Data Science Natural Language Processing NLP Papers Summary

Day 114: NLP Papers Summary – A Summarization System For Scientific Documents

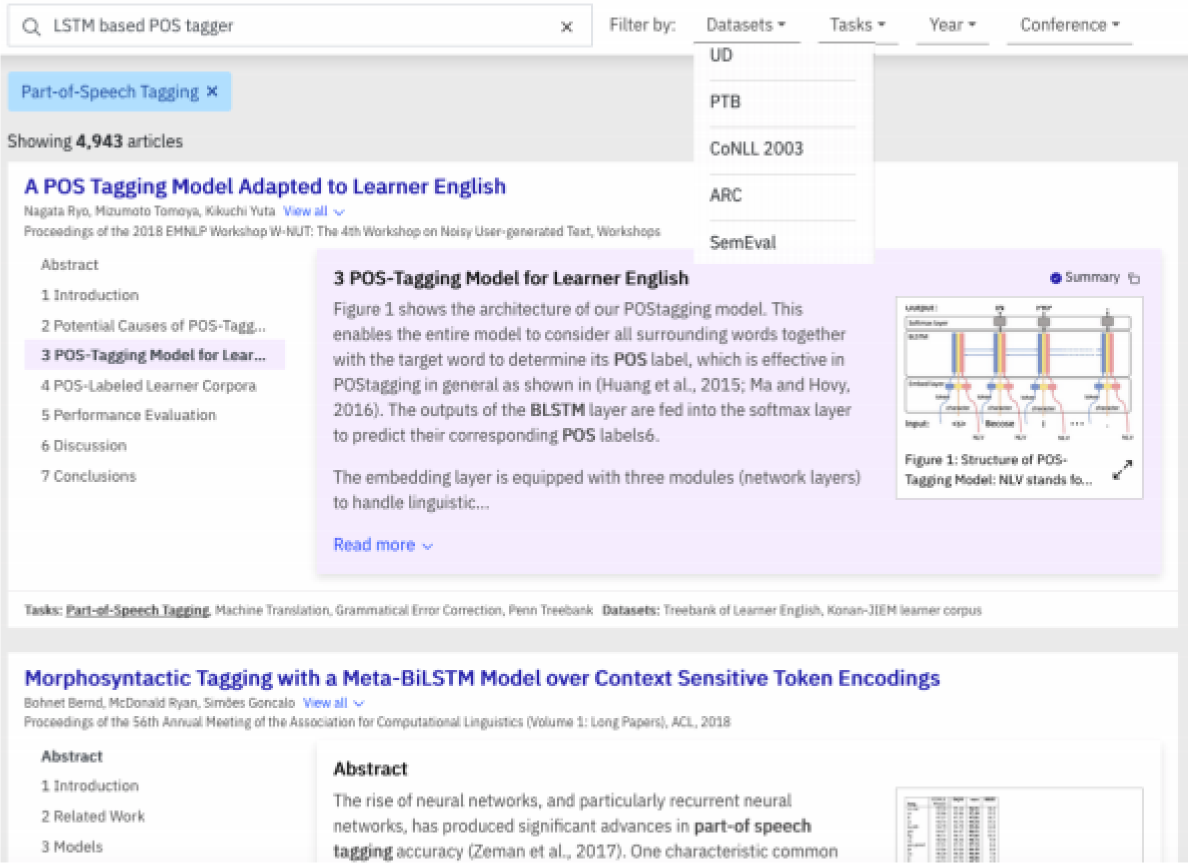
By Ryan 23rd April 2020 No Comments

Objective and Contribution

Proposed IBM Science Summariser for summarising computer science research papers. The system can identify different scenarios such as discovery, exploration, and understanding of scientific documents. The proposed system summarises research papers in two ways: either in free-text query or by choosing categorised values such as scientific tasks, datasets, and more. The proposed system ingested 270,000 papers.

The IBM Science Summariser produces summaries that focuses on user's queries (query-focused summarisation). It summarises various sections of the paper independently, allowing users to focus on relevant sections only. This allows for the interaction between user's queries and various entities in the paper.

