This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/09/11/the-words-put-in-your-mouth-sign-to-text-ai-and-fluent-error/.

Słowa, które włożono w twoje usta: tłumaczenie AI z języka migowego na tekst i płynny błąd

Heather Grizzle · 11 września 2026

Sztuczna inteligencja tłumacząca język migowy na tekst potrafi stworzyć płynne zdania, których osoba Głucha nigdy nie wypowiedziała w języku migowym, a następnie zapisać je jako oficjalny zapis. Dlaczego wyniki BLEU nie są w stanie dostrzec błędu, który rzekomo mają naprawiać.

Spis treści

1. W którą stronę biegnie tłumaczenie
2. Metryka nie widzi tej porażki
3. Pewność jako sygnał treningowy, nie jako ujawnienie informacji
4. Co dane o pozie przekazują, a czego nie
5. Benchmarki to nie wdrożenia
6. Czego powinien wymagać nabywca
7. Jak twierdzenie dotarło do języka angielskiego
8. Ograniczenia tej analizy
9. Źródła

Pewne siebie błędne tłumaczenie w systemach zamiany języka migowego na tekst oraz benchmarki, które tego nie dostrzegają

On 9 September 2026 the Chinese technology outlet Quantum Bit reported a method from vivo AI Lab called CAPO-SLT, from a paper titled "CAPO-SLT: Confidence-Aware Policy Optimization for Stable Sign Language Translation Generation", posted on OpenReview, where papers sit while under peer review. The problem it names is the one that should worry anyone buying this technology. A translation system can produce a sentence that is fluent, grammatical and entirely plausible while having no sup-

port in what the signer actually did, and because the sentence reads well, nothing about it announces the error. As the reporting puts it, a single word that is locally plausible but lacking visual support can pull an entire translation off course, with the error accumulating across the sentence. The method adjusts the bounds within which reinforcement learning is allowed to update the model, token by token, according to how confident the previous policy was: tighter bounds where confidence is high, near 0.2, and looser ones where it is low, up to 0.3, with additional caps on tokens carrying negative advantage. The aim is to stop the model from locking in confident guesses and to leave room for uncertain ones to be pulled back toward the visual evidence.

Zgłoszone wyniki są węższe, niż sugeruje nagłówkowe ujęcie, a różnica ta ma znaczenie. W chińskim benchmarku CSL-Daily zgłoszone zyski są mierzone względem dwóch nazwanych systemów, które także działają na podstawie danych o pozie: mniej więcej 1,26 BLEU-4 nad Geo-Sign i 3,07 BLEU-4 nad Uni-Sign, przy czym twierdzenie dotyczy najlepszych wyników wśród systemów opartych wyłącznie na pozie, a nie najlepszych w całej dziedzinie. W amerykańskim benchmarku języka migowego How2Sign zgłoszone wartości to 41,4 BLEU-1, 15,2 BLEU-4 i 34,9 ROUGE-L, czyli zyski rzędu około punktu lub mniej w stosunku do Uni-Sign. Sama praca nie mogła zostać odczytana na potrzeby tej analizy, ponieważ repozytorium, które ją hostuje, wymaga kroku weryfikacji przeglądarki, dlatego poniżej przeanalizowano problem, który praca nazywa, kierunek tłumaczenia, w jakim działa, oraz rodzaj przedstawianych dowodów, zamiast oceniać sam system. Warto przyjrzeć się tym kwestiom samym w sobie, ponieważ problem jest rzeczywisty, kierunek jest tym, który instytucje obsługują najmniej ostrożnie, a ten rodzaj dowodów nie jest w stanie wykazać tego, czego się od niego oczekuje.

W którą stronę biegnie tłumaczenie

Most public argument about signed language AI concerns output to Deaf people: an avatar on a departure board, a signing video on a website, a robot in a classroom. Sign-to-text runs the other way. The input is a Deaf person signing and the output is text that everyone else reads, and that changes what a mistake is. When an avatar signs badly, a Deaf person receives a poor translation and can usually tell that something is wrong. When a sign-to-text system fabricates a fluent sentence, the Deaf person's words have been replaced by words they did not say, and those words go to a cli-

nician, a caseworker, an investigator or a court reporter, who has no way of knowing anything happened.

