SANCL: Multimodal Review Helpfulness Prediction with Selective Attention and Natural Contrastive Learning

Anonymous ACL submission

Abstract

With the boom of e-commerce, Multimodal Review Helpfulness Prediction (MRHP) that identifies the helpfulness score of multimodal product reviews has become a research hotspot. Previous work on this task focuses on attention-based modality fusion, information integration, and relation modeling, which primarily exposes the following drawbacks: 1) the model may fail to capture the really essential information due to its indiscriminate attention formulation; 2) lack appropriate modeling methods that take full advantage of correlation among provided data. In this paper, we propose SANCL: Selective Attention and Natural Contrastive Learning for MRHP. SANCL adopts a probe-based strategy to enforce high attention weights on the regions of greater significance. It also constructs a contrastive learning framework based on natural matching properties in the dataset. Experimental results on two benchmark datasets with three categories show that SANCL achieves state-of-the-art baseline performance with lower memory consumption.

1 Introduction

We have witnessed an acceleration towards an e-commerce boom that has transpired over the past decades (Vulkan, 2020). In the virtual bazaar, countless deals are made between mutually invisible sellers and customers from time to time. For customers, it may be their biggest headache to determine whether they should pay for a good when being overwhelmed by tempting advertisements, as they can hardly learn about the true information about a product in face of the seller’s meticulous promotion without any references. In this situation, reviews in e-shops that can provide justification information, are thus of great value to customers. However, the quality of reviews under a certain product page can be disparate—many customers are willing to leave informative feedback on the product, while many others arbitrarily write a few words and even paste irrelevant messages in their comments. Therefore, from the perspective of online shopping platforms, they would be welcome and attractive to customers if they provide a service that can intelligently filter and place the most helpful reviews at the top position. The task in the machine learning field to solve this problem is Review Helpfulness Prediction (RHP) (Tang et al., 2013).

As the thriving of multimodal learning research and the handy accessibility of multimodal data in this Internet era, the latest progress incorporated image (vision modality) information into the review helpfulness prediction (RHP) (Liu et al., 2021) as Multimodal RHP (MRHP). Although previous work attained excellent results in MRHP, there are still some drawbacks. First, the attention mechanism in these works for representation learning follows the most basic setting—it directly computes the correlation, e.g., the similarity of feature vectors, among multimodal and multi-domain data is an essential factor in modeling (Xu et al., 2020; Chen et al., 2019). Nevertheless, existing studies (Xu et al., 2020; Liu et al., 2021) simply quantified them in similarity metrics, such as cosine value, for direct classification use. Though gained appreciative re-
sults, we believe they can be better utilized through the contrastive learning scheme to enhance the quality of learned representations.

In this paper, we propose a novel framework, SANCL, to incorporate these two points. In SANCL we first generate a special “probe” mask that highlights the key sentences from the product and review text. The mask then attends the computation attention modules to help focus more on those task-related sentences. Then we construct a contrastive learning framework to learn better modality representations with internal correlations of data. Based on contrastive predictive coding (CPC) (Oord et al., 2018), the framework is composed of two feature spaces (domains). Each domain takes specific combinations of projected representations as input, according to their relation types from our analysis. Through optimization over the contrastive score, the multimodal and multi-domain representations can learn from the inherent relations. Our contribution can be summarized as follows:

- We design a selective attention approach, including the probe mask generation and mask-based attention computation, for the information aggregation in MRHP tasks.

- We analyze the characteristics and relations in multimodal reviews and formulate a contrastive learning framework to refine the learned representations.

- Extensive experiments on three publicly available datasets show our approach achieves state-of-the-art performance with lower memory consumption.

2 Related Work

In this section, we briefly recap some relevant work in the field of review helpfulness prediction and multimodal contrastive learning.

