Information Retrieval (CS60092)

Project Proposals

Ashutosh Kumar Singh (19CS30008), Ashwamegh Rathore (19CS30009), Nakul Aggarwal (19CS10044), Suhas Jain (19CS30048)

Preference 1:
Query-by-Example for Scientific Article Retrieval (Project 9)

Motivation and Relevance

We realized over a few group sessions that in research, it is naturally difficult to express information requirements as simple keyword queries because they can have a very broad meaning and might not be able to retrieve a specific section from the whole bundle of documents. A far more effective way to understand the needs of the user is by asking the user to give a research paper deemed as relevant to his aspirations. Currently, researchers primarily rely on keyword search of online indices such as *Google Scholar* and *PubMed* to help them combat the overload of information. We therefore think that an IR model that asks the user to specify his query as a *paper* (along with an additional set of facets or aspects to focus on) rather than as a set of words can achieve the same intention with greater effectiveness. We want to give any researcher the freedom to ask – *"I came across the paper XYZ during my research and am extremely inclined towards the results and mathematical background given by the author. Can you please get me some more papers that I might find relevant to my research?"*

Problem Statement

Envision and develop a model that is able to retrieve scientific papers analogous to a *query* scientific paper, along specifically chosen rhetorical structure elements (facets/aspects), like results, keywords, relevant works etc. So we frame our task as one of retrieving scientific papers given a ***query paper*** and additional information indicating the ***query facet(s)***.

Literature Review

We reviewed many papers on QBE for Scientific Article Retrieval (and other related topics) to discover various baselines to solve the above problem, out of which we found the objective of the following two very closely related to the underlying gist of the given project statement.

Arini et al. [EG11] Given a small set of papers Q that they refer to as the query set, Arini et. al. seek to return a set A of additional papers that are related to the concept defined by the query. Intuitively, a paper that cites all of the articles in Q is likely to represent related research. Likewise, a paper that is cited by every article in Q might contain relevant background information. Since, it is restrictive to require the papers in A to have a direct citation to or from every article in the query set, *they select a set A that maximizes a more general notion of* ***influence*** *to and from the papers in Q.* They have quantified the inference relation among the papers in the form of a weighted directed acyclic graph. They have proposed various statistical techniques to sample, sort and diversify the retrieved set of relevant documents.

Upadhyay et al. [Upa+20] They define and solve a novel academic search task, called aspect-based retrieval, which allows the user to specify the aspect along with the query to retrieve a ranked list of relevant documents. Their primary idea is to estimate a language model for the aspect as well as the query using a **domain-specific knowledge base** and use a mixture of the two to determine the relevance of the article. Inspired from this work, we are thinking of modeling the relevance of a document as the probability of generating the query from the document. *Given a query Q (generally as a research paper of interest) and an aspect A, a document D is deemed as relevant on the basis of both Q and A.* The authors have further described a probabilistic solution to this problem with various baselines.

Plan

– We first intend to shortlist the best performing models among the reviewed papers so that we can possibly build an ensemble information retrieval engine that is able to combine the results produced by all the high-performing technologies.
– We will study the dataset provided to us more deeply. It is possible that for varying methodologies the requisite pre-processing tasks are also different. In that case, we will have to tweak the dataset accordingly.
– Next and the most crucial part is the implementation. We would first prefer recognizing the independent components in the entire project so that each group member can work in parallel, to increase the throughput. The knowledge of the tools necessary for translating the theoretical postulations into code can be gathered concurrently while implementing.
– Finally we will train the model on the entire dataset and test it. The results will be documented and compared with those obtained by the authors.
– The exercise will be wrapped up by documenting the motivation, theoretical formulation, results and future works in a brief project report and also a presentation (as required) to explain the gist, results and possible extensions in the project.

Preference 2:
Evidence Retrieval for Fact Verification (Project 5)

Motivation and Relevance

The constantly growing online textual information and the rise in the popularity of social media have been accompanied by the spread of fake news and false claims. However, manual fact-checking is time consuming and intractable on a large scale. The ability to automatically perform fact-checking is critical to minimize negative social impact. Wikipedia is a unique knowledge source and we are interested in using it as a constantly evolving source of detailed information that could facilitate intelligent machines - if we (and the machines) are able to leverage its powers.
However, the selection of relevant evidence sentences for accurate fact-checking and explainability remains a challenge. For the retrieval of relevant evidence from a corpus of documents, existing systems typically utilize traditional sparse retrieval which may have poor recall. With the newer dense retrieval models have proven effective in question answering as these models can better capture the latent semantic content of text.

Problem Statement

Fact verification is a two-step process. First, we retrieve supporting or refuting evidence related to a claim. Then based on the set of evidence snippets, the task is to determine whether the claim is true or false. In this project, we are interested in the first step, i.e., evidence retrieval. We are interested in developing an IR system which when provided a claim, retrieves a set of documents and then relevant sentences from those documents which support or refute the claim.

Literature Review

Thorne et al. [Tho+18a] introduce a publicly available dataset for verification against textual sources called FEVER. It consists of 185,445 claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. They also provide the baseline and oracle supported scores for this task.
Zhou et al. [Zho+19] focus more on the claim verification part and introduce novelty by using a fully-connected evidence graph based approach that utilizes different aggregators to collect multievidence information.
Thorne et al. [Tho+18b] provide an overview of the FEVER shared task, a competition which challenged participants to classify whether human-written factoid claims could be supported or refuted using evidence retrieved from Wikipedia. It sheds light on the retrieval systems and verification techniques developed by various participants.
Nie et al. [NCB18] realise that because of huge number of documents, conducting semantic matching between all the documents from the collection with the claim is computationally intractable. Hence, we should start the retrieval by applying a keyword matching step to narrow down the search space and then move towards semantic matching.
Chen et al. [Che+17] suggest an approach that combines a search component based on bigram hashing and TF-IDF

matching with a multi-layer recurrent neural network model trained to detect answers in Wikipedia paragraphs. Chakrabarty et al. [CAM18] use multiple methods to retrieve documents from the web and took a union of all the documents to increase the coverage to 94.4% from the baseline of 55.3%.

Plan

The pipeline of the process will consist of two key steps: document retrieval and sentence extraction. To increase the coverage of our document retrieval step we use different tools and APIs to retrieve Wikipedia pages. Some of these tools and APIs are: Google Custom Search API, Wikipedia Python API, MediaWiki API, Tantivy, Bing Search API etc. We will try to device parsing techniques and entity recognition algorithms to identify key phrases and tokens from the claim query before passing them to the APIs.
In the document retrieval phase we mostly rely on word-matching however in sentence extraction, now that our search space is narrowed, we can try out an amalgamation of semantic-matching and word-matching. To do the semantic matching we can try out different metrics like TF-IDF vector similarity or some natural language inference techniques.
After this we will extract the top k sentences that support or refute the claim and we will track precision@k and recall@k. Also note that recall is the most important factor in this step because the FEVER score counts a prediction as true if a complete set of evidence is retrieved. We evaluate the correctness of the evidence retrieved by computing the F1-score of all the predicted sentences in comparison to the human annotated sentences for those claims requiring evidence on our complete pipeline system. We shall also use the OFEVER metric described in the FEVER dataset paper.

Preference 3:
Efficient and Fast Image Retrieval (Project 7)

Motivation and Relevance

Recent visual recognition methods typically train multiclass classifiers using image datasets labeled with a predefined set of discrete classes. However, such classifiers are not capable of capturing semantic relationships among visual categories since they are trained in the discrete label space. For example, discrete classifiers treat the three classes cat, dog and bicycle as unrelated and distinct categories. As a result, they cannot encode the fact that the two classes cat and dog are semantically more similar than that between cat and bicycle. Furthermore, to recognize a new category, the discrete classifiers need to be retrained on a sufficient amount of training examples of the new class. The lack of semantic information transfer substantially limits the visual recognition methods to scale up to large numbers of classes. To address these issues, visual-semantic embedding have been proposed to leverage the semantic knowledge from text data.
This also helps in IR related search tasks where the user might not know the exact label of the entity it wants to search for but can still get the desired result from searching for labels of semantically similar entities.

Problem Statement

In content-based image retrieval, it is often claimed that visually similar images are clustered in this feature space. However, there are two major problems with this approach: 1) Visual similarity does not always correspond to semantic similarity. 2) The classification objective does not enforce a high distance between different classes, so that the nearest neighbors of some images may belong to completely different classes. Hierarchy-based semantic embeddings overcome these issues. The task is to learn semantic embeddings (more details are provided in paper) using the CIFAR-100 dataset.

Literature Review

Li et al. [Li+17] propose the structured discriminative and difference constraints to learn visual-semantic embeddings. They exploit the discriminative constraints to capture the intra and inter-class relationships of image embeddings. They also align the difference vector between a pair of image embeddings with that of the corresponding word embeddings. The difference constraints help regularize image embeddings to preserve the semantic relationships among word embeddings.
Frome et al. [Fro+13] proposed an approach to leverage semantic knowledge learned in the text domain, and

transfer it to a model trained for visual object recognition. They present a new deep visual-semantic embedding model trained to identify visual objects using both labeled image data as well as semantic information gleaned from unannotated text.

Plan

- We first want to do a more detailed literature review of the papers which have a focus on semantic embedding and do a comparative study of which models are worth looking into and might have useful input on how we can structure our model.
- Next we will perform an iterative approach of trying different models and methods and then make observations as to what gives good result for the particular use case we are dealing with.
- After finalising the approach we are going to use we will train and test it extensively and make sure there are no gaps remaining and if there is scope for any more optimisation.
- After that of the time permits we will move on to making an indexing strategy which can help us implement a fast and semantically sound query processing.

References

[EG11] Khalid El-Arini and Carlos Guestrin. "Beyond keyword search: discovering relevant scientific literature". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 439–447.

[Fro+13] Andrea Frome et al. "DeViSE: A Deep Visual-Semantic Embedding Model". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

[Che+17] Danqi Chen et al. *Reading Wikipedia to Answer Open-Domain Questions*. 2017. arXiv: 1704.00051 [cs.CL].

[Li+17] Dong Li et al. *Learning Structured Semantic Embeddings for Visual Recognition*. 2017. arXiv: 1706.01237 [cs.CV].

[CAM18] Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan. "Robust Document Retrieval and Individual Evidence Modeling for Fact Extraction and Verification." In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 127–131. DOI: 10.18653/v1/W18-5521. URL: https://aclanthology.org/W18-5521.

[NCB18] Yixin Nie, Haonan Chen, and Mohit Bansal. *Combining Fