Large Language Models

Philipp Koehn

24 October 2024

Recall: Statistical Machine Translation

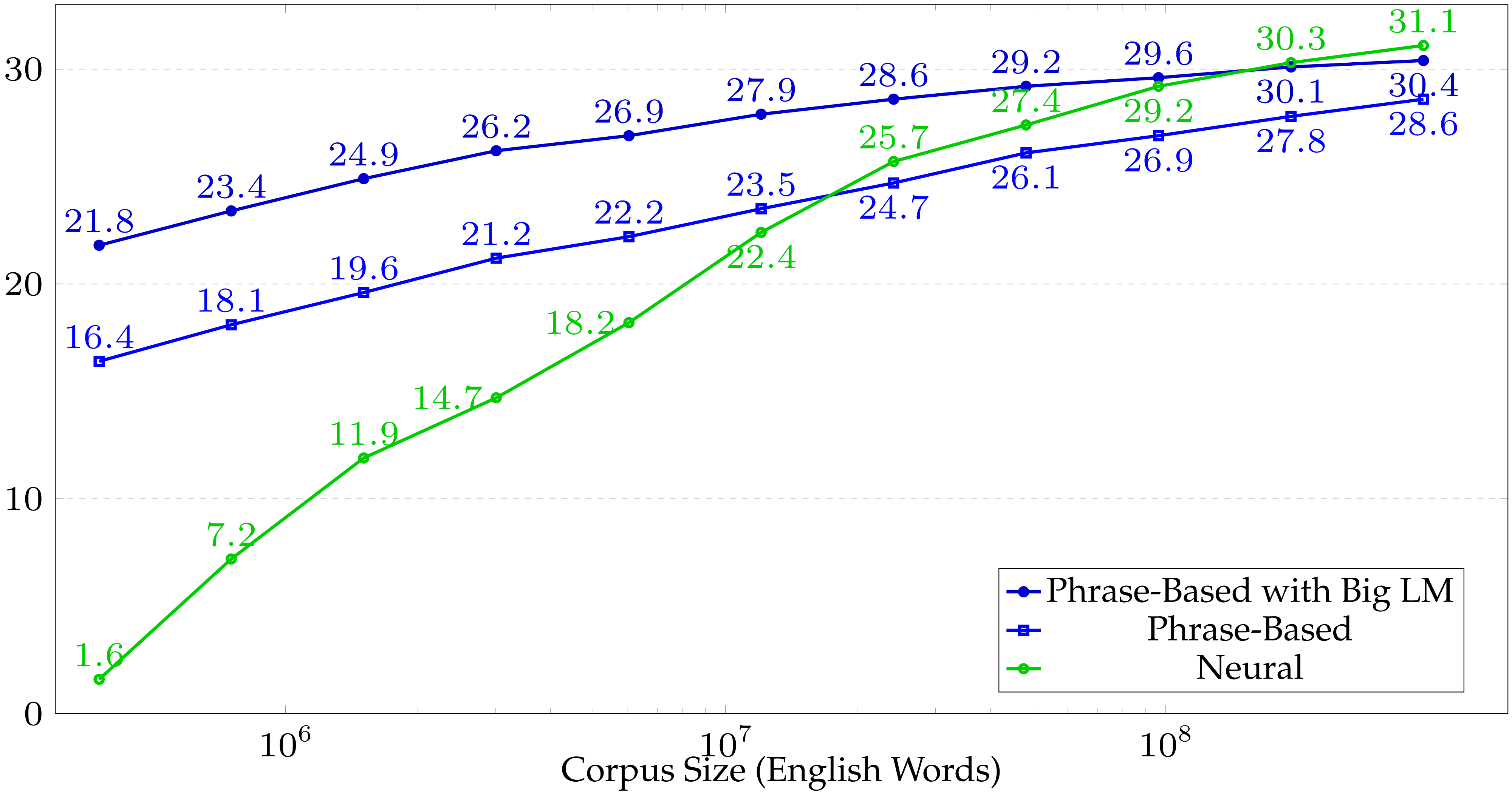
- Statistical Machine translation

$$\text{argmax}_e p(e|f) = \text{argmax}_e\, p(f|e)\, p(e)$$

- Combination of translation model $p(f|e)$ and language model $p(e)$
  - translation model ensures correct meaning
  - language model ensures fluency

