Edit Distances

and their application to downstream tasks, in research and commercial contexts

CENTRE FOR TRANSLATION STUDIES

UNIVERSITY OF SURREY

Félix do Carmo & Diptesh Kanojia

f.docarmo@surrey.ac.uk

d.kanojia@surrey.ac.uk

Agenda

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» **Part one:**
- What are edit distances and their different implementations and applications.

» **Part two:**
- Demonstration of an exercise which simulates a sequence of edit steps with increasing complexity, analyzed with TER.

Discussion time and short break

» **Part three:**
- Computational perspective
- Exercise prepared with Python packages, which includes edit distances and similarity metrics (strsimpy). Participants can reproduce the sequence of the exercise in a Python notebook that will be shared, or to watch the step-by-step demonstration.

» **Part four:**
- Implications for research and commercial applications of edit distances in cases such as use of MQM in QE and APE.

» **Part five:**
- Discussion and Q&A.

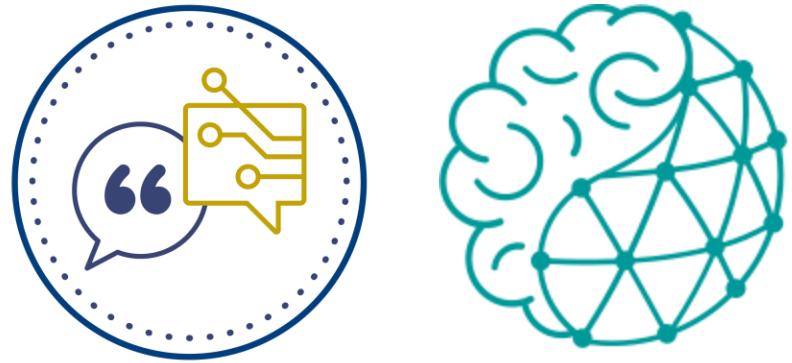
Abstract

- *The tutorial will describe TER, Levenshtein and Damerau's distances and several other edit distances, and disassemble them into their essential components. We will discuss the centrality of four editing actions: insert, delete, replace and move words, and show their implementations in openly available packages and toolkits.*

- *The application of edit distances in downstream tasks often assumes that these accurately represent work done by post-editors and real errors that need to be corrected in MT output. We will discuss how imperfect edit distances are in capturing the details of this error correction work and the implications for researchers and for commercial applications of these uses of edit distances. In terms of commercial applications, we will discuss their integration in computer-assisted translation tools and how the perception of the connection between edit distances and post-editor effort affects the definition of translator rates.*

Part 1:

Edit distances and their different implementations and applications



Edit distances

- What are “edit distances”?
- What are these used for?

Edit distance

2 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

In computational linguistics and computer science, **edit distance** is a string metric, i.e. a way of quantifying how dissimilar two strings (e.g., words) are to one another, that is measured by counting the minimum number of operations required to transform one string into the other. Edit distances find applications in natural language processing, where automatic spelling correction can determine candidate corrections for a misspelled word by selecting words from a dictionary that have a low distance to the word in question. In bioinformatics, it can be used to quantify the similarity of DNA sequences, which can be viewed as strings of the letters A, C, G and T.

Different definitions of an edit distance use different sets of like operations. Levenshtein distance operations are the removal, insertion, or substitution of a character in the string. Being the most common metric, the term *Levenshtein distance* is often used interchangeably with *edit distance*.[1]

Source: https://en.wikipedia.org/wiki/Edit_distance

Edit distances and translation

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

- Translation Edit Rate (TER): Snover et al 2006
- Human-Targeted Edit Rate (HTER): Snover et al 2006 & Snover et al 2009b
- Translation Edit Rate + (TERplus): Snover et al 2009a

- Word Error Rate (WER): Tillmann et al 1997
- Number, Edition error, Recognition error (NER) : Romero-Fresco & Perez 2011

- To identify errors of MT
- To identify effort and evaluate post-editors/translators

Elements of editing

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

- 4 operations or actions:
  - Deleting
  - Inserting
  - Replacing
  - Moving
- Different edit distances use/estimate different operations.

Types of edit distance [edit]

Different types of edit distance allow different sets of string operations. For instance:

Algorithm	Operations Allowed			
	Insertions	Deletions	Substitutions	Transposition
Levenshtein Distance	✓	✓	✓	
Longest Common Subsequence (LCS)	✓	✓		
Hamming Distance			✓	
Damerau–Levenshtein Distance	✓	✓	✓	✓
Jaro distance				✓

Some edit distances are defined as a parameterizable metric calculated with a specific set of allowed edit operations, and each operation is assigned a cost (possibly infinite). This is further generalized by DNA sequence alignment algorithms such as the Smith–Waterman algorithm, which make an operation's cost depend on where it is applied.

Source: https://en.wikipedia.org/wiki/Edit_distance

4 simple operations

- Deletion

  Example 1:

  a) A word compound is a complex thing.
  b) A word **[X]** is a complex thing.

- Insertion

  Example 2:

  a) A word is a complex thing.
  b) A word **[compound]** is a complex thing.

- Replacement (substitution)

  Example 3:

  a) A word is a complex thing.
  b) A **[compound]** is a complex thing.

- Movement (transposition / shift)

  Example 4:

  a) A word compound is a complex thing.
  b) A **[compound word]** is a complex thing.

2 simple and 2 complex operations

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Deletion and Insertion

- Changes in positions (index) of words
  - In one of the strings, there is an empty position
- One operation is the opposite of the other (to reverse a deletion, you do an insertion)
- They result in a change in the total number of words in a string
- They always have cost 1

» Replacement and Movement

- Change in form only (position remains the same): replacement
  - Opposite of replacement is replacement by the previous form
- Change in position only (form remains the same): movement
  - Opposite of movement is movement in opposite direction
- They may not result in a change in the total number of words in a sentence
- They may have a higher cost

Economical/efficient operations

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Insertion and deletion are “primary”, incomplete operations.
  - *You cannot edit with only deletions or only insertions.*

» Replacement and movement are “secondary” and more complex operations
  - Both can be described as simultaneous deletions and insertions
    - Replacement: deletion of one form and insertion of a different form in the same position
    - Movement: deletion of one form in one position and insertion of the same form in a different position
  - *You can describe editing only with replacement*

» Do we need this complexity? Recall the “minimum number of operations” (shortest path)
  - These metrics describe the most economical/efficient processes
  - Secondary operations are more economical than primary ones
  - The most detailed the metric, the more efficient it is in identifying the shortest path
  - *(If post-editors are only trained to do deletions and insertions, they are not trained to be efficient.)*

The most complex operation: movement

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Example 4:

1. *A word compound is a complex thing.*
2. *A compound word is a complex thing.*

   Reality: 1 movement (transposition: move "word" 1 position forward)

   Estimate: 1 transposition ("word" 1 position forward OR "compound" 1position back") OR 2 replacements OR 2 deletions + 2 insertions

» Example 5:

1. *A blue trailer big truck.*
2. *A big blue trailer truck.*

   Reality: 1 movement ("big" 2 positions back)

   Estimate: 1 movement ("big" 2 positions back OR "blue trailer" 1 position forward); OR 3 replacements OR 3 deletions + 3 insertions

» Example 6:

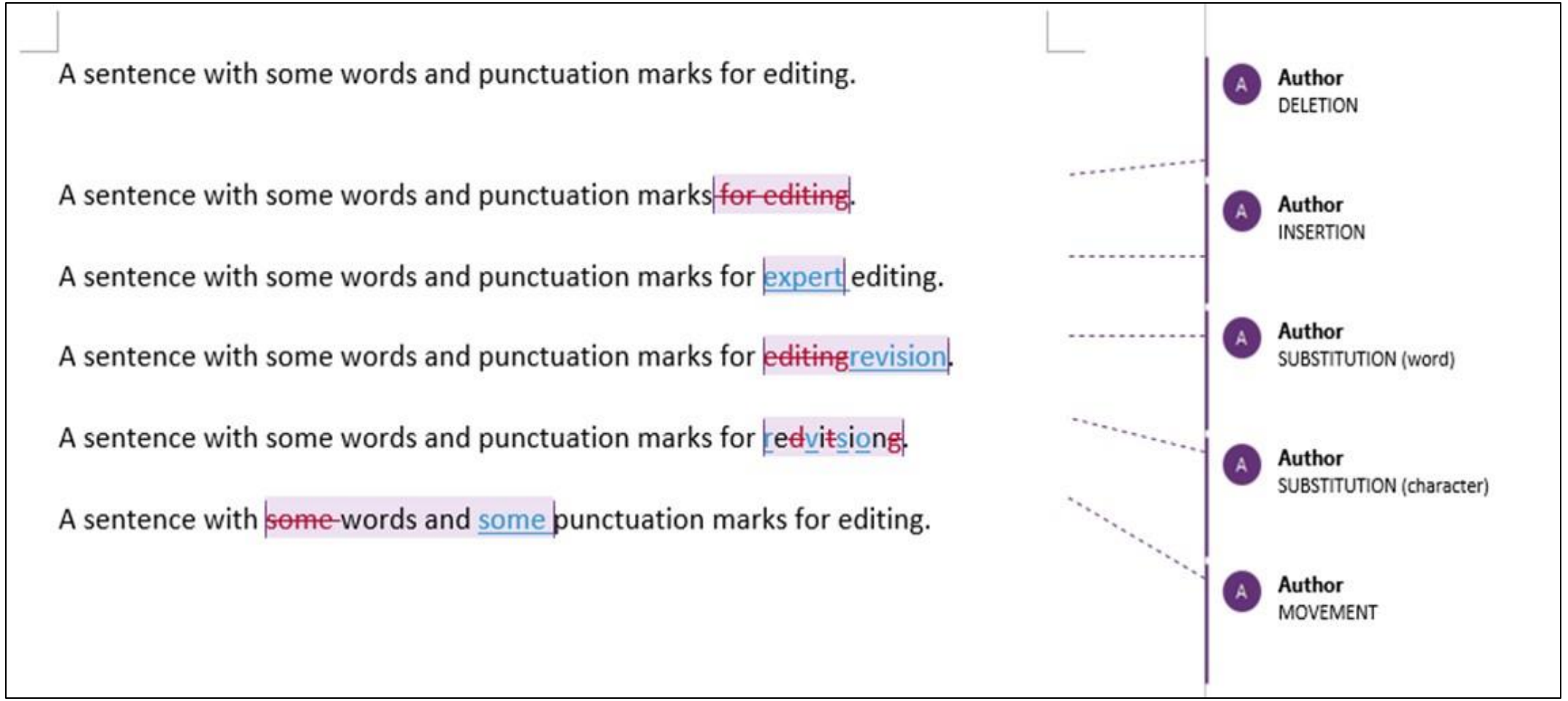
1. *Tomorrow I will do that.*
2. *I will do that tomorrow.*

   Reality: 1 movement (≠ transposition / shift) ("tomorrow" to the end of the sentence: 4 positions forward)

