszukiwarka internetowa i silniki rekomendacji - z których żaden nie zbliżył nas znacząco do ogólnej sztucznej inteligencji.

Wyszukiwarka internetowa oczywiście zmieniła jednak świat; to jedna z największych historii sukcesu AI. Wyszukiwarka Google, Bing i inne są niesamowicie potężnymi i fantastycznie użytecznymi elementami inżynierii, wspieranymi przez sztuczną inteligencję, które niemal natychmiast znajdują dopasowania wśród miliardów dokumentów internetowych. Być może zaskakujące jest to, że chociaż wszystkie są zasilane przez sztuczną inteligencję, nie mają prawie nic wspólnego z rodzajem zautomatyzowanego, syntetycznego odczytu maszynowego, o który apelowaliśmy. Chcemy maszyn, które potrafią zrozumieć, co czytają. Wyszukiwarki nie. Weź wyszukiwarkę Google. W algorytmie Google istnieją dwie podstawowe koncepcje, jedna stara i jedna, którą firma Google zapoczątkowała. Żadne z nich nie zależy od tego, czy system rozumie dokumenty. Pierwszy, starszy pomysł był używany w programach do wyszukiwania dokumentów od wczesnych lat sześćdziesiątych, na długo przed Google lub internetem: dopasowujesz słowa w zapytaniu do słów w dokumencie. Chcesz wyszukać

przepisy na kardamon? Nie ma problemu - po prostu znajdź wszystkie strony zawierające słowa „przepis" i „kardamon". Nie trzeba rozumieć, że kardamon jest przyprawą, nie trzeba rozumieć, jak pachnie, jak smakuje, ani wiedzieć nic o historii tego, jak jest pozyskiwany ze strąków lub w jakich kuchniach najczęściej go używa. Chcesz znaleźć instrukcje dotyczące budowania samolotów? Po prostu dopasuj kilka słów, takich jak „model", „samolot" i „jak", a otrzymasz wiele przydatnych trafień, nawet jeśli maszyna nie ma pojęcia, czym właściwie jest samolot, nie mówiąc już o tym, czym jest udźwig i opór lub powody, dla których prawdopodobnie wolałbyś latać komercyjnie niż jeździć na modelu w skali. Drugi, bardziej innowacyjny pomysł - słynny algorytm PageRank - polegał na tym, że program mógł wykorzystać zbiorową mądrość sieci do oceny, które strony internetowe są wysokiej jakości, sprawdzając, które strony otrzymały wiele linków, w szczególności linki z innych wysokiej jakości stron. Ten wgląd katapultował Google ponad wszystkie inne wyszukiwarki internetowe tamtych czasów. Ale dopasowywanie słów nie ma wiele wspólnego ze zrozumieniem tekstów, podobnie jak liczenie linków przychodzących z innych stron. Powodem, dla którego wyszukiwarka Google działa tak samo dobrze, jak bez żadnego wyrafinowanego odczytu, jest to, że wymagana jest niewielka precyzja. Wyszukiwarka nie musi wczytywać się głęboko, aby rozpoznać, czy jakiś traktat o uprawnieniach prezydenckich skłania się w lewo, czy w prawo; użytkownik może to rozgryźć. Wyszukiwarka Google musi tylko ustalić, czy dany dokument dotyczy właściwego ogólnego tematu. Zwykle można uzyskać całkiem dobre pojęcie o temacie dokumentu, po prostu patrząc na słowa i krótkie frazy, które się w nim znajdują. Jeśli ma uprawnienia „prezesa" i „wykonawcy", użytkownik prawdopodobnie będzie zadowolony z posiadania łącza; jeśli chodzi o Kardashianów, prawdopodobnie nie ma to znaczenia. Jeśli dokument wspomina „George", „Martha" i „Bitwa pod Yorktown", wyszukiwarka Google może odgadnąć, że dokument dotyczy Jerzego Waszyngtona, nawet jeśli nic nie wie o małżeństwach czy wojnach rewolucyjnych.

