Experiments with the semantic similarity measure between sentences for the LexRank Text Summarization System

Chernenko, Tatjana

Institute for Computational Linguistics of University Heidelberg
Im Neuenheimer Feld 325, 69120 Heidelberg, Germany
http://www.cl.uni-heidelberg.de/

Automatic Textsummarization
Prof. Dr. Katja Markert
WS 2017/2018

Abstract. The project aims to re-implement LexRank-based text summarization system [1], which utilizes a stochastic graph-based method for computing relative importance of textual units for Natural Language Processing. In this model, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences. The objective is to explore the influence of replacing the cosine similarity measure with a combination of features from ECNU [3], a new system for semantic similarity between sentences. Both systems perform good, showing high ROUGE scores (61.7% ROUGE-su4 on the three multi-document clusters) for the short summaries of three sentences. The experiments with features used in ECNU system [3] allow us to achieve 62.87% average ROUGE-su4.

Keywords: Text Summarization · LexRank · ECNU · Semantic Similarity Measure.

1 Introduction

LexRank provides a stochastic graph-based method for computing relative importance of textual units for the problem of the extractive multi-document text summarization. It calculates the sentence importance utilizing the concept of eigenvector centrality in a graph representation of sentences. In this model, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences [1]. The project explores the possibilities of replacing the cosine similarity with the best practices of the SemEval-2017 Task 1 [2] for calculating the sentence semantic similarity. The best overall system is from ENCU [3]. It presents different feature engineering and deep learning models and shows the performance of their combinations. The ensemble of 3 machine learning algorithms and 4 deep learning models by averaging these 7 scores (EN-seven), achieves the best results. It promotes the

performance to 78.51 % on STS 2016 and to 85.18 % on STS 2017 English Datasets.
The proposed project implements Gradient Boosting Regression method with features based on n-gram overlap and kernel similarity of bags-of-words.

2 Implementation

The implementation of LexRank and modified LexRank [1] are described in the following Subsections.

2.1 LexRank Implementation

LexRank is based on the concept of ”prestige” (or ”centrality”) in social networks, where a cluster of documents can be viewed as a network of sentences. The sentences that are similar to many of the other sentences in a cluster are more central (or salient) to the topic [1]. To implement the similarity, we represent each sentence as an N-dimensional vector, where N is the number of all possible words in the language, values are the occurrences of the word times the idf of the word. The idf-modified cosine similarity measures the similarity between two vectors. Sentences are represented as vector $TF \cdot IDF$ metrics.

We construct a cosine similarity matrix which shows similarities between the sentence pairs. This matrix can be also represented as a graph where each edge shows the cosine similarity between a pair of sentences. Then we compute sentence centrality as described in Section 3.2 [1] and use a power method to compute the stationary distribution [1]. Class LexRankSummarizer of the program implements the above described operations. The project leaves the implementation of the reranker, that penalizes the sentences already included in the summary, for the future work.

2.2 Implementation of some ECNU features and learning algorithms

The overall architecture of ECNU system consists of three modules. The first module, Traditional NLP Module, extracts two kinds of NLP features (Sentence Pair Matching Features and Single Sentence Features). These NLP-based similarity scores act as features to build regressors to make prediction. The second module, Deep Learning Module, trains end-to-end neural networks to obtain similarity scores. The third Ensemble Module averages the above two modules to get a final score. Providing the basic architecture for the future implementation of all above mentioned modules, this project considers the first NLP Module of ECNU system. In this section, we give the description of the implemented approach.

[1] `https://gitlab.cl.uni-heidelberg.de/chernenko/automatic_textsummarization_ws_17_18/tree/master`

The feature engineering part provides two kinds of features. The sentence pair matching features directly calculate the similarity of two sentences. This type of features is represented in our project by the Weighted N-gram Overlap of lemmatized words. The single sentence features represent each sentence in the same vector space to calculate the sentence similarity. We design BOW feature: each sentence is viewed as a Bag-of-Words (BOW) and each word is weighted by its IDF value [3]. Such simple BOW Features with kernel functions are effective for sentence semantic similarity (Table 4 [3]). The learning part utilizes Gradient Boosting Regression algorithm implemented in scikit-learn toolkit to make prediction. The project also provides the architecture for the future construction of all ECNU modules, allowing to experiment with all the features and ensembles of algorithms.

2.3 Modified LexRank: LexRank with ECNU

The next step of the implementation replaces the idf-modified cosine similarity function of LexRank with described in Section 2.2 ECNU features and learning algorithm (Weighted N-gram Overlap of lemmatized words, BOW feature and GB Regression learning algorithm).

3 Evaluation Data

Test data for the experiments are taken from 2003 summarization evaluations of Document Understanding Conferences (DUC2003). Task 2 Dataset involves generic summarization of 30 news documents clusters and provides 10 different human judges. For the development and testing of LexRank we use all the documents from DUC2003 Task 2 dataset (30 multi-document clusters, each of them consists of 10 documents to one topic) and get 30 summaries. Modified LexRank works with three document clusters of DUC2003 Task 2 data (d30020t, d30034t, d31010t) and provides three text summarizations. The projects uses the 3-sentences limit of the length of the summaries. The above-mentioned documents clusters (d30020t, d30034t, d30034t) are used for the automatic evaluation of LexRank and modified LexRank.

