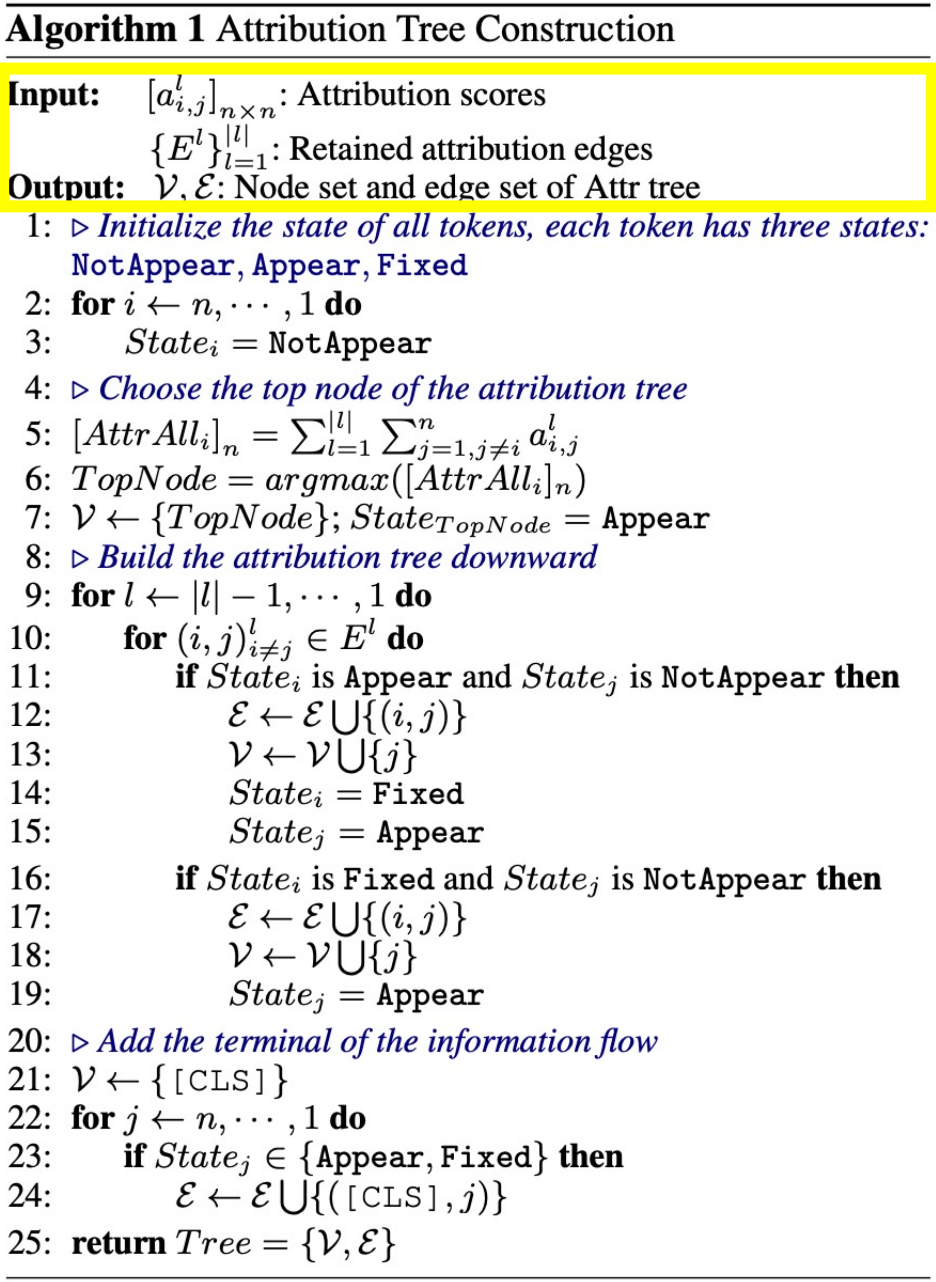
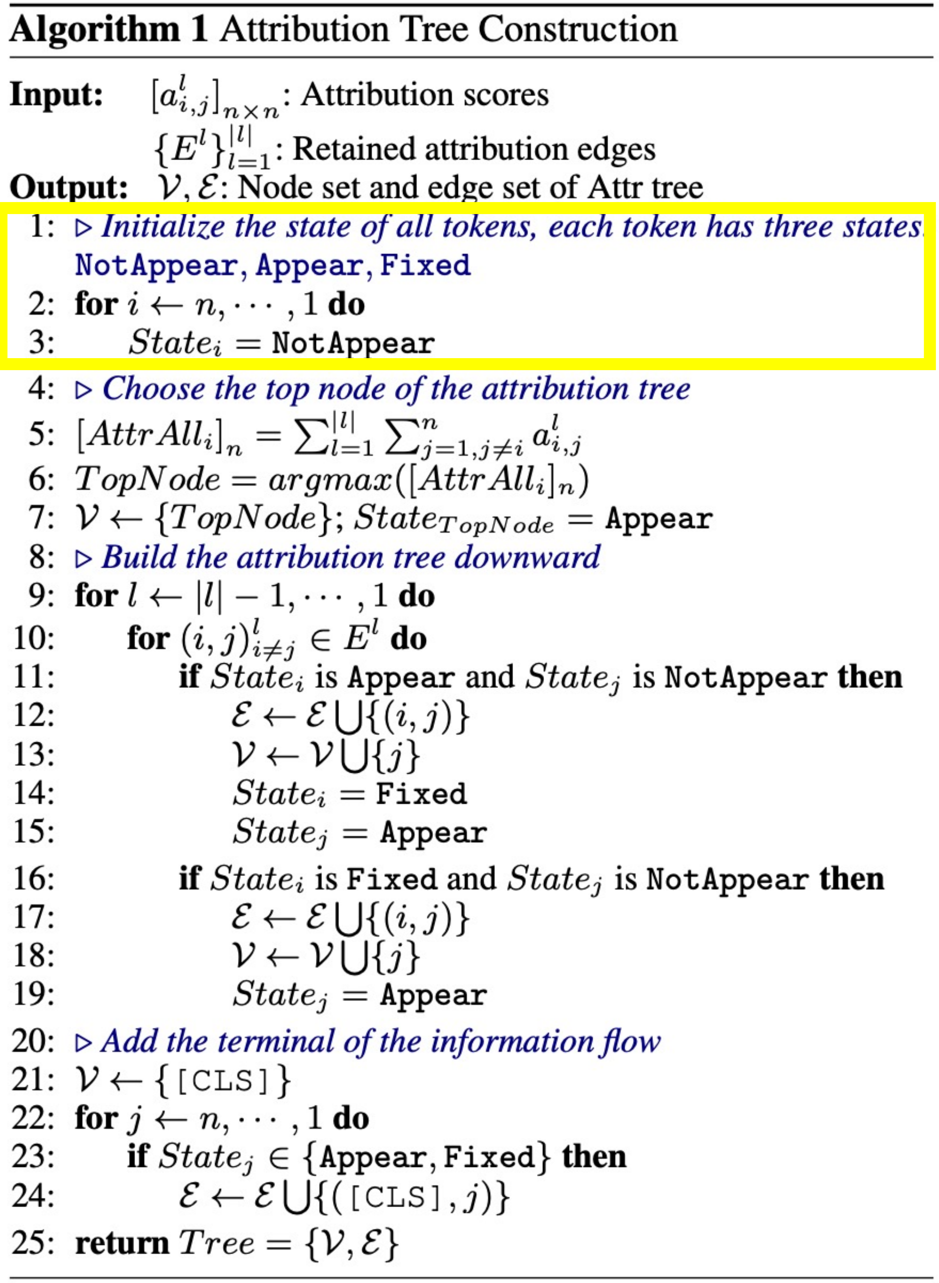


Algorithm 1 Attribution Tree Construction

```
Input:  $[a^l_{i,j}]_{n \times n}$: Attribution scores
        $\{E^l\}^{|l|}_{l=1}$: Retained attribution edges
Output: $\mathcal{V}, \mathcal{E}$: Node set and edge set of Attr tree
▷ Initialize the state of all tokens, each token has three states:
  NotAppear, Appear, Fixed
for $i \leftarrow n, \cdots, 1$ do
    $State_i$ = NotAppear
▷ Choose the top node of the attribution tree
$[AttrAll_i]_n = \sum_{l=1}^{|l|} \sum_{j=1, j \neq i}^{n} a^l_{i,j}$
$TopNode = argmax([AttrAll_i]_n)$
$\mathcal{V} \leftarrow \{TopNode\}$; $State_{TopNode}$ = Appear
▷ Build the attribution tree downward
for $l \leftarrow |l| - 1, \cdots, 1$ do
    for $(i, j)^l_{i \neq j} \in E^l$ do
        if $State_i$ is Appear and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_i$ = Fixed
            $State_j$ = Appear
        if $State_i$ is Fixed and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_j$ = Appear
▷ Add the terminal of the information flow
$\mathcal{V} \leftarrow \{$[CLS]$\}$
for $j \leftarrow n, \cdots, 1$ do
    if $State_j \in \{$Appear, Fixed$\}$ then
        $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{($[CLS]$, j)\}$
return $Tree = \{\mathcal{V}, \mathcal{E}\}$
```

Sentence: Seldom has a movie so closely matched the spirit of a man and his work

Model input is the attribution scores between each token in the sentence and the edges which tokens are related



Algorithm 1 Attribution Tree Construction

Input: $[a^l_{i,j}]_{n \times n}$: Attribution scores

$\{E^l\}^{|l|}_{l=1}$: Retained attribution edges

Output: $\mathcal{V}, \mathcal{E}$: Node set and edge set of Attr tree

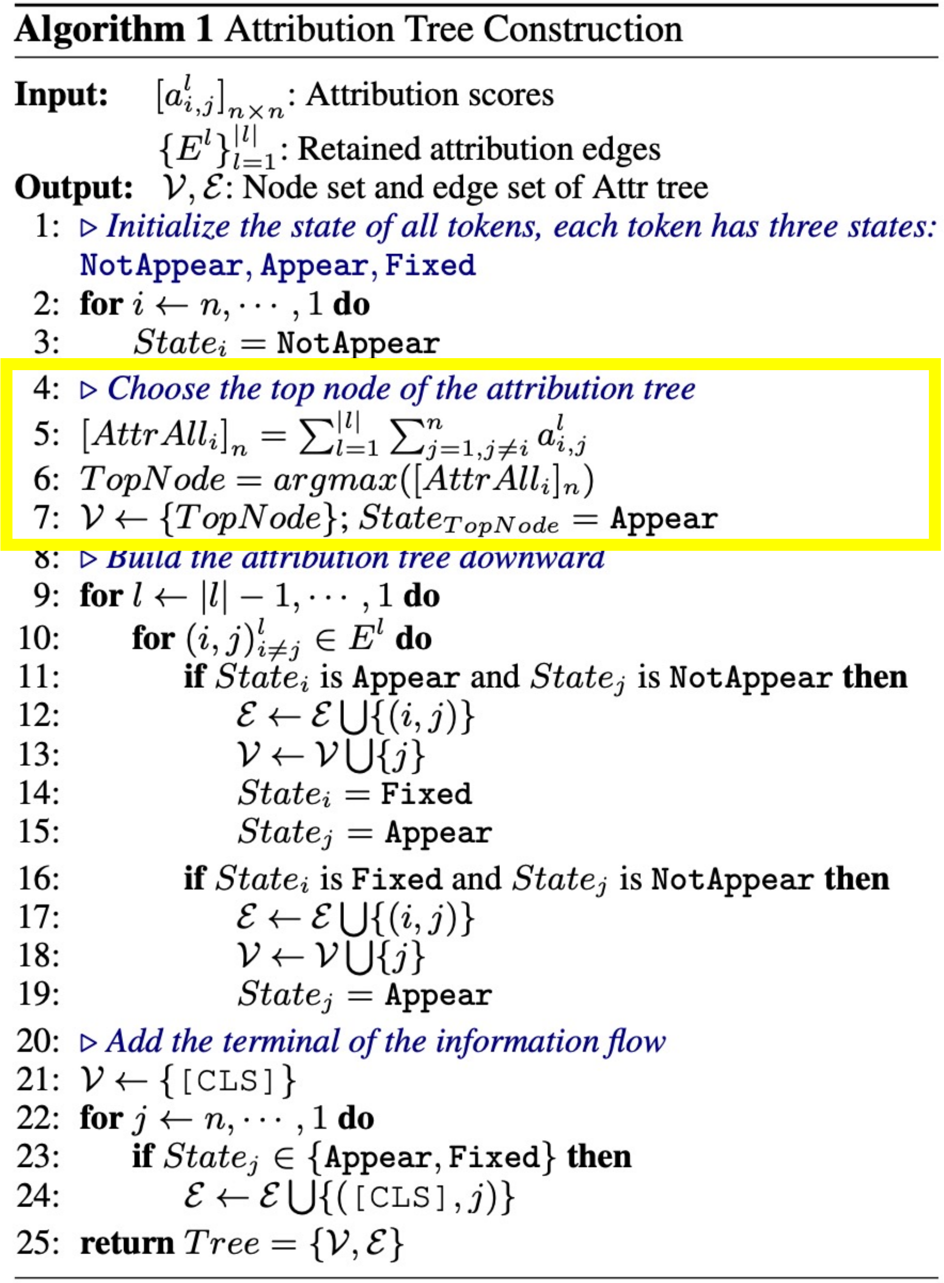
```
▷ Initialize the state of all tokens, each token has three states: NotAppear, Appear, Fixed
for $i \leftarrow n, \cdots, 1$ do
    $State_i = \texttt{NotAppear}$
▷ Choose the top node of the attribution tree
$[AttrAll_i]_n = \sum^{|l|}_{l=1} \sum^{n}_{j=1, j \neq i} a^l_{i,j}$
$TopNode = argmax([AttrAll_i]_n)$
$\mathcal{V} \leftarrow \{TopNode\}$; $State_{TopNode} = \texttt{Appear}$
▷ Build the attribution tree downward
for $l \leftarrow |l| - 1, \cdots, 1$ do
    for $(i, j)^l_{i \neq j} \in E^l$ do
        if $State_i$ is Appear and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_i = \texttt{Fixed}$
            $State_j = \texttt{Appear}$
        if $State_i$ is Fixed and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_j = \texttt{Appear}$
▷ Add the terminal of the information flow
$\mathcal{V} \leftarrow \{\texttt{[CLS]}\}$
for $j \leftarrow n, \cdots, 1$ do
    if $State_j \in \{\texttt{Appear}, \texttt{Fixed}\}$ then
        $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(\texttt{[CLS]}, j)\}$
return $Tree = \{\mathcal{V}, \mathcal{E}\}$
```

Status NotAppear Tokens

seldom
has
a
movie
so

closely
matched
the
spirit
of

a
man
and
his
work

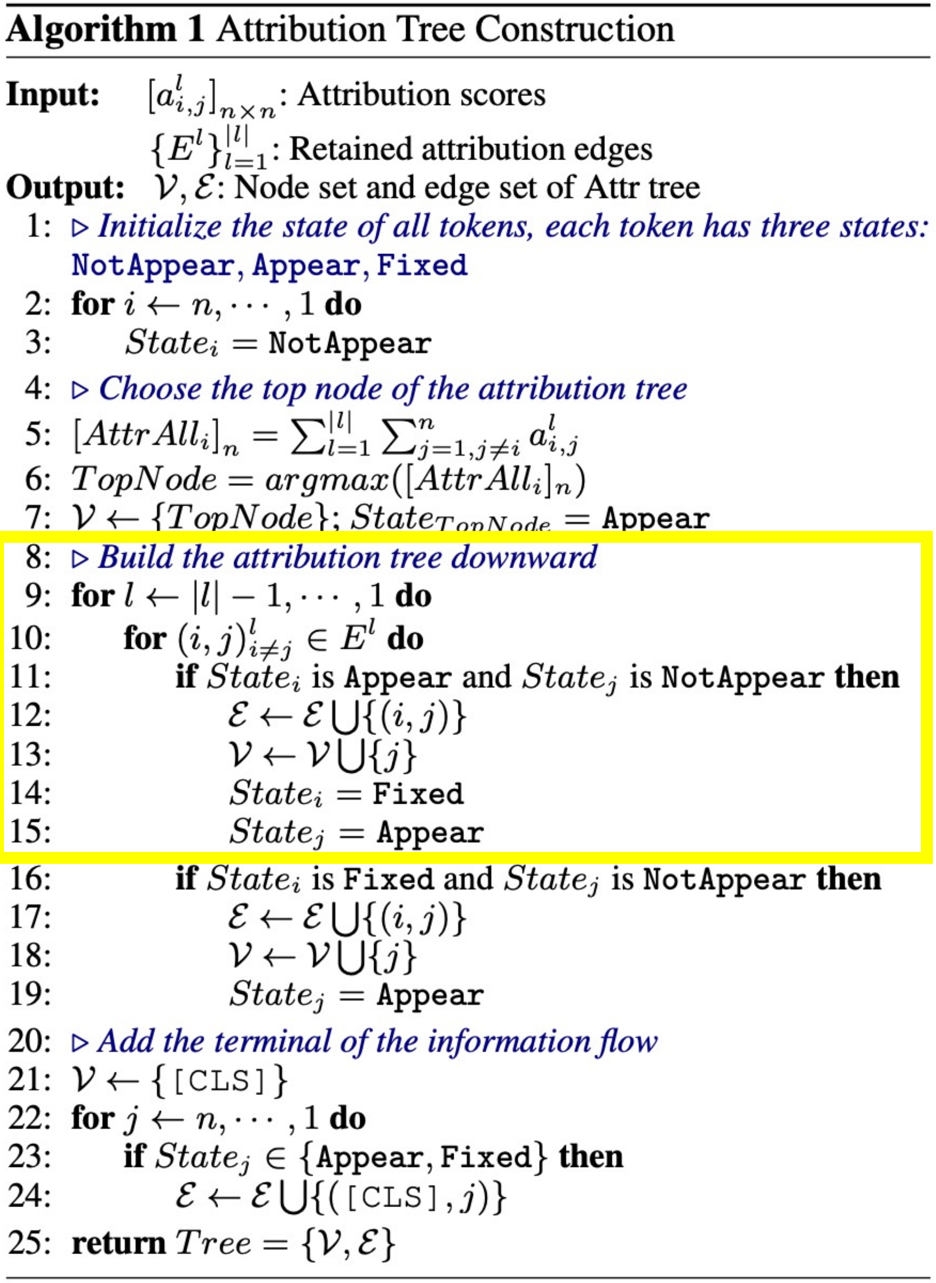


Algorithm 1 Attribution Tree Construction

```
Input:  $[a^l_{i,j}]_{n \times n}$: Attribution scores
        $\{E^l\}^{|l|}_{l=1}$: Retained attribution edges
Output: $\mathcal{V}, \mathcal{E}$: Node set and edge set of Attr tree
▷ Initialize the state of all tokens, each token has three states: NotAppear, Appear, Fixed
for $i \leftarrow n, \cdots, 1$ do
    $State_i = $ NotAppear
▷ Choose the top node of the attribution tree
$[AttrAll_i]_n = \sum^{|l|}_{l=1} \sum^{n}_{j=1, j \neq i} a^l_{i,j}$
$TopNode = argmax([AttrAll_i]_n)$
$\mathcal{V} \leftarrow \{TopNode\}$; $State_{TopNode} = $ Appear
▷ Build the attribution tree downward
for $l \leftarrow |l| - 1, \cdots, 1$ do
    for $(i,j)^l_{i \neq j} \in E^l$ do
        if $State_i$ is Appear and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i,j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_i = $ Fixed
            $State_j = $ Appear
        if $State_i$ is Fixed and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i,j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_j = $ Appear
▷ Add the terminal of the information flow
$\mathcal{V} \leftarrow \{$[CLS]$\}$
for $j \leftarrow n, \cdots, 1$ do
    if $State_j \in \{$Appear, Fixed$\}$ then
        $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{($[CLS]$, j)\}$
return $Tree = \{\mathcal{V}, \mathcal{E}\}$
```

Attribution Tree

matched



Algorithm 1 Attribution Tree Construction

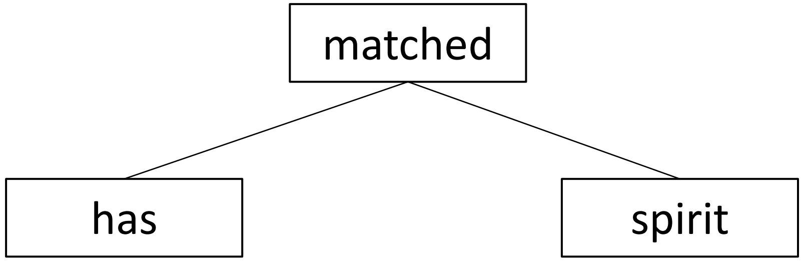
Input: $[a^l_{i,j}]_{n\times n}$: Attribution scores

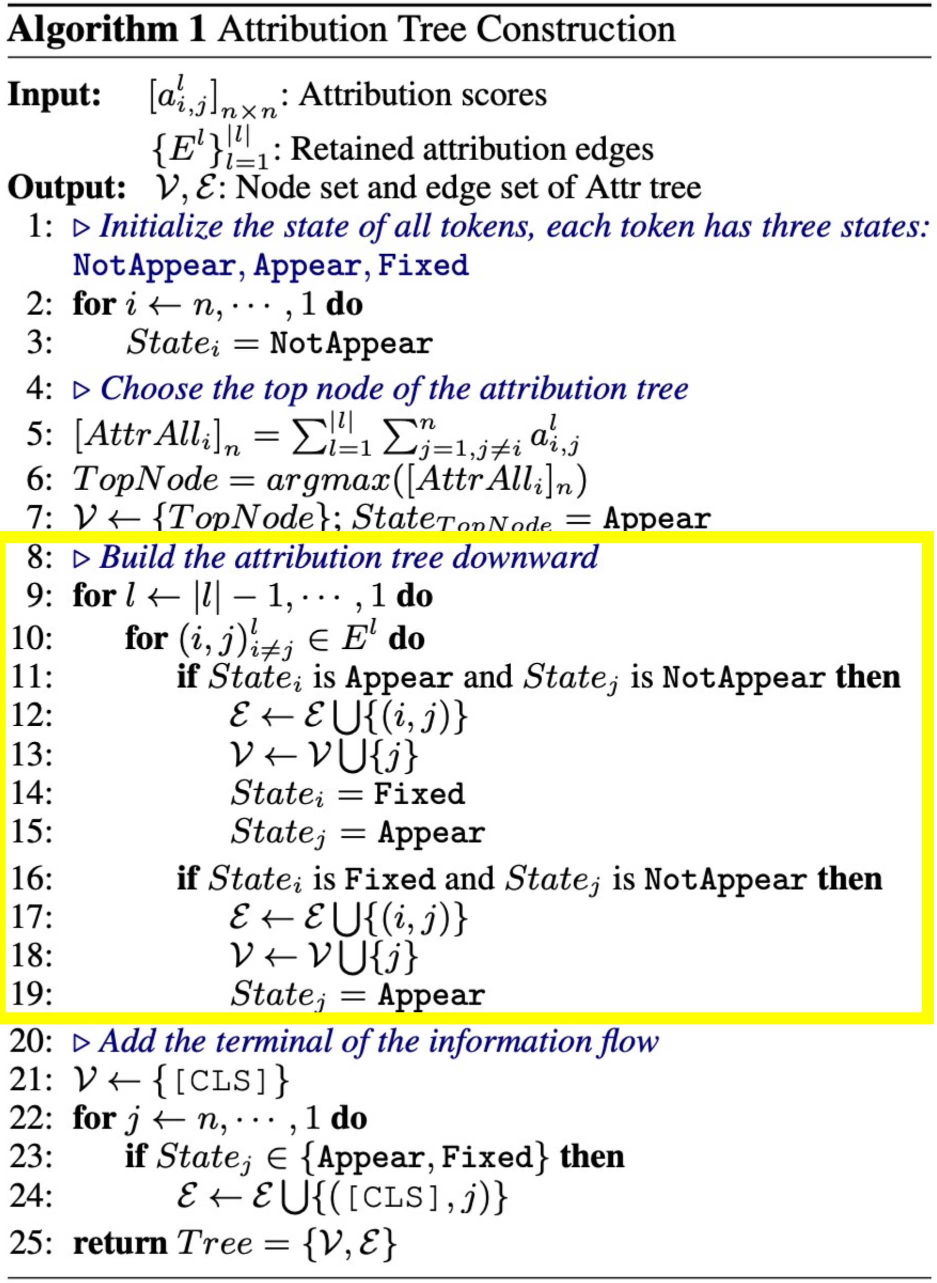
$\{E^l\}^{|l|}_{l=1}$: Retained attribution edges

Output: $\mathcal{V}, \mathcal{E}$: Node set and edge set of Attr tree

```
▷ Initialize the state of all tokens, each token has three states: NotAppear, Appear, Fixed
for $i \leftarrow n, \cdots, 1$ do
    $State_i$ = NotAppear
▷ Choose the top node of the attribution tree
$[AttrAll_i]_n = \sum_{l=1}^{|l|} \sum_{j=1, j\neq i}^{n} a^l_{i,j}$
$TopNode = argmax([AttrAll_i]_n)$
$\mathcal{V} \leftarrow \{TopNode\}$; $State_{TopNode}$ = Appear
▷ Build the attribution tree downward
for $l \leftarrow |l|-1, \cdots, 1$ do
    for $(i,j)^l_{i\neq j} \in E^l$ do
        if $State_i$ is Appear and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i,j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_i$ = Fixed
            $State_j$ = Appear
        if $State_i$ is Fixed and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i,j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_j$ = Appear
▷ Add the terminal of the information flow
$\mathcal{V} \leftarrow \{$[CLS]$\}$
for $j \leftarrow n, \cdots, 1$ do
    if $State_j \in \{$Appear, Fixed$\}$ then
        $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{($[CLS]$, j)\}$
return $Tree = \{\mathcal{V}, \mathcal{E}\}$
```