Google nie zawsze jest tak powierzchowny. Czasami udaje mu się zinterpretować zapytania i dać w pełni uformowane odpowiedzi, a nie tylko długie listy linków. To trochę bliższe czytaniu, ale tylko trochę, ponieważ Google na ogół czyta tylko zapytania, a nie same dokumenty. Jeśli zapytasz „Jaka jest stolica Missisipi?" Google poprawnie analizuje Twoje pytanie i wyszukuje odpowiedź („Jackson") w utworzonej wcześniej tabeli. Jeśli zapytasz „Ile wynosi 1,36 euro w rupii", parsowanie jest znowu poprawne i system może, po zapoznaniu się z inną tabelą (tym razem z kursami wymiany), poprawnie obliczyć, że „1,36 euro = 110,14 rupii indyjskich". W większości przypadków, gdy Google zwraca tego rodzaju odpowiedź, jest ona zwykle wiarygodna (system przypuszczalnie robi to tylko wtedy, gdy jego wskaźniki sugerują, że odpowiedzi są prawdopodobnie poprawne). Ale wciąż daleko mu do ideału, a błędy, które popełnia, dają dobrą wskazówkę na temat tego, co się dzieje. Na przykład w kwietniu 2018 r. zapytano wyszukiwarkę Google „Kto jest obecnie w Sądzie Najwyższym?" i otrzymano raczej niepełną odpowiedź „John Roberts", tylko jeden członek spośród dziewięciu. Jako bonus Google udostępnił listę siedmiu innych sędziów, których „ludzie również szukają": Anthony Kennedy, Samuel Alito, Clarence Thomas, Stephen Breyer, Ruth Bader Ginsburg i Antonin Scalia. Wszyscy ci ludzie byli oczywiście na dworze, ale Scalia nie żył. Następca Scalii, Neil Gorsuch, oraz niedawno mianowani Elena Kagan i Sonia Sotomayor były nieobecne na liście Google. To prawie tak, jakby Google całkowicie pominął słowo „obecnie". Wracając do naszego wcześniejszego punktu dotyczącego syntezy, ostateczny system odczytu maszynowego skompilowałby swoją odpowiedź, czytając Google News i aktualizując swoją listę, gdy pojawią się zmiany; lub przynajmniej konsultując się z Wikipedią (którą ludzie dość regularnie aktualizują) i wydobywając obecnych sędziów. Wygląda na to, że Google tego nie robi. Zamiast tego, jak najlepiej możemy powiedzieć, jest to po prostu przyglądanie się regularnościom statystycznym (Alito i Scalia pojawiają się w wielu poszukiwaniach wymiaru sprawiedliwości), a nie autentyczne czytanie i rozumienie ich źródeł. Aby wziąć inny przykład, spróbowano zapytać Google: „Kiedy zbudowano pierwszy most?" i otrzymano następujące wyniki na górze wyników: Mosty żelazne i

stalowe są dziś używane i większość głównych [sic] światowych rzek jest przecinana przez tego typu mosty. Zdjęcie przedstawia pierwszy na świecie żelazny most. Został zbudowany w Telford w 1779 roku przez Abrahama Darby'ego (trzeci) i był pierwszą dużą budowlą w historii zbudowaną z żelaza. Słowa „pierwszy" i „most" pasują do naszego zapytania, ale pierwszy zbudowany most nie był żelazny, a „pierwszy żelazny most" nie równa się „pierwszemu mostowi"; Google był wyłączony przez tysiące lat. A faktem jest, że ponad dekadę po ich wprowadzeniu, wyszukiwania, w których Google odczytuje pytanie i udziela bezpośredniej odpowiedzi, nadal pozostają w znacznej mniejszości. Kiedy otrzymujesz linki, a nie odpowiedzi, jest to generalnie znak, że Google polega tylko na słowach kluczowych i liczeniu linków, a nie na prawdziwym zrozumieniu. Firmy takie jak Google i Amazon oczywiście stale ulepszają swoje produkty i dość łatwo jest ręcznie zakodować system, aby poprawnie wymienić obecny zestaw sędziów Sądu Najwyższego; małe przyrostowe ulepszenia będą kontynuowane. To, czego nie widzimy na horyzoncie, to ogólne rozwiązanie wielu rodzajów wyzwań, które podnieśliśmy. Kilka lat temu zobaczyliśmy sprytnego mema na Facebooku: zdjęcie Baracka Obamy z podpisem „W zeszłym roku powiedziałeś nam, że masz 50 lat; teraz mówisz, że masz 51 lat. Co to jest, Barack Obama? Dwie różne wypowiedzi, wypowiedziane w różnym czasie, mogą być prawdziwe. Jeśli jesteś człowiekiem, rozumiesz żart. Jeśli jesteś maszyną zajmującą się niewiele więcej niż dopasowywaniem słów kluczowych, jesteś zgubiony.

