NLP Final Project Proposal: Keyword Tagger

By Leeya Howley, Vishesh Goyal, Chan Hyun Yoo, Corina Luca

Introduction

In an era of rapidly expanding digital content, efficiently organizing and retrieving information from vast databases is essential. Our project focuses on identifying effective methods for extracting keywords from texts, central to automating data organization and retrieval. We will experiment optimally deriving relevant keywords to enhance information systems' accuracy and usability. By implementing keyword tagging algorithms, we aim to improve metadata generation, enabling more precise document filtering in databases. This effort seeks to facilitate better information management and accessibility, addressing the growing challenge of handling large data volumes efficiently.

Description/Evaluation

The core task of our project is to generate a list of keywords for a given document, with these keywords ranked according to their importance or relevance to the document's content. To evaluate the effectiveness and accuracy of our keyword generation algorithm, we have devised a comprehensive evaluation plan that benchmarks the algorithm's output against a curated set of documents that have been manually tagged with keywords. This comparison between the algorithm-generated keywords and the manually tagged "answer key" serves as the primary method for assessing the algorithm's performance.

To ensure robustness and reliability in our evaluation, we will employ cross-validation techniques. Cross-validation involves dividing the dataset into several parts, using some for training the algorithm, developing, and testing. We will be using the following documents respectively for each subcorpora: the *Inspec* database comprised of 2000 short documents from science journal abstracts, *DUC 2001* is comprised of 308 mid-length news articles organized in 30 topics, and *NUS* which contains 211 long scientific conference papers (4~12 pages). In summary, this method helps in assessing the algorithm's generalizability and performance consistency across different subsets of data.

The evaluation of our keyword tagging algorithm will heavily rely on rank-aware evaluation metrics that are particularly suitable for tasks involving ranking or ordering of results. Specifically, we will utilize the Mean Reciprocal Rank (MRR), which provides insight into the position of the first relevant keyword; the Mean Average Precision (MAP), which gives an overall precision score across all documents by taking into account the order of the keywords; and the Normalized Discounted Cumulative Gain (nDCG), which measures the gain of the keywords based on their positions in the result list, penalizing highly ranked irrelevant keywords. These metrics together will offer a multi-faceted view of the algorithm's effectiveness in not just identifying relevant keywords, but also in ranking them in order of importance, providing a thorough and nuanced evaluation of the keyword tagger's performance.

Research Papers

The paper "A News Story Categorization System" by Philip J. Hayes, Laura E. Knecht, and Monica J. Cellio describes a pilot commercial application employing natural language processing techniques to categorize news stories into broad topics without full semantic or syntactic analysis. Utilizing pattern-matching techniques and knowledge-based rules to identify relevant fragments within the text, the system achieves categorization accuracy slightly lower than human performance. The search for words and phrases is organized around pattern sets, a collection of patterns that represent specific words and phrases associated with a concept, used to identify relevant content within a text for categorization or tagging. With an average accuracy of 93% on a sample of 500 stories, the paper highlights the system's efficiency and potential for practical application in improving text processing tasks like routing and archiving news stories. Overall, these techniques can effectively automate the task of text categorization, providing a solid foundation for developing a keyword tagger.

Anette Hulth's paper, “Improved Automatic Keyword Extraction Given More Linguistic Knowledge,” tackles the similar aim of our project to effectively extract keywords from documents. Hulth provides an interesting contrast from our intended strategy, suggesting exploring beyond purely statistical approaches (like TF-IDF and graph-based algorithms): she emphasizes the idea of adding linguistic knowledge (such as syntactic features) to improve the precision of keyword extraction. Such research could prove an interesting comparison to our project or even outline future methods of improvement to our work. Regardless of term selection strategy, between n-grams, NP-chunks, or POS tag pattern matching, Hulth found that the inclusion of POS tags as a feature improves results. Her study involves the Inspec database abstracts, as we intend to use, offering a shared ground for comparing results.

Kazi Saidul Hasan and Vincent Ng's “Automatic Keyphrase Extraction: A Survey of the State of the Art” describes a survey of different supervised and unsupervised methods for keyword extraction and identifies common errors in the field. The article provides helpful information about features of different types of corpora, where length, structure, and topic can affect the success of the keyword extraction results. These are useful details to keep in mind when creating our system and working with different texts. Hasan and Ng discuss common evaluation methods of precision, recall, and F-score, but highlight a key issue that these methods may unfairly penalize systems for semantic equivalents or variants of keywords. They suggest checking your system against external resources to improve the system's “understanding” of document topics. The paper outlines four common errors of keyword extraction: overgeneration, infrequency, redundancy, and evaluation errors, and again suggests the incorporation of external resources to address the lack of background knowledge the system has. While likely out of the scope of our project, deep diving into these errors and connecting to external sources would be an interesting continuation of our intended work.