For evaluation, we use the automatic summary evaluation metric ROUGE (using Pyrouge Toolkit), a recall-based metric for fixed-length summaries which is based on n-gram co-occurrence. The project utilizes human judges from DUC2003 Task 2 as reference summaries and provides ROUGE-1, ROUGE-2, ROUGE-3, RIUGE-4 and ROUGE-su4 scores for three document clusters.

4 Results

We monitor LexRank and modified LexRank performance on the three DUC2003 multi-document clusters, using four different human annotations for each cluster as reference (peer7.2 DUC2003 Task 2 Data). We show the results in Tables 1-4.

Table 1: Performance of LexRank and Modified LexRank on DUC2003 Task 2 multi-document clusters (d30020t document cluster).

d30020t document cluster	LexRank	Modified LexRank
ROUGE-1 Recall	0.59562	0.59562
ROUGE-1 Precision	0.98428	0.98428
ROUGE-1 F-score	0.74214	0.74214
ROUGE-2 Recall	0.43041	0.43041
ROUGE-2 Precision	0.71215	0.71215
ROUGE-2 F-score	0.53654	0.53654
ROUGE-3 Recall	0.21203	0.21203
ROUGE-3 Precision	0.35127	0.35127
ROUGE-3 F-score	0.26444	0.26444
ROUGE-4 Recall	0.1177	0.1177
ROUGE-4 Precision	0.19524	0.19524
ROUGE-4 F-score	0.14686	0.14686
ROUGE-su4 Recall	0.53642	0.53642
ROUGE-su4 Precision	0.8894	0.8894
ROUGE-su4 F-score	**0.66922**	**0.66922**

LexRank Summary for d30020t document cluster

BEIJING (AP) _ China hopes to win its first Asian Games gold medal in the martial art of taekwondo next month in Thailand, the official newspaper China Sports Daily said Tuesday. The team of six women and three men will be China's first to compete in taekwondo at the Asian Games, the report said. China only began developing taekwondo in 1995, a year after the last Asian Games, the newspaper said.

Modified LexRank Summary for d30020t document cluster

BEIJING (AP) _ China hopes to win its first Asian Games gold medal in the martial art of taekwondo next month in Thailand, the official newspaper China Sports Daily said Tuesday. The team of six women and three men will be China's first to compete in taekwondo at the Asian Games, the report said. China only began developing taekwondo in 1995, a year after the last Asian Games, the newspaper said.

Table 2: Performance of LexRank and Modified LexRank on DUC2003 Task 2 multi-document clusters (d30034t document cluster).

d30034t document cluster	LexRank	Modified LexRank
ROUGE-1 Recall	0.64639	0.52737
ROUGE-1 Precision	0.94898	0.97903
ROUGE-1 F-score	0.76899	0.68549
ROUGE-2 Recall	0.47737	0.40383
ROUGE-2 Precision	0.70141	0.75081
ROUGE-2 F-score	0.5681	0.52518
ROUGE-3 Recall	0.26853	0.22537
ROUGE-3 Precision	0.35127	0.35127
ROUGE-3 F-score	0.31967	0.29325
ROUGE-4 Recall	0.17511	0.14585
ROUGE-4 Precision	0.25771	0.27199
ROUGE-4 F-score	0.20853	0.18988
ROUGE-su4 Recall	0.56925	0.48574
ROUGE-su4 Precision	0.83754	0.90537
ROUGE-su4 F-score	**0.67781**	**0.63226**

LexRank Summary for d30034t document cluster
JAKARTA, Indonesia (AP) _ Assailants killed three soldiers and a civilian in the disputed territory of East Timor, Indonesia's official Antara news agency reported Sunday. Meanwhile, military officials said an Indonesian soldier was killed early Friday in Manufahi, about 40 kilometers (25 miles) southeast of Dili, East Timor capital. Indonesia annexed East Timor, a former Portuguese colony, after invading during a 1975 civil war that broke out when Portugal colonizers left.

Modified LexRank Summary for d30034t document cluster
Meanwhile, Aisyah Amini, chairperson of the Parliament's commission dealing with foreign affairs, also criticized the proposal. "The East Timor issue is not personal affair of Suharto, but the problem of the nation," she said. Portugal and East Timorese pro-independence groups rejected that proposal and insisted a referendum be held inside the territory to decide its future.