A co z „wirtualnymi asystentami", takimi jak Siri, Cortana, Asystent Google i Alexa? Plusem jest to, że często podejmują działania, a nie tylko podają listę linków; w przeciwieństwie do wyszukiwarki Google, zostały one od początku zaprojektowane tak, aby interpretować zapytania użytkowników nie jako zbiory losowych słów kluczowych, ale jako rzeczywiste pytania. Ale po kilku latach wszystkie są chybione, skuteczne w niektórych dziedzinach i słabe w innych. Na przykład, wszyscy są całkiem dobrzy w pytaniach „faktoidów" - „Kto wygrał World Series w 1957?"; każdy z nich ma również kieszenie o wyraźnej sile. Asystent Google jest dobry w udzielaniu wskazówek i kupowaniu biletów do kina. Siri jest dobra w udzielaniu wskazówek i dokonywaniu rezerwacji. Alexa jest dobra w matematyce, całkiem przyzwoita w opowiadaniu gotowych dowcipów i (nic dziwnego) dobra w zamawianiu rzeczy od Amazona. Ale poza ich szczególnymi obszarami siły nigdy nie wiadomo, czego się spodziewać. Niedawno pisarka Mona Bushnell próbowała zapytać wszystkie cztery programy o drogę do najbliższego lotniska. Asystent Google dał jej listę biur podróży. Siri dał jej wskazówki do bazy hydroplanów. Cortana dała jej listę stron z biletami lotniczymi, takich jak Expedia. Podczas niedawnej przejażdżki, Alexa uzyskała 100% punktów w pytaniach takich jak Czy Donald Trump jest osobą?, Czy Audi jest pojazdem? i Czy Edsel jest pojazdem?, ale zbombardowane pytaniami typu Czy Audi może używać gazu? , Czy Audi może przejechać z Nowego Jorku do Kalifornii? i Czy rekin to pojazd? Albo weźmy ten przykład, wysłany niedawno do Gary'ego na Twitterze: zrzut ekranu z czyjąś próbą poproszenia Siri o „najbliższą restaurację fast food, która nie była McDonald's". Siri posłusznie wymyśliła listę trzech pobliskich restauracji i wszystkie serwowały fast foody - ale każda z nich była McDonaldem; słowo „nie" zostało całkowicie zlekceważone. WolframAlpha, wprowadzony w 2009 roku jako „pierwszy na świecie silnik wiedzy obliczeniowej", nie jest lepszy. Ma ogromne wbudowane bazy danych wszelkiego rodzaju informacji naukowych, technologicznych, matematycznych, spisowych i socjologicznych, a także zbiór technik wykorzystywania tych informacji do odpowiadania na pytania, ale jego zdolność do łączenia wszystkich tych informacji jest wciąż niepewna. Jego siłą są pytania matematyczne typu „Jaka jest waga stopy sześciennej złota?" „Jak daleko jest Biloxi w stanie Missisipi od Kalkuty?" oraz „Jaka jest objętość dwudziestościanu o długości krawędzi 2,3 metra?" (odpowiednio „547 kg", „8781 mil" i „26,5 m,sup>3</sup>"). Ale granice jego zrozumienia nie są trudne do osiągnięcia. Jeśli zapytasz „Jak daleko jest granica Meksyku od San Diego?" dostajesz „1144 mile", co jest całkowicie błędne. WolframAlpha ignoruje słowo „granica" i zamiast tego zwraca odległość od San Diego do geograficznego centrum Meksyku. Jeśli nieco przeformułujesz pytanie o objętość

dwudziestościanu, zastępując słowa „o krawędziach o długości 2,3 metra" słowami „którego krawędzie mają 2,3 metra długości", WolframAlpha nie rozpoznaje już, że pytanie dotyczy objętości; wszystko, co otrzymujesz, to ogólna informacja, że dwudziestościany mają 30 krawędzi, 20 wierzchołków i 12 ścian, bez wzmianki o objętości. WolframAlpha może ci powiedzieć, kiedy urodziła się Ella Fitzgerald i kiedy umarła; ale jeśli zapytasz „Czy Ella Fitzgerald żyła w 1960 roku?", błędnie zinterpretuje to pytanie jako „Czy Ella Fitzgerald żyje?" i odpowiada „Nie". Ale czekaj, mówisz, co z Watsonem, który był tak dobry w odpowiadaniu na pytania, że pokonał dwóch ludzkich mistrzów w Jeopardy! To prawda, ale niestety Watson nie jest tak ogólnie potężny, jak mogłoby się wydawać. Prawie 95 procent odpowiedzi Jeopardy! jak się okazuje, są tytułami stron Wikipedii. Zwycięstwo w Jeopardy! to często tylko kwestia znalezienia odpowiedniego artykułu. Od tego rodzaju inteligentnego wyszukiwania informacji do systemu, który potrafi naprawdę myśleć i rozumować, jest daleka droga. Do tej pory IBM nawet nie przekształcił Watsona w solidnego wirtualnego asystenta. Kiedy ostatnio szukaliśmy czegoś takiego na stronie IBM, wszystko, co mogliśmy znaleźć, to przestarzałe demo Watson Assistant, które koncentrowało się wąsko na symulowanych samochodach, w żaden sposób nie dorównując bardziej wszechstronnym ofertom Apple, Google, Microsoft, lub Amazon. Wirtualni asystenci, tacy jak Siri i Alexa, z pewnością zaczynają być użyteczni, ale mają przed sobą długą drogę. I, co najważniejsze, podobnie jak w przypadku wyszukiwarki Google, zachodzi bardzo niewielka synteza. O ile wiemy, bardzo niewielu z nich kiedykolwiek próbuje połączyć informacje w elastyczny sposób z wielu źródeł, a nawet z jednego źródła z wieloma zdaniami, tak jak robiłeś to wcześniej, gdy czytałeś o Almanzo i o Elli Fitzgerald. Prawda jest taka, że żaden obecny system AI nie może powielić tego, co zrobiłeś w tych przypadkach, integrując serię zdań i rekonstruując zarówno to, co zostało powiedziane, jak i to, co nie zostało powiedziane. Jeśli podążasz za tym, co mówimy, jesteś człowiekiem, a nie maszyną. Pewnego dnia możesz poprosić Alexę o porównanie relacji prezydenta z The Wall Street Journal z relacjami z The Washington Post lub zapytać, czy twój lekarz rodzinny mógł przeoczyć coś w twoich najnowszych listach przebojów, ale na razie to tylko fantazja. Lepiej trzymaj się pytania Alexy o pogodę. Co nam zostaje? Mieszanina wirtualnych asystentów, często przydatnych, nigdy w pełni niezawodnych - z których żaden nie jest w stanie zrobić tego, co my, ludzie, za każdym razem, gdy czytamy książkę. Sześćdziesiąt lat w historii sztucznej inteligencji komputery wciąż są funkcjonalnymi analfabetami.

