Modern Talking in Key Point Analysis: Key Point Matching using Pretrained Encoders

Key Point Analysis Shared Task 2021

Jan Heinrich Reimer Thi Kim Hanh Luu Max Henze Yamen Ajjour

Martin Luther University Halle-Wittenberg, Germany

November 11, 2021

Key Point Matching

- Arguments influence daily decisions [Bar+20]
- Large amount of information on the Web
- Need to summarize $\rightarrow$ key points
- Find matching key points for arguments

Example

Argument:	Sex selection can lead to gender imbalance by distorting the natural male-female sex ratio.	
Key Point:	Sex selection can lead to gender imbalance	$\rightarrow$ *match*
Key Point:	It is unethical/unhealthy for parents to intervene	$\rightarrow$ *no match*

Baseline: Token Overlap

Example

Argument: Sex selection can lead to gender imbalance by distorting the natural male-female sex ratio.

Key Point: Sex selection can lead to gender imbalance

Approach

- Key points are sampled from arguments $\rightarrow$ similar vocabulary
- Count tokens that appear in argument and key point

$$\text{score}_{\text{arg,kp}} = \frac{|\{t : t \in \text{tokens}_{\text{arg}} \wedge t \in \text{tokens}_{\text{kp}}\}|}{\min\{|\text{tokens}_{\text{arg}}|, |\text{tokens}_{\text{kp}}|\}}$$

- Rule-based, no training

Preprocessing

- Stemming, synonyms, antonyms[1] $\rightsquigarrow$ generalization
- Stop words (without `not`) $\rightsquigarrow$ less noise/confusion

[1] Using NLTK [Bir06] and WordNet [Mil95]

Transformers: BERT and RoBERTa

- Pretrained encoder models:
  - BERT [Dev+18]
  - RoBERTa [Liu+19]
- Train for sentence pair regression:

BERT	`[CLS]`	`argument`	`[SEP]`		`key point`	`[SEP]`
RoBERTa	`<s>`	`argument`	`</s>`	`<s>`	`key point`	`</s>`

- Fine-tune pretrained model with ArgKP-2021 training data

Why RoBERTa?

- Trained on $10\times$ more data than BERT
- Larger batches, learning rates, step sizes $\rightarrow$ longer training
- Often outperforms BERT [Liu+19]

Transformers: BERT and RoBERTa (cont.)

Parameters and Implementation

- Simple Transformers library[2]
  `ClassificationModel(..., args={"regression": True})`
- Pretrained models
  - BERT-Base and RoBERTa-Base
  - 12 hidden layers of size 768, 12 attention heads with dropout 0.1
- Fine-tuning
  - Batch size 32, 1 epoch
  - Learning rate $2 \cdot 10^{-5}$, warmup proportion 6 %
  - No weight decay, no early stopping, no oversampling, skip missing labels

[2] `https://simpletransformers.ai/`

Evaluation: Mean Average Precision

Strict Labels

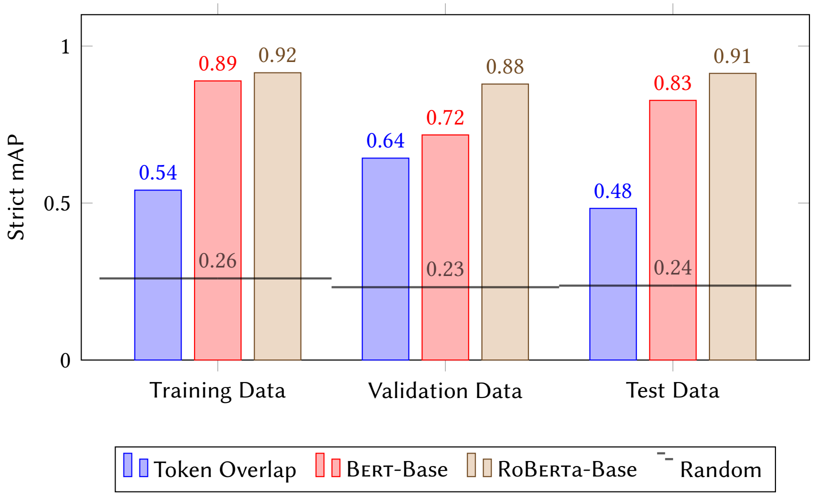

Figure: Mean average precision of the match label for different approaches and baselines under the strict label setting.

Evaluation: Mean Average Precision

Relaxed Labels

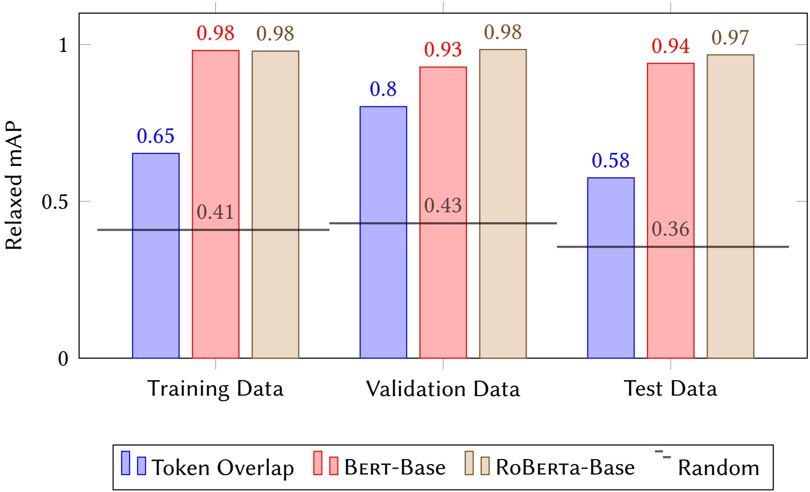

Figure: Mean average precision of the match label for different approaches and baselines under the relaxed label setting.

Error Analysis

- RoBERTa generalizes better than BERT
- BERT: some uncertain pairs (prediction around 0.5)
  $\rightarrow$ Example from training set without matching key points
  RoBERTa does predict correctly
- Difficulties wih very long arguments
  $\rightarrow$ Example from training set with 6.5$\times$ more tokens than key point
- Both predict non-matching pairs better than matching pairs
  (likely because of imbalanced training data)

Table: Falsely predicted pairs from the ArgKP-2021 dataset.

Argument	**Key point**	**True**	**BERT**	**RoBERTa**
School uniforms can be less comfortable than students' regular clothes.	School uniforms are expensive	0	0.48	0.03
affirmative action discriminates the majority, preventing skilled workers from gaining employment over someone less qualified but considered to be a member of a protected minority group.	Affirmative action reduces quality	1	-0.05	0.03

Conclusion

- Strong, rule-based baseline (twice as good as random)
- BERT an RoBERTa models better for context understanding
- Scores on test set

  mAP strict: **0.913**

  mAP relaxed: **0.967**
- Hyperparameter tuning is important

Future Work

- Ensemble with RoBERTa and overlap baseline
- Improved, more robust language models [Sun+21]
- Avanced textual oversampling to balance training data

Thank you!

References

- Bar-Haim, Roy et al. (2020). "From Arguments to Key Points: Towards Automatic Argument Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky et al. Association for Computational Linguistics, pp. 4029–4039.
- Bird, Steven (2006). "NLTK: the natural language toolkit". In: *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 69–72.
- Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.
- Liu, Yinhan et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692. arXiv: `1907.11692`.
- Miller, George A (1995). "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11, pp. 39–41.
- Sun, Yu et al. (2021). "ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation". In: *CoRR* abs/2107.02137. arXiv: `2107.02137`.