The article "Simple Unsupervised Keyphrase Extraction using Sentence Embeddings" by Kamil Bennani-Smires, Claudiu Musat, Andreaa Hossmann, Michael Baeriswyl, and Martin Jaggi introduces an innovative approach to keyphrase extraction that leverages sentence embeddings for unsupervised learning using an unsupervised method, EmbedRank. This method contrasts with traditional techniques by focusing on the semantic relationship between phrases and the overall document context, promising a more nuanced and effective identification of relevant keywords. This strategy could significantly complement our project's objectives by offering an alternative, potentially more efficient pathway to enhance our keyword extraction framework and improve the precision of information retrieval systems.

In Youngsam Kim, et al's article "Applying Graph-based Keyword Extraction to Document Retrieval" the authors describe keyword extraction based on the PageRank algorithm that uses graphs to achieve greater results (evaluated by Mean Average Precision) than traditional TF-IDF models. In the graph-based approach, words in a document are filtered based on POS tags and then made into vertices in the graph. Edges between vertices are created depending on the proximity of words to each other on the document. Using the graph, stop-words are removed and TF-IDF weighting can be used on the results of the keyword extraction. In evaluation, the experiment found graph-based keyword extraction to be slightly better than TF-IDF frequency analysis, but combining the two produced the best results. Using this article, we can create a template for our approach in graph-based keyword extraction in creating a graph and filtering words based on POS tags. Additionally, the article provides potential suggestions for steps and enhancements we can take in our project.

Strategy For Solving the Problem

We plan to implement two approaches and compare the results of each. For the first and simpler approach, we will use TF-IDF scoring and a curated stop-words list, measuring frequency and significance of words in order to extract and sort a list of keywords for given documents. We plan to reach a simple baseline of identifying the most frequent nouns in the document, which are likely to be relevant keywords. Step by step, this approach involves tokenizing and preprocessing of our input documents, the calculation of term frequency and inverse document frequency, and then using these to find the TF-IDF of each word in each document. This simpler approach can then be evaluated against the answer key, and we can compare the performance against our more complicated second approach.

Our second approach would be to employ graph-based algorithms to leverage its capabilities to capture the semantic relationship between words and phrases. With the documents represented as graphs, we will establish edge connections based on their co-occurrence and semantic similarities. Then, we would be going through graph-based algorithms such as TextRank, PageRank, TopicRank, SingleRank, PositionRank, and Word Attraction Rank and finetune each algorithm to optimize the performance of the individual algorithm. Extracted keywords will be compared against manually annotated keywords and be used to calculate rank-aware evaluation metrics such as Mean Reciprocal Rank (MRR) or Mean Average Precision (MAP) to assess the performance of our algorithm. After analyzing the strength and weakness of each algorithm we will come up with a final algorithm and fine tune its parameters based on evaluation results to enhance keyword extraction performance.

We will include error analysis in our approaches by taking a deep dive into our development set versus the answer key in order to reveal issues for improvement.

Collaboration Plan

In line with the project guidelines, we have assigned proposed roles of focus to each team member according to our strengths:

- Writing papers in English - Leeya Howley
- Programming / Software Development - Vishesh Goyal
- Theoretical Issues (computer science, math, linguistics) - Chan Hyun Yoo
- Evaluation (testing and measuring success) - Corina Luca

The components of our project will be broadly separated into strategic approaches. Corina and Leeya will focus on TF-IDF analysis while Vishesh and Chan Hyun will tackle graph based algorithms. The keyword extraction performance of each approach can be evaluated against each other and illuminate potential areas for improvement.

Bibliography

Anette Hulth. 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223.

Peggy M. Andersen, Philip J. Hayes, Steven P. Weinstein, Alison K. Huettner, Linda M. Schmandt, and Irene B. Nirenburg. 1992. Automatic Extraction of Facts from Press Releases to Generate News Stories. In *Third Conference on Applied Natural Language Processing*, pages 170–177, Trento, Italy. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273, Baltimore, Maryland. Association for Computational Linguistics.

Bennani-Smires, K., Musat, C. C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In *Proceedings of the Conference on Computational Natural Language Learning*.

Kim, Y., Kim, M., Cattle, A., Otmakhova, Y., Park, S., & Shin, H. (2013). Applying Graph-based Keyword Extraction to Document Retrieval. *International Joint Conference on Natural Language Processing.*