

Natural language processing for personalised language learning



Cambridge ALTA
Institute for Automated Language Teaching and Assessment

❖ Zheng Yuan ❖ July 1, 2022 ❖ RIKEN AIP

A bit of myself



Cambridge, UK

Mphil & PhD @ Cambridge



London, UK

BSc Eng @ QMUL



Beijing, China



Xianning, China

“City of Osmanthus”



Guangzhou, China

A bit of myself



Research Associate

Department of Computer Science and
Technology, University of Cambridge



Postdoctoral Teaching Fellow

Trinity College, University of Cambridge



Praeceptor

Corpus Christi College, University of
Cambridge



A bit of myself



Assistant Professor

Department of Informatics, King's College
London



Visiting Researcher
University of Cambridge

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

A bit of myself



Volunteer teaching

China & Kenya





Systemic inequalities do exist!

‘One-size-fits-all’ learning, teaching and assessment

However, not every learner has the same interest and aptitude to learn



BBC Sign in Home News Sport Weather iPlayer Sounds

NEWS

Home Coronavirus Brexit UK World Business Politics Tech Science Health Family & Education

England Local News Regions Leicester

Lack of diverse curriculum 'hampers BAME students', university study finds

Personalised education

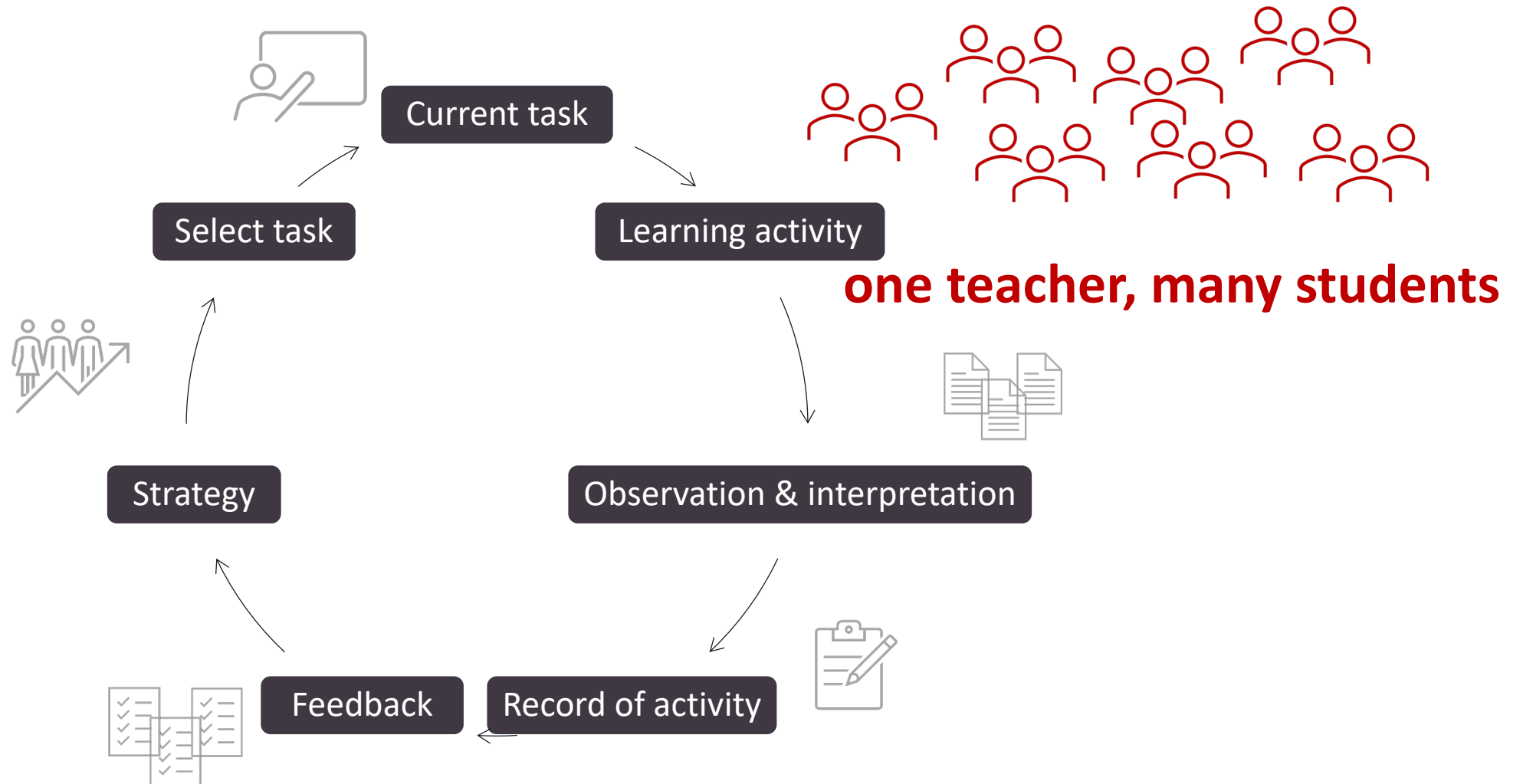
Address the distinct learning needs, abilities, or cultural backgrounds of individuals

Enhance the experience of learning and teaching

Provide equal opportunities for learners and teachers worldwide

Personalised education: Challenges?

Classroom learning and teaching



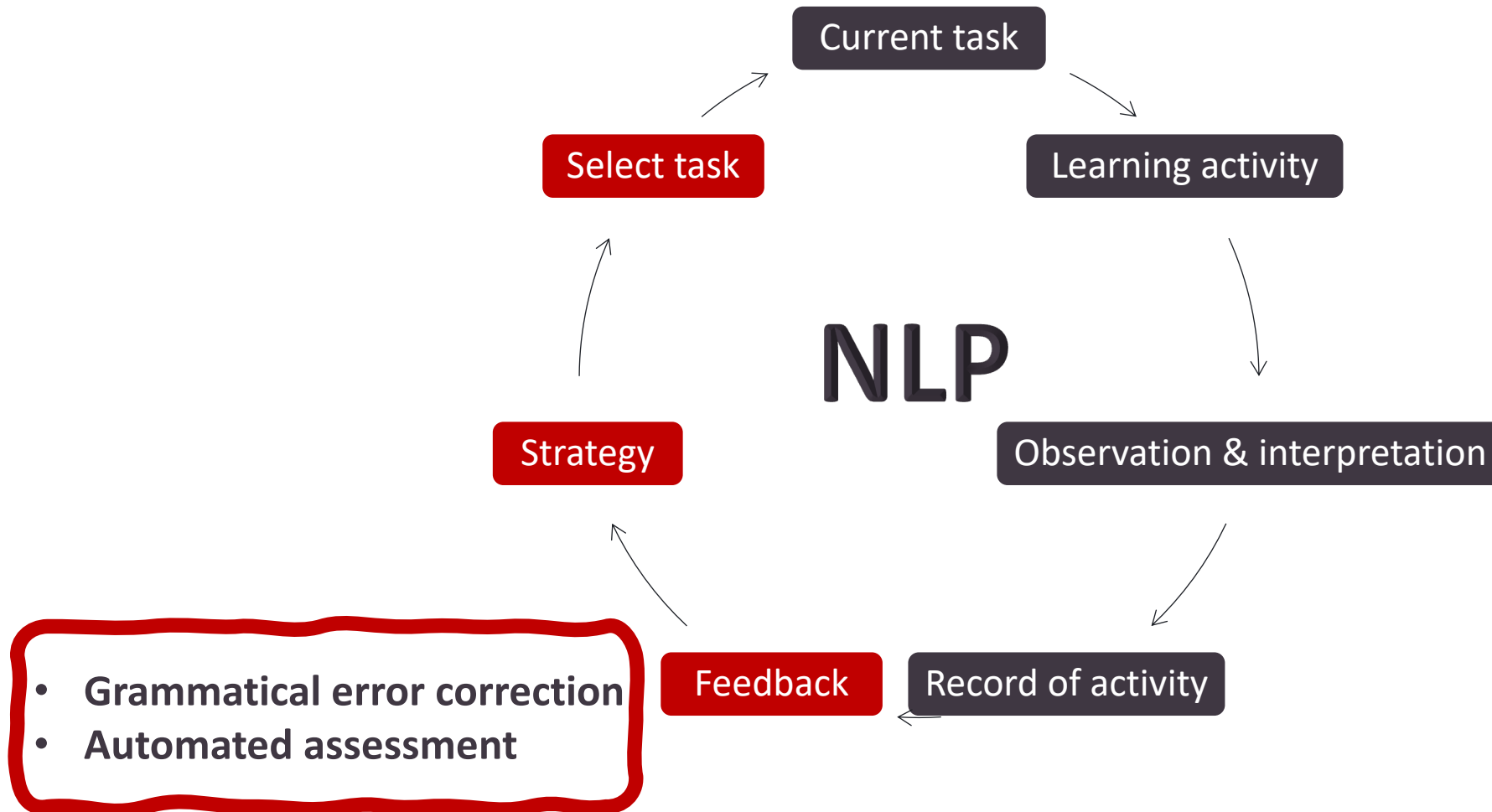
Personalised education:

Can natural language processing (NLP) help?

