FONDAZIONE BRUNO KESSLER

Quality Estimation in support of Automatic Post-Editing

Marco Turchi
Fondazione Bruno Kessler, Trento, Italy
turchi@fbk.eu

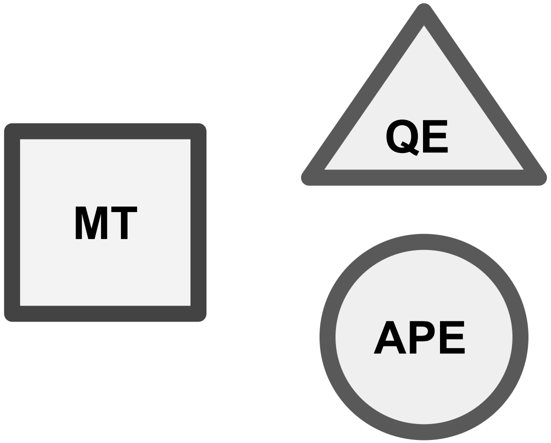
In collaboration with Amirhossein Tebbifakhr and Matteo Negri

Outline

- Motivation
- Previous Work
- Effort-aware APE
- Conclusion

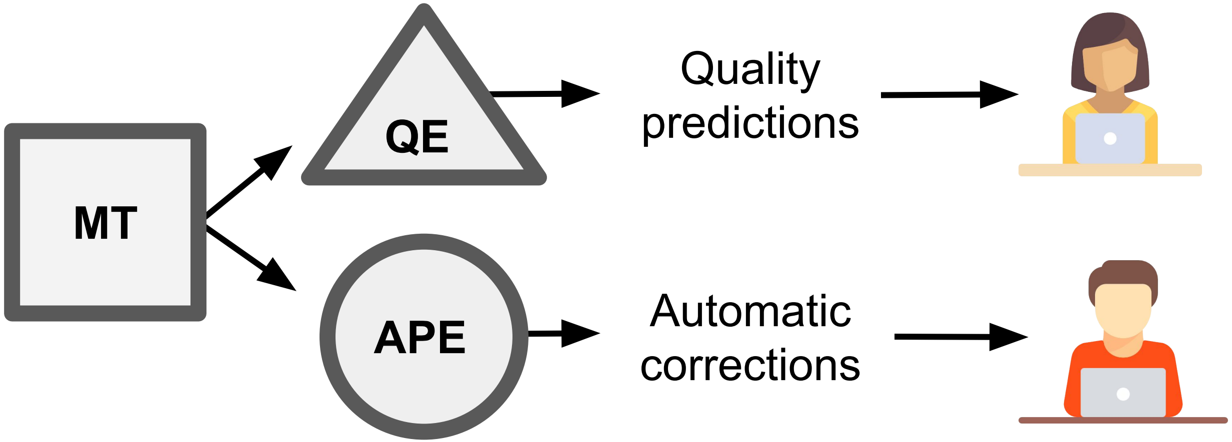
Motivation

- QE and APE: two ancillary MT tasks...



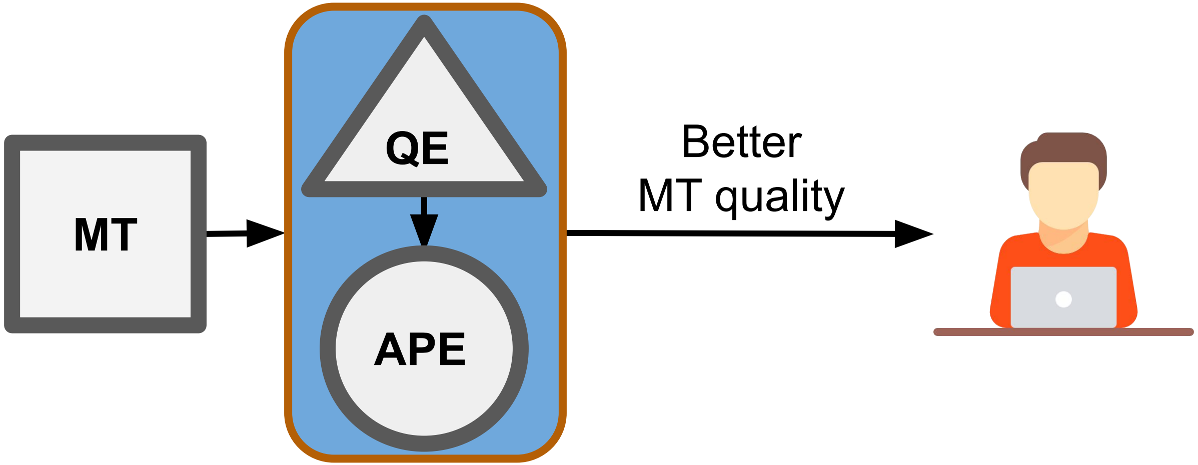
Motivation

- QE and APE: two ancillary MT tasks...
- ...mostly explored separately



Motivation

- QE and APE: two ancillary MT tasks...
- ...mostly explored separately
- Can we combine them to get better translations?



Quality Estimation (QE)

A supervised learning task:

- Predict MT quality at run-time (without references)
- Learn from *(src, mt, quality_label)* triplets
- Assign *quality_label* to *(src, mt)* test pairs
    - Granularity: word, phrase, sentence, document
    - Label: Post-editing time/effort, binary/Likert scores, ranking
    - Approaches: regression, classification, ranking

Automatic Post-editing (APE)

A “monolingual translation” task:

- Correct MT errors
- Learn from *(src, mt, post-edited MT)* triplets
- Produce *post-edited MT* given *(src, mt)* test pairs
  - Approaches: phrase-based MT, neural MT

Issues in APE

SRC: *Ape decoding is not always perfect*

MT*: La decodifica Ape non è sempre perfetta*

- **Wrong corrections**
  - APE*: La decodifica delle scimmie non è sempre perfetta*

Issues in APE

SRC: *Ape decoding is not always perfect*

MT*: La decodifica Ape non è sempre perfetta*

- **Wrong corrections**
  - APE*: La decodifica delle scimmie non è sempre perfetta*
- **Unnecessary corrections**
  - APE: *Non sempre la decodifica Ape è priva di errori*

Issues in APE

Automatic evaluation metrics penalize both!

- **Wrong corrections**
  - APE: *La decodifica delle scimmie non è sempre perfetta*
- **Unnecessary corrections**
  - APE: *Non sempre la decodifica Ape è priva di errori*

Issues in APE

- Ideal scenario:
  - Limiting wrong and unnecessary edits
    - In particular, when the *mt* is perfect
  - Fixing all the errors
    - Improving the number of corrected sentences

Outline

- Motivation
- Previous Work
- Effort-aware APE
- Conclusion

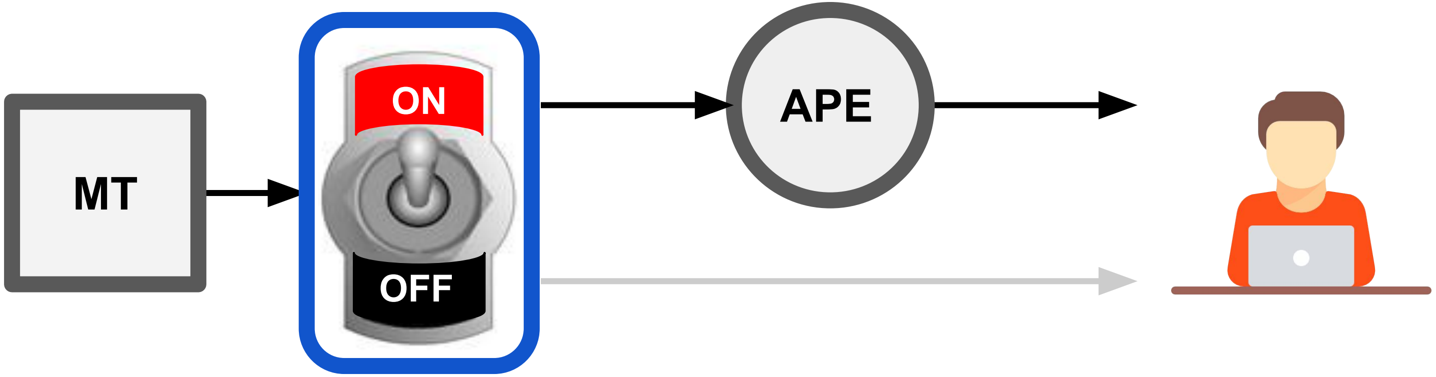
Combining QE & APE

Three strategies

- QE as activator: suggests whether to run APE or not
- QE as guidance: informs APE decoding
- QE as selector: chooses between MT and APE