Głębokie uczenie nie rozwiąże tego problemu, podobnie jak ściśle powiązany trend uczenia się „od końca do końca", w którym sztuczna inteligencja jest szkolona do przekształcania danych wejściowych bezpośrednio w wyniki, bez żadnych pośrednich podsystemów. Na przykład, podczas gdy tradycyjne podejście do prowadzenia pojazdów dzieliłoby rzeczy na podsystemy, takie jak percepcja, przewidywanie i podejmowanie decyzji (być może wykorzystujące uczenie głębokie jako element w niektórych z tych podsystemów), system kompleksowy zrezygnowałby z podsystemów zamiast tego zbudował system cardrivingu, który pobiera obrazy z kamer jako dane wejściowe i zwraca jako dane wyjściowe, korekty przyspieszenia i sterowania - bez żadnych pośrednich podsystemów do określania, gdzie znajdują się różne obiekty i jak się poruszają, czego można oczekiwać od innych kierowców rób i nie rób i tak dalej. Kiedy to działa, może być bardzo skuteczne i prostsze do wdrożenia niż bardziej ustrukturyzowane alternatywy; systemy typu end-to-end często wymagają stosunkowo niewielkiej pracy ludzkiej. A czasami są najlepszym dostępnym rozwiązaniem. Jak wyjaśniono w artykule New York Times Magazine w Tłumaczu Google, kompleksowe systemy uczenia głębokiego znacznie poprawiły stan wiedzy w zakresie tłumaczenia maszynowego, zastępując wcześniejsze podejścia. W dzisiejszych czasach, jeśli chcesz zbudować program do, powiedzmy, tłumaczeń między francuskim i angielskim, zacząłbyś od zebrania ogromnego zbioru dokumentów, które istnieją zarówno w wersji francuskiej, jak i angielskiej, zwanych bitexts (wymawiane „bye-texts"), jak obrady parlamentu kanadyjskiego, które zgodnie z prawem muszą być publikowane w obu językach. Na podstawie takich danych Tłumacz Google może automatycznie uczyć się powiązań między angielskimi słowami i frazami a ich francuskimi

odpowiednikami, bez wcześniejszej znajomości języka francuskiego lub angielskiego ani żadnej wcześniejszej wiedzy o zawiłościach gramatyki francuskiej. Nawet sceptycy są zdumieni. Problem w tym, że jeden rozmiar nie pasuje do wszystkich. Tłumaczenie maszynowe okazuje się niezwykle dobrym rozwiązaniem dla metod typu end-to-end, częściowo ze względu na łatwą dostępność dużej ilości odpowiednich danych, a częściowo dlatego, że generalnie istnieje mniej lub bardziej wyraźna zgodność między angielskimi słowami i francuskie słowa. (W większości przypadków właściwe francuskie słowo jest jedną z opcji, które można znaleźć w słowniku francusko-angielskim, a przez większość czasu relacje między kolejnością słów w dwóch językach mają dość standardowe wzorce.) Ale wiele innych aspektów rozumienia języka jest znacznie słabiej dopasowanych. Odpowiadanie na pytania jest znacznie bardziej otwarte, po części dlatego, że słowa w prawidłowej odpowiedzi na pytanie mogą nie mieć oczywistego związku ze słowami tekstu. Tymczasem nie istnieje baza danych pytań i odpowiedzi o rozmiarach porównywalnych z francusko-angielskim postępowaniem parlamentarnym. Nawet gdyby tak było, wszechświat pytań i odpowiedzi jest tak ogromny, że jakakolwiek baza danych byłaby zaledwie malutką próbką wszystkich możliwości. Jak wyjaśniono wcześniej, stwarza to poważne problemy dla głębokiego uczenia się: im bardziej system głębokiego uczenia musi odchodzić od swojego zestawu szkoleniowego, tym więcej wpada w kłopoty. I, prawdę mówiąc, nawet w przypadku tłumaczenia maszynowego, kompleksowe podejście wciąż ma swoje ograniczenia. Często (ale nie zawsze) są w porządku, jeśli chodzi o zrozumienie, ale dopasowanie słów, fraz i tak dalej nie zawsze wystarcza. Gdy uzyskanie właściwego tłumaczenia zależy od głębszego zrozumienia, systemy się psują. Jeśli podasz Tłumaczowi Google francuskie zdanie „Je mange un avocat pour le déjeuner", co w rzeczywistości oznacza „jem awokado na lunch", otrzymasz tłumaczenie „jem prawnika na lunch". Francuskie słowo avocat oznacza zarówno „awokado", jak i „prawnik", a ponieważ ludzie dużo częściej piszą o prawnikach niż o awokado (szczególnie w obradach kanadyjskiego parlamentu), Tłumacz Google ma częstsze znaczenie, poświęcając sens statystyce . We wspaniałym artykule w The Atlantic Douglas Hofstadter opisał ograniczenia Tłumacza Google:

My, ludzie, wiemy wiele rzeczy na temat par, domów, rzeczy osobistych, dumy, rywalizacji, zazdrości, prywatności i wielu innych niematerialnych, które prowadzą do takich dziwactw, jak para małżeńska mająca ręczniki wyhaftowane „jego" i „jej". Tłumacz Google nie zna takich sytuacji. Tłumacz Google nie zna sytuacji, kropka. Zna tylko ciągi złożone ze słów złożonych z liter. Chodzi o ultraszybkie przetwarzanie fragmentów tekstu, a nie o myślenie, wyobrażanie sobie, zapamiętywanie lub rozumienie. Nawet nie wie, że słowa oznaczają rzeczy.

