This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/09/11/consistently-wrong-apple-discosign-and-sign-language-ai/.

始终如一的错误:Apple DiscoSign 与手语人工智能

Heather Grizzle · 2026年9月11日

Apple 和 *Gallaudet* 的研究人员打造了具备记忆功能的手语人工智能。当系统保持状态时,一个早期的错误就不再是噪音,而会变成系统性的问题。

目录

1. 论文声称了什么,以及它对自身的说明
2. 一致性不等于准确性
3. 新指标衡量的是什么,以及是谁验证了它们
4. 字符注释仍然不是语言
5. 那个未被点名的数据集
6. 当系统具备记忆时,评估会发生哪些变化
7. 本分析的局限性
8. 资料来源

A paper posted to arXiv on 2 September 2026 and accepted to the main conference at EMNLP describes DiscoSign, a framework for translating English text into American Sign Language gloss while keeping track of what has already been said. Its authors are researchers at Apple and at Gallaudet University, and aggregators reported it appearing on Apple's machine learning research site on 11 September, which is where most readers outside the field will have encountered it. The technical claim is narrow and well stated. Most text-to-sign systems translate one sentence at a time and have no memory between sentences, so a referent placed in one location can be silently reassigned in the next, the same concept can be rendered by three different signs in three consecutive sentences, and the rhetorical structures that organise signed discourse cannot be chosen because the system cannot see the discourse. DiscoSign

adds explicit registries that persist across sentences, enforces them as constraints on the model, and then verifies and corrects the output against them.

This is a real advance on a real defect, and the paper is candid about what it does not do. The governance significance lies one step further on. When a system carries state across a document, the unit that can fail is no longer the sentence. Errors stop being independent events and start being properties of the whole passage, and the evaluation practices the field has built, which score sentences and average them, cannot see that. The risk is not that the research overstates itself. It is that the words "discourse-aware" will arrive in procurement documents long before anyone has agreed how to test whether discourse is actually being handled.

论文声称了什么,以及它对自身的说明

DiscoSign处理三种现象。第一种是空间共指:在ASL中,实体被指派到手语空间中的位置,随后通过指向该位置来回指,因此系统必须记住Alice被安放在左侧、Bob被安放在右侧。第二种是问答从句(Question-Answer Clauses),即打手语者提出一个修辞性问题并立即作答的话题—述题结构,这种结构在某些话语位置是恰当的,在另一些位置则显得生硬。第三种是概念—字符注释一致性,意味着一篇文档中出现五次的某个概念每次都应以相同的手势呈现,而不是在近义手势之间来回切换。每一种现象都由一个模块处理,该模块维护一个登记表,将其作为硬性约束提交给语言模型,然后检查返回的输出并就地修正违反之处。

The paper reports that this substantially improves spatial consistency and entity tracking against sentence-level baselines while holding single-sentence quality steady, and it introduces its own metrics because, as it says, standard machine translation metrics "fail to capture discourse-level quality." Its limitations section is unusually direct. The framework "addresses manual sign production through glosses but does not incorporate non-manual markers (NMMs) such as facial expressions and prosody, which carry critical grammatical and affective information in sign languages." One of its three datasets "lacks ground truth ASL annotations," so parts of the discourse-level evaluation rest on automatic metrics and back-translation. The rule that decides when a Question-Answer Clause is appropriate "cannot represent the optionality" of a structure that is sometimes a matter of style. The authors also state that they "collaborated with members of the signing community" and include Gallaudet co-authors, which is not the norm in this literature and should be said plainly.

一致性不等于准确性

The mechanism that makes DiscoSign work is also what makes it a governance problem, and this is the part that does not follow from reading the abstract. A registry enforces that a decision made early is applied throughout. If the decision is right, the whole passage is coherent. If the decision is wrong, the whole passage is wrong the same way. A referent assigned to the wrong location in the second sentence does not produce one bad sentence; it produces a document in which every later reference points confidently to the wrong person. A concept mapped to an inexact sign is not an isolated slip; it is that sign, repeated, with the system's consistency machinery actively preventing the correction that variation might otherwise have introduced. The paper's own description of the pipeline says that "discourse-level errors introduced at the gloss stage propagate through the entire pipeline."

句子级系统的失败是嘈杂的,而嘈杂的失败有一个未被充分重视的特点:读者会察觉到它。不一致性作为一种缺陷是可辨识的。而一个始终如一地犯错的系统,看起来就像是一个言之有意的系统。对于接收输出而非将其与原文核对的聋人读者而言,一个稳定但错误的指代对象根本不像是一个错误;它只是文档所写的内容而已。这颠覆了通常的假设,即更好的内部连贯性更安全。连贯性在降低错误可见度的同时,提高了犯错的代价,而任何逐句评分的平均值都无法显示这一点。

新指标衡量的是什么,以及是谁验证了它们

论文在这一点上颇为谨慎,这份谨慎值得细读,因为它决定了下游可以负责任地提出何种主张的上限。新指标衡量的是空间索引一致性、概念–字符注释映射的稳定性,以及问答从句使用的恰当性。三者中有两者衡量的是系统是否始终如一地做同一件事,这是就输出本身而言的属性,并非衡量聋人读者是否理解了这段文字。为验证这些指标,两名ASL流利的评审员对三十二个故事、共一百五十二个句子按五分制打分。空间一致性与他们的判断呈中等程度相关。概念–字符注释一致性也呈相关。问答从句恰当性指标未达到统计显著性,论文将此归因于评审员之间在何为此类从句这一问题上存在分歧,两者之间的加权一致性被描述为中等程度。

For a research contribution validating a new metric, two expert raters is a reasonable design and the honesty about the negative result is creditable. As a basis for institutional deployment it is nowhere near sufficient, and the paper does not claim otherwise. Three distinctions matter if this work travels into procurement. First,

correlating an automatic metric with expert ratings of the same property is not the same as measuring whether Deaf readers comprehend the document, which nobody has yet done at discourse level. Second, "ASL-fluent" is not a synonym for Deaf, and the paper does not report the evaluators' deafness, certification or background. Third, when qualified raters disagree about whether a grammatical structure is even present, an automated score asserting that its use was appropriate is not yet an assurance artifact, whatever number it produces.

字符注释仍然不是语言

将此报道为手语人工智能已超越逐句翻译的说法,超出了实际所建成的内容。DiscoSign生成的是字符注释,一种代替手动手势的书面标签序列,而作者也直接说明未纳入非手动标记。在ASL中,论文试图保留的大量话语结构恰恰存在于那些渠道之中。话题标记、构成该框架所建模结构本身的问句与答句之间的边界、标示所表达视角归属者的角色转换、将某一指代与空间中某一位置相联系的目光注视:这些都由面部、头部和身体承载。字符串可以记录某个指代对象是索引j;它无法记录使该指代得以落实的目光注视。

The authors see this and say something useful about it, which is that the state their registries hold is close to the information a downstream visual system would need in order to produce non-manual markers, and that deriving those annotations from the registry is a natural next step. That is a sober description of unfinished work. It is not the same as a system that produces coherent signed discourse, and the distance between the two is where a procurement officer reading a press summary will be misled. A buyer told that a vendor's pipeline is discourse-aware is entitled to ask which stage that describes, and whether anything downstream of the gloss preserves it.

那个未被点名的数据集

The paper lists three data sources: ASL STEM Wiki, Aesop's Fables from Project Gutenberg, and, in its own words, "a licensed dataset." The first two are identified and checkable. The third is not named. This is ordinary practice and very likely reflects a commercial licence rather than anything untoward, but it is worth registering as a standing question rather than passing over it, because in signed language work the provenance of a corpus is a question about people. Signed data is recorded from signers whose faces and bodies are the data. Whether those signers were consented

for machine translation research, whether they were compensated, and what the licence permits downstream are not answerable when the source is unnamed. A field that increasingly turns on the quality and provenance of its corpora will eventually be asked these questions by regulators and by Deaf organisations, and naming the source is the cheapest possible way to be ready for that.

当系统具备记忆时,评估会发生哪些变化

If discourse-aware systems become the norm, and the direction of travel suggests they will, then anyone evaluating them needs instruments that operate at the level the system now operates at. Six requirements follow from what this paper shows. The first is persistence under length: consistency measured across three sentences says nothing about a system holding a dozen referents across a page of prose, so tests must state document length and referent density rather than reporting an average. The second is propagation: an evaluation should report how far a single early error travels, measured as affected downstream sentences, not diluted across a document as a fractional score. The third is worst case alongside mean, because an institution′s exposure is determined by the passage that failed, not the average passage.

The fourth is recovery: a system that maintains state should be tested on whether it can revise an assignment when later text contradicts an earlier inference, and on what happens when a document legitimately reassigns a referent. The fifth is human correction at the right level, since a reviewer checking sentence by sentence will approve a passage whose error is a property of the whole, which means review procedures written for sentence-level output do not transfer. The sixth is comprehension by Deaf readers, measured on whole documents against a human signed comparator, which remains the only evidence that any of the internal consistency being measured actually reaches the person it is for. NCG′s position, which this paper does not contradict, is that internal metrics are an input to assurance and never a substitute for validated output, and the evidence required rises with the stakes of the setting in which the output is used.

本分析的局限性

This reading rests on the arXiv version of the paper as posted, together with public coverage of its appearance on Apple′s research site, which was not independently verified. DiscoSign is research, not a product: nothing here describes a deployed

system, no vendor is making claims about it, and no NCG instrument has been applied to it. The criticisms above are not criticisms of the authors, who state most of these limitations themselves and did the field a service by publishing metrics that make discourse failure measurable at all. The argument is about what happens next, when a genuine research advance is compressed into a capability claim by people who did not read the limitations section.

这篇论文的有益之处在于,它使一个隐藏的问题变得可见。句子级系统一直以来在话语层面都存在失败;只是此前没有工具能够察觉这一点。如今这样的工具才刚刚起步,而它伴随而来的一项发现,更应被当作警示而非仅仅是结果来看待。让一个系统保持一致,并不能使其正确。它只会使其要么从头到尾正确,要么从头到尾错误,并且在任何人读到输出的只言片语之前,就已经决定了会是哪一种。

资料来源

1. Vasileios Baltatzis, Mert Inan, Connor Gillis, Raja Kushalnagar, Lorna Quandt, Leah Findlater and Colin Lea, "DiscoSign: Discourse-Aware Text to Sign Language Gloss Translation," arXiv:2609.02796, 2 September 2026, accepted at EMNLP 2026 Main Conference. https://arxiv.org/abs/2609.02796
2. DiscoSign, full text (HTML). https://arxiv.org/html/2609.02796
3. AI Global Wire, report of the paper's appearance on Apple Machine Learning Research, 11 September 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.

引用方式

Grizzle, H. M. (2026, September 11). 始终如一的错误:Apple DiscoSign 与手语人工智能. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003304