Position-Aware Tagging for Aspect Sentiment Triplet Extraction (ASTE): A Summary

Nada Alaoui

October 4, 2025

Abstract

This document summarizes the paper "Position-Aware Tagging for Aspect Sentiment Triplet Extraction (ASTE)," which introduces a novel joint, end-to-end model for ASTE. Moving beyond traditional pipeline approaches, the core innovation lies in a position-aware tagging scheme that simultaneously extracts target entities, their associated sentiments, and corresponding opinion spans. This scheme enriches standard BIOES tags with sentiment polarity and relative offsets to opinion span boundaries, enabling the model to capture intricate interactions between triplet elements in a single sequence tagging pass. The proposed JET (Jointly Extract the Triplets) model, utilizing this tagging scheme, demonstrates superior performance compared to existing pipeline methods across various datasets, establishing a new state-of-the-art in ASTE. Further analysis reveals the robustness of the approach across different span lengths and strictness levels, the importance of its core features (distance and opinion quality scores), and the complementary nature of its variants, which can be leveraged through ensemble methods for enhanced recall. The work significantly advances the field of Aspect-Based Sentiment Analysis by providing a more comprehensive and effective solution for granular sentiment understanding.

1 Core Idea: Joint Extraction via Position-Aware Tagging

The paper's central idea is to move away from traditional pipeline approaches for Aspect Sentiment Triplet Extraction (ASTE) and instead propose a joint, end-to-end model that leverages a novel position-aware tagging scheme [16].

Traditional pipeline methods break the ASTE task into multiple stages (e.g., first extract aspects, then sentiments, then link them to opinion spans). The authors observe that the three elements of an ASTE triplet—the target entity, its associated sentiment, and the opinion span explaining the sentiment—are highly interconnected. They argue that capturing these rich interactions jointly, rather than sequentially, will lead to better performance.

The main challenge for a joint sequence tagging approach is how to effectively design tags that can encode all three pieces of information and their relationships within a single sequence. This is where the "position-aware tagging scheme" comes in.

2 Most Important Concepts Explained

2.1 Aspect Sentiment Triplet Extraction (ASTE)

- **Definition:** The task of extracting triplets of (Target Entity, Associated Sentiment, Opinion Span).
- **Components:**
  - **Target Entity:** The specific object or aspect being discussed (e.g., ”food,” ”vegan options,” ”salad”).
  - **Associated Sentiment:** The polarity of the sentiment towards the target (e.g., positive +, neutral 0, negative -).
  - **Opinion Span:** The words or phrases that express or explain the sentiment for that specific target (e.g., ”so so,” ”excited,” ”cheap with fresh salmon”).
- **Example (from the paper):** In ”food was so so but excited to see many vegan options”
  - Triplet 1: (”food”, 0, ”so so”)
  - Triplet 2: (”vegan options”, +, ”excited”)
- **Significance:** This task provides a more granular and comprehensive understanding of sentiment compared to simpler aspect-based sentiment analysis, as it explicitly links the sentiment to the specific words that justify it.

2.2 Limitations of Existing Approaches (Pipeline & BIOES)

- **Pipeline:** Prior work often used multi-stage pipelines, e.g., first extracting targets with sentiment, then using a separate classifier to pair them with opinion spans. This struggles to capture the inherent relationships between the triplet elements.
- **BIOES Tagging:** Standard sequence tagging schemes like BIOES (Begin, Inside, Outside, End, Single) [5] are excellent for identifying spans (like named entities or aspect terms). However, they lack the expressiveness to directly encode:
  - The sentiment polarity for a specific target.
  - The connection between a target and its corresponding (potentially distant) opinion span. For instance, knowing ”vegan” is the start of a target doesn’t tell you where its opinion ”excited” is located. This typically necessitates a separate linking step.

2.3 Position-Aware Tagging Scheme

- **Innovation:** The core contribution. It extends the standard BIOES tags (B and S) to include additional information:

- B (Begin) and S (Single-word target) are enriched to $B^{\varepsilon}_{j,k}$ and $S^{\varepsilon}_{j,k}$.
- ε (Epsilon): Denotes the sentiment polarity for the target (+, 0, -).
- j, k (Offsets): Indicate the relative positions (distances) of the opinion span's left and right ends, measured from the starting position of the target.
  * Positive offsets: Opinion span is to the right of the target.
  * Negative offsets: Opinion span is to the left of the target.

