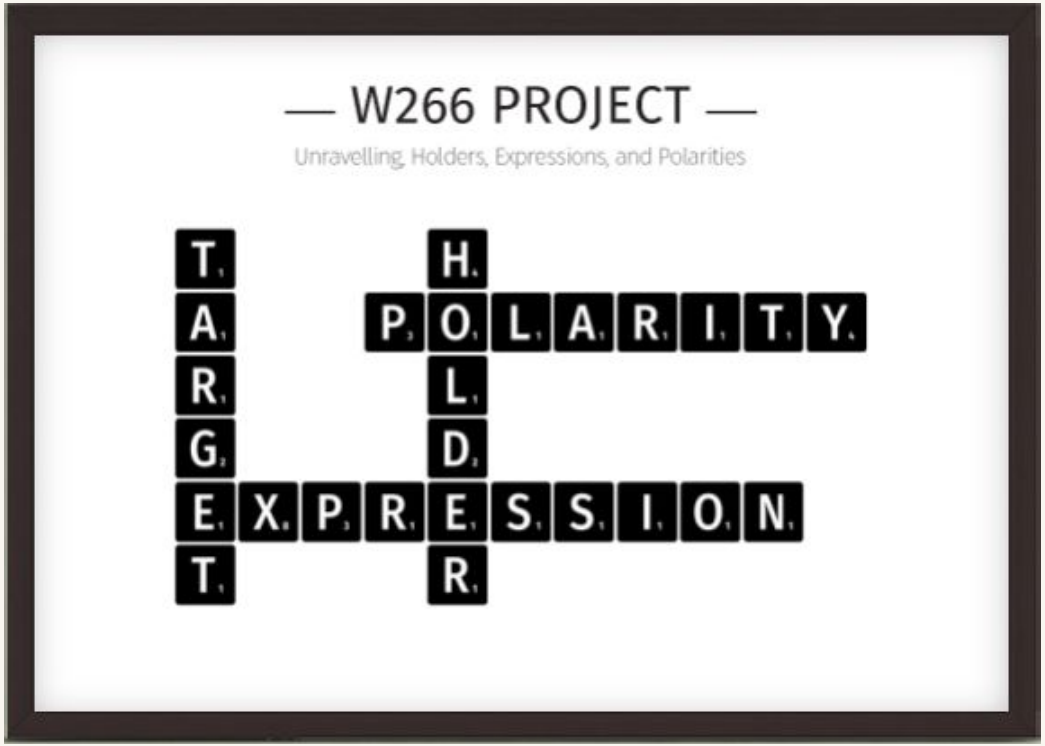
# Gridlocked Opinions:
## A Tagging Scheme Unravelling *Targets*, *Holders*, *Expressions*, and *Polarities*

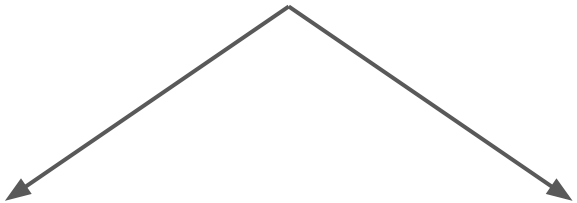



W266
Javier Rondon, Dominic Lim

# SemEval 2022 Task 10: Structured Sentiment Analysis

(holder, target, expression, polarity)

“The room was great, but I prefer the penthouse”



(“-”, room, “was great”, “Positive”) (“I”, “penthouse”, “prefer”,”Positive”)

# SemEval 2022 Task 10: Structured Sentiment Analysis (continued)

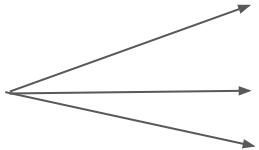
- **Objective: Propose a simplified architecture to predict all possible quadruples of holder, target, expression and polarity.**