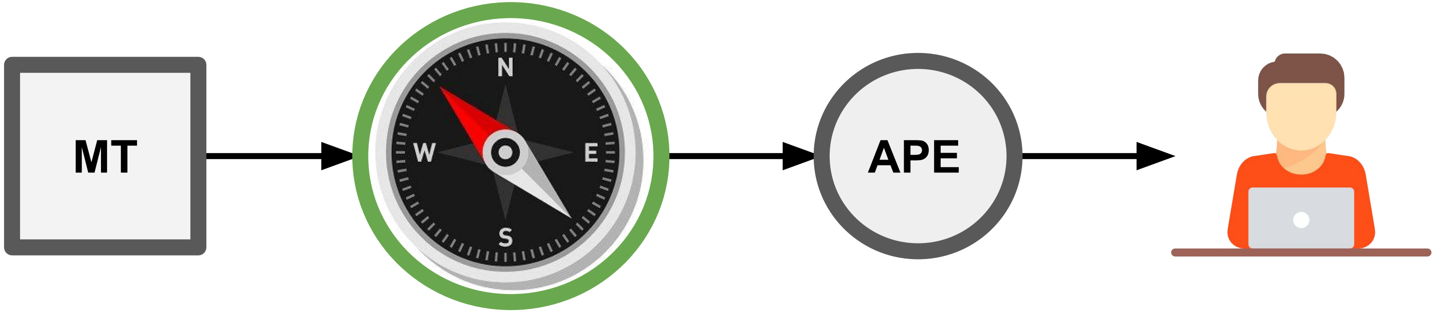
QE as activator

Triggers APE when QE score is below a threshold



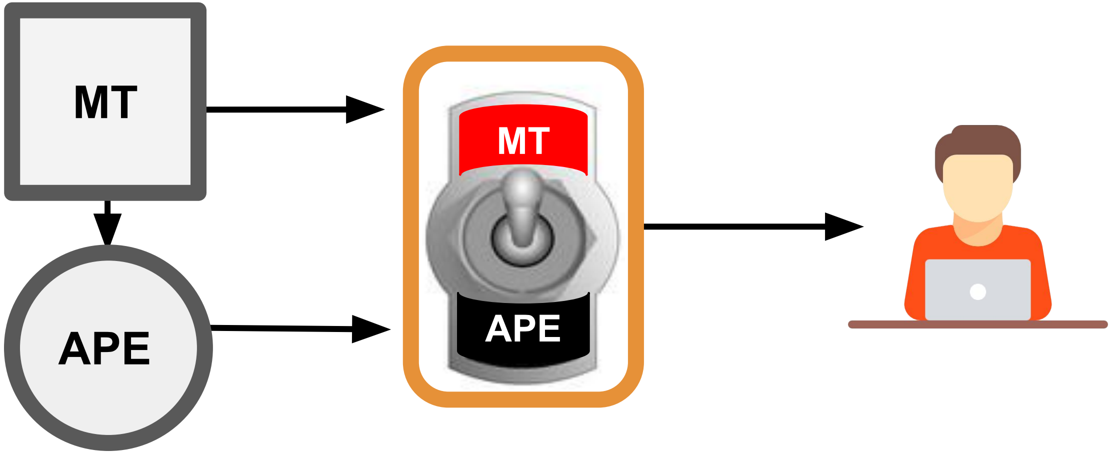
QE as guidance

Indicates which MT tokens have to be kept/changed



QE as selector

Chooses between raw MT and APE output



Experiments: data

- English-German
  - WMT`16 QE/APE data set
  - Domain: information technology
  - *(src, mt, post-edited MT)* triplets
    - *mt*: phrase-based system
    - *post-edited MT*: professional translators
  - Training: 12K, Dev: 1K, Test: 2K

Experiments: QE systems

- Best QE systems at WMT`16
  - Sentence-level [Kozlova et al., 2016]
    - Used for QE as activator
  - Word-level: [Martins et al., 2016] *
    - Used for QE as guidance, selector
- ORACLE labels: released by QE task organizers

* Thanks to Unbabel for providing us with the QE word level predictions

Experiments: APE systems

- Best APE submissions at WMT`16
  - Phrase-based: [Chatterjee et al., 2016]
  - Neural: [Junczys-Dowmunt and Grundkiewicz, 2016]
    - Used for QE as activator, selector
- *Ad-hoc* system
  - Neural "guided decoder" [Chatterjee et al. 2017]
    - Used for QE as guidance

QE as **activator**

Triggers APE...

...if the predicted MT quality...

...is below a threshold

QE as activator

Triggers APE...

- Phrase-based/Neural

...if the predicted MT quality...

...is below a threshold

QE as activator

Triggers APE...

- Phrase-based/Neural

...if the predicted MT quality...

- Sentence-level

...is below a threshold

QE as activator

Triggers APE...

- Phrase-based/Neural

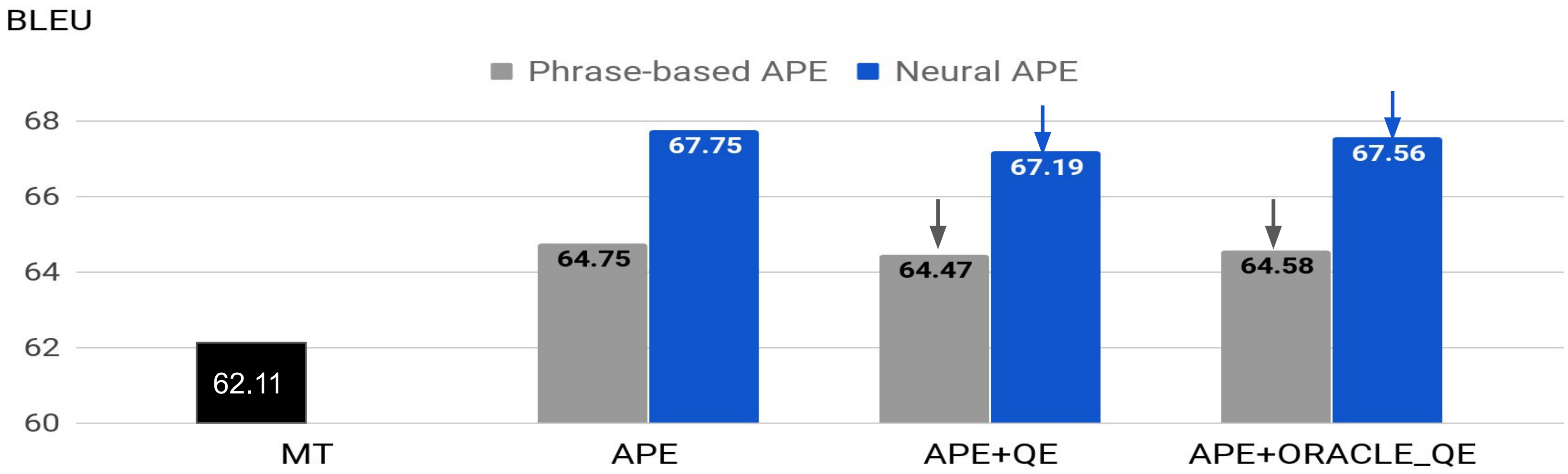
...if the predicted MT quality...

- Sentence-level

...is below a threshold

- Estimated on dev data (TER=10)

QE as activator results



Performance drop wrt APE without QE

- Sentence-level QE too coarse-grained?

QE as guidance

Informs APE…

...with quality labels…

...about MT tokens to be kept/changed

QE as guidance

Informs APE…

- Phrase-based/Neural

...with quality labels…

...about MT tokens to be kept/changed

QE as guidance

Informs APE…

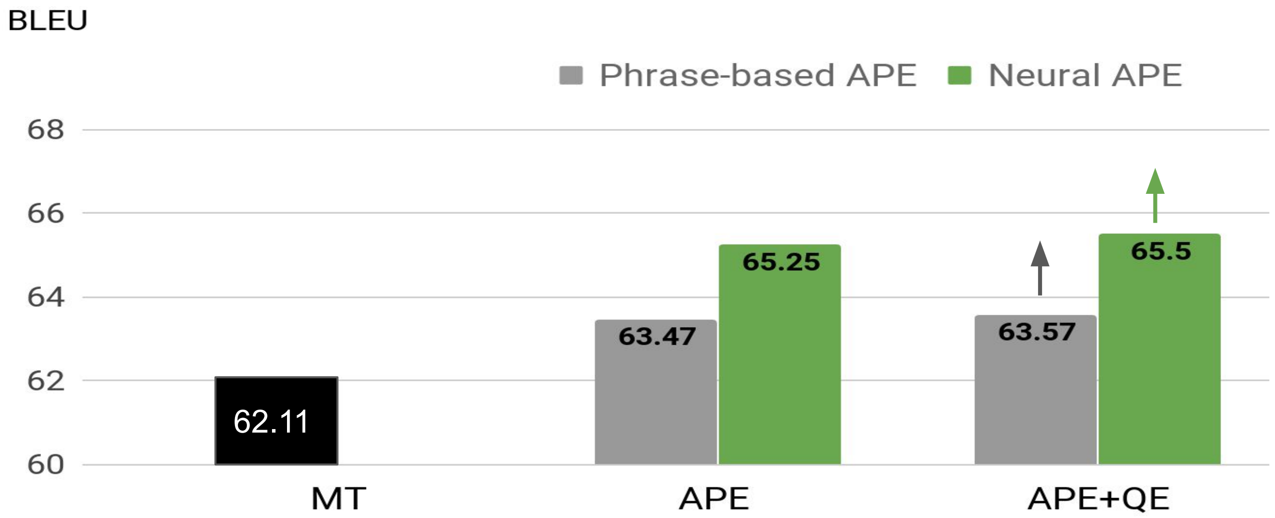
- Phrase-based/Neural

...with quality labels…

- Word-level (“good”/”bad”)

