Introduction to Machine Learning:

Machine Learning for Natural Language Processing

- Zheng Yuan
- October 23, 2023
- The Australian National University

A bit of myself

- Assistant Professor at King’s College London
- Visiting Researcher at University of Cambridge
- Fellow at Trinity College, University of Cambridge

A bit of myself

- Assistant Professor at King’s College London
- Visiting Researcher at University of Cambridge
- Fellow at Trinity College, University of Cambridge
- *Currently visiting ANU School of Computing*

A bit of my work

- Machine Learning and Deep Learning for **Natural Language Processing (NLP)**
- Real-world applications in education, healthcare, creativity, social media and finance

What is Natural Language Processing?

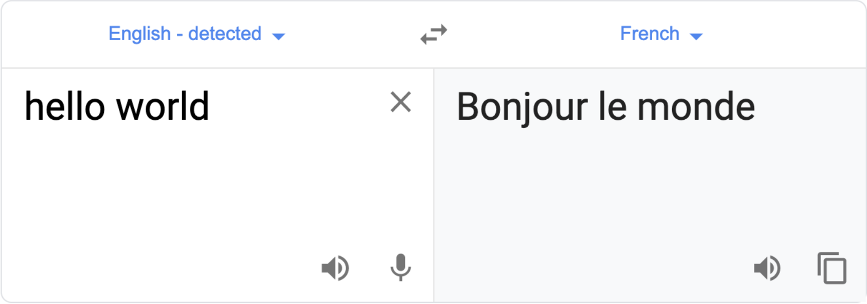
- NLP is the intersection of computer science, linguistics and machine learning
- The field focuses on communication between computers and humans in natural language
- NLP is all about making computers understand and generate human language

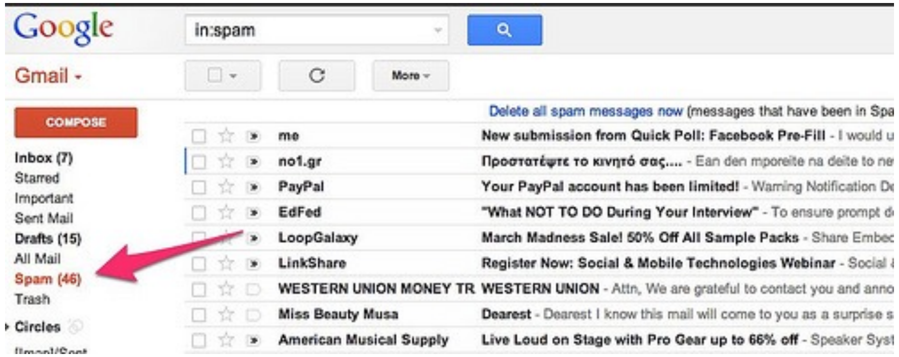
Natural language understanding (NLU)

Natural language generation (NLG)

ML & NLP applications

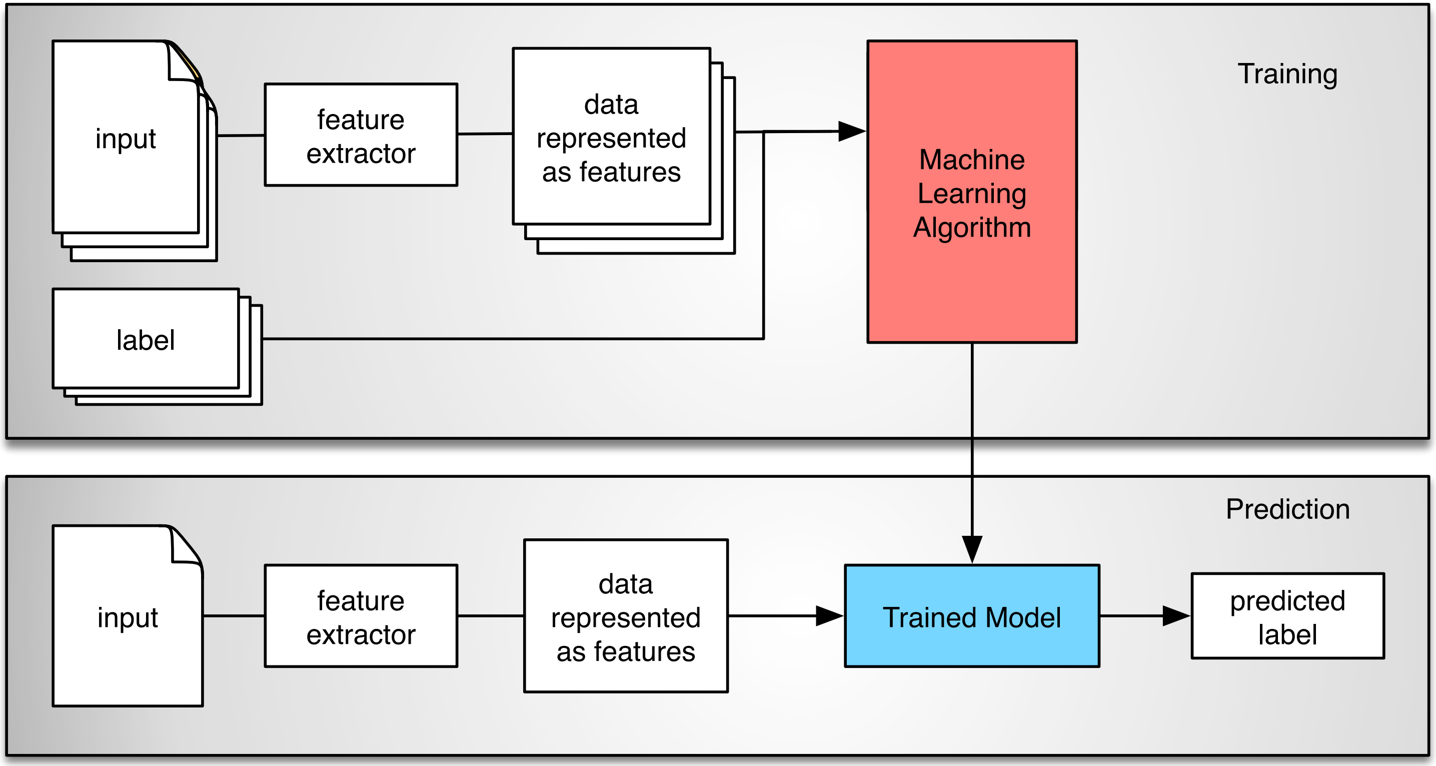
- Google translate - *machine translation*
- Spam filtering - *binary classification*
- Text prediction - *language modelling*
- Sentiment analysis
- Question answering
- And much more …





Supervised learning

Virtually all existing systems are learning-based and state-of-the-art systems are all supervised



Grammatical Error Correction

Task

- Input:

  Nowadays, there are many people that are learning foreign language. Is it worth to learn a foreign language? […] people who know how to speak a foreign language have more opportunities to get a job in important companies […] It could allow you to communicate with people, know different cultures …

Task

- Detection:

  Nowadays, there are many people that are learning foreign language. Is it worth to learn a foreign language? […] people who know how to speak a foreign language have more opportunities to get a job in important companies […] It could allow you to communicate with people, know different cultures …

Task

- Detection:

  Nowadays, there are many people that are learning foreign language. Is it worth to learn a foreign language? […] people who know how to speak a foreign language have more opportunities to get a job in important companies […] It could allow you to communicate with people, know different cultures …

- Correction:

  Nowadays, there are many people ~~that~~who are learning foreign ~~language~~languages. Is it worth ~~to learn~~learning a foreign language? […] people who know how to speak a foreign language have more opportunities to get a job in ~~important~~big companies […] It could allow you to communicate with people, get to know different cultures …