Ta asymetria stanowi problem zarządzania. Jedyną osobą w pomieszczeniu, która może wykryć błąd, jest ta, której znaczenie zostało podmienione, a to często osoba, która nigdy nie widzi wyniku. Tekst powstały w ten sposób nie znika: zostaje wpisany do notatki przyjęcia, zeznania, oceny pracowniczej, akt sprawy. Gdy już tam trafi, niesie ze sobą powagę dokumentu urzędowego, a nie status tłumaczenia maszynowego, a osoba Głucha musi kwestionować dokument, który rzekomo ją cytuje. Płynny błąd w tym kierunku nie jest nieudanym udogodnieniem dostępu. Jest przypisaniem wypowiedzi, a instytucje praktycznie nie dysponują mechanizmem jego wycofania.

Metryka nie widzi tej porażki

Every number reported for this method, and for essentially every system like it, is BLEU or ROUGE. Both work by comparing the machine's output with a reference translation and measuring overlap of word sequences. They were built for spoken-language machine translation and they measure resemblance to a reference, which is not the same as fidelity to what the signer expressed. A fabricated sentence that happens to share common phrasing with the reference scores well. A correct translation phrased differently scores badly. The failure mode the method is designed to fix, fluent output unsupported by the visual evidence, is precisely the failure these metrics are least able to detect, because fluency is what they reward.

This is not a complaint from outside the field. A paper from the Centre for Vision, Speech and Signal Processing at the University of Surrey, posted this month, evaluates six translation systems and concludes that "gains in BLEU-4 are not on their own evidence of better sign language understanding," noting that the low-resource multimodal setting "allows models to exploit spurious correlations and spoken-language priors, rather than learning stronger sign representations." Their alternative protocol, which tests whether the content of the passage survived, reorders the field and reveals a gap between systems that BLEU-4 could not see at all. Read alongside that finding, a reported improvement in BLEU-4 is evidence that output resembles references more closely. It is not yet evidence that the system understood the signing, and it is certainly not evidence that confident fabrication has been reduced.

Wielkość zgłoszonego zysku ma tu również znaczenie. Poprawy rzędu mniej więcej jednego do trzech punktów BLEU-4 względem nazwanych punktów odniesienia to

zwyczajny przyrostowy postęp, a są one przedstawiane jako dowód na to, że system fabrykuje mniej. W zestawieniu z ustaleniem z Surrey jest to najsłabsza część argumentacji: zysk kilku punktów w metryce, co do której sama dziedzina twierdzi, że nie śledzi rozumienia języka migowego, nie może wykazać, że ograniczono pewne siebie fabrykowanie. Twierdzenie może nadal być prawdziwe. Jego wykazanie wymagałoby innego instrumentu, a praca, która jest w trakcie recenzji, a nie opublikowana, jest miejscem, w którym takie wykazanie musiałoby się znaleźć.

Pewność jako sygnał treningowy, nie jako ujawnienie informacji

The most interesting thing about the approach, taken at face value, is that it treats the model's own uncertainty as information worth acting on. That is the right instinct and it points at the thing procurement should be asking for, though not quite in the form the method provides. A confidence value used inside training changes how the model learns. It does not reach the person reading the output, and it does not change what the output looks like when the model is uncertain. The system still produces a fluent sentence, because producing fluent sentences is what it does. Nothing in the text says that three of its words were guesses.

Two things would change that, and neither is exotic. The first is abstention: the system should be able to decline. A sign-to-text system that can say "this passage was not clearly captured" and mark the gap is considerably safer than one that always returns a complete sentence, because a visible gap invites a human to fill it and a fluent invention does not. The second is exposure: confidence, where the system has it, should be visible at the point of use, and low-confidence spans should be marked in the text that reaches the reader. A vendor stating 95 percent accuracy is describing an average, and an average tells a buyer nothing about the character of the remaining five percent. Five percent noise is an inconvenience. Five percent confident fabrication, distributed unpredictably through a medical history or a witness statement, is a different product altogether, and the one figure most vendors publish cannot distinguish between them.