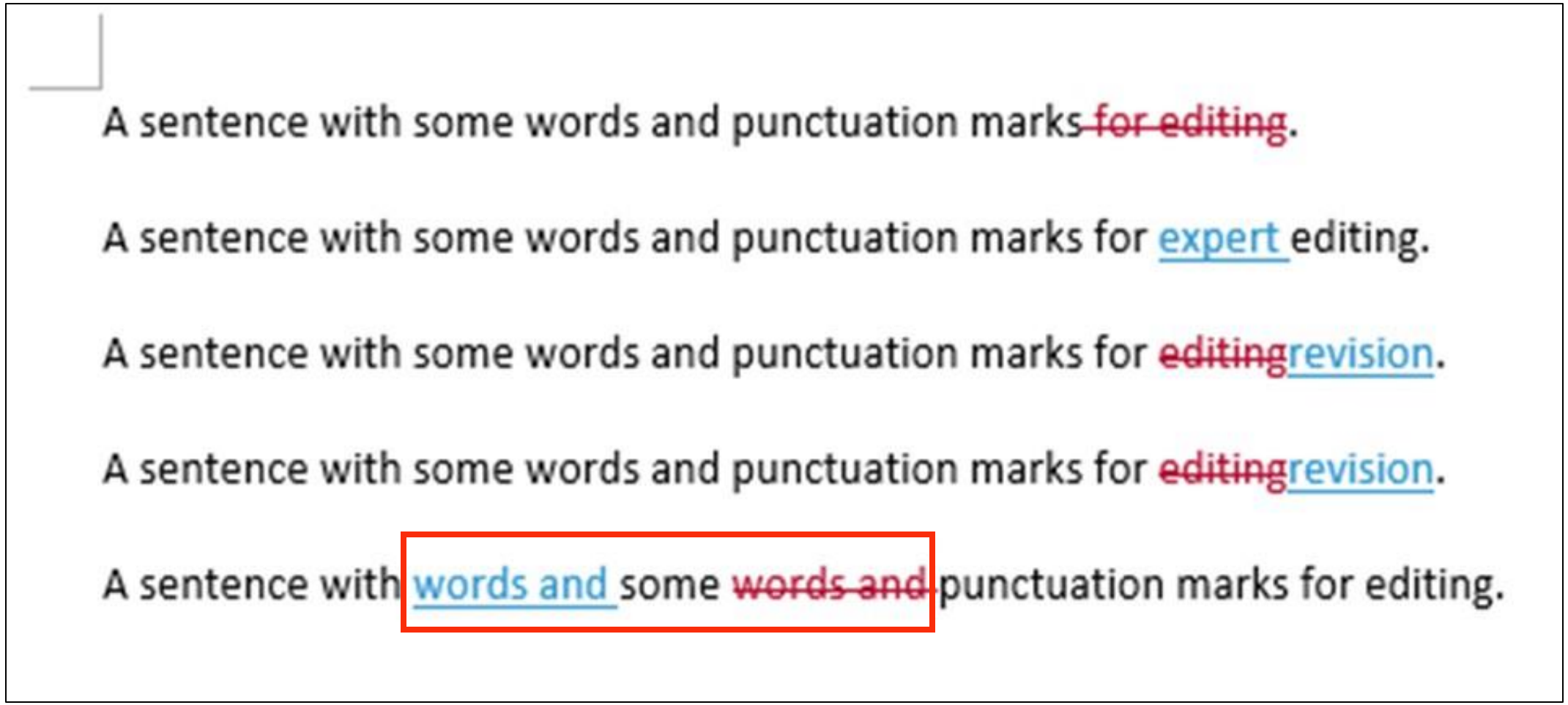
   Estimate: 1 movement ("tomorrow" 4 positions forward OR "I will do that" 4 positions back); OR 5 replacements (all the words in the sentence); OR 5 deletions + 5 insertions

- It has been shown that estimating movement is an NP-complete task (Shapira & Storer 2007)

Difference between reality and estimate



Difference between reality and estimate



Edit distances are imperfect proxies for editing

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» If you want to capture what the translator (post-editor) did, edit distances are inaccurate estimates

» They may mis-identify the position(s) and the form(s) which were edited.

» Recall that there are many fundamental differences between the various edit distances available
  - Different code implementations estimate edit operations differently
    - Scores may vary a lot (part 3 of the tutorial)

Difference between TER and HTER

- HTER – Human-targeted (or human-mediated) Translation Edit Rate (Snover et al 2006)

- HTER is not a TER applied to human data (PE)
  - In HTER, as used by Snover et al (2006), monolingual human annotators choose, from several references, the one that preserves most of the MT output, and edit it, if necessary, to reduce the distance even further as long as that kept the semantic relationship to the other references.
  - In HTER, the semantic relationship to the source text may get lost, so this is not PE.
- 'Human-targeted' has been interpreted as meaning something like 'targeted at a human process', giving access to the process of creation of a PE version, when in fact it means "targeted [by human annotators] for this system output" (Snover et al., 2006, 2).

- HTER is the exactly the same method, same error rate, as TER.
- *If the method is the same, why would we need a different name (a different metric) just because we are measuring a different thing?*

Difference between TER/HTER and TERplus

» TERplus (or TERp) (Snover et al., 2009)

- Sub-types of substitution:
  - replacing whole words (normal substitution)
  - replacing words by one of its variants (stem matching)
  - replacing with a synonym (synonym matching)
  - and replacing phrases (phrase substitution or paraphrase).
- Cost for first 3 subtypes of substitution is 1
  - Cost of substitution by paraphrase is not fixed, estimated by a combination of the probabilities of that paraphrase and the amount of edits needed to align the two phrases.
- Relaxation of the criteria to identify a shift.
- TERp is capped at cost 1 per string/edit, but TER is not capped.
  - When a one-word string becomes a string with three words:
    - TERp presents an edit score of 100.00 (which means 'all words in the string [1.00 or 100%] were edited')
    - TER presents an edit score of 300.00 (which means that the number of words tripled).

Difference between error (rates) and edit (rates)

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Footnote in Snover et al. (2006) explains that the 'E' in TER should be understood as an 'edit', not as an 'error'
  - There may be different edits for the same error

» But the difference is also in what you are measuring (hypothesis) against what (reference).
  - MT against PE: identify the errors in the MT output against a reference
  - PE against MT: estimate the edits the post-editor did

» Default setting in TERcom (Snover et al 2006)

» TER(mt,pe)
  - MT output as the hypothesis and the PE version as the reference

  - A deletion will mean "a word that the MT output has missed"
  - It does not mean "a word that the translator has missed"
  - TER is an MT "error rate"

Difference between error (rates) and edit (rates)

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» If you want to estimate what the translator has done, you should invert the hypothesis and the reference

» HER(pe,mt)
- Human Edit Rate (do Carmo, 2021)
- Use PE as the hypothesis and compare it to the MT as the reference
- In this case, a deletion is “a word that the translator has deleted”
- HER is an “edit rate”

» HER (as an inversion of TER) affects mainly deletions and insertions
- Inversion of replacements should still be replacements
- Inversion of movements should also be movements

Edit distances are statistical in nature

- Account for variance and error margin

- Let's look very briefly at the computational side of all of this, as an introduction to the more detailed work we will do in part 3 of the tutorial

How is ‘distance' captured, computationally?

» **Example -> “cat”::“chat”**

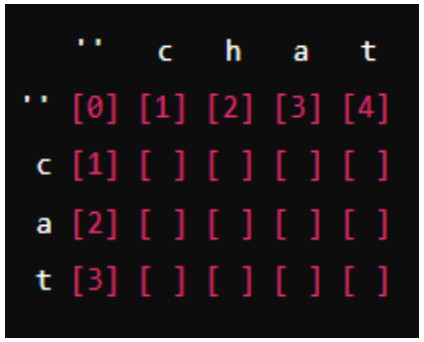
» **Aim:** fill the matrix with minimum edit distance value for each character pair in “cat” and “chat”

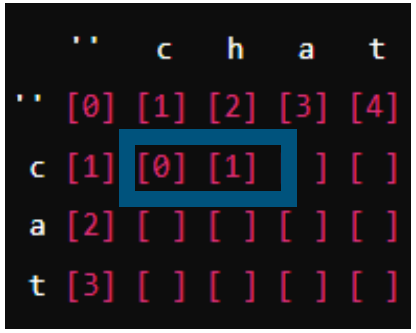
- Matrix initialization (*m*+1) X (*n*+1)
  - ***m*** -> source length; ***n*** -> target length
- Recursive computation for each cell
  - compute cost for each operation- insertion, deletion, substitution, and select the minimum.
- For each computation,
  - c equals c; value 0
  - c does not equal h; value 1

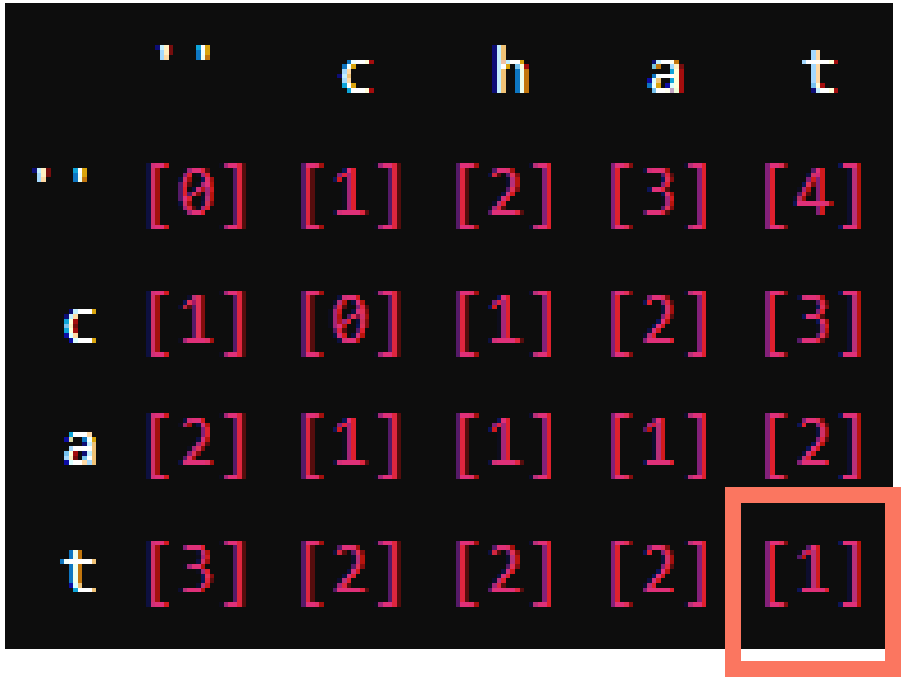
» Dynamic Programming

» Considers all possible ways to transform one string into another.