Attribution Tree

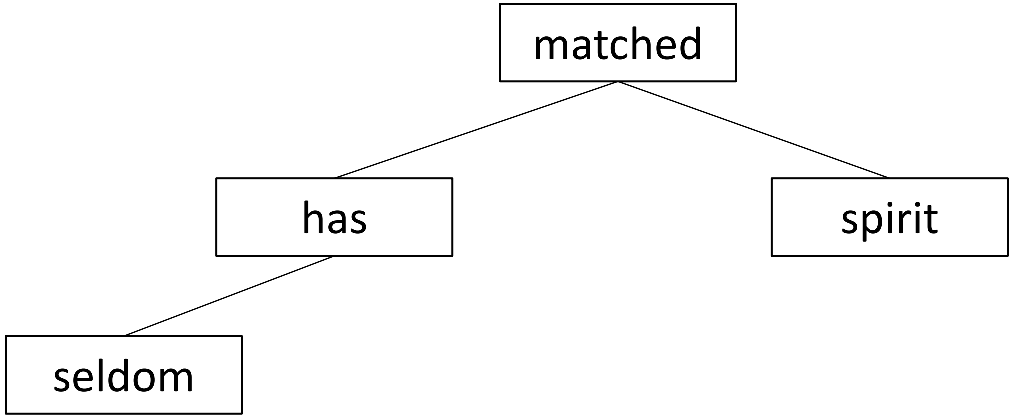


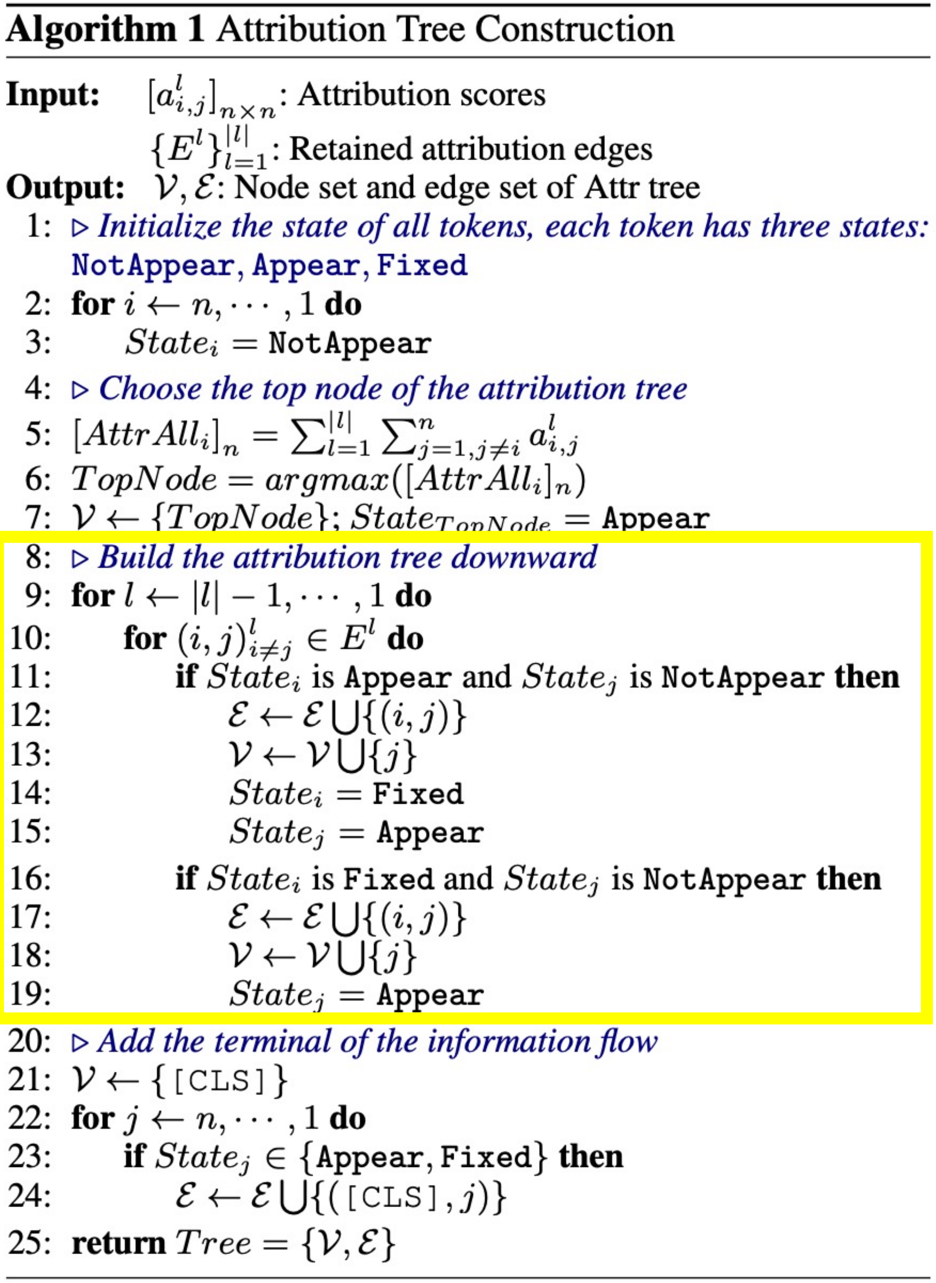


Algorithm 1 Attribution Tree Construction

```
Input:  $[a^l_{i,j}]_{n\times n}$: Attribution scores
        $\{E^l\}^{|l|}_{l=1}$: Retained attribution edges
Output: $\mathcal{V}, \mathcal{E}$: Node set and edge set of Attr tree
▷ Initialize the state of all tokens, each token has three states:
NotAppear, Appear, Fixed
for $i \leftarrow n, \cdots, 1$ do
    $State_i$ = NotAppear
▷ Choose the top node of the attribution tree
$[AttrAll_i]_n = \sum^{|l|}_{l=1} \sum^{n}_{j=1, j\neq i} a^l_{i,j}$
$TopNode = argmax([AttrAll_i]_n)$
$\mathcal{V} \leftarrow \{TopNode\}$; $State_{TopNode}$ = Appear
▷ Build the attribution tree downward
for $l \leftarrow |l| - 1, \cdots, 1$ do
    for $(i, j)^l_{i\neq j} \in E^l$ do
        if $State_i$ is Appear and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_i$ = Fixed
            $State_j$ = Appear
        if $State_i$ is Fixed and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_j$ = Appear
▷ Add the terminal of the information flow
$\mathcal{V} \leftarrow \{$[CLS]$\}$
for $j \leftarrow n, \cdots, 1$ do
    if $State_j \in \{$Appear, Fixed$\}$ then
        $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{($[CLS]$, j)\}$
return $Tree = \{\mathcal{V}, \mathcal{E}\}$
```

Attribution Tree

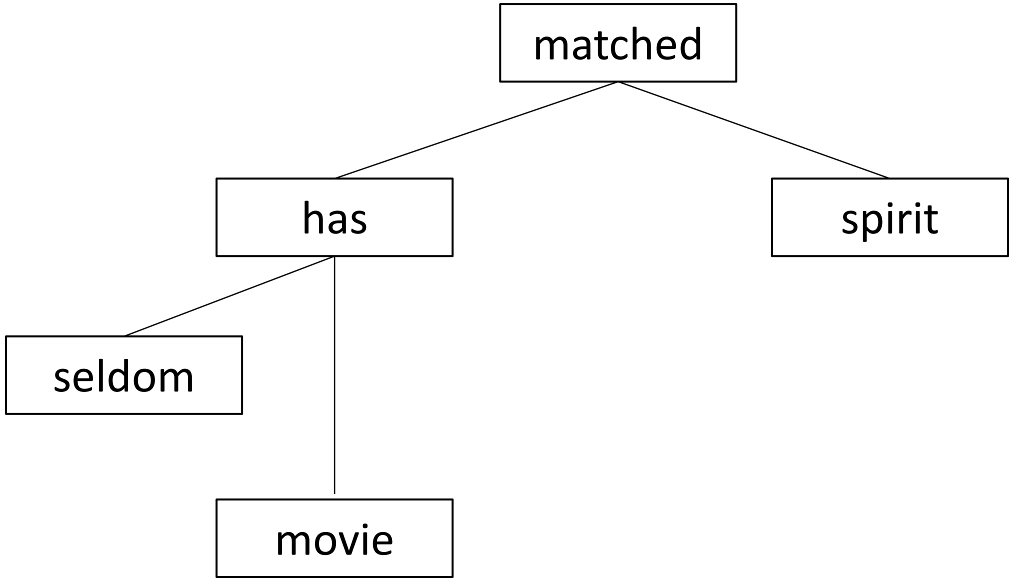


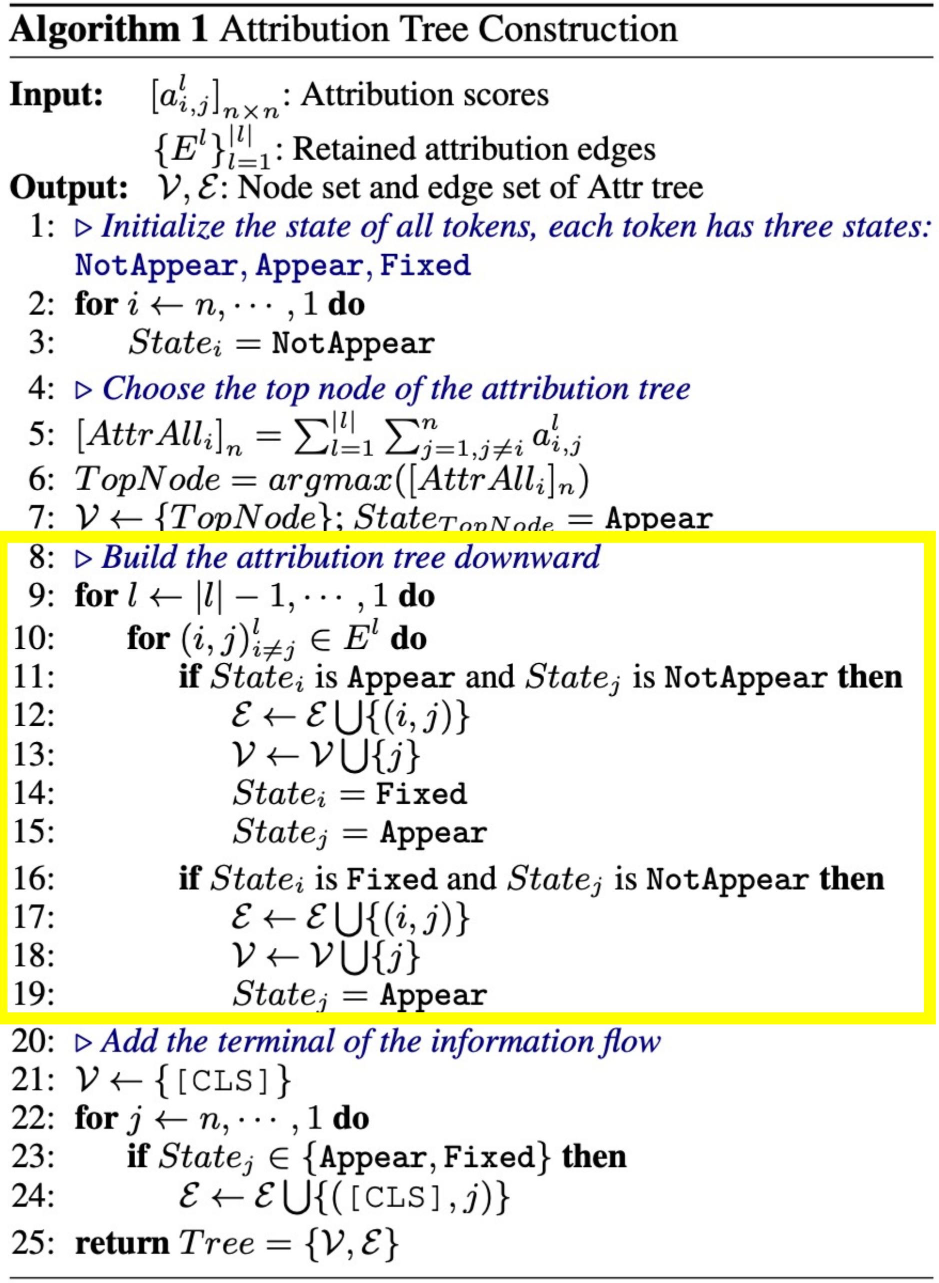


Algorithm 1 Attribution Tree Construction

```
Input: $[a^l_{i,j}]_{n \times n}$: Attribution scores
       $\{E^l\}_{l=1}^{|l|}$: Retained attribution edges
Output: $\mathcal{V}, \mathcal{E}$: Node set and edge set of Attr tree
▷ Initialize the state of all tokens, each token has three states: NotAppear, Appear, Fixed
for $i \leftarrow n, \cdots, 1$ do
    $State_i$ = NotAppear
▷ Choose the top node of the attribution tree
$[AttrAll_i]_n = \sum_{l=1}^{|l|} \sum_{j=1, j \neq i}^{n} a^l_{i,j}$
$TopNode = argmax([AttrAll_i]_n)$
$\mathcal{V} \leftarrow \{TopNode\}$; $State_{TopNode}$ = Appear
▷ Build the attribution tree downward
for $l \leftarrow |l| - 1, \cdots, 1$ do
    for $(i, j)^l_{i \neq j} \in E^l$ do
        if $State_i$ is Appear and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_i$ = Fixed
            $State_j$ = Appear
        if $State_i$ is Fixed and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_j$ = Appear
▷ Add the terminal of the information flow
$\mathcal{V} \leftarrow \{$[CLS]$\}$
for $j \leftarrow n, \cdots, 1$ do
    if $State_j \in \{$Appear, Fixed$\}$ then
        $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{($[CLS]$, j)\}$
return $Tree = \{\mathcal{V}, \mathcal{E}\}$
```