- Methodology
  - Reviewed Sem-Eval 2022 submissions to find opportunities for improvement
    - Prior research in this topic considered solutions based on Sequence Labelling and Dependency Graph Parsing
    - Evaluated on “Sentiment Graph $F_1$” scores
    - Used provided baseline models: Dependency Graph Parsing and Sequence-Labeling pipeline

  - Modified the Grid Tagging Scheme solution from ISCAS team and assessed the performance of our solution

# Grid Tagging Scheme (GTS)

- We extended GTS with additional tags to extract holders, implicit holders and targets

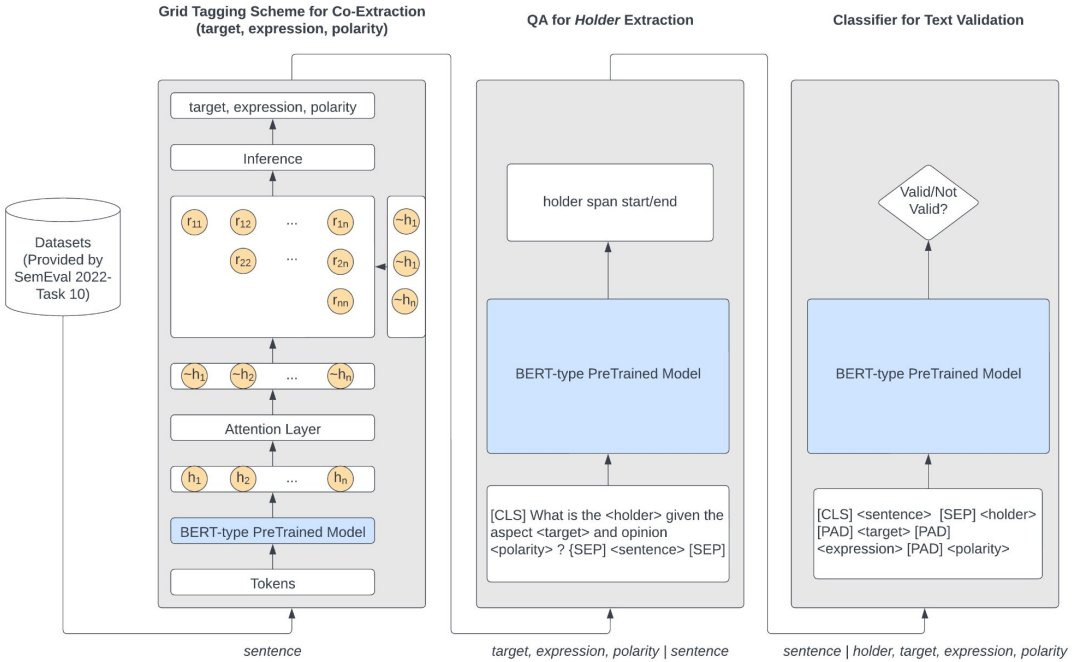


| Tags | Meaning |
|---|---|
| Target | words in pair (wi,wj) belong to the same target term |
| Expression | words in pair (wi,wj) belong to the same opinion term |
| Holder | words in pair (wi,wj) belong to the same holder term |
| IA | implicit target term |
| IO | implicit expression term |
| IH | implicit holder term |
| IHIA | implicit holder and target term |
| Positive | the words belong to a target, |
| Negative | expression and holder tuple |
| Neutral | and form a positive/neutral/negative relation |
| 0 | No relation |

