Natural Language Processing

Natural Language Generation

Yangfeng Ji

Information and Language Processing Lab

Department of Computer Science

University of Virginia

https://uvanlp.org/

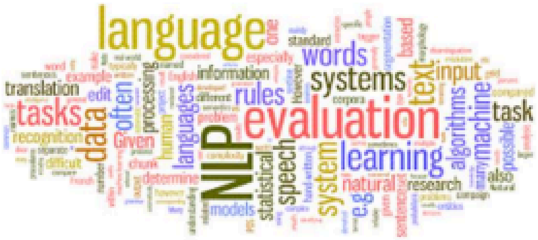
Outline

- Section I: Overview
- Section II: Decoding Algorithms
- Section III: Evaluation Strategies

Section I

Overview

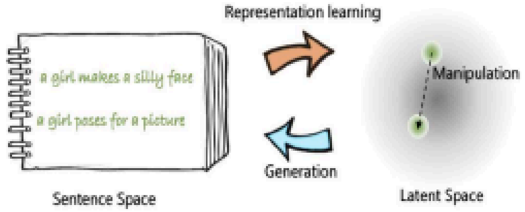
Text Generation



A sub-field in

natural language processing

Building software systems to produce ***coherent***, ***readable*** and **useful** written or spoken text.



Produces explanations, summaries, answers to questions, poems, dialogs, programs, …

(Ji et al., 2020)

Example: Machine Translation

Translate texts from one language to another language

[User] We will discuss several issues today!

[System] Nous discuterons plusieurs questions aujourd'hui.

[System] 我们今天将讨论几个问题！

[System] Bugun cok sayıda sorunu tartısacağız.

Example: Conversational Systems

Siri, Alexa, Google Assistant …

[USER] Where is my next appointment and am I free for lunch?

[Agent] Your next meeting is at 10:30 at City Center. Did you want me to book a place for lunch in downtown ?

Example: Document Summarization

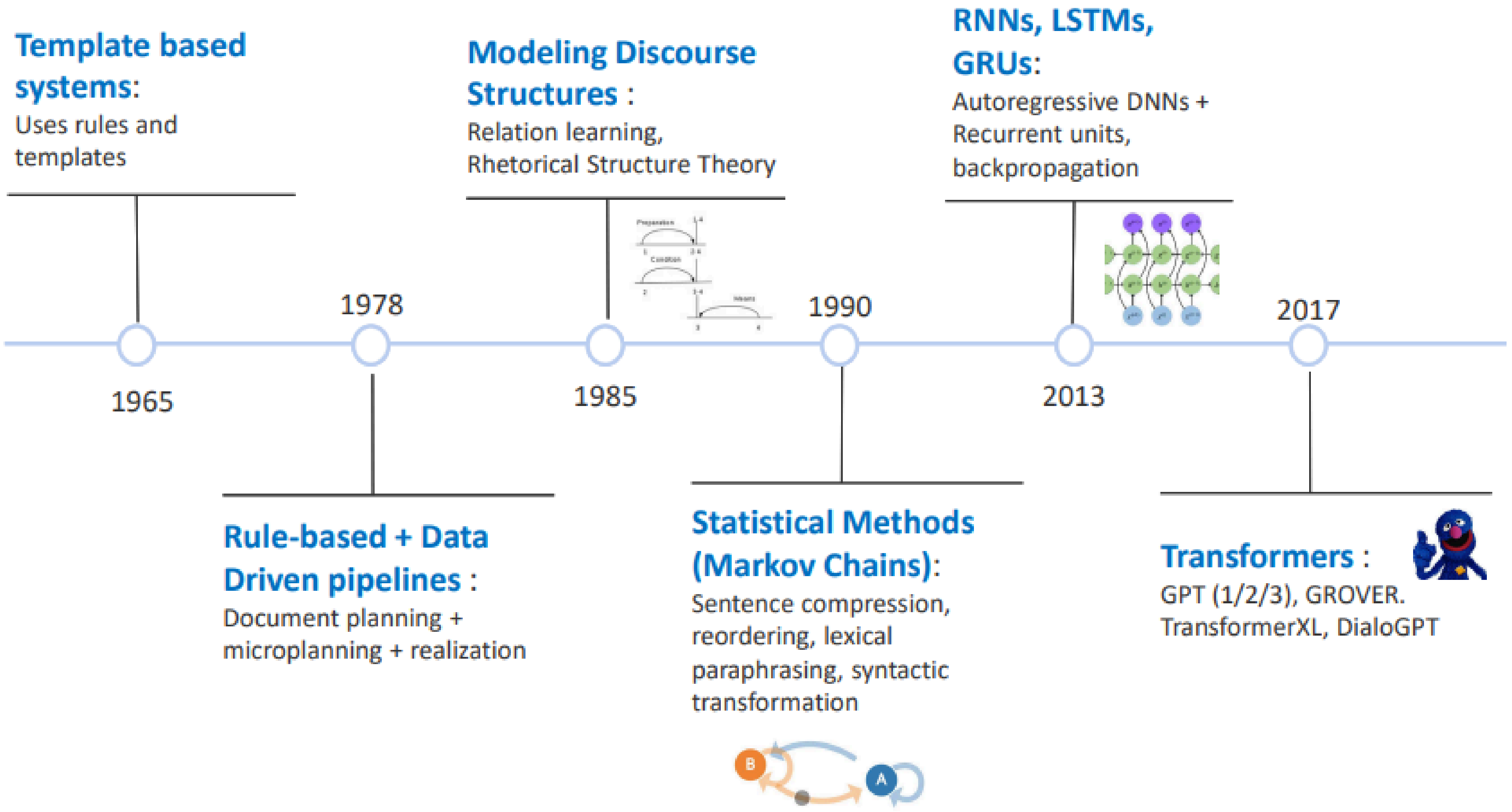
Extract or summarize the key information from one or multiple documents.

