Semantic Annotation with RescoredESA: Rescoring Concept Features Generated From Explicit Semantic Analysis

Zhuoren Jiang Dalian Maritime University Miao Chen*

Xiaozhong Liu Indiana University Bloomington Indiana University Bloomington

INTRODUCTION

Annotating text with a set of concepts reveals semantically important entities and thus is an important step in semantic annotation. This study we try to extend a semantic annotation approach, Explicit Semantic Analysis (ESA), by enriching it with context and knowledge base.

ESA formula:

$$v_E = v_d \cdot M^T$$

RescoredESA:

$$e_i + \alpha \left(\frac{\sum_{j=1}^{K} LS(C_{E_i}, C_{I_j})}{k} \right)$$

$$r_i = \frac{1 + \alpha}{k}$$

THE RESCORED ESA APPROACH

Text

"I'm looking for a replacement radio/tape player for a 1984 Toyota Tercel..."



In-text Wikipedia Concepts

String	Matched Wikipedia concept
Radio	[radio]
Tape	[Cassette deck]
Toyota Tercel	[Toyota Tercel]



Wikipedia Concepts from ESA

[Jimmy Van Ostrand] [Ostrander, Minnesota] [USS Tercel (AM-386)] [Toyota Paseo] [Toyota Tercel] [Ostrander, Ohio] [John Ostrander] [Toyota Origin] [Toyota Raum] [T. C. Ostrander]

 $\mathcal{IS}(C_{Ei}, C_{Ii})$

ESA vector

vector

In-text Wikipedia

ESA vector

 C_{E1}

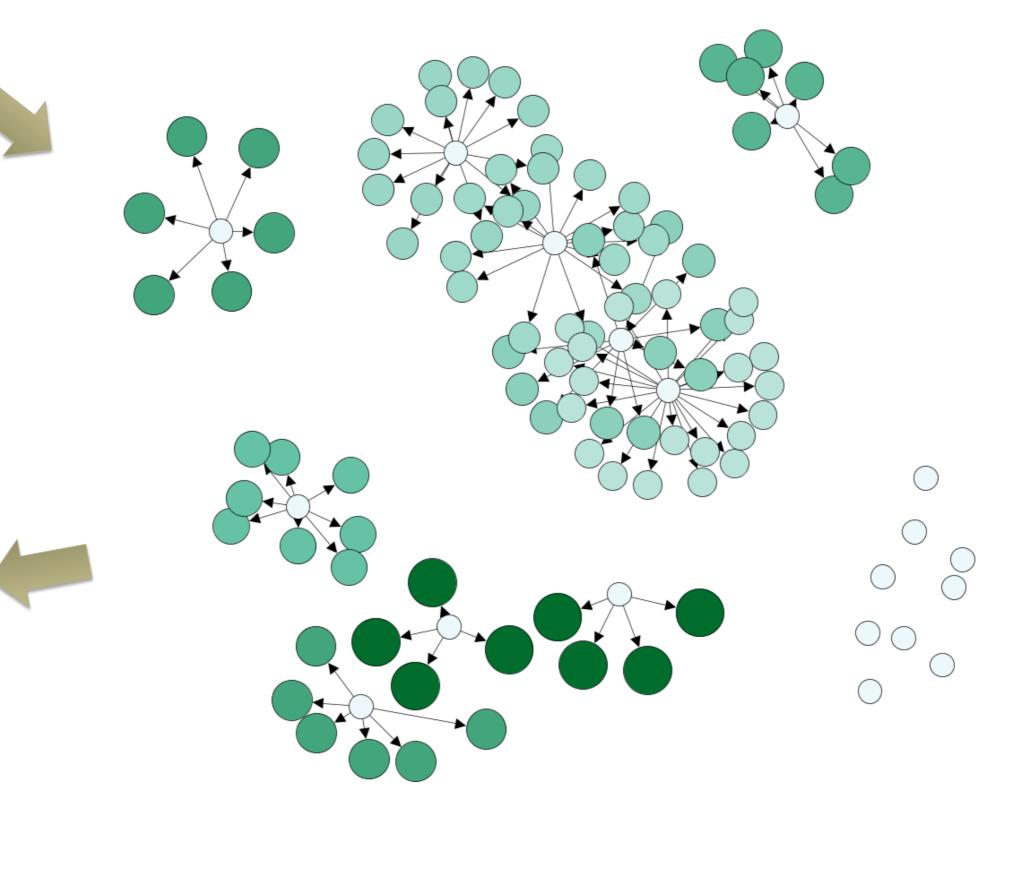
 C_{E2}

 $E = [C_{E1}, C_{E2}, ..., C_{En}]$

 $I = [C_{I1}, C_{I2}, ..., C_{Ik}]$

In-text Wikipedia vector

Wikipedia Concept Network



Wikipedia concepts from RescoredESA

[Toyota Yaris] [Toyota Corolla] [Toyota Starlet] [Butter-fly doors] [Toyota Platz] [Geo Prizm] [Toyota Tercel] [Toyota E engine] [Toyota A engine] [Toyota Sprinter]

EXPERIMENT

In order to examine performance of RescoredESA, we use the 20 newsgroups data, a widely used news data set for text classification. RescoredESA, along with other comparable approaches, are evaluated using their performance on classifying 20 newsgroup - that is, first representing text with these approaches and then perform text classification based on the representation results.

RESULTS

The results show that for the three text categorization methods other than BalanceWinnow, the general trend is that: Concat (BOW + RescoredESA) > BOW > InText > RescoredESA > ESA. For BalancedWinnow, the order is similar but with ESA and RescoredESA reversed: Concat(BOW + RescoredESA) > BOW > InText > ESA > Rescore- dESA.

Parameter α Maximum Entropy → Naive Bayes → SVM 0.75 -0.70 -0.65 0.60 0.55 -0.50 -0.40

Parameter y

