Beyond Trigram: Learn to Distinguish Fake Article
11-761 Project Report

Yanran Hao
xxx@andrew.cmu.edu

Yulan Huang
yyy@andrew.cmu.edu

Ang Lu
alu1@andrew.cmu.edu

Heqing Ya
heqingy@andrew.cmu.edu

April 24, 2016

Abstract

1 Introduction

2 Feature Generation

3 Experiments

3.1 Data Description

The training set is 1000 articles, 500 real and 500 fake, of varying length. The development set is 200 articles, 100 fake, 100 real. Articles in development set is truncated to meet the length distribution in Table 3.1. Besides these two sets, we also use a 100 million word corpus of Broadcast News articles as external source for generation of specific feature.

3.2 Data Preprocessing

Articles from training set are truncated following the document length distribution in Table 3.1. The number of truncated training articles is 10065. The sentence per article distribution of original training set, truncated

#sentence	#article
1	20
2	10
3	10
4	10
5	10
7	10
10	10
15	10
20	10

Table 1: Article length distribution of dev set

training set and development set is shown in Figure 1.

3.3 Classifier Choice

Popular classifiers for binary classification task are chosen as candidates, including KNN, logistic regression, SVM, gradient boosting (xgboost). Testing individual feature on development set and cross validation on training set, xgboost outperforms all the other algorithm by about 5%. Also it is very fast comparing to SVM. So we choose xgboost
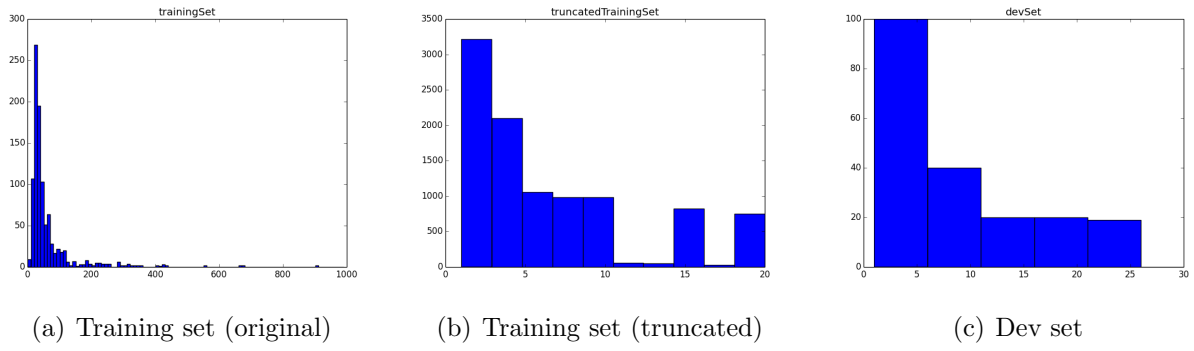
1

(a) Training set (original) (b) Training set (truncated) (c) Dev set

Figure 1: Article length distribution of different sets

as the final classier.

3.4   Result

4   Conclusions