Neural vs. Statistical Machine Translation

BLEU Scores with Varying Amounts of Training Data

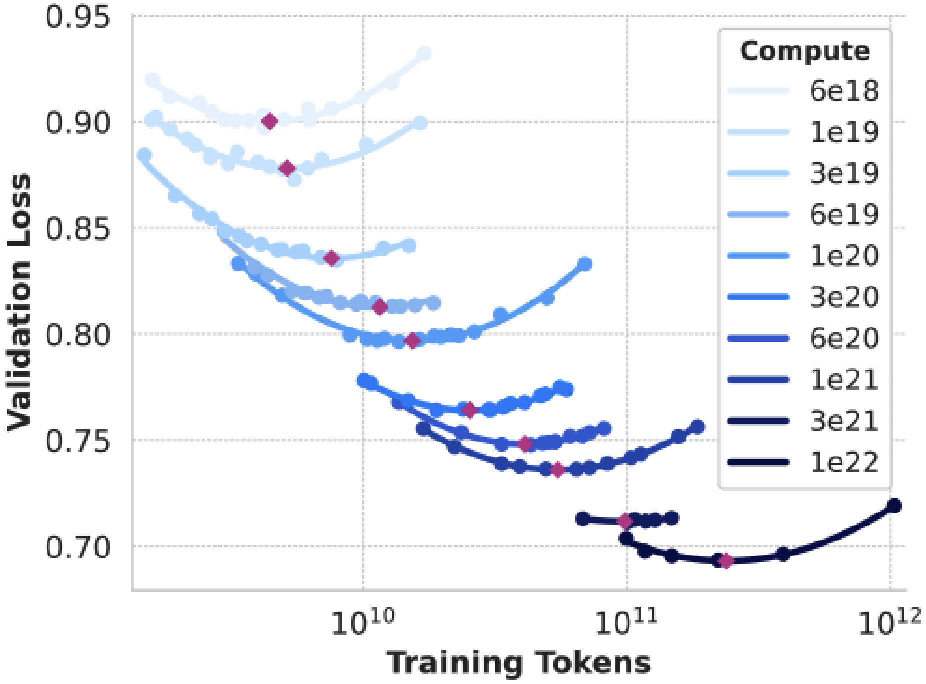


[from *Six Challenges for Neural Machine Translation*, 2017, Koehn and Knowles]

What Happened to the LM in MT?

- Edinburgh SMT system 2013: 126 billion token LM [Durrani et al., 2013]
- Fusion model: merge predictions from MT and LM [Gulcehre et al., 2015]
- Backtranslation: synthesize source side of monolingual data [Sennrich et al., 2017]
- mBART: Monolingual pretraining [Liu et al., 2020]
- None of them used data at the scale used in SMT
  LLMs finally do that now (since 2022)

Training



- Scaling laws: more data → bigger models → better performance
- Today: trillions of words → 10s to 100s of billions of parameters
- Llama3 405B: trained on 16,384 GPUs — available open source

Instruction Training

- Examples of requests and responses constructed by human annotators

- May be collected from actual user requests and edited by experts

- May be generated from existing data sets

Question Answering	**Summarization**	**Translation**
What is the highest mountain in the world? The highest mountain in the world is Mount Everest.	*Summarize the following paragraph into one sentence.* The Federal Reserve paused its campaign of interest rate increases for the first time in more than a year. But officials suggested that rates would rise more in 2023, as inflation remains ”well above” the central bank’s target. Summary: No interest rate rise for now but maybe later in the year.	*Translate from English to German.* English: My name is Ozymandias, King of Kings; Look on my Works, ye Mighty, and despair! German: Mein Name ist Ozymandias, König der Könige; Schau auf meine Werke, du Mächtiger, und verzweifle!

Adapting LLMs to MT

- Convert parallel data into chat format

> *Translate the following sentence from German to English.*
> *German: Das Haus ist groß.*
> *English: The house is big.*

- Use it as instruction data for fine-tuning large language model

- Our work: fine-tuning various LLMs with QLoRA [Zhang et al., 2023]
  - Llama2 and BLOOMZ show best results
  - fine-tuning with QLoRa sufficient

- Adapting for document-level translation [Wu et al., 2024]
- Adapting for specific domains [Zheng et al., 2024]
- BigTranslate: Adapting LLM for MT for over 100 Languages [Yang et al., 2023]

ALMA

- Advanced Language Model-based trAnslator [Xu et al., 2024]
- Two stage fine tuning
  - fine-tune on monolingual data for languages involved
  - fine-tune on parallel data
- Not much parallel data needed (only WMT development sets)
- Building on Llama2-7B, better translation performance than GPT3.5

