Semantic Sentence Matching via Interacting Syntax Graphs

Chen Xu¹, Jun Xu¹,²*, Zhenghua Dong³, Ji-Rong Wen¹,²
¹Gaoling School of Artificial Intelligence, Renmin University of China
²Beijing Key Laboratory of Big Data Management and Analysis Methods
³Huawei Noah’s Ark Lab
{xc_chen, junxu, jrwen}@ruc.edu.cn
dongzhenhua@huawei.com

Abstract

Studies have shown that the sentence’s syntactic structures are important for semantic sentence matching. A typical approach is encoding each sentence’s syntactic structure into an embedding vector, which can be combined with other features to predict the final matching scores. Though successes have been observed, embedding the whole syntactic structures as one vector inevitably overlooks the fine-grained syntax matching patterns, e.g. the alignment of specific term dependencies relations in the two inputted sentences. In this paper, we formalize the task of semantic sentence matching as a problem of graph matching in which each sentence is represented as a directed graph according to its syntactic structures. The syntax matching patterns (i.e. similar syntactic structures) between two sentences, therefore, can be extracted as the sub-graph structure alignments. The proposed method, referred to as Interacted Syntax Graphs (ISG), represents two sentences’ syntactic alignments as well as their semantic matching signals into one association graph. After that, the neural quadratic assignment programming (QAP) is adapted to extract syntactic matching patterns from the association graph. In this way, the syntactic structures fully interact in a fine granularity during the matching process. Experimental results on three public datasets demonstrated that ISG can outperform the state-of-the-art baselines effectively and efficiently. The empirical analysis also showed that ISG can match sentences in an interpretable way.

1 Introduction

Matching two natural language sentences has become a fundamental technique in information retrieval (IR) and natural language processing (NLP). Extensive research efforts have been devoted to the task (Li and Xu, 2014; Xu et al., 2020). Recently, researchers found that the sentences’ rich syntactic information helps match. Some studies utilize the implicit syntax-encoding methods to learn the sentence embedding based on its syntactic structures (Chen et al., 2016, 2017). Other studies directly utilize the different syntactic tags as syntactic features and fuse them with word features (Mou et al., 2016; Chen et al., 2018; Liu et al., 2020).

All these approaches separately represent the syntactic information of the two input sentences as two coarse-grained embedding vectors, ignoring the fine-grained matching patterns between the syntactic structures.

In sentence matching, The fine-grained syntax structures are lost during the process of representing the syntactic information as a vector. Studies in (Xu et al., 2020) also verified that representation-based methods will lose the fine-grained matching signals. Figure 1 illustrates two semantically similar sentences from the SNLI dataset where sentences $X$ = “Some men are playing a sport”, $Y$ = “A soccer game with multiple males playing”). The words and POS tags are shown in the nodes, and syntactic dependencies are shown as edges.

Figure 1: A semantically similar sentence pair ($X$ = “Some men are playing a sport”, $Y$ = “A soccer game with multiple males playing”) from SNLI. The words and POS tags are shown in the nodes, and syntactic dependencies are shown as edges.

We can see that though the overall syntactic struc-
turers of the two sentences are very different, they still contain matched sub-graphs: “men(NNS) \text{nsubj} \text{playing}(VBG) \text{obj} \text{sport}(NN)” from $X$ and “game(NN) \text{obj} \text{playing}(VBG) \text{nsubj} \text{males}(NNS)” from $Y$. The matched sub-graphs provide a crucial syntactic matching signal for downstream models to make high-precision matching decisions. Existing approaches, however, inevitably overlook the fine-grained sub-graph matching signal because the sub-graph details are lost during the embedding of the whole syntax graphs. Specifically, the sub-graph details denote the combination of first-order (e.g. Part-of-Speech) and second-order (e.g. word dependency) syntactic information.

In this paper, we propose to formalize the task of sentence matching as matching two directed graphs where each graph corresponds to one sentence, and its edges and nodes correspond to the word-word syntactic dependencies and the words’ syntax and semantic information, respectively. The sentence matching, therefore, becomes a process of first extracting the sub-graph structure alignments and then summarizing them into the final matching score.

A neural model called Interacted Syntax Graphs (ISG) is developed for conducting the matching. Specifically, given two sentences, ISG employs a pre-trained language model (PLM) and a syntactic parser to get the word embeddings and the syntactic structures as the initialization features. Then, ISG fuses these semantic and syntactic features into an associate graph. Sub-graph matching patterns can be extracted based on the associate graph structure, by using Lawler’s quadratic assignment programming (QAP) (Cho et al., 2010). Finally, a matching classifier is used to merge the matching patterns and semantic vectors outputted from the PLM, resulting in the final matching score.

ISG offers several advantages, including accurate extraction and fusion of the syntactic graph matching signals, ease in interpretation, and high matching accuracy. The contributions of this paper can be summarized as follows: (1) We highlight the importance of the fine-grained matching patterns from syntax graphs in semantic sentence matching. A novel matching model called ISG is proposed; (2) Experimental results based on three available benchmarks showed that the matching accuracy of ISG outperformed the state-of-the-art baselines; (3) Analysis showed that ISG can discriminate important syntactic and semantic matching patterns in an interpretable way.

