NOVARA
CONSULTING GROUP

Ten dokument został przetłumaczony z języka angielskiego przez maszynę i może zawierać błędy. Oryginał w języku angielskim znajduje się pod adresem https://www.novaracg.com/2026/09/11/consistently-wrong-apple-discosign-and-sign-language-ai/.

Konsekwentnie błędny: Apple DiscoSign i sztuczna inteligencja dla języka migowego

Heather Grizzle · 11 września 2026

Badacze z Apple i Gallaudet stworzyli sztuczną inteligencję dla języka migowego, która pamięta. Gdy system utrzymuje stan, jeden wczesny błąd przestaje być hałaśliwy i staje się systemowy.

Spis treści

1. Co twierdzi artykuł, a co mówi sam o sobie
2. Spójność to nie to samo co dokładność
3. Co mierzą nowe metryki i co je zwalidowało
4. Gloss wciąż nie jest językiem
5. Zbiór danych, którego nie nazwano
6. Co zmienia się w ewaluacji, gdy system pamięta
7. Ograniczenia tej analizy
8. Źródła

Artykuł opublikowany w arXiv 2 września 2026 roku i przyjęty na główną konferencję EMNLP opisuje DiscoSign, framework do tłumaczenia tekstu angielskiego na glosy amerykańskiego języka migowego (ASL), który zachowuje pamięć o tym, co już zostało powiedziane. Jego autorami są badacze z Apple i Gallaudet University, a agregatorzy informacji donieśli, że pojawił się on na stronie Apple poświęconej badaniom nad uczeniem maszynowym 11 września — i to właśnie tam większość czytelników spoza tej dziedziny mogła się z nim zetknąć. Twierdzenie techniczne jest wąskie i dobrze sformułowane. Większość systemów tekst-na-migowy tłumaczy jedno zdanie na raz i nie ma pamięci między zdaniami, więc referent umieszczony w jednym miejscu może zostać po cichu przypisany na nowo w kolejnym, to samo pojęcie może zostać oddane

trzema różnymi znakami w trzech kolejnych zdaniach, a struktury retoryczne organizujące dyskurs migowy nie mogą zostać wybrane, ponieważ system nie widzi dyskursu. DiscoSign dodaje jawne rejestry, które utrzymują się w wielu zdaniach, wymusza je jako ograniczenia na modelu, a następnie weryfikuje i koryguje wynik względem nich.

To rzeczywisty postęp w rozwiązaniu rzeczywistej wady, a artykuł szczerze mówi o tym, czego nie robi. Znaczenie dla nadzoru leży o krok dalej. Gdy system przenosi stan przez cały dokument, jednostką, która może zawieść, przestaje być zdanie. Błędy przestają być zdarzeniami niezależnymi i stają się właściwościami całego fragmentu, a praktyki ewaluacyjne wypracowane przez tę dziedzinę, które oceniają zdania i uśredniają wyniki, nie są w stanie tego dostrzec. Ryzyko nie polega na tym, że badania przeceniają swoje znaczenie. Polega na tym, że słowa „świadomy dyskursu" trafią do dokumentów zamówień publicznych na długo zanim ktokolwiek uzgodni, jak testować, czy dyskurs jest rzeczywiście obsługiwany.

Co twierdzi artykuł, a co mówi sam o sobie

DiscoSign zajmuje się trzema zjawiskami. Pierwsze to koreferencja przestrzenna: w ASL bytom przypisuje się lokalizacje w przestrzeni migowej, a później odnosi się do nich, wskazując z powrotem na te miejsca, więc system musi pamiętać, że Alicja została umieszczona po lewej, a Bob po prawej. Drugie to klauzule pytanie-odpowiedź (Question-Answer Clauses), struktury temat-komentarz, w których osoba migająca zadaje pytanie retoryczne i natychmiast na nie odpowiada, co jest odpowiednie w niektórych pozycjach dyskursu, a niezręczne w innych. Trzecie to spójność pojęcie-gloss, co oznacza, że pojęcie pojawiające się pięć razy w dokumencie powinno być oddane za każdym razem tym samym znakiem, a nie na przemian bliskoznacznymi. Każdym z nich zajmuje się moduł, który utrzymuje rejestr, przedstawia go modelowi językowemu jako twarde ograniczenie, a następnie sprawdza zwrócony wynik i koryguje naruszenia na miejscu.

Artykuł podaje, że rozwiązanie to znacząco poprawia spójność przestrzenną i śledzenie encji w porównaniu z bazowymi modelami na poziomie zdań, przy jednoczesnym utrzymaniu jakości pojedynczych zdań na stałym poziomie, i wprowadza własne miary, ponieważ — jak stwierdza — standardowe miary tłumaczenia maszynowego „nie są w stanie uchwycić jakości na poziomie dyskursu". Sekcja ograniczeń jest niezwykle bezpośrednia. Framework „zajmuje się manualną produkcją znaków poprzez glosy, ale nie uwzględnia niemanualnych markerów (NMM), takich jak mimika twarzy i pro-

zodia, które niosą kluczowe informacje gramatyczne i emocjonalne w językach migowych". Jeden z trzech zbiorów danych „nie posiada adnotacji ASL uznanych za wzorcowe (ground truth)", więc części ewaluacji na poziomie dyskursu opierają się na automatycznych miarach i tłumaczeniu zwrotnym. Reguła decydująca o tym, kiedy klauzula pytająco-odpowiadająca jest odpowiednia, „nie jest w stanie odzwierciedlić opcjonalności" struktury, która czasem jest kwestią stylu. Autorzy stwierdzają również, że „współpracowali z członkami społeczności osób posługujących się językiem migowym" i wymieniają współautorów z Gallaudet University, co nie jest normą w tej literaturze i powinno zostać jasno powiedziane.

Spójność to nie to samo co dokładność

Mechanizm, który sprawia, że DiscoSign działa, jest jednocześnie tym, co czyni go problemem z punktu widzenia nadzoru, i to jest część, której nie da się wywnioskować z samego streszczenia. Rejestr wymusza, aby decyzja podjęta wcześniej była stosowana konsekwentnie w całym tekście. Jeśli decyzja jest słuszna, cały fragment jest spójny. Jeśli decyzja jest błędna, cały fragment jest błędny w ten sam sposób. Referent przypisany do niewłaściwej lokalizacji w drugim zdaniu nie skutkuje jednym złym zdaniem; skutkuje dokumentem, w którym każde późniejsze odniesienie z przekonaniem wskazuje na niewłaściwą osobę. Pojęcie zmapowane na niedokładny znak nie jest odosobnionym potknięciem; to ten sam znak, powtarzany, przy czym mechanizm zapewniający spójność systemu aktywnie zapobiega korekcie, którą w innym przypadku mogłaby wprowadzić zmienność. Własny opis potoku przetwarzania w artykule mówi, że „błędy na poziomie dyskursu wprowadzone na etapie glosowania rozprzestrzeniają się przez cały potok".

Systemy na poziomie zdania zawodzą hałaśliwie, a hałaśliwe zawodzenie ma niedocenianą właściwość: czytelnik to zauważa. Niespójność jest czytelna jako wada. System, który myli się konsekwentnie, wygląda jak system, który mówi to, co ma na myśli. Dla czytelnika Głuchego, który odbiera wynik, zamiast sprawdzać go względem źródła, stabilne błędne odniesienie wcale nie jest oczywistym błędem; jest po prostu tym, co mówi dokument. To odwraca zwykłe założenie, że lepsza wewnętrzna spójność jest bezpieczniejsza. Spójność podnosi koszt błędu w tym samym momencie, w którym obniża jego widoczność, a żadna średnia z ocen poszczególnych zdań tego nie pokaże.

Co mierzą nowe metryki i co je zwalidowało

Artykuł jest w tym miejscu ostrożny, a ta ostrożność zasługuje na uważną lekturę, ponieważ wyznacza pułap tego, co można odpowiedzialnie twierdzić na dalszych etapach. Nowe metryki mierzą spójność indeksu przestrzennego, stabilność mapowania pojęcie-gloss oraz odpowiedniość użycia klauzul pytanie-odpowiedź. Dwie z trzech mierzą, czy system robił to samo przez cały czas, co jest właściwością wyniku rozpatrywaną samą w sobie, a nie miarą tego, czy czytelnik Głuchy zrozumiał fragment. Aby zwalidować metryki, dwoje biegłych w ASL ewaluatorów oceniło trzydzieści dwie historie, sto pięćdziesiąt dwa zdania, w skali pięciopunktowej. Spójność przestrzenna korelowała umiarkowanie z ich ocenami. Spójność pojęcie-gloss korelowała. Metryka odpowiedniości dla klauzul pytanie-odpowiedź nie osiągnęła istotności statystycznej, a artykuł przypisuje to temu, że ewaluatorzy różnili się w ogóle co do tego, co liczy się jako taka klauzula, przy czym ważona zgodność między nimi została określona jako umiarkowana.