Prompt Engineering

- How a task is presented to the language model matters
- Black art: often unclear what is in the training data
- *“Imagine you are...”*: creates a style or sets a context for generation
- Detailed instructions
  - *“Translate every word.”*
  - *“Do not provide any explanations.”*
  - *“Do not drop or add any information.”*

Translation Guidance by Specification

- Language models allow text generation with specifications
  - *“Write in the style of William Shakespeare.”*
  - *“Use formal language.”*
  - *“Write in the style of a marketing message.”*
  - *“Use the following terminology in your translation: ...”*

- Unclear, if this requires any additional model training
  (especially in the case of a very specific style, e.g., house style of a company)

- Alternatively, provide example of target language text in the expected style

Chain of Thought

- Language model has very limited working memory

- Complex reasoning may require several inference steps

What is 56 times 204? Think step by step.

I first multiply 50 and 200, which is 10,000.
Then I multiply 6 and 200, which is 1,200.
I first multiply 50 and 4, which is 200.
Then I multiply 6 and 4, which is 24.
Adding together 10,000 and 1,200 and 200 and 24 gives 11,424.

- Intermediate steps create working memory

Decomposition of Tasks

Write an outline for an argument that LLMs are great.
Outline:

1. LLMs have emerged as general problem solvers
2. LLMs have become publicly available
3. Many people find LLMs useful 4. ...

Write a paragraph about ”LLMs have emerged as general problem solvers”.

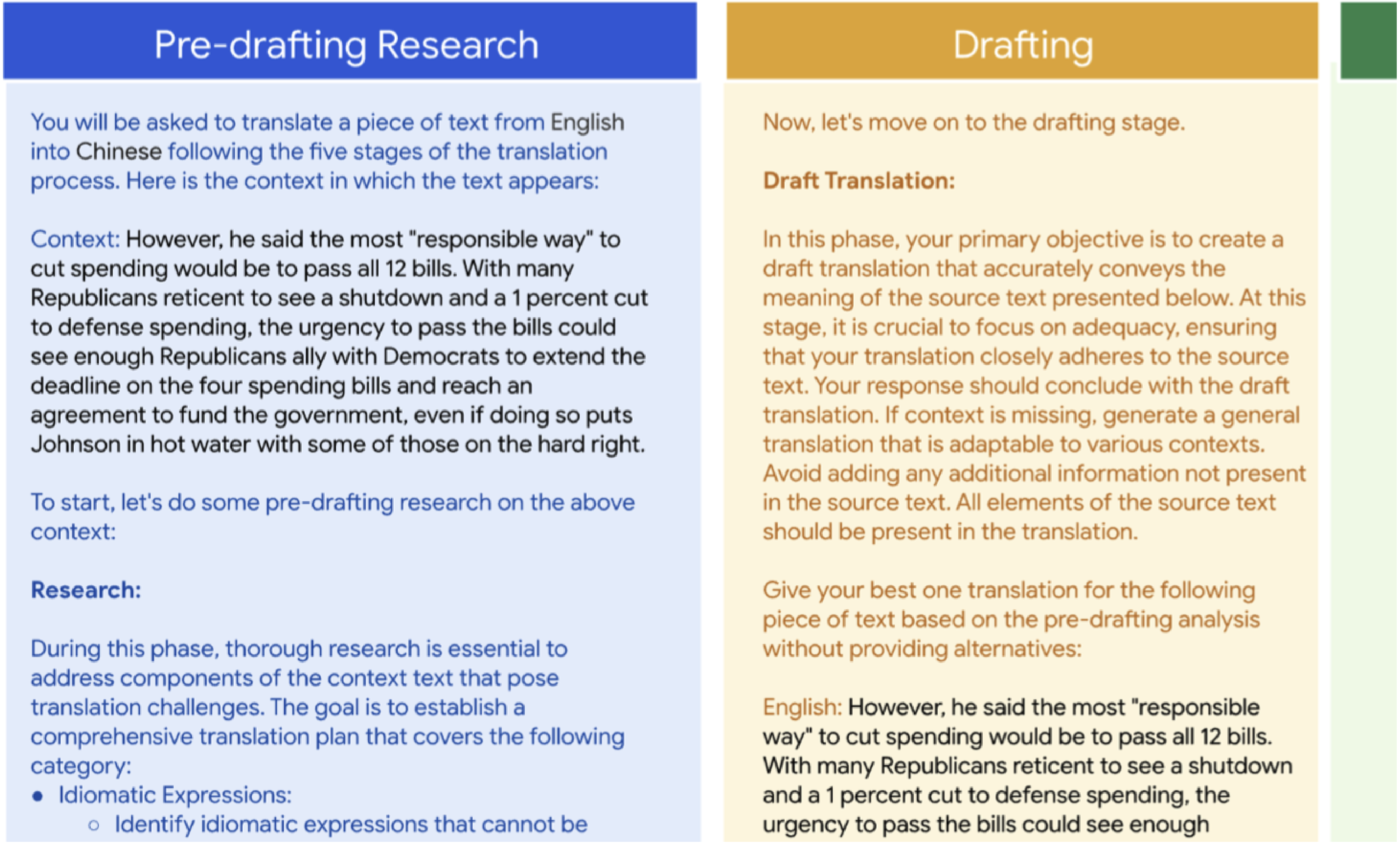
Language model (LLMs) have rapidly emerged as versatile general problem solvers across various domains. Leveraging vast datasets and sophisticated algorithms, LLMs exhibit remarkable capabilities in understanding, generating,...