...about MT tokens to be kept/changed

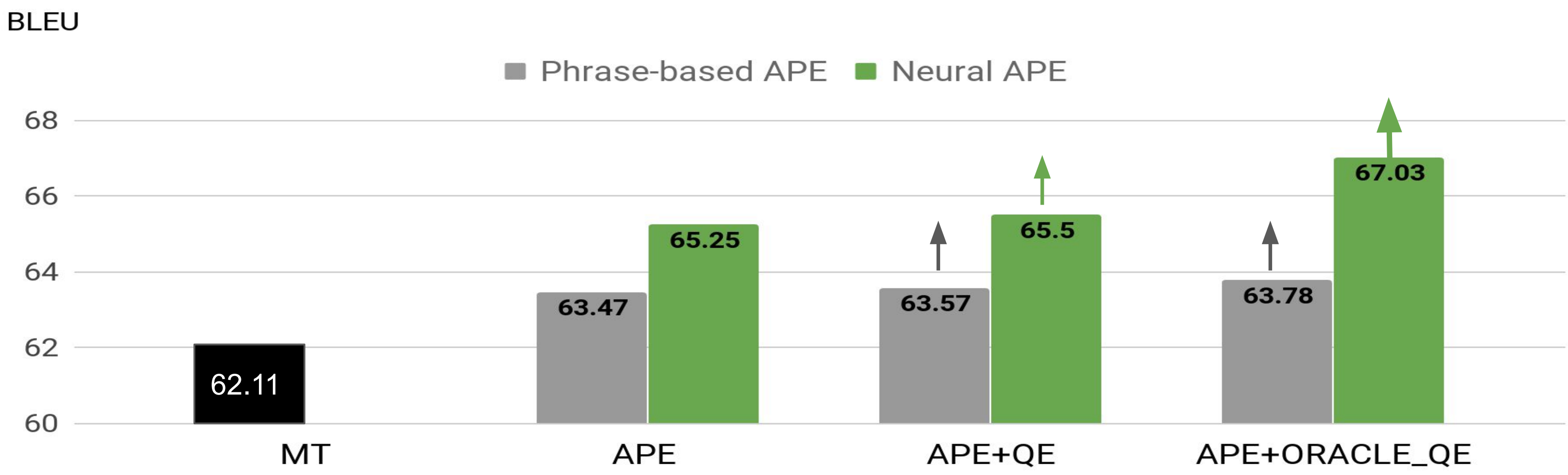
QE as guidance results



Small gain wrt APE without QE

- Larger for neural APE (+0.25 BLEU)

QE as guidance results



Small gain wrt APE without QE

- Larger for neural APE (+0.25 BLEU)
- <u>Room for improvement with better predictions (+1.78 wrt NAPE)</u>

QE as selector

Selects APE…

...if the predicted quality…

...is better than MT

QE as selector

Selects APE…

- Phrase-based/Neural

...if the predicted quality…

...is better than MT

QE as selector

Selects APE…

- Phrase-based/Neural

...if the predicted quality…

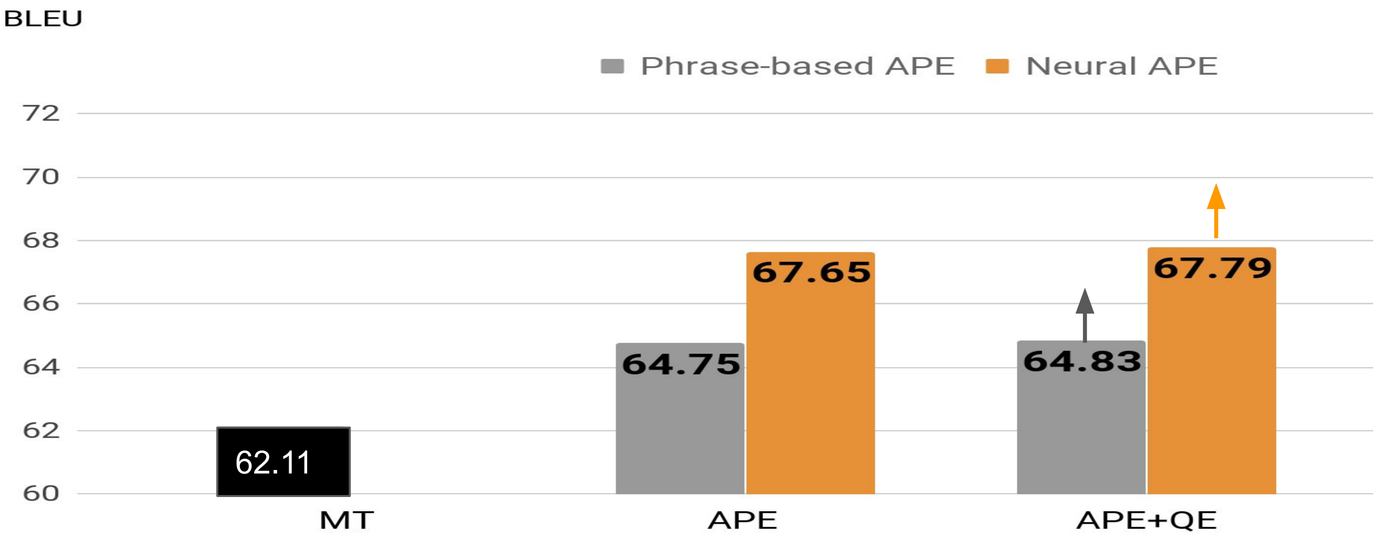
- Word-level

...is better than MT

QE as selector (word-level)

- **Word-level** QE
  - Annotate both MT and APE
  - Replace MT tokens if MT="bad" and APE="good"

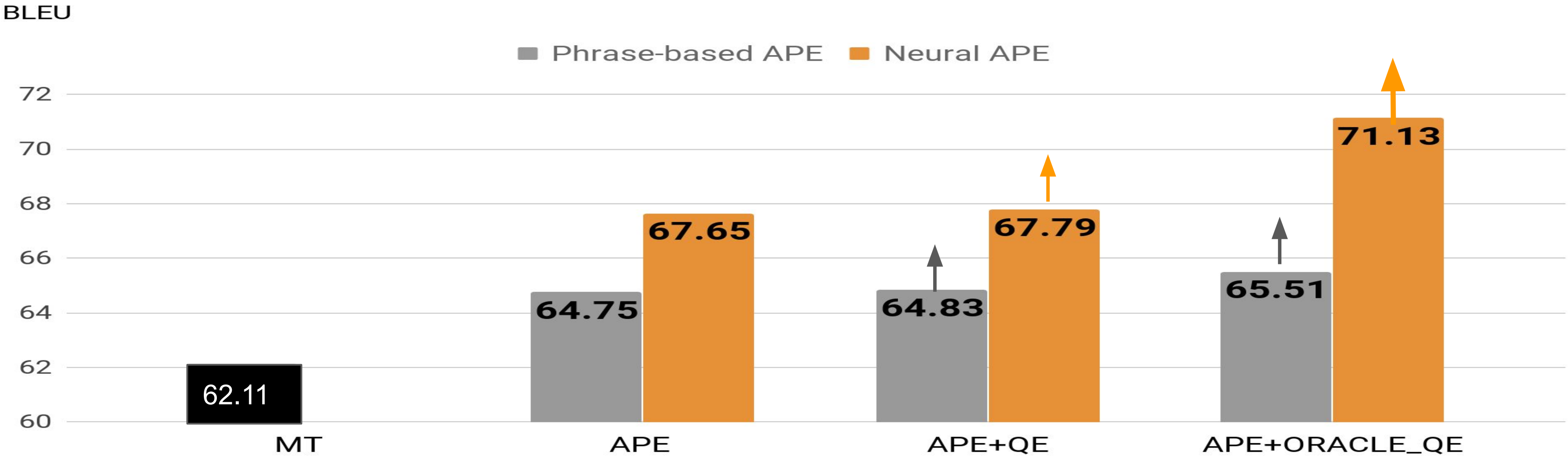
QE as selector (word-level) results



Small gain, both for phrase-based and neural APE

- Larger for neural APE

QE as selector (word-level) results



Small gain, both for phrase-based and neural APE

- Larger for neural APE
- Room for improvement with better predictions (+3.34 wrt NAPE)

Quick Summary

- Pro:
  - QE seems to able to support APE

- Cons:
  - Need of Oracle QE to see large gains
  - APE not aware of QE information
  - All results on top of a phrase based MT system

Outline

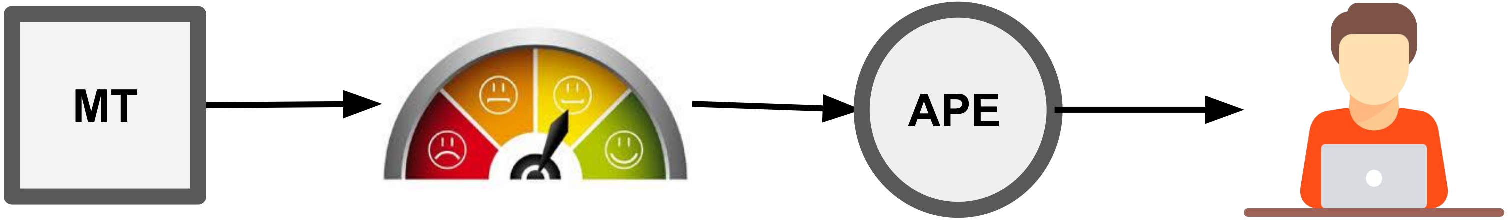
- Motivation
- Previous Work
- Effort-aware APE
- Conclusion

Effort-Aware APE

- QE as activator + QE as guidance
- QE as effort indicator:

Effort-Aware APE

- QE as activator + QE as guidance
- QE as effort indicator:



Effort-Aware APE

- QE as effort indicator:
    - Informs the APE about the effort needed to fix the errors
    - Prepends an effort tag in front of *src* and *mt*

Effort-Aware APE

- QE as effort indicator:
  - Informs the APE about the effort needed to fix the errors
  - Prepends an effort tag in front of *src* and *mt*

SRC: *Ape decoding is not always perfect*

MT*: La decodifica Ape non è sempre perfetta*

Effort-Aware APE

- QE as effort indicator:
  - Informs the APE about the effort needed to fix the errors
  - Prepends an effort tag in front of *src* and *mt*

SRC: *<no_postedits> Ape decoding is not always perfect*

MT*: <no_postedits> La decodifica Ape non è sempre perfetta*

Effort Token

- No Post-edit
- Light Post-edit
- Heavy Post-edit

Effort-Aware APE

- QE as effort indicator vs QE as activator
  - Diff: Always routes sentences to APE

Effort-Aware APE

- QE as effort indicator vs QE as activator
  - Diff: Always routes sentences to APE
- QE as effort indicator vs QE as guidance
  - Diff: APE aware of QE info

Experiments: data

- WMT`19 QE/APE data set
- Neural MT outputs

- English-German
  - Training: 13K, Dev: 1K, Test: 1K
- English Russian
  - Training: 15K, Dev: 1K, Test: 1K

Experiments: QE systems

- At training time
    - Effort token obtained by arbitrary thresholding the TER
        - No Post-edit (TER = 0)
        - Light Post-edit (0< TER < 40)
        - Heavy Post-edit (TER >= 40)

Experiments: QE systems

- A test time
  - There is not the *pe* to compute the TER
  - Predicting the effort token

Experiments: QE systems

- How to compute the effort token
  - BERT:
    - Building a classifier that predicts the 3 tags

Experiments: QE systems

- How to compute the effort token
  - BERT:
    - Building a classifier that predicts the 3 tags
  - Nearest neighbour:
    - Using the label of the most similar <*src*, *mt, pe*> triplet in the training data

Experiments: APE systems

- Neural *FBK* system
  - Multi-source APE
  - Dual Transformer
  - Ad-hoc pre-processing of the German data
  - Training on artificial data
  - Fine-tuning on in-domain data

QE as effort indicator

Informs APE…

- Neural

...with quality labels…

- Effort token (“No”/“Light”/“Heavy”)

...about the effort to correct the MT

Token Prediction Performance

- Tokens distribution:

	En-De	En-Ru
NO	281	621
Light	615	219
Heavy	104	160

Token Prediction Performance

- Tokens distribution:

	En-De	En-Ru
NO	281	621
Light	615	219
Heavy	104	160

- Prediction Performance:

Accuracy	**En-De**	**En-Ru**
BERT	52	51
N-N	65	64

Token Prediction Performance

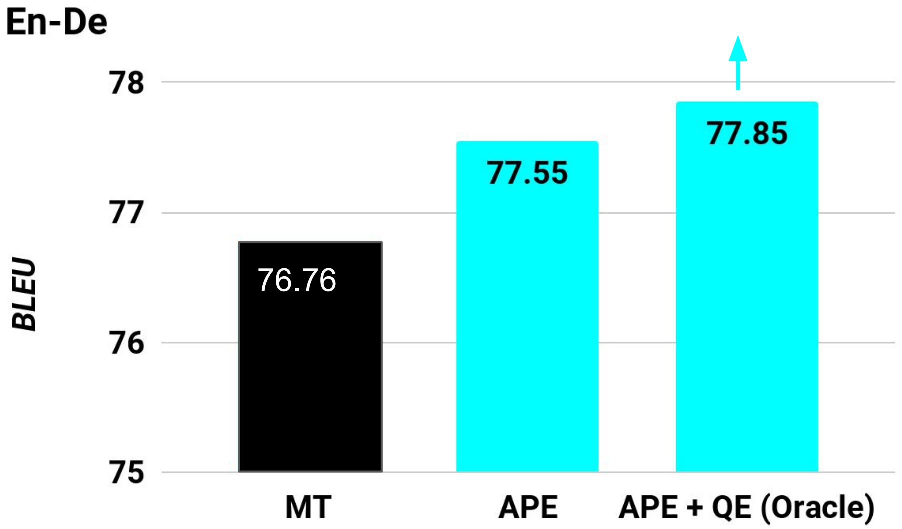
- Tokens distribution:

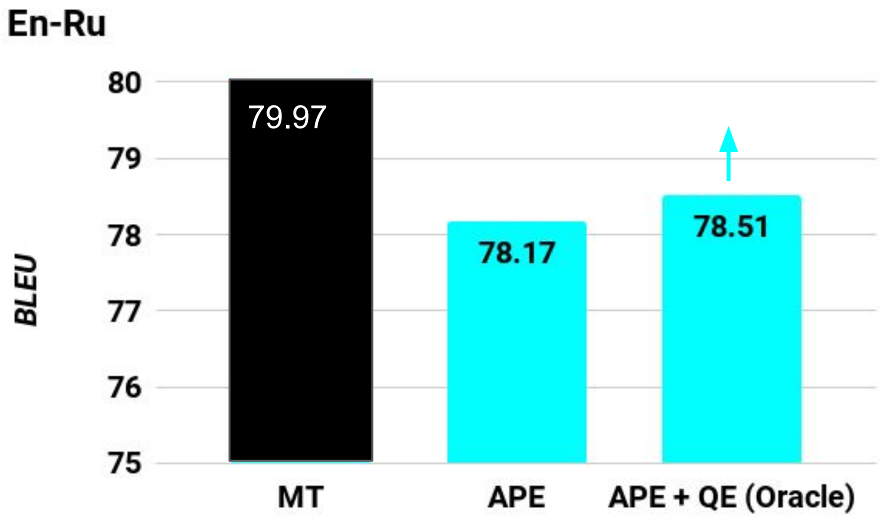
	En-De	En-Ru
NO	281	621
Light	615	219
Heavy	104	160

- Prediction Performance:

Accuracy	**En-De**	**En-Ru**
BERT	52	51
N-N	65	64

QE as effort indicator results

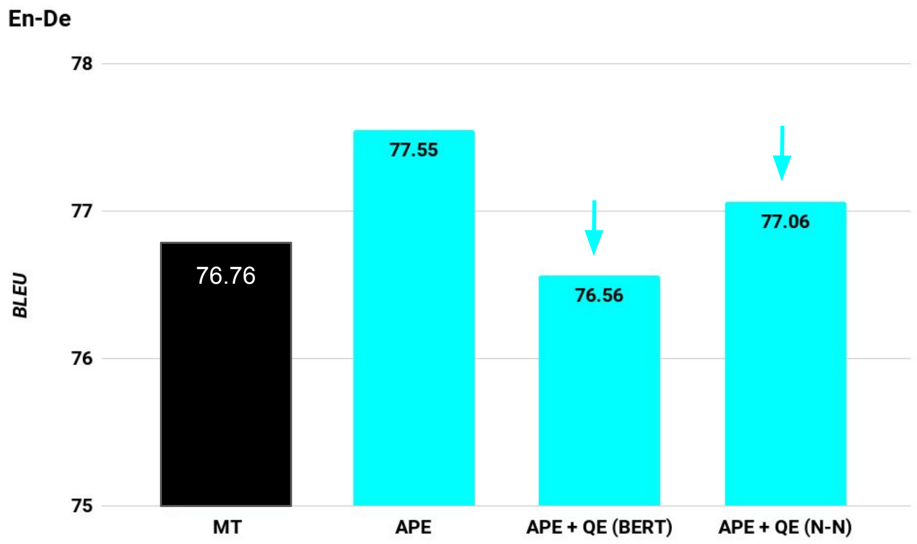


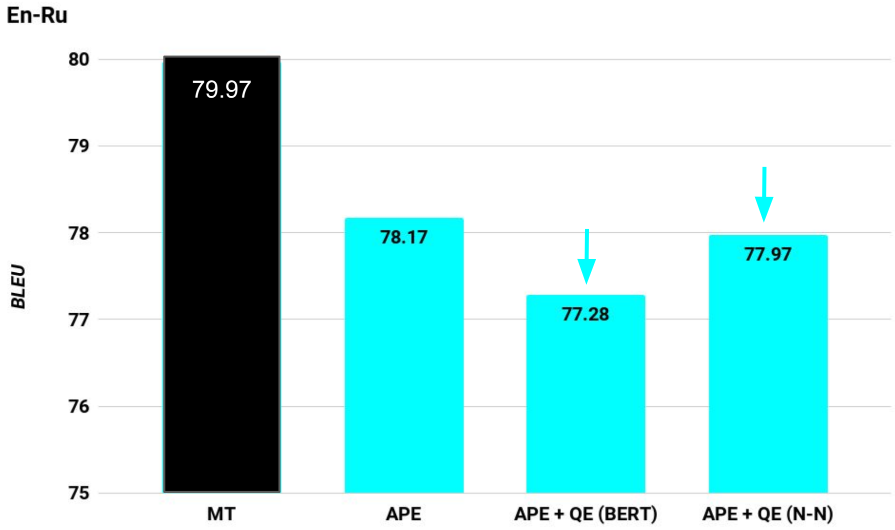


Adding the oracle token:

- Shows small improvements when using the Oracle token
- … but when the token is predicted?

QE as effort indicator results

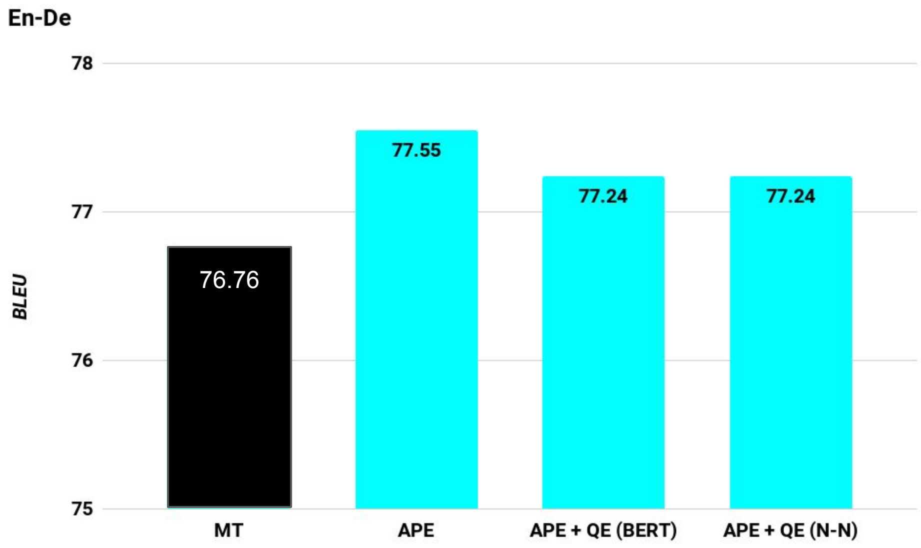


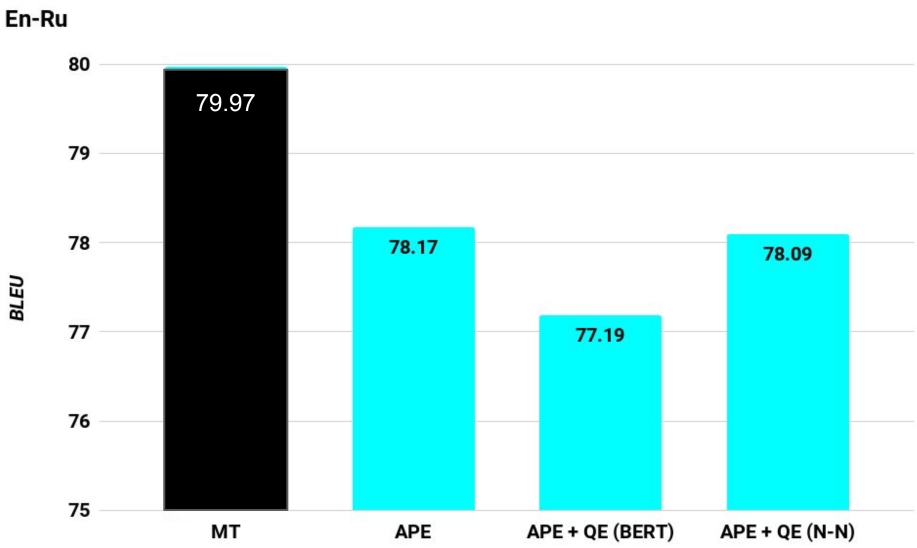


Adding the predicted token:

- Does not improve over APE without token
- Using N-N better than BERT

QE as effort indicator results





Robustify the predictor adding wrong labels in the dev

- Helps in improving the performance ...
- … but still below the APE without token

Let’s summarise

- Adding the token results in:
  - Small BLEU improvements only with the Oracle
  - APE is sensitive to the quality of the QE labels

- So ...

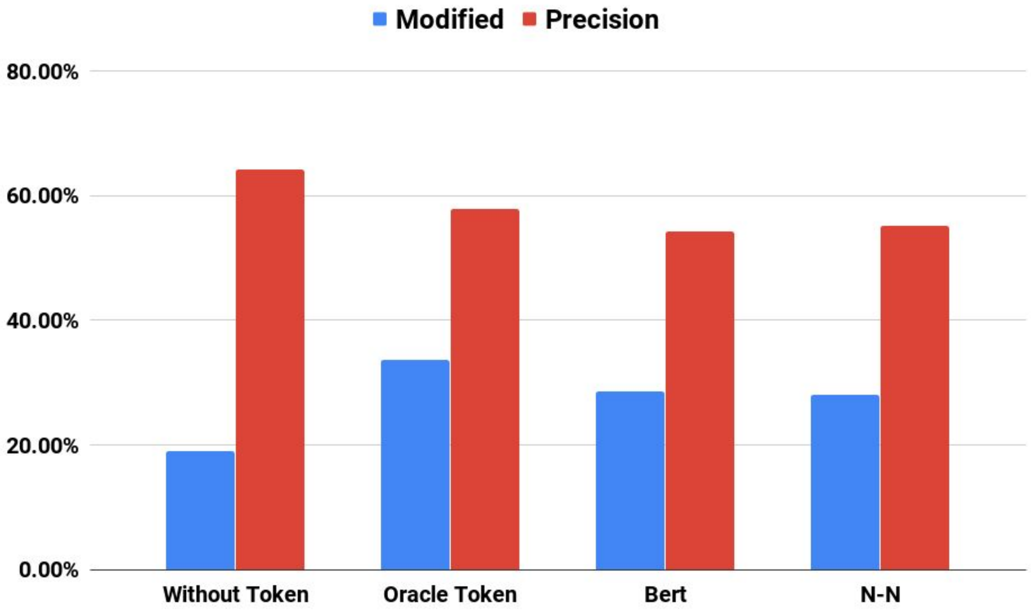
Let’s summarise

- Adding the
    - Small BL … e
    - APE is s
- So ...

Further Analysis

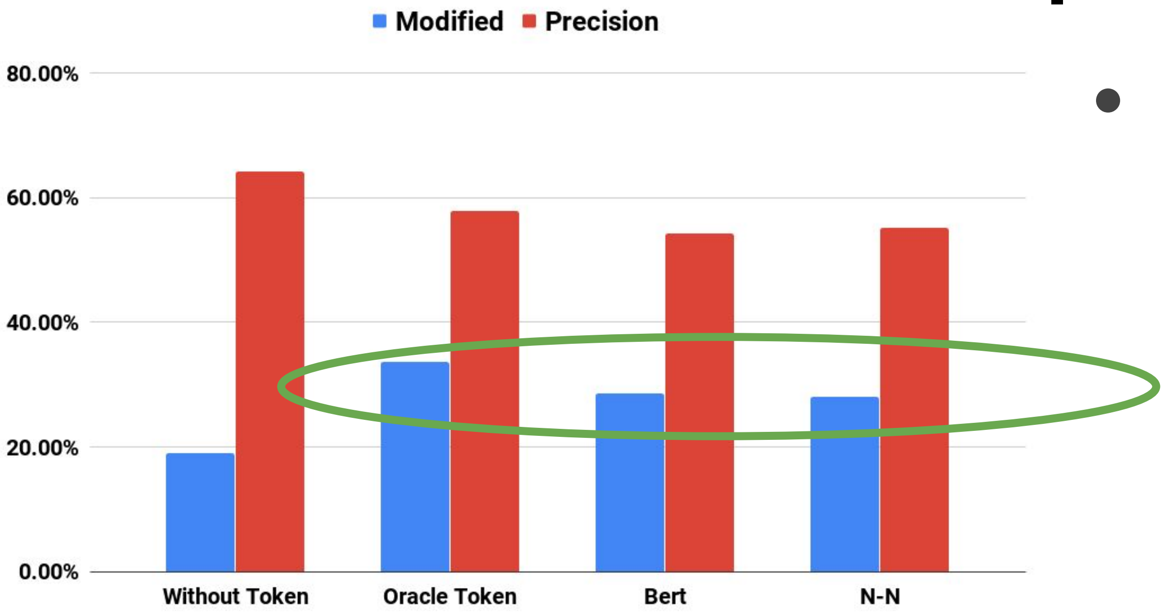
- Does the effort token help?
- How are the edits distributed?
- How does the performance change according to the token?