- Value on bottom-right corner is ‘edit distance’







Applications of Edit Distance in NLP

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

- **Edit distance** is a key concept applied across text comparison tasks in NLP
  - as a key metric calculating the minimum number of operations required to transform one text sequence into another.
- **Focus on MT**
  - Translation accuracy
  - Post-editing workflows, and accuracy
  - Evaluating translation/post-editing quality
- **Applications beyond MT** in NLP
  - *Cognate Detection* – between cross-lingual text written in same/similar script.
  - *Spell Checking* – between misspelled word and dictionary entries
  - *Text Normalization* – non-standard forms of words to standardized forms
  - *Document Similarity* – similarity between documents using distance.

original:

```
This part of the
document has stayed the
same from version to
version.  It shouldn't
be shown if it doesn't
change.  Otherwise, that
would not be helping to
compress the size of the
changes.

This paragraph contains
text that is outdated.
It will be deleted in the
near future.

It is important to spell
check this dokument. On
the other hand, a
misspelled word isn't
the end of the world.
Nothing in the rest of
this paragraph needs to
be changed. Things can
be added after it.
```

new:

```
This is an important
notice! It should
therefore be located at
the beginning of this
document!

This part of the
document has stayed the
same from version to
version.  It shouldn't
be shown if it doesn't
change.  Otherwise, that
would not be helping to
compress the size of the
changes.

It is important to spell
check this document. On
the other hand, a
misspelled word isn't
the end of the world.
Nothing in the rest of
this paragraph needs to
be changed. Things can
be added after it.

This paragraph contains
important new additions
to this document.
```

The command `diff original new` produces the following *normal diff output*:

```
0a1,6
> This is an important
> notice! It should
> therefore be located at
> the beginning of this
> document!
>
11,15d16
< This paragraph contains
< text that is outdated.
< It will be deleted in the
< near future.
<
17c18
< check this dokument. On
---
> check this document. On
24a26,29
>
> This paragraph contains
> important new additions
> to this document.
```

- A key application example, **diff**, in computer science.

Image Source: Wikipedia

Applications in NLP beyond MT

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» **Spell Checking**

- *e.g.,* misspelled word "hte" is compared to dictionary entries like "the", "hat", and "hit", with "the" *having the smallest edit distance* (1 substitution).
- *e.g.,* In word processors, typing "recieve" will prompt a correction to "receive", as edit distance identifies "receive" as the closest match (1 transposition).

» **Text Normalization**

- *e.g.,* "u r awesome" is normalized to "you are awesome" using edit distance to match non-standard abbreviations with their full forms.

» **Document Similarity**

- Two documents discussing similar topics are compared at a sentence or paragraph level, using edit distance to quantify how much one text must be altered to resemble the other.

- A specific use case is **plagiarism detection**, edit distance can measure how much content was copied or slightly altered between two texts.

Applications in MT Evaluation and Correction

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» **High Level:** Most metrics compare machine translated output with human-generated reference.

» Statistical metrics

- BLEU: n-gram precision b/w **h**ypothesis and **r**eference.
- METEOR: combines unigram precision, recall and alignment b/w ***h*** and ***r***.
- TER: measures number of edits (insertions, deletions, substitutions) needed to transform ***h*** to ***r***.
- chrF : measures character-level precision, recall, and F1-score b/w ***h*** and ***r***.

» Translation Workflows

- Phrase alignment
- Quality feedback during MT model training
- Handling translation and domain divergence

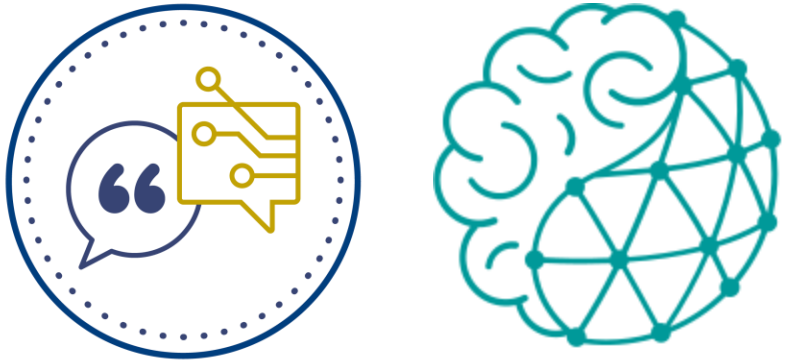
» Post-Translation Workflows

- Evaluating translation accuracy
  - *e.g.*, Quality Estimation data curation workflow
- Evaluating (automatic) post-editing quality

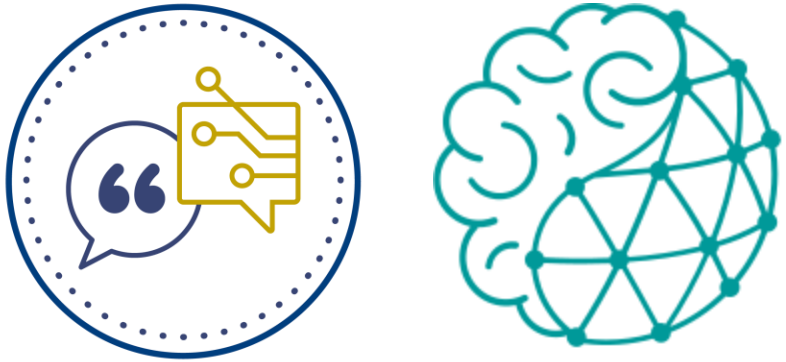
Conclusion of part 1

- There are many different edit distances
- Different implementations may record different edits
  - TER has three well-known implementations (TERCOM, SacreBLEU, PyTER).
  - We discuss all three implementation in this tutorial (Parts 2 and 3)
- TER is an error rate: you use MT as hypothesis and PE as reference
  - It identifies errors in the MT output
- To have an edit rate, you should use PE as hypothesis and MT as reference: HER
  - It estimates the edits done by PE
- All edit distances are estimates, of the shortest path
- Check the purpose and the accuracy you need for your purposes
- Applied to many downstream tasks in NLP which what we discuss in Part 4.

Any questions?



Part 2:



Analysing an incrementally-complex sequence of edits

How accurate is TER?

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Does TER, as an edit rate (HER), reveal the edits performed by the post-editors?

» Does it measure edits or words edited?

- If you delete a two-word phrase, does that count as one edit, or two words edited?

» Does it identify the words that are deleted, inserted, replaced and moved correctly?

1. *A blue trailer big truck.*
2. *A big **blue trailer** truck.*

- If the translator chose to move “blue trailer”, it may be because there is a syntactic/semantic unit here, and I may want to capture this cohesion.
  - This unit may always need to be together
  - This edit may be repeated later again.

How accurate is TER?

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Go to: https://github.com/surrey-nlp/AMTA-EditDistances-tutorial

» See the two txt files there:
- Experiment-UNEDITED.txt contains a 15-word sentence repeated 50 times, formatted to be processed by TERcom
- Experiment-EDITED.txt contains the same 15-word sentence, after applying a sequence of 50 incrementally-complex edits.

» NOTE: After the break, in part 3 of this tutorial, you will be able to use a Python notebook to apply different edit distances and similarity metrics to extend this exercise.

Method

- After creating the “edited” file with a growing sequence of edits, we measured the edit distance using TERcom
  - TERcom is a free tool to estimate TER, available at: http://www.cs.umd. edu/~snover/tercom/
  - We used version 0.7.25, in Java code
  - The command we used was:
    - java -jar tercom.7.25. jar -N -s -r C:/…/tercom/files/**File-Ref.txt** -h C:/…/tercom/files/**File-Hyp.txt** -n C:/…tercom/outputs/Exp1.Out.txt.

- To use TERcom to estimate an edit rate (HER), we used the Experiment-EDITED.txt as the hypothesis and the Experiment-UNEDITED.txt as the reference.

- TERcom outputs different reports, with different levels of detail, which are worth analysing

How does TERcom estimate the edits?

Dynamic programing (recursively)

- Count the number of words in *h* and *r*
  - Mind tokenization in non-latin based languages
- Create an index for each unit (word in position 1, etc.)
- Compare words and positions
- Identify missing words in either *h* (deletion) or *r* (insertion)
- Identify new forms in *r* with the same index (position) as in *h* (replacement)
- Then, by a greedy search algorithm, test moving units in *r*, one position at a time, and compare them to *h*
  - Identify same unit with different index
  - Do this by words first and then by phrases (n-grams)
  - Mark this as movement (shift) when it reduces the number of insertions and deletions
    - Economical and efficient
  - This is what makes movement so hard to estimate
  - Movement is the only operation that identifies actions affecting more than one word at a time
- All operations, even movement of two or three words, have cost 1
- Numbers of edits are normalised by total of words in the string

Outputs of TERcom

» Summary reports (.sum files)

- TER(mt,pe)

Sent Id	Ins	Del	Sub	Shft	WdSh	NumEr	NumWd	TER
Sentence1	**1**	0	0	0	0	1	7	**14.286**
Sentence2	0	**1**	0	0	0	1	9	**11.111**
Sentence3	0	0	**1**	0	0	1	7	14.286
Sentence4	0	0	0	**1**	1	1	17	5.882
TOTAL	1	1	1	1	1	4	40	10.000

- HER(pe,mt)

Sent Id	Ins	Del	Sub	Shft	WdSh	NumEr	NumWd	HER
Sentence1	0	**1**	0	0	0	1	8	**12.500**
Sentence2	**1**	0	0	0	0	1	8	**12.500**
Sentence3	0	0	**1**	0	0	1	7	14.286
Sentence4	0	0	0	**1**	1	1	17	5.882
TOTAL	1	1	1	1	1	4	40	10.000

Outputs of TERcom

» XML report

TER(mt,pe)

```
<seg segid="Sentence4">
  <hyp id="1" refid="" wrd_cnt="17.0" num_errs="1.0">
    "In","In",C,0
    "this","this",C,0
    "sentence","sentence",C,0
    ",",",",C,0
    "all","all",C,0
    "words","words",C,-2
    "are","are",C,0
    "correct","correct",C,0
    ",",",",C,0
    "but","but",C,0
    "one","one",C,0
    "is","is",C,0
    "in","in",C,0
    "the","the",C,0
    "wrong","wrong",C,0
    "position","position",C,0
    ".",".",C,0
  </hyp>
</seg>
```

HER(pe,mt)

```
<seg segid="Sentence4">
  <hyp id="1" refid="" wrd_cnt="17.0" num_errs="1.0">
    "In","In",C,0
    "this","this",C,0
    "sentence","sentence",C,0
    ",",",",C,0
    "all","all",C,0
    "are","are",C,0
    "correct","correct",C,0
    "words","words",C,2
    ",",",",C,0
    "but","but",C,0
    "one","one",C,0
    "is","is",C,0
    "in","in",C,0
    "the","the",C,0
    "wrong","wrong",C,0
    "position","position",C,0
    ".",".",C,0
  </hyp>
</seg>
```

Outputs of TERcom

» Report in .pra files

```
Sentence ID: Sentence1:1
Original Ref: This sentence has a redundant {superfluous} word .
Original Hyp: This sentence has a redundant word .
Hyp After Shift: This sentence has a redundant word .
Alignment: (     D  )
NumShifts: 0
Score: 0.125 (1.0/8.0)
Sentence ID: Sentence2:1
Original Ref: In this sentence , a [] is missing .
Original Hyp: In this sentence , a word is missing .
Hyp After Shift: In this sentence , a word is missing .
Alignment: (     I   )
NumShifts: 0
Score: 0.125 (1.0/8.0)
Sentence ID: Sentence3:1
Original Ref: This sentence has a incorrect word .
Original Hyp: This sentence has a _corrected_ word .
Hyp After Shift: This sentence has a corrected word .
Alignment: (    S  )
NumShifts: 0
Score: 0.14285714285714285 (1.0/7.0)
Sentence ID: Sentence4:1
Original Ref: In this sentence , all are correct words , but one is in the wrong position .
Original Hyp: In this sentence , all words are correct , but one is in the wrong position .
Hyp After Shift: In this sentence , all are correct words , but one is in the wrong position .
Alignment: (                  )
NumShifts: 1
  [5, 5, 7/7] ([words])
Score: 0.058823529411764705 (1.0/17.0
```