2 Related Work

Machine learning models have been widely used for matching natural language sentences (Li and Xu, 2014; Xu et al., 2020). Among them, the representative methods include DSSM (Huang et al., 2013) and its extensions (Wang et al., 2017a; Shen et al., 2014; Kim et al., 2019; Yang et al., 2019). Representative interaction-based models include ARC-II (Hu et al., 2014), MatchPyramid (Pang et al., 2016), etc. Recently, the pre-trained language model has been adapted to conducting matching (Devlin et al., 2019; Liu et al., 2019). These models always focused on superficial matching signals and ignore explicit NLP knowledge.

Recently, there is a trend to utilize explicit NLP knowledge to improve sentence representation. For example, HIM (Chen et al., 2017) used the syntactic dependencies to enhance the sentence representations, see also (Chen et al., 2016; Liu et al., 2018; Tymoshenko and Moschitti, 2018). The NLP knowledge-enhanced matching models have also adapted to the interaction-based models. For example, MIX (Chen et al., 2018) utilizes POS and named-entity tags as prior features. Recently, (Sachan et al., 2021; Bai et al., 2021; Zhang et al., 2020b) found that syntax can help PLM capture more information and achieve impressive results for NLP tasks. However, these models overlook the fine-grained syntactic matching patterns.

Graph matching problem (GM) has also been adopted for discovering the patterns between different graphs (Loiola et al., 2007). The key is to learn a practical affinity function with given two structures. In early work, most on seeking approximate affinity function and Euclid distance together with the Gaussian kernel is applied (Cho et al., 2010; Leordeanu et al., 2012). Recently, quadratic assignment programming (QAP) (Cho et al., 2010) has a wide application in Graph Matching (GM) because of its great performance. The affinity function in QAP can be learned with the manners of unsupervised (Leordeanu et al., 2012), semi-supervised (Leordeanu et al., 2011), or supervised (Loiola et al., 2007). Recently, deep graph matching has been applied for GM on images (Zanfir and Sminchisescu, 2018; Wang et al., 2021) and the matching accuracy has been achieved. Moreover, Graph-based models have also been used for sentence matching (Yao et al., 2019; Zhang et al., 2020a; Sachan et al., 2021).
3 Problem Formulation

3.1 Sentence matching

The matching of a pair of natural language sentences can be formally described as follows: suppose that $\Z$ is the set of labels which is defined by a specific matching task. In the paraphrase identification (PI) tasks, $\Z = \{0, 1\}$, where ‘0’ and ‘1’ respectively denote the relationship of “dissimilar” and “similar”; in natural language inference (NLI) $\Z = \{0, 1, 2\}$, where 0, 1, 2 respectively indicate “contradiction”, “neutral”, and “entailment”. A set of training instances $D = \{(X_i, Y_i, z_i)\}_{i=1}^n$ is given where each sample $(X, Y, z) \in D$ consists of a sentence pair $(X, Y)$ and its ground-truth matching label $z$. Moreover, the $X, Y$ are two sequences of words: $X = \{x_1, x_2, \ldots, x_{|X|}\}$ and $Y = \{y_1, y_2, \ldots, y_{|Y|}\}$, where the $x_i$ and $y_j$ denote the $i$-th and $j$-th words in $X$ and $Y$, $t_X$ and $t_Y$ are the number of words (lengths) of $X$ and $Y$, respectively.

3.2 Quadratic assignment programming

Quadratic assignment programming (QAP) is a type of combinatorial optimization problems (Loiola et al., 2007), originally designed for the facilities-location problems. Suppose to assign $N$ facilities to $N$ locations, with the cost $f_{ij}, d_{kl}$ of the affinities between the facilities $(i, j)$ and locations $(k, l)$, plus the costs $b_{i\phi(i)}$ of assigning a facility $i$ to a certain location $\phi(i)$. The objective of QAP is assigning each facility to a location such that the total assignment cost is minimized. Lawler (1963) introduced a general form of QAP as in Equation (1):

$$\min_{\phi \in S} \sum_{i=1}^{N} \sum_{j=1}^{N} c_{ij\phi(i)\phi(j)} + \sum_{i=1}^{N} b_{i\phi(i)},$$

where $S$ is the set of all permutations $\phi : N \rightarrow N$, $\forall i, j, k, l, c_{ijkl} := f_{ij}d_{kl}$ if $i \neq j, k \neq l$, otherwise $c_{kikk} := f_{ik}d_{kk} + b_{ik}$. The formulation has been widely applied to graph matching, which involves establishing node correspondences between two graphs based on the linear and quadratic structure affinity (Leordeanu and Hebert, 2005).

3.3 Sentence matching over syntax graphs

This paper proposes to adopt QAP for conducting sentence matching, by regarding the words in one sentence as the “facilities” and words in another sentence as the “locations”, and their differences in syntactic structures and semantics as the “assignment costs”. In this way, QAP enables the matching model to involve not only the linear syntactic structure (e.g. word attribute structure) costs which correspond to assigning the “facilities” to certain “locations”, but also the quadratic syntactic structures (e.g. word-word relation structure) costs which correspond the affinities between the assigning “facilities” and “locations”.

