[Data Science](#)[Natural Language Processing](#)[NLP Papers Summary](#)

Day 188: NLP Papers Summary – A Supervised Approach To Extractive Summarisation Of Scientific Papers

By Ryan 6th July 2020 No Comments

Objective and Contribution

Introduced a 10K CSPubSum summarisation dataset. We also developed multiple models for the dataset and observed that models which capture local and global context perform the best. We introduced a new summarisation feature, AbstractROUGE, which increases summarisation performance and HighlightROUGE, which can be used to extend our dataset.



CSPubSum and CSPubSumExt

We created a 10148 computer science extractive summarisation dataset. The publications are collected from ScienceDirect and are grouped into 27 domains. Each paper in the dataset has a title, abstract, author written highlight statements and author defined keywords. The highlight statements are the gold summary statements. See the figure below for an example.

Paper Title Statistical estimation of the names of HTTPS servers with domain name graphs
Highlights we present the domain name graph (DNG), which is a formal expression that can keep track of cname chains and characterize the dynamic and diverse nature of DNS mechanisms and deployments. We develop a framework called service-flow map (sfmap) that works on top of the DNG.sfmap estimates the hostname of an HTTPS server when given a pair of client and server IP addresses. It can statistically estimate the hostname even when associating DNS queries are unobserved due to caching mechanisms, etc through extensive analysis using real packet traces, we demonstrate that the sfmap framework establishes good estimation accuracies and can outperform the state-of-the art technique called dn-hunter. We also identify the optimized setting of the sfmap framework. The experiment results suggest that the success of the sfmap lies in the fact that it can complement incomplete DNS information by leveraging the graph structure. To cope with large-scale measurement data, we introduce techniques to make the sfmap framework scalable. We validate the effectiveness of the approach using large-scale traffic data collected at a gateway point of internet access links .
Summary Statements Highlighted in Context from Section of Main Text Contributions: in this work, we present a novel methodology that aims to infer the hostnames of HTTPS flows, given the three research challenges shown above. The key contributions of this work are summarized as follows. We present the domain name graph (DNG), which is a formal expression that can keep track of cname chains (challenge 1) and characterize the dynamic and diverse nature of DNS mechanisms and deployments (challenge 3). We develop a framework called service-flow map (sfmap) that works on top of the DNG. sfmap estimates the hostname of an https server when given a pair of client and server IP addresses. It can statistically estimate the hostname even when associating DNS queries are unobserved due to caching mechanisms, etc (challenge 2). Through extensive analysis using real packet traces , we demonstrate that the sfmap framework establishes good estimation accuracies and can outperform the state-of-the art technique called dn-hunter, [2]. We also identify the optimized setting of the sfmap framework. The experiment results suggest that the success of the sfmap lies in the fact that it can complement incomplete DNS information by leveraging the graph structure. To cope with large-scale measurement data, we introduce techniques to make the sfmap framework scalable. We validate the effectiveness of the approach using large-scale traffic data collected at a gateway point of internet access links. The remainder of this paper is organized as follows: section2 summarizes the related work. [...]

Table 1: An example of a document with summary statements highlighted in context.

We created two different version of the dataset: CSPubSum and CSPubSumExt. The summary statistics of the two datasets are shown in the figure below. The CSPubSum consists of positive and negative examples for each paper. The positive examples are highlight statements

whereas the negative examples are randomly sampled from the bottom 10% of sentences based on ROUGE-L. The test set consists of 150 full papers and it's the set we used to evaluate the quality of summaries. The CS PubSumExt is where we used HighlightROUGE to find sentences similar to the highlights from the full paper. This allows us to extend the dataset to 263K instances for training set and 131K instances for test set.

	#documents	#instances
CSPubSum Train	10148	85490
CSPubSumExt Train	10148	263440
CSPubSum Test	150	N/A
CSPubSumExt Test	10148	131720

Table 2: The CSPubSum and CSPubSumExt datasets as described in Section 2.2. Instances are items of training data.


HighlightROUGE and AbstractROUGE

The HighlightROUGE is used to generate more training data. It takes the gold summary and the text from the research papers and finds sentences that yield the best ROUGE-L score in relation to the highlights. We selected the top 20 sentences as positive instances and the bottom 20 sentences as negative instances. Note that we excluded extracting sentences from the abstracts as there are already a summary.

AbstractROUGE is a new summarisation feature that measures the ROUGE-L score between the sentence and the abstract. The idea is that sentences which are good summaries of the abstract are also likely to be good summaries of the highlights.

