Generative AI towards Scientific Discovery

Jiahuan Pei

Overview

AI

Research Experience

New Opportunities

Natural Languages

- NLP & IR
- Generative language models
- Diversity & Uncertainty

Natural Sciences

- Computational Molecular Biology
- Neuroscience

What is a language model?

- Probability distribution over words or word sequences
  - Given word sequence as context (observation) in a given “language”, how likely is a word can be generate (prediction) ?

$$P(w_t \mid \text{context})\, \forall t \in V.$$

- An example

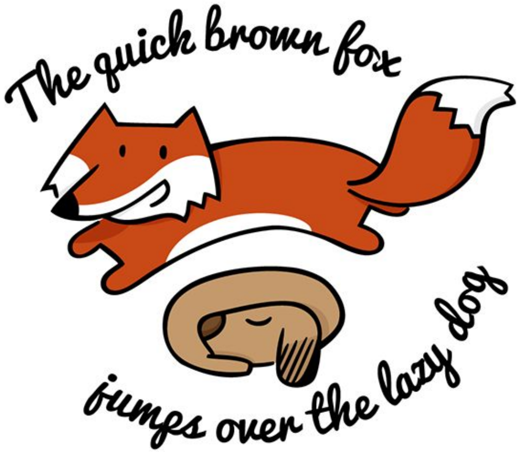
  p1 = P(“a quick brown fox”)

  p2 = P(“fox quick a brown”)

  p3 = P(“un chien quick brown”)

  p4 = P(“un chien brun rapide”)

  - NLP: p1 > p2 > p3 > p4
  - IR: p1==p2



Multi-task Dialogue Generation

- Unified context-to-text generation
  - 3 Tasks: NLU, DPL, NLG
  - 2 Types of Models
    - Causal language model

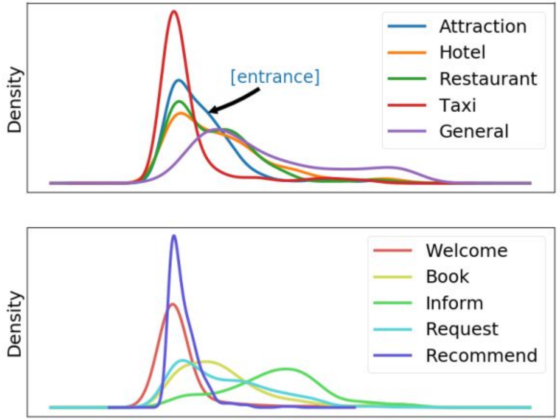
$$p_\theta(X^i_{1:n}) = \prod_{j=1}^{n}(x^i_j|X^i_{1:j-1}).$$

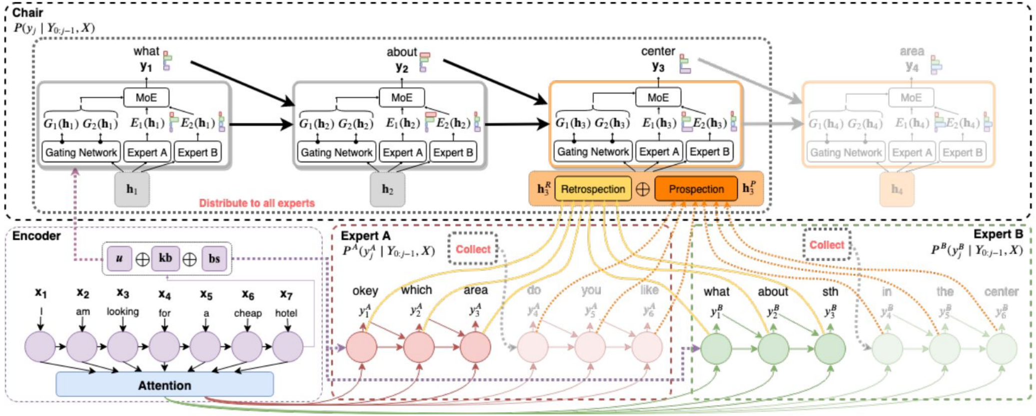
    - Conditional casual language model

$$p_\theta(Y_{1:m}|H_{1:n}) = \prod_{j=1}^{m} p_\theta(y_j|Y_{1:j-1}, H_{1:n}).$$

Yan, G., **Pei, J.**, Ren, P., et al (2022, July). ReMeDi: Resources for Multi-domain, Multi-service, Medical Dialogues. In SIGIR 2022.