Jako wkład badawczy walidujący nową miarę, dwóch ekspertów oceniających to rozsądny projekt, a uczciwość w kwestii negatywnego wyniku zasługuje na uznanie. Jako podstawa do wdrożenia instytucjonalnego jest to dalece niewystarczające, a artykuł nie twierdzi inaczej. Trzy rozróżnienia mają znaczenie, jeśli praca ta trafi do zamówień publicznych. Po pierwsze, skorelowanie automatycznej miary z ocenami ekspertów dotyczącymi tej samej właściwości nie jest tym samym, co zmierzenie, czy osoby Głuche rozumieją dokument — czego nikt jeszcze nie zrobił na poziomie dyskursu. Po drugie, „biegły w ASL” nie jest synonimem „Głuchy”, a artykuł nie podaje, czy osoby oceniające były Głuche, posiadały certyfikaty ani jakie było ich pochodzenie. Po trzecie, gdy wykwalifikowani oceniający nie zgadzają się co do tego, czy dana struktura gramatyczna w ogóle występuje, automatyczny wynik stwierdzający, że jej użycie było odpowiednie, nie jest jeszcze artefaktem zapewniającym jakość, niezależnie od tego, jaką liczbę generuje.

Gloss wciąż nie jest językiem

Relacje przedstawiające to jako sztuczną inteligencję dla języka migowego wykraczającą poza tłumaczenie zdanie po zdaniu sięgają dalej niż to, co faktycznie zbudowano. DiscoSign produkuje gloss, sekwencję pisemnych etykiet zastępującą znaki ręczne, a autorzy wprost mówią, że znaczniki niemanualne nie zostały uwzględnione. W ASL znaczna część struktury dyskursu, którą artykuł stara się zachować, mieści się właśnie

w tych kanałach. Zaznaczanie tematu, granica między częścią pytającą a odpowiadającą tych samych struktur, które modeluje framework, zmiany ról sygnalizujące, czyja perspektywa jest wyrażana, kierunek wzroku wiążący odniesienie z powrotem z lokalizacją w przestrzeni — wszystko to jest przenoszone przez twarz, głowę i ciało. Ciąg glossów może zapisać, że odniesienie to index-j; nie może zapisać spojrzenia, które sprawia, że odniesienie trafia do celu.

Autorzy dostrzegają to i mówią na ten temat coś pożytecznego, mianowicie że stan przechowywany w ich rejestrach jest zbliżony do informacji, jakich potrzebowałby system wizualny niższego szczebla, aby wytworzyć niemanualne markery, i że wyprowadzenie tych adnotacji z rejestru jest naturalnym kolejnym krokiem. Jest to trzeźwy opis niedokończonej pracy. Nie jest to jednak tożsame z systemem, który tworzy spójny dyskurs migowy, a odległość między tymi dwoma rzeczami to miejsce, w którym urzędnik odpowiedzialny za zamówienia publiczne, czytając streszczenie prasowe, zostanie wprowadzony w błąd. Nabywca, któremu powiedziano, że potok dostawcy jest „świadomy dyskursu", ma prawo zapytać, którego etapu to dotyczy i czy cokolwiek na dalszych etapach po glosowaniu to zachowuje.

Zbiór danych, którego nie nazwano

Artykuł wymienia trzy źródła danych: ASL STEM Wiki, Bajki Ezopa z Project Gutenberg oraz, jak to ujęto, „licencjonowany zbiór danych". Dwa pierwsze są zidentyfikowane i możliwe do zweryfikowania. Trzeci nie ma podanej nazwy. Jest to zwyczajna praktyka i najprawdopodobniej odzwierciedla licencję komercyjną, a nie coś niepokojącego, warto jednak odnotować to jako otwarte pytanie, a nie pominąć, ponieważ w pracy nad językiem migowym pochodzenie korpusu jest pytaniem o ludzi. Dane migowe są nagrywane od osób migających, których twarze i ciała stanowią dane. Nie da się odpowiedzieć na pytania, czy osoby te wyraziły zgodę na badania nad tłumaczeniem maszynowym, czy zostały wynagrodzone i co licencja pozwala robić z danymi na dalszych etapach, gdy źródło nie jest nazwane. Dziedzina, która w coraz większym stopniu opiera się na jakości i pochodzeniu swoich korpusów, prędzej czy później zostanie zapytana o to przez regulatorów i organizacje osób Głuchych, a nazwanie źródła to najtańszy możliwy sposób, by być na to przygotowanym.