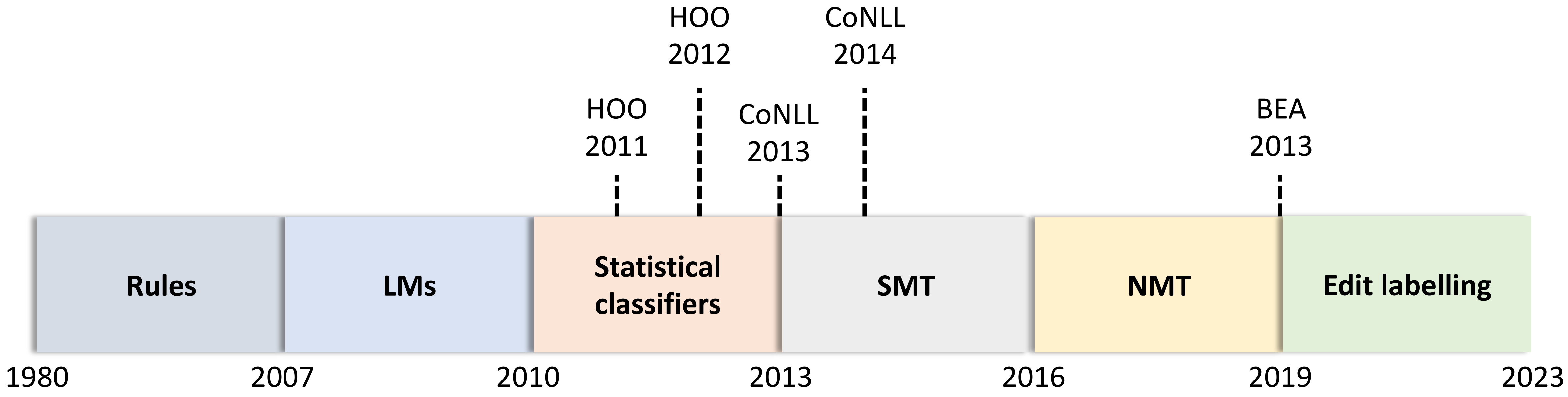
Motivations

- GEC is the task of automatically detecting and correcting errors in text
- For language learners
  - An estimated 1.5 billion people are learning English
  - Billions are also learning other languages
  - Teachers cannot possibly correct everything!
- For native speakers
  - Academic essays and publications
  - Business emails and marketing
  - Search bar queries (e.g. Google, Amazon, eBay)
- De-noise text for other NLP applications; e.g. translation, generation

Challenges

- Alternative corrections are possible
  - In conclude -> In conclusion OR To conclude
- Errors may interact
  - ε Book is good -> The Book is good -> The book is good
- Some error types are harder to correct than others
  - Function words vs. Content words
  - in -> at home vs. look at -> watch TV
- Error distributions differ significantly among users/domains

ML approaches to GEC



- Paradigm shift roughly every 3 years
- Next shift: Generative large language models (e.g. ChatGPT)?
- Shared tasks greatly contributed to progress

Language models

- In context, some words are more probable than others
  - They sell a big variety of products.
  - They sell a wide variety of products.
  - They sell a great variety of products.
  - They sell a large variety of products.
- Use this property to correct improbable sequences

Language models

- Example:
  - I often work in home.
- Approach
  - Train a language model from native, correct text
  - Define/generate a confusion set: {in, at, from, on, …}
  - Score each sentence to find which is best
    - I often work in home. 284.1275
    - I often work at home. 98.49942
    - I often work from home. 55.42596
    - I often work on home. 315.6587

Language models

- Advantages
  - Only require (lots of) native text; e.g. Wikipedia
  - Can detect all error types, including semantic errors
  - Effective in a low resource setting
  - Versatile
- Disadvantages
  - Probability is not grammaticality; e.g. I is the ninth letter of the alphabet.
  - Rare words; e.g. paraklausithyron
  - Generating confusion sets can be hard
    - E.g. I ate the big ___ .

Statistical classifiers

- Example: Predict the correct form of every verb
  - They were eat ice-cream when I arrive.

1. Define labels
2. Define features
3. Use machine learning to predict label from features

Statistical classifiers

- Example: Predict the correct form of every verb
  - They were eat ice-cream when I arrive.

1. Define labels
2. Define features
3. Use machine learning to predict label from features

Statistical classifiers

- Example: Predict the correct form of every verb
  - They were <u>eat</u> ice-cream when I <u>arrive</u>.