Does the effort token help?



- 28% of data has TER == 0
  - 72% should be modified

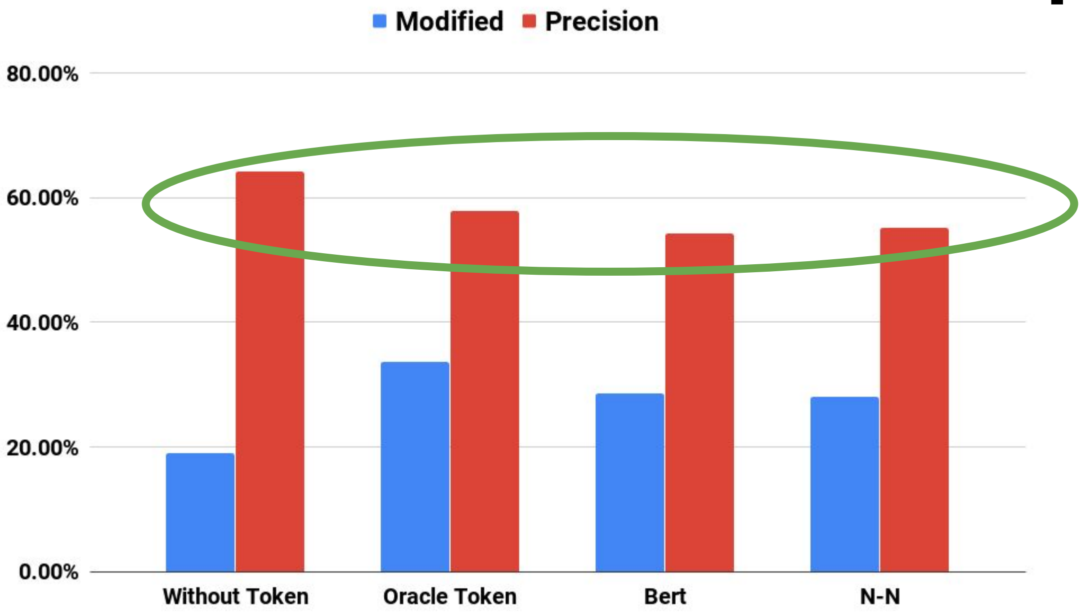
Does the effort token help?



- 28% of data has TER == 0
  - 72% should be modified

- Effort-aware APE applies more changes

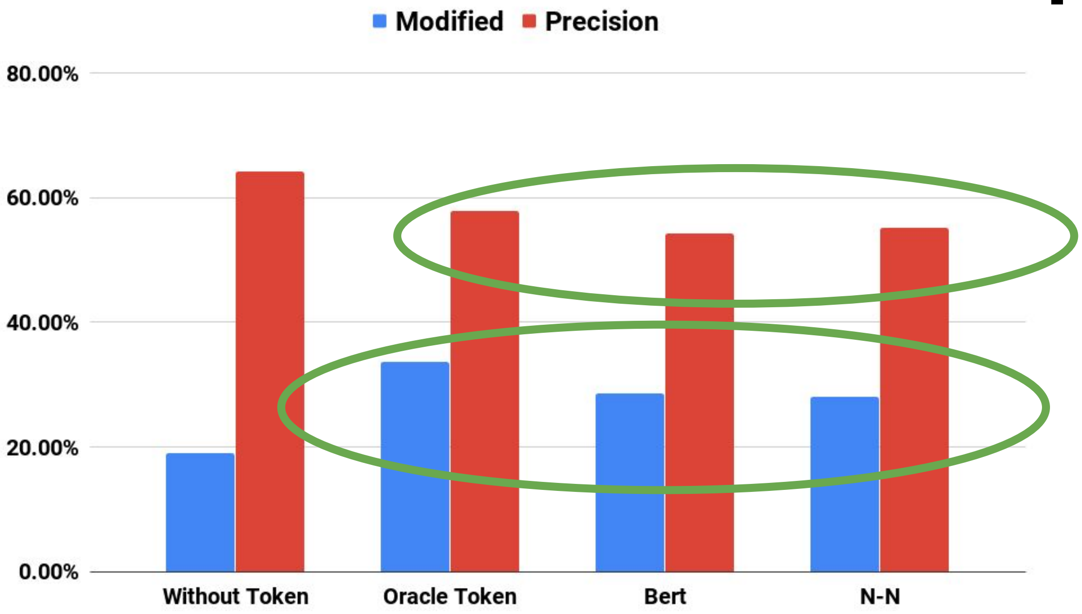
Does the effort token help?



- 28% of data has TER == 0
  - 72% should be modified

- Effort-aware APE applies more changes
- ... at the cost of a small precision drop

Does the effort token help?



- 28% of data has TER == 0
    - 72% should be modified

- System with predicted tokens not far from Oracle both in precision and sentence modifies

Further Analysis

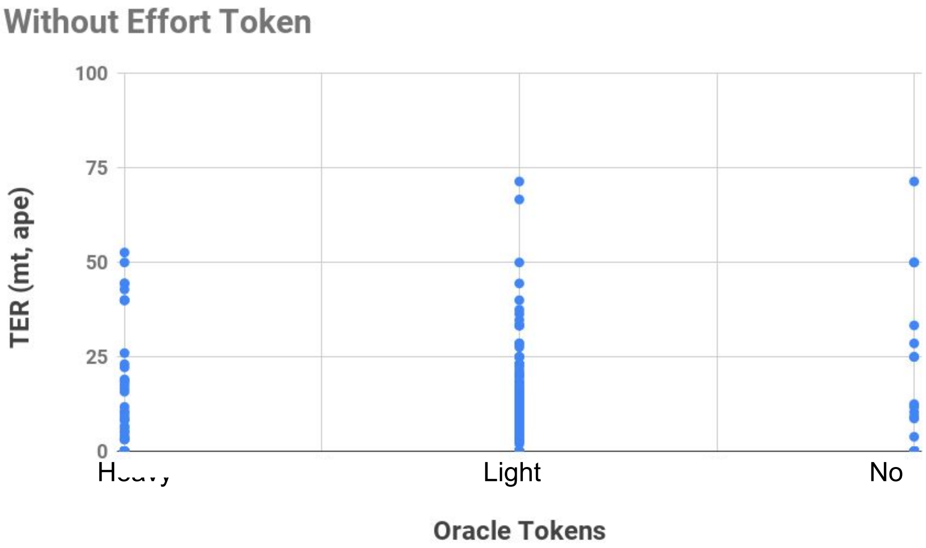
- Does the effort token help?

YES!!!