Figure 2 gives an illustrative example of formulating the sentence matching as graph matching and finding matching patterns in the affinity graph, with a sentence pair $X$ = “Mary likes flour food” (length $|X| = 4$) and $Y$ = “Mary loves noodles” (length $|Y| = 3$), using the parsed POS tags and syntactic dependencies shown in Figure 2(a) which are represented as the node and edge weights, respectively.

As shown in Figure 2(a), we firstly constructed two directed graph $(G_X, G_Y)$ based on the two sentences $(X, Y)$’s syntax. Specifically, the word embedding and word syntax (i.e., POS) can be encoded as node features, and the word-word syntax (i.e., dependencies) can be encoded as edge features. The fine-grained syntax-based matching signals can be viewed as the alignments between
sub-graphs of two sentence graphs \((G_X, G_Y)\).

At the same time, to capture syntax-based patterns effectively, we further construct the affinity matrix \(K\) or the association graph \(G\) based on the sentence graphs \((G_X, G_Y)\). The weights of nodes and edges in the association graph are corresponding to the diagonal and off-diagonal elements in the affinity matrix (Figure 2(b,c), respectively).

Specifically, the weights of nodes describe the word semantic similarities and POS (word attribute) affinities, and the weights of edges describe the syntactic dependency (word-word relation) affinities. For example, the node \(x_1 y_1 = (\text{"Mary"}, \text{"Mary"})\) could have the weight of, for example, \(1.0 + 1.0 = 2.0\) where the first 1.0 denoting the semantic similarity, and the second 1.0 denoting the similarity between the POS tags.

As an example for the edges corresponding to the off-diagonal elements in affinity matrix \(K\), there could be an edge with weight, for example, 1.0 between node \(x_1 y_1 = (\text{"Mary"}, \text{"Mary"})\) and node \(x_2 y_2 = (\text{"likes"}, \text{"loves"})\) because the dependency relation between \(x_1 = \text{"Mary"}\) and \(x_2 = \text{"likes"}\) is “nsubj”, while the dependency relation between \(y_1 = \text{"Mary"}\) and \(y_2 = \text{"loves"}\) is also “nsubj”. Similarly, the other edges can also be created.

Then the QAP is applied to learn the semantic and syntactic matching patterns in affinity matrix \(K\). Formally, a relaxed form of QAP can be shown as Equation (2):

\[
\max_S \text{vec}(S)^T K \text{vec}(S), \tag{2}
\]

where \(S\) matrix encodes the word-word correspondence; \(\text{vec}(S)\) is \(S\)'s column-vectorized notation, and \(K \in \mathbb{R}^{1 \times 1 \times 1 \times 1 \times 1} \). The \(S\) matrix can be regarded as an aggregated matching patterns of syntax and semantics.

4 Proposed model: ISG

In this section, we present an efficient implementation of interacting syntax graphs (ISG) with quadratic assignment programming (QAP) (Lawler, 1963; Cho et al., 2010). Figure 3 illustrates the model architecture, which can be divided into sentence characteristic initialization, QAP component, and matching classifier.

4.1 Sentence characteristic initialization

In this component, the inputted natural language sentence pair \((X, Y)\) is processed with a pre-trained language model (PLM) and an NLP parser, generating the semantic features and syntactic structures.

**Semantic features** Given a pair \((X, Y)\), the semantic matching vector (e.g., “[CLS]” vector of BERT), \(v_x \in \mathbb{R}^d\) and the words embeddings \(F_X \in \mathbb{R}^{l_X \times d}\) and \(F_Y \in \mathbb{R}^{l_Y \times d}\) consists of the semantic features:

\[
(v_x, F^X, F^Y) = \text{PLM}(X, Y; \theta_p),
\]

where \(d\) denotes the size of the feature vector, PLM could be BERT or other PLM models, and \(\theta_p\) denotes the parameters of PLM.

**Syntactic structures** Generally speaking, there are two types of structures: the word attribute structure (WAS) which reflects the attributes of the word, and the word-word relation structure (WRS) which defines the relationship between two words.

The WAS attributes can be further categorized...
and this paper only considers POS attributes. Given any sentence $X = \{x_1, \ldots, x_{tX}\}$, the sequence of WAS attributes could be $\{a_{x_1}, a_{x_2}, \ldots, a_{x_{tX}}\}$.

The WRS attributes can also be further categorized and this paper considers syntactic dependency. Given any sentence $X = \{x_1, \ldots, x_{tX}\}$, the WRS parsing results (a dependency parsing graph) can be represented as two incidence matrices:

$$(I^X, H^X) = \text{Parse}(X),$$

where $I^X \in \mathbb{R}^{tX \times eX}$ records the output-links and $H^X \in \mathbb{R}^{tX \times eX}$ records the in-links, $eX$ denotes the edge number of WRS. The elements of these two matrices are defined as: if k-th edge links from word $x_i$ to $x_j$ (its type also denoted as $e_k(x_i, x_j)$), $I^X(i, k) = H^X(j, k) = 1$, and note that in order to reduce the noise from the dependencies, we also set $I^X(j, k) = H^X(i, k) = 1$. Otherwise, $I^X(i, k) = H^X(j, k) = I^X(j, k) = H^X(i, k) = 0$.