Review Helpfulness Prediction Customer reviews play an important role in helping customers investigate products before determining whether to purchase. (Zhu and Zhang, 2010; Diaz and Ng, 2018; Gamzu et al., 2021). Support vector regression (SVM) was first employed to automatically judge the review helpfulness (Kim et al., 2006; Zhang and Varadarajan, 2006; Tsur and Rappoport, 2009). Later, linear regression (Lu et al., 2010; Ghose and Ipeirotis, 2010), extended tensor factorization (Moghaddam et al., 2012), and probabilistic matrix factorization models (Tang et al., 2013) have been applied to integrate complicated constraints into the learning process. With the development of deep learning, deep neural networks (Lee and Choeh, 2014; Fan et al., 2018; Chen et al., 2018) have been utilized to model the sophisticated elements in this task. Recently, Qu et al. (2020) proposed a graph neural network to capture the intrinsic relationship between the products and their reviews. However, most existing studies only focus on the text of reviews, neglecting the images that usually exist in online reviews. This paper takes advantage of the images and proposes a novel contrastive learning framework with a selective attention mechanism to learn expressive multimodal features.

Multimodal Representation Learning The foremost problem of multimodal tasks lies in multimodal representation learning (Baltrušaitis et al., 2018). The concept of multimodal representation learning covers many techniques, such as multimodal fusion (Vielzeuf et al., 2018; Wang et al., 2020; Mai et al., 2020; Han et al., 2021a), multimodal contrastive learning (Yuan et al., 2021; Han et al., 2021b), etc. Attention-based architectures are the basic routine in multimodal fusion, but the formulations are similar. In this paper, knowing about the particularity of MRHP and its dataset, we devise a novel attention mechanism to better aggregate information in textual data. Additionally, we also upgrade the application of contrastive learning. Unlike the ordinary treatment that divides samples into positive and negative groups according to “from myself” or “not from myself” (Cui et al., 2020; Liang et al., 2020), we extract contrastive pairs according to the natural correlation in the dataset and construct the framework of two feature spaces termed as domains.

3 Method

In this section, we first introduce the problem definition of Multimodal Review Helpfulness Prediction (MRHP). Then we elaborate on the model architecture and processing pipeline of our method.

3.1 Problem Definition

Given a collection of product descriptions $\mathcal{P} = \{P_1, P_2, ..., P_N\}$ and associated reviews $\mathcal{R} = \{R_1, R_2, ..., R_N\}$ gleaned from an e-shopping
our high quality, handmade, 100%…
I would rate zero stars if possible. The product I received was nothing like what was pictured …

Figure 1: The overview of SANCL. The output layer is omitted. Features in red boxes \((S_t^{ii}, S_t^{ii}, S_t^{pr}, S_t^{pr})\) are used in final helpfulness score prediction.

3.2 Overview
The overall architecture of SANCL is depicted in Figure 1. We first generate a probe mask for each review according to the corresponding product name and review text as shown in Figure 2. The probe mask highlights the sentences that mention the product, which then participates in the computation of selective attention to produce text representations. For images, we feed the features extracted by pre-trained visual neural networks to two self-attention modules to produce image representations. Then we project these representations of each modality in both product description and customer review into two shared spaces (domains). We finally develop a contrastive learning module to compute the cross-modality and review-product contrastive scores, which further improves the quality of representations output from attention modules.

3.3 Input Encoding

Context-aware Textual Representation For both review and product text, we initialize the token representations with GloVe (Pennington et al., 2014)\(^1\) or pre-trained models as \(E_t = \{e_1^t, e_2^t, ..., e_l^t\} \in \mathbb{R}^{l \times d_{v}}\), where \(l\) is the length (number of tokens) of a given sentence and \(d_{v}\) is the embedding dimension. We then send these embeddings to a uni-directional Gated Recurrent Unit (GRU) (Cho et al., 2014), yielding token-wise and sequence representations \(H_t = \{h_1^t, ..., h_l^t\}\) and \(h_t^{seq}\):

\[
H_t, h_t^{seq} = \text{GRU}(E_t; \theta_t).
\]

where \(\theta_t\) is the parameters in GRU.

Visual Feature Extraction We apply Faster R-CNN (Ren et al., 2015) on raw images and yield the hidden representations \(E_v = \{e_1^v, e_2^v, ..., e_n^v\} \in \mathbb{R}^{n \times d_{v}}\) in the last layer before the classifier to map the Regions of Interest (RoI) in an image to a hidden space, where \(n\) is the number of hot regions detected in the image and \(d_{v}\) is the vector lengths of hidden representations. Then same as Liu et al. (2021), we feed them into a self-attention module that outputs the encoded image representations \(H_v = \{h_1^v, h_2^v, ..., h_n^v\}\).