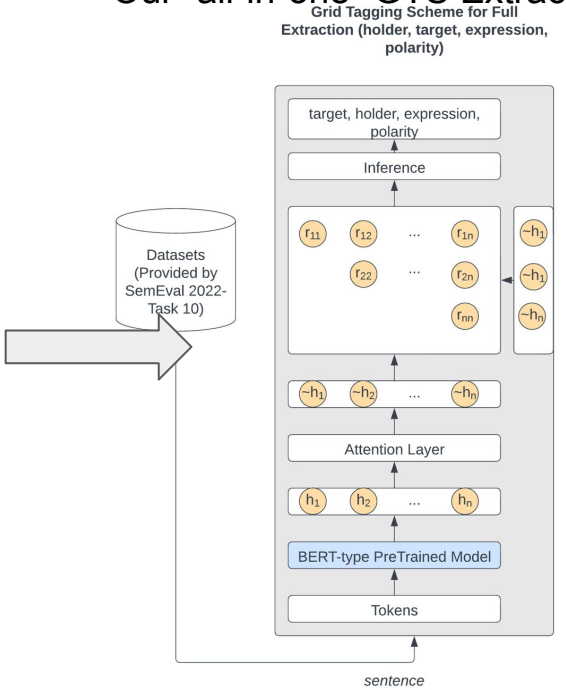
| [CLS] | Fantastic | food | and | breathtaking | view | |
|---|---|---|---|---|---|---|
| Implicit Holder | Positive | 0 | 0 | Positive | 0 | [CLS] |
| | Expression | Positive | 0 | 0 | 0 | Fantastic |
| | | Target | 0 | 0 | 0 | food |
| | | | 0 | 0 | 0 | and |
| | | | | Expression | Positive | breathtaking |
| | | | | | Target | view |

# Architecture / Proposed Solution

## Original ISCAS Extraction/Validation Pipeline



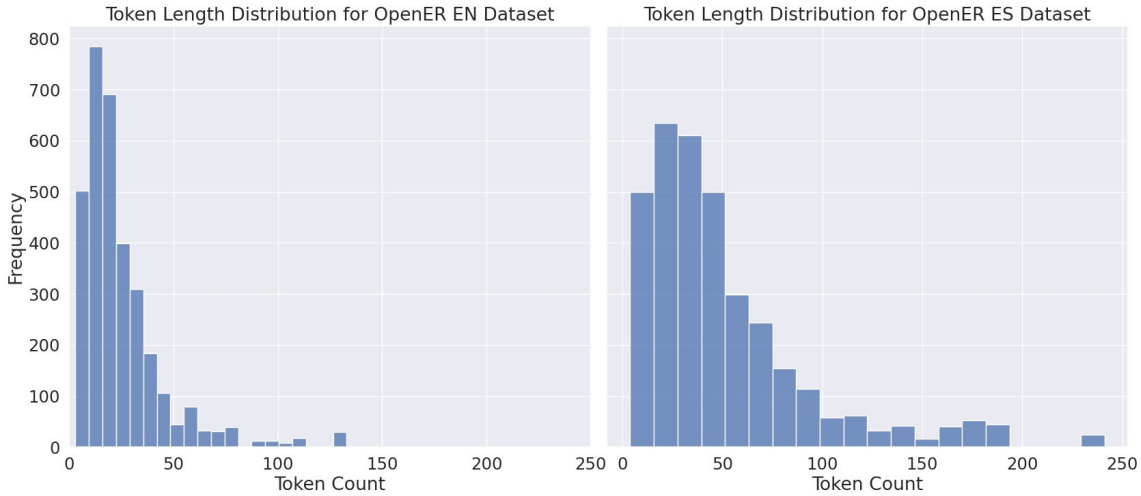

## Our “all-in-one” GTS Extraction



# $\text{OpenER}_{\text{EN}}$ / $\text{OpenER}_{\text{ES}}$ Datasets

- Annotated hotel reviews in English and Spanish
- Opinions with Missing Holder and Missing Targets
- Pronounced positive skew in distribution of tokens per review

| | Datasets | # of Reviews | # of Holders | # of Targets | # of Expressions | # of Polarities |
|---|---|---|---|---|---|---|
| **OpenER-EN** | Train | 1,744 | 266 | 2,679 | 2,884 | 2,884 |
| | Dev | 249 | 49 | 371 | 400 | 400 |
| **OpenER-ES** | Train | 1,438 | 176 | 2,748 | 3,042 | 3,042 |
| | Dev | 206 | 23 | 363 | 387 | 387 |