Multi-expert Dialogue Response Generation

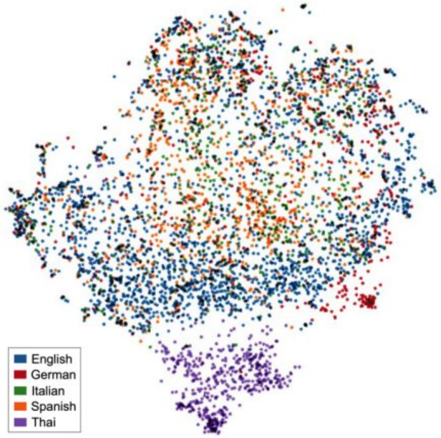


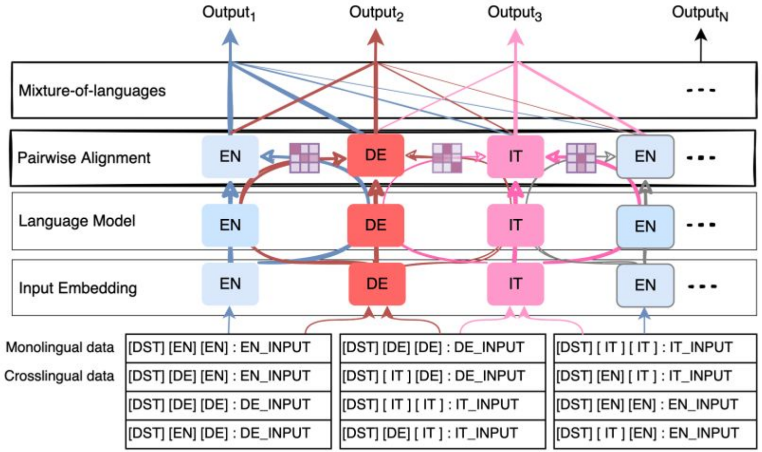


- A chair-expert model
- Retrospective and prospective collaboration mechanisms
- A global-local learning scheme

J. Pei, P. Ren, C. Monz, M. de Rijke. MoGNet: Retrospective and Prospective Mixture-of-Generators for Task-oriented Dialogue Response Generation. ECAI 2020. (first author, CCF B conference).

Multi-lingual Dialogue Generation

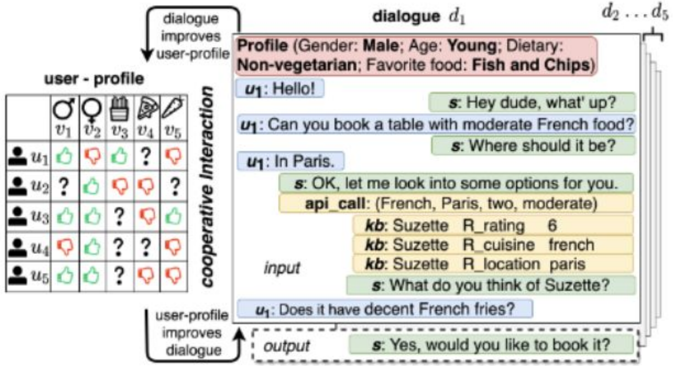


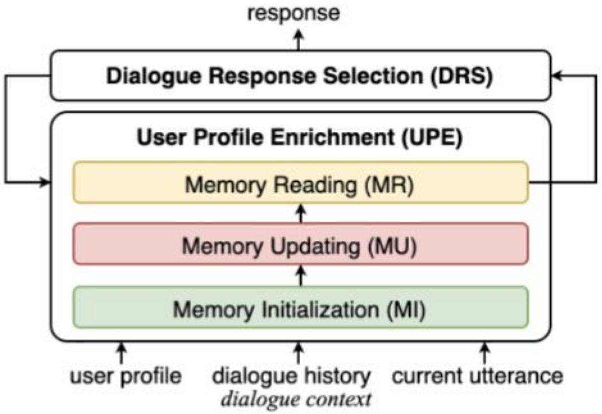


- A unified generation framework with mixture-of-language routing for Multilingual TDSs.
- Benefits from
  - Multilingual data argumentation;
  - Language characteristic modelling;
  - Mixture-of-language routing.

J. Pei, G. Yan, P. Ren, and M. de Rijke. Mixture-of-languages routing for multilingual task-oriented dialogue systems. TOIS 2023 (Under review).