In this paper, we used the Stanford CoreNLP parser (Manning et al., 2014) for getting POS, and syntactic dependencies. Note that other syntactic structures can be also used, such as named-entity and semantic dependencies (Wang et al., 2019b).

### 4.2 QAP component

Based on the word embeddings and parsed syntactic structures, the QAP component first constructs an association graph (affinity matrix) and then solves the QAP problem, achieving the permutation which represents the word matching between the two sentences.

#### 4.2.1 Learned affinity matrix construction

Following the practices in (Zhou and De la Torre, 2015), the QAP sparse affinity matrix $K^l \in \mathbb{R}^{tX \times tY}$, referred to as the learned affinity matrix, can be factorized as

$$K^l = \text{diag}(\text{vec}(P)) + (I^X \otimes_K I^Y)\text{diag}(\text{vec}(R))(H^X \otimes_K H^Y)^T,$$

where operator $\text{diag}(\cdot)$ builds a diagonal matrix from input vector, $I^X, H^X, I^Y, H^Y$ are sentences $X$ and $Y$’s parsing results, as described in Section 4.1, $\otimes_K$ denotes Kronecker product, and $P$ and $R$ encode the WAS, word embedding similarity and WRS similarity matrix, respectively and they are defined as:

$$P = (1 - \alpha)U^X\Lambda_\text{a}U^Y^T + \alpha F^X\Lambda_\text{f}F^Y^T, R = L^X\Lambda_\text{r}L^Y^T,$$

where $\Lambda_\text{a}, \Lambda_\text{f}, \Lambda_\text{r}$ are learn-able parameters for affinity metric, $\alpha$ is the trade-off coefficient for POS affinities and word-word similarities, and $U^X \in \mathbb{R}^{tX \times d}, U^Y \in \mathbb{R}^{tY \times d}$ are the WAS sequence embeddings of $X, Y$ and the edge representations $L^X \in \mathbb{R}^{eX \times d}, L^Y \in \mathbb{R}^{eY \times d}$ are built by its edge sequence embeddings. Note that all the aforementioned operations for constructing $K^l$ allow back propagation, and we adopt the GPU implementation provided by (Wang et al., 2019a). A more detailed QAP factorization is given in Appendix A.

#### 4.2.2 Solving the permutation vector

Due to the high compute cost for solving the permutation vector through the learned affinity matrix $K^l$, we adopt the GCN method implemented by Wang et al. (2019a) to approximate the QAP problem into a linear assignment programming(LAP) problem, which can be solved in an efficient way for both time and space.

Specifically, we build the association graph $G = \{v^{(0)}, A\}$ with its initial node embedding $v^{(0)}$ and its sparse adjacent matrix $A$ from the learned affinity matrix $K^l$. Then we can apply GCN method to updated the node embedding for k-th GCN layer, $k = 1, 2, \ldots, G_k$. The key idea is to encode the quadratic structure (WRS) to the linear structure (WAS). The permutation matrix $S$ can be regarded as the last layer of the node features:

$$\text{vec}(S) = v^{(G_k)}, \quad v^{(k+1)} = AWf(v^{(k)}; \theta_k) + v^{(k)},$$

(4)

where the $f(\cdot)$ is a MLP projection function at the k-th layer is parameterized by $\theta_k$ and the k-th layer node embedding of association graph denotes as: $v^{(k)} \in \mathbb{R}^{tX \times ty \times tX \times ty}$, with the initial embeddings $v^{(0)} \in \mathbb{R}^{tX \times ty \times 1}$ taken from the diagonal elements of $K^l$. The GCN projection matrix $W \in \mathbb{R}^{tX \times ty \times tX \times ty}$ comes from the off-diagonal elements.

$$v^{(0)}(i, a) = K^l(ia, ia), \quad W(ia, jb) = K^l(ia, jb),$$

for all $i, j \in tX, a, b \in ty$.

As for the adjacent matrix $A$, in order to control its sparsity, we introduce a hyper-parameter $\gamma$ to generate the sparse adjacent matrix of association graph $G$ from the projection matrix $W$:

$$A(ia, jb) = \begin{cases} 1 & \text{if } W(ia, jb) \geq \gamma, \\ 0 & \text{otherwise}. \end{cases}$$

### 4.3 Matching classifier

Given the semantic matching feature $v_s$ and QAP permutation vector $v^{(G_k)}$, the matching score $\hat{x}$ can
be obtained by the MLP parameterized by $\theta_m$:
\[
\hat{z}(X, Y) = \text{MLP}(|v_{sk}|v^{(G_k)}; \theta_m),
\]
where $'|'$ denotes the concatenation operation, $\hat{z}(X, Y) = [\hat{z}_1, \cdots, \hat{z}_|Z|]$ and $\hat{z}_k$ denotes the probability of $k$-th category. The last layer is softmax so that the output is a probability distribution.