NLP for personalised language learning

- Language education
 - A great need for people who can communicate in multiple languages
 - Millions of people are learning English as a second language
- Computer-assisted language learning (CALL)

NLP for personalised language learning



Feedback

Nowadays, there are many people ^{who} that are learning foreign ^{languages} language.
Is it worth to learn a foreign language?
^{learning}

Firstly, there are more multinational companies that need people to speak other languages, so that means that people who know how to speak a foreign language have more opportunities to get a job in important companies, or will have more chances of being promoted.

[...]

Finally, learning another language give the learner the ability to step inside the mind and context of that other culture. It could allow you to communicate with people, know different cultures ...

Learning foreign languages will bring you many benefits. However, if you have the possibility of learning a new language, do it...

Grammatical error correction

- GEC is the task of automatically detecting and correcting errors in learner text
- A difficult theoretical problem
- An important practical application with high educational and commercial value

Input:

Nowadays, there are many people that are learning foreign language.

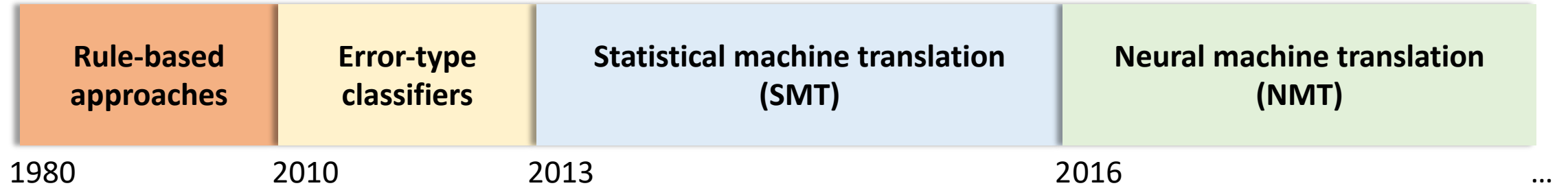
Detection:

*Nowadays, there are many people **that** are learning foreign **language**.*

Correction:

*Nowadays, there are many people ~~that~~**who** are learning foreign **languages**.*

Grammatical error correction



- Hand-coded rules

E.g. **informations** -> **information**

in the other hand -> **on** the other hand

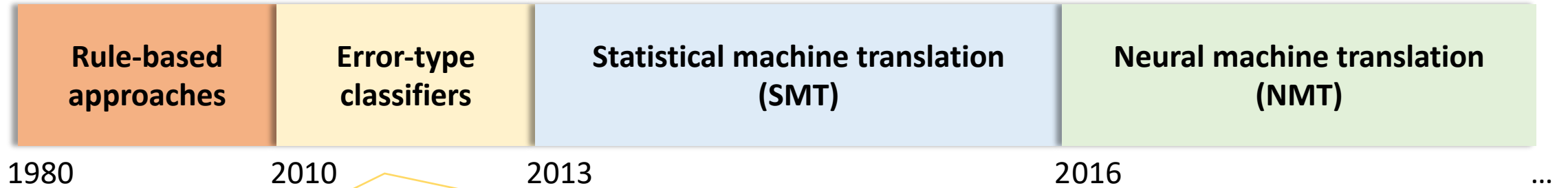


Can be very effective



Hard to scale and maintain

Grammatical error correction



- Data-driven
- A classification task (e.g. for articles and prepositions)

E.g. article classifier

- For each noun phrase
 - Features: previous 3-gram, next 3-gram, head noun, ...
 - Classes: no article, definite article, indefinite article

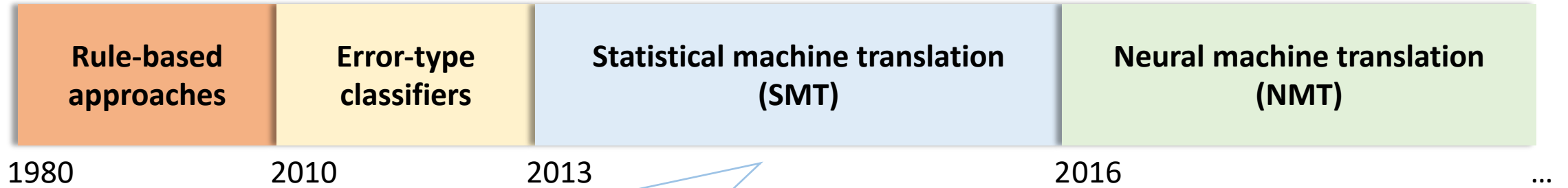


More flexible than rules



Target single error types

Grammatical error correction



- “Monolingual” machine translation
- Translate from “incorrect” into “correct” English

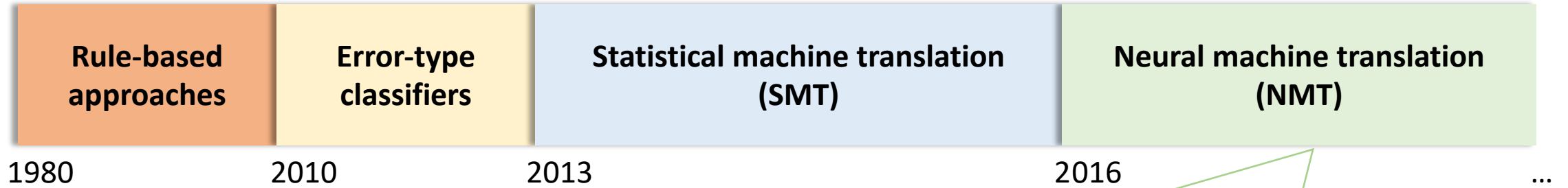
*There **is** **hundred** of ways **that** an idea can originate **from** .*

*There **are** **hundreds** of ways **in which** an idea can originate .*

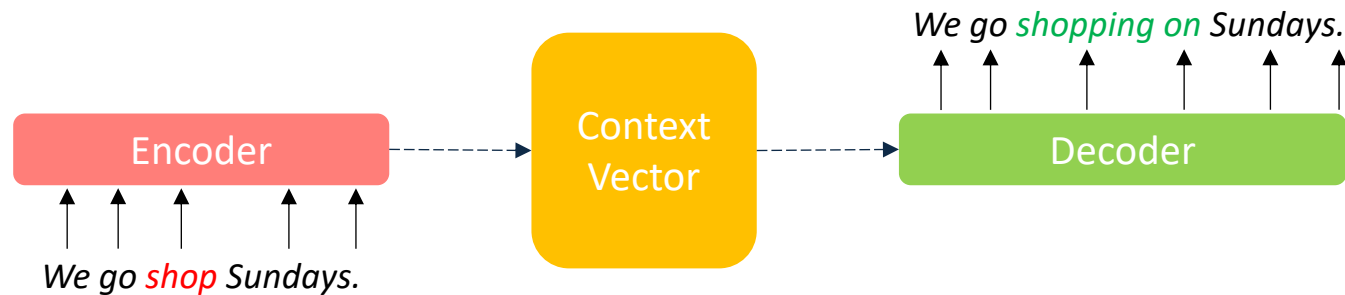
Grammatical error correction: SMT

- Adapting SMT for GEC:
 - Generating artificial errors (Yuan and Felice, 2013; Felice and Yuan, 2014; Rei et al., 2017; Yuan et al., 2019)
 - Adding GEC-specific features: character-level Levenshtein distance (Felice et al., 2014; Yuan et al., 2016)
 - Introducing additional large-scale language models on monolingual native data (Yuan 2017)
 - N-best list re-ranking (Felice et al., 2014; Yuan et al., 2016; Yannakoudakis et al., 2017; Yuan et al., 2019; Yuan et al., 2021)
- SMT-based GEC became a dominant approach in the field
 - Our system won an international competition in 2014 (Felice et al., 2014)
 - Many researchers have subsequently built on this work

Grammatical error correction



- Same concept as SMT but with neural networks
- Encoder-decoder



Grammatical error correction: NMT

- We presented the first work using NMT for GEC (Yuan and Briscoe, 2016)
- Our NMT-based system ranked 1st amongst academic entries in an international competition (Yuan et al., 2019)
- State-of-the-art in multiple languages
 - E.g. English, Chinese, Spanish, Arabic, Russian
- Companies are using this technology to build their commercial products

Feedback

*Nowadays, there are many people that are learning foreign language.
Is it worth to learn a foreign language?*

Firstly, there are more multinational companies that need people to speak other languages, so that means that people who know how to speak a foreign language have more opportunities to get a job in important companies, or will have more chances of being promoted.