Personalized Dialogue Response Selection

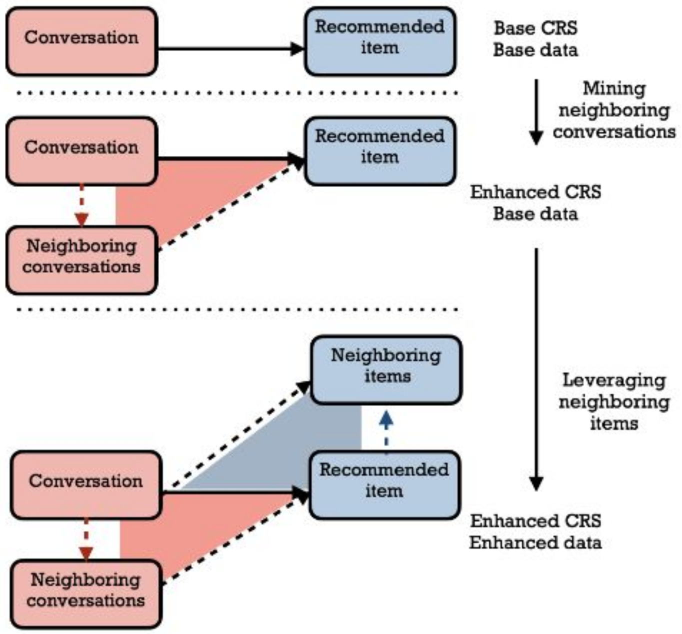


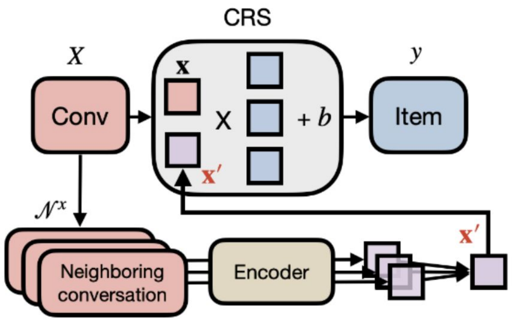


- A closed loop cooperative paradigm
  - Dialogue helps to enrich the user-item interactions.
  - Enriched user-item interactions help to select a better response.
- A learning algorithm with multiple hops

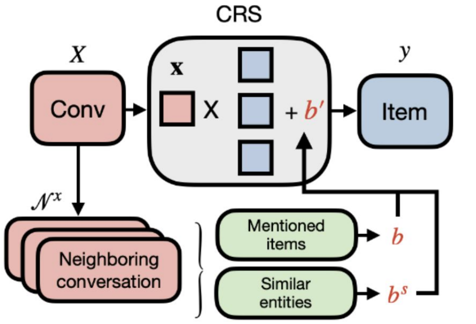
J. Pei, P. Ren, C. Monz, M. de Rijke. A Cooperative Memory Network for Personalized Task-oriented Dialogue Systems with Incomplete User Profiles. TheWebConf 2021. (first author, CCF A conference).

Neighboring Augmentation for Conversational Recommendation





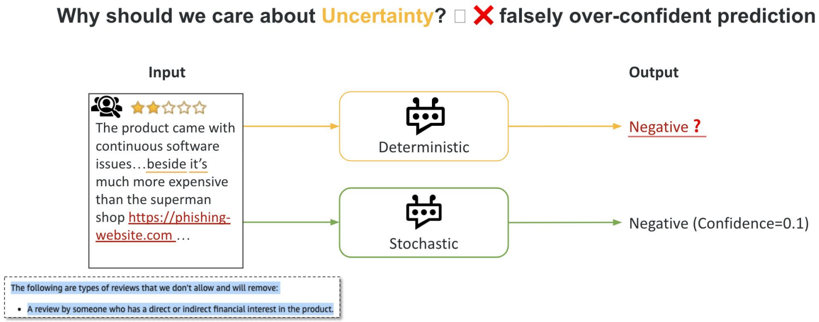
(b) CRS with enhanced user representation.

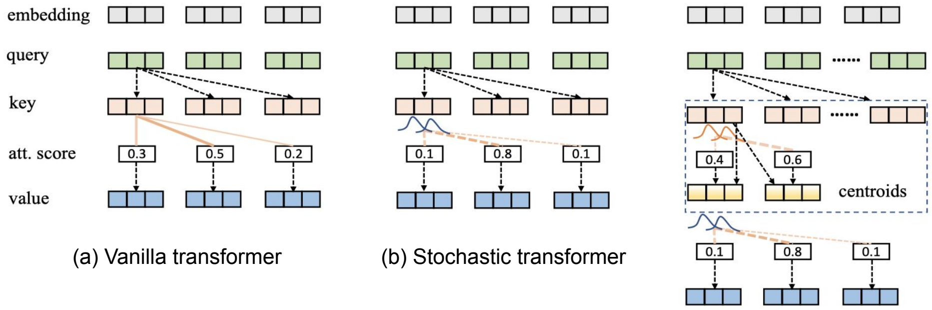


(c) CRS with enhanced user preference.

Liu, Y., **Pei, J.**, de Rijke, M. Augmentation with Neighboring Information for Conversational Recommendation (Submitted to SIGIR 2023)