- **How it works (Example from paper):**
  - For "food": Tagged as $S^{0}_{2,3}$
    * S: Single-word target.
    * 0: Neutral sentiment.
    * 2,3: Opinion span starts 2 words to the right of "food" and ends 3 words to the right of "food" (i.e., "so so").
  - For "vegan options" (first word "vegan"): Tagged as $B^{+}_{-4,-4}$
    * B: Beginning of a multi-word target.
    * +: Positive sentiment.
    * -4,-4: Opinion span starts 4 words to the left of "vegan" and ends 4 words to the left of "vegan" (i.e., "excited").

- **Benefits:**
  - **Joint Extraction:** A single sequence tagging pass can extract all three triplet elements simultaneously.
  - **Structural Information:** Explicitly encodes the connection and relative position between a target and its opinion span, even if they are far apart.
  - **Rich Interactions:** Allows the model to learn and leverage the interplay between target, sentiment, and opinion span during training.
  - **One-to-one correspondence (Theorem 2.1):** The paper proves that this scheme allows a unique mapping between a tag sequence and a set of non-overlapping triplets (each with one opinion span), ensuring its theoretical soundness.
  - **Handles Overlapping Opinion Spans:** Can deal with complex cases where opinion spans for different targets might overlap, which was difficult for previous methods.

2.4 JET (Jointly Extract the Triplets) Model

This is the proposed neural architecture that utilizes the position-aware tagging scheme. While the paper outlines a simple LSTM-based neural architecture for learning feature representations, the core innovation lies in how these representations are used with the novel tagging scheme to compute "factorized feature scores." These scores help capture the complex interactions among the elements within a triplet, allowing the model to jointly predict all parts.

In essence, the paper argues that by cleverly designing the output labels to encode more information (sentiment and opinion span position), they can transform a complex multi-stage problem into a single, more effective sequence tagging task. This leads to better performance because the model can learn the interdependencies between targets, sentiments, and opinion spans from the start.

3 Complete Explanation of How JET's Scoring System Works

3.1 The Big Picture

JET needs to look at a sentence and decide how to label each word. There are many possible ways to label a sentence, so JET picks the labeling with the highest probability.

3.2 Step 1: The Probability Formula (CRF)

The model uses a conditional random field (CRF) framework [7]:

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp(s(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}'} \exp(s(\mathbf{x}, \mathbf{y}'))}$$

- $\mathbf{x}$ = input sentence (like "food was delicious")
- $\mathbf{y}$ = ONE complete way to label the entire sentence
- $\mathbf{y}'$ = all other possible ways to label it
- This formula converts raw scores into probabilities (0 to 1).
- The labeling with highest probability wins!

Process: Score each possible labeling → Convert to probabilities → Pick the highest.

3.3 Step 2: How We Calculate the Score $s(\mathbf{x}, \mathbf{y})$

$$s(\mathbf{x}, \mathbf{y}) = \sum_i \psi(\overline{y}_i, \overline{y}_{i+1}) + \sum_i \Phi(y_i)$$

The score has TWO parts that we add together:

- **Part A: Transition Scores** $\sum_i \psi(\overline{y}_i, \overline{y}_{i+1})$
  - Looks at pairs of adjacent labels.
  - Asks: "Do these two labels make sense next to each other?"

- Example: B→I (good) vs B→B (weird).

- **Part B: Individual Label Scores** $\sum_i \Phi(y_i)$
  - Looks at each word's label individually.
  - Asks: "Does this label make sense for this word?"
  - This is where JET's innovation happens!

The summation $\sum_i$ means: "For every position i in the sentence, calculate this score and add them all up."

3.4 Step 3: Individual Label Scores - The Innovation!

Basic score:

$$\Phi(y_i) = f_t(h_i)_{\overline{y}_i}$$

- h_i = BiLSTM's understanding of word i in context [4, 3].
- $f_t(h_i)$ returns scores for [B, I, O, E, S].
- $\overline{y}_i$ subscript means: pick the score that matches our label's sub-tag.