Summary

High Quality Content by WIKIPEDIA articles! Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. Resulting summary report allows individual users, so as professional information consumers, to quickly familiarize themselves with information contained in a large cluster of documents. In such a way, multi-document summarization systems are complementing the news aggregators performing the next step down the road of coping with information overload.

History of NLG

The timeline of NLG evolution



Template-based Generation

An example of generation template

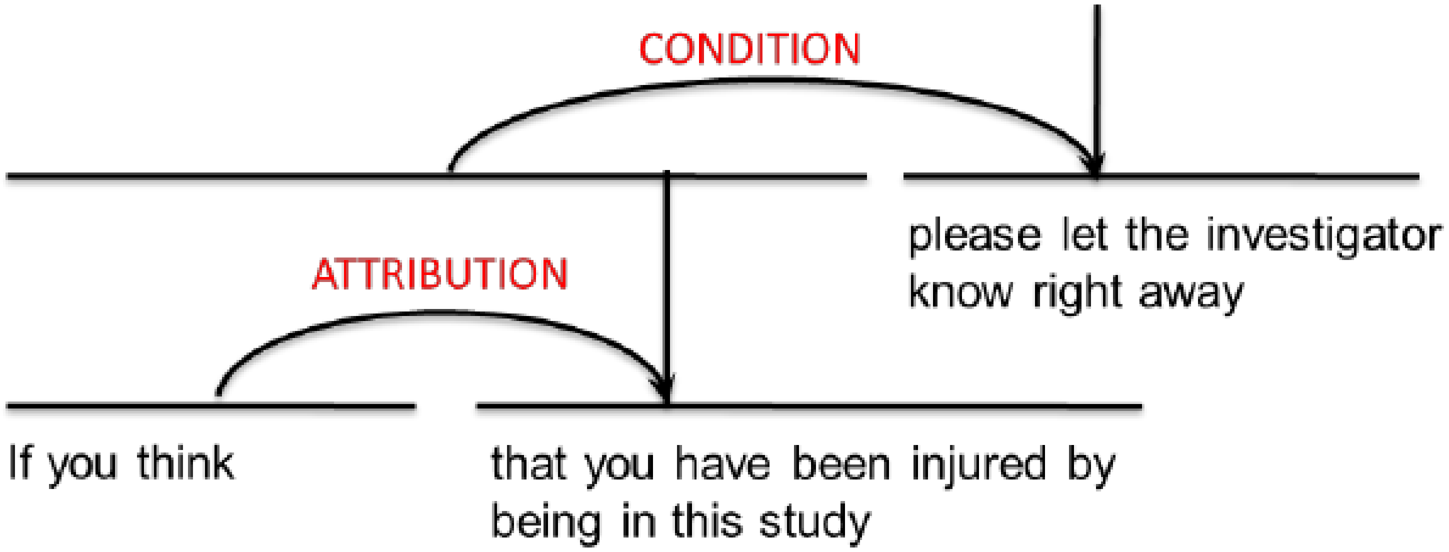
```
((EVAL member)
 (TEMPLATE verb-form
           ((process "be")
            (person (member person))
            (number (member number))
            (gender (member gender))) )
 (EVAL class)
 (PUNC "." left) )
```

Figure 2: A member-class Template.