First set of edits: simple (1 word)

		EDITED		TERCOM SCORES						
	ACTIONS	Actions	Words	I	D	S	Sh	WdSh	NumEr	HER
1	Delete 1 word	1	1		1				1	6.67%
2	Insert 1 word	1	1	1					1	6.67%
3	Replace 1 word	1	1			1			1	6.67%
4	Move 1 word 1 position forward	1	1				1	1	1	6.67%
5	Move 1 word 1 position back	1	1				1	1	1	6.67%
6	Move 1 word 2 positions forward	1	1				1	1	1	6.67%
7	Move 1 word 2 positions back	1	1				1	1	1	6.67%

Second set of edits: simple (1 phrase)

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

		EDITED		TERCOM SCORES						
	ACTIONS	Actions	Words	I	D	S	Sh	WdSh	NumEr	HER
8	Delete 1 phrase (2 words)	1	2		*2				2	13.33%
9	Delete 1 phrase (3 words)	1	3		*3				3	20.00%
10	Insert 1 phrase (2 words)	1	2	*2					2	13.33%
11	Insert 1 phrase (3 words)	1	3	*3					3	20.00%
12	Replace 1 phrase (2 words)	1	2			*2			2	13.33%
13	Replace 1 phrase (3 words)	1	3			*3			3	20.00%
14	Move 1 phrase (2 words) 1 position forward	1	2				1	*1	1	6.67%
15	Move 1 phrase (2 words) 1 position back	1	2				1	2	1	6.67%
16	Move 1 phrase (3 words) 1 position forward	1	3				1	*1	1	6.67%
17	Move 1 phrase (3 words) 1 position back	1	3				1	*1	1	6.67%
18	Move 1 phrase (2 words) 2 positions forward	1	2				1	2	1	6.67%
19	Move 1 phrase (2 words) 2 positions back	1	2				1	2	1	6.67%
20	Move 1 phrase (3 words) 2 positions forward	1	3				1	*2	1	6.67%
21	Move 1 phrase (3 words) 2 positions back	1	3				1	*2	1	6.67%

Third set of edits: complex (2 edits per sentence)

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

		EDITED		TERCOM SCORES						
	ACTIONS	Actions	Words	I	D	S	Sh	WdSh	NumEr	HER
22	Delete 1 + 1 word (dift positions)	2	2		2				2	13.33%
23	Insert 1 + 1 word (dift positions)	2	2	2					2	13.33%
24	Replace 1 + 1 word (dift positions)	2	2			2			2	13.33%
25	Move 1 + 1 word one pos. fwd (dift positions)	2	2				2	2	2	13.33%
26	Delete 1 word + Insert 1 word (dift positions)	2	2	1	1				2	13.33%
27	Delete 1 word + Replace 1 word (dift positions)	2	2		1	1			2	13.33%
28	Delete 1 word + Move 1 word one pos. fwd (dift positions)	2	2		1		1	1	2	13.33%
29	Insert 1 word + Delete 1 phrase (2wd) (dift positions)	2	3	1	*2				3	20.00%
30	Insert 1 word + Replace 1 phrase (2wd) (dift positions)	2	3	1		*2			3	20.00%
31	Insert 1 word + Move 1 phrase (2wd) one pos. fwd (dift positions)	2	3	1			1	*1	2	13.33%
32	Replace 1 word + Delete 1 phrase (3wd) (dift positions)	2	4		*3	1			4	26.67%
33	Replace 1 word + Insert 1 phrase (3wd) (dift positions)	2	4	*3		1			4	26.67%
34	Replace 1 word + Move 1 phrase (3wd) one pos. (dift positions)	2	4			1	1	*1	2	13.33%
35	Delete 1 phrase (2wd) + Insert 1 phrase (2wd)	2	4	*2	*2				4	26.67%
36	Delete 1 phrase (2wd) + Replace 1 phrase (2wd)	2	4		*2	*2			4	26.67%
37	Delete 1 phrase (2wd) + Move 1 phrase (2wd) 2 positions	2	4		*2		1	2	*3	20.00%

Fourth set of edits: complex (3 or more per sentence)

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

		EDITED		TERCOM SCORES						
	ACTIONS	Actions	Words	I	D	S	Sh	WdSh	NumEr	HER
38	Delete 1 word + Insert 1 word + Replace 1 word	3	3	1	1	*2			*4	26.67%
39	Delete 1 word + Insert 1 word + Move 1 word 5 positions	3	3	*	*	*1	*2	*2	3	20.00%
40	Delete 1 word + Replace 1 word + Move 1 word 7 positions	3	3		1	1	1	1	3	20.00%
41	Replace 1 word + Insert 1 phrase (2wd)	2	3	*2		1			3	20.00%
42	Replace 1 phrase (2wd) + Insert 1 word	2	3	1		*2			3	20.00%
43	Replace 1 phrase (2wd) + Insert 1 phrase (2wd)	2	4	*2		*2			4	26.67%
44	Insert 1 word + Delete 1 phrase (2wd) + Replace 1 word	3	4	*	*1	*2	*1	*1	4	26.67%
45	Insert 1 phrase (2wd) + Delete 1 word + Replace 1 word	3	4	*2	1	1			4	26.67%
46	Insert 1 word + Delete 1 word + Replace 1 phrase (2wd)	3	4	1	1	*2			4	26.67%
47	Insert 1 ph (2wd) + Delete 1 ph (2wd) + Replace 1 ph (2wd)	3	6	*2	*2	*2			6	40.00%
48	Insert 1 word + Delete 1 ph (2wd) + Replace 1 word + Move 1 word	4	5	*	1	*2	*2	*2	5	33.33%
49	Insert 1 ph (2wd) + Delete 1 word + Replace 1 word + Move 1ph (2wd)	4	6	*2	1	1	1	2	*5	33.33%
50	Insert 1 word + Delete 1 word + Replace 1 ph (2wd) + Move 1 word	4	5	*	*	*3	*2	*2	5	33.33%

What if we only need the global scores?

TOTALS	I	D	S	Sh	Edits	Words	NumEr	NumWd	HER
HER scores	30	31	38	27			126	750	16.80%

What if we only need the global scores?

TOTALS	I	D	S	Sh	Edits	Words	NumEr	NumWd	HER
HER scores	30	31	38	27			126	750	16.80%
Performed	23	23	23	23					

What if we only need the global scores?

TOTALS	I	D	S	Sh	Edits	Words	NumEr	NumWd	HER
HER scores	30	31	38	27			126	750	16.80%
Performed	23	23	23	23					
Over estimated	130%	135%	165%	117%					

Secondary operations are at the extremes of over-estimation:

- Replacement can describe all editing
- Movement is the most efficient of the four actions, and the most complex to identify

What if we only need the global scores?

TOTALS	I	D	S	Sh	Edits	Words	NumEr	NumWd	HER
HER scores	30	31	38	27			126	750	16.80%
Performed	23	23	23	23	92	142		750	
Over estimated	130%	135%	165%	117%					

TER scores do not reflect the number of edited words, nor the number of edits, but a combination of both.

What if we only need the global scores?

TOTALS	I	D	S	Sh	Edits	Words	NumEr	NumWd	HER
HER scores	30	31	38	27			126	750	16.80%
Performed	23	23	23	23	92	142		750	
Over estimated	130%	135%	165%	117%					

Rate/edits	Rate/words
12.27%	18.93%

- TER/HER, as implemented in TERcom, is “biased” towards replacements and against movements.
- TER/HER, as implemented in TERcom, overestimates the number of edits, but it underestimates the number of edited words.

Conclusion of part 2:

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

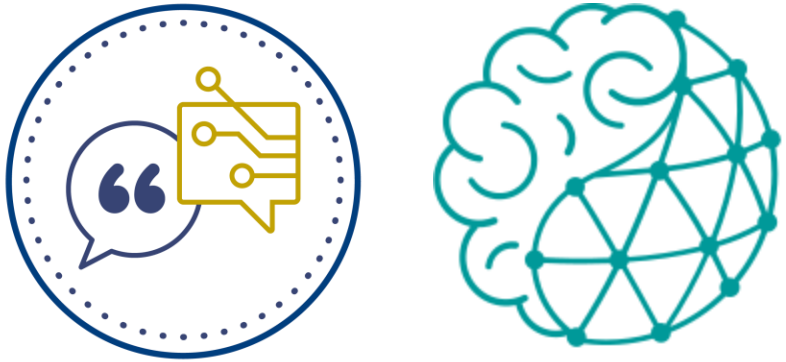
» TERcom is a very detailed tool to estimate edit rates
  - Unfortunately, Java and Perl only, but we are not aware of other tools that provide this level of detail (showing which word was edited and how)
» However, it does not capture the breadth and depth of complex editing
  - Note these examples are very simple an systematic; real editing is much messier
» There is a bias towards replacement and against movement
  - An analysis of primary and secondary actions helps us understand the "bias" in the estimates of editing

» In the next part of the tutorial, we will delve into the technical side of edit distances
» In the last part, we will discuss implications of the use of edit distances in downstream tasks

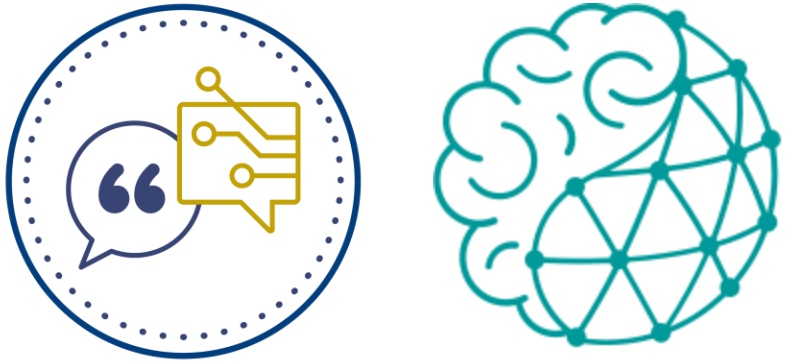
Any questions?

Break:

(20 minutes)



Part 3: Building a Computational Perspective



Please click the link to start

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Link to Google COLAB programming notebook

» https://colab.research.google.com/drive/1JSGLzUxK6GPrjJybSA2WTkn1rfkPzE37?usp=sharing

Note: both links lead to the same programming notebook.