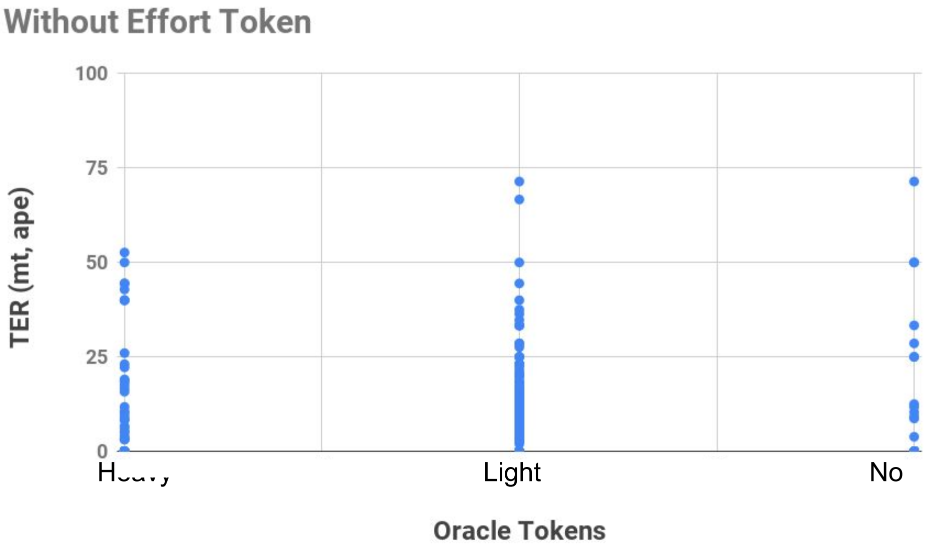
Further Analysis

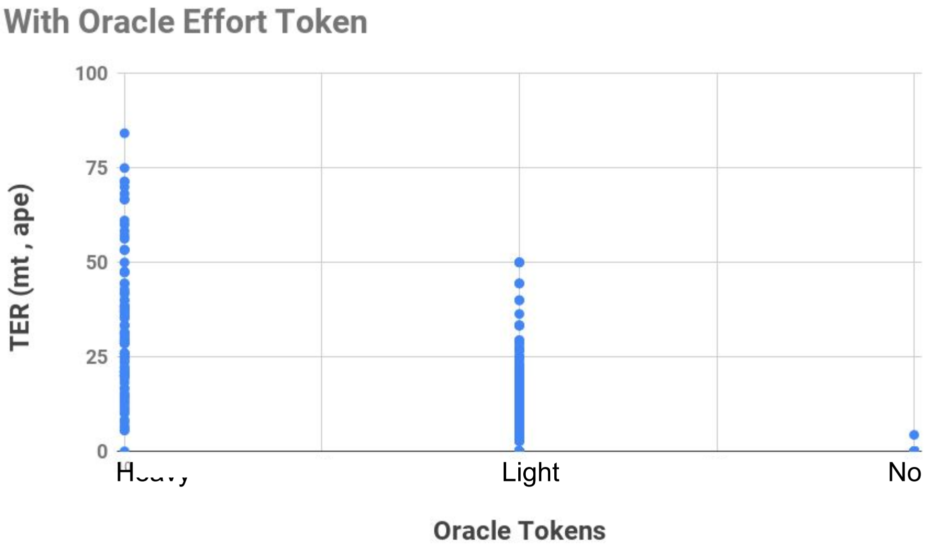
- Does the effort token help?
- How are the edits distributed?

How are the APE edits distributed?

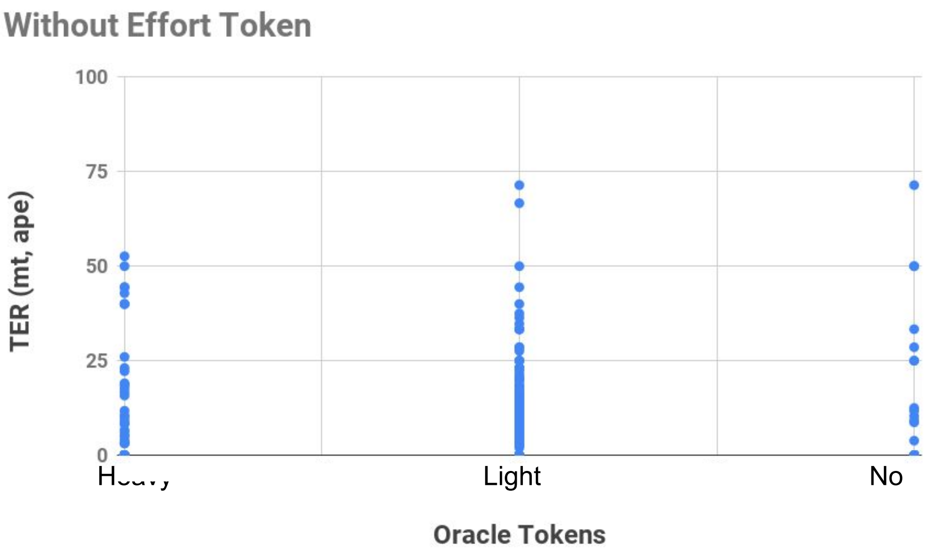


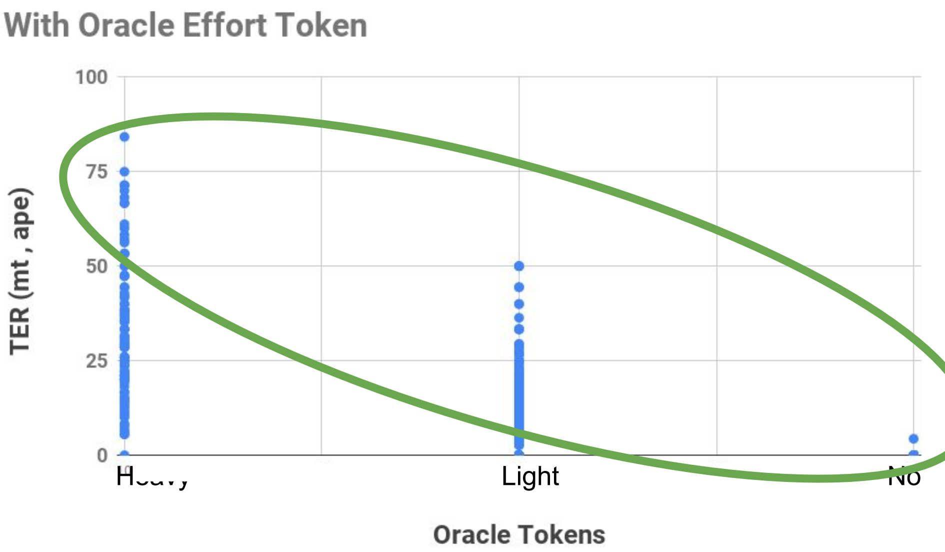
How are the APE edits distributed?





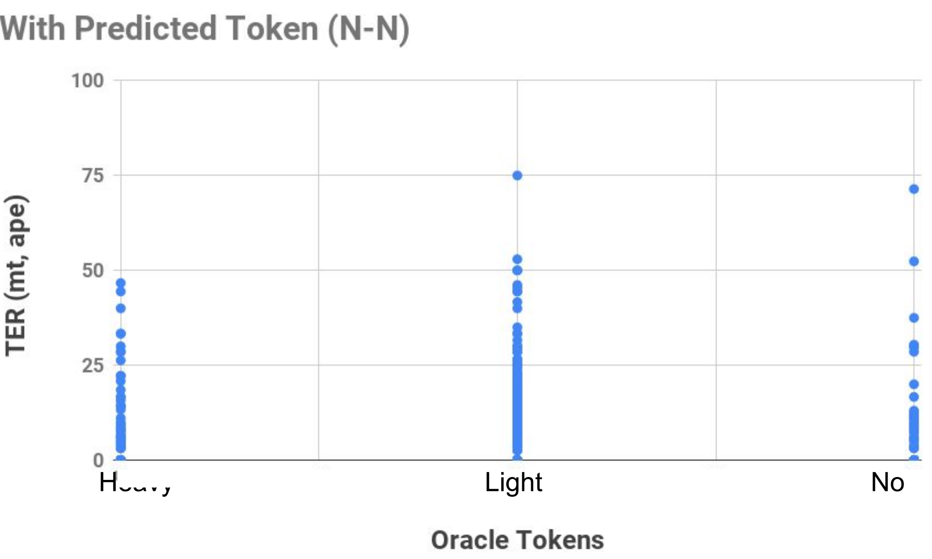
How are the APE edits distributed?

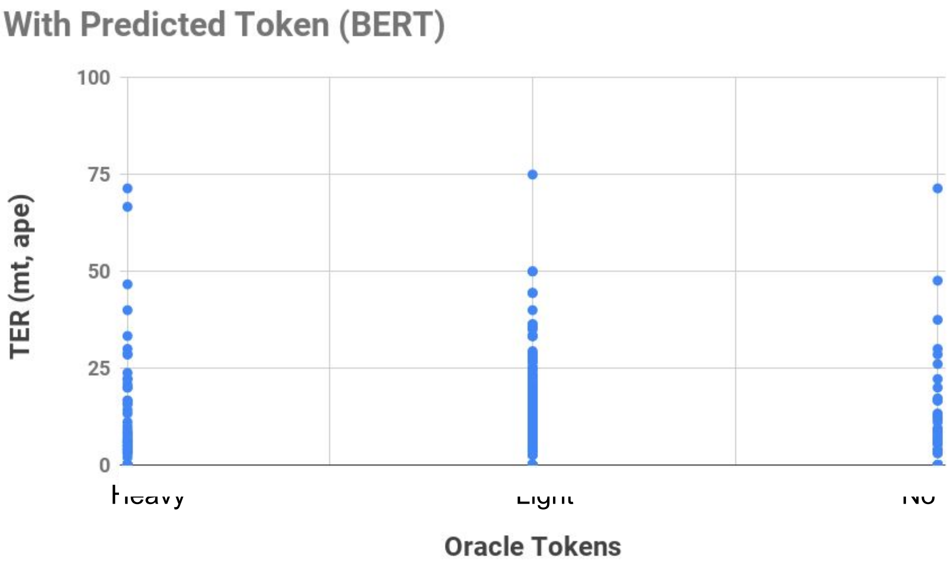




- Edits depend on the token
- Small Bleu variance, but better scenario

How are the APE edits distributed?





- Predicted tokens do not reflect the same trend
- Partial benefit from using them

Further Analysis

- Does the effort token help?
- How are the edits distributed?

