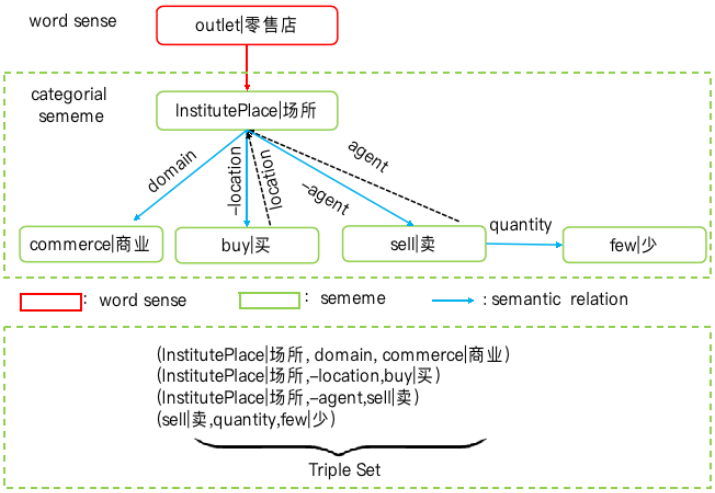
Sememe Tree Prediction for English-Chinese Word Pairs

Baoju Liu, Lei Hou, Juanzi Li
Department of Computer Science and Technology, Tsinghua University, China 100084

Introduction

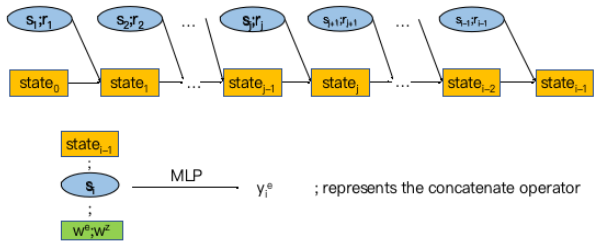
- Sememes: the minimum unambiguous indivisible semantic units of human languages in the field of linguistics.
- Sememe knowledge has improved many natural language processing tasks, e.g., word sense disambiguation with sememes, pre-train word embedding enhancement, semantic composition modeling, event detection, relation extraction augement, sentiment analysis and textual adversarial attack.
- HowNet : Chinese and English bilingual sememe base. 2214 sememes, 116 semantic relations. Each word sense in HowNet is described by a sememe tree.

Problem 1. The hypothesis space of tree is too large

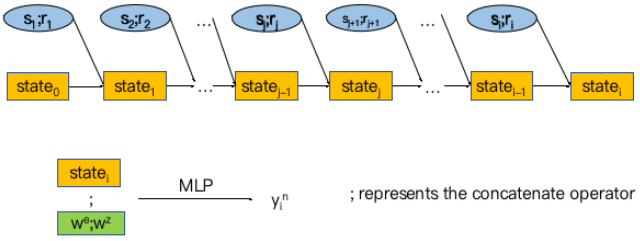


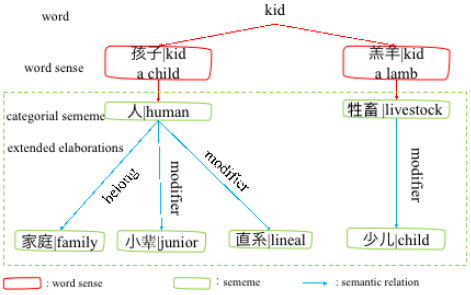
Path Generation : construct two classifiers to generate trees in a depth-first manner

- Edge Generator: the path (root to current node)



- Node Generator: the path (root to current node)





- Tree Generator

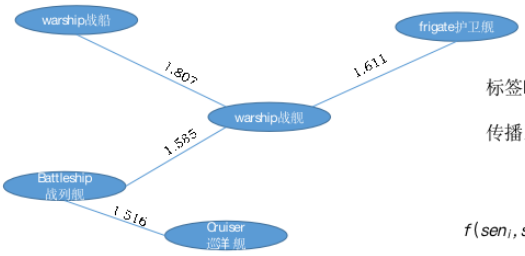
Algorithm 1: Tree Generator

Input: word sense sen, a sememe tree $tree$, a node s_i, the path from root to the node P_{s_i}
Output: a sememe tree $tree$

```
E=Edge Generator(sen, s_i, P_{s_i})) generate a list of all correct head-edges
Sort the output correct results according to the score from high to low
if E_0 == "null" then
    return tree
else
    for e_j in E do
        N_i = Node Generator(w_s, e_j, P_{s_i} + e_j) generate a list of all correct tail nodes
        for n_k in N_i do
            Tree Generator(sen, tree, n_k, P_{s_i} + e_j + n_k)
        end
    end
end
```

Label Propagation : words with similar semantics are more likely to have similar sememe trees

- Word Sense Graph



边的权重：相似度

$$cos(w_i^e, w_j^e) + cos(w_i^z, w_j^z)$$

标签label：associate each triplet with a dimension

传播函数：

$$\tilde{l}_i = g\Big(\bigtimes_{sen_j \in N_i} f(sen_i, sen_j)\, l_j\Big)$$

$$f(sen_i, sen_j) = \frac{e^{a(cos(w_i^e, w_j^e) + cos(w_i^z, w_j^z))}}{\sum_{sen_j \in N_i} e^{a(cos(w_i^e, w_j^e) + cos(w_i^z, w_j^z))}}$$

Algorithm 2: Triple Set to a Tree

Input: a triple set T, a categorial sememe $\bar{s}_{sen}$
Output: a sememe tree $tree$

```
Build a tree with root node s̄_sen
Create a set N_tree = {s̄_sen} represents the set of all nodes of the tree
l_past = 0
while len(N_tree) > l_past do
    l_past = len(N_tree)
    for (h,r,t) in T do
        if (h ∈ N_tree) then
            N_tree.add(t)
            add(h,r,t) to tree at node h
        end
    end
    return tree
end
```

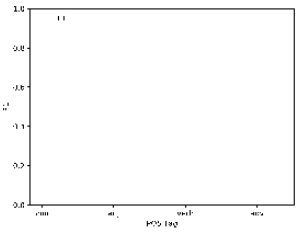
Experiments

DataSets

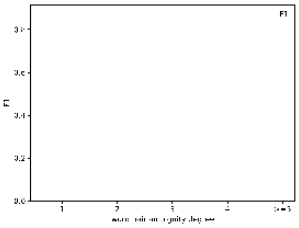
Words		Word Pairs		
Chinese	English	Total	With vectors	With vectors and sememe trees
104,027	118,347	208,276	93,081	44393

Experimental results

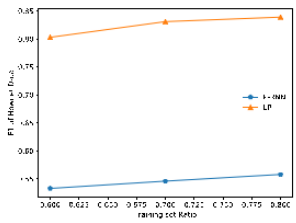
method	Edge generator	Node generator	P-RNN	LP
P	0.824	0.765	0.576	0.818
R	0.853	0.739	0.542	0.863
F	0.838	0.752	0.558	0.840



词性



歧义度



训练数据大小