(McRoy et al., 2000)

Linguistic-informed Generation

Using discourse structures or syntactic structures for generation



Rhetorical Structure Theory (RST) characterizes how different text units are semantically organized together to form a single coherent text

NLG Modeling

- Auto-regressive models
  - RNN LMs
  - GPT
- Sequence-to-sequence models
  - LSTM-based encoder-decoder
  - BART
- Copying mechanism
  - Pointer generator

Section II

Decoding Algorithms

Greedy Decoding

At each step, pick the word with the largest prediction probability

$$w_t \leftarrow \text{argmax } P(W_t \mid W_{1:t-1})$$

This often produces short and common responses

Context:

```
This is the best coffee I ever had, do you want to give it a try?
```

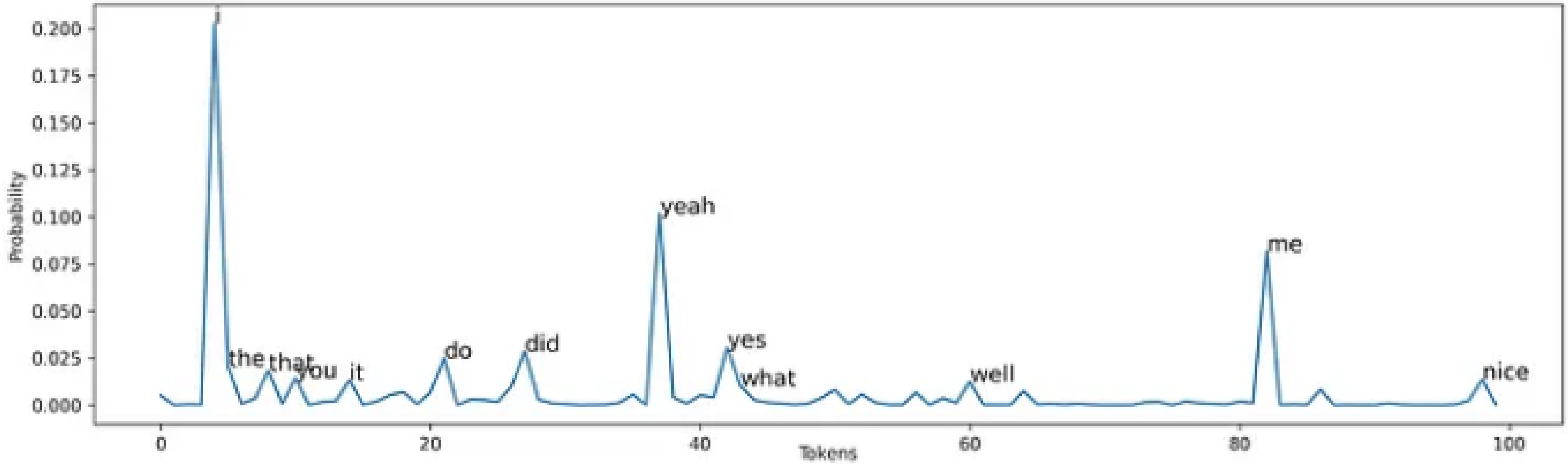
Response:

```
Okay.
```

Random Decoding/Sampling

Randomly pick a word, and the chance is proportion to its prediction probability

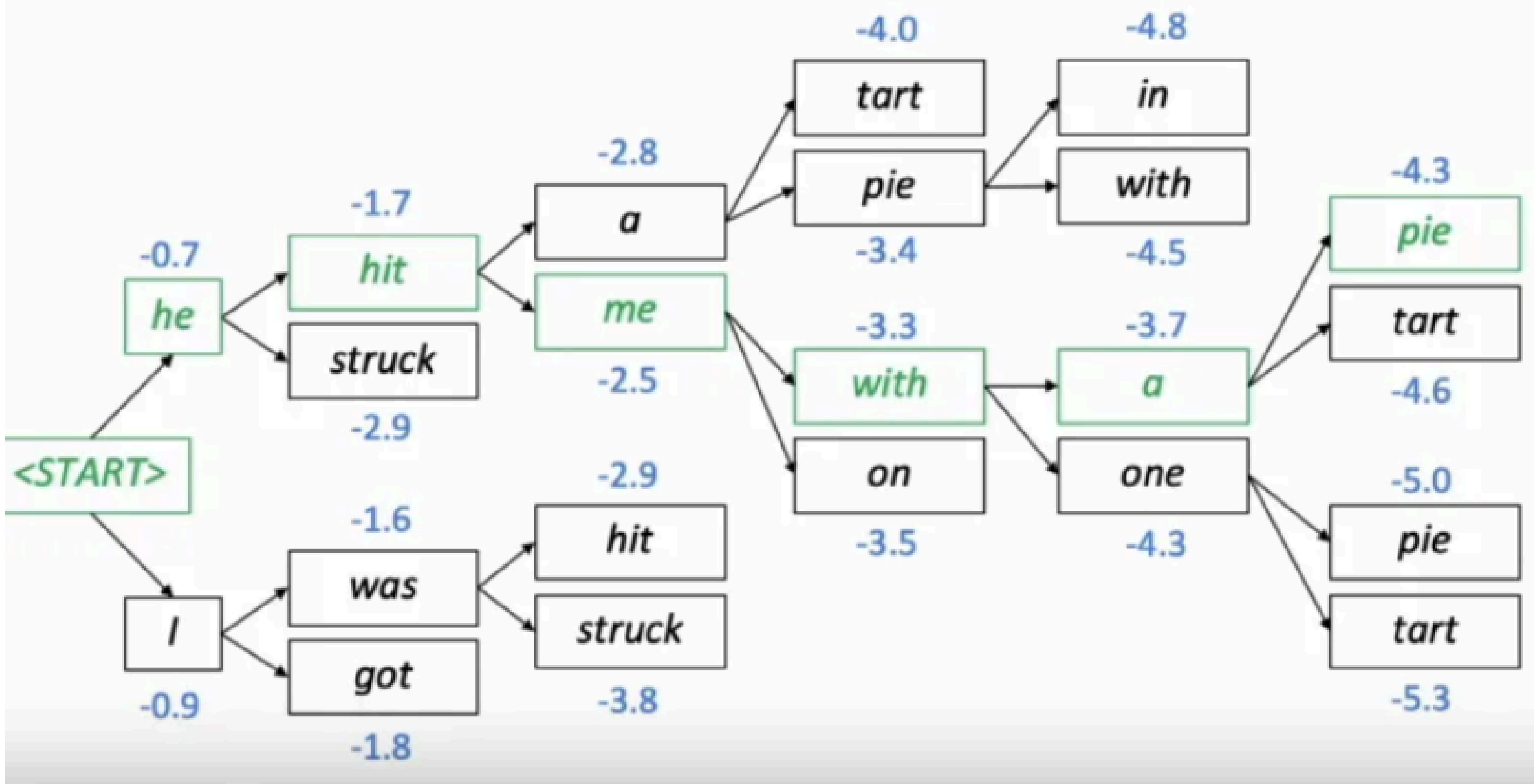
$$w_t \sim P(W_t \mid W_{1:t-1})$$



(Phy, 2020)