1. Define labels
   - Six different tags for main verbs
   - Multi-class classification

Tag	Meaning	Example 1	Example 2
VB	base form	eat	arrive
VBD	past tense	ate	arrived
VBG	gerund/present participle	eating	arriving
VBN	past participle	eaten	arrived
VBP	non-3rd person singular present	eat	arrive
VBZ	3rd person singular present	eats	arrives

Statistical classifiers

- Example: Predict the correct form of every verb
  - They were eat ice-cream when I arrive.

1. Define labels
2. Define features
3. Use machine learning to predict label from features

Statistical classifiers

- Example: Predict the correct form of every verb
  - They were eat ice-cream when I arrive.

1. Define labels
2. Define features
   - Any other features?

Features	Example 1	Example 2
PrecededByTo?	N	N
IsAuxiliary?	N	N
Lemma	eat	arrive
Ngrams (unigram, bigram, trigram)	“eat”, “were eat”, “eat ice-cream”, “They were eat”, “were eat ice-cream”, “eat ice-cream when”	“arrive”, “I arrive”, “arrive .”, “when I arrive”, “I arrive .”
MainVerb?	Y	N

Statistical classifiers

- Example: Predict the correct form of every verb
  - They were eat ice-cream when I arrive.

1. Define labels
2. Define features
3. Use machine learning to predict label from features
   I. Train a model on data
   II. Model learns how to weight feature importance
   III. Model outputs a label which indicates corrected form type

Statistical classifiers

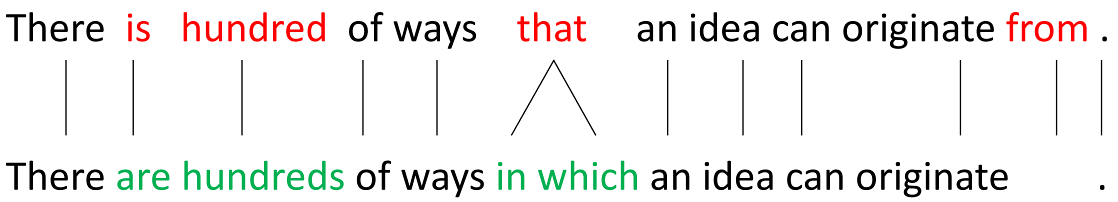
- Common classification techniques:
  - Naive Bayes
  - Logistic regression
  - Maximum entropy models
  - Support Vector Machines
  - ...
- Training data:
  - Native text (correct)
  - Non-native error-annotated data
  - Hybrid datasets

Statistical classifiers

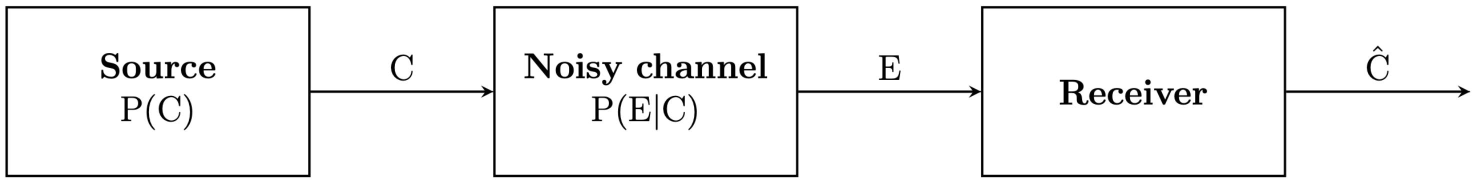
- Advantages
  - More flexible than rule-based systems
  - Only requires native data (but annotated data helps)
- Disadvantages
  - Feature engineering can be complicated
  - Works better for small confusion sets (e.g. function words)
  - Only targets single error types
  - Classifier order matters