Figure below showcase the user-interface of the IBM Science Summariser. Users pose their queries (or use the filters on metadata fields). Relevant papers are then returned with summarisation results. Each section is clearly shown with entities accurately highlighted.



LSTM based POS tagger

Showing **4,943** articles

A POS Tagging Model Adapted to Learner English

Nagata Ryo, Mizumoto Tomoya, Kikuchi Yuta

Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, Workshops

Abstract
1 Introduction
2 Potential Causes of POS-Tagg...
3 POS-Tagging Model for Lear...
4 POS-Labeled Learner Corpora
5 Performance Evaluation
6 Discussion
7 Conclusions

3 POS-Tagging Model for Learner English

Figure 1 shows the architecture of our POStagging model. This enables the entire model to consider all surrounding words together with the target word to determine its **POS** label, which is effective in POStagging in general as shown in (Huang et al., 2015; Ma and Hovy, 2016). The outputs of the **BLSTM** layer are fed into the softmax layer to predict their corresponding **POS** labels6.

The embedding layer is equipped with three modules (network layers) to handle linguistic...

Figure 1: Structure of POS-Tagging Model: NLV stands fo...

Tasks: Part-of-Speech Tagging, Machine Translation, Grammatical Error Correction, Penn Treebank **Datasets:** Treebank of Learner English, Konan-JIEM learner corpus

Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings

Bohnet Bernd, McDonald Ryan, Simões Goncalo

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL, 2018

Abstract
1 Introduction
2 Related Work
3 Models

Abstract

The rise of neural networks, and particularly recurrent neural networks, has produced significant advances in **part-of speech tagging** accuracy (Zeman et al., 2017). One characteristic common

Figure 2: IBM Science Summarizer UI.

Summarisation of Scientific Articles – What, Why, How?

WHAT DOES A SUMMARISATION SYSTEM FOR SCIENTIFIC PAPERS CONSISTS OF?

1. Extracting the structure
2. Extracting tables and figures from PDF
3. Identify important entities
4. Generating useful summary

WHY IS THIS NEEDED?

Below are the pain-points of academic researchers:

1. Keeping up-to-date with current work
2. Preparing research project / grant request
3. Preparing related works when writing a paper
4. Checking novelty of an idea

The first pain point tend to happen daily / weekly, with information overload and lots of time spent reading papers. Pain points 2 – 4 are important but less frequent.

HOW DO RESEARCHERS SEARCH AND READ RESEARCH PAPERS?

1. Researchers search by either keywords, entities (such as task name, dataset name, or models etc) or citation. For example, "state of the art results for SQUAD"
2. Read the title –> abstract. However, researchers mentioned that abstract is not informative enough to determine relevancy

System Overview

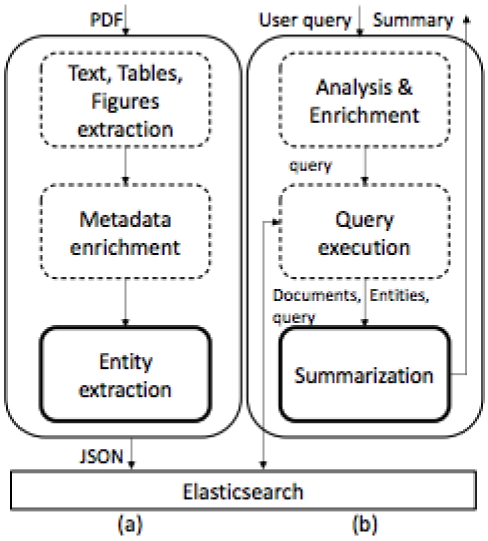


Figure 1: IBM Science Summarizer Framework.