Top-k Decoding

The decoding illustration with $k = 2$

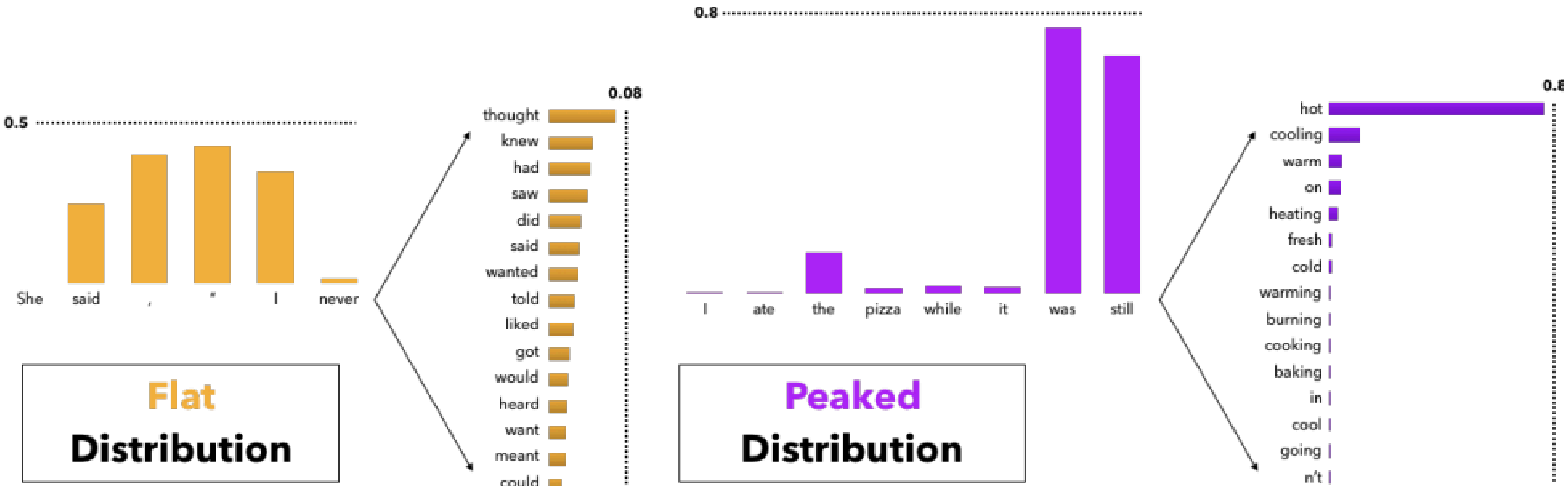


(Chen, 2020)

Top-p Decoding

Identify the top words that their probability accumulation is larger than p

$$\{w_1^{(1)}, \ldots, w_1^{(k)}\} \leftarrow \sum_{w^{(1)},\ldots,w^{(k)} \in V^{(p)}} P(w^{(i)} \mid w_{1:t-1}) > p$$



This is also called *Nucleus Sampling*.

Sampling Temperature

Assume $\{\alpha_k\}_{k=1}^{|V|}$ are the logits, we have the prediction probability as

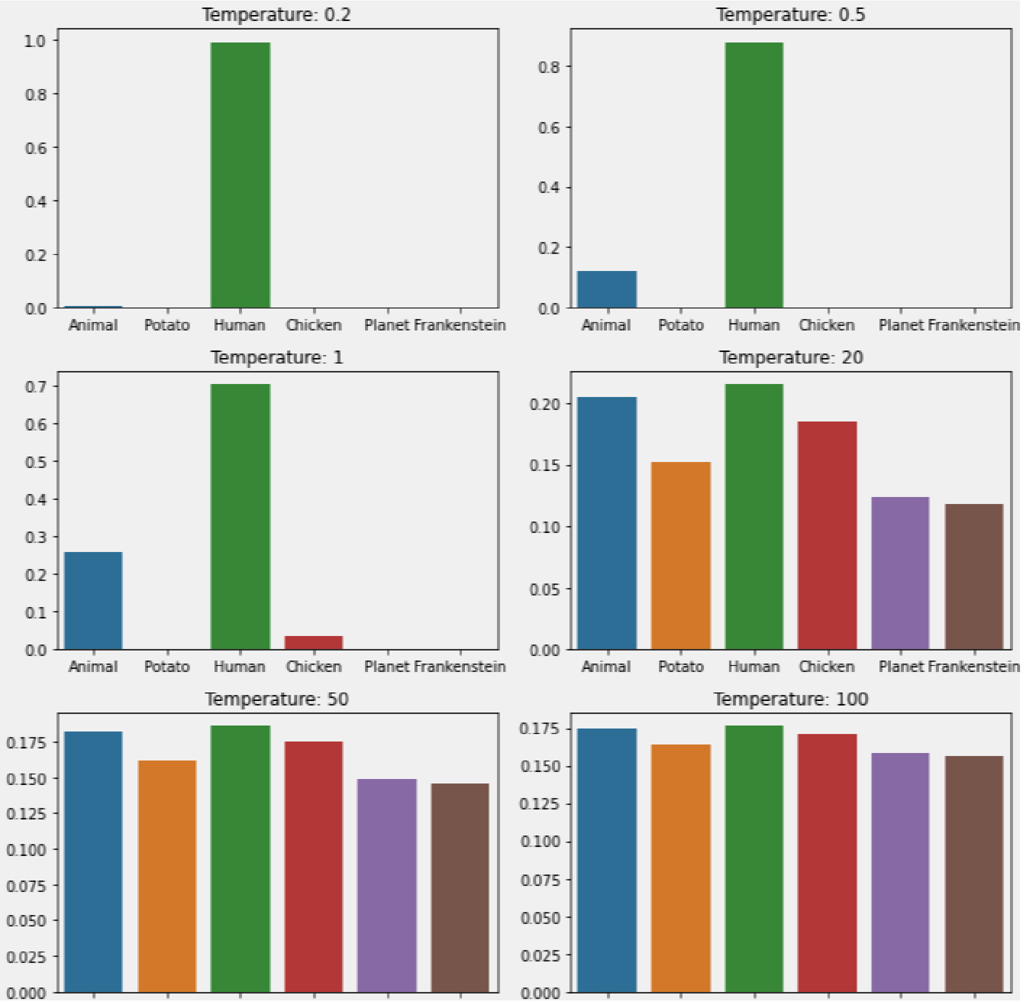
$$\text{softmax}(\alpha_k) = \frac{\exp(\alpha_k)}{\sum \exp(\alpha_k)}$$

With temperature τ, we have

$$\text{softmax}(\alpha_k/\tau) = \frac{\exp(\alpha_k/\tau)}{\sum \exp(\alpha_k/\tau)}$$

Lower temperature will lead to more deterministic sampling results.

Sampling Temperature (II)

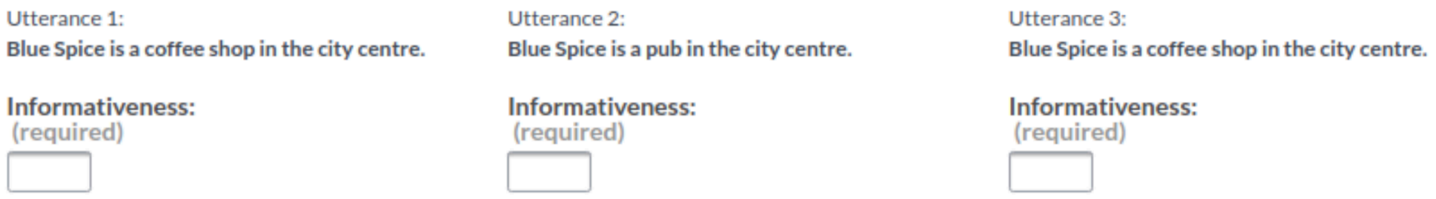


Section III

Evaluation Strategies

Human Evaluation

- Evaluation dimensions
  - Examples: fluency, coherence, correctness, factuality, etc.
- Format of the evaluation
  - Single sample evaluation with a Likert scale
  - Pairwise comparison
  - Ranking



(Celikyilmaz et al., 2021)

Concerns of Human Evaluation

There are some factors that make human evaluation results hard to reproduce by other researchers

- Number of participants
- Education background of participants
- Question design
  - Framing of the questions (Schoch et al., 2022)
- ...

Automatic Evaluation: BLEU

BLEU is originally designed to evaluate machine translation results

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where

- p_n: n-gram precision
- w_n: weight for n-gram precision, usually, $w_n = 1/4$
- BP: brevity penalty
- N: the largest length of n-gram, usually, $N = 4$

(Papineni et al., 2002)

n-Gram Precision

Candidate:

```
the the the the the the the
```

Reference:

```
The cat is on the mat
```

- Uni-gram precision: $2/7$
- Bi-gram precision: $0/7$

Automatic Evaluation: BLEU

The brevity penalty is introduced to penalize shorter generated (translated) text

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

where

- c is the length of the generated text
- r is the length of the reference text

Automatic Evaluation: ROUGE

ROUGE is originally designed for evaluating document summary

$$\text{ROUGE-}N = \frac{\sum_{S\in\mathcal{S}}\sum_{\text{n-gram}\in S}\text{Matched-count}(\text{n-gram})}{\sum_{S\in\mathcal{S}}\sum_{\text{n-gram}\in S}\text{Count}(\text{n-gram})}$$

This metric is defined by counting the matched n-grams from a generated summary.