Uncertainty Estimation: Stochastic Transformers



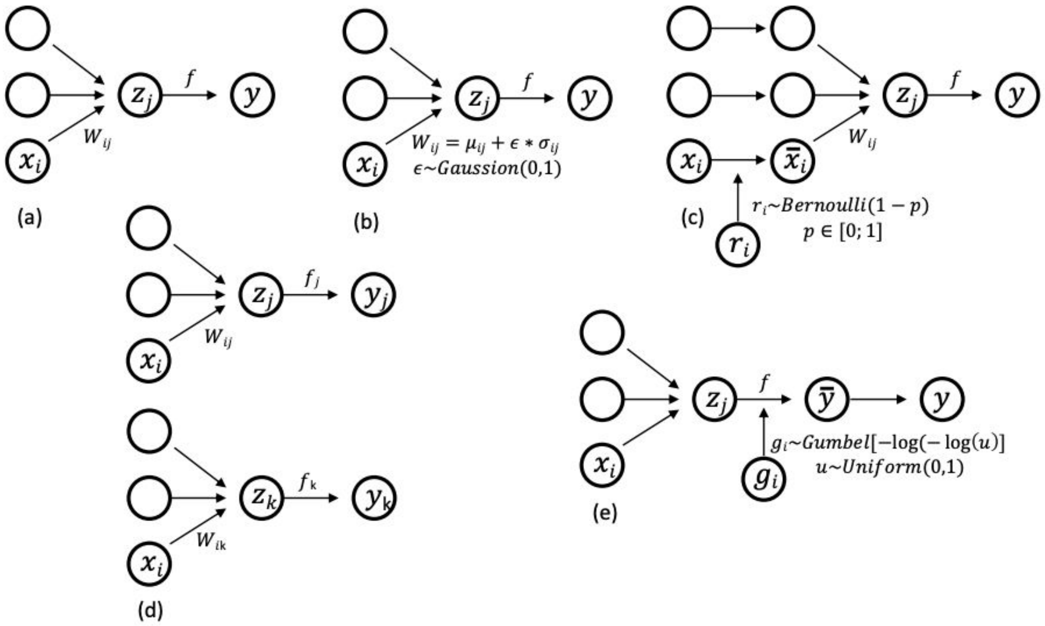


(a) Vanilla transformer (b) Stochastic transformer (c) Hierarchical stochastic transformer

- Enable transformers with uncertainty estimation while retain the original predictive performance.
- STO-TRANS has difficulties in the trade-off between in-domain and out-of-domain performance.

Pei, J., Wang, C., Szarvas, G. Transformer uncertainty estimation with hierarchical stochastic attention. AAAI 2022.

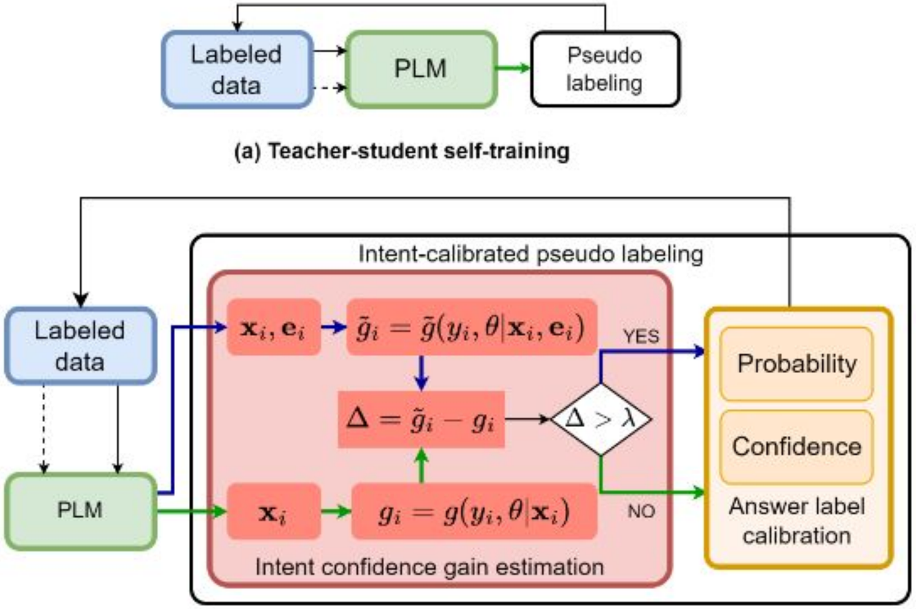
Uncertainty Estimation: Stochastic Transformers



(a) Deterministic neural network;
(b) Bayesian neural network;
(c) Variational dropout;
(d) Ensemble;
(e) Gumbel-Softmax trick

Pei, J., Wang, C., Szarvas, G. Transformer uncertainty estimation with hierarchical stochastic attention. AAAI 2022.