Mimo całego poczynionego postępu większość pisemnej wiedzy na świecie pozostaje zasadniczo niedostępna, nawet jeśli jest cyfrowa i internetowa, ponieważ jest w formie, której maszyny nie rozumieją. Na przykład elektroniczna dokumentacja zdrowotna jest wypełniona tak zwanym tekstem nieustrukturyzowanym, takimi jak notatki lekarskie, e-maile, artykuły prasowe i dokumenty z edytora tekstu, które nie pasują do tabeli. Prawdziwy system odczytu maszynowego byłby w stanie zanurkować, przeglądając notatki lekarza w poszukiwaniu ważnych informacji, które są rejestrowane w badaniach krwi i zapisach przyjęć. Ale problem jest tak daleko poza tym, co może zrobić obecna sztuczna inteligencja, że notatki wielu lekarzy nigdy nie są szczegółowo czytane. Narzędzia AI dla radiologii zaczynają być badane; są w stanie patrzeć na obrazy i odróżnić guzy od zdrowych tkanek, ale nie mamy jeszcze możliwości zautomatyzowania innej części tego, co robi prawdziwy radiolog, czyli łączenia obrazów z historią pacjenta. Zdolność do zrozumienia nieustrukturyzowanego tekstu jest obecnie poważnym wąskim gardłem w szerokim zakresie potencjalnych komercyjnych zastosowań sztucznej inteligencji. Nie potrafimy jeszcze zautomatyzować procesu czytania umów prawnych, artykułów naukowych czy raportów finansowych, ponieważ każdy z nich składa się z fragmentu tekstu, którego AI wciąż nie jest w stanie ogarnąć. Chociaż obecne narzędzia automatycznie wyciągają podstawowe informacje z nawet najtrudniejszego tekstu, duża część treści jest zazwyczaj pomijana.

Coraz bardziej wyszukane wersje dopasowywania tekstu i liczenia linków pomagają - trochę - ale po prostu nie prowadzą nas do programów, które naprawdę potrafią czytać i rozumieć. Oczywiście sytuacja nie jest lepsza dla rozumienia języka mówionego (czasami nazywanego rozumieniem dialogu). Jeszcze większe wyzwania staną przed skomputeryzowanym asystentem lekarza, który będzie próbował przełożyć mowę na notatki medyczne (aby lekarze mogli spędzać więcej czasu z pacjentami, a mniej na swoich laptopach). Rozważ ten prosty fragment dialogu, który przesłał nam dr Vik Moharir:

LEKARZ: Czy odczuwasz ból w klatce piersiowej przy jakimkolwiek wysiłku?

PACJENT: Cóż, w zeszłym tygodniu ścinałem podwórko i czułem się, jakby siedział na mnie słoń. [Wskazując na klatkę piersiową]

Dla osoby jest oczywiste, że odpowiedź na pytanie lekarza brzmi „tak"; cięcie podwórka należy do kategorii wysiłku i wnioskujemy, że pacjent doświadczył bólu na podstawie naszej wiedzy, że słonie są ciężkie, a zmiażdżenie przez ciężkie przedmioty jest bolesne. Automatycznie wnioskujemy również, że słowo „czuć" jest używane w przenośni, a nie dosłownie, biorąc pod uwagę ilość szkód, jakie zadałby słoń. Dla maszyny, o ile wcześniej nie było dużo konkretów o słoniach, to prawdopodobnie tylko trochę gadania o dużych ssakach i pracach na podwórku.