Table 3: Performance of LexRank and Modified LexRank on DUC2003 Task 2 multi-document clusters (d31010t document cluster).

d31010t document cluster	LexRank	Modified LexRank
ROUGE-1 Recall	0.42223	0.47454
ROUGE-1 Precision	0.95359	0.98833
ROUGE-1 F-score	0.5853	0.64121
ROUGE-2 Recall	0.27656	0.34722
ROUGE-2 Precision	0.62606	0.72461
ROUGE-2 F-score	0.38365	0.46948
ROUGE-3 Recall	0.11299	0.13549
ROUGE-3 Precision	0.25638	0.28333
ROUGE-3 F-score	0.15685	0.18332
ROUGE-4 Recall	0.17511	0.14585
ROUGE-4 Precision	0.06294	0.0465
ROUGE-4 F-score	0.08744	0.06296
ROUGE-su4 Recall	0.36285	0.43194
ROUGE-su4 Precision	0.82468	0.90449
ROUGE-su4 F-score	**0.50396**	**0.58467**

LexRank Summary for d31010t document cluster

The cold front arrived on Nov. 16 with temperatures dipping as low as minus 26 Celsius (minus 4 Fahrenheit). The cold front arrived on Nov. 16 with temperatures dipping as low as minus 26 Celsius (minus 4 Fahrenheit). On Friday, a man identified as Adam S., 47, was found frozen to death in a Warsaw park.

Modified LexRank Summary for d31010t document cluster

The relentless parade of Pacific storms will plague the Northwest as pockets of cold air aloft rotate into the region. Spates of gusty winds and heavy rain will accompany intense showers in coastal areas. It already has claimed 20 more lives than the number of cold-related deaths in all of last winter, police say.

Table 4: Average performance of LexRank and Modified LexRank on DUC2003 Task 2 multi-document clusters (three clusters) .

Average	LexRank	Modified LexRank
ROUGE-1 Recall	0.55475	0.53251
ROUGE-1 Precision	0.96228	0.98388
ROUGE-1 F-score	0.69881	0.68961
ROUGE-su4 Recall	0.48951	0.48470
ROUGE-su4 Precision	0.85054	0.89975
ROUGE-su4 F-score	**0.61700**	**0.62872**

Unigram-based ROUGE score (ROUGE-1) has been shown to agree with human judgments most [4]. As we see from the Tables 1-4, both LexRank and

Modified LexRank reach hight results in ROUGE-1 scores, showing in average 96-98% in ROUGE-1 Precision and 53-55% in ROUGE-1 Recall. In the first case (d30020t cluster), both systems produce the same summary text. In the second case (d30034t cluster), standard LexRank reaches 12% higher ROUGE-1 Recall and 4% smaller ROUGE-1 Precision. The manual evaluation of the produced summaries shows that systems focus on different aspects of the texts. While standard LexRank talks about killed people and explains the historical aspects of the accident, modified LexRank focuses on the reaction and proposals of the officials to the accident, missing the accident and proposals of the government itself, possibly because of the summary limit of three sentences. Both summaries provide important information. The evaluation in this case could possibly profit from the increased sentences limit. In the third case (d1010t document cluster) modified LexRank performs better in both Recall and Precision metrics. As mentioned in Section 2.1, the project implements LexRank without a reranker, that's for now why we can ignore the redundancy of the standard LexRank, producing two equal sentences. Modified LexRank also doesn't contain reranker, but avoids redundancy in this case. We can also notice that modified LexRank provides more general information for the whole Northwest area and the whole number of the killed people, while standard LexRank gives exact information about the temperature in an unknown specific region and focuses on the death of a specific person in a concrete place. Conceptually, both systems perform well both in terms of automatic and manual evaluation and could profit from the evaluation on more document clusters with an increased sentences limit and reranker to provide the fair comparison with 665-byte human judges (see Section 5 for possible improvements).

5 Discussion and possible improvements

As we see in Section 4, both systems perform well in terms of ROUGE-1, ROUGE-su4 scores and manual evaluations. Modified version of LexRank, utilizing two ECNU features and learning algorithm, reaches in average 1.1% improvement in ROUGE-su4 score and leaves place for the experiments with other ECNU modules. The evaluation could possibly profit from an increased sentences limit and implemented reranker module to avoid redundancy and achieve a better information coverage. The evaluation on the whole DUC2003 Task 2 Dataset would be also interesting. However, for the modified LexRank it is connected with the increased running time of the code. While standard LexRank needs a couple of minutes for the summarization of one cluster of ten documents, modified version of LexRank, calculating weighted N-gram overlap, BOW features and using Gradient Boosting Regression, takes more than 6-7 hours for one document cluster (depending on the system). Other possible improvement is implementing and testing LexRank for other languages.

6 Conclusion

The main contributions of the project are working implementations of LexRank, producing informative summaries with the given sentences limit, and it's modified version, utilizing novel practices of one of the best performing system of the SemEval-2017 shared task [2]. The project provides architecture for the future implementations and experiments and shows the ways for the possible improvements.

References

1. Erkan G., Radev D.R. (2017). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization.
2. Cer D. et al. (2017) SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation
3. Tian J.et al. (2017). ECNU at SemEval-2017 Task 1: Leverage Kernel-based Traditional NLP features and Neural Networks to Build a Universal Model for Multilingual and Cross-lingual Semantic Textual Similarity.
4. Lin C. and Hovy E. (2000). Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics.