CSE 4392 Special Topics
Natural Language Processing

Language Models

2026 Spring

An Example

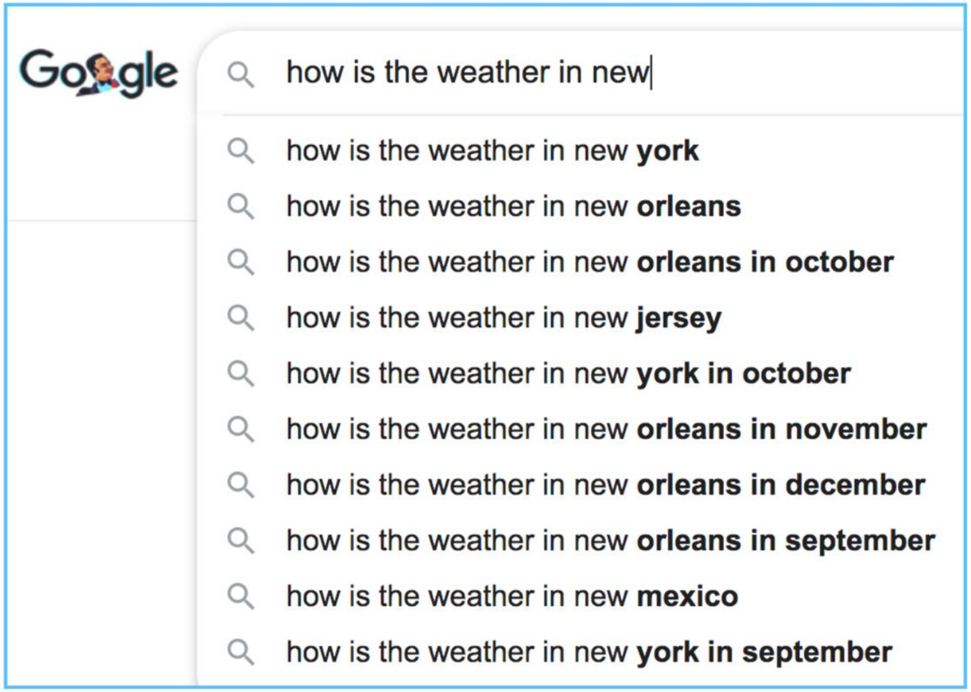
Today in Arlington, TX, it’s 45F and sunny.

vs.

Today in Arlington, TX, it’s 45F and blue.

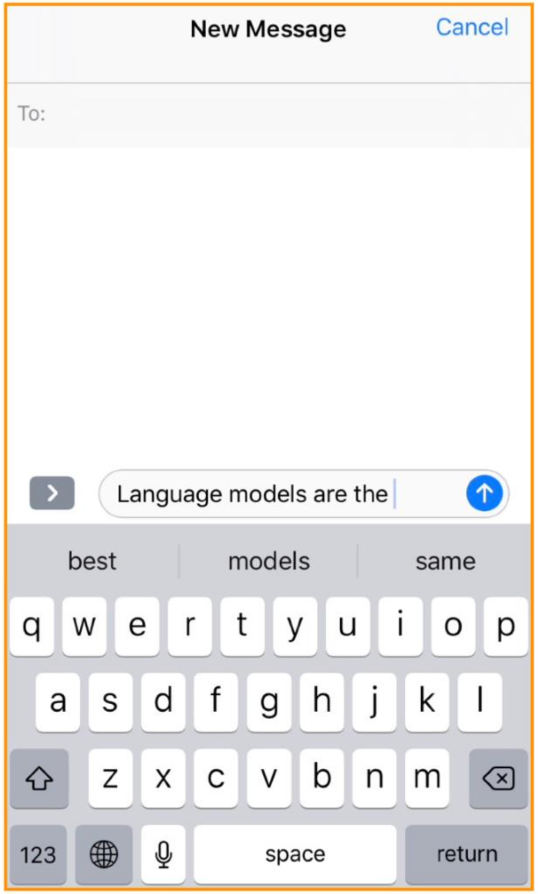
- Both are grammatical
- But which is more likely?

Language Models are Everywhere



how is the weather in new

- how is the weather in new **york**
- how is the weather in new **orleans**
- how is the weather in new **orleans in october**
- how is the weather in new **jersey**
- how is the weather in new **york in october**
- how is the weather in new **orleans in november**
- how is the weather in new **orleans in december**
- how is the weather in new **orleans in september**
- how is the weather in new **mexico**
- how is the weather in new **york in september**