Jak wpadliśmy w ten bałagan? Głębokie uczenie jest bardzo skuteczne w uczeniu się korelacji, takich jak korelacje między obrazami lub dźwiękami a etykietami. Jednak głębokie uczenie się ma trudności ze zrozumieniem, w jaki sposób przedmioty, takie jak zdania, odnoszą się do ich części (takich jak słowa i frazy). Czemu? Brakuje w nim tego, co językoznawcy nazywają kompozycyjnością: sposobu konstruowania znaczenia zdania złożonego ze znaczenia jego części. Na przykład w zdaniu Księżyc znajduje się 240 000 mil od Ziemi, słowo księżyc oznacza jeden konkretny obiekt astronomiczny, ziemia oznacza inny, mila oznacza jednostkę odległości, 240 000 oznacza liczbę, a następnie, ze względu na sposób wyrażenia a zdania działają kompozycyjnie w języku angielskim, 240 000 mil oznacza określoną długość, a zdanie Księżyc znajduje się 240 000 mil od ziemi zapewnia, że odległość między dwoma ciałami niebieskimi jest właśnie tą długością. Co zaskakujące, uczenie głębokie nie ma tak naprawdę żadnego bezpośredniego sposobu radzenia sobie z kompozycyjnością; ma po prostu mnóstwo odosobnionych bitów informacji zwanych funkcjami, bez żadnej struktury. Może nauczyć się, że psy w kapeluszu mają ogony i nogi, ale nie wie, jak odnoszą się one do cyklu życia psa. Głębokie uczenie nie rozpoznaje psa jako zwierzęcia złożonego z części takich jak głowa, ogon i cztery nogi, a nawet tego, czym jest zwierzę, nie mówiąc już o tym, czym jest głowa i jak koncepcja głowy różni się w zależności od żaby, psa i ludzi, różniących się szczegółami, ale mających wspólny stosunek do ciał. Głębokie uczenie nie rozpoznaje również, że zdanie takie jak Księżyc znajduje się 240 000 mil od ziemi zawiera zwroty, które odnoszą się do dwóch ciał niebieskich i długości. Weźmy inny przykład, kiedy poprosiliśmy Tłumacza Google o przetłumaczenie „Elektryk, do którego dzwoniliśmy, aby naprawić telefon, pracuje w niedziele" na francuski, otrzymaliśmy odpowiedź L'électricien que nous avons appelé pour réparer le téléphone fonctionne le dimanche. Jeśli znasz francuski, wiesz, że to nie jest w porządku. W szczególności słowo działa ma dwa tłumaczenia w języku francuskim: travaille, co oznacza trudy i fonctionne, co oznacza, że działa prawidłowo. Google użyło słowa fonctionne zamiast travaille, nie pojmując, jak człowiek by to zrobił, że „praca w niedziele" to coś, co w kontekście odnosi się do elektryka i że jeśli mówisz o osobie pracującej, powinieneś używać czasownik travaille. Pod względem gramatycznym podmiotem czasownika jest tutaj elektryk, a nie telefon. Znaczenie zdania jako całości jest funkcją tego, jak części są połączone, a Google tak naprawdę tego nie rozumie. Jego sukces w wielu przypadkach skłania nas do myślenia, że system rozumie więcej niż w rzeczywistości, ale prawda (po raz kolejny ilustruje iluzoryczną lukę w postępie) jest taka, że w jego tłumaczeniach jest bardzo mało głębi. Pokrewnym i nie mniej krytycznym problemem jest to, że głębokie uczenie nie ma dobrego sposobu na włączenie wiedzy podstawowej, co widzieliśmy wcześniej w rozdziale 3. Jeśli uczysz się

powiązać obraz z etykietą, nie ma znaczenia, w jaki sposób Ty to zrób. Dopóki to działa, nikt nie przejmuje się wewnętrznymi szczegółami systemu, ponieważ liczy się tylko to, że masz odpowiednią etykietę dla danego obrazu. Całe zadanie jest często stosunkowo odizolowane od większości tego, co znasz. Język prawie nigdy taki nie jest. Praktycznie każde zdanie, które napotykamy, wymaga wyciągania wniosków na temat tego, jak szeroki zakres wiedzy podstawowej jest powiązany z tym, co czytamy. Głębokiemu uczeniu brakuje bezpośredniego sposobu przedstawienia tej wiedzy, nie mówiąc już o wnioskowaniu na jej temat w kontekście rozumienia zdania. I wreszcie, uczenie głębokie dotyczy statycznych tłumaczeń, od danych wejściowych do etykiety (obraz kota do etykiety kota), ale czytanie jest procesem dynamicznym. Kiedy używasz statystyk, aby przetłumaczyć historię, która zaczyna się Je mange une pomme na „Jem jabłko", nie musisz wiedzieć, co oznacza każde zdanie, jeśli potrafisz rozpoznać, że w poprzednich bitextach je było kojarzone z „I, mange with eat" , une z i pomme z jabłkiem. W większości przypadków program do tłumaczenia maszynowego może wymyślić coś pożytecznego, po prostu przerzucając jedno zdanie na raz, bez zrozumienia znaczenia fragmentu jako całości. Kiedy czytasz opowiadanie lub esej, robisz coś zupełnie innego. Twoim celem nie jest stworzenie kolekcji statystycznie prawdopodobnych dopasowań; ma zrekonstruować świat, którym autor próbował się z tobą podzielić. Kiedy czytasz historię Almanzo, możesz najpierw zdecydować, że historia zawiera trzech głównych bohaterów (Almanzo, jego ojciec i pan Thompson), a następnie zaczynasz uzupełniać niektóre szczegóły dotyczące tych postaci (Almanzo jest chłopcem , jego ojciec jest dorosły itp.), a także zaczynasz próbować ustalić niektóre z wydarzeń, które miały miejsce (Almanzo znalazł portfel, Almanzo zapytał pana Thompsona, czy portfel należy do niego i tak dalej). Robisz coś podobnego (w dużej mierze nieświadomie) za każdym razem, gdy wchodzisz do pokoju, oglądasz film lub czytasz historię. Ty decydujesz, jakie istoty tam są, jaki jest ich związek między sobą i tak dalej. Używając języka psychologii poznawczej, czytając dowolny tekst, budujesz poznawczy model znaczenia tego, co mówi tekst. Może to być tak proste, jak skompilowanie tego, co Daniel Kahneman i nieżyjąca już Anne Treisman nazwali plikiem obiektowym – zapisem pojedynczego obiektu i jego właściwości – lub tak złożone, jak pełne zrozumienie skomplikowanego scenariusza. Kiedy czytasz fragment z Farmer Boy, stopniowo budujesz mentalną reprezentację — wewnętrzną w twoim mózgu — wszystkich ludzi, przedmiotów i zdarzeń z historii oraz relacji między nimi: Almanzo, portfel i pan Thompson a także wydarzenia, w których Almanzo rozmawiał z panem Thompsonem, pan Thompson krzyczał i klepał się po kieszeni, pan Thompson wyrywał portfel Almanzo i tak dalej. Dopiero po przeczytaniu tekstu i skonstruowaniu modelu poznawczego robisz wszystko, co robisz z narracją – odpowiadasz na pytania na jej temat, tłumaczysz na rosyjski, streszczasz, parodiujesz, ilustrujesz lub po prostu pamiętasz na później. Google Translate, plakatowe dziecko wąskiej sztucznej inteligencji, omija cały proces budowania i używania modelu poznawczego; nigdy nie musi niczego rozumować ani śledzić; robi to, co robi całkiem dobrze, ale obejmuje tylko najmniejszy wycinek tego, o czym naprawdę jest czytanie. Nigdy nie buduje modelu poznawczego opowieści, bo nie może. Nie można zapytać systemu głębokiego uczenia „co by się stało, gdyby pan Thompson szukał portfela i znalazł wybrzuszenie w miejscu, w którym spodziewał się znaleźć portfel", ponieważ nie jest to nawet część paradygmatu. Statystyki nie zastąpią zrozumienia w świecie rzeczywistym. Problem nie polega tylko na tym, że tu czy tam występuje błąd losowy, ale na tym, że istnieje fundamentalna rozbieżność między rodzajem analizy statystycznej, która jest wystarczająca do tłumaczenia, a konstrukcją modelu poznawczego, która byłaby wymagana, gdyby systemy rzeczywiście pojęły to, co mają. próbują czytać.