Translating Step-by-Step

[Briakou et al.; Google, 2024]

- Inspired by the work process of professional translators
  - pre-drafting research: find hard translation problems
  - drafting: translation with focus on accuracy
  - refinement: improve fluency
  - proofreading: check for errors
- All these steps contribute to the quality of the translation

Translating Step-by-Step



Pre-drafting Research

You will be asked to translate a piece of text from English into Chinese following the five stages of the translation process. Here is the context in which the text appears:

Context: However, he said the most "responsible way" to cut spending would be to pass all 12 bills. With many Republicans reticent to see a shutdown and a 1 percent cut to defense spending, the urgency to pass the bills could see enough Republicans ally with Democrats to extend the deadline on the four spending bills and reach an agreement to fund the government, even if doing so puts Johnson in hot water with some of those on the hard right.

To start, let's do some pre-drafting research on the above context:

Research:

During this phase, thorough research is essential to address components of the context text that pose translation challenges. The goal is to establish a comprehensive translation plan that covers the following category:

- Idiomatic Expressions:
  - ○ Identify idiomatic expressions that cannot be

Drafting

Now, let's move on to the drafting stage.

Draft Translation:

In this phase, your primary objective is to create a draft translation that accurately conveys the meaning of the source text presented below. At this stage, it is crucial to focus on adequacy, ensuring that your translation closely adheres to the source text. Your response should conclude with the draft translation. If context is missing, generate a general translation that is adaptable to various contexts. Avoid adding any additional information not present in the source text. All elements of the source text should be present in the translation.

Give your best one translation for the following piece of text based on the pre-drafting analysis without providing alternatives:

English: However, he said the most "responsible way" to cut spending would be to pass all 12 bills. With many Republicans reticent to see a shutdown and a 1 percent cut to defense spending, the urgency to pass the bills could see enough

in-context learning

In-Context Learning

- Problem
  - language models are trained on very diverse language usage
  - it may be confused on what it is expected to do
- Solution: provide examples (“shots”) of the task in the prompt
- This has been shown to be successful even for new tasks

Multi-Shot Translation

- Provide examples in the prompt

> Translate from German to English. Here are some examples.
> German: Ein Hund bellt. English: A dog barks.
> German: Ein Schwein grunzt. English: A pig grunts.
> German: Eine Katze miaut. English: A cat meows.
> German: Ein Wolf heult. English: A wolf howls.
>
> Now translate the following sentence.
> German: Ein Vogel singt. English:

- This is the standard approach when prompting language models

Provide Text as Style Guidance

- We want to translate in a particular style, e.g., patents

> Translate in the style of a patent.
> Here is some example text of the style: According to an aspect of this invention, a method includes detecting a syntactic chunk in a first string in a first language, assigning a syntactic label to the detected syntactic chunk in the first string, aligning the detected syntactic chunk in the first string to a syntactic chunk in a second language string, said aligning based...
>
> Translate from German to English.
> German: Eine oder mehrere der folgenden Funktionen können ebenfalls enthalten sein.
> English:

Specify Terminology

- A common constraint on translation is company-specific terminology

- For example, legal domain
  - *Rechtswissenschaft* = *jurisprudence* (not *law*)
  - *Kläger* = *plaintiff* (not *prosecutor*)
  - *Strafe* = *sentence* (not *penalty*)

- Provide them in the prompt

> Translate from German to English.
> Use the following terminology in the translation...

- In reality not so simple: need to distinguish technical and casual use of terms