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/.

Le parole messe in bocca: l'IA da segni a testo e l'errore fluente

Heather Grizzle · September 11, 2026

L'IA di traduzione dal segno al testo può inventare frasi fluenti che una persona Deaf non ha mai firmato, per poi archiviarle come verbale ufficiale. Perché i punteggi BLEU non riescono a cogliere il fallimento che pretendono di risolvere.

Contenuti

1. In quale direzione va la traduzione
2. La metrica non riesce a vedere il fallimento
3. La fiducia è un segnale di addestramento, non una divulgazione
4. Ciò che l'input di posa porta con sé, e ciò che non porta
5. I benchmark non sono implementazioni reali
6. Ciò che un acquirente dovrebbe esigere
7. Come l'affermazione è arrivata in inglese
8. I limiti di questa analisi
9. Fonti

Traduzioni errate ma sicure di sé nei sistemi da lingua dei segni a testo, e i benchmark che non riescono a vederle

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 support 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.

I risultati riportati sono più circoscritti di quanto suggerisca l'impostazione del titolo, e la differenza è rilevante. Sul benchmark cinese CSL-Daily, i guadagni riportati sono misurati rispetto a due sistemi specifici che operano anch'essi a partire da input di posa: circa 1,26 BLEU-4 in più rispetto a Geo-Sign e 3,07 BLEU-4 in più rispetto a Uni-Sign, con l'affermazione che si tratta dei migliori punteggi tra i sistemi basati solo sulla posa, non dei migliori in assoluto nel settore. Sul benchmark della lingua dei segni americana How2Sign, i dati riportati sono 41,4 BLEU-1, 15,2 BLEU-4 e 34,9 ROUGE-L, guadagni di circa un punto o meno rispetto a Uni-Sign. L'articolo stesso non ha potuto essere letto ai fini di questa analisi, poiché il repository che lo ospita richiede una verifica tramite browser; di conseguenza quanto segue esamina il problema individuato dal lavoro, la direzione di traduzione in cui opera e il tipo di evidenza offerta, piuttosto che valutare il sistema. Questi aspetti meritano di essere esaminati di per sé, perché il problema è reale, la direzione è quella che le istituzioni gestiscono con minor cura, e il tipo di evidenza non può dimostrare ciò che gli si chiede di dimostrare.

In quale direzione va la traduzione

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

L'asimmetria è il problema di governance. L'unica persona nella stanza in grado di individuare l'errore è quella il cui significato è stato sostituito, ed è spesso proprio la

persona che non vede mai l'output. Il testo prodotto in questo modo non svanisce: viene digitato in una nota di accettazione, in una dichiarazione, in una valutazione delle prestazioni, in un fascicolo. Una volta lì, assume l'autorità del documento ufficiale anziché lo status di una traduzione automatica, e la persona Deaf deve contestare un documento che pretende di citarla. Un errore fluente in questa direzione non è un fallimento dell'accomodamento di accessibilità. È un'attribuzione, e le istituzioni non dispongono quasi di alcun meccanismo per ritrattarla.

La metrica non riesce a vedere il fallimento

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.

Anche l'entità del guadagno riportato ha qui importanza. Miglioramenti di circa uno-tre punti BLEU-4 rispetto a sistemi di riferimento nominati sono progressi incrementali ordinari, e vengono presentati come evidenza che un sistema inventa meno. Letto alla luce del risultato di Surrey, questo è il punto più debole della tesi: un guadagno di pochi punti in una metrica che, secondo lo stesso settore, non riflette la comprensione

della lingua dei segni non può dimostrare che l'invenzione sicura sia stata ridotta. L'affermazione potrebbe comunque essere vera. Dimostrarla richiederebbe uno strumento diverso, e l'articolo, che è in fase di revisione e non pubblicato, è il luogo in cui tale dimostrazione dovrebbe essere cercata.

La fiducia è un segnale di addestramento, non una divulgazione

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.

Ciò che l'input di posa porta con sé, e ciò che non porta

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 person'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 nega-

te 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.

I benchmark non sono implementazioni reali

CSL-Daily è un corpus di lingua dei segni cinese registrato in laboratorio su argomenti quotidiani. How2Sign è un ampio insieme di video didattici in lingua dei segni americana. Entrambi sono risorse di ricerca serie e nessuno dei due assomiglia alle condizioni in cui un'istituzione utilizzerebbe effettivamente un sistema da segni a testo: una persona che soffre, che segna da un letto d'ospedale a un'angolazione per cui la telecamera non era pensata, in una variante regionale, con una mano immobilizzata da una cannula, interrotta, angosciata, o che passa da un registro all'altro. Il dato riportato per How2Sign, 15,2 BLEU-4, merita di essere tenuto a mente proprio per questo motivo. Sul più aperto dei due benchmark, in condizioni favorevoli, lo stato dell'arte di questo settore produce un output la cui sovrapposizione con una traduzione di riferimento umana resta modesta. Si tratta di un risultato di ricerca, e ragionevole. Non è una base su cui ci si dovrebbe affidare per registrare ciò che un paziente Deaf ha detto.

Ciò che un acquirente dovrebbe esigere

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.

Come l’affermazione è arrivata in inglese

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.

Vale la pena nominarlo, perché è così che la maggior parte dei responsabili degli acquisti incontrerà ormai le affermazioni su questa tecnologia. Un’affermazione di un fornitore entra nella letteratura come articolo sottoposto a revisione, viene riportata accuratamente da una testata di settore, viene riscritta da un aggregatore per il recupero, viene recuperata da un assistente e arriva in una nota informativa con i numeri irrigiditi e la citazione scomparsa. Ogni passaggio è difendibile preso singolarmente. Il risultato è una cifra senza denominatore, senza sistema di riferimento e senza documento alle spalle, presentata con più sicurezza di quanta ne rivendicassero gli autori originali. La disciplina che protegge qui un acquirente è la stessa che questo articolo chiede alla tecnologia: risalire l’affermazione fino alla cosa da cui proviene, e dire chiaramente ciò che non si è potuto leggere.

I limiti di questa analisi

Il resoconto di CAPO-SLT qui presentato si basa su un reportage di settore in lingua cinese sul lavoro, che nomina l’articolo e lo collega. L’articolo stesso, ospitato su un re-

pository che richiede una verifica tramite browser, non ha potuto essere letto ai fini di questa analisi, quindi le cifre citate sono così come riportate anziché verificate rispetto alla fonte, e nulla di quanto sopra dovrebbe essere letto come una valutazione del sistema o del vivo AI Lab, il cui problema dichiarato è quello giusto e il cui metodo potrebbe benissimo fare ciò che afferma. L'argomentazione si regge sull'enunciazione del problema, che il settore ha ora nominato apertamente, e sul lavoro pubblicato da altri riguardo a ciò che i benchmark standard possono e non possono mostrare. Nessuno strumento NCG è stato applicato a questo sistema, e il punto più ampio non dipende affatto da questo singolo metodo.

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.

Fonti

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.

Come citare

Grizzle, H. M. (2026, September 11). Le parole messe in bocca: l'IA da segni a testo e l'errore fluente. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003302