### 4.4 Learning the model parameters

ISG has parameters to determine, including $\Theta = \{\theta_p, \theta_k, \Lambda_u, \Lambda_f, \theta_m \mid k = 1, 2, \cdots, G_k\}$. In the training phase, given a set of sentence pairs with ground truth labels $D = \{(X_i, Y_i, z_i)\}_{i=1}^N$, the learning algorithm aims to minimize the matching loss $L_m$, which measures the differences between the prediction $\hat{z}$ and ground-truth $z$, regularized by the affinity regularizer $R$ which forces the learned affinity matrix $K^I$ and the original parsed affinity matrix $K$ the being similar. Formally, the loss $L$ that being minimized is:
\[
L = L_m(\hat{z}, z) + \lambda_a R(K, K^I) + \mu_r \|\theta\|^2,
\]
where $\|\theta\|^2$ is the $\ell_2$ regularizer, $\lambda_a, \lambda_r$ denote the trade-off coefficient of affinity regularizer and $\ell_2$ regularizer.

**Matching loss** The matching loss $L_m$ is learned by minimizing the cross-entropy loss between the labels and the predicted results:
\[
L_m = - \sum_{(X,Y,z) \in D} \sum_{k=1}^{|Z|} z_k \log \hat{z}_k,
\]

**Affinity regularizer** The affinity regularizer $R$ aims to force the structure affinities respectively correspond to the parsed syntactic structure and that of learned from neural network to be similar. Thus the $R$ is learned to minimize the KL-divergence between the learned affinity matrix $K^I$ and parsed affinity matrix $K$:
\[
R = \sum_{(X,Y) \in D} KL(K||K^I),
\]
where the parsed affinity matrix $K$ is defined as follows: the diagonal elements $K(ia, ia)$ will be 1 if the matched words have identical attribute, otherwise 0. And the off-diagonal element $K(ia, jb)$ will be 1 if the word pair $(x_i, x_j)$ and $(y_a, y_b)$ have identical word-word relation, otherwise 0.

### 4.5 Time complexity of online matching

At the online time, ISG needs to process the sentence pairs with PLM, parse them with the NLP parser, solve the QAP and finally calculate the matching score. The online time complexity for typical PLM (Devlin et al., 2019; Liu et al., 2019) and NLP parser (Manning et al., 2014; Wang et al., 2019b) is of $O(|t_x + ty|^2 \times d)$ and $O(|t_x|^2 + |ty|^2 \times d)$, where $d$ is the embedding dimension of each word.

At the online matching, the time complexity of the relaxed QAP is related to GCN, which is of $O(G_k m t + G_k n t^2)$ on the association graph, where $n = t_x t_y$ is the total number of nodes, $m$ is the total number of edges, $G_k$ is the number of layers, and $t$ is the dimension of the node hidden features. Note that the hyper-parameter $\gamma$ controls the sparsity of the edges (as mentioned in Section 4.2), we can adjust $\gamma$ so that $m \ll t_x t_y$ and therefore reduce the time complexity of the relaxed QAP to $O(G_k t_x t_y t^2)$, which is more efficient than the original QAP (Wang et al., 2019a). Therefore, the total time complexity of ISG is $O(|t_x + ty|^2 \times d + G_k t_x t_y t^2)$, which is comparable with the underlying PLM.

### 5 Experiments

We conducted experiments to verify the effectiveness of the proposed approach. The source code and experiments are available at the link.\(^1\)

#### 5.1 Experimental Settings

The experiments were conducted on three large scale publicly available benchmarks:

**Quora Question Pairs (QQP):**\(^2\) a large public dataset for paraphrase identification. QQP contains 404k labeled sentence pairs. We used the same data split as in (Wang et al., 2017b).

**SNLI:**\(^3\) a well-known dataset for natural language inference (NLI). SNLI contains 570k labeled sentence pairs. Following the practices in (Bowman et al., 2015), we used the same data split way. **SciTail:**\(^4\) another NLI dataset based on science exams and web. Its label only contains two classes: “entailment” or “neutral”. The dataset contains 27k sentence pairs.

\(^1\)https://github.com/XuChen0427/Semantic-Sentence-Matching-via-Interacting-Syntax-Graphs
\(^2\)https://www.kaggle.com/c/quora-question-pairs
\(^3\)https://nlp.stanford.edu/projects/snli
\(^4\)http://data.allenai.org/scitail/
Several state-of-the-art baselines which conduct the matching without utilizing syntactic structures were chosen as the baselines, including DIIN (Gong et al., 2018), MwAN (Tan et al., 2018), BIMPM (Wang et al., 2017a), CSRAN (Kim et al., 2019), DecAtt (Parikh et al., 2016), CAFE (Tay et al., 2018), and DGEM (Khot et al., 2018), RE2 (Yang et al., 2019), and the BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019). Some models are task-adopted (e.g. DGEM is for NLI task), thus they are missing on some datasets. ISG was compared with the baselines DDR-match (Yu et al., 2020, 2022) that applied unsupervised assignment problems to conduct sentence matching. ISG was also compared with the baselines that utilize syntactic structures like HIM (Chen et al., 2017), which uses the constituency tree to improve local word representation, and Sembert (Zhang et al., 2020b), Syntext-bert (Bai et al., 2021) that applied different syntax to improve PLM performance. To make a fair comparison, we cannot reproduce their results due to the source codes are unavailable. We conducted matching scores based on the learned embeddings. We found that ISG will still perform all of the baselines. The results also indicated that though PLM (e.g. BERT, RoBERTa) achieved SOTA accuracy, ISG can still get improvements by incorporating the syntactic information.