The system (figure above) has two components:

1. Ingestion pipeline and search engine (Elasticsearch)

2. Summarisation

INGESTION PIPELINE

The system contains 270,000 papers from arXiv and ACL. The pipeline consists of 3 main steps:

1. Extracting paper's text, tables and figures
2. Metadata enrichment with annotations and entities
3. Entity Extraction

The system uses Science-Parse to extract PDF text, tables and figures. Science-Parse supports figures and table extraction into an image file (and its caption text). Figure and table references in text paragraphs are detected. We also extracted tasks, datasets and metrics. The output is return in JSON format. Elasticsearch is used to index the papers where we index its title, abstract text, sections text and some metadata.

The system has three types of entities: task, dataset, and metric. Both dictionary-based and learning-based approach are implemented. The dictionary-based are manually created using paperswithcode website. To cover all evolving topics, the learning-based approach is taken where we analyse the entire paper to extract the three types of entities. This was treated as a text entailment task whereby the paper contents is the text and the targeting Task-Dataset-Metric (TDM) triples as hypothesis. This approach forces the model to learn the similarity patterns between the text and the triples. Overall, the system has indexed 872 tasks, 345 datasets, and 62 metrics from the entire corpus.

SUMMARISATION

The summary can be generic or query-focused. The language can be quite different between sections and so sections are summarise independently and these section-based summaries are then composed together into one summary. The inputs of the summarisation are query (optional), entities, and the relevant papers returned by search engine. The summarisation is broken down into multiple steps:

1. Query Handling
2. Pre-processing

3. Summarisation

If query Q is given, it can either be very precise or verbose. If it's short and precise, we would expand it using query expansion, which transforms Q into 100 unigram terms (obtained from analysing top papers that are returned from the Q). If Q is verbose, a fixed-point weighting schema is used to rank the query terms. If no Q, keyphrases of the paper are used as proxy for the query.

In terms of pre-processing, we perform sentence tokenisation, word tokenisation, lowercased and removal of stop words. Each sentence is then transform into uni-grams and bi-grams BoW representations.

In terms of summarisation, we used a SOTA unsupervised, extractive, query focused summarisation algorithm. The algorithm takes in the paper section, query Q, desired summary length (10 sentences), and a set of entities that are linked to the query. The generated summary is a subset of sentences from the paper section selected through an unsupervised optimisation scheme. This sentence selection is posed as a multi-criteria optimisation problem, where several summary quality objectives are considered. These summary qualities are:

1. *Query saliency*. Does the summary contains many query related terms (cosine similarity)?
2. *Entities coverage*. Does the entities covered in the summary match with our set of entities?
3. *Diversity*.
4. *Text coverage*. How much does the summary covers the paper section?
5. *Sentence length*. We want the summaries to bias towards longer sentences, which are assumed to be more informative.

Human Evaluation

EVALUATION SETUP

We approached 12 authors and asked them to evaluate summaries of two papers they have co-authored. This gives us a total of 24 papers. For each paper, we produced two types of summaries: section-based summary and section-agnostic summary (treating the paper content as flat text). This is for us to assess the benefit of section-bases summarisation. This us a total of 48 summaries to be evaluated.

The authors are required to perform 3 tasks per summary:

1. For each sentence, determine whether the sentence should be included as part of the summary (binary measure of the precision)
2. How well each sections of the paper is covered in the summary (measure of recall, 1 – 5 scale, 3 being good)
3. Evaluate the overall quality of the summary (1 – 5 scale, 3 being good)

RESULTS

The results are shown in the figure below. For task 2, section-based summary scored higher in 68% of the papers. The average score for section-based summaries is 3.32 which highlights the quality of section-based summaries.

Conclusion and Future Work

As future work, the IBM Science Summariser plan on adding support for more entities and ingest more papers. More qualitative study is being conducted to assess its usage and quality of summaries, including automatic evaluation of summaries.

Source: https://www.aclweb.org/anthology/D19-3036.pdf

Ryan

Data Scientist

Previous Post

Day 113: NLP Papers Summary - On Extractive and Abstractive Neural Document Summarization with Transformer Language Models

Next Post

Day 115: NLP Papers Summary - SCIBERT: A Pretrained Language Model for Scientific Text

Data Science Natural Language Processing NLP Papers Summary

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

Ryan
30th 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