Large Language Models and NER

Better Results with Less Work

Rosamond Elizabeth Thalken, David Mimno, and Matthew Wilkens
presented by: Rebecca M. M. Hicke

Goal

Identify entities and relationships between identities in text segments.

Dale, striking a trail, turned his back to the fading afterglow and strode down the valley.

Common Methods

Language is variable and ambiguous!

- Regular Expressions
- Manual annotation
- Gazetteers (lists of entities)
- "Traditional" NER

Brittle

Time intensive

Lists never complete

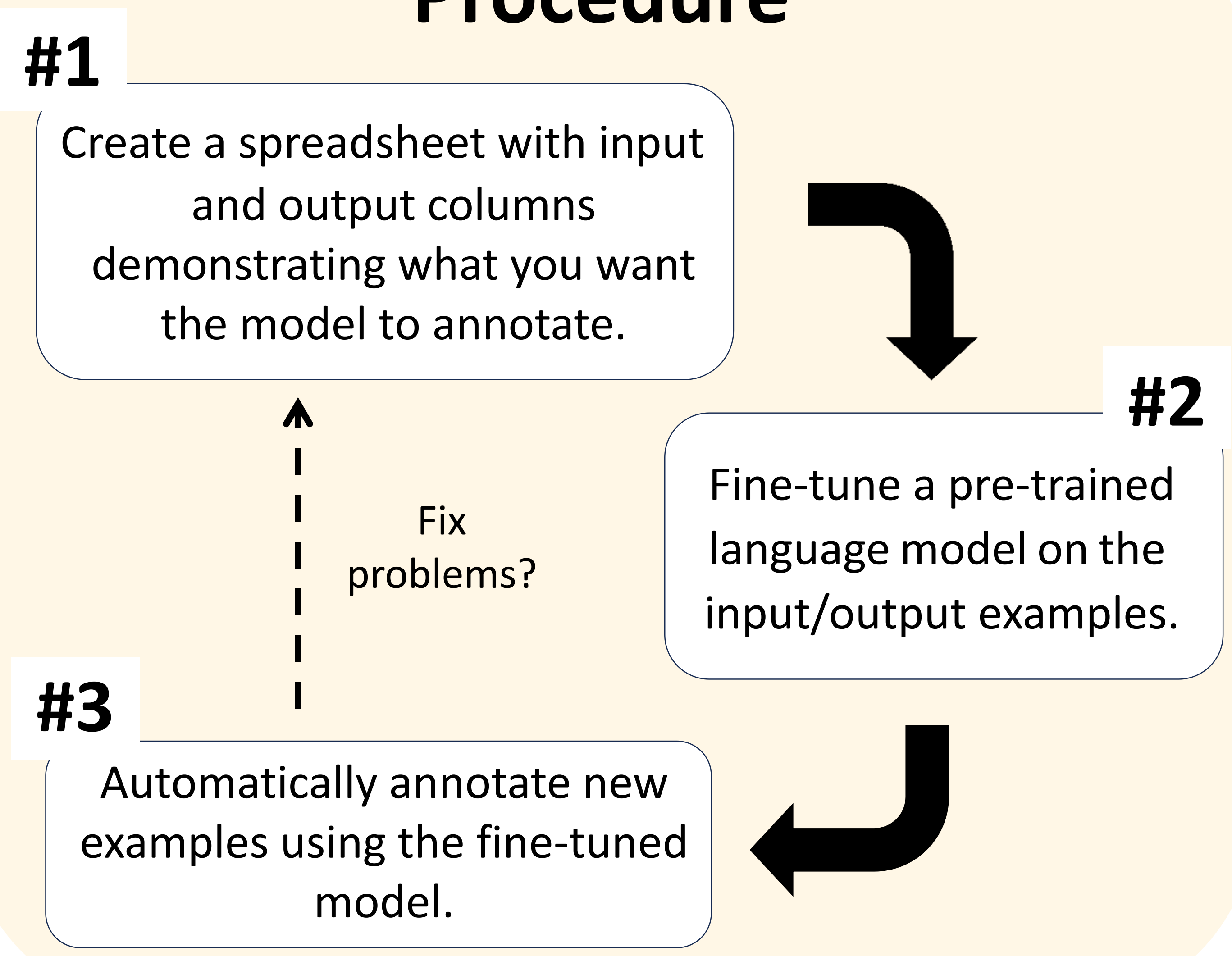
Doesn't adapt well to new kinds of text

Our Solution

Text-to-text generative large language models can be trained to identify entities or other patterns in text. These models possess a flexibility that the alternative methods don't.

They're better at doing what you want and not what you said.

Procedure



#1 Create a spreadsheet with input and output columns demonstrating what you want the model to annotate.

#2 Fine-tune a pre-trained language model on the input/output examples.

#3 Automatically annotate new examples using the fine-tuned model.

Fix problems?

Examples

Here are some interesting examples of inputs and outputs from our own experiments.

```
I believe  C.        Genus = Chrisops,
impar Rond.     →    Epithet = impar,
                       Author = Rond
```

The LLM has learned that `C.` is an abbreviation for `Chrisops`.

```
Even then -- in [the palace of
[the Sultan: 1] [himself: 1]: 2] --
[the three guardian priests: 3]
still kept [their: 3] watch in
secret.
```

An impressively accurate coreference annotation.

```
… there passed            … there passed
from [[our: 4]            from [our: 5]
highways: 5] [a    vs.    highways a
picturesque               picturesque
figure: 6].               figure.
```

A mistake: the LLM (right) has missed the nuances that the human annotator (left) has captured.

Sidebar: LLMs love punctuation!

We find that performance improves when we use an explicitly formatted structure for the output.

```
I believe C. impar Rond.
Ann. M, C. Gen. VII,           Genus = Chrysops,
460, to be the same as    →    Epithet = impar,
dispar; I have seen the        Author = Rond.
types.

His coats were                 [His: 1] coats were
execrable; his hat        →    execrable; [his: 1] hat
not to be handled.             not to be handled.
```

Check out this tutorial on using LLMs for NER!