Cross-lingual Knowledge Validation Based Taxonomy Derivation from Heterogeneous OnlineWikis

Zhigang Wang¹, Juanzi Li², Shuangjie Li¹, Mingyang Li¹, Jie Tang¹, Kuo Zhang², Kun Zhang²
1 TSINGHUA UNIVERSITY, BEIJING, CHINA
2 SOGOU INCORPORATION, BEIJING, CHINA

Introduction

- Creating KBs based on the crowd-sourced wikis has attracted significant research interest in the field of intelligent Web.
- However, the user-generated subsumption relations in the wikis and the semantic taxonomic relations in the KBs are not exactly the same.
- Current taxonomy derivation approaches include:
  - The heuristic-based methods
  - The corpus-based methods

- Here, we systematically study the problem of cross-lingual knowledge validation based taxonomy derivation from heterogeneous online wikis.
- The problem of cross-lingual taxonomic relation prediction is at the heart of our work.

Approach

Given two wikis \( W_1, W_2 \) in different languages (English and Chinese here) and the set of cross-lingual links \( CL \), **Cross-lingual Taxonomy Derivation** is a cross-lingual knowledge validation based boosting process, by simultaneously learning four taxonomic prediction functions \( f_{en}, f_{zh}, g_{en} \) and \( g_{zh} \) in \( T \) iterations.

1. Weak Classifier
   - We utilize the binary classifier for the basic learner and use the Decision Tree as our implementation.

2. Boosting Model
   - **Active Set**: the set of training data.
   - **Pool**: the set of all labeled data.
   - **Unknown Data Set**: the set of unlabeled data.

Learning Process.
- Train a hypothesis on current active set.
- Re-weight the weight vector.
- Predict \( U \) using current classifier and validate the results using \( CL \).
- Expand \( P \) and update \( U \).
- Resample \( A \) with the constant size.

Experiments

Comparison Methods

- **Heuristic Linking (HL)**: only uses the linguistic heuristic features, and trains the taxonomic relation prediction functions separately using the decision tree model.
- **Decision Tree (DT)**: uses both the linguistic heuristic features and the structural features, and trains the taxonomic relation prediction functions separately using the decision tree model.
- **Adaptive Boosting (AdaBoost)**: uses the same basic learner, and iteratively trains the taxonomic relation prediction functions using the real AdaBoost model.

Conclusion and Future Work

- DAB gives a new way for language processing tasks using cross-language resources.
- The future work contains automatically learning more cross-lingual validation rules and conducting more experiments in other languages.

References