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 → 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 → match
Key Point: It is unethical/unhealthy for parents to intervene → 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

$$score_{arg,kp} = \frac{|\{t : t \in tokens_{arg} \land t \in tokens_{kp}\}|}{\min\{|tokens_{arg}|, |tokens_{kp}|\}}$$

- Rule-based, no training

Preprocessing

- Stemming, synonyms, antonyms$^1$ $\rightarrow$ generalization
- Stop words (without not) $\rightarrow$ 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\textsuperscript{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

\textsuperscript{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)
  → Example from training set without matching key points
  RoBERTa does predict correctly
- **Difficulties with very long arguments**
  → 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.

<table>
<thead>
<tr>
<th>Argument</th>
<th>Key point</th>
<th>True</th>
<th>BERT</th>
<th>RoBERTa</th>
</tr>
</thead>
<tbody>
<tr>
<td>School uniforms can be less comfortable than students’ regular clothes.</td>
<td>School uniforms are expensive</td>
<td>0</td>
<td>0.48</td>
<td>0.03</td>
</tr>
<tr>
<td>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.</td>
<td>Affirmative action reduces quality</td>
<td>1</td>
<td>-0.05</td>
<td>0.03</td>
</tr>
</tbody>
</table>
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]
▶ Advanced textual oversampling to balance training data

Thank you!
References