Attribution Tree

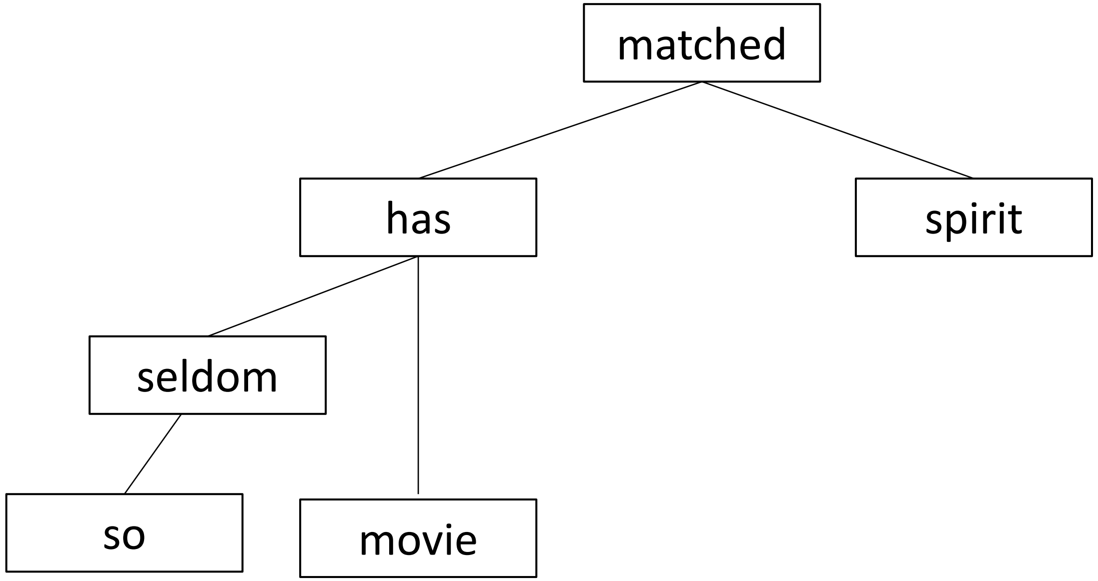


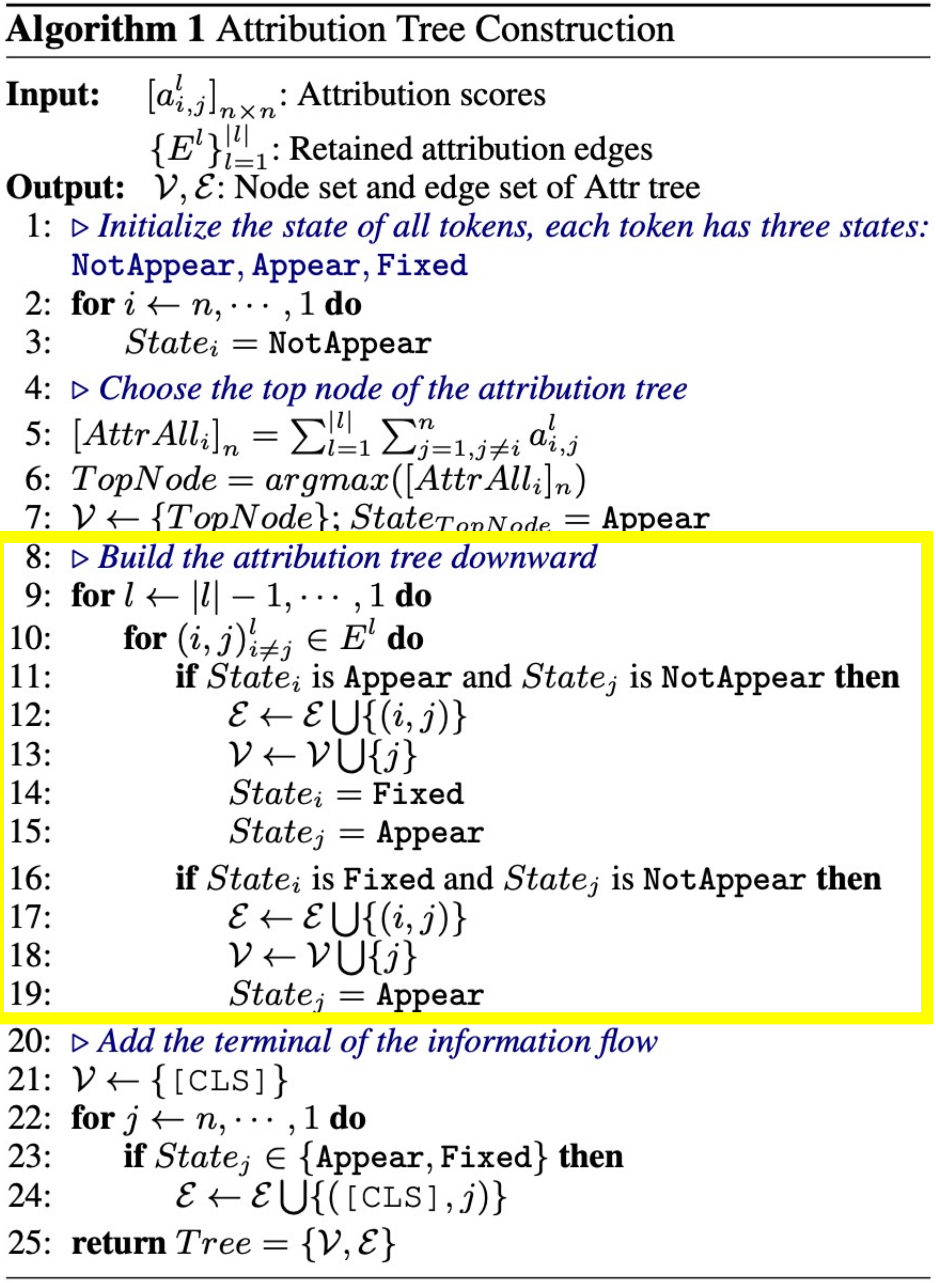


Algorithm 1 Attribution Tree Construction

```
Input:  $[a^l_{i,j}]_{n \times n}$: Attribution scores
        $\{E^l\}_{l=1}^{|l|}$: Retained attribution edges
Output: $\mathcal{V}, \mathcal{E}$: Node set and edge set of Attr tree
$\triangleright$ Initialize the state of all tokens, each token has three states:
  NotAppear, Appear, Fixed
for $i \leftarrow n, \cdots, 1$ do
    $State_i$ = NotAppear
$\triangleright$ Choose the top node of the attribution tree
$[AttrAll_i]_n = \sum_{l=1}^{|l|} \sum_{j=1, j \neq i}^{n} a^l_{i,j}$
$TopNode = argmax([AttrAll_i]_n)$
$\mathcal{V} \leftarrow \{TopNode\}$; $State_{TopNode}$ = Appear
$\triangleright$ Build the attribution tree downward
for $l \leftarrow |l| - 1, \cdots, 1$ do
    for $(i, j)^l_{i \neq j} \in E^l$ do
        if $State_i$ is Appear and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_i$ = Fixed
            $State_j$ = Appear
        if $State_i$ is Fixed and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i, j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_j$ = Appear
$\triangleright$ Add the terminal of the information flow
$\mathcal{V} \leftarrow \{$[CLS]$\}$
for $j \leftarrow n, \cdots, 1$ do
    if $State_j \in \{$Appear, Fixed$\}$ then
        $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{($[CLS]$, j)\}$
return $Tree = \{\mathcal{V}, \mathcal{E}\}$
```

