Keyword Extraction from Research Abstracts

1 Introduction

Most popular topic and keyword extraction methods are frequency and engineered-features based. Statistical topic modeling methods such as Latent Dirichlet Allocation excel when extracting a set of meaningful and relevant words from text. These methods however, require a large enough amount of data and suffer when presented with limited text as is only the abstract of a paper. Keyword extraction algorithms such as RAKE rely on co-ocurrace frequency of candidate keywords. We propose a data-driven method to automatically extract relevant and informative words from text based on semantic relationships. In this work, we aim to measure the effectiveness of a keyword extraction method that relies on selecting candidate keywords based on semantic relations, in particular cosine similarity of chunked terms. We further provide a performance evaluation analysis based on author-specified keywords. We describe the derivation of this method, as well as research its applications.

1.1 Problem Description

In order to fascilitate inter-disciplinary research within the University of Texas at El Paso, we propose a system that intelligently recommends which communities of research and practice faculty members should be members of based on their publication data. Popular text classification models are trained using frequency based feature vectors from a fixed vocabulary, as well as hand engineered features such as title and paper keyword words. We propose an automatic feature extraction scheme that produces dense feature vectors of word embedding n-grams. These feature n-grams are considered keywords, being especially informative within the text. We analyse and compare our method's prediction performance using these keyword feature vectors, as well as discuss results on a subjective basis based on prior knowledge of the documents' contents.

2 Keyword and Topic Extraction

The purpose of keyword and topic extraction is to retrieve a set of words from an input text such that these words concisely describe the main ideas inherent in the document. This extracted set of words should be small relative to the document's length.

3 Keyword Extraction: Finding Relevant Words Within Text

4 Implementation Details

In the following subsections, we describe necessary data processing and dataset dependencies of our method.

4.1 Embeddings Dataset

Language models are effective capturing semantic relations of words in vector embedding space. A popular approach to language modeling is to learn the model a using neural network. Along with learning a model of a language, a mapping of words from strings to vectors is also learned. This mapping is then used to transform words as strings to dense vectors in Euclidean space, where semantic relations are captured. The embedding size is a meta parameter of a language model, and is specified before learning. Larger embeddings usually lead to better model performance and better semantic relationship capturing. Due to memory limitations, we work with embeddings of size 50.

We use the GloVe word embedding dataset from Stanford's Natural Language Processing department for mapping our text words to embedding vectors. We also normalize all embeddings to have unit norm, reducing cosine distance computations dot product operations.

4.2 Data Preprocessing

We remove all non alpha-numeric characters from every document.

4.3 Dataset Description

We consider a collection of ACM paper abstracts, along with their corresponding author-specified keywords. Each abstract corresponds to one of three communities of research: Cyber-Security, Smart Cities, and Undergraduate Research.

5 Performance Evaluation

We compare our proposed keyword extraction methods with other popular keyword extraction techniques. We test with a set of ACM paper abstract along with their corresponding paper keywords. We consider the paper keywords as the ground truth, the correct words that should always be extracted with any method. This test is done for validation purposes i.e. quantifying our method's performance. It should be noted that paper keywords are sometimes not explicitly found within the abstract.

In order to evaluate keyword performance metric e.g. recall and fmeasure, we compute the abstract's **mean token metric percentage**. For example, consider a keyword *w*, made up of three tokens i.e. single words a, b, and c. If only two of the three tokens are extracted e.g. b and c, then the token recall percentage is 66. We compute this metric for all keywords in the abstract, and average them. This is the abstract's mean token recall. We compute this for recall, precision, and fmeaure.

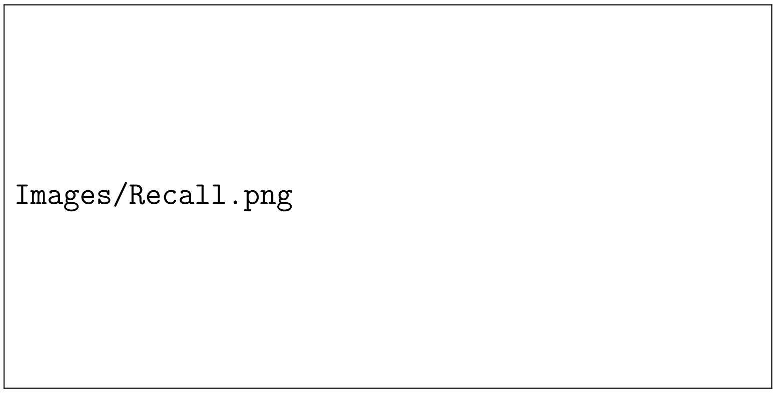

Figure 1: Mean Token Recall

Images/Precision.png

Figure 2: Mean Token Precision

Images/Fmeasure.png

Figure 3: Mean Token F-measure

6 Application: Automatic Keyword Extraction in Research Abstracts

One application of our method is to extract keywords and topics from research abstracts. In an effort to fascilitate interdisciplinary research collaboration within various communities of practice in the UTEP, we are creating systems to recommend and assign faculty members to research communities based on the member's research preferences and areas of expertise.