Other variants

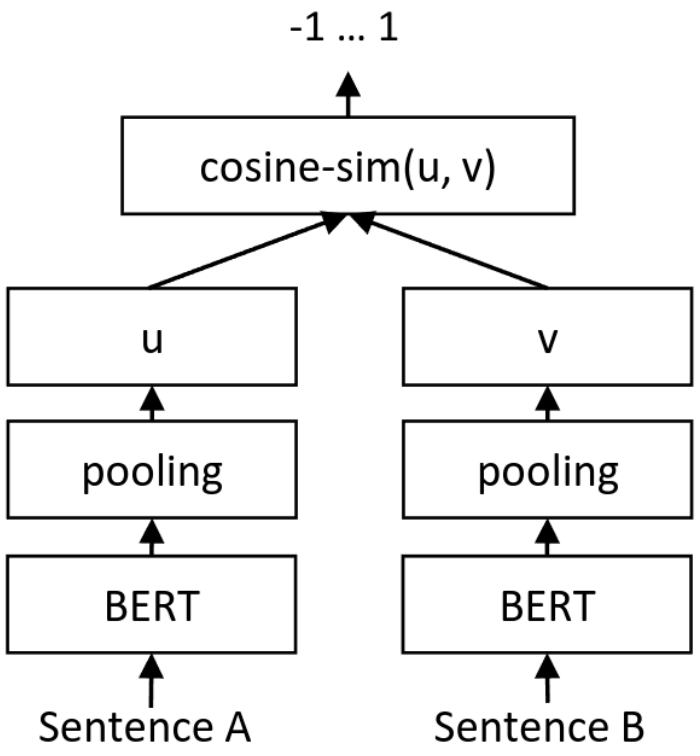
- ROUGE-L
- ROUGE-S

(Lin et al., 2004)

Automatic Evaluation: Neural network based

Consider evaluation as a similarity measurement problem by computing a score based on the generated text and the reference text.

Reimers and Gurevych, 2019



NLG Evaluation

One of the efforts on building a unified evaluation platform

The 💎 GEM Benchmark: Natural Language Generation, its Evaluation and Metrics

**Sebastian Gehrmann,[9,*] Tosin Adewumi,[20,21] Karmanya Aggarwal,[14]
Pawan Sasanka Ammanamanchi,[15] Aremu Anuoluwapo,[21,38] Antoine Bosselut,[28]
Khyathi Raghavi Chandu,[2] Miruna Clinciu,[7,11,35] Dipanjan Das,[9] Kaustubh D. Dhole,[1]
Wanyu Du,[42] Esin Durmus,[5] Ondřej Dušek,[3] Chris Emezue,[21,30] Varun Gangal,[2]
Cristina Garbacea,[39] Tatsunori Hashimoto,[28] Yufang Hou,[13] Yacine Jernite,[12] Harsh Jhamtani,[2]
Yangfeng Ji,[42] Shailza Jolly,[6,29] Mihir Kale,[9] Dhruv Kumar,[44] Faisal Ladhak,[4] Aman Madaan,[2]
Mounica Maddela,[8] Khyati Mahajan,[34] Saad Mahamood,[32] Bodhisattwa Prasad Majumder,[37]
Pedro Henrique Martins,[16] Angelina McMillan-Major,[43] Simon Mille,[26] Emiel van Miltenburg,[31]
Moin Nadeem,[22] Shashi Narayan,[9] Vitaly Nikolaev,[9] Rubungo Andre Niyongabo,[21,36]
Salomey Osei,[19,21] Ankur Parikh,[9] Laura Perez-Beltrachini,[35] Niranjan Ramesh Rao,[24]
Vikas Raunak,[23] Juan Diego Rodriguez,[41] Sashank Santhanam,[34] João Sedoc,[25]
Thibault Sellam,[9] Samira Shaikh,[34] Anastasia Shimorina,[33] Marco
Antonio Sobrevilla Cabezudo,[40] Hendrik Strobelt,[13] Nishant Subramani,[17,21] Wei Xu,[8]
Diyi Yang,[8] Akhila Yerukola,[27] Jiawei Zhou[10]**

Thank You!