[...]

Finally, learning another language give the learner the ability to step inside the mind and context of that other culture. It could allow you to communicate with people, know different cultures ...

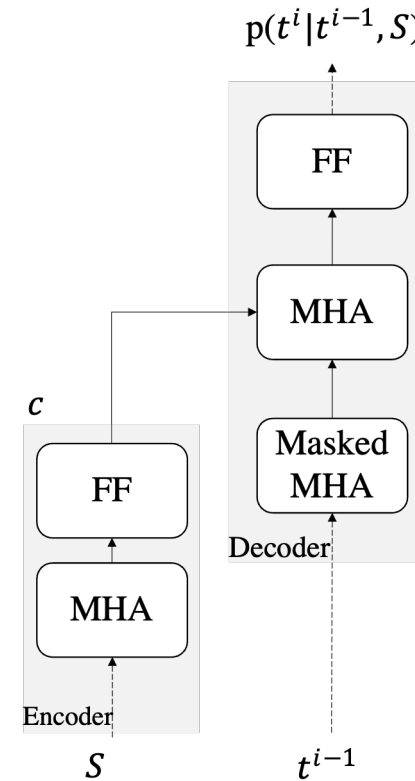
*Learning foreign languages will bring you many benefits. **Therefore** However, if you have the possibility of learning a new language, do it...*

Contextual GEC

- GEC systems normally operate at the sentence level, but context matters:
 - *The man went to the café. He ~~order~~ordered a coffee.*
 - *I see the cows in the field. ~~It is~~They are sleeping.*
- We propose contextual GEC (Yuan and Bryant, 2021)

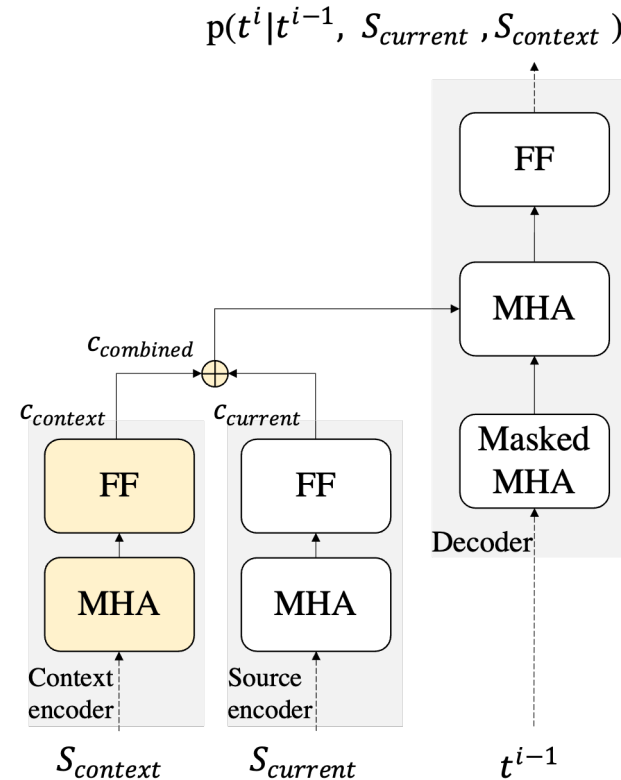
Contextual GEC

- Transformer NMT-based
 - Single-encoder
 - $\langle \text{context} \rangle [\text{SEP}] \langle \text{sent} \rangle$



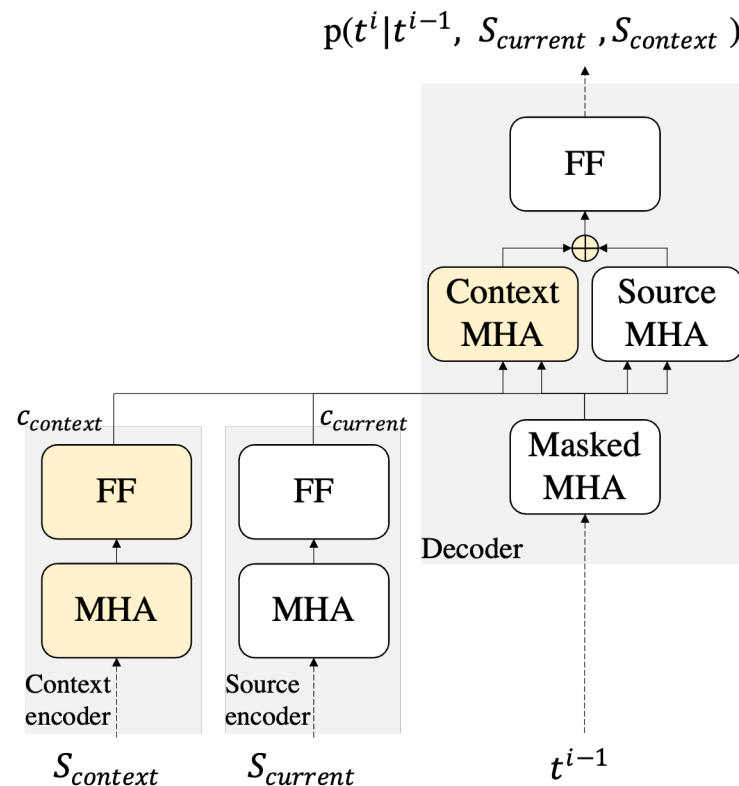
Contextual GEC

- Transformer NMT-based
 - Single-encoder
 - $\langle \text{context} \rangle [\text{SEP}] \langle \text{sent} \rangle$
 - Multi-encoder (encoder side)
 - Combine context and sent representation
 - Multi-encoder (decoder side)
 - Combine context and sent attention output



Contextual GEC

- Transformer NMT-based
 - Single-encoder
 - $\langle \text{context} \rangle [\text{SEP}] \langle \text{sent} \rangle$
 - Multi-encoder (encoder side)
 - Combine context and sent representation



Contextual GEC

| | BEA | FCE | CoNLL-2014 |
|--------------|-------------------|-------------------|-------------------|
| Baseline | 52.9 | 57.8 | 48.3 |
| SingleEnc | 53.5 | 57.4 | 48.3 |
| MultiEnc-enc | 56.5 | 59.2 | 50.5 |
| MultiEnc-dec | 56.6 (3.7) | 59.6 (1.8) | 51.6 (3.3) |

F_{0.5} (evaluated by ERRANT)