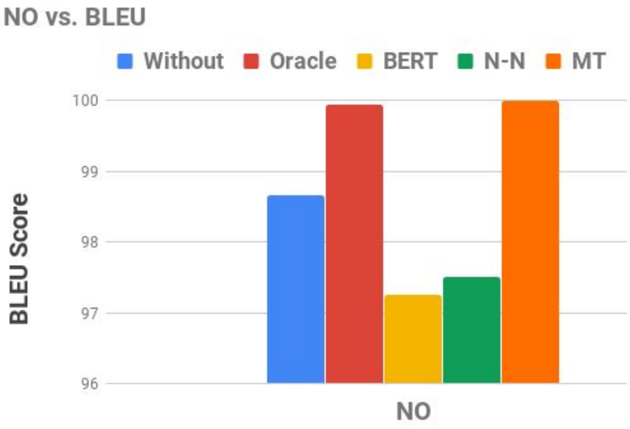
More friendly distribution for human

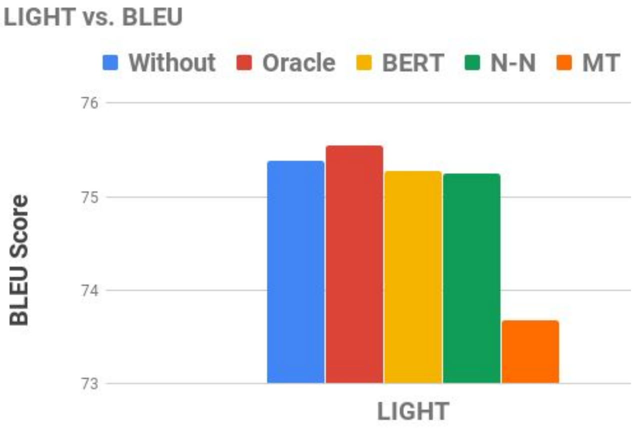
post-editing

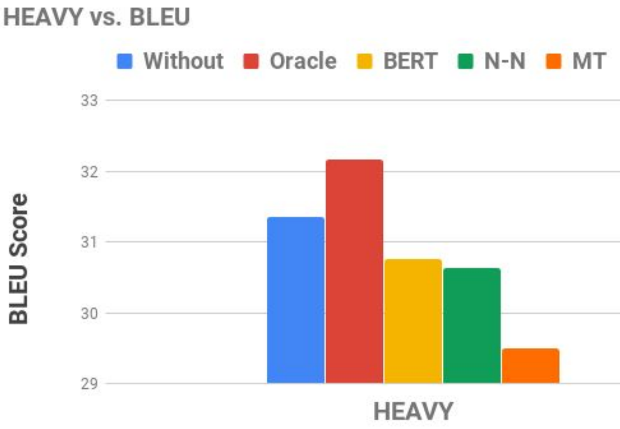
Further Analysis

- Does the effort token help?
- How are the edits distributed?
- How does the performance change according to the token?

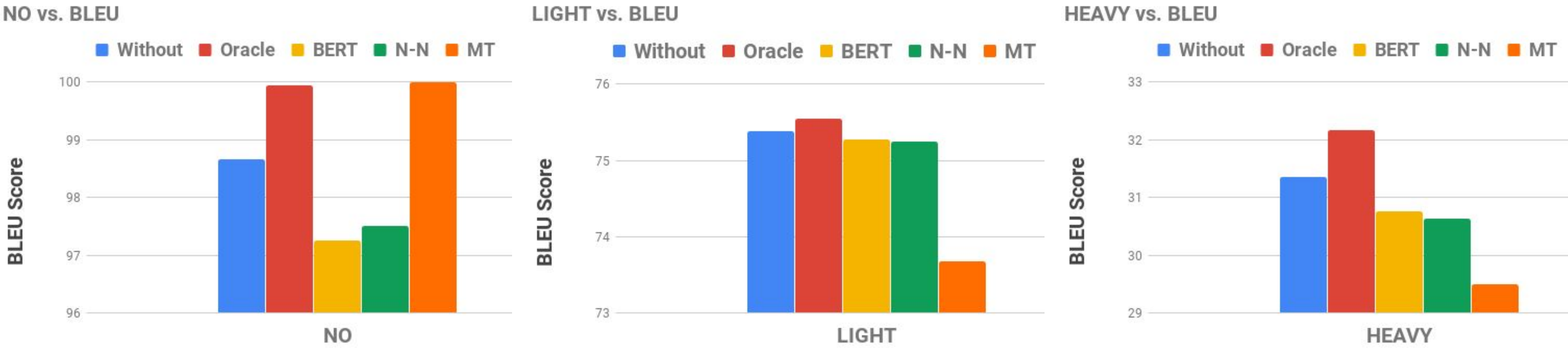
Performance vs Effort Token





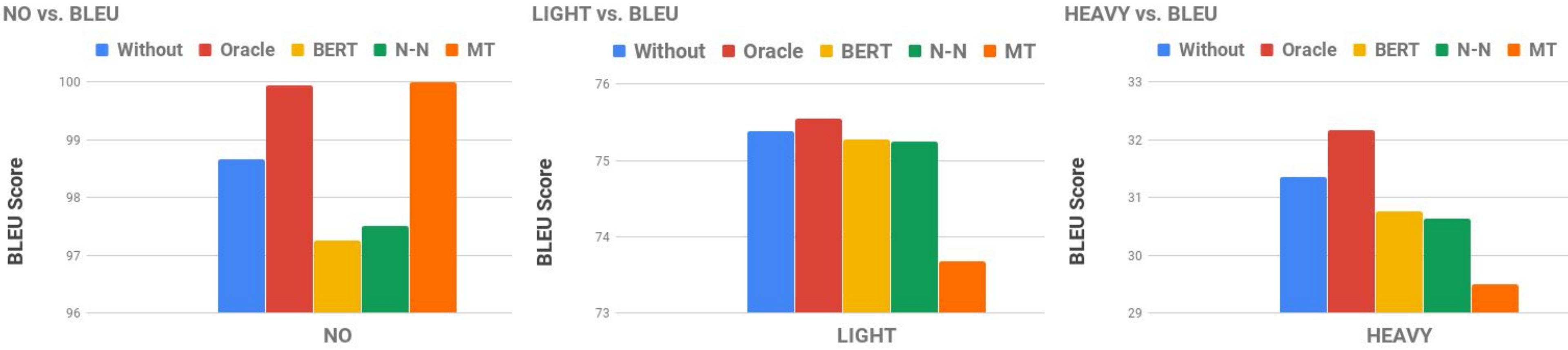


Performance vs Effort Token



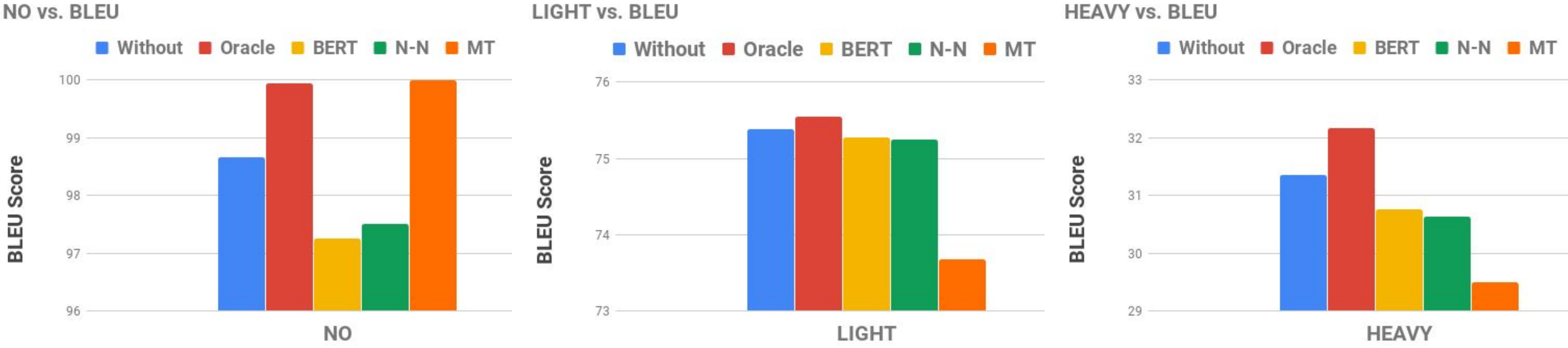
- All systems better than MT for “Light” and “Heavy”

Performance vs Effort Token



- All systems better than MT for “Light” and “Heavy”
- Oracle outperforms the others everywhere

Performance vs Effort Token



- All systems better than MT for “Light” and “Heavy”
- Oracle outperforms the others everywhere
- BERT and N-N reasonable good only for “Light”

Further Analysis

- Does the effort token help?
- How are the edits distributed?
- How does the performance change according to the token?

Oracle outperforms the “without token”

Outline

- Motivation
- Previous Work
- Effort-aware APE
- Conclusion

Conclusions

- Present a novel approach based on the effort token

- Using predicted tokens not encouraging

- Adding the Oracle token presents:
    - Small BLEU improvements
    - Better edits distribution
    - More changes, at the cost of small drop in precision

Conclusions

- Can QE support APE?
  - In theory: yes
  - In practice: not yet

- Room for improvement conditioned to:
  - More reliable QE predictions
  - More robust APE models

Quality Estimation in support of Automatic Post-Editing

Marco Turchi
Fondazione Bruno Kessler, Trento, Italy
turchi@fbk.eu

In collaboration with Amirhossein Tebbifakhr and Matteo Negri