Comparing Distances: Section One

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

Example	Description	Levenshtein	Damerau-Levenshtein	LCS	N-gram (n=2)	N-gram (n=4)	N-gram (n=5)
1	**Delete 1 word** Deleted: 'menu'	5	5	5	0.0515	0.0567	0.0598
3	**Replace 1 word** Substituted: 'Configurações' ➔ 'Configuração'	3	3	5	0.0309	0.0335	0.0351
8	**Delete 1 phrase (2 words)** Deleted: 'menu principal'	15	15	15	0.1598	0.1572	0.1567
9	**Delete 1 phrase (3 words)** Deleted: 'a partir do'	12	12	12	0.1237	0.1289	0.1320
10	**Insert 1 phrase (2 words)** Inserted: 'a opção'	8	8	8	0.0762	0.0786	0.0800
13	**Replace 1 phrase (3 words)** Substituted: 'menu principal' ➔ 'da primeira lista'	13	13	18	0.1340	0.1263	0.1216
20	**Move 1 phrase (3 words) 2 positions forward** Moved: 'configurações de tela' moved two positions forward	40	40	42	0.4124	0.4227	0.4268
28	**Delete 1 word + Move 1 word one position forward (different positions)** Deleted: 'pode' Moved: 'selecionando Configurações' ➔ 'Configurações selecionando'	29	29	31	0.2990	0.3041	0.3113
36	**Delete 1 phrase (2 words) + Replace 1 phrase (2 words)** Deleted: 'as configurações' Substituted: 'menu principal' ➔ 'primeiro conjunto'	31	31	38	0.3247	0.3273	0.3258
43	**Replace 1 phrase (2 words) + Insert 1 phrase (2 words)** Inserted: 'as opções' Substituted: 'menu principal' ➔ 'primeiro conjunto'	24	24	31	0.2227	0.2250	0.2236

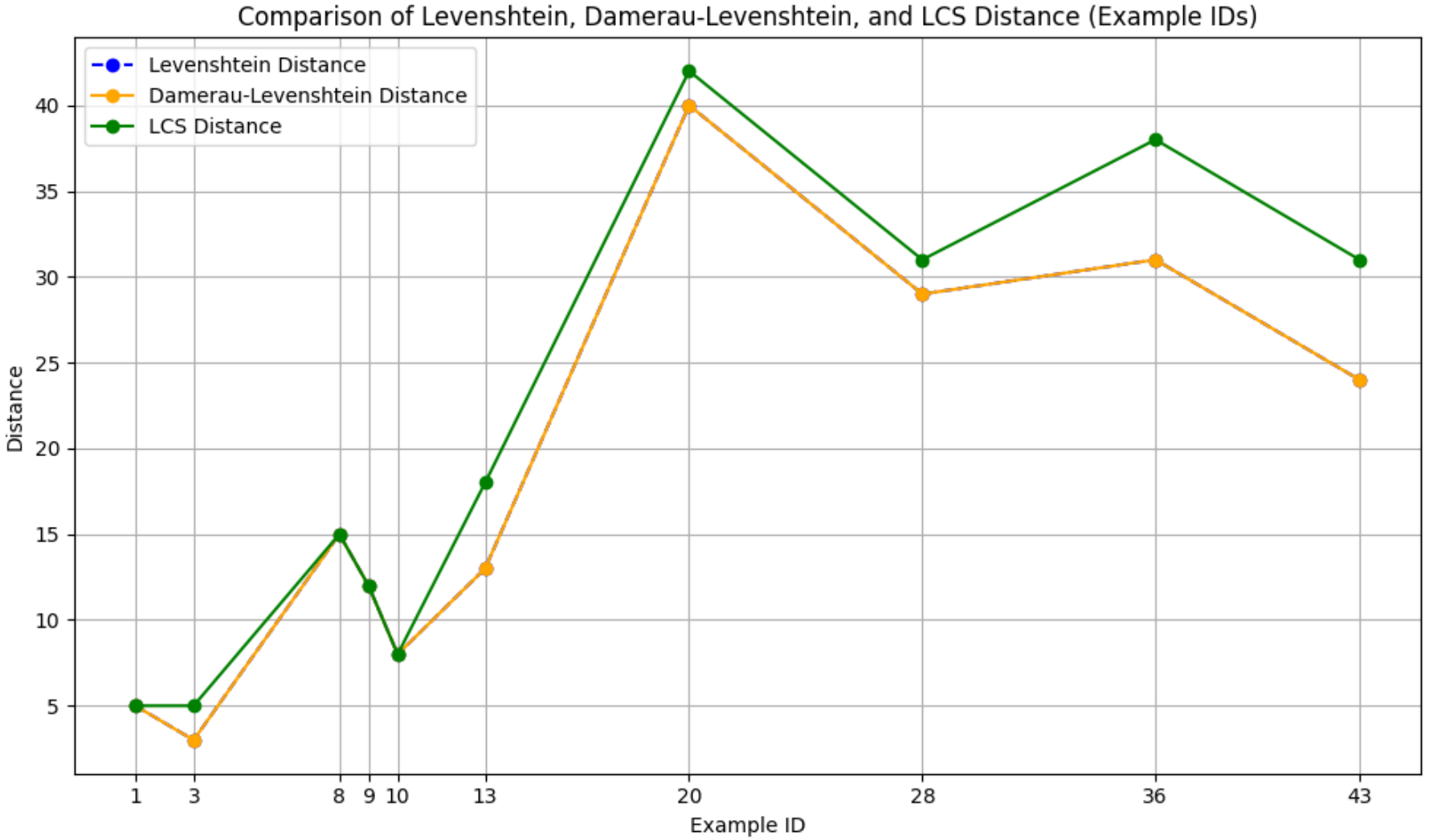
Comparing Distances-based metrics: Section Two

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

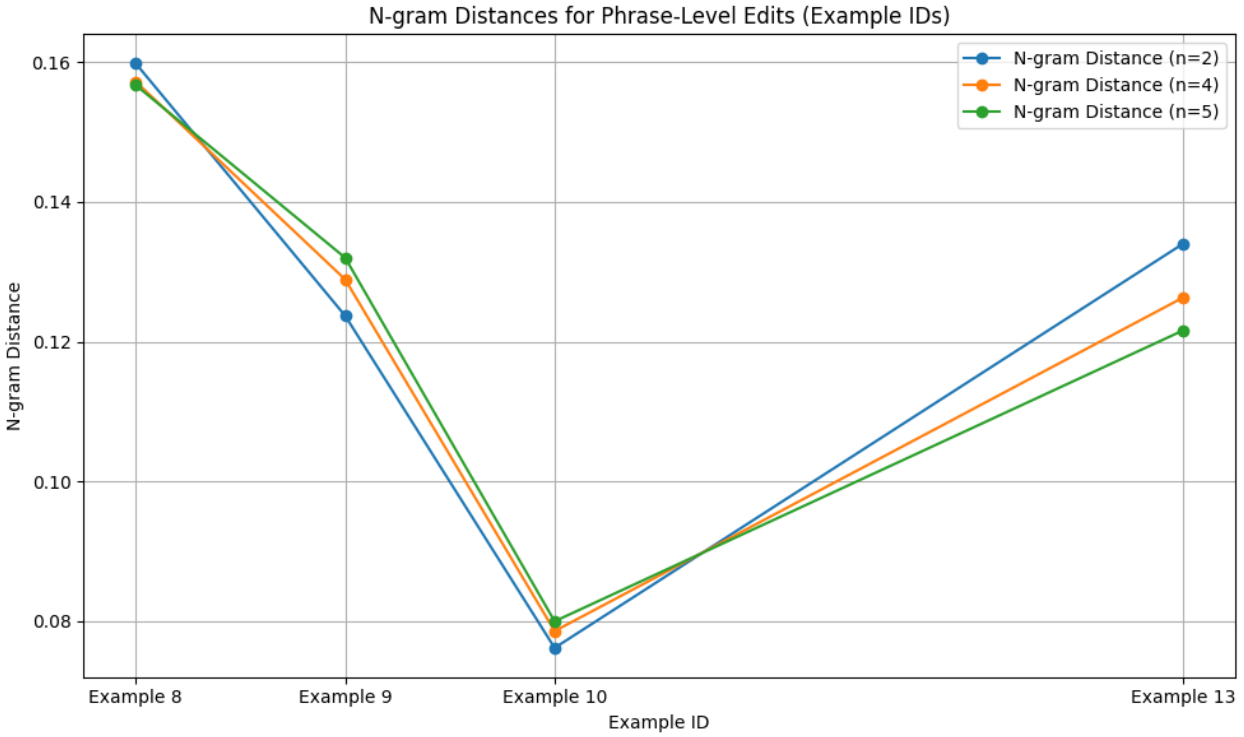
Example	Description	Insert	Delete	Substitute	Shifts	Total Edits	pyTER Score	SB TER	BLEU Score	chrF Score
1	**Delete 1 word** Deleted: 'menu'	1	0	0	0	1	0.0515	0.0714	0.8293	0.9289
3	**Replace 1 word** Substituted: 'Configurações' ➔ 'Configuração'	0	0	1	0	1	0.0309	0.0714	0.8003	0.9343
8	**Delete 1 phrase (2 words)** Deleted: 'menu principal'	2	0	1	0	3	0.1546	0.2143	0.7979	0.8575
9	**Delete 1 phrase (3 words)** Deleted: 'a partir do'	3	0	0	0	3	0.1237	0.2143	0.6499	0.8778
10	**Insert 1 phrase (2 words)** Inserted: 'a opção'	0	2	0	0	2	0.0825	0.1429	0.7625	0.9650
13	**Replace 1 phrase (3 words)** Substituted: 'menu principal' ➔ 'da primeira lista'	0	0	3	0	3	0.1031	0.2143	0.7166	0.8438
20	**Move 1 phrase (3 words) 2 positions forward** Moved: 'configurações de tela' moved two positions forward	0	0	0	1	1	0.0103	0.0714	0.6128	0.9169
28	**Delete 1 word + Move 1 word one position forward (different positions)** Deleted: 'pode' Moved: 'selecionando Configurações' ➔ 'Configurações selecionando'	1	0	0	1	2	0.1134	0.1429	0.6094	0.8367
36	**Delete 1 phrase (2 words) + Replace 1 phrase (2 words)** Deleted: 'as configurações' Substituted: 'menu principal' ➔ 'primeiro conjunto'	2	0	2	0	4	0.2990	0.2857	0.5078	0.6726
43	**Replace 1 phrase (2 words) + Insert 1 phrase (2 words)** Inserted: 'as opções' Substituted: 'menu principal' ➔ 'primeiro conjunto'	0	2	2	0	4	0.2268	0.2857	0.5749	0.8091