Language models are the

And Many Applications

- Predicting words is important in many situations
  - Machine translation

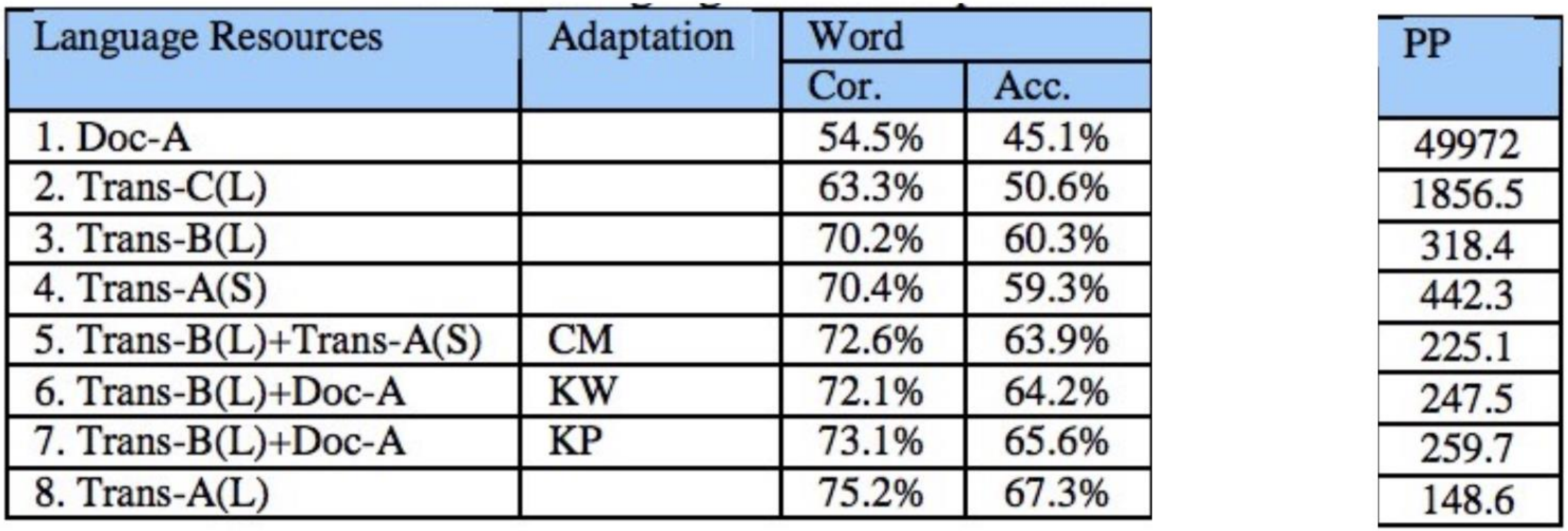
    P(a **smooth** finish) > *P*(a **flat** finish)

  - Speech recognition/Spell checking

    P(high school **principal**) > *P*(high school **principle**)

  - Information extraction, question answering

Impact on Downstream Applications



Language Resources	Adaptation	Word		PP
		Cor.	Acc.	
1. Doc-A		54.5%	45.1%	49972
2. Trans-C(L)		63.3%	50.6%	1856.5
3. Trans-B(L)		70.2%	60.3%	318.4
4. Trans-A(S)		70.4%	59.3%	442.3
5. Trans-B(L)+Trans-A(S)	CM	72.6%	63.9%	225.1
6. Trans-B(L)+Doc-A	KW	72.1%	64.2%	247.5
7. Trans-B(L)+Doc-A	KP	73.1%	65.6%	259.7
8. Trans-A(L)		75.2%	67.3%	148.6

documents vs transcripts (Miki et al. 2006)

New Approach to Language Modeling Reduces Speech Recognition Errors by Up to 15%

Ankur Gandhe

Principal, Applied Scientist

Alexa Speech group, Amazon

Quiz

- Below is the sizes of the various corpora used in the previous example of speech recognition.

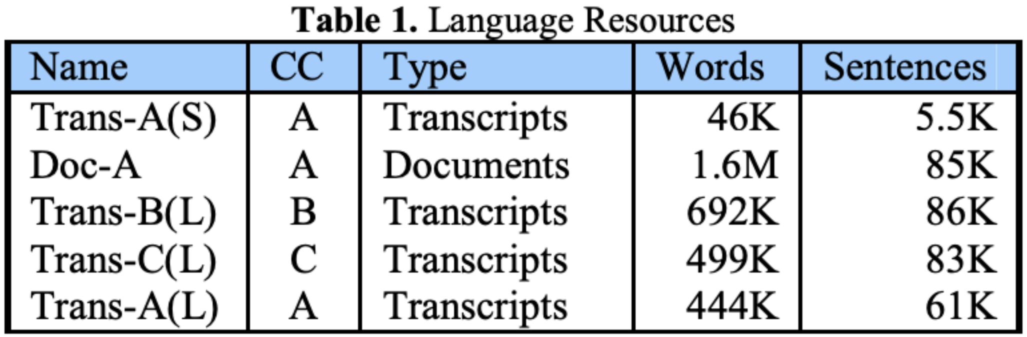


Table 1. Language Resources

Name	CC	Type	Words	Sentences
Trans-A(S)	A	Transcripts	46K	5.5K
Doc-A	A	Documents	1.6M	85K
Trans-B(L)	B	Transcripts	692K	86K
Trans-C(L)	C	Transcripts	499K	83K
Trans-A(L)	A	Transcripts	444K	61K

- Doc-A is call summaries collected at call center A, and Trans-A(S) is the *small* call transcripts collected at call center A, which is much smaller than Doc A. Why is the speech recognition model trained on Trans-A(S) much more accurate than Doc-A (59% vs 45%)?

What is a Language Model?

- Probabilistic model of a sequence of words.
  - How likely is a given phrase/sentence/paragraph/ document?

- Joint probability distribution:

$$P(w_1, w_2, \ldots, w_\mathrm{n})$$

Quiz

- Which of the following can be considered a sequence of words?
- a) a document
- b) a sentence
- c) a paragraph
- d) all of the above