# Experiment Design and Results

- Experimented “all-in-one” GTS with different pre-trained BERT variant models
- Models trained on NVIDIA A10 GPU (24 GB PCIE)
- Training times in excess of 8 hours
- Best models are **all-RoBERTa-large** and **XLM-RoBERTa**

| Dataset | | Language Model | $F_1$ Score | Precision | Recall |
|---|---|---|---|---|---|
| **OpenER-EN** | Dev | BERT large uncased | 0.66 | 0.67 | 0.65 |
| | Test | | 0.62 | 0.65 | 0.59 |
| **OpenER-EN** | Dev | BERT_review | 0.65 | 0.69 | 0.62 |
| | Test | | 0.63 | 0.66 | 0.6 |
| **OpenER-EN** | Dev | all-RoBERTa-large-v1 | 0.66 | 0.7 | 0.62 |
| | Test | | **0.66** | **0.68** | **0.64** |
| **OpenER-ES** | Dev | XLM_RoBERTa_large | 0.67 | 0.74 | 0.62 |
| | Test | | **0.61** | **0.71** | **0.54** |
| **OpenER-ES** | Dev | RoBERTa-large-bne | 0.61 | 0.64 | 0.58 |
| | Test | | 0.58 | 0.62 | 0.54 |
| **OpenER-ES** | Dev | distilbert-base | 0.48 | 0.63 | 0.48 |
| | Test | -multilingual-cased | 0.37 | 0.58 | 0.28 |

**Results:**

- Best Test Sentiment $F_1$ score of **0.66** and **0.61** for EN and ES, respectively
- Compelling Performance compared to provided baseline and ISCAS teams
- Addtl hyper-parameter tuning could tighten gap between ISCAS and “all-in-one” model

| Model | OpenER-EN | OpenER-ES |
|---|---|---|
| Graph Baseline | 0.521 | 0.495 |
| Seq Baseline | 0.329 | 0.24 |
| ISCAS GTS | 0.71 | 0.669 |
| **Our results** | **0.66** | **0.61** |

# Example of Incorrect extraction

**Label:**


```
{
  "sent_id":"opener_en/kaf/hotel/english00214_f731285c2d232cf15e5cdae66ab186b1-6",
  "text":"The first floor 24 hr bar was well run and no matter what time of day
          there was always a waiter on hand to serve .",
  "opinions":[
     {
        "Source":[[],[]],
        "Target":[["24 hr bar"],["16:25"]],
        "Polar_expression":[["well run"],["30:38"]],
        "Polarity":"Positive",
        "Intensity":"Standard"
         ....
```


**Prediction:**


```
  "sent_id": "opener_en/kaf/hotel/english00214_f731285c2d232cf15e5cdae66ab186b1-6",
  "text": "The first floor 24 hr bar was well run and no matter what time of day
           there was always a waiter on hand to serve .",
  "opinions": [
    {
      "Source": [[],[]],
      "Target": [["bar"],["22:25"]],
      "Polar_expression": [["well run"],["30:38"]],
      "Polarity": "Positive"
    }
  "opinions": [
    {
      "Source": [[],[]],
      "Target": [["The first floor"],["0:15"]],
      "Polar_expression": [["24 hr"],["16:21"]],
      "Polarity": "Positive"
```


- The model correctly classified the target but struggled on the span start/end. i.e “24 hr bar” vs “bar”
- Predicted target is still useful for sentiment analysis
- Limitation:
  - Token length (24) > median token length/review (18)
  - Models trained on deeper networks (# hops) might help to disentangle longer and more complex reviews

# Conclusion

- Experimented “all-in-one” Grid Tagging Scheme with different BERT variant models
  - **all-RoBERTa-large** and **XLM-RoBERTa** produced best results for EN, ES datasets

- Produced Compelling Results
  - $F_1$ scores of 0.66 and 0.61 for OpeNER-EN and OpeNER-ES dataset, respectively
  - Considerably outperformed Dep Graph and Sequence baseline models
  - Only slightly underperformed original ISCAS submission despite significantly fewer computing/resource demands

- Performance gains could be realized with additional hyper parameter tuning.