Statistical machine translation

- GEC can be viewed as a translation from "incorrect" into "correct" English



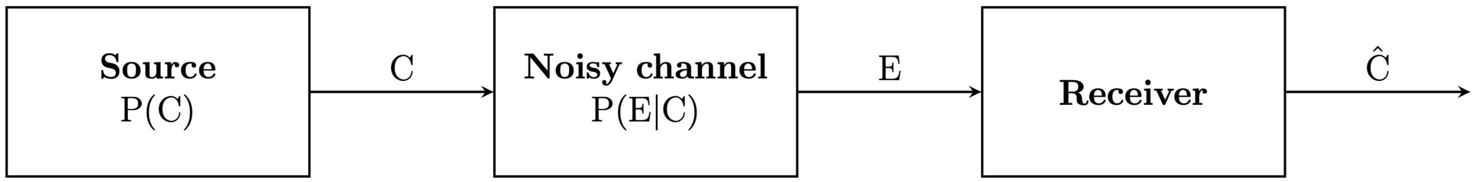
- SMT is inspired by the noisy channel model (Shannon, 1948)



- Requires a parallel corpus of original → corrected sentences

Statistical machine translation

1. Align sentences at the word level
2. Extract phrase mappings into a phrase table
3. Generate translations using the phrase table and a language model (i.e. decoding)



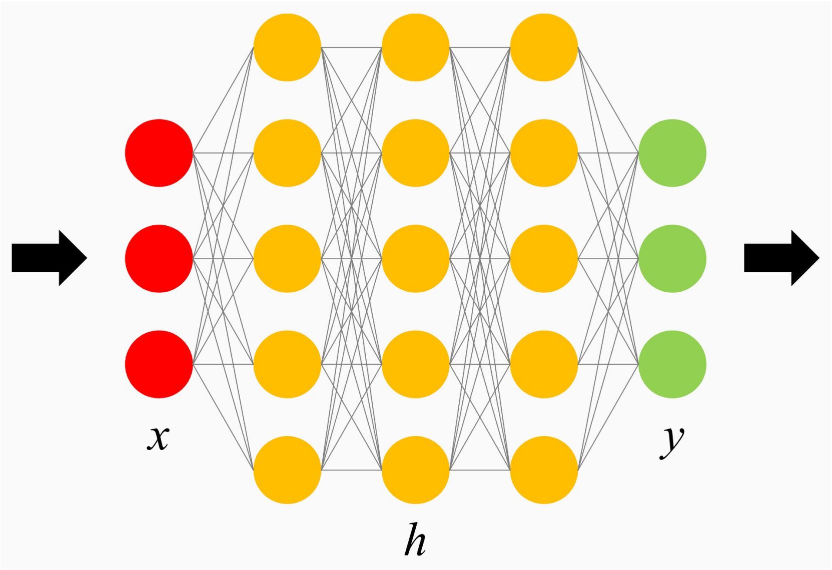
$$\hat{C} = \arg\max_C P(C|E) = \arg\max_C \frac{P(E|C)P(C)}{P(E)} = \arg\max_C P(E|C)P(C)$$

Neural machine translation

- Same concept as SMT but with neural networks

Deep neural networks

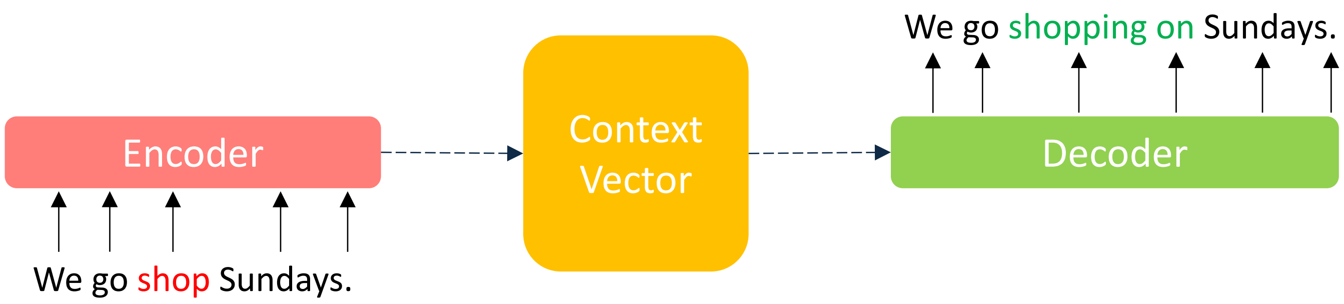
- General models that map an input *x* to an output *y* via a number of hidden states *h*