For position-aware tags (like $B^{+}_{2,3}$ or $S^{0}_{-1,1}$), we add 3 MORE scores:

$$\Phi(y_i) += f_s([g_{i+j,i+k}; \overleftarrow{h_i}])_{\varepsilon} + f_o(g_{i+j,i+k}) + f_r(j,k)$$

- **Score 1: Sentiment Score** $f_s([g_{i+j,i+k}; \overleftarrow{h_i}])_{\varepsilon}$
  - $g_{i+j,i+k}$ = representation of the opinion span.
  - $\overleftarrow{h_i}$ = backward context of the target.
  - $[;]$ = concatenation of both representations.
  - ε subscript = pick score for the sentiment (positive/negative/neutral).
  - Asks: "Do target + opinion together suggest this sentiment?"
- **Score 2: Opinion Quality Score** $f_o(g_{i+j,i+k})$
  - Takes only the opinion span representation.
  - Returns single score for "how good are these opinion words?"
  - "delicious" = high score, "the table" = low score.
- **Score 3: Distance Score** $f_r(j,k) = W_r \times w_r[\min(j,k)] + b_r$
  - j, k = offset distances to opinion span.
  - $w_r[\min(j,k)]$ = learned embedding for this distance.
  - W_r, b_r = learnable parameters.
  - Asks: "Is this a reasonable distance between target and opinion?"

3.5 Training vs Prediction - Important Distinction!

- **During Training (Learning Phase):**
  - We KNOW the correct labels y_i from training data.
  - We use these known labels to pick scores and teach the model.
  - Goal: Make correct labelings get high scores.
- **During Prediction (Testing Phase):**
  - We DON'T know the correct labels - that's what we're finding!
  - We calculate scores for ALL possible complete sentence labelings.
  - Pick the complete labeling with highest probability.

3.6 The Complete Process

1. Try every possible way to label the entire sentence.
2. For each possible labeling: Calculate total score using transition scores + individual label scores.
3. For position-aware tags: Add the 3 special scores (sentiment + opinion quality + distance).
4. Convert all scores to probabilities using the CRF formula.
5. Pick the labeling with highest probability.

Why This Works Better: Instead of finding targets and opinions separately then matching them, JET considers ALL the relationships (target quality + opinion quality + sentiment + distance) simultaneously when scoring each possible labeling.

3.7 Handling Multiple Opinion Spans - The JET^o Variant

- **The Problem with JET^t:**
  - JET^t (target-focused) can handle multiple targets sharing the same opinion span.
  - BUT it cannot handle one target associated with multiple different opinion spans.
  - Example issue: "food" $\rightarrow$ "delicious" AND "fresh" (one target, multiple opinions).
- **The Solution: JET^o (Opinion-Focused Approach):**
  - **Core Innovation:** Flip the tagging perspective - tag opinion spans instead of targets.
  - **How it works:** Use the same $B^{\varepsilon}_{j,k}$ and $S^{\varepsilon}_{j,k}$ tags, but now:
    * B/S encode the opinion span (not the target).

  - * j, k offsets point to where the target is located.
    * ε remains the sentiment polarity.
  - **Benefits:**
    * Multiple opinions can now point to the same target.
    * Each opinion gets its own tag with offsets back to the shared target.
    * Solves the "one target, multiple opinions" limitation.

- **Model Variants:**
  - JET^t: Tags targets, points to opinions (original approach).
  - JET^o: Tags opinions, points to targets (variant for multiple opinions).

3.8 Training and Inference

- **Training Process:**
  - **Loss Function:** $L = -\sum \log p(\text{correct\_labels — sentence})$.
  - **Goal:** Maximize probability of correct labelings on training data.
  - Uses standard neural CRF training procedures.

- **Inference (Prediction):**
  - **Method:** Viterbi-like MAP inference (finds most probable complete labeling).
  - **Process:** Consider all possible ways to label the sentence → Pick highest probability.

- **Computational Complexity:**
  - **Challenge:** Position-aware tags create $O(M^2)$ possible labels per position, where M = maximum opinion span length (set to 2-6 words).
  - **Time Complexity:** $O(nM^2)$ where n = sentence length.
  - **Efficiency:** Since $M \ll n$ (M≤6, n≤80), this is better than previous $O(n^2)$ methods.
  - **Why it works:** Opinion spans are typically short, keeping M small and manageable.

Key Insight: The position-aware scheme remains computationally efficient because opinion spans are naturally short, making the expanded label space manageable while providing richer representational power.

4 Experiments

4.1 Data - Dataset Improvements

- **Problem with Original ASTE-Data-V1:**
  - Missing triplets where one opinion span applies to multiple targets.
  - Example: "Best service and atmosphere" should create TWO triplets: ("service", +, "Best") and ("atmosphere", +, "Best").
  - V1 dataset was incomplete, missing many such cases.
- **Solution: ASTE-Data-V2:**
  - Refined dataset including all missing triplets.
  - More complete and accurate for training/testing.
- **Four Datasets Used:**
  - 14Rest, 15Rest, 16Rest: Restaurant domain (different years).
  - 14Lap: Laptop domain.
  - All derived from SemEval competition data [14, 12, 13].

4.2 Baseline Models - The Competition

All baselines use pipeline approaches:

1. RINANTE+: LSTM-CRF → two-stage matching with classifier.
2. CMLA+: Attention mechanism → two-stage matching [15].
3. Li-unified-R: Multi-layer LSTM → two-stage matching [8].
4. Peng et al. (2019): Graph Convolutional Networks → two-stage matching [10].

Key Difference: Baselines extract targets/opinions separately, then match them in a second stage. JET does everything jointly in one step.

4.3 Experimental Setup

- **Technical Configuration:**
  - Word Embeddings: 300d GloVe [11] + 100d offset embeddings.
  - Architecture: BiLSTM with hidden size 300.
  - Advanced Version: Added BERT contextualized embeddings [1].
  - Training: 20 epochs max, Adam optimizer [6].
- **Evaluation Metric:**
  - **Strict F1 Score:** Triplet only counts as correct if ALL three components are exactly right: Target boundaries + Opinion boundaries + Sentiment polarity.

4.4 Main Results - Key Findings

1. **JET vs Pipeline Baselines:**
   - Both JET^t and JET^o significantly outperform all pipeline methods.
   - Best improvement: ∼7 F1 points over strongest baseline.
   - **Why it works:** Joint modeling captures target-opinion-sentiment interactions better than sequential approaches.
2. **Effect of Maximum Offset Distance (M):**
   - **Pattern:** F1 score increases as M increases (optimal around M=5-6).
   - **Explanation:** Larger M allows model to connect targets and opinions that are farther apart.
   - Example: If target-opinion distance is 4 words but M=2, model can't connect them.
3. **Performance Characteristics:**
   - JET's Strength: +15 precision points improvement (fewer wrong predictions).
   - Trade-off: Maintains acceptable recall scores.
   - Result: Better overall F1 scores due to much higher precision.
4. **Dataset Performance Differences:**
   - Better: 14Rest, 16Rest datasets.
   - Lower: 14Lap, 15Rest datasets.
   - Cause: Sentiment distribution differences between training and test sets.
5. **State-of-the-Art Results:**
   - JET + BERT: Achieves new best performance on all four datasets.
   - Best Model: JET^o (M=6) + BERT with highest F1 scores.

Core Success Factors:

1. Joint extraction eliminates error propagation from multi-stage pipelines.
2. Position-aware tagging directly encodes target-opinion relationships.
3. Flexible distance modeling (M parameter) captures various target-opinion separations.
4. Higher precision through better relationship modeling.

5 Analysis

5.1 Robustness Analysis - Testing Model Limits

5.1.1 Length Analysis

What they tested: Model performance on different lengths of:

- Targets: 1 word ("food") vs 3 words ("vegan food options")
- Opinion spans: 1 word ("great") vs 4 words ("absolutely amazing quality")
- Offsets: Distance 1 (adjacent) vs distance 6 (far apart)

Key Findings:

- JET^o strengths: Short targets ($\leq$3 words), single-word opinions, long distances (offset $\geq$4).
- Both models struggle with: Longer spans (performance drops as length increases).
- **Why longer spans are harder:** More words = more boundary decisions = more opportunities for errors in marking where spans begin/end.

5.1.2 Evaluation Method Analysis

Three evaluation strictness levels:

1. (T, O, S): Everything must be EXACTLY correct (main results).
2. (T_p, O, S): Target boundaries can OVERLAP with correct answer.
3. (T, O_p, S): Opinion boundaries can OVERLAP with correct answer.

Key Finding:

- JET^t: Better at exact opinion span boundaries (because it explicitly tags targets).
- JET^o: Better at exact target span boundaries (because it explicitly tags opinions).
- **Pattern:** Each model excels at what it doesn't directly tag.

5.2 Qualitative Analysis - Real Examples

Example 1: "Food is fresh and hot ready to eat" Gold triplets: ("Food", +, "fresh") AND ("Food", +, "hot ready")

Model Predictions:

- Peng et al.: Predicts wrong opinion span "hot ready"
- JET^t: Only finds 1 triplet (LIMITATION: can't handle same target → multiple opinions)

- JET^o: Gets both triplets correctly (STRENGTH: designed for this case)

Example 2: "with a quaint bar and good food" Gold triplets: ("bar", +, "quaint") AND ("food", +, "good")

Model Predictions:

- Peng et al.: Finds all spans but connects them wrong (each target to both opinions)
- JET^t & JET^o: Both get exact connections right

Key Insight: Pipeline methods struggle with target-opinion matching, while JET's joint approach gets connections right.

5.3 Ablation Study - Component Importance

Testing what happens when removing components:

1. **+char embedding:** Add character-level word features
   - **Result:** Minimal improvement (data too sparse).
2. **-offset features:** Remove $f_r(j, k)$ distance scoring
   - **Impact:** F1 drops significantly (especially JET^t).
   - **Conclusion:** Distance information is the "glue" connecting targets and opinions.
3. **-opinion features:** Remove f_o(opinion_span) quality scoring
   - **Impact:** F1 drops consistently across all models.
   - **Conclusion:** Opinion quality assessment is essential.

Key Takeaway: Both distance modeling AND opinion quality scoring are critical for performance.

5.4 Ensemble Analysis - Combining Complementary Strengths

5.4.1 The Complementary Models Concept

- JET^t STRENGTH: Different targets $\rightarrow$ same opinion.
- JET^t WEAKNESS: Same target $\rightarrow$ different opinions.
- JET^o STRENGTH: Same target $\rightarrow$ different opinions.
- JET^o WEAKNESS: Different targets $\rightarrow$ same opinion.
- **Hypothesis:** Combining both should capture more triplets.

5.4.2 Two Ensemble Methods

1. **JET$^{o\to t}$:** Start with JET^t predictions, add non-overlapping JET^o predictions.
2. **JET$^{t\to o}$:** Start with JET^o predictions, add non-overlapping JET^t predictions.

Overlap Definition (for avoiding duplicates): Two triplets overlap if BOTH conditions are true:

- Their target spans share word positions AND
- Their opinion spans share word positions.

Results:

- **Recall improvement:** Ensemble finds more total correct triplets by combining each model's strengths.
- **Mixed F1 performance:** Works well on 3/4 datasets, struggles on 16Rest.
- **Confirmation:** The models DO complement each other as hypothesized.

Core Analysis Insights:

1. Model specialization: Each JET variant excels at different triplet patterns.
2. Component importance: Distance and opinion quality features are both essential.
3. Complementary design: JET^t and JET^o have opposite strengths/weaknesses.
4. Ensemble potential: Combining models can improve recall by capturing missed triplets.

6 Related Work - Where ASTE Fits in the Research Landscape

This section explains how ASTE relates to other sentiment analysis tasks. Think of it as a family tree of research problems:

6.1 The Big Parent Field: Aspect-Based Sentiment Analysis (ABSA)

Goal: Understand sentiment toward specific aspects/targets in text

- Example: "The food was great but service was terrible"
- ABSA asks: What aspects? What sentiments toward each?

6.2 The Task Family Tree

1. **Target Extraction (The Simplest Child):**
   - **Goal:** Just find the targets/aspects being discussed.
   - **Example:** "food", "service".
   - **Methods:** Sequence labeling with CRF.
   - **What's missing:** No sentiment, no opinion words.
2. **Aspect Sentiment Analysis (The Middle Child):** Two sub-versions:
   - **Version A:** Given a target, predict its sentiment.
     - Input: "food" + sentence → Output: positive.
     - Methods: Neural networks, attention mechanisms.
   - **Version B:** Find targets AND their sentiments jointly.
     - Output: ("food", positive), ("service", negative).
     - Methods: Sequence labeling approaches.
3. **Target-Opinion Co-extraction (The Advanced Child):**
   - **Goal:** Find targets AND opinion words that express sentiment.
   - **Example:** ("food", "great"), ("service", "terrible").
   - **Methods:** Sequence labeling.
   - **What's missing:** Sentiment polarity isn't explicitly extracted here. You know what words express the opinion, but not if it's positive, negative, or neutral.
4. **Aspect Sentiment Triplet Extraction (ASTE) - Our Current Focus (The Most Complex Child):**
   - **Goal:** Extract (Target, Sentiment, Opinion Span) triplets.
   - **Example:** ("food", +, "great"), ("service", -, "terrible").
   - **Why it's harder:** Combines all previous subtasks. It needs to:
     (a) Identify targets.
     (b) Identify opinion spans.
     (c) Determine sentiment for each target.
     (d) Correctly link each target, its sentiment, and its opinion span(s).
   - **Previous Methods for ASTE:**
     - **Pipeline Approaches (JET's Competition):** These typically break ASTE into multiple steps, e.g.:
       (a) Extracting aspect terms and their sentiments.
       (b) Extracting opinion terms.

(c) Then, a separate step to match or link these extracted components into triplets.

– **Issue with Pipelines:** Error propagation from one stage to the next. If the first stage misses an aspect, the later stages can't find its sentiment or opinion. They also struggle to model the inherent, complex relationships between all three elements of a triplet.

6.3 JET's Contribution in the Landscape

- **Joint Modeling:** JET stands out by performing all these steps simultaneously in an end-to-end fashion.
- **Position-Aware Tagging:** The key innovation is the novel tagging scheme which directly encodes the sentiment and the relative position of the opinion span within the tags themselves. This allows the model to learn the complex relationships between target, sentiment, and opinion span at once, rather than trying to link them post-hoc.
- **Efficiency:** Despite the complex task, the method remains computationally efficient due to the nature of span lengths and the CRF inference.

The "Related Work" section positions ASTE as a crucial and challenging task in sentiment analysis, and highlights how JET's joint, position-aware tagging approach addresses the limitations of previous pipeline methods, pushing the state-of-the-art.

7 Conclusion and Future Work

7.1 Summary of Contributions

- Introduced a novel, general **position-aware tagging scheme** for ASTE.
- The scheme effectively encodes **target entity, sentiment polarity, and opinion span offsets** into a single sequence tagging framework.
- Developed **JET (Jointly Extract the Triplets) model**, an end-to-end architecture leveraging this tagging scheme.
- Demonstrated **state-of-the-art performance** across multiple benchmark datasets, significantly outperforming existing pipeline methods.
- Provided **in-depth analysis** confirming the robustness across various span lengths and evaluation strictness levels.
- Showcased the **critical importance of distance and opinion quality features** through ablation studies.
- Illustrated the **complementary nature of JET^t and JET^o variants** and the potential for ensemble methods to improve recall.

7.2 Future Work

- **Handling Discontinuous Spans:** The current scheme assumes continuous target and opinion spans. Real-world text often has discontinuous phrases (e.g., ”service was, to be honest, not good”). Extending the tagging scheme to handle such cases would be a significant advancement.

- **Nested Entities/Opinions:** Exploring how to manage scenarios where opinion spans might themselves contain targets or other complex nested structures.

- **More Sophisticated Span Representations:** Investigating richer ways to represent opinion and target spans beyond simple concatenation or average pooling, potentially using graph neural networks or more advanced attention mechanisms [9, 17].

- **Multi-sentence Triplets:** Currently, the model operates sentence-by-sentence. Extending it to extract triplets where the target, sentiment, or opinion span might cross sentence boundaries (e.g., ”The food was delicious. I highly recommend it.”)

- **Few-shot/Zero-shot ASTE:** Adapting the model to work effectively with limited labeled data or in completely new domains without retraining, perhaps using meta-learning or transfer learning techniques.

- **Integration with Commonsense Knowledge:** Enhancing the model with external knowledge bases to better understand implicit sentiments or subtle opinion expressions.

- **Further Ensemble Exploration:** Developing more advanced or dynamic ensemble strategies [2] beyond simple non-overlapping additions to maximize the benefits of complementary models.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Thomas G Dietterich. Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

[3] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. In *Neural Networks*, volume 18, pages 602–610, 2005.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[5] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. In *arXiv preprint arXiv:1508.01991*, 2015.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[8] Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6714–6721, 2019.

[9] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1105–1116, 2016.

[10] Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. Knowing what, how and why to explain: A study on aspect-level sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5253–5260, 2019.

[11] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[12] Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, 2015.

[13] Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, and Mohammad Daoud. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, pages 19–30, 2016.

[14] Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, 2014.

[15] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 3316–3322, 2017.

[16] Lu Xu, Hao Li, Wei Lu, and Lidong Bing. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, 2020.

[17] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of*

the 55th Annual Meeting of the Association for Computational Linguistics, pages 1227–1236, 2017.