Intent-calibrated Self-training for Answer Selection



(a) Teacher-student self-training

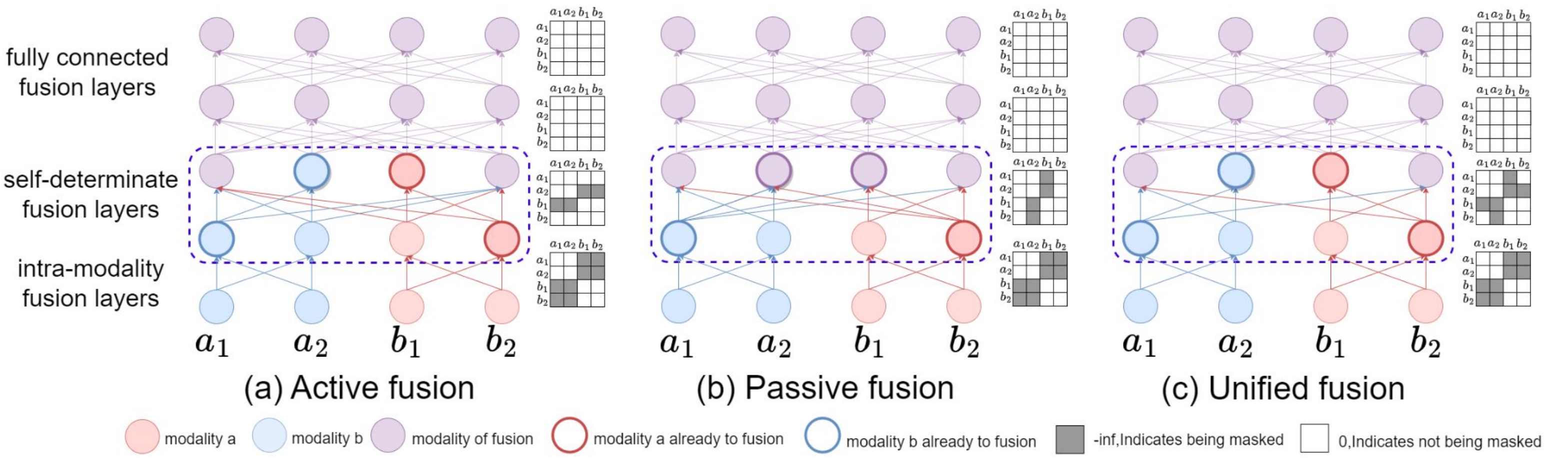
$$\Delta = \tilde{g}(y_i, \beta|\mathbf{x}_i, \mathbf{e}_i) - g(y_i, \beta|\mathbf{x}_i)$$

$$\begin{aligned} g(y_i, \beta|\mathbf{x}_i) &= \mathbf{H}[y_i|\mathbf{x}_i]] - \mathbf{E}_{P(\beta)}[\mathbf{H}[y_i|\mathbf{x}_i;\beta]] \\ &\approx \mathbf{H}[\mathbf{E}_{P(\beta)}[y_i|\mathbf{x}_i;\beta]] - \mathbf{E}_{P(\beta)}[\mathbf{H}[y_i|\mathbf{x}_i;\beta]] \\ &\approx -\frac{1}{T}\sum_{t=1}^{T} p_t \cdot \log \frac{1}{T}\sum_{t=1}^{T} p_t \\ &\quad + \frac{1}{T}\sum_{t=1}^{T} p_t \cdot \log p_t. \end{aligned}$$

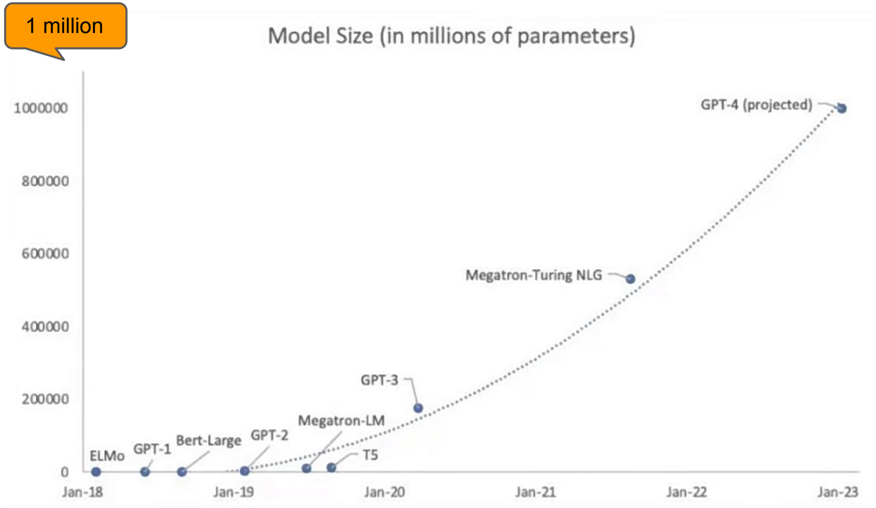
$$\begin{aligned} \tilde{g}(y_i, \beta|\mathbf{x}_i, \mathbf{e}_i) &= \mathbf{H}[y_i|\mathbf{x}_i, \mathbf{e}_i] - \mathbf{E}_{P(\beta)}[\mathbf{H}[y_i|\mathbf{x}_i, \mathbf{e}_i;\beta]] \\ &\approx \mathbf{H}[\mathbf{E}_{P(\beta)}[y_i|\mathbf{x}_i, \mathbf{e}_i;\beta]] - \mathbf{E}_{P(\beta)}[\mathbf{H}[y_i|\mathbf{x}_i, \mathbf{e}_i;\beta]] \\ &\approx -\frac{1}{T}\sum_{t=1}^{T} p_i^e \tilde{p}_t \cdot \log \frac{1}{T}\sum_{t=1}^{T} p_i^e \tilde{p}_t \\ &\quad + \frac{1}{T}\sum_{t=1}^{T} p_i^e \tilde{p}_t \cdot \log p_i^e \tilde{p}_t. \end{aligned}$$

