NOVARA
CONSULTING GROUP

As Palavras Postas na Sua Boca: IA de Língua Gestual para Texto e o Erro Fluente

Heather Grizzle · 11 de setembro de 2026

A IA de língua gestual para texto pode inventar frases fluentes que uma pessoa Deaf nunca gestualizou, e depois arquivá-las como registo. Por que razão as pontuações BLEU não conseguem ver a falha que afirmam corrigir.

Conteúdo

Erro de tradução confiante em sistemas de língua gestual para texto, e os benchmarks que não o conseguem ver

Em 9 de setembro de 2026, o meio de comunicação tecnológico chinês Quantum Bit noticiou um método do vivo AI Lab chamado CAPO-SLT, proveniente de um artigo intitulado "CAPO-SLT: Confidence-Aware Policy Optimization for Stable Sign Language Translation Generation", publicado no OpenReview, onde os artigos permanecem enquanto estão sujeitos a revisão por pares. O problema que identifica é aquele que deveria preocupar qualquer pessoa que compre esta tecnologia. Um sistema de tradução

pode produzir uma frase fluente, gramaticalmente correta e inteiramente plausível, sem qualquer sustentação naquilo que o sinalizante realmente fez, e como a frase soa bem, nada nela denuncia o erro. Como refere a reportagem, uma única palavra que é localmente plausível mas carece de sustentação visual pode desviar toda uma tradução, com o erro a acumular-se ao longo da frase. O método ajusta os limites dentro dos quais a aprendizagem por reforço pode atualizar o modelo, token a token, de acordo com o grau de confiança da política anterior: limites mais apertados onde a confiança é elevada, próximos de 0,2, e mais soltos onde é baixa, até 0,3, com limites adicionais em tokens que carregam vantagem negativa. O objetivo é impedir que o modelo se fixe em suposições confiantes e deixar margem para que as incertas sejam reconduzidas à evidência visual.

Os resultados relatados são mais restritos do que a apresentação do título sugere, e essa diferença é importante. No benchmark chinês CSL-Daily, os ganhos reportados são medidos em relação a dois sistemas nomeados que também trabalham a partir de entrada de pose: aproximadamente 1,26 BLEU-4 sobre o Geo-Sign e 3,07 BLEU-4 sobre o Uni-Sign, sendo a afirmação a de obter as melhores pontuações entre os sistemas baseados apenas em pose, e não as melhores do campo em geral. No benchmark de Língua Gestual Americana How2Sign, os valores relatados são 41,4 BLEU-1, 15,2 BLEU-4 e 34,9 ROUGE-L, ganhos de cerca de um ponto ou menos em relação ao Uni-Sign. O próprio artigo não pôde ser lido para esta análise, porque o repositório que o aloja exige um passo de verificação no navegador, pelo que se segue um exame do problema que o trabalho identifica, do sentido de tradução em que opera e do tipo de evidência que está a ser oferecida, em vez de uma avaliação do sistema. Vale a pena examinar estes aspetos por si só, porque o problema é real, o sentido é aquele que as instituições tratam com menos cuidado, e o tipo de evidência não consegue demonstrar aquilo que se lhe pede que demonstre.

Em que sentido corre a tradução

A maior parte do debate público sobre IA em língua gestual diz respeito à produção destinada a pessoas Deaf: um avatar num painel de partidas, um vídeo a gesticular num site, um robô numa sala de aula. A tradução de língua gestual para texto funciona no sentido inverso. A entrada é uma pessoa Deaf a gesticular e a saída é texto que toda a gente lê, e isso muda o que constitui um erro. Quando um avatar gesticula mal, uma pessoa Deaf recebe uma tradução fraca e geralmente consegue perceber que algo

está errado. Quando um sistema de língua gestual para texto fabrica uma frase fluente, as palavras da pessoa Deaf foram substituídas por palavras que ela não disse, e essas palavras chegam a um clínico, a um gestor de casos, a um investigador ou a um estenógrafo do tribunal, que não tem forma de saber que algo aconteceu.

A assimetria é o problema de governação. A única pessoa na sala capaz de detetar o erro é aquela cujo significado foi substituído, e essa pessoa é muitas vezes aquela que nunca vê o resultado. O texto produzido desta forma não se evapora: é digitado numa nota de admissão, numa declaração, numa avaliação de desempenho, num processo. Uma vez ali, carrega a autoridade do registo em vez do estatuto de uma tradução automática, e a pessoa Deaf tem de contestar um documento que pretende citá-la. Um erro fluente neste sentido não é uma falha de acomodação de acessibilidade. É uma atribuição, e as instituições quase não têm mecanismos para a retratar.

A métrica não consegue ver a falha

Todos os números relatados para este método, e praticamente para qualquer sistema semelhante, são BLEU ou ROUGE. Ambos funcionam comparando a saída da máquina com uma tradução de referência e medindo a sobreposição de sequências de palavras. Foram construídos para tradução automática de línguas orais e medem semelhança com uma referência, o que não é o mesmo que fidelidade ao que o sinalizante exprimiu. Uma frase fabricada que por acaso partilha formulações comuns com a referência pontua bem. Uma tradução correta formulada de forma diferente pontua mal. O modo de falha que o método pretende corrigir, uma saída fluente sem sustentação na evidência visual, é precisamente aquele que estas métricas menos conseguem detetar, porque a fluência é o que elas recompensam.

Esta não é uma crítica vinda de fora do campo. Um artigo do Centre for Vision, Speech and Signal Processing da Universidade de Surrey, publicado este mês, avalia seis sistemas de tradução e conclui que "ganhos em BLEU-4 não são, por si só, prova de melhor compreensão da língua gestual", observando que o cenário multimodal de baixos recursos "permite que os modelos explorem correlações espúrias e priors da língua oral, em vez de aprenderem representações mais sólidas dos gestos". O seu protocolo alternativo, que testa se o conteúdo da passagem sobreviveu, reordena o campo e revela uma lacuna entre sistemas que o BLEU-4 não conseguia ver de todo. Lido em conjunto com essa constatação, uma melhoria relatada em BLEU-4 é prova de que a saída se assemelha mais às referências. Ainda não é prova de que o sistema compreen-

deu a gesticulação, e certamente não é prova de que a fabricação confiante foi reduzida.

A dimensão do ganho relatado também importa aqui. Melhorias de aproximadamente um a três pontos de BLEU-4 sobre referências nomeadas são progresso incremental normal, e estão a ser apresentadas como prova de que um sistema fabrica menos. Lida à luz da constatação de Surrey, essa é a parte mais fraca do argumento: um ganho de poucos pontos numa métrica que o próprio campo diz não acompanhar a compreensão da língua gestual não pode estabelecer que a fabricação confiante foi reduzida. A afirmação pode ainda assim ser verdadeira. Demonstrá-la exigiria um instrumento diferente, e o artigo, que está em revisão em vez de publicado, é onde essa demonstração teria de ser encontrada.

A confiança é um sinal de treino, não uma divulgação

O aspeto mais interessante da abordagem, tomado à letra, é que trata a própria incerteza do modelo como informação sobre a qual vale a pena agir. Esse é o instinto certo e aponta para aquilo que a área de aquisições deveria exigir, embora não exatamente na forma que o método fornece. Um valor de confiança usado dentro do treino muda a forma como o modelo aprende. Não chega à pessoa que lê a saída, e não muda o aspeto da saída quando o modelo está incerto. O sistema continua a produzir uma frase fluente, porque produzir frases fluentes é o que faz. Nada no texto indica que três das suas palavras foram suposições.

Duas coisas mudariam isso, e nenhuma delas é exótica. A primeira é a abstenção: o sistema deveria poder recusar-se. Um sistema de língua gestual para texto que consiga dizer "esta passagem não foi claramente captada" e assinalar a lacuna é consideravelmente mais seguro do que um que devolve sempre uma frase completa, porque uma lacuna visível convida um humano a preenchê-la, ao passo que uma invenção fluente não convida a nada. A segunda é a exposição: a confiança, onde o sistema a tiver, deveria ser visível no ponto de utilização, e os trechos de baixa confiança deveriam ser assinalados no texto que chega ao leitor. Um fornecedor que declara 95 por cento de precisão está a descrever uma média, e uma média não diz nada a um comprador sobre a natureza dos restantes cinco por cento. Cinco por cento de ruído é um incómodo. Cinco por cento de fabricação confiante, distribuída de forma imprevisível ao longo de um histórico clínico ou de um depoimento de testemunha, é um produto com-

pletamente diferente, e o único valor que a maioria dos fornecedores publica não consegue distinguir entre os dois.

O que a entrada de pose transmite e o que não transmite

O sistema relatado trabalha a partir de dados de pose, ou seja, pontos-chave corporais rastreados em vez de vídeo bruto. Há aqui uma vantagem real de privacidade, uma vez que os pontos-chave identificam menos do que a filmagem do rosto de um sinalizante, e isso importa num campo em que os dados são o corpo de uma pessoa. Mas também levanta uma questão que não pode ser respondida sem o artigo: quais os pontos-chave. As línguas gestuais transportam gramática no rosto. Um abanar de cabeça pode negar a frase que acompanha, a posição das sobrancelhas pode marcar a diferença entre uma afirmação e uma pergunta, e os padrões da boca desambiguam gestos que as mãos representam de forma idêntica. Se um conjunto de pontos-chave amostra o rosto de forma esparsa, um sistema pode ser estruturalmente incapaz de ver as características que invertem um significado, e a falha que isso produz é do tipo fluente: uma frase declarativa limpa onde o sinalizante negou, ou uma afirmação onde o sinalizante perguntou. Num contexto clínico ou jurídico, a diferença entre uma frase e a sua negação é a totalidade do conteúdo.

Os benchmarks não são implementações

O CSL-Daily é um corpus de Língua Gestual Chinesa gravado em laboratório sobre temas do quotidiano. O How2Sign é um vasto conjunto de vídeos instrucionais em Língua Gestual Americana. Ambos são recursos de investigação sérios e nenhum se assemelha às condições em que uma instituição usaria efetivamente a tradução de língua gestual para texto: uma pessoa com dores, a gesticular a partir de uma cama de hospital num ângulo para o qual a câmara não foi concebida, numa variedade regional, com uma mão imobilizada por uma cânula, interrompida, angustiada, ou a alternar entre registos. O valor relatado para o How2Sign, 15,2 BLEU-4, vale a pena reter por essa razão. No mais aberto dos dois benchmarks, em domínio, sob condições favoráveis, o estado deste campo produz uma saída cuja sobreposição com uma tradução de referência humana permanece modesta. Esse é um resultado de investigação, e razoável. Não é uma base sobre a qual se deva confiar para registar o que um paciente Deaf disse.

O que um comprador deveria exigir

Seguem-se seis requisitos, específicos deste sentido de tradução. O primeiro é a abstenção com uma lacuna visível, de modo que a incerteza apareça como lacuna e não como prosa. O segundo é a marcação de confiança no texto entregue, disponível ao leitor e não apenas ao modelo. O terceiro é a retrotradução para o sinalizante antes de qualquer registo, para que a pessoa Deaf veja, na sua própria língua, o que o sistema está prestes a dizer em seu nome, e possa impedi-lo. O quarto é a proveniência no registo: o texto derivado de tradução automática deve ser rotulado como tal onde quer que seja armazenado, com a fonte preservada, para que uma disputa posterior seja sobre a saída de uma máquina e não sobre a credibilidade da pessoa.

O quinto é uma via de correção que chegue ao próprio registo e não apenas ao serviço de apoio do fornecedor, porque uma frase não corrigida num processo sobrevive a qualquer ticket. O sexto é a evidência adequada à afirmação: fidelidade semântica medida por protocolos que testam se o conteúdo sobreviveu, testes adversariais sobre negação, perguntas e mudança de papel, desagregação por variedade de gesticulação e condições de captação, e compreensão verificada com sinalizantes Deaf em vez de inferida a partir da sobreposição com a referência. A posição da NCG é a de que a evidência exigida aumenta com o que está em jogo no contexto, e que um sistema cujos erros entram num registo oficial atribuído a uma pessoa se situa no topo dessa escala, e não a meio.

Como a afirmação chegou ao inglês

Há aqui uma segunda história, e ela pertence a um artigo sobre evidência. A reportagem do Quantum Bit está devidamente sustentada: nomeia o método, descreve o que faz, relata ganhos face a referências nomeadas e liga ao artigo. Quando isto chegou aos leitores de língua inglesa, já tinha passado por um site cuja assinatura é um "departamento editorial de IA", que se descreve como um portal de otimização para motores generativos, ou seja, conteúdo escrito para ser recuperado e repetido por ferramentas de pesquisa de IA, e que classifica os seus próprios artigos por qualidade. Essa versão eliminou a ligação ao artigo e converteu ganhos medidos face a referências específicas em pontuações absolutas fixas. Nada nisto é fraudulento. É simplesmente um resumo de um resumo, otimizado para ser captado por máquinas, e aquilo que discretamente removeu foi o caminho de volta à evidência.

Vale a pena assinalar isto porque é assim que a maioria dos responsáveis por aquisições irá agora encontrar afirmações sobre esta tecnologia. A afirmação de um fornecedor entra na literatura como um artigo submetido, é relatada com exatidão por um meio especializado, é reescrita por um agregador para efeitos de recuperação, é recuperada por um assistente e chega a uma nota informativa com os números já cristalizados e a citação desaparecida. Cada passo é defensável por si só. O resultado é um valor sem denominador, sem referência e sem documento por trás, apresentado com mais confiança do que os autores originais alguma vez afirmaram. A disciplina que protege um comprador aqui é a mesma que este artigo exige da tecnologia: seguir a afirmação até à sua origem, e dizer claramente aquilo que não se conseguiu ler.

Os limites desta análise

O relato do CAPO-SLT aqui apresentado assenta na reportagem especializada em língua chinesa sobre o trabalho, que nomeia o artigo e liga a ele. O próprio artigo, alojado num repositório que exige um passo de verificação no navegador, não pôde ser lido para esta análise, pelo que os valores citados são conforme relatados e não verificados face à fonte, e nada acima deve ser lido como uma avaliação do sistema ou do vivo AI Lab, cujo problema declarado é o correto e cujo método pode muito bem fazer o que afirma. O argumento assenta na formulação do problema, que o campo agora nomeou abertamente, e no trabalho publicado por outros sobre aquilo que os benchmarks padrão conseguem e não conseguem mostrar. Nenhum instrumento da NCG foi aplicado a este sistema, e o ponto mais amplo não depende de forma alguma deste método específico.

Vale a pena terminar com o que aqui é animador, porque há algo. Um laboratório de investigação comercial identificou publicamente a fabricação confiante como o defeito central da tradução de língua gestual para texto, em vez de relatar mais um ganho incremental e deixar a falha por nomear. Essa é a versão honesta do problema, e nomeá-lo é a condição prévia para o medir. O risco agora é o habitual: que a nomeação circule, a correção seja pressuposta, e algures uma instituição digite uma frase fluente no processo de uma pessoa Deaf, na crença de que o problema foi resolvido por um artigo que nunca leu.

Fontes

1. Quantum Bit via 36kr, "CAPO-SLT alcança as melhores pontuações em três métricas chinesas," relatando o método do vivo AI Lab, 9 de setembro de 2026. https://36kr.com/p/3975775861813507
2. vivo AI Lab, "CAPO-SLT: Confidence-Aware Policy Optimization for Stable Sign Language Translation Generation," OpenReview (em revisão; não lido para esta análise). https://openreview.net/forum?id=l4bCsrSkYx
3. Zgeo (departamento editorial de IA), relato reescrito do anterior, 9 de setembro de 2026. https://www.zgeo.com.cn/news/capo-slt-sign-language-translation-vivo-ai-lab
4. Oline Ranum, Edward Fish, Simon Hadfield e Richard Bowden, "Beyond BLEU: A Case for Redefining Sign Language Translation Benchmarks," Universidade de Surrey (CVSSP), arXiv:2609.03734, setembro de 2026. https://arxiv.org/abs/2609.03734
5. "RVLF: A Reinforcing Vision-Language Framework for Gloss-Free Sign Language Translation," arXiv:2512.07273 (resultados relatados sem glosa no CSL-Daily). https://arxiv.org/abs/2512.07273

Como citar

Grizzle, H. M. (2026, September 11). The Words Put in Your Mouth: Sign-to-Text AI and Fluent Error. Novara Consulting Group. https://www.novaracg.com/pt/2026/09/11/the-words-put-in-your-mouth-sign-to-text-ai-and-fluent-error/