Methodology

We experimented with two different sentence encoding methods: average word embeddings and RNN encoding. We have also selected 8 different summariser features to help encode the local and global context of each sentence:

1. *AbstractROUGE*.
2. *Location*. We assign integer location based on 7 different sections of the paper  ight, Abstract, Introduction, Results / Discussion / Analysis, Method, Conclusion, all else

3. *Numeric Count*. Measure the number of numbers in a sentence. The idea is that sentences with heavy maths are unlikely to be good summaries
4. *Title Score*. Measure the overlap between the non-stopwords of each sentence and the title of the paper
5. *Keyphrase Score*. Measure how many author-defined keywords appear in the sentence. The idea is that important sentences will contain more keywords
6. *TFIDF*. TFIDF was calculated for each word and averaged over the sentence. We ignored stopwords
7. *Document TFIDF*. Same as TFIDF except the count of words in a sentence is the TF and the count of words in the rest of the paper is the background corpus, which allows us to measure how important a word is in a sentence relative to the rest of the document
8. *Sentence Length*. The idea is that short sentences are very unlikely to be good summaries

MODELS

Our models could take in any combination of the four possible inputs:

1. The sentence encoded with RNN (S)
2. Vector representation of the abstract (A)
3. The 8 features from previous section (F)
4. Average non-stopword word embeddings in the sentence (Word2Vec)

We experimented with 7 different models as listed below:

1. *Single Feature Models*. Model that only use one feature (we exclude sentence length, numeric count, and section)
2. *FNet*. A single layer NN that uses all 8 features to classify each sentence
3. *Word2Vec* and *Word2VecAF*. Both are single layer networks where Word2Vec takes in sentence average vector and Word2VecAF takes in the sentence average vector, abstract average vector, and handcrafted features
4. *SNet*. Feed the sentence vectors into bidirectional RNN with LSTM
5. *SFNet*. Processes the sentence with LSTM and passes the output to a fully connected layer with dropout. The handcrafted features are used as a separate inputs to a fully connected


layer. The outputs of the LSTM and features hidden layer are concatenated and output the binary prediction

6. *SAFNet*. Extend SFNet by encoding abstract too. This is shown in the figure below.

7. *SAF + F and S+F Ensemblers*. The ensemble methods use weighted average of the output of two different models. SAF + F is the ensemble of SAFNet and FNet and S+F is the ensemble of SNet and FNet

Results

MOST RELEVANT SECTIONS TO A SUMMARY

First of, we want to understand which sections contribute the most to our gold summary. To do this, we compute the ROUGE-L score of each sentence against the gold summary and average sentence-level ROUGE-L scores by section. In addition, there are also many occurrences where the authors directly copy the sentences from within the main text into the highlight statements. The ROUGE score and Copy/Paste score is captured in the figure below. The title has the highest ROUGE score which it's as expected. However, surprisingly, the introductory  the third-lowest ROUGE score, however, it does have the second highest Copy/Paste score. We

believe this contradictory results of the introduction section is due to the length of the section. The introduction section is long and this is bad for the ROUGE score as it contains more sentences that are not good for the summaries. However, more sentences in introduction section also means that there are more potential sentences to be use as highlights as demonstrated by the high Copy/Paste score.

MODEL PERFORMANCE AND ERROR ANALYSIS

The figure 3 below showcase the ROUGE-L score of the different models. Our ensemble models significantly outperformed the baseline models showcasing the effectiveness of our sentence encoding and features.





In figure 4 below, we compared the performance of all the models we developed in this paper. We found that architectures that uses sentence encoding and our handcrafted features performed the best by both ROUGE scores and test set accuracy. LSTM was able to outperformed the average word embeddings method which tells us that the ordering of words in a sentence is important. Another observation is that the highest accuracy result does not translate to the highest ROUGE score although they are strongly correlated. SAFNet achieved the highest accuracy on CSPubSumExt but underperformed AbstractROUGE summariser on CSPubSum. We manually examined 100 sentences from CSPubSumExt that were misclassified by SAFNet. We found that the primary reasons of false positives was lack of context and long range dependency. Other reasons for false positives include mislabelled and including maths heavy sentences.

The primary reason for false negatives are mislabelled data and failure to recognise entailment, observation or conclusion. Overall, a high accuracy does not equates to high ROUGE scores and this is most likely due to overfitting to the training data that has mislabelled examples.





EFFECT OF USING ROUGE-L TO GENERATE MORE DATA

The figure 5 below showcase the performance difference between three selected models on CSPubSumExt (full data) and CSPubSum (low data). Across all three models, we see a consistent improvement when trained on full data suggesting that increasing training data using ROUGE-L does improves the summarisation performance.





EFFECT OF ABSTRACTROUGE METRIC ON SUMMARISER PERFORMANCE

Figure 6 below showcase the performance of 4 models trained with and without AbstractROUGE. We observed that AbstractROUGE does improve the performance of summarisation techniques and that sentence encoding and features engineering lead to a more stable model.





Conclusion and Future Work

Potential future work involves developing model to better capture the global context and the cross-sentence dependencies.

Source: <https://www.aclweb.org/anthology/K17-1021.pdf>

Ryan

Data Scientist

Previous Post



Next Post

< Day 187: Learn
NLP With Me –
Embeddings of
Language,
Knowledge
Representation,
and Reasoning

Day 189: Learning
PyTorch - PyTorch
Lightning
Introduction

[Data Science](#) [Natural Language Processing](#) [NLP Papers Summary](#)

Day 365: NLP Papers Summary – A Survey on Knowledge Graph Embedding



Ryan

20th December 2020



[Data Science](#) [Implementation](#) [Natural Language Processing](#)

Day 364: Ryan's PhD Journey – OpenKE-PyTorch Library Analysis + code snippets for 11 KE models

Ryan

29th December 2020





[Data Science](#) [Ryan's PhD Journey](#)

Day 363: Ryan's PhD Journey – Literature Review – Knowledge Acquisition – 1st Passes XIV

Ryan

28th December 2020