Jednym z zaskakująco trudnym wyzwaniem dla głębokiego uczenia się (choć nie w przypadku klasycznych podejść do sztucznej inteligencji) jest po prostu zrozumienie słowa „nie". Pamiętasz porażkę Siri z „Znajdź restaurację typu fast food, która nie jest McDonaldem"? Osoba zadająca pytanie prawdopodobnie chciała uzyskać odpowiedź w stylu „The Burger King przy 321 Elm Street, Wendy przy 57 Main Street i IHOP przy 523 Spring Street". Ale nie ma nic w Wendy's, Burger King czy IHOP, co jest

szczególnie kojarzone ze słowem „nie” i nie zdarza się tak często, że ktoś odnosi się do któregoś z nich jako nie do McDonalda, więc brutalne statystyki nie pomagają zrobiliby z pokrewnym królem i królową. Można sobie wyobrazić kilka statystycznych sztuczek w celu rozwiązania tego konkretnego problemu (identyfikacja restauracji), ale pełne potraktowanie wszystkich sposobów, w których nie można użyć, wykracza daleko poza zakres obecnych podejść. To, czego naprawdę potrzebuje ta dziedzina, to podstawa tradycyjnych operacji obliczeniowych, z których zbudowane są bazy danych i klasyczna sztuczna inteligencja: budowanie listy (restauracje typu fast food w określonej okolicy), a następnie wykluczanie elementów, które należą do innej listy (lista różnych franczyz McDonald’s). Ale głębokie uczenie zostało zbudowane wokół unikania właśnie tego rodzaju obliczeń. Listy są podstawowe i wszechobecne w programach komputerowych i istnieją od ponad pięćdziesięciu lat (pierwszy główny język programowania AI, LISP, został dosłownie zbudowany wokół list), a jednak nie są one nawet częścią struktury głębokiego uczenia się. Zrozumienie zapytania ze słowem nieobecnym staje się zatem ćwiczeniem polegającym na wbijaniu kwadratowych kołków w okrągłe otwory.