- Different architectures and applications under the deep learning umbrella

Neural machine translation

- Same concept as SMT but with neural networks
- Sequence-to-sequence model based on the *encoder-decoder* framework



Machine translation

- Advantages
  - Corrects all error types simultaneously
  - Handles interacting errors
  - Does not require feature engineering or expert knowledge
  - Single end-to-end model
  - State of the art (Transformer NMT)
- Disadvantages
  - Requires (lots of) parallel training data
  - Can take a long time/lots of resources to train
  - Uninterpretable
  - Hard to customise

Neural edit labelling

- Predict edit label for every word

They	likes	to	eat	the	ice-cream	.
KEEP	REPLACE	KEEP	KEEP	DELETE	KEEP	KEEP

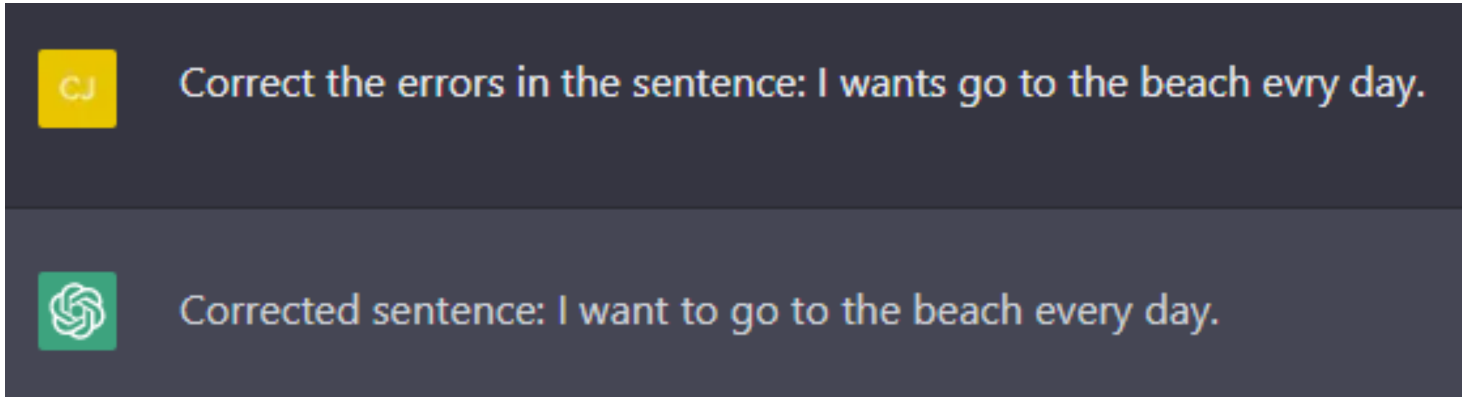
- Essentially a classifier for every word
- Same principle as sequence labelling
  - Requires labelled data
  - Fine-tune various pretrained neural language models
  - Choice of labels is an open question
    - E.g. binary (correct/incorrect) vs. detailed labels (>5,000?)

Neural edit labelling

- Advantages
  - Handles all error types (depending on label set)
  - Single end-to-end model
  - More efficient than neural translation
  - Somewhat interpretable
  - State of the art
- Disadvantages
  - Requires (lots of) parallel training data
  - May not handle multi-token or interacting errors well
  - Requires engineering the size/scope of the label set