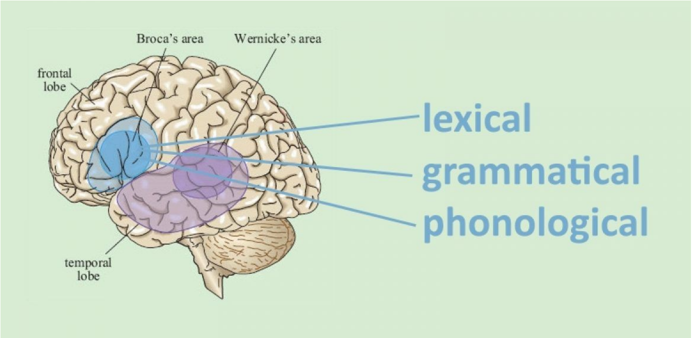
Deng, W., **Pei, J.**, Ren, P. Intent-calibrated Self-training for Answer Selection in Open-domain Dialogues (Submitted to TACL 2023)

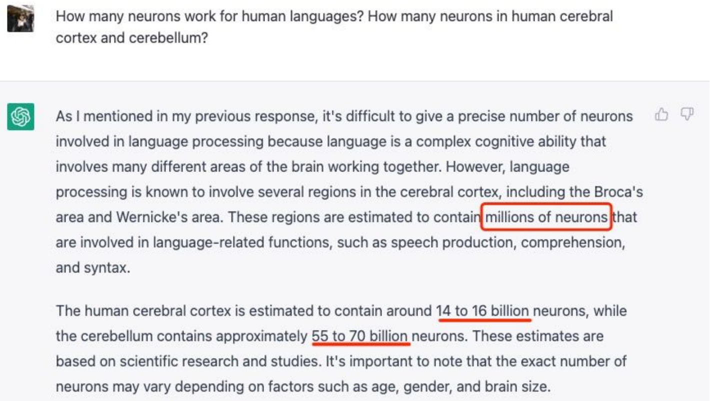
Self-determinate Multi-modality Fusion for Transformers



🔥 Milestone: AI-Generated Content (AIGC)







Overview

AI

Research Experience

Interdisciplinary Opportunities

Natural Languages

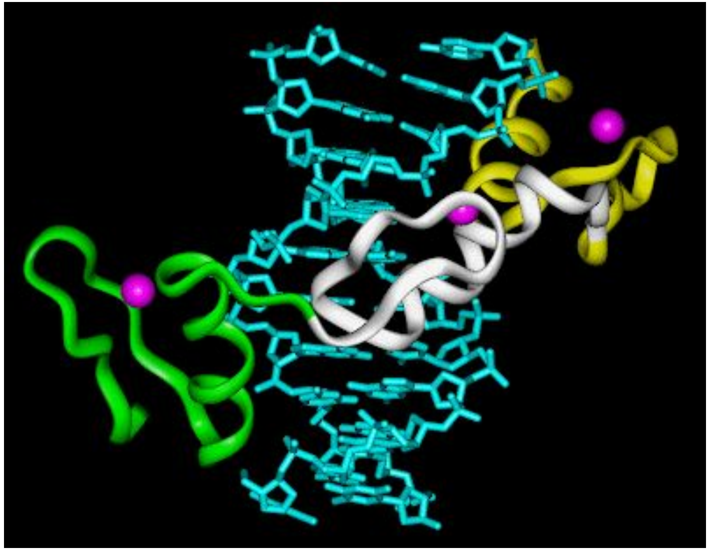
- NLP & IR
- Generative language models
- Diversity & Uncertainty

Natural Sciences

- Computational Molecular Biology
- Neuroscience

Example I: AI for Computational Molecular Biology

- 95% of diseases are not treatable, and drug development is getting ever harder and more expensive — developing a single drug today takes on average 10 years and costs 1.3B$.



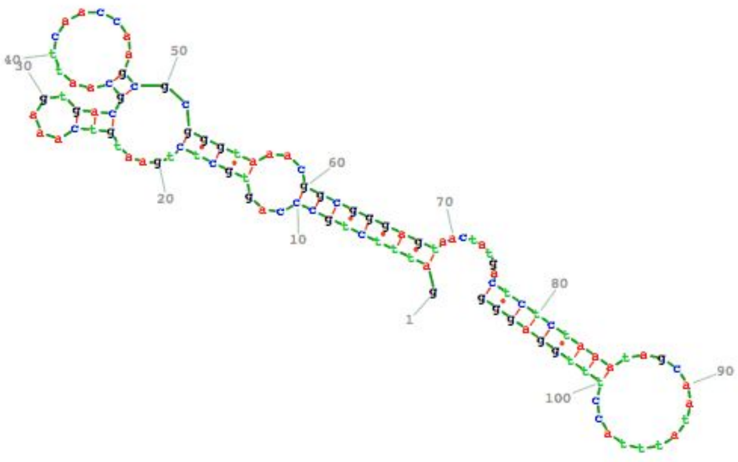
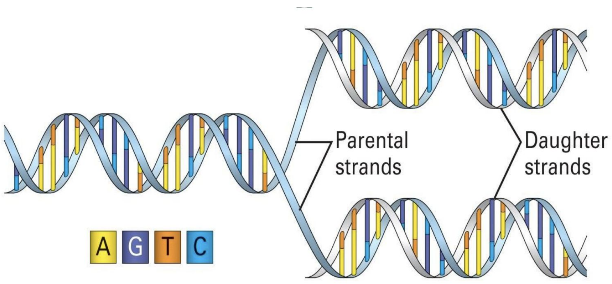