Co dane o pozie przekazują, a czego nie

The reported system works from pose data, meaning tracked body keypoints rather than raw video. There is a real privacy advantage in that, since keypoints are less identifying than footage of a signer's face, and it matters in a field where the data is a per-

son's body. But it also raises a question that cannot be answered without the paper: which keypoints. Signed languages carry grammar on the face. A headshake can negate the sentence it accompanies, eyebrow position can mark the difference between a statement and a question, and mouth patterns disambiguate signs that the hands render identically. If a keypoint set samples the face sparsely, a system may be structurally unable to see the features that reverse a meaning, and the failure this produces is the fluent kind: a clean declarative sentence where the signer negated, or a statement where the signer asked. In a clinical or legal setting, the difference between a sentence and its negation is the whole of the content.

Benchmarki to nie wdrożenia

CSL-Daily to korpus chińskiego języka migowego nagrany w laboratorium na tematy codzienne. How2Sign to duży zbiór instruktażowych nagrań wideo w amerykańskim języku migowym. Oba są poważnymi zasobami badawczymi, lecz żaden nie przypomina warunków, w jakich instytucja rzeczywiście używałaby zamiany migania na tekst: osoby cierpiącej, migającej ze szpitalnego łóżka pod kątem, do którego kamera nie była zaprojektowana, w odmianie regionalnej, z ręką unieruchomioną przez kaniulę, przerywanej, zaniepokojonej lub przechodzącej między rejestrami. Zgłoszony wynik How2Sign, 15,2 BLEU-4, warto zapamiętać z tego właśnie powodu. W tym bardziej otwartym domenowo z dwóch benchmarków, w sprzyjających warunkach, stan tej dziedziny daje wynik, którego nakładanie się z tłumaczeniem referencyjnym wykonanym przez człowieka pozostaje umiarkowane. To jest wynik badawczy i rozsądny. Nie jest to podstawa, na której cokolwiek powinno polegać przy zapisywaniu tego, co powiedział Głuchy pacjent.

Czego powinien wymagać nabywca

Six requirements follow, and they are specific to this direction of translation. The first is abstention with a visible gap, so that uncertainty appears as a gap rather than as prose. The second is confidence marking in the delivered text, available to the reader and not only to the model. The third is back-translation to the signer before anything is recorded, so that the Deaf person sees, in their own language, what the system is about to say in their name, and can stop it. The fourth is provenance in the record: text derived from machine translation should be labelled as such wherever it is stored,

with the source retained, so that a later dispute is about a machine's output rather than about the person's credibility.

The fifth is a correction route that reaches the record itself rather than the vendor's support desk, because an uncorrected sentence in a case file outlives any ticket. The sixth is evidence appropriate to the claim: semantic fidelity measured by protocols that test whether content survived, adversarial testing on negation, questions and role shift, disaggregation by signing variety and capture conditions, and comprehension checked with Deaf signers rather than inferred from reference overlap. NCG's position is that the evidence required rises with the stakes of the setting, and that a system whose errors enter an official record attributed to a person sits at the top of that scale, not in the middle of it.

Jak twierdzenie dotarło do języka angielskiego

There is a second story here, and it belongs in a piece about evidence. The Quantum Bit report is properly sourced: it names the method, describes what it does, reports gains against named baselines and links the paper. By the time this reached English-speaking readers, it had passed through a site whose byline is an "AI editorial department", which describes itself as a portal for generative engine optimisation, meaning content written to be retrieved and repeated by AI search tools, and which scores its own articles for quality. That version dropped the link to the paper and converted gains measured against particular baselines into flat absolute scores. Nothing about it is fraudulent. It is simply a summary of a summary, optimised to be picked up by machines, and the thing it quietly removed was the route back to the evidence.