We also note that ISG outperformed the baselines that utilize the syntactic structures for matching. Comparing ISG with these models, we found that these baseline models all encode the syntactic structures as sentence features to enrich their representations, while ISG incorporates the syntactic and semantic matching patterns through a graph matching task and aggregates them through the affinity matrix. Moreover, we also compared the graph matching methods that separately encode two graphs into two embeddings using GCN methods and conducted matching scores based on the learned embeddings. We found that ISG will still outperform the representation-based graph matching methods. The results demonstrated that the ISG is more effective to utilize syntactic matching signals.

5.2 Experimental results

Table 1 reports the matching accuracy of the proposed ISG and the baselines on the three datasets. The '-' means the number is not available. The accuracy of baselines is according to the numbers reported. For our methods, the averaged numbers over 5 runs are reported, with the standard deviations in parentheses. From the results, we can see that different versions of the proposed ISG outperformed all of the baselines. The results also indicated that though PLM (e.g. BERT,RoBERTa) achieved SOTA accuracy, ISG can still get improvements by incorporating the syntactic information.

We also investigated the online time complexity of ISG. Figure 4 reports the impacts of association graph sparsity on ISG-BERT\textsubscript{BASE} on the Scitail test-set, where the sparsity (calculated as the fraction of edge number and square of node number in association graph) is from \([2e - 4\%, 3e - 3\%]\). The sparsity was adjusted by changing the hyper-parameters $\gamma$.

Figure 4(a) illustrates that the inference time of ISG will decrease with the increase of the association graph sparsity. Moreover, the inference time of ISG-BERT\textsubscript{BASE} is about 2.5 times that of the underlying PLM, and about 0.7 times that of BERT\textsubscript{LARGE}. The results verified the time complexity analysis conclusion in Section 4.5.

Figure 4(b) shows the accuracy curves of ISG.
Table 1: Performance comparisons on Quora Question Pairs, SNLI and SciTail. The ±numbers in brackets mean 1-std deviations. The * denotes the models are our implementations and are trained among same settings.

<table>
<thead>
<tr>
<th>Models without syntactic structures</th>
<th>QQP:Acc(%)</th>
<th>SNLI:Acc(%)</th>
<th>SciTail:Acc(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGEM (Khot et al., 2018)</td>
<td>-</td>
<td>-</td>
<td>77.3</td>
</tr>
<tr>
<td>DecAtt (Parikh et al., 2016)</td>
<td>-</td>
<td>82.5</td>
<td>81.7</td>
</tr>
<tr>
<td>CAFE (Tay et al., 2018)</td>
<td>-</td>
<td>88.5</td>
<td>83.3</td>
</tr>
<tr>
<td>BIMPM (Wang et al., 2017a)</td>
<td>88.7</td>
<td>88.8</td>
<td>85.4</td>
</tr>
<tr>
<td>DHIN (Gong et al., 2018)</td>
<td>89.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MwAN (Tan et al., 2018)</td>
<td>89.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CSRAN (Kim et al., 2019)</td>
<td>89.2</td>
<td>88.7</td>
<td>-</td>
</tr>
<tr>
<td>RE2 (Yang et al., 2019)</td>
<td>89.2</td>
<td>89.0</td>
<td>86.6</td>
</tr>
<tr>
<td>DDR-Match(BERT)* (Yu et al., 2022)</td>
<td>89.6</td>
<td>89.2</td>
<td>90.3</td>
</tr>
<tr>
<td>BERT$_{BASE}$ (Devlin et al., 2019)</td>
<td>89.4</td>
<td>89.0</td>
<td>89.5</td>
</tr>
<tr>
<td>BERT$_{LARGE}$ (Devlin et al., 2019)</td>
<td>89.6</td>
<td>89.2</td>
<td>90.6</td>
</tr>
<tr>
<td>RoBERTa*$_{LARGE}$ (Liu et al., 2019)</td>
<td>90.0</td>
<td>90.1</td>
<td>91.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models with syntactic structures</th>
<th>QQP:Acc(%)</th>
<th>SNLI:Acc(%)</th>
<th>SciTail:Acc(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIM (Chen et al., 2017)</td>
<td>88.7</td>
<td>88.6</td>
<td>71.6</td>
</tr>
<tr>
<td>SemBERT$_{BASE}$ (Zhang et al., 2020b)</td>
<td>89.8</td>
<td>90.2</td>
<td>92.1</td>
</tr>
<tr>
<td>SemBERT$_{LARGE}$ (Zhang et al., 2020b)</td>
<td>90.7</td>
<td>91.0</td>
<td>92.1</td>
</tr>
<tr>
<td>SyntaxBERT$_{BASE}$ (Bai et al., 2021)</td>
<td>89.6</td>
<td>87.8</td>
<td>-</td>
</tr>
<tr>
<td>SyntaxBERT$_{LARGE}$ (Bai et al., 2021)</td>
<td>89.5</td>
<td>89.0</td>
<td>-</td>
</tr>
<tr>
<td>ISG(representation-GCN)-BERT$_{BASE}$</td>
<td>90.1</td>
<td>89.7</td>
<td>90.6</td>
</tr>
<tr>
<td>ISG(representation-GCN)-BERT$_{LARGE}$</td>
<td>90.2</td>
<td>89.9</td>
<td>91.7</td>
</tr>
<tr>
<td>ISG(representation-GCN)-RoBERTa*$_{LARGE}$</td>
<td>90.8</td>
<td>90.4</td>
<td>92.3</td>
</tr>
<tr>
<td>Ours(ISG-BERT$_{BASE}$)*</td>
<td>90.5 (+0.14)</td>
<td>90.0 (+0.16)</td>
<td>90.8 (+0.26)</td>
</tr>
<tr>
<td>Ours(ISG-BERT$_{LARGE}$)*</td>
<td>90.8 (+0.08)</td>
<td>90.4 (+0.03)</td>
<td>91.9 (+0.24)</td>
</tr>
<tr>
<td>Ours(ISG-RoBERTa$_{LARGE}$)*</td>
<td>91.4 (+0.1)</td>
<td>91.2 (+0.08)</td>
<td>93.3 (+0.2)</td>
</tr>
</tbody>
</table>