Language generation

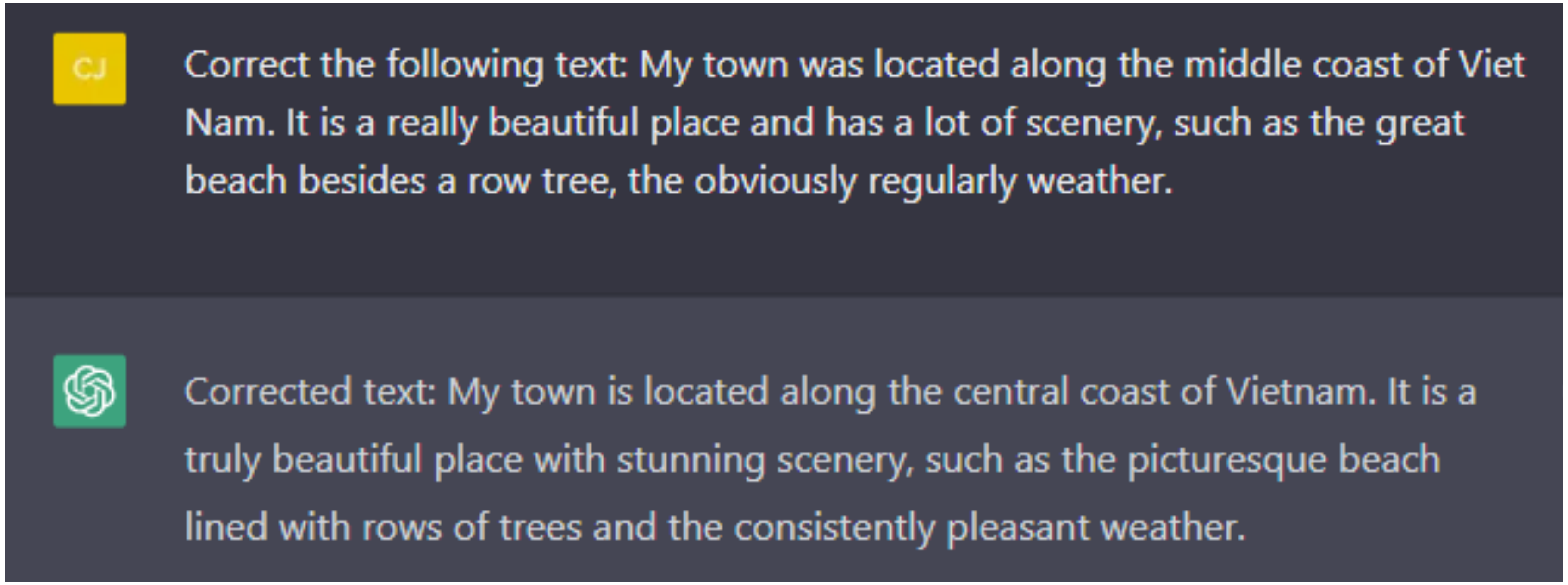
- ChatGPT



- Train a large language model (LLM) on a very large amount of data and fine-tune on hundreds of language generation tasks
- Many models are available
  - Bloom, Cohere, Google T5, Meta OPT, GPT*

Language generation

- Advantages
  - Versatile
  - Impressively fluent output



Language generation

- Disadvantages
  - Not all “corrections” are errors
    - really beautiful -> truly beautiful
    - has a lot of scenery -> stunning scenery
  - Inconsistent output
    - The same input can give different output

Language generation

- Disadvantages
  - Prompting matters

Input	Output
Correct the text: I wants go to the beach evry day.	I wants to go to the beach every day.
Correct **the errors in** the text: I wants go to the beach evry day.	I wants go to the beach every day.
Fix the errors in the text: I wants go to the beach evry day.	I want to go to the beach every dayθ

Future challenges

- System combination
  - Which approaches have complimentary strengths?
- Training data selection
  - Optimise the most discriminative training data
- Unsupervised approaches
  - Human-annotated corpora are expensive to create
- Domain generalisation
  - Language learning vs. business vs. documentation vs. poetry
- Improved evaluation
  - Users want n-best edits

Future challenges

- Feedback Comment Generation
  - Explainable GEC
- Multilingual GEC
  - Develop systems for other languages
- Contextual GEC
  - Move beyond the sentence-level
- Semantic errors
  - Systems weak on idioms, multi-word expressions, collocations
- Personalised systems
  - Adapt to user first language and ability level

Questions / Comments?

https://www.cl.cam.ac.uk/~zy249/

zheng.yuan@kcl.ac.uk