Image produced with RNAmovies [J. Waldispuehl]

All Life depends on 3 critical molecules

- DNAs (Deoxyribonucleic acid)
  - Hold information on how cell works
- RNAs (Ribonucleic acid)
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body’s major components (e.g. hair, skin, etc.)
  - “Workhorses” of the cell

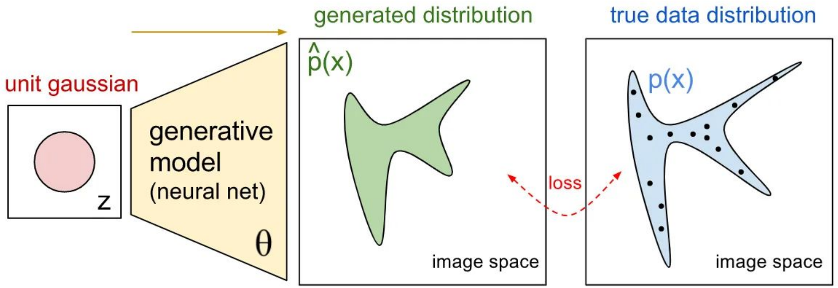
Representation of the life codes



- DNA
  - The structure and the four genomic letters code (Adenine, Guanine, Thymine, and Cytosine)
- RNA
  - It is usually only a single strand.
  - T(hyamine) is replaced by U(racil)
  - Several types of RNA exist for different functions in the cell.
- Proteins
  - are polypeptides (strings of amino acid residues)
  - Represented using strings of letters from an alphabet of 20: AEGLV…WKKLAG
  - Typical length 50…1000 residues

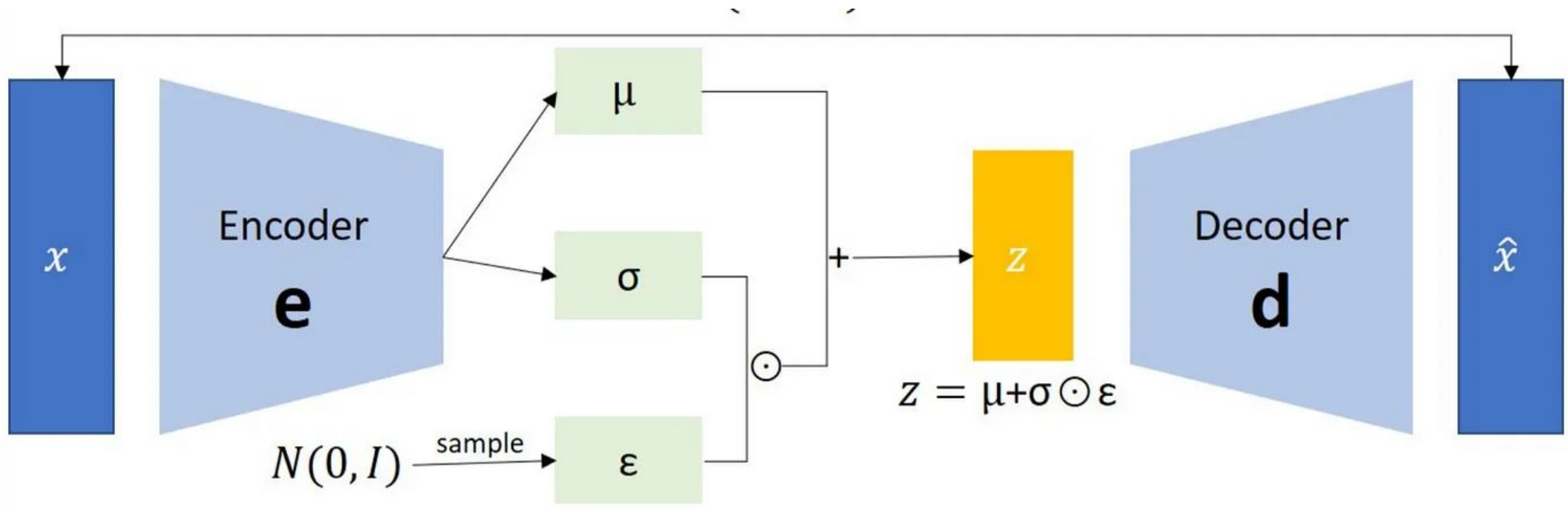
Generative models

- Goal:
  - To use these observations to build a p_{model} that can accurately mimic the observations produced by p_{data}.
  - maximize likelihood estimation / minimize some divergence