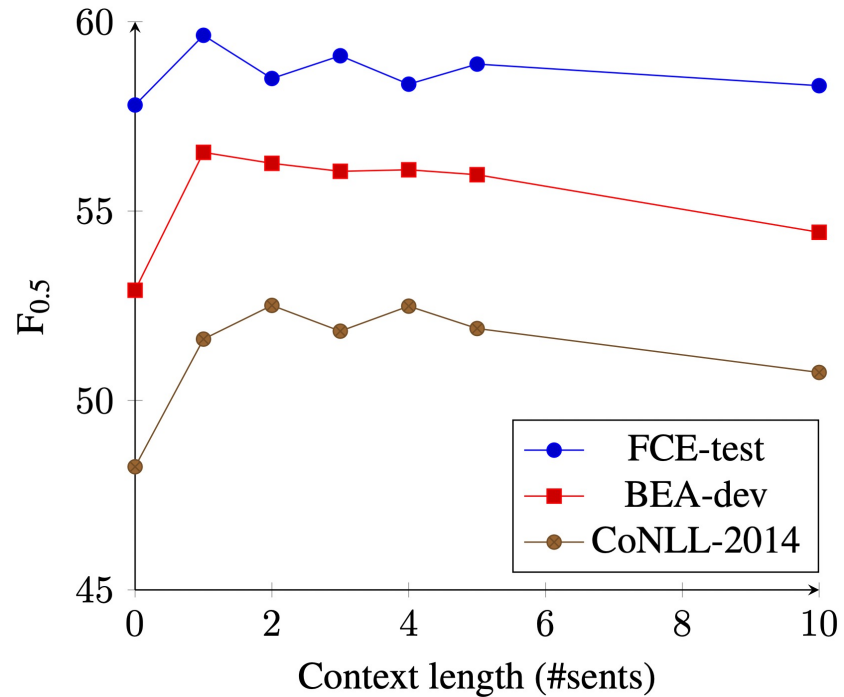
BEA: full range of first languages and ability levels

FCE: upper-intermediate learner English

CoNLL: advanced undergraduate university students from an Asian language background

- SingleEnc context concatenation is not very effective
- MultiEnc-dec is the most effective
- Context helps some datasets more than others

Contextual GEC



- 1 sent context is usually enough
- Errors do not tend to have long dependencies

Contextual GEC

- Context can help most error types
- The biggest gains for subject-verb agreement, noun number, pronoun, and verb tense errors
- Examples:

| | |
|----------------|--------------------------------------|
| <i>Context</i> | <i>Then we <u>went</u> to Taxco.</i> |
|----------------|--------------------------------------|

| | |
|---------------|--|
| <i>Source</i> | <i>We stay in a very luxurious hotel.</i> |
|---------------|--|

| | |
|-----------------|--|
| <i>Baseline</i> | <i>We stay in a very luxurious hotel.</i> |
|-----------------|--|

| | |
|-----------------------|--|
| <i>Contextual GEC</i> | <i>We stayed in a very luxurious hotel.</i> |
|-----------------------|--|

| | |
|----------------|--|
| <i>Context</i> | <i>My favourite sport is volleyball. When I am on the beach I like playing with my sister in the sand and then we go in the sea.</i> |
|----------------|--|

| | |
|---------------|---------------------------------|
| <i>Source</i> | <i>It is very funny.</i> |
|---------------|---------------------------------|

| | |
|-----------------|---------------------------------|
| <i>Baseline</i> | <i>It is very funny.</i> |
|-----------------|---------------------------------|

| | |
|-----------------------|--------------------------------|
| <i>Contextual GEC</i> | <i>It is great fun.</i> |
|-----------------------|--------------------------------|

Feedback

*Nowadays, there are many people that are learning foreign language.
Is it worth to learn a foreign language?*

Firstly, there are more multinational companies that need people to speak other languages, so that means that people who know how to speak a foreign language have more opportunities to get a job in important companies, or will have more chances of being promoted.

[...]

*Finally, learning another language give the learner the ability to step inside the mind and context of that other culture. It could allow you to communicate with people, know different cultures ...
get to know*

Learning foreign languages will bring you many benefits. However, if you have the possibility of learning a new language, do it...

Semantic errors

- Multiword expressions (MWEs) are challenging for language learners
 - *After some years, she ~~get~~~~loved~~fell in love with him.*
 - *I'm writing to ~~inform~~give you some advice on...*
- Current GEC systems fail to take them into consideration
- We propose incorporating MWE information to GEC (Taslimipoor et al., 2022)

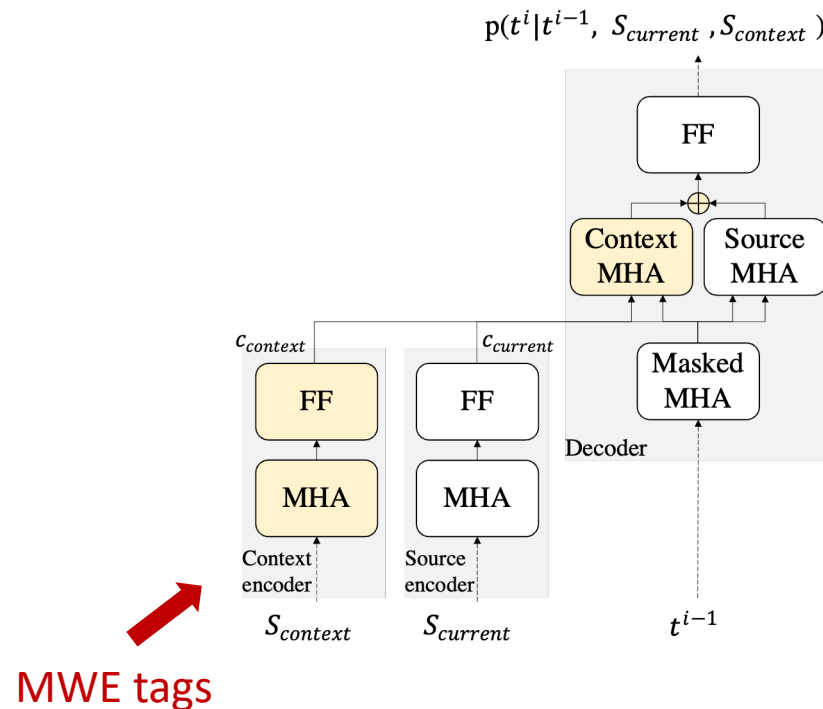
Semantic errors

- Incorporating MWE identification into two GEC systems:
 - MWE identification
 - MTLB-STRUCT (Taslimipoor et al., 2020)
 - winning system at the 2020 shared task
 - using ELECTRA pre-trained model for sequence labelling
 - STRUSLE dataset

| | MWE | | | MWE (verbal) | | |
|-------------------|------|------|----------------|--------------|------|----------------|
| | P | R | F ₁ | P | R | F ₁ |
| Liu et al. (2021) | 82.0 | 64.3 | 72.0 | - | - | 63.9 |
| Our system | 90.7 | 66.8 | 76.7 | 65.2 | 68.2 | 66.7 |

Semantic errors

- Incorporating MWE identification into two GEC systems:
 - Multi-encoder (decoder side) (Yuan et al., 2021)



Semantic errors

- Incorporating MWE identification into two GEC systems:
 - BART-based GEC (Katsumata and Komachi, 2020)
 - Using special MWE tokens to mark the spans of MWEs
 - Fine-tuning on MWE-annotated GEC data

Source:

... they also [MWE] made talks [/MWE] about the earth's problems ...

Target:

... they also [MWE] gave talks [/MWE] about the earth's problems ...

Semantic errors

| | Baseline | MWE-GEC |
|-------------------|----------|---------|
| Multi-encoder GEC | 49.5 | 51.1 |
| BART-based GEC | 51.1 | 51.5 |

F0.5 (evaluated by ERRANT)

BEA: full range of first languages and ability levels

- Adding MWE information improves both GEC system performance

Semantic errors

| | MWE type | Baseline | MWE-GEC |
|-------------------|----------|----------|-------------|
| Multi-encoder GEC | VID | 42.0 | 45.1 |
| | VPC.full | 32.5 | 43.5 |
| BART-based GEC | VID | 42.0 | 53.4 |
| | VPC.full | 29.7 | 41.3 |

VID: verbal idiomatic expressions, e.g. go bananas

VPC.full: fully non-compositional verb particle constructions, e.g. give up

- The highest improvement is in the case of VPC.full and VID
- Future work benefits from more detailed annotation of learner errors related to their understanding of MWEs

GED for GEC

- Grammatical error detection (GED)
 - Binary
 - Predict whether a word is correct or incorrect in context
 - Multi-class
 - Predict whether a word is correct or incorrect as well as **its error type** in context

| Original | | Yesterday | I | spend | quite | long | time | for | shopping | . |
|------------|----------|-----------|---|--------------|-------|--------|------|--------|----------|---|
| Detection | Binary | C | C | I | C | I | C | I | C | C |
| | 4-class | C | C | R | C | M | C | U | C | C |
| | 25-class | C | C | VERB:TENSE | C | DET | C | PREP | C | C |
| | 55-class | C | C | R:VERB:TENSE | C | M:DET | C | U:PREP | C | C |
| Correction | | Yesterday | I | spent | quite | a long | time | | shopping | . |