Warto to nazwać, ponieważ w ten sposób większość osób zajmujących się zamówieniami będzie teraz napotykać twierdzenia o tej technologii. Twierdzenie dostawcy wchodzi do literatury jako zgłoszona praca, jest rzetelnie zrelacjonowane przez serwis branżowy, przepisane przez agregator pod kątem wyszukiwalności, pobrane przez asystenta i trafia do notatki informacyjnej z liczbami utwardzonymi i bez cytowania. Każdy krok sam w sobie da się obronić. Wynikiem jest liczba bez mianownika, bez punktu odniesienia i bez dokumentu za nią stojącego, przedstawiona z większą pewnością, niż twierdzili sami autorzy. Dyscyplina, która chroni tu nabywcę, jest tą samą, której ten artykuł wymaga od technologii: prześledzić twierdzenie z powrotem do jego źródła i powiedzieć wprost, czego nie udało się przeczytać.

Ograniczenia tej analizy

Przedstawiony tu opis CAPO-SLT opiera się na chińskojęzycznym doniesieniu branżowym dotyczącym tej pracy, które wymienia ją i linkuje. Sama praca, hostowana w repozytorium wymagającym kroku weryfikacji przeglądarki, nie mogła zostać odczytana na potrzeby tej analizy, więc przytoczone liczby są podane tak, jak zostały zgłoszone, a nie zweryfikowane ze źródłem, i nic powyżej nie powinno być odczytywane jako ocena systemu ani vivo AI Lab, którego sformułowany problem jest właściwy, a metoda być może rzeczywiście robi to, co deklaruje. Argumentacja opiera się na sformułowaniu problemu, które dziedzina ta teraz otwarcie nazwała, oraz na opublikowanych pracach innych autorów dotyczących tego, co standardowe benchmarki mogą, a czego nie mogą wykazać. Żaden instrument NCG nie został zastosowany do tego systemu, a szersza teza w ogóle nie zależy od tej jednej metody.

It is worth ending on what is encouraging here, because there is something. A commercial research lab has publicly identified confident fabrication as the central defect in sign-to-text translation rather than reporting another incremental gain and leaving the failure unnamed. That is the honest version of the problem, and naming it is the precondition for measuring it. The risk now is the familiar one: that the naming travels, the fix is assumed, and an institution somewhere types a fluent sentence into a Deaf person's file in the belief that the problem was solved by a paper it never read.

Źródła

1. Quantum Bit via 36kr, "CAPO-SLT achieves top scores on three Chinese metrics," reporting vivo AI Lab's method, 9 September 2026. https://36kr.com/p/3975775861813507
2. vivo AI Lab, "CAPO-SLT: Confidence-Aware Policy Optimization for Stable Sign Language Translation Generation," OpenReview (under review; not read for this analysis). https://openreview.net/forum?id=l4bCsrSkYx
3. Zgeo (AI editorial department), rewritten account of the above, 9 September 2026. https://www.zgeo.com.cn/news/capo-slt-sign-language-translation-vivo-ai-lab
4. Oline Ranum, Edward Fish, Simon Hadfield and Richard Bowden, "Beyond BLEU: A Case for Redefining Sign Language Translation Benchmarks," University of Surrey (CVSSP), arXiv:2609.03734, September 2026. https://arxiv.org/abs/2609.03734
5. "RVLF: A Reinforcing Vision-Language Framework for Gloss-Free Sign Language

Translation," arXiv:2512.07273 (reported CSL-Daily gloss-free results). https://arxiv.org/abs/2512.07273

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.

SUPPORT NCG

NCG does this work without vendor money, so our findings stay ours. If it has been useful to you, you can support NCG's independent work here. NCG is a for-profit company, not a charity. Contributions are not tax-deductible and buy no service or influence.

Jak cytować

Grizzle, H. M. (2026, September 11). Słowa, które włożono w twoje usta: tłumaczenie AI z języka migowego na tekst i płynny błąd. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003302