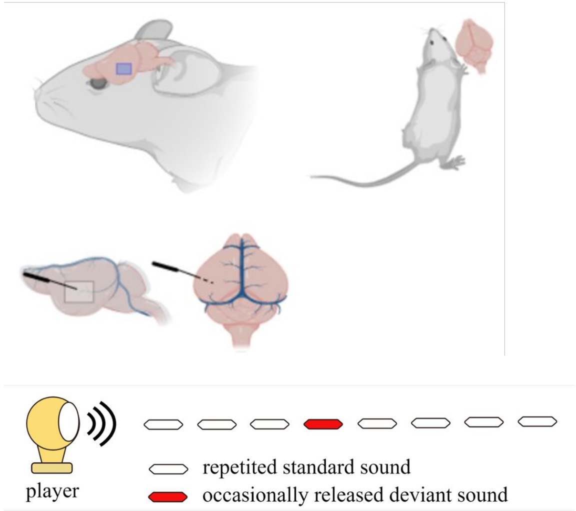


- Examples
  - VAE (Variational Auto-Encoder)
  - Generative Adversarial Network (GAN)
  - Normalizing Flow (NF)

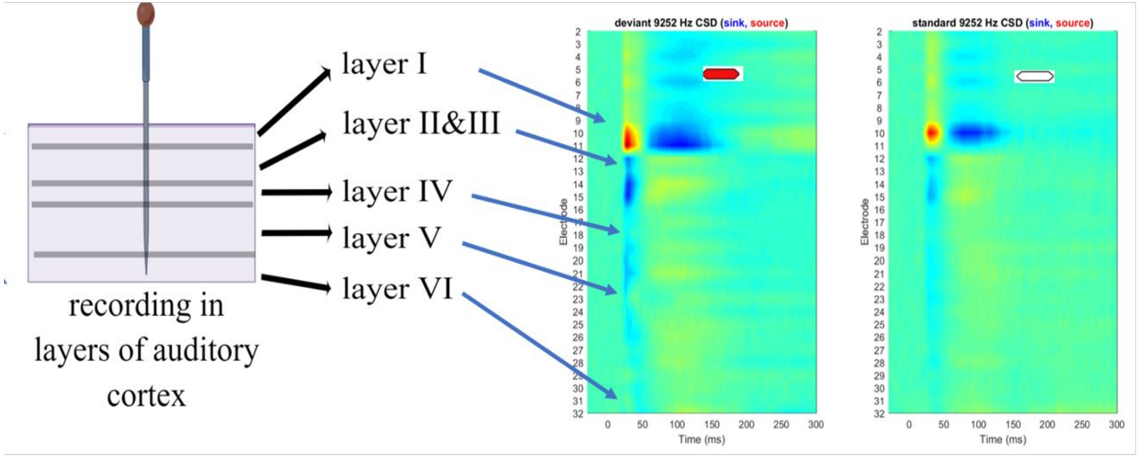
VAE (Variational Auto-Encoder)



Example II: AI for Neuroscience



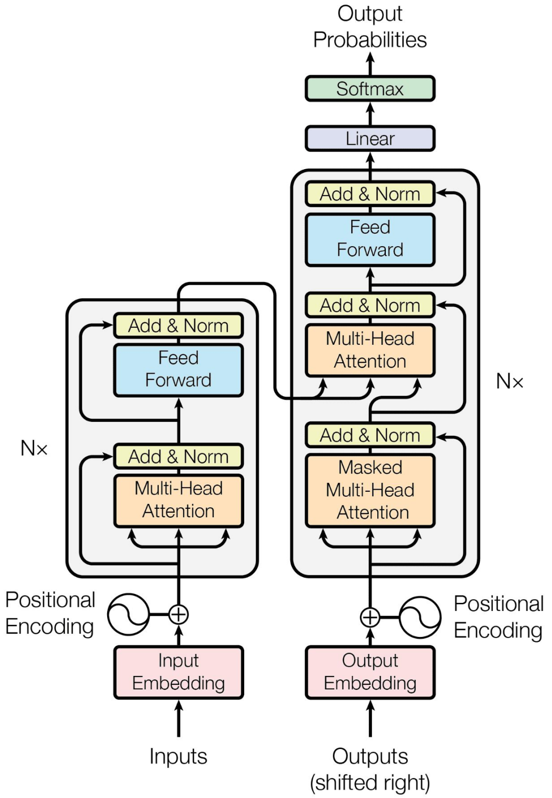
Current source density



Auditory Change Detection in Rats, Tiantian Yang, Piia Astikainen, Centre for Interdisciplinary Brain Research, Department of Psychology, University of Jyväskylä, Jyväskylä, Finland

Space-Time Modeling with Transformers

- Auditory Change Detection
  - Representation of membrane potential
    - Changes with time
    - Changes with layer (space)
  - Detection of the difference between patterns
    - Repetitive standard sound
    - Occasionally released deviant sound



Thank you for your attention!

Q & A