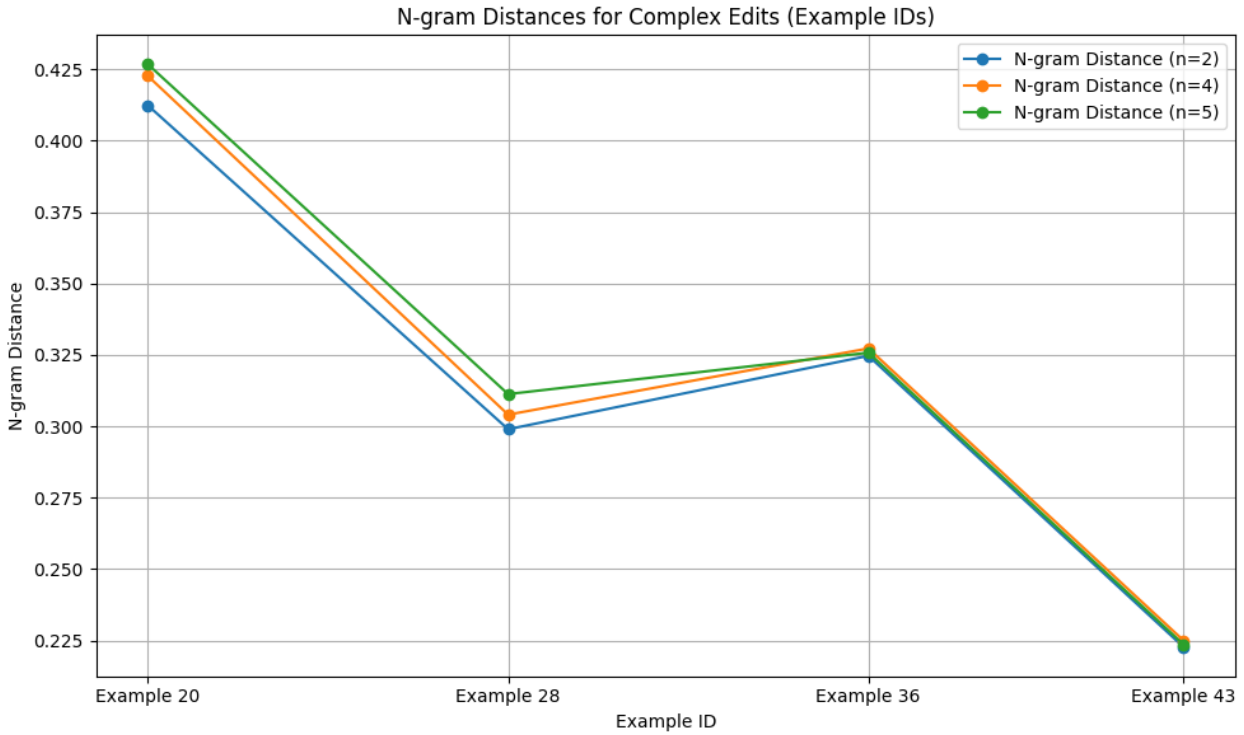
Comparing Distances: LD, DLD, LCS

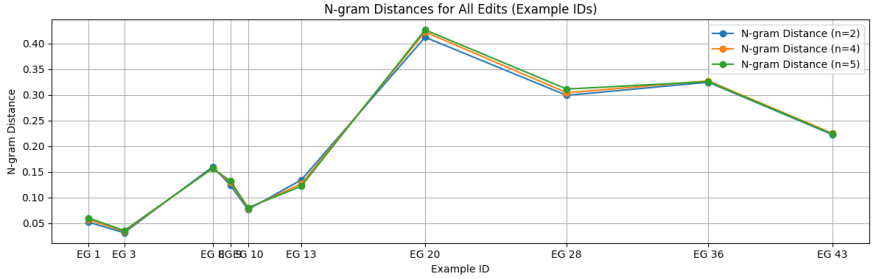


- Lower Distances for Simple Edits
- LCS tends to get higher values, particularly in examples with more complex edits (e.g., Example 20; shows disc. b/w LCS and others). LCS focuses on the order of sequences.
- LD and DLD provide comparable results when no transpositions are present (despite transposition in EG 20)
- Complex edits and phrase movements lead to a greater divergence between the metrics

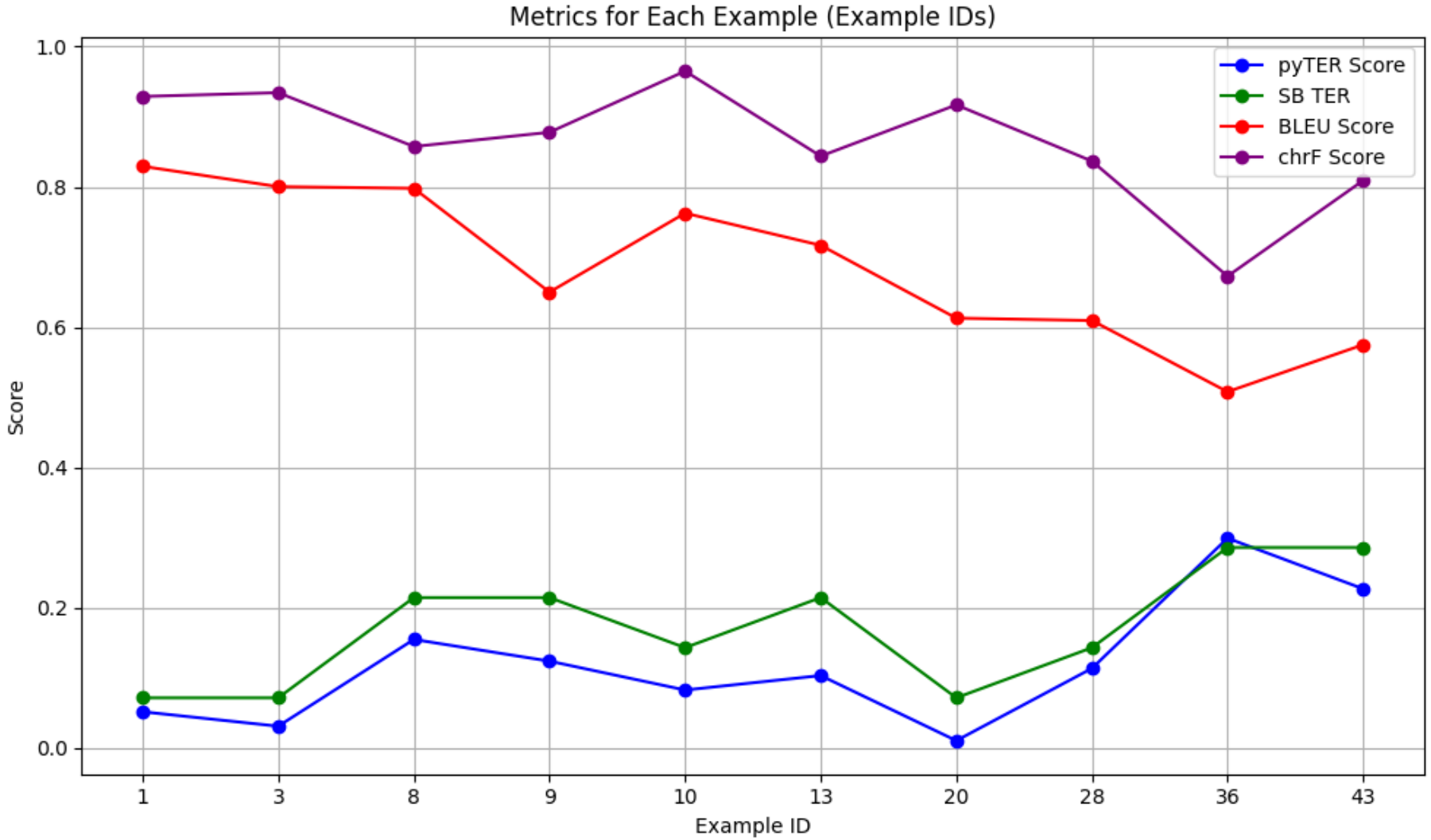
Comparing Distances: *n-gram* (2, 4, and 5)







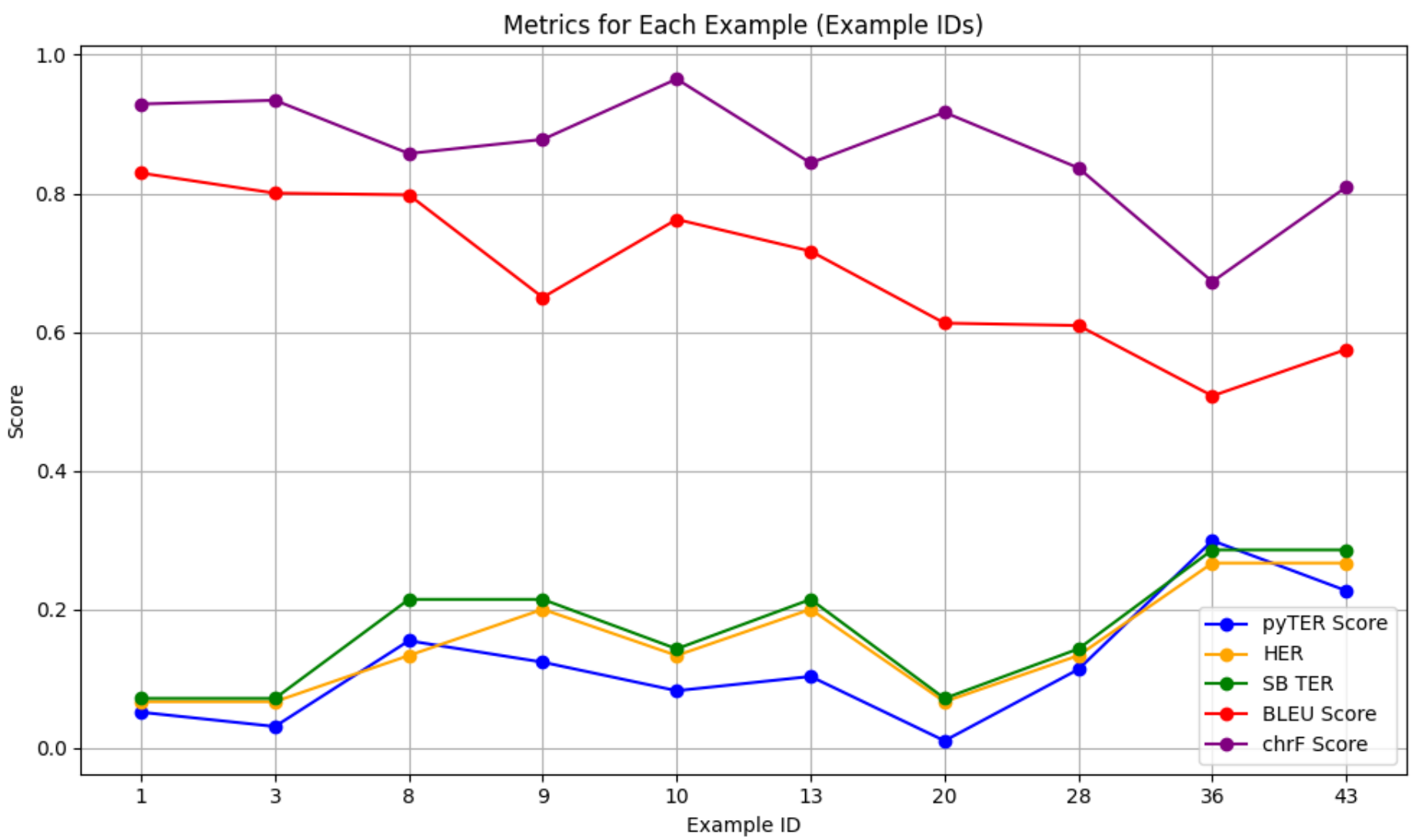
Comparing Statistical Metrics



Please interpret the graph exclusively, first, for chrF and BLEU.