Attribution Tree

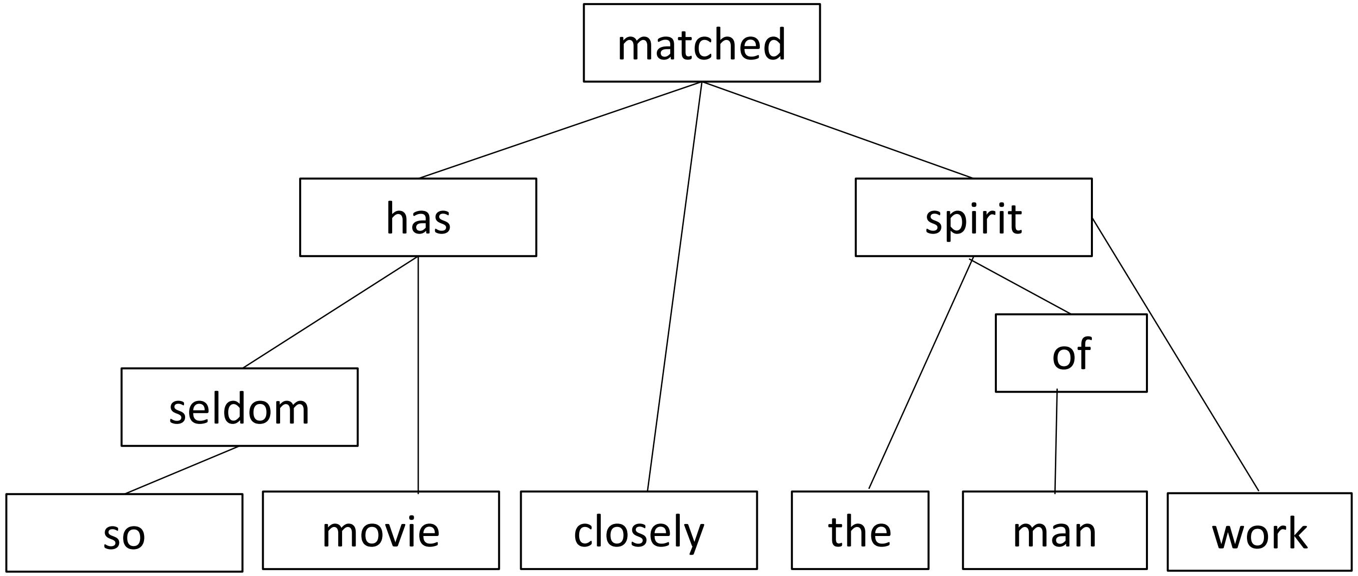


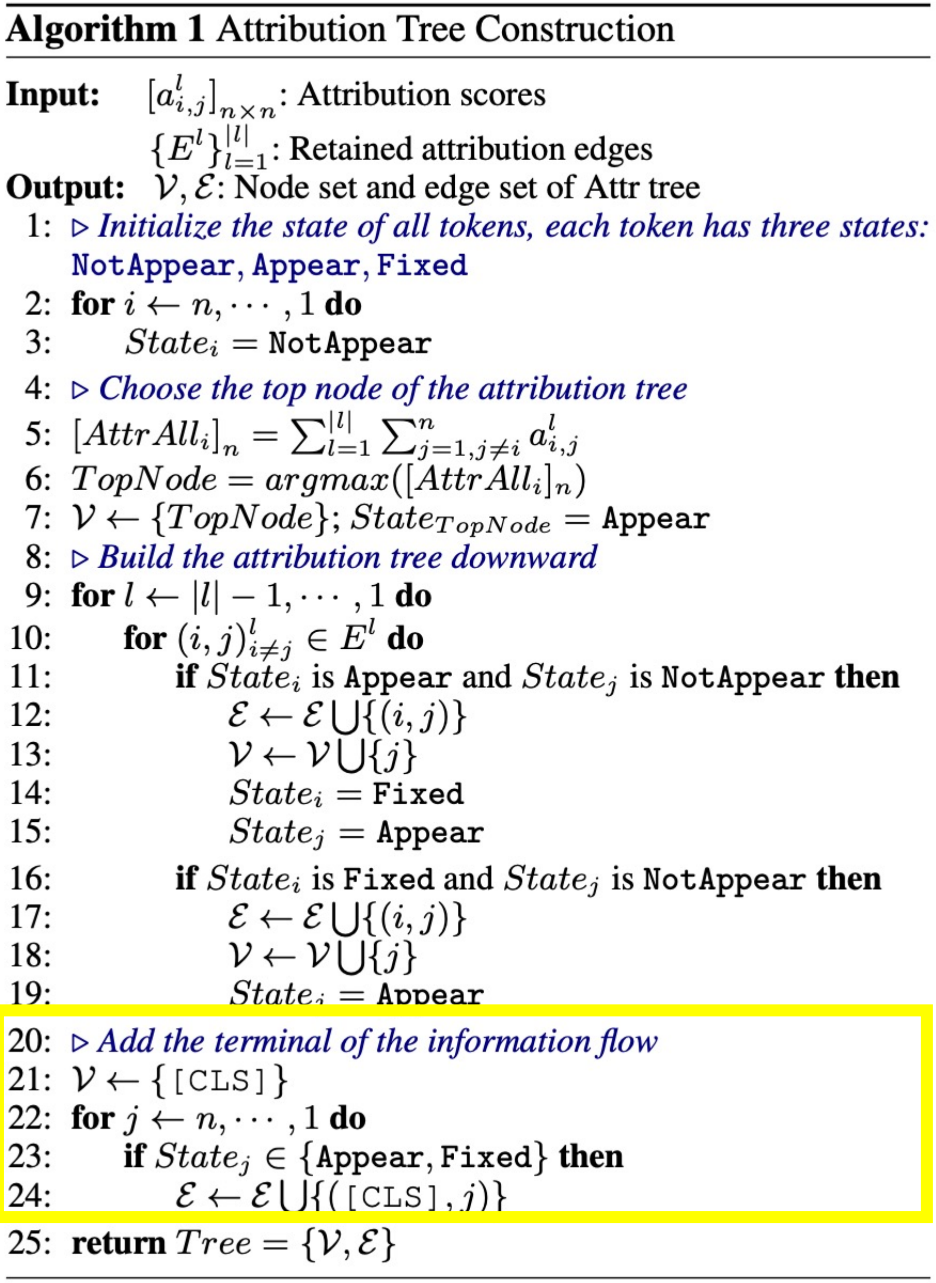


Algorithm 1 Attribution Tree Construction

```
Input:  $[a^l_{i,j}]_{n \times n}$: Attribution scores
        $\{E^l\}^{|l|}_{l=1}$: Retained attribution edges
Output: $\mathcal{V}, \mathcal{E}$: Node set and edge set of Attr tree
▷ Initialize the state of all tokens, each token has three states:
  NotAppear, Appear, Fixed
for $i \leftarrow n, \cdots, 1$ do
    $State_i$ = NotAppear
▷ Choose the top node of the attribution tree
$[AttrAll_i]_n = \sum_{l=1}^{|l|} \sum_{j=1, j \neq i}^{n} a^l_{i,j}$
$TopNode = argmax([AttrAll_i]_n)$
$\mathcal{V} \leftarrow \{TopNode\}$; $State_{TopNode}$ = Appear
▷ Build the attribution tree downward
for $l \leftarrow |l| - 1, \cdots, 1$ do
    for $(i,j)^l_{i \neq j} \in E^l$ do
        if $State_i$ is Appear and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i,j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_i$ = Fixed
            $State_j$ = Appear
        if $State_i$ is Fixed and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i,j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_j$ = Appear
▷ Add the terminal of the information flow
$\mathcal{V} \leftarrow \{$[CLS]$\}$
for $j \leftarrow n, \cdots, 1$ do
    if $State_j \in \{$Appear, Fixed$\}$ then
        $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{($[CLS]$, j)\}$
return $Tree = \{\mathcal{V}, \mathcal{E}\}$
```

Attribution Tree

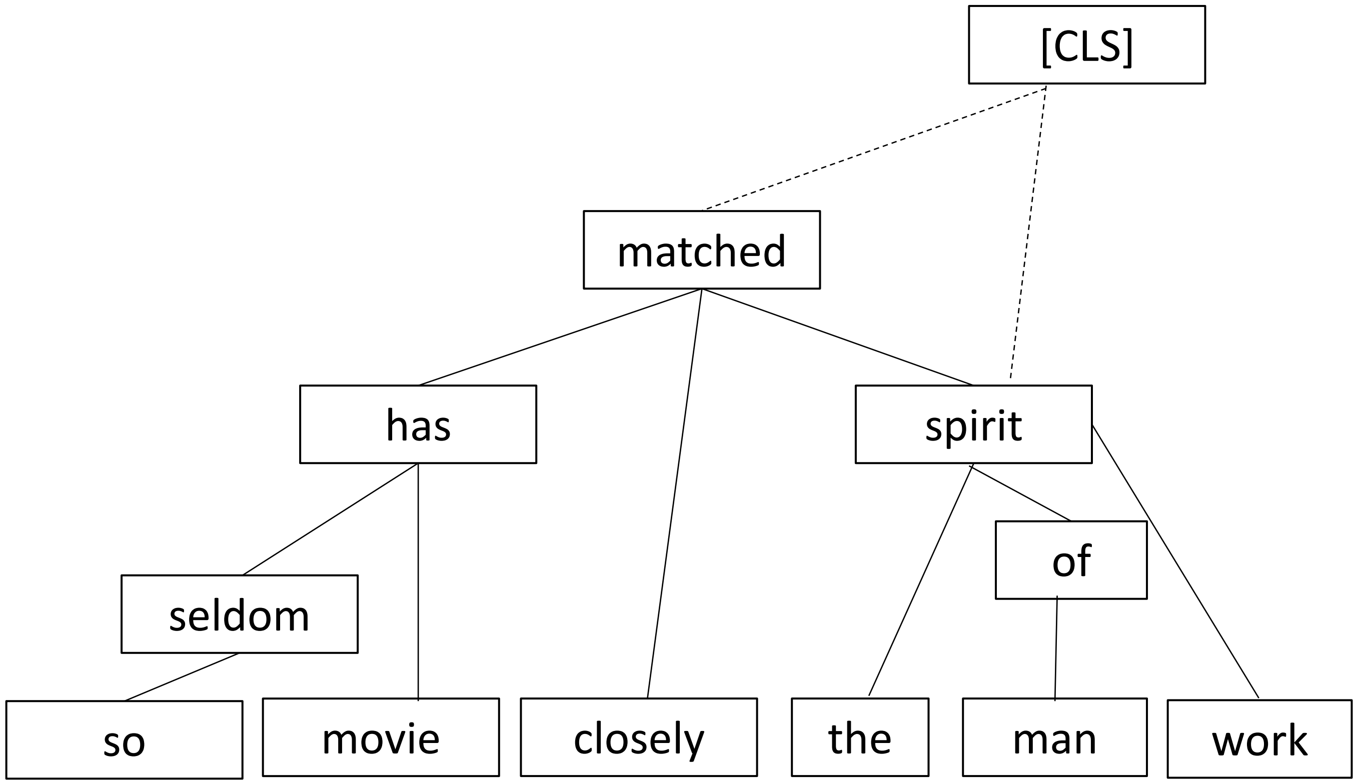




Algorithm 1 Attribution Tree Construction

```
Input: $[a^l_{i,j}]_{n\times n}$: Attribution scores
       $\{E^l\}^{|l|}_{l=1}$: Retained attribution edges
Output: $\mathcal{V}, \mathcal{E}$: Node set and edge set of Attr tree
▷ Initialize the state of all tokens, each token has three states: NotAppear, Appear, Fixed
for $i \leftarrow n, \cdots, 1$ do
    $State_i$ = NotAppear
▷ Choose the top node of the attribution tree
$[AttrAll_i]_n = \sum_{l=1}^{|l|} \sum_{j=1,j\neq i}^{n} a^l_{i,j}$
$TopNode = argmax([AttrAll_i]_n)$
$\mathcal{V} \leftarrow \{TopNode\}$; $State_{TopNode}$ = Appear
▷ Build the attribution tree downward
for $l \leftarrow |l|-1, \cdots, 1$ do
    for $(i,j)^l_{i\neq j} \in E^l$ do
        if $State_i$ is Appear and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i,j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_i$ = Fixed
            $State_j$ = Appear
        if $State_i$ is Fixed and $State_j$ is NotAppear then
            $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{(i,j)\}$
            $\mathcal{V} \leftarrow \mathcal{V} \bigcup \{j\}$
            $State_j$ = Appear
▷ Add the terminal of the information flow
$\mathcal{V} \leftarrow \{$[CLS]$\}$
for $j \leftarrow n, \cdots, 1$ do
    if $State_j \in \{$Appear, Fixed$\}$ then
        $\mathcal{E} \leftarrow \mathcal{E} \bigcup \{($[CLS]$, j)\}$
return $Tree = \{\mathcal{V}, \mathcal{E}\}$
```

Attribution Tree



[CLS]
matched
has
spirit
seldom
of
so
movie
closely
the
man
work