Ryerson University
Data Science & Analytics Graduate Studies

2018-2019 M.Sc. in Data Science and Analytics

Synonym Detection with Knowledge Bases

Imran Barolia **Supervisor: Professor Pawel Pralat**

Objective

For this study existing knowledge bases is used to get synonyms seed from WordNet and parameter is trained using bilinear scoring function. This is resulted in model predicting more word that could be similar but might not be part of synonyms seed or knowledge base.

Background

This paper aims to identify language inconsistencies within each tweet. Since identifying these synonyms from knowledge bases (WordNet) is difficult because it could be specific to domain corpus and may not be used widely in general discussion or might not exists in knowledge base at all.

Since existing knowledge base is manually curated by human and are limited, maintaining these knowledge bases on regular bases is very difficult and costly. But since these are developed by human, it provides very strong evidence if two words are synonyms.

Methodology

$$df(s) = \frac{|\text{Synonyms List}[s]|}{|\text{Tweets}|}$$

$$\text{Score}(u,v) = x_u W x_v^t$$

Bilinear scoring function

$$p(u|v) = \frac{\exp(x_u^T x_v + x_u^T c_v)}{Z}$$

Conditional probability

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Results

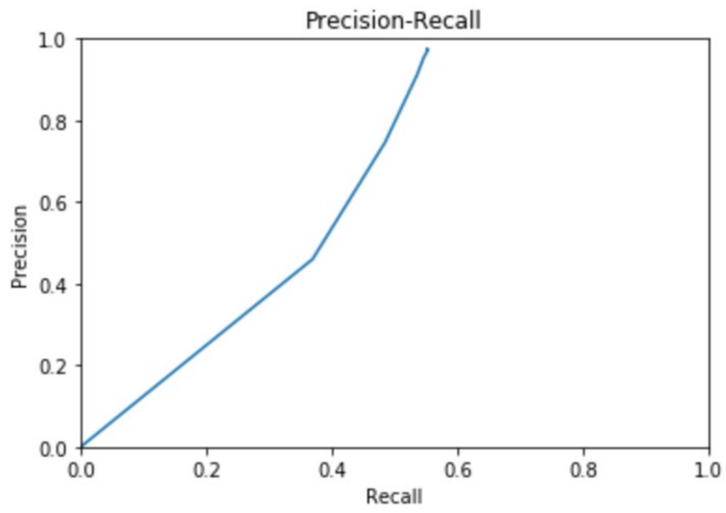
1- If the synonym that have been generated is actually synonym/similar (WordNet) then True Positive.

2- If the synonym that have been generated is not a synonym/similar (WordNet) then False Positive.

3- If initial seed has not generated any score for candidate synonym then False Negative.

4- Rest generated candidate synonym are True Neg

	Predicted: NO	Predicted: YES
Actual: NO	TN=7581334	FP=2612
Actual: YES	FN=8178	TP=7756



Conclusions

As we know that distributed approach usually gets high precision but incorporating scoring model, this study shows increase in precision score.

Future work of this study is to generalize these approaches to other domain and build a strategy to auto detect domain level corpus and apply these strategy without fist filtering domain data.