Now, let us interpret the graph for TER scores.

Comparing Statistical Metrics: Adding HER

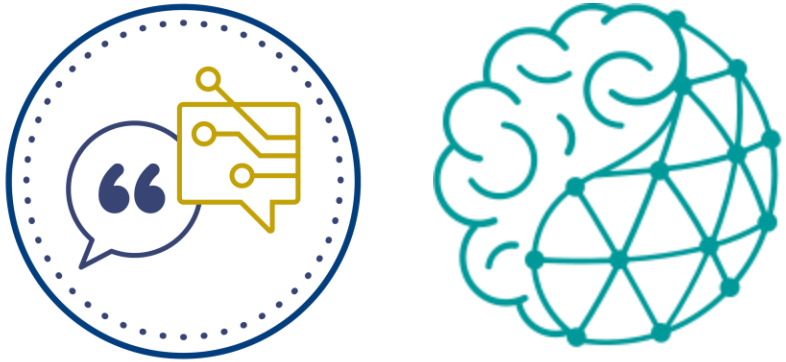


HER scores added to the same plot.

Any questions?

Part 4:

Implications for research and commercial applications of edit distances



Applications of Edit Distances

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Research
  - Machine Translation
  - Downstream tasks: Cognate Detection, Spell Checking, Automatic Post-Editing and Quality Estimation
  - Improving/expanding edit distances to increase their descriptive power

» Commercial/professional
  - Evaluate MT errors
  - Measure editing activity
  - Assess translators
  - Add new services

Post-MT: Statistical Evaluation & Challenges

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

Lack of Semantic Understanding	Metrics like BLEU, TER, and chrF focus on surface-level token matching, often failing to capture semantic meaning or sentence nuances.
Sensitivity to Synonyms and Paraphrases	Statistical metrics penalize variations in wording, even when the meaning is preserved, leading to artificially lower scores for valid translations.
Bias Towards Specific Sentence Lengths	Metrics like BLEU can disproportionately penalize translations that are shorter or longer than the reference, irrespective of the translation's quality.
Misalignment with Human Judgment	Human evaluators prioritize fluency, readability, and meaning, while statistical metrics focus on structural similarity, leading to discrepancies between automated scores and human assessment.
Overreliance on Word Overlap	Metrics like BLEU and TER rely heavily on n-gram matching, which does not account for syntactic flexibility or contextual appropriateness, limiting their ability to evaluate translations with high fluency but different word choices.

Post-MT: Predictive Evaluation

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

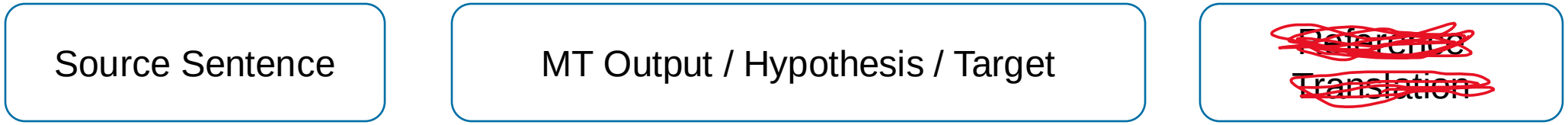
People-Centred AI
UNIVERSITY OF SURREY

» Quality Estimation for Machine Translation

Quality Estimation (QE) helps evaluate machine translation output without the need of a *reference translation.*

Statistical methods like BLEU, chrF, TER, need a reference translation to compare the output and provide a score between 0-100, ‘vaguely’ indicating the quality of translation.

However, at sentence-level, QE predicts the mean Direct Assessment (DA) score between 0-100, ‘vaguely’ indicating translation quality.



Why go training models if statistical methods exist? Human correlation, Bad references, Multiple possible references, Idiomatic Translation, …

Post-MT: Recent Efforts in Evaluation

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

Large (embeddings) Encoder-based

COMET22 – Ensemble between the Estimator model + multitasking to predict word-level tags.

COMET23 – Use of XLM-R-XXL (10.7B, 44GB) model with 12 models in the ensemble.

Sindhujan et. al. (2023) use **a single TransQuest with an InfoXLM-Large computational model** and,
achieve 2nd best in most Indic language pairs, and best score for English-Tamil sentence-level QE.

Large Language Model (LLM)-based

Large Language Models Are State-of-the-Art Evaluators of Translation Quality

Tom Kocmi and **Christian Federmann**
Microsoft, One Microsoft Way, Redmond, WA-98052, USA
{tomkocmi,chrife}@microsoft.com

**Predicting Perfect Quality Segments in MT Output with Fine-Tuned OpenAI LLM:
Is it possible to capture editing distance patterns from historical data?**

Serge Gladkoff [1], **Gleb Erofeev** [1], **Lifeng Han**[2], and **Goran Nenadic**[2]
[1] Logrus Global, Translation & Localization
[2] The University of Manchester, UK

Towards Making the Most of LLM for Translation Quality Estimation

Hui Huang, Shuangzhi Wu, Xinnian Liang, Bing Wang, Yanrui Shi, Peihao Wu, Muyun Yang & Tiejun Zhao

Post-MT: Automatic Correction

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Automatic Post-editing for Machine Translation

» Creation of computational APE model(s) which can identify and correct errors in the Machine Translation (MT) output using the gold-standard[1] and synthetic data, *while avoiding overcorrection*.

- Cope with systematic errors of an MT system when decoding is inaccessible
- Improve MT output by exploiting *information unavailable to the decoder*
- Provide professional translators with improved MT output quality to reduce (human) post-editing effort
- Adapt the output of a general-purpose MT system to the lexicon/style requested in a specific application domain.

- Bhattacharyya, Pushpak, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. "Findings of the WMT 2023 Shared Task on Automatic Post-Editing." In *Proceedings of the Eighth Conference on Machine Translation*, pp. 672-681. 2023.

- Bhattacharyya, Pushpak, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. "Findings of the WMT 2023 Shared Task on Automatic Post-Editing." In *Proceedings of the Eighth Conference on Machine Translation*, pp. 672-681. 2023.

Post-MT: Merging Evaluation and Correction

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Quality Estimation-informed Automatic Post Editing

- Quality Estimation (QE) can provide
  - *segment-level scores*, or
  - *word-level error annotation*
    - including *Multidimensional Quality Metrics (MQM)-based severity level of the error, or*
  - *document-level estimation of quality*

» *However, this analysis of output is limited to evaluation of hypothesis only*

- Automatic Post-editing (APE) can help with
  - *Automatic identification,*
  - *and correction of the errors in output.*

» *However, it's output is limited to a corrected version of the hypothesis*

» *Prone to overcorrection*

» APE is *evidently* prone to overcorrection which means it correct parts of hypothesis which do not need any correction. The minimum-edits rule is not enforced during the computational process.

» **QE-informed APE** (Deoghare et al., 2023) emerges as a viable direction to make more informed automatic edits to the hypothesis.

- Subtask 3 within the QE shared task at WMT 2024.

‘Edit Distance’ Applications: MT Research

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Translation workflow

» Quality Estimation to improve MT

- QE can be use within translation workflow to improve machine translation (Specia et al., 2018)
- *HTER (Human-Targeted TER)*
  Measures the edit distance between machine translation (MT) output and a reference.
- Applied at the segment level, HTER can be predicted using computational QE models.

» Word-Level QE

- *Local Context Similarity*
  Word-level QE predicts which words in the MT output *need editing*.

» Document-Level QE

- *Overall Quality Assessment*
  Document-level QE aggregates segment-level quality estimates (e.g., TER) across the entire document.

» Post-translation workflow

» Post-Editing

The process of editing MT output to improve its quality, likely ***manually***. Edit distance metrics help measure the difference between the raw MT and the post-edited text.

» Automatic Post-Editing (APE)

APE systems can leverage edit distance metrics like Levenshtein or Damerau-Levenshtein to identify and correct errors in MT output, such as typographical errors or word transpositions.

Use of TER for APE evaluation is quite standard for computational research.