Figure 5: Cross sentence word-word similarity matrix and syntactic affinity matrices for two pairs: example(“this gas is oxygen”, “oxygen gas is given off by plants”), which is from SciTail training set. Darker colors means higher similarities or affinities values.

which first increases in $[9e-4\%, 3e-3\%]$ and then dropped. We conclude that even association graph became sparse, ISG still constantly outperformed BERT$_{BASE}$ and outperformed BERT$_{LARGE}$ at some point. The results demonstrated that the QAP is efficient and will not delay the online time.

5.3 Empirical Analysis

5.3.1 Ablation Study

Firstly, we respectively set the WAS(POS) features $U^X$, $U^Y$, WRS(syntactic dependencies) features $L^X$, $L^Y$ and semantic features $v_s$, $F^X$, $F^Y$ to zero vectors, to investigate their effects. Table 2 reports the accuracy of the ISG variation on the SciTail test data under BERT$_{BASE}$, where each variation is denoted as, for example, “ISG-w/o WRS” which means the WRS features were set zeros. Similar phenomenons have also been observed on the other two datasets, with other PLMs.

Compared ISG-BERT$_{BASE}$ with its variations, we can see that the matching performances dropped with large margins if the semantic features were set as zeros, indicating that only considering the syntax patterns did not work well. We also ob-
Table 2: Ablation study on SciTail test set.

<table>
<thead>
<tr>
<th>Ablation Study Model</th>
<th>Acc(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT\textit{BASE} (Devlin et al., 2019)</td>
<td>89.5 (±0.28)</td>
</tr>
<tr>
<td>ISG-w/o semantic and WRS</td>
<td>68.1 (±0.29)</td>
</tr>
<tr>
<td>ISG-w/o semantic and WAS</td>
<td>67.8 (±0.28)</td>
</tr>
<tr>
<td>ISG-w/o semantic</td>
<td>68.5 (±0.26)</td>
</tr>
<tr>
<td>ISG-w/o WAS</td>
<td>90.2 (±0.27)</td>
</tr>
<tr>
<td>ISG-w/o WRS</td>
<td>90.4 (±0.28)</td>
</tr>
<tr>
<td>ISG</td>
<td>90.8 (±0.26)</td>
</tr>
</tbody>
</table>

Table 3: Ablation study for different syntactic structure on SciTail test set.

<table>
<thead>
<tr>
<th>Ablation Study Model</th>
<th>Acc(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT\textit{BASE} (Devlin et al., 2019)</td>
<td>89.5 (±0.28)</td>
</tr>
<tr>
<td>ISG NER&amp;Syntactic dependencies</td>
<td>90.4 (±0.25)</td>
</tr>
<tr>
<td>ISG NER&amp;Semantic dependencies</td>
<td>90.2 (±0.17)</td>
</tr>
<tr>
<td>ISG POS&amp;Syntactic dependencies</td>
<td>90.8 (±0.26)</td>
</tr>
<tr>
<td>ISG POS&amp;Semantic dependencies</td>
<td>90.6 (±0.21)</td>
</tr>
</tbody>
</table>

served that the performances dropped when the WAS and WRS features were set to zeros. The bad performances were caused by removing the WAS and WRS features, indicating that syntax matching patterns are effective for sentence matching.