Derived from ERRANT

GED for GEC

- Multi-class GED
 - Treat GED as a sequence labelling (or token classification) task
 - Fine-tune pre-trained language representation models
 - BERT
 - XLNet
 - ELECTRA

GED for GEC

- Token-level $F_{0.5}$ for incorrect (I) labels

| System | BEA | FCE | CoNLL-2014-a | CoNLL-2014-b |
|----------------------------------|-------------|-------------|--------------|--------------|
| GED (BERT) | 59.2 | 67.9 | 45.6 | 53.8 |
| GED (XLNet) | 63.2 | 69.8 | 48.6 | 58.7 |
| GED (ELECTRA) | 65.5 | 72.9 | 51.2 | 64.7 |
| <i>Bell et al. (2019)</i> | 48.5 | 57.3 | 36.9 | 46.3 |
| <i>Kaneko and Komachi (2019)</i> | - | 61.7 | 35.3 | 43.7 |

- New state of the art
- ELECTRA's discriminative objective makes it more closely-related:
 - A discriminator (rather than a generator)
 - Aims to detect replaced tokens

GED for GEC

- GED (ELECTRA) binarised and macro-averaged $F_{0.5}$

| Mode | BEA-dev | | FCE-test | |
|----------|-------------|-------|-------------|-------|
| | binarised | macro | binarised | macro |
| Binary | 65.5 | 80.4 | 72.9 | 83.5 |
| 4-class | 66.1 | 67.1 | 72.6 | 71.0 |
| 25-class | 63.1 | 47.3 | 72.1 | 54.6 |
| 55-class | 65.8 | 33.0 | 73.9 | 34.9 |

- Lower macro-averaged scores for multi-class GED
- Adding more error types does not significantly affect the performance of binarised detection

binarised $F_{0.5}$: the score for detecting any non-C labels regardless of class

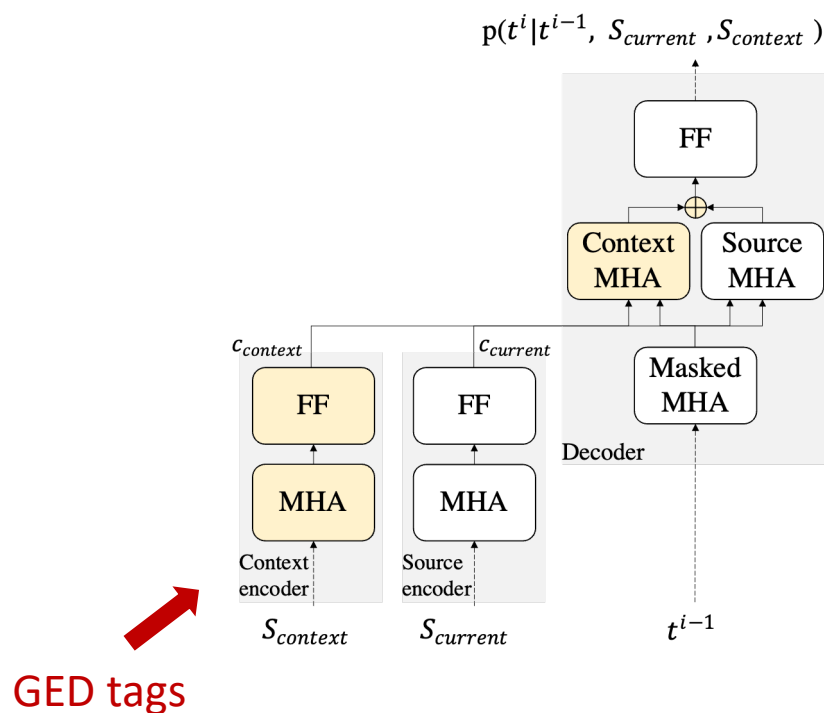
macro-averaged $F_{0.5}$: the average $F_{0.5}$ across all classes

GED for GEC

- Two ways of using GED to improve GEC (Yuan et al., 2021)

1) As auxiliary input

2) For re-ranking



Original:

Yesterday I *spend* quite *long* time *for* shopping.

N-best system output:

1st Yesterday I *spent* quite *long* time shopping.

...

3rd Yesterday I *spent* quite *a* long time *for* shopping.

...

6th Yesterday I *spent* quite *a* long time shopping.

GED for GEC

- Using GED as auxiliary input

| | | |
|----------|----------|--------------|
| Baseline | | 49.80 |
| Binary | + GED | 52.20 |
| | + Oracle | 63.68 |
| 4-class | + GED | 52.84 |
| | + Oracle | 67.86 |
| 25-class | + GED | 51.78 |
| | + Oracle | 68.36 |
| 55-class | + GED | 51.52 |
| | + Oracle | 70.24 |

- Adding GED yields a consistent statistically significant improvement
- Our best system uses the 4-class GED
- The system benefits the most when the finest and most granular level of error type information is provided
- We expect further performance gains with better multi-class GED

F0.5 (evaluated by ERRANT)

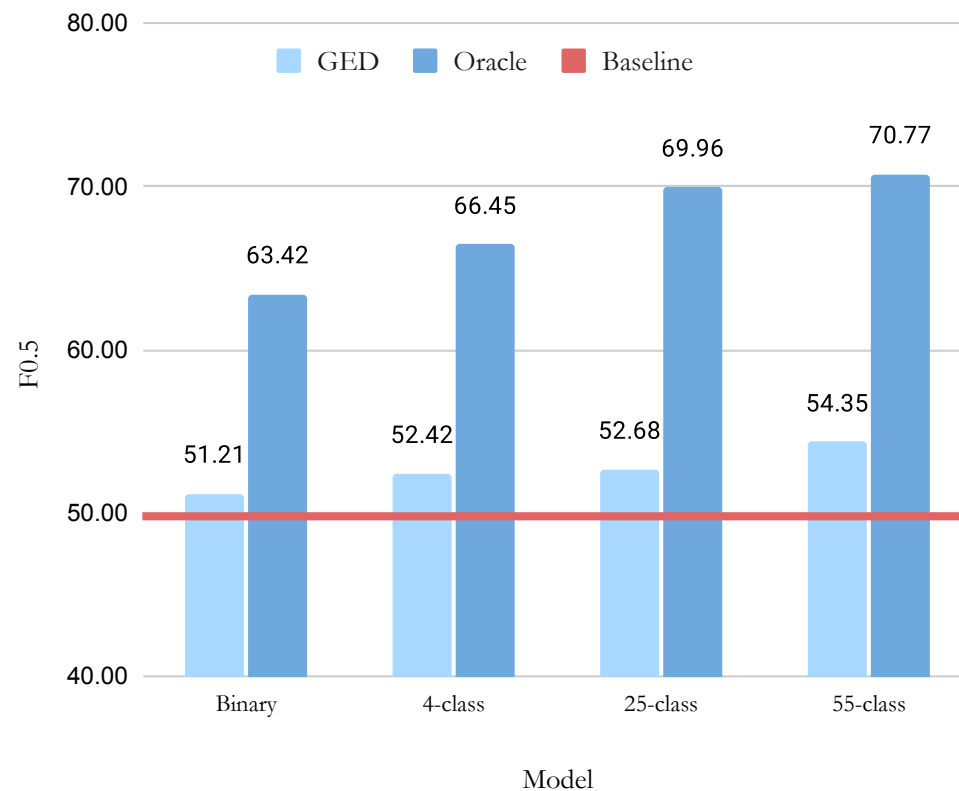
BEA: full range of first languages and ability levels

GED for GEC

- Using GED for N-best list re-ranking
 1. generate an N-best list of candidate hypotheses
 2. align each hypothesis with the source sentence using ERRANT
 3. convert the edit spans to token-based detection labels
- Using the minimum Hamming distance to ensure the maximal overlap between the GEC hypothesis and the GED predictions
- No need to use any other features

GED for GEC

- Using GED for N-best list re-ranking
 - Using GED for re-ranking GEC output improves the results consistently and significantly
 - The best re-ranking model uses 55-class GED
 - The performance boost using GED predictions is still far from what the model can achieve using oracle



GED for GEC

- Constrained setting

| | BEA | CoNLL |
|-------------------------------|-------------|-------------|
| Baseline | 47.4 | 43.2 |
| + GED input | 51.8 | 47.0 |
| + GED re-ranking | 58.5 | 54.4 |
| Raheja and Alikaniotis (2020) | 49.1 | 47.1 |
| Kaneko et al. (2020) | 54.8 | 53.6 |

BEA: full range of first languages and ability levels

CoNLL: advanced undergraduate university students from an Asian language background

Feedback

B

Nowadays, there are many people ^{who} that are learning foreign ^{languages} language.