A do tego dochodzi problem niejednoznaczności. Języki ludzkie są pełne niejasności. Słowa mają wiele znaczeń: praca (jako czasownik) może oznaczać albo pracę. A te są stosunkowo jasne; Lista wszystkich różnych znaczeń słów, takich jak in lub take, wypełnia wiele kolumn dobrego słownika. Rzeczywiście, większość słów, z wyjątkiem bardzo technicznych, ma wiele znaczeń. Również struktura gramatyczna fraz jest często niejednoznaczna. Czy zdanie „Ludzie mogą łowić” oznacza, że ludzie mogą łowić ryby, czy też (jak w Steinbeck’s Cannery Row) ludzie pakują sardynki i tuńczyka do puszek? Słowa takie jak zaimki często wprowadzają dalsze niejasności. Jeśli powiesz, że Sam nie mógł podnieść Harry’ego, ponieważ był zbyt ciężki, to w zasadzie mógłby być Samem lub Harrym. To, co jest niesamowite w nas, ludziach, czytelnikach, to to, że w 99 procentach czasu nawet nie zauważamy tych niejasności. Zamiast się pogubić, szybko i bez świadomego wysiłku wracamy na właściwą drogę, aby je zinterpretować, jeśli taki istnieje. Załóżmy, że słyszysz zdanie, w którym Elsie próbowała skontaktować się z ciotką przez telefon, ale nie odebrała. Chociaż zdanie jest logicznie niejednoznaczne, nie ma wątpliwości, co ono oznacza. Nigdy nie przychodzi ci do głowy świadomie zastanawiać się, czy sądzony środek odbywał się w postępowaniu sądowym (jak w Sądzie karnym osądził Abe Ginsburga za kradzież), czy też dotarcie oznacza fizyczne dotarcie do celu (jak w łodzi dotarł do brzegu), czy też na telefon oznacza, że ciocia balansowała niepewnie na górze telefonu (jak w kępie kurzu na telefonie), czy też słowo, które ona w zdaniu, na które nie odpowiedziała, odnosi się do samej Elsie (tak jakby się skończyło zdanie z ale nie otrzymała odpowiedzi). Zamiast tego od razu skupiasz się na właściwej interpretacji. Teraz spróbuj zdobyć maszynę, która zrobi to wszystko. W niektórych przypadkach pomocne mogą być proste statystyki. Słowo „sądzony” oznacza znacznie częściej podejmowaną próbę niż postępowanie sądowe. Wyrażenie przez telefon oznacza częstsze używanie telefonu do komunikacji niż fizyczne używanie telefonu na górze, chociaż są wyjątki. Kiedy po czasowniku zasięg następuje osoba, a słowo telefon jest w pobliżu w zdaniu, prawdopodobnie oznacza to pomyślnie nawiązaną komunikację. Jednak w wielu przypadkach statystyki nie pomogą Ci znaleźć właściwego rozwiązania. Zamiast tego często nie ma sposobu na rozwiązanie danej niejednoznaczności bez faktycznego zrozumienia, co się dzieje. W zdaniu, które czyta Elsie próbowała skontaktować się z ciotką przez telefon, ale nie odpowiedziała, liczy się wiedza ogólna wraz z rozumowaniem. Znajomość kontekstu sprawia, że dla czytelnika staje się oczywiste, że Elsie nie odebrałaby własnego telefonu. Logika podpowiada, że to musi być jej ciotka. Nikt nie musi nas uczyć tego rodzaju wnioskowania w szkole, ponieważ wiemy, jak to robić instynktownie; wynika to naturalnie z tego, jak w pierwszej kolejności interpretujemy świat. Głębokie uczenie nie może nawet zacząć rozwiązywać tego rodzaju problemu.

Niestety, jak dotąd nic tak naprawdę nie zadziałało. Klasyczne techniki sztucznej inteligencji, które były powszechne na długo przed tym, jak uczenie głębokie stało się popularne, są znacznie lepsze w kompozycyjności i są użytecznym narzędziem do budowania modeli kognitywnych, ale jak dotąd nie

były tak dobre jak uczenie głębokie w uczeniu się od dane, a język jest zbyt złożony, aby ręcznie zakodować wszystko, czego potrzebujesz. Klasyczne systemy AI często używają szablonów. Na przykład szablon [MIEJSCE1 to ODLEGŁOŚĆ od MIEJSCA2] może być dopasowany do zdania Księżyc znajduje się 240 000 mil od Ziemi i użyty do zidentyfikowania tego jako zdania określającego odległość między dwoma miejscami. Jednak każdy szablon musi być ręcznie zakodowany, a w chwili napotkania nowego zdania, które różni się od poprzedniego (np. Księżyc leży około 240 000 mil od Ziemi lub Księżyc okrąża Ziemię w odległości 240 000 mil), system zaczyna się psuć. A szablony same w sobie prawie nic nie pomagają w rozwiązaniu łamigłówek polegających na łączeniu wiedzy o języku z wiedzą o świecie w celu rozwiązania niejednoznaczności. Jak dotąd dziedzina rozumienia języka naturalnego znalazła się między dwoma stołkami: jeden, głębokie uczenie, jest fantastyczny w uczeniu się, ale kiepski w kompozycyjności i konstruowaniu modeli kognitywnych; druga, klasyczna sztuczna inteligencja, łączy kompozycyjność i konstrukcję modeli poznawczych, ale w najlepszym przypadku jest przeciętna w nauce. I obu brakuje głównej rzeczy, do której dążyliśmy w tej części: zdrowego rozsądku. Nie możesz budować wiarygodnych modeli poznawczych złożonych tekstów, chyba że wiesz dużo o tym, jak działa świat, o ludziach, miejscach i przedmiotach oraz o tym, jak wchodzą w interakcje. Bez tego ogromna większość tego, co przeczytasz, nie miałaby żadnego sensu. Prawdziwym powodem, dla którego komputery nie potrafią czytać, jest brak nawet podstawowej wiedzy o tym, jak działa świat. Niestety nabycie zdrowego rozsądku jest znacznie trudniejsze niż mogłoby się wydawać. Jak zobaczymy, potrzeba, aby maszyny nabrały zdrowego rozsądku, jest również znacznie bardziej wszechobecna, niż można by sobie wyobrazić. Jeśli jest to paląca kwestia dla języka, prawdopodobnie jest jeszcze bardziej nagląca dla robotyki.