\(^1\)We used glove.840B.300d in our experiments.
3.4 Probe-based Selective Attention (PSA)

Having gained elementary encoded multimodal representations, interactions between parallel review pieces and corresponding product descriptions are required to form the product-aware review representations. Previous work primarily formulated these interactions as token-wise description–review attention (Fan et al., 2019; Qu et al., 2020). Though succinct and effective, this token-by-token or sentence-by-sentence computation scheme may neglect distinct the relative importance among sentences and missing really task-related information. Because only through loss back-propagation without any re-weighting operation, it can not always be ensured that larger weights will be put on those key sentences. To mitigate this issue, we propose the selective attention approach. It first generates a special “probe” mask then performs discriminate attention based on that.

**Probe Mask Generation** The probe mask should reflect the position (i.e., in which sentence) where the product is mentioned in a review. An example of the generation process is displayed in Figure 2. We first retrieve the core words from the product name by looking up its dependency tree and picking the lemmatized form of the words around the root. Next, we use coreference resolution to identify all coreference clusters in the review.

![Figure 2: An example of mask generation](image)

There are three possible resolution results: (1) A cluster containing the core word of the product name; (2) At least one cluster exists but the core word is missing in all clusters; (3) No coreference cluster exists. For (1) and (3), we do not require extra steps as the existence of entity clusters can be confirmed. For (2), we are still uncertain whether an entity cluster is in the text. We devise a simple rule to tackle this situation—we regard the first cluster as product name mention cluster, based on our observation that the first repeatedly mentioned pronouns in a review are more likely to refer to the product. After locating these product name mentions, we create the probe mask \( M \in \mathbb{R}^{1 \times l} \) by assigning 1 to the positions of those mentioned sentences and 0 to others. The process is summarized in Algorithm 1.

**Algorithm 1: Probe Mask Generation**

| Input: Review sentences \( R \), product name \( P \) |
| Output: Probe mask \( M \) |

\[
\begin{align*}
\text{# core words and coreference clusters extraction:} & & & \hat{R} \leftarrow \text{Lemmatise}(R), \hat{P} \leftarrow \text{Lemmatise}(P); \\
\text{# core words extraction:} & & & T \leftarrow \text{DependencyParse}(\hat{P}); \\
& & & W \leftarrow \text{FindWordsNearRoot}(T); \\
\text{clusters} & & & \leftarrow \text{FindCoreferenceCluster}(\hat{R}) \\
\text{mask generation:} & & & M \leftarrow \text{ZeroInit}(R, \text{size}) \\
\quad \text{if} & C = \emptyset & & \text{return} M \\
\quad & \text{foreach} c \in \text{clusters} & & \text{do} \\
\quad & \quad & & \text{if} w \in W \text{ in } c \text{ then} \\
\quad & \quad & & \quad \text{gold_cluster} = c \\
\quad & \quad & & \text{end} \\
\quad & \quad & & \text{if} \text{gold_cluster} = \emptyset \text{ then} \\
\quad & \quad & & \quad \text{gold_cluster} = \text{clusters}[0] \\
\quad & \quad & & \text{end} \\
\quad & \quad & & \text{foreach} \text{sent} \in \hat{R} \text{ do} \\
\quad & \quad & & \quad \text{if} w \in \text{gold_cluster in } \text{sent} \text{ then} \\
\quad & \quad & & \quad \quad M[\text{sent.start: sent.end}] \leftarrow \text{True} \\
\quad & \quad & & \text{end} \\
\quad & \text{end} \\
\text{end} \\
\text{return } M
\end{align*}
\]

**Selective Attention with Probe Mask** There are three steps to acquire product-aware review representations—self-attention, cross-text attention, and pooling, among which the first and last steps take advantage of probe masks generated. We first transform the probe mask to a new form:

\[
M' = \alpha M + \beta (1 - M), \quad (2)
\]

where \( \alpha > \beta > 0 \) since we expect the mask could help focus more on the sentences where the product is mentioned. This effect embodies in the self attention computation of the review text \( \mathbf{H}_r \), where the fundamental attention weights are computed as:

\[
\mathbf{A} = \text{softmax}(\mathbf{W_H}_r), \quad (3)
\]

We renew the original attention matrix \( \mathbf{A} \in \mathbb{R}^{l \times l} \):

\[
\mathbf{A}' = (M')^T M' \odot \mathbf{A}, \quad (4)
\]
where $W_{m,s,i}$ and $b_{m,s,i}$ are weights and biases in the $i$-th layer of the projection network. Note that the data in the same modality and domain share the same network parameters. In the succeeding content, we are going to describe details of the two contrastive-learning domains, mainly concerning how to pick positive and negative samples for contrastive learning and training.

**Inner Instance (II) Domain** In the inner instance domain, we separate positive and negative pairs according to how similar the representations between image and text are in a single training instance. First, from the sellers’ perspective, the text and image of a product should match well so as to attract customers. Thus we mark text-image pairs of product descriptions as positive ones (the set of these pairs is denoted as $S^+_t$). Therefore, we mark the former as positive (the set is denoted as $S^+_i$) and the latter as negative ones (the collection is denoted as $S^-_i$). Besides, from our observation, reviews that achieve high helpfulness scores possess a high similarity between its text and the attached image.

**Product-Review (PR) Domain** The semantic matching property also exists between product descriptions and their associated reviews. As helpfulness is dependent on how well a review is pertinent to the theme of the product, we argue that review pieces of high helpfulness scores ($S^+_{pr}$) should match the product introduction both visually and literally, while those low-score pieces ($S^-_{pr}$) match the introduction poorly in both modalities.

**Multi-domain Contrastive Predictive Coding (MCPC)** In contrastive predictive coding (Oord et al., 2018), we need to compute contrastive scores for every sample pair. According to the common approach (Yuan et al., 2021; Han et al., 2021b), exponential function is chosen as the score function:

$$
\varphi(A, B) = \exp \left( \frac{\text{norm}(A^T)\text{norm}(B)}{\tau} \right),
$$

where $\text{norm}(\ast)$ is the $l_2$-norm function, $\tau$ is the temperature hyper-parameter, for simplicity we keep its value 1.0 in our experiments. By noise contrastive estimation (Gutmann and Hyvärinen, 2010), in the inner instance domain the score is
We select all review-related representations where \( r \) and \( W \) are the text-image pair from the instance, i.e., a review piece or product description. The summation is over \( S_{ii}^t \) and \( S_{ii}^p \) because instances counted here are from both product descriptions and review pieces. Similarly in product-review domain the score is:

\[
c_{pc}^{m}_{pr} = - \sum_{S_{m,j} \in S_{pr}^t} \log \frac{\varphi(S_{m,j}^t, S_{m,j}^p)}{\sum_{S_k \in (S_{pr}^t \cup S_{pr}^p)} \varphi(S_{m,k}^t, S_{m,k}^p)} \quad (12)
\]

where \( S_{m,j}^r \) is the representation of modality \( m \) in review \( r \) from the positive review set \( S_{pr}^t \) and \( S_{m,j}^p \) is the counterpart of the corresponding product.

### 3.6 Prediction and Training

We select all review-related representations in the common spaces of two domains \((S_{ii}^{t,r}, S_{ii}^{t,r}, S_{ii}^{t,r}, S_{ii}^{t,r})\) and concatenate them as features for final prediction (F). After concatenating these features, a linear layer takes them as input and outputs the helpfulness score predictions \( \xi_r \):

\[
F = \text{concat}([S_{ii}^{t,r}, S_{ii}^{t,r}, S_{ii}^{t,r}, S_{ii}^{t,r}]) \quad (14)
\]

\[
\xi_r = W \xi + b \quad (15)
\]

where \( W \) and \( b \) are the weight matrix and bias in the output layer. Same as Liu et al. (2021), we adopt the standard pairwise ranking loss as the task loss:

\[
L_{task} = \sum_i \max(0, \gamma - \xi_{r,+} + \xi_{r,-}) \quad (16)
\]

where \( r^+, r^- \) are an arbitrary pair of review pieces under product \( P_t \), \( \gamma \) is a scaling factor. Contrastive losses make up the auxiliary loss:

\[
L_{aux} = c_{pc}^{i} + c_{pc}^{pr} \quad (17)
\]

Hence the total loss for training is (\( \kappa \) is a hyper-parameter to adjust the effect of auxiliary loss):

\[
L = L_{task} + \kappa L_{aux} \quad (18)
\]
Table 1: Results on three datasets; all reported metrics are the average of five runs; "∗" are from the open-source code in Liu et al. (2021); "†" represent the results significantly outperforms PRHNet and MCR with p-value < 0.05 based on paired t-test.

In both settings, we also test our method with BERT (Devlin et al., 2018) encoder and compare that to the respective SOTA models on BERT. In addition, we test and record basic BERT performance (BERT+a double linear layers).

### 4.3 Metrics

As MRHP is a ranking task, the metrics for comparison are as well ranking-customized. After sorting all prediction-truth scores in descending order, the Mean Average Precision (MAP) computes the mean precision of top-1 to top-K samples. K is usually large enough to ensure top-K can encompass the entire collection of reviews under every product. The Normalized Discounted Cumulative Gain (NDCG-N) (Järvelin and Kekäläinen, 2017; Diaz and Ng, 2018) purely reckons the gain value over top-N predictions (N is 3 and 5 in our experiments), which simulates the real circumstances of a typical customer who would always read the topmost reviews.

### 5 Results and Analysis

In this section, we will compare our approach with several advanced baselines and explore how it improves in the multimodal helpfulness prediction task.

#### 5.1 Performance Comparison

We list the performance of our model and baselines in Table 1. Notably, SANCL consistently outperforms all the baselines in both text-only and multimodal, BERT and Glove initialization settings. These outcomes initially demonstrate the efficacy of our method in MRHP tasks. It is surprising that we cannot gain significant performance boost by replacing Glove with BERT as the text encoder. We speculate the reason is that Glove embeddings are expressive enough for this task.

Moreover, it can be claimed that SANCL is a lightweight model compared to the multimodal SOTA, since the model size and GPU memory consumption of SANCL are much lower than MCR. The total number of parameters is 2.63M in MCR and 1.41M (exclude the embedding layer) in SANCL, which indicates a double efficiency. The average GPU memory usage of SANCL during the training on Amazon-MRHP Home & Kitchen is around 2.4G, while MCR occupies an average of 13.7G GPU memory during training, which is 4.7 times higher than SANCL.

#### 5.2 Ablation Study

To verify the benefits of our proposed method, we carry out comprehensive ablation experiments on the Amazon electronics dataset, including the selective attention and contrastive learning components. In selective attention, we first replace learned attentive weights with a fixed value of 0.5, since we find most β values in our experiments are around 0.5. Next, we remove the entire selective attention module and only preserve the primitive attention computation. The
### Table 2: Ablation study of SANCL on the Electronics dataset.

<table>
<thead>
<tr>
<th>Description</th>
<th>MAP</th>
<th>N-3</th>
<th>N-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SANCL</td>
<td>56.19</td>
<td>46.98</td>
<td>49.92</td>
</tr>
<tr>
<td>Attention</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o learned $\beta$ (fixed at 0.5)</td>
<td>55.61</td>
<td>46.37</td>
<td>49.58</td>
</tr>
<tr>
<td>w/o probe mask</td>
<td>55.43</td>
<td>46.11</td>
<td>49.45</td>
</tr>
<tr>
<td>Contrastive learning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o $cpc_{ii}$</td>
<td>55.54</td>
<td>46.29</td>
<td>49.23</td>
</tr>
<tr>
<td>w/o $cpc_{pr}$</td>
<td>55.81</td>
<td>46.40</td>
<td>49.47</td>
</tr>
<tr>
<td>w/o $cpc_{ii}$ and $cpc_{pr}$</td>
<td>55.35</td>
<td>46.28</td>
<td>49.09</td>
</tr>
</tbody>
</table>

### 5.3 Case Study

To understand how our model deals with samples in-depth, we randomly picks up a test product-review instance from the test set of Amazon Home & Kitchen to explain how SANCL works and, as shown in Table 3.

In this example, the customer bought the pins to fix the edge of his sofa. Instead of photoing pins themselves, the customer only presented the tidy sofa after installing the pins. We first visualize the attention weights in test time, as shown in Fig. 3. Note that only the first sentence in the review contains the elements in the co-reference clusters, which we have emphasized with italics and underline in Table 3. Consistently, we observe the significant larger weights in the region of first sentence (row/column 1-19) while the rest region’s weights are much smaller. We also ran MCR and collect its prediction on this example, and it is clear that MCR commits a severe error here, probably caused by the direct classification on the unimodal cosine similarity. In our approach, as we carefully analyze and classify the positive and negative pairs in the multi-domain contrastive learning framework, the huge semantic similarity between review text and image and between product description and review text, indicated by the high CPC scores $S_{ii}^{cv}$, assists the model to correctly predict the score.

### Table 3: Examples from the Amazon Home & Kitchen test set.

| Product Name: Twisty Pins for Upholstery, Slipcovers and Bedskirts 50/pkg |
| Review (Helpfulness Score: 4): I bought these to pin the loose material on a sofa cover and they worked like a charm. The sofa cover definitely looks form fitting now. |
| Predictions: SANCL: 4.5291 MCR: -1.0832 |
| CPC score: $cpc_{ii} = 0.82$, $cpc_{pr} = 0.76$, $cpc_{pr}^{v} = 0.21$ |

### 6 Conclusion

We propose a novel framework, SANCL, for the task of multimodal review helpfulness prediction (MRHP) in this paper. We first present a selective attention mechanism, which purposefully aggregates information from these crucial sentences in the review text by generating the probe mask that exerts re-normalization on the attention weights and pooling stage. We then build up a multi-domain natural contrastive learning framework in our model. It exploits the natural relations among the data from different fields and modalities in the dataset to enhance the model’s capacity of multimodal representation learning. Results of comprehensive experiments and analyses demonstrate the superiority of our model over the comparable baselines and the efficacy of the novel components.
References


A Dataset Specification

Specifications of the two datasets are in Table 4 and 5 below.

<table>
<thead>
<tr>
<th></th>
<th>Amazon-MRHP (Products/Reviews)</th>
<th>Lazada-MRHP (Products/Reviews)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
<td><strong>Cloth. &amp; Jew.</strong></td>
<td><strong>Elec.</strong></td>
</tr>
<tr>
<td><strong>Train</strong></td>
<td>12074/277308</td>
<td>10564/240505</td>
</tr>
<tr>
<td><strong>Dev</strong></td>
<td>3019/122148</td>
<td>2641/84402</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>3966/87492</td>
<td>3327/9750</td>
</tr>
</tbody>
</table>

Table 4: Statistics of the Amazon-MRHP dataset.

<table>
<thead>
<tr>
<th></th>
<th>Amazon-MRHP (Products/Reviews)</th>
<th>Lazada-MRHP (Products/Reviews)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
<td><strong>Cloth. &amp; Jew.</strong></td>
<td><strong>Elec.</strong></td>
</tr>
<tr>
<td><strong>Train</strong></td>
<td>6596/104093</td>
<td>3848/41828</td>
</tr>
<tr>
<td><strong>Dev</strong></td>
<td>1649/26139</td>
<td>963/10565</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>2062/32274</td>
<td>1204/12661</td>
</tr>
</tbody>
</table>

Table 5: Statistics of the Lazada-MRHP dataset.

B Hyperparameter Search

The optimal hyperparameter settings are provided in Table 6 and 7.

<table>
<thead>
<tr>
<th></th>
<th>Amazon-MRHP Hyperparameters</th>
<th>Lazada-MRHP Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
<td><strong>Cloth. &amp; Jew.</strong></td>
<td><strong>Elec.</strong></td>
</tr>
<tr>
<td><strong>learning rate</strong></td>
<td>$1e^{-4}$</td>
<td>$5e^{-5}$</td>
</tr>
<tr>
<td><strong>text embedding dim</strong></td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td><strong>text embedding dropout</strong></td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>image embedding dim</strong></td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td><strong>LSTM hidden dim</strong></td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td><strong>shared space hidden</strong></td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td><strong>$\kappa$</strong></td>
<td>0.25</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>batch size</strong></td>
<td>32</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 6: Hyperparameters for all categories using glove-300d embeddings.

Table 7: Hyperparameters for all categories using BERT as encoder.

C Language Tools

For coreference resolution, we use neuralcoref, an extension that can be placed on SpaCy processors. For BERT model, we use the huggingface transformers package to load.