A Detailed Settings for the Experiments in Section 2.1

A.1 Dataset Description

We summarize the statistics of the datasets used in Section 2 in Table 5.

Dataset	Training	Testing	# classes
SNLI	549k	10k	3
MultiNLI	393k	10k	3
QuoraQP	384k	10k	2
MSRP	4k	2k	2
SICK	5k	5k	2/3
ByteDance	256k	32k	3

Table 5: Information about the datasets.

For SICK, both `entailment_label` and `relatedness_score` are provided. We use the sentence pairs with `relatedness_score` greater than 3.6 as `duplicated`, and otherwise `not_duplicated`. This threshold gives roughly 50% of positive pairs and 50% negative pairs.

For ByteDance, since no existing dataset partition is available, we randomly divide the dataset into a training set, a validation set, and a testing set in a ratio of 8:1:1. We use the sentences in English during our experiments.

A.2 Features Used in Unlexicalized

We list the 15 features we used in method **Unlexicalized** in Section 2.1. We use 3 types of unlexicalized features (Bowman et al., 2015):

- The BLEU score of both sentences, using n-gram length from 1 to 4, which are totally 4 features.
- The length difference between the two sentences, as one real-valued feature.
- The number and percentage of overlap words between both sentences over all words and over just nouns, verbs, adjectives and adverbs, which are totally 10 features.

A.3 Features Used in Advanced

We list the features we used in method **Advanced** in Section 2.1. As mentioned above, if we use a node to represent a sentence and add an undirected edge if two sentences are compared in the dataset, the whole dataset can be viewed as a graph as illustrated in Figure 3. To classify the edges in the graph, we use 3 types of graph-based features:

- The origin and extended leakage features: degrees of both nodes, number of 2-hop and 3-hop paths between the two nodes, number of 2-hop and 3-hop neighbors of both nodes, which are totally 8 features.
- The element-wise product and dot product of Deepwalk (Perozzi et al., 2014) embedding of the two nodes, all together as 65 features.
- The resource allocation index, Jaccard coefficient, preferential attachment score and Adamic-Adar index (Zhou et al., 2009; Liben-Nowell and Kleinberg, 2007) of both two nodes, which are totally 4 features.

B Proof for the Theorems

B.1 Derivation of Equation (1)

Here we present the derivation of Equation (1).

Proof.

$$
\begin{aligned}
&P_{\widehat{\mathscr{D}}}(Y=1|l)\\
&=P(Y=1|S=Y,l)\\
&=\frac{P(Y=1,S=1|l)}{P(Y=1,S=1|l)+P(Y=0,S=0|l)}\\
&=\frac{P(Y=1|l)P(S=1|l)}{P(Y=1|l)P(S=1|l)+P(Y=0|l)P(S=0|l)}\\
&=\frac{P(Y=1)P(S=1|l)}{P(Y=1)P(S=1|l)+P(Y=0)P(S=0|l)}.
\end{aligned}
$$

By solving the above equation, we have the result in Equation (1). □

B.2 Proof of Theorem 1

Here we present the proof for Theorem 1, *i.e.*, the unbiased expectation theorem.

Proof.

$$
\begin{aligned}
&E_{x,y,l\sim\widehat{\mathscr{D}}}\Big[w\Delta\big(f(x,l),y\big)\Big]\\
=&\int\frac{P(S=Y)}{P(S=y|l)}\Delta(f(x,l),y)dP_{\widehat{\mathscr{D}}}(x,y,l)\\
=&\int\Delta(f(x,l),y)\frac{P(S=Y)}{P(S=y|l)}dP(x,y,l|S=Y)\\
=&\int\Delta(f(x,l),y)\frac{P(S=Y)}{P(S=y|l)}\frac{P(S=y|x,y,l)dP(x,y,l)}{P(S=Y)}\\
=&\int\Delta(f(x,l),y)dP(x,y,l)\\
=&E_{x,y,l\sim\mathscr{D}}\Big[\Delta\big(f(x,l),y\big)\Big].
\end{aligned}
$$

□

As illustrated above, by adding specific weights to the samples, we can obtain the loss unbiased to the leakage neutral distribution $\mathscr{D}$. The unbiased loss can be used for both training and evaluation.