Is it worth to learn a foreign language?
^{learning}

Firstly, there are more multinational companies that need people to speak other languages, so that means that people who know how to speak a foreign language have more opportunities to get a job in important companies, or will have more chances of being promoted.

^{big}
[...]

Finally, learning another language ^{gives} give the learner the ability to step inside the mind and context of that other culture. It could allow you to communicate with people, know different cultures ...

^{get to know} ^{Therefore}
Learning foreign languages will bring you many benefits. However, if you have the possibility of learning a new language, do it...




Automated assessment

- Mimic the judgment of teachers evaluating the quality of learner writing
- Used in the classroom as well as by well-established test providers
- Extremely valuable to both learners and teachers








Automated assessment

Feature-based systems

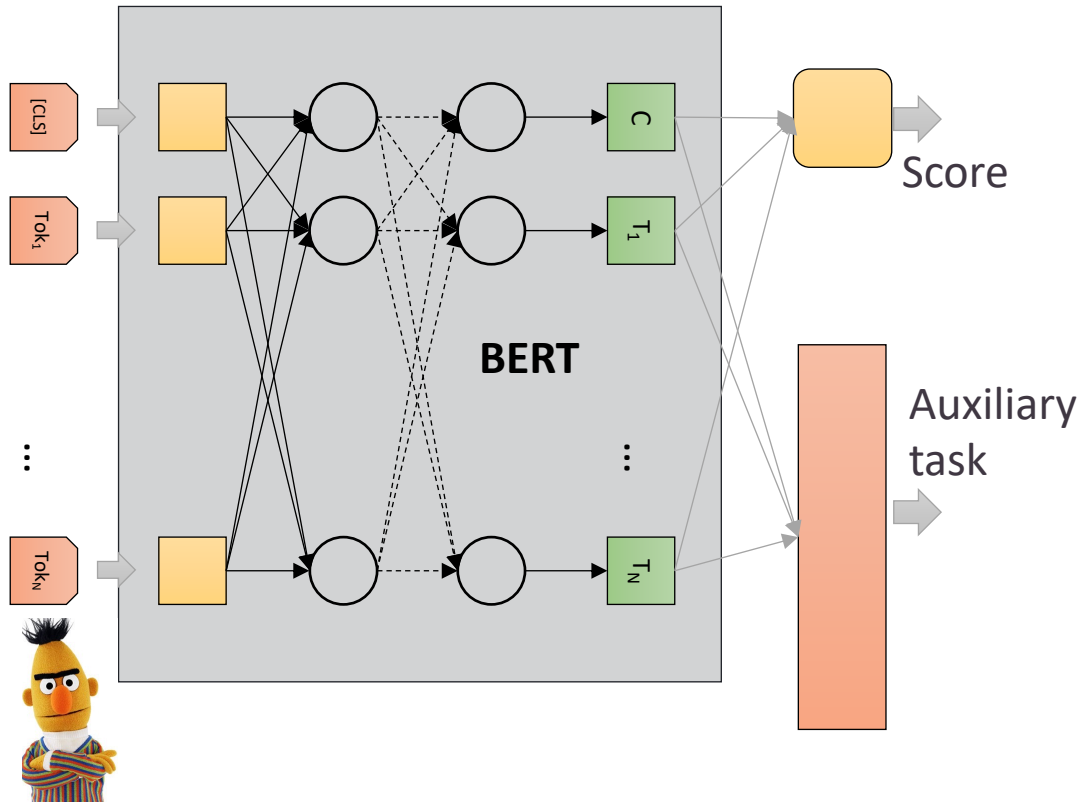
-  Massive manual feature engineering
-  No or little generalisation
-  Easy to be fooled

Neural systems

-  No need for feature engineering
-  Can be easily adapted to new data/task
-  Robust
-  *Lots of annotated data*
-  *Computational cost and hardware requirements*

Neural AA systems are still in infancy!

Automated assessment



- We presented the **first** neural multi-tasking learning AA system (Andersen et al., 2021)
 - New **state-of-the-art**
 - Using pre-trained Transformer-based language model
 - BERT
 - Multi-task learning
 - multiple NLP objectives: POS, GR, GED, L1, gender
 - Optimising a combination of the main loss and an auxiliary loss
 - $Loss = \lambda \times Loss_{main} + (1 - \lambda) \times Loss_{aux}$

Automated assessment

| | Pearson | Spearman | RMSE (0-40 scale) |
|--|--------------|--------------|-------------------|
| <i>Human-human agreement (upper bound)</i> | <i>0.797</i> | <i>0.793</i> | <i>3.72</i> |
| SOTA feature-based system | 0.782 | 0.786 | 3.73 |
| Neural AA system | 0.811 | 0.807 | 3.68 |

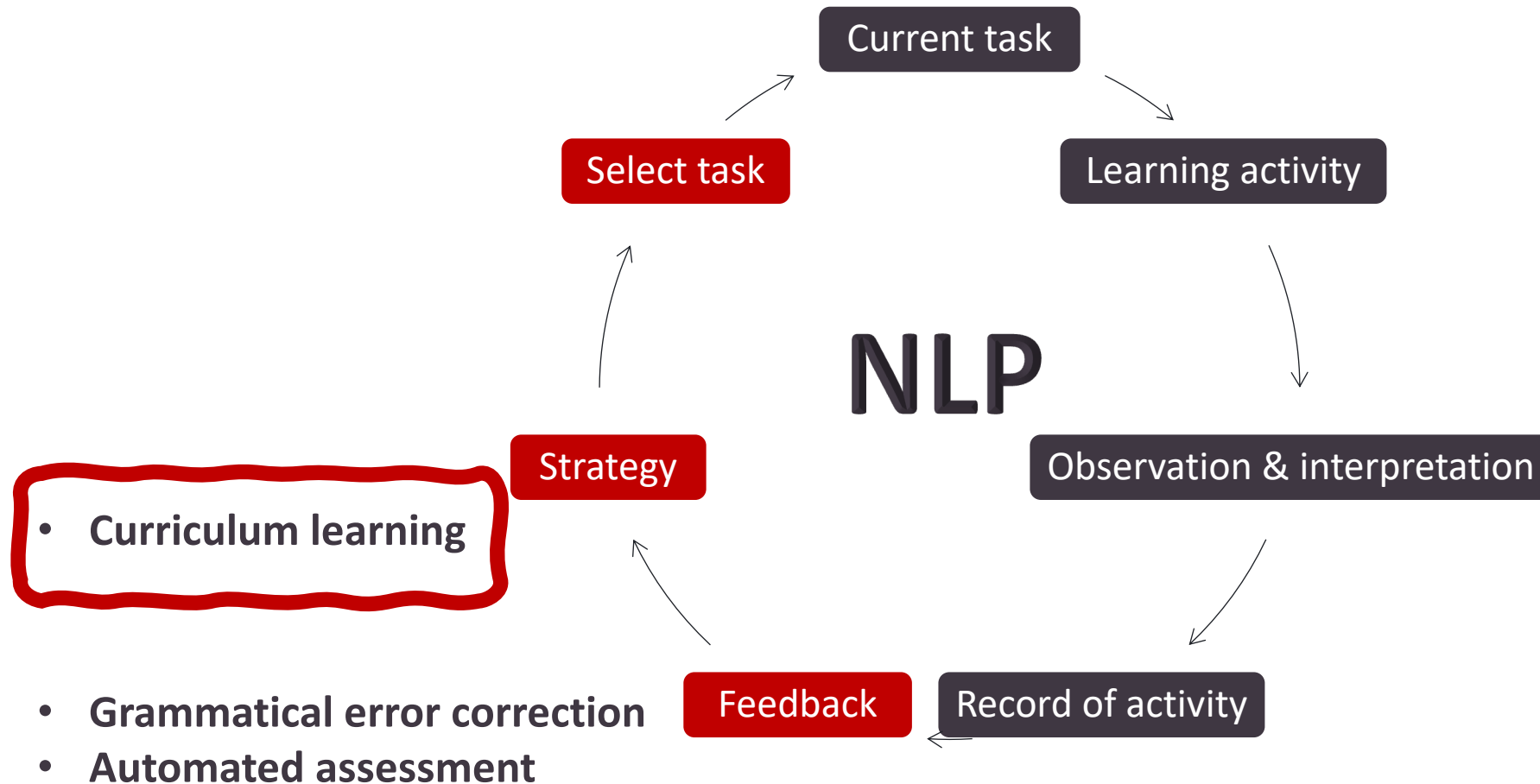
- The CLC FCE dataset: 1,244 exam scripts written by candidates sitting the Cambridge ESOL First Certificate in English (FCE) examination in 2000 and 2001
- Our best neural system can even beat human examiners

Automated assessment

| Aux. task | Pearson | Spearman | RMSE (0-40 scale) |
|-----------|--------------|--------------|-------------------|
| none | 0.779 | 0.773 | 3.87 |
| + POS | 0.760 | 0.760 | 4.11 |
| + GR | 0.784 | 0.777 | 3.79 |
| + GED | 0.788 | 0.779 | 3.81 |
| + L1 | 0.781 | 0.779 | 3.81 |
| + gender | 0.774 | 0.768 | 3.85 |

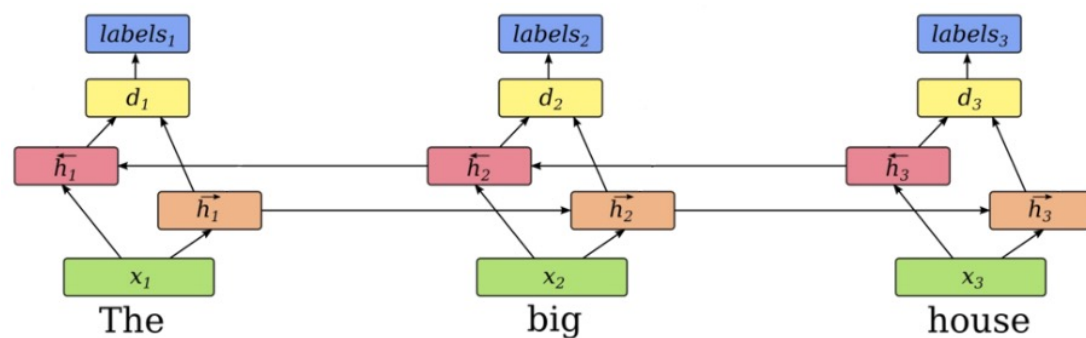
- Adding auxiliary training objectives helps
- Demographic information (L1, gender) helps
 - Need for personalised systems to provide better feedback for learning
 - Removing unwanted bias to ensure fairness in assessment

NLP for personalised language learning



Learner performance prediction

- Predict what errors each learner will make in the future based on their learning history
- LSTM-based neural sequence modelling (Yuan, 2018)
- Our approach generalises well across languages



| | English | Spanish | French |
|-------------|---------|---------|--------|
| Baseline | 0.19 | 0.17 | 0.28 |
| Yuan (2018) | 0.48 | 0.44 | 0.51 |

Duolingo SLAM dataset
 F_1 : the higher the better

Learner performance prediction

- Language-specific learner features:
 - Extracting the CEFR level for each word from the English Vocabulary Profile (EVP)
 - A1, A2, B1, B2, C1 and C2, with A1 being the lowest and C2 the highest
- Do not yield improvement:

| English | F_1 |
|------------------|-------|
| W/o CEFR feature | 0.48 |
| W CEFR feature | 0.47 |

present · *verb* [T]  /prɪˈzent/

+ Word family

+ present (GIVE)

B2 to give, provide or make known

+ present (TV/RADIO)

B2 to introduce a television or radio show

+ present (PLAY/FILM)

B2 to show a new play or film

+ present a danger/difficulty/problem, etc.

C1 to cause a danger/difficulty/problem, etc.

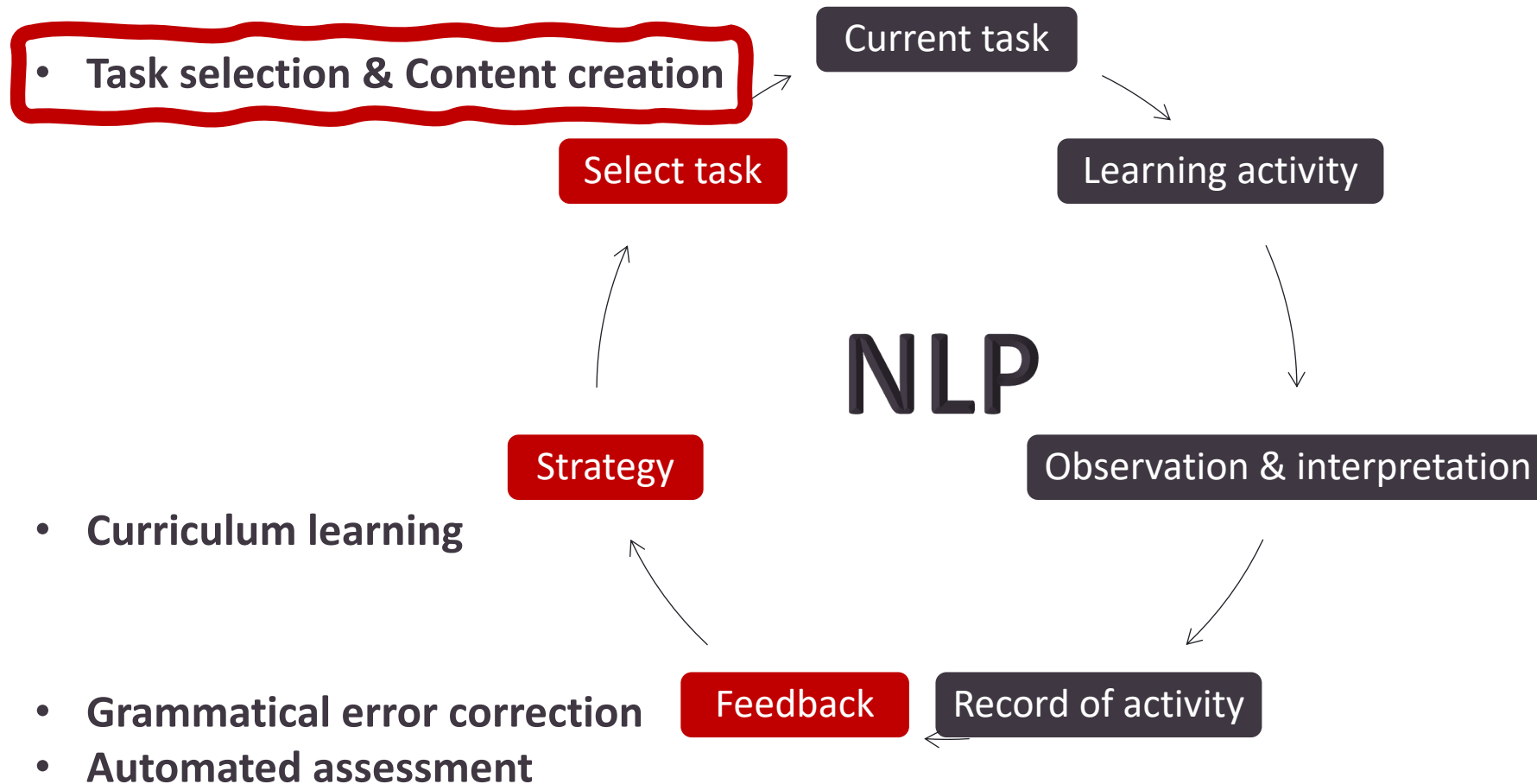
+ present (INFORMATION)

C2 to give people information in a formal way

+ present (OPPORTUNITY)

C2 If an opportunity presents itself, it becomes possible.

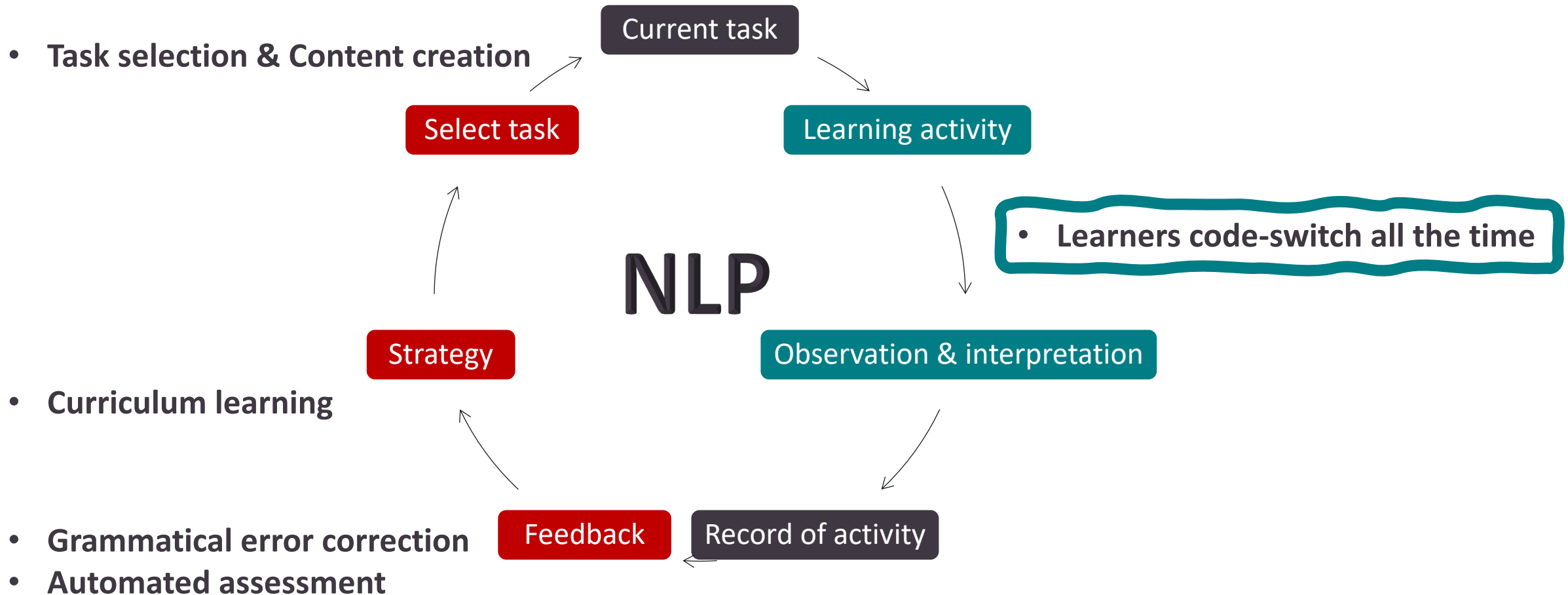
NLP for personalised language learning



Task selection

- Identify appropriate content to meet individual needs
 - Word-sense disambiguation (Yuan and Strohmaier, 2021)
 - **Zero-shot learning** for multilingual and cross-lingual NLP
 - **“the best current system”** - an independent international evaluation
- Manipulate the content so that it is at the appropriate level
 - Complex word identification & text simplification (Yuan et al., 2021)

NLP for personalised language learning



Code-switching in language education

- Code-switching (CSW) is the phenomenon where speakers use two or more languages in a single discourse or utterance
- Examples:
 - [English-*Japanese*] Even more, you can learn how to pronounce correctly. *それに, モチベーション高まります。*
'Even more, you can learn how to pronounce correctly. *Additionally, it increases motivation.*'
 - [English-*German*] My hobby is *fussball spielen*.
'My hobby is *playing football.*'

Code-switching in language education

- Over the past 50 years, language teachers have been asked to discourage students from mixing their first and second languages
- In recent years, CSW has become an increasingly recognised natural product of multilingualism

Code-switching in language education

- CSW has been shown to offer many pedagogical benefits
 - accelerating students' confidence
 - increasing their access to content
 - improving their engagement, participation and classroom rapport
- Plurilingual competence is to be encouraged, developed and rewarded
- Unfortunately, current educational technologies are not yet able to keep up with this 'multilingual turn' in education

Code-switching in language education

- Most systems are developed exclusively for monolingual data
- Existing educational NLP systems are not expected to deal with CSW input
- Any use of non-standard English is penalised as it is seen as lack of second-language proficiency
 - Online AA systems reject submissions in which non-English content is above a small threshold

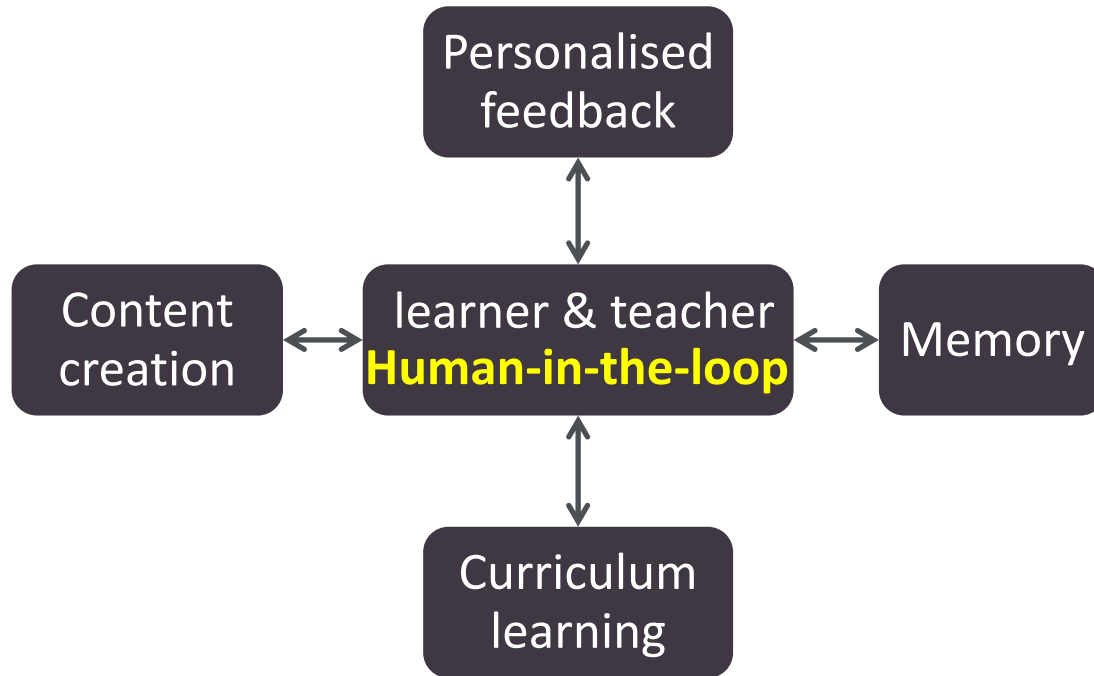
Code-switching in language education

- Most non-English words are either **ignored** (a) or flagged as **errors** (b), as systems are not able to distinguish genuine CSW from real errors
 - a. [*Chinese-English*] 当下倾盆大雨时, we will stay home and our dad will tell us Halloween stories.
'*When it is pouring*, we will stay at home and our dad will tell us Halloween stories.'
 - b. [*English-Spanish*] That was everything related to this *situación*.
'That was everything related to this *situation*.'
- Systems fail to identify errors in the neighbouring words, as soon as non-English words are involved

Code-switching in language education

- New designs for feedback systems
 - Personalised systems that are capable of catering for different first languages and proficiency levels. E.g.:
 - *Learner's sentence: 'My hobby is fussball spielen'*
 - *Feedback: 'Did you mean "My hobby is **playing football**"?'.*
- This kind of information could then be fed back into content creation and adaptive learning
- A very difficult task

NLP for personalised language learning



- The ultimate goal is not to replace human teachers but to
 - Enhance the learning experience
 - Maximise the vital human aspects of teaching

- Fully-funded PhD studentships at King's College London
- Workshop on *Multilingualism in education: A holistic approach to building technologies for language learners*

Thank you!

zheng.yuan@kcl.ac.uk

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