Moreover, we conducted the experiments with different WAS and WRS. Specifically, we respectively utilize the POS and named-entity(NE) as WAS and respectively utilize the syntactic dependencies and semantic dependencies as WRS. For semantic dependencies parsing, we follow Wang et al. (2019b). Table 3 also reports the accuracy of the ISG-BERT\textit{BASE} variation on the SciTail test data.

Compared to the ISG variations, we can see that the matching accuracy is different for different WAS and WRS. The best and worst performance are caused by POS&syntactic dependencies and NER&Semantic dependencies, respectively. However, we can observe that all of these variations outperform the BERT baseline, which indicates the effectiveness of ISG in different WAS and WRS.

An experiment on the robustness of ISG’s parameters can be found in Appendix B.

5.3.2 Matching Visualization of ISG

We conducted experiments to investigate how the ISG matched two sentences, using a real example from Scitail. The experiment was conducted based on the results of ISG-RoBERTa.

Figure 5(a) illustrated the word-word similarity matrix of these two sentences, based on the word embeddings outputted by RoBERTa, where the darker colors denote the higher similarities. Figure 5(b) illustrated the affinities between POS and dependencies in two sentences. Based on the similarities and affinity matrices, ISG solved the QAP and achieved a new correspondence matrix in Figure 5(c). The POS, word semantic similarities, and dependencies affinities correspond to the node weights and edge weights in the association graph.

The example illustrates the example that is from the SciTail training set: \((X = \text{“this gas is oxygen”}, Y = \text{“oxygen gas is given off by plants”})\) whose ground truth label is “neutral”. The example illustrates how the ISG conducts sentence matching. Comparing word-word similarities by RoBERTa (Figure 5(a)) and that of ISG (Figure 5(c)), we can see that RoBERTa’s results show high similarities between words in two sentences. On the other hand, ISG can find fine-grained syntactic patterns, for example, the matching patterns of POS “VB” and “NN” and dependency “nsubj”. And ISG will output the “dissimilar” result due to its low syntactic matching pattern scores.

The analysis clearly showed that ISG can well learn different syntax-based matching signals, and make them into good interactions. The results also showed what syntactic matching patterns are important and how two sentences were matched with the association graph.

6 Conclusion

This presents a novel sentence matching model which formulates sentence matching as a problem of syntax graph matching, referred to as ISG. Based on the constructed association graph, ISG explicitly aligns the syntactic sub-graphs as fine-grained matching signals. Neural QAP is adopted to learn the rich matching patterns from the training data. ISG offers several advantages: explicitly interacting with the syntactic structures in fine granularity, high matching accuracy, and the ability to interpret. Experiments on three publicly available benchmarks verified the effectiveness, robustness, and interpretability of ISG.

Acknowledgments

This work was funded by the National Key R&D Program of China (2019YFE0198200), National Natural Science Foundation of China (61872338, 61832017), and Beijing Outstanding Young Scientist Program NO. BJJWZYJH01201910020098.
References


Appendix A: An intuitive example on affinity matrix factorization

Figure 6 gives a working example of factorizing the affinity matrix $\mathbf{K}^l \in \mathbb{R}^{l \times t_y \times t_y}$ in Equation (3) (Zhou and De la Torre, 2015):

$$\mathbf{K}^l = \text{diag} (\text{vec} (\mathbf{P})) + (\mathbf{I}^X \otimes \mathbf{1}^Y) \text{diag} (\text{vec} (\mathbf{R})) (\mathbf{H}^X \otimes \mathbf{H}^Y)^T,$$

(9)

with the aforementioned example sentence pair: (“Mary likes flour food”, “Mary love noodles”). As shown in Figure 6(a), the words, POS and syntactic dependencies are represented in the nodes and edges, respectively.

The diagonal and off-diagonal elements in the affinity matrix (Figure 6(d)) represent the affinity of sentence linear structures and quadratic structures, respectively. According to Zhou and De la Torre (2015), the affinity matrix $\mathbf{K}^l$ can be factorized into six matrices $\mathbf{P}, \mathbf{R}, \mathbf{I}^X, \mathbf{1}^Y, \mathbf{H}^X, \mathbf{H}^Y$ (shown in Figure 6(b,c)) and defined in Section 4.

Appendix B: Robustness of ISG

ISG has a set of important hyper-parameters $\lambda_a$ which trade-off the affinity regularizer $R_a$ and matching loss $L_m$. We conducted experiments on the Scitail test set with BERT$_{BASE}$ as the encoder to test the sensitivity of these hyper-parameters. Figure 7 illustrates the performance changes w.r.t. $\lambda_a$ in terms of accuracy, where $\lambda_a \in [3e-3, 1.1e-2]$. We can see that ISG performed best when $\lambda_a \approx 8e-3$. However, the performance changes were not severe (from 90.6% to 90.8% in terms of accuracy). We conclude that (1) the introduction of the affinity regularizer enables ISG to have some tolerances to the errors caused by the NLP parser, which inevitably occurs in real-world applications; (2) ISG is robust and not sensitive to the $\lambda_a$.

Figure 7: Accuracy curve of ISG-BERT$_{BASE}$ w.r.t. $\lambda_a$ (trade-off coefficient for affinity regularizer) on the Scitail test set.