Co zmienia się w ewaluacji, gdy system pamięta

Jeśli systemy świadome dyskursu staną się normą — a kierunek rozwoju wskazuje, że tak się stanie — to każdy, kto je ocenia, potrzebuje narzędzi działających na poziomie, na którym system obecnie funkcjonuje. Z tego, co pokazuje ten artykuł, wynika sześć wymogów. Pierwszy to trwałość w miarę długości tekstu: spójność mierzona na trzech zdaniach nic nie mówi o tym, czy system utrzymuje kilkanaście referentów na całej stronie prozy, więc testy muszą podawać długość dokumentu i gęstość referentów, a nie raportować średnią. Drugi to propagacja: ewaluacja powinna raportować, jak daleko sięga skutek pojedynczego wczesnego błędu, mierzony jako liczba dotkniętych nim kolejnych zdań, a nie rozcieńczony w całym dokumencie jako ułamkowy wynik. Trzeci to najgorszy przypadek obok średniej, ponieważ narażenie instytucji jest determinowane przez fragment, który zawiódł, a nie przez przeciętny fragment.

Czwarty to odzyskiwanie: system utrzymujący stan powinien być testowany pod kątem tego, czy potrafi zmienić przypisanie, gdy późniejszy tekst zaprzecza wcześniejszemu wnioskowi, oraz co się dzieje, gdy dokument w sposób uzasadniony przypisuje referenta na nowo. Piąty to korekta ludzka na właściwym poziomie, ponieważ recenzent sprawdzający zdanie po zdaniu zaakceptuje fragment, którego błąd jest właściwością całości, co oznacza, że procedury recenzji napisane dla wyników na poziomie zdań nie przenoszą się na ten kontekst. Szósty to zrozumienie przez czytelników Głuchych, mierzone na całych dokumentach względem ludzkiego komparatora migowego, co pozostaje jedynym dowodem na to, że cała mierzona wewnętrzna spójność rzeczywiście dociera do osoby, dla której jest przeznaczona. Stanowisko Novara Consulting Group, któremu ten artykuł nie zaprzecza, jest takie, że miary wewnętrzne są elementem wejściowym dla procesu zapewniania jakości i nigdy nie zastępują zwalidowanego wyniku, a wymagane dowody rosną wraz ze stawką sytuacji, w której wynik jest wykorzystywany.

Ograniczenia tej analizy

Niniejsza analiza opiera się na wersji artykułu opublikowanej w arXiv, wraz z publicznymi doniesieniami o jego pojawieniu się na stronie badawczej Apple, które nie zostały niezależnie zweryfikowane. DiscoSign to badania, nie produkt: nic tutaj nie opisuje wdrożonego systemu, żaden dostawca nie składa na jego temat żadnych oświadczeń, a żadne narzędzie Novara Consulting Group nie zostało do niego zastosowane. Powyższe uwagi krytyczne nie są krytyką autorów, którzy sami wymieniają większość tych



ograniczeń i wyświadczyli tej dziedzinie przysługę, publikując miary, które w ogóle czynią błąd na poziomie dyskursu mierzalnym. Argument dotyczy tego, co dzieje się dalej, gdy prawdziwy postęp badawczy zostaje skompresowany do twierdzenia o możliwościach przez osoby, które nie przeczytały sekcji ograniczeń.

Wartość tego artykułu polega na tym, że czyni widocznym ukryty problem. Systemy na poziomie zdania zawodziły w kwestii dyskursu przez cały czas; po prostu nie było narzędzia, które mogłoby to dostrzec. Teraz pojawia się jego początek, a przychodzi z odkryciem, które należy odczytywać w równym stopniu jako ostrzeżenie, co jako wynik. Uczynienie systemu spójnym nie czyni go poprawnym. Sprawia, że jest poprawny albo niepoprawny od początku do końca, i decyduje o tym, zanim ktokolwiek przeczyta jedno słowo wyniku, które z tych dwóch będzie prawdą.

Źródła

1. Vasileios Baltatzis, Mert Inan, Connor Gillis, Raja Kushalnagar, Lorna Quandt, Leah Findlater i Colin Lea, „DiscoSign: Discourse-Aware Text to Sign Language Gloss Translation,” arXiv:2609.02796, 2 września 2026, przyjęty na główną konferencję EMNLP 2026. https://arxiv.org/abs/2609.02796
2. DiscoSign, pełny tekst (HTML). https://arxiv.org/html/2609.02796
3. AI Global Wire, doniesienie o pojawieniu się artykułu na stronie Apple Machine Learning Research, 11 września 2026. https://aiglobalwire.com/article/discosign-discourse-aware-text-to-sign-language-gloss-translation-dbf89917-bb51-4793-be6d-29ed6727d3e0

TIPS

Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

Jak cytować

Grizzle, H. M. (2026, September 11). Konsekwentnie błędny: Apple DiscoSign i sztuczna inteligencja dla języka migowego. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003304