Other NLP tasks

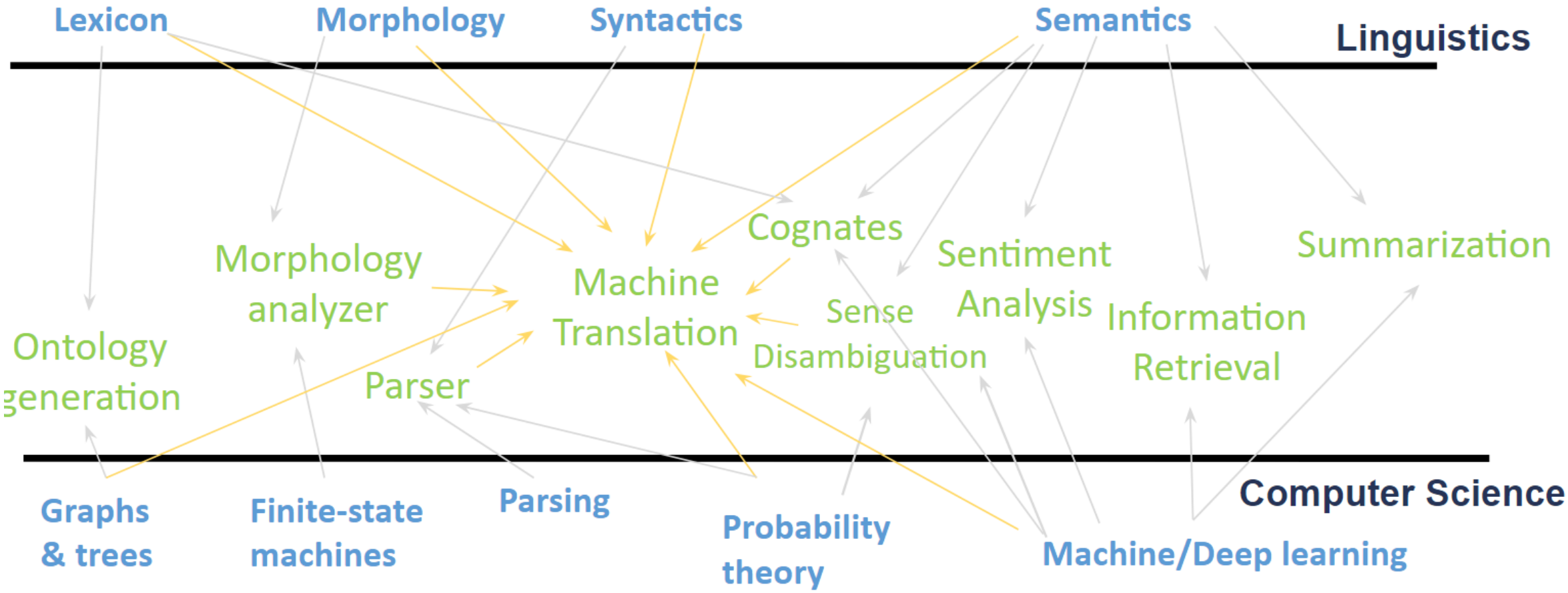


Image credits: Prof Pushpak Bhattacharyya

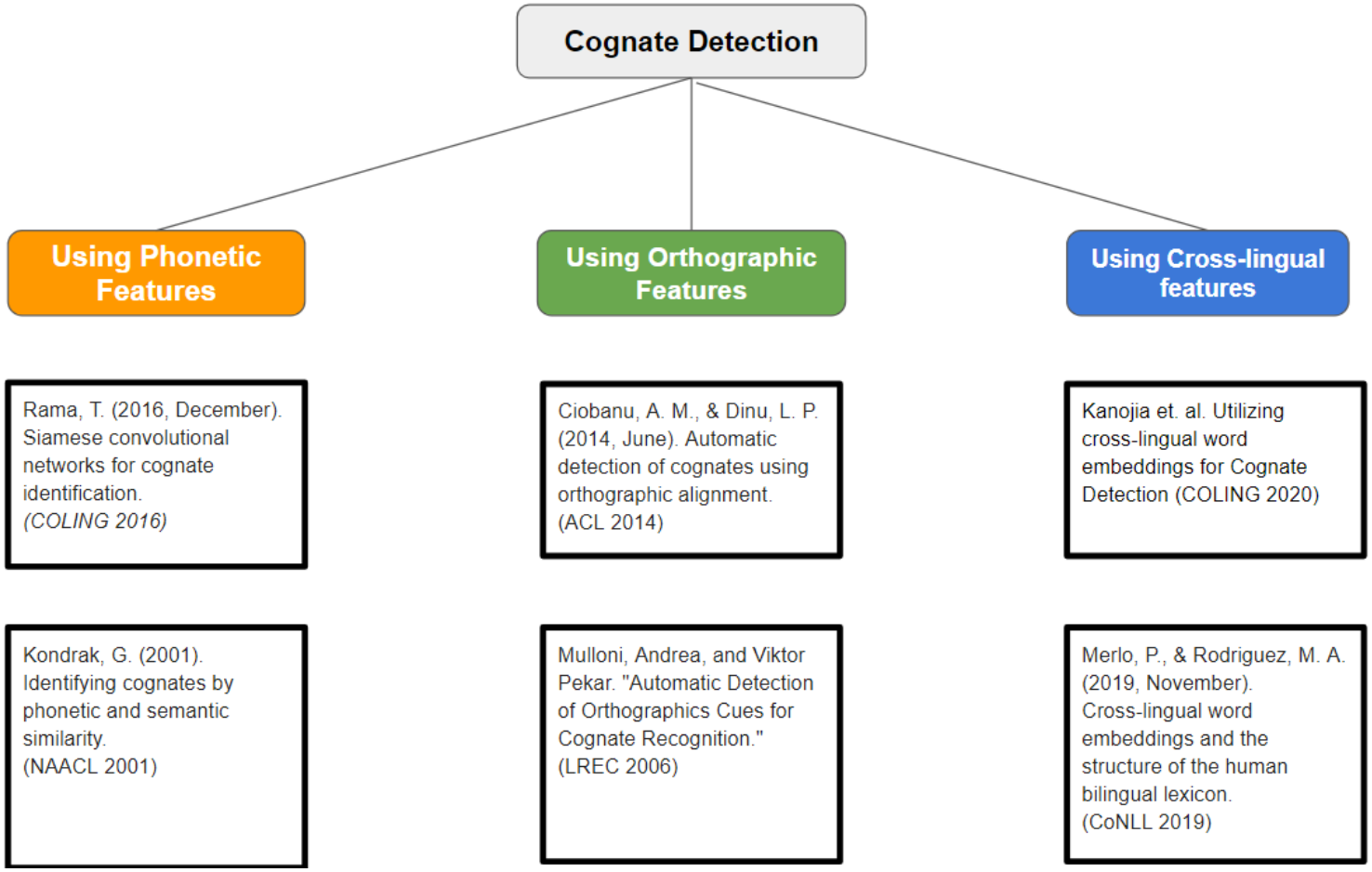
Other NLP tasks: Detection of Shared Vocabulary

Cognates represent a *large chunk of the shared vocabulary* among language pairs.

Previously, the task of Cognate Detection has shown to help the downstream tasks of Machine Translation via word alignment (Kondrak, 2005)

Cognitive Psycholinguistics-based features have also shown to improve various NLP tasks (Mishra et. al., 2016) (features from gaze/eye-tracking)

- Including the Cognate Detection task (Kanojia et al., 2021) (for Hi-Mr)



Researching editing

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Many examples of research in Translation Studies uses edit distances (mainly TER/HTER)
  - Evaluation of quality (MT and human)
  - Applied as error typologies (MQM etc)

» In reality, editing is not neat. It can even be considered “inefficient”.
  - It does not follow the “minimum distance”
  - Edits can be incomplete, repetitive, there are overlaps and backtracks
    - You may first move a word, and then change it (replace it)
    - You may replace several words at the same time in several sentences, and then make small local adjustments to that word and the following ones
  - Why? Because of the mental processes of translation, back and forth between two languages
  - Mental processes are aim for other things rather than the “shortest path”, like “accuracy” or “quality”

» Measuring the “shortest path” does not help learn the process

Research: Improve/expand edit distances

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Explore technical costs/weights associated with the process
- Characters, words, phrases
- Primary vs secondary operations
- Frequent edits
- Time-consuming edits
- Semantic considerations (e.g. paraphrases)

» Research linguistic knowledge
- Not just semantic substitution
- Semantic deletions, movements, etc.
- Syntactic changes after first edit

» Collaborative work with Translation Process Research teams
- Research the mental editing process
  - For how many hours can you read and edit with no loss of attention?
  - What do you do first? Delete superfluous words?
- Combine with user activity logs
  - Or even replace by proper user activity logs (the actual actions)

Applications in commercial settings

» Acquiring knowledge on the process
- This extends data retrieved as Translation Memories (product data) with process data
- *! Ethical questions here*

» Measuring effort
- Pressure on productivity
- Cognitive effort still not accounted for

» Influence on rates
- Words edited (plus time?)
- Still, a reduction of current rates

» Changes to workflows and role of humans in workflows with increasing automation
- Automated evaluation
- Integration of APE and QE
- All validated by edit distances, under the assumption of their reliability

Research & commercial aplications: ethics

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

» Ethical concerns:
- Capturing specialised know-how
- Reducing the value of human work
- Generated knowledge is usually not returned to the generators of the initial value (translators/post-editors)

» Research on process automation
- Reflect on professional consequences of the use of these instruments in distributed workplaces

» Involve as many stakeholders as you can in your research
- Learn from them
- Explore how to return the acquired knowledge to the creators of the knowledge

Conclusions of the tutorial

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

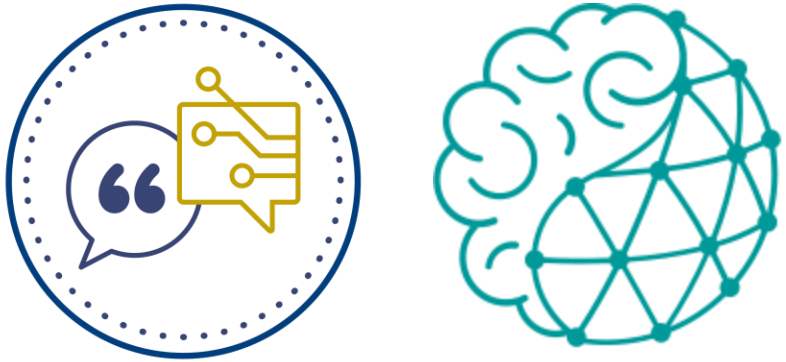
People-Centred AI
UNIVERSITY OF SURREY

» Edit distances are metrics
  - They are very useful to measure comparable things
  - Make sure you use them consistently
» Edit distances are statistical
  - They say a lot about patterns of behaviour
  - But we do not know how much they sat about linguistic and process dimensions

» Research and teaching
  - Detailed understanding of edit distances as research instruments
  - Detailed understanding of language data and translation/editing processes
» Use in downstream tasks
  - May be blunt instruments
  - *Ethical concerns*

Part 5:

Discussion



References

Deoghare, S., Kanojia, D., Blain, F., Ranasinghe, T. and Bhattacharyya, P., 2023, December. Quality Estimation-Assisted Automatic Post-Editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1686-1698).

Specia, L., Scarton, C. and Paetzold, G.H., 2022. *Quality estimation for machine translation*. Springer Nature.

do Carmo, F., Shterionov, D., Moorkens, J., Wagner, J., Hossari, M., Paquin, E., Schmidtke, D., Groves, D. and Way, A., 2021. A review of the state-of-the-art in automatic post-editing. *Machine Translation*, *35*, pp.101-143.

do Carmo, F., 2021. Editing actions: a missing link between translation process research and machine translation research. *Explorations in empirical translation process research*, pp.3-38.

Kanojia, D., Sharma, P., Ghodekar, S., Bhattacharyya, P., Haffari, G. and Kulkarni, M., 2021. Cognition-aware cognate detection. *arXiv preprint arXiv:2112.08087*.

Specia, L. and Shah, K., 2018. Machine translation quality estimation: Applications and future perspectives. *Translation quality assessment: from principles to practice*, pp.201-235.

Popović, M., 2018. Error classification and analysis for machine translation quality assessment. *Translation quality assessment: From principles to practice*, pp.129-158.

Mishra, A., Kanojia, D. and Bhattacharyya, P., 2016, March. Predicting readers' sarcasm understandability by modeling gaze behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1).

Romero-Fresco, P. and Pérez, J.M., 2015. Accuracy rate in live subtitling: The NER model. *Audiovisual translation in a global context: Mapping an ever-changing landscape*, pp.28-50.

Burchardt, A., 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*.

References

Snover, M.G., Madnani, N., Dorr, B. and Schwartz, R., 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, *23*, pp.117-127.

Snover, M., Madnani, N., Dorr, B. and Schwartz, R., 2009, March. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the fourth workshop on statistical machine translation* (pp. 259-268).

Shapira, D. and Storer, J.A., 2007. Edit distance with move operations. *Journal of discrete algorithms*, *5*(2), pp.380-392.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J., 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* (pp. 223-231).

Kondrak, G., 2005. Cognates and word alignment in bitexts. In *Proceedings of Machine Translation Summit X: Papers* (pp. 305-312).

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A. and Sawaf, H., 1997, September. Accelerated DP based search for statistical translation. In *Eurospeech* (pp. 2667-2670).

Levenshtein, V., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

Damerau, F.J., 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), pp.171-176.

python-string-similarity: https://github.com/luozhouyang/python-string-similarity#readme

SacreBLEU: https://github.com/mjpost/sacrebleu

PyTER (Python 3): https://pypi.org/project/pyter3/

TERCOM: https://github.com/jhclark/tercom

Thank you for your attention.

CENTRE FOR TRANSLATION STUDIES

UNIVERSITY OF SURREY

For any queries, feel free to contact us:

f.docarmo@surrey.ac.uk

and

d.kanojia@surrey.ac.uk

All materials will be available at:

https://github.com/surrey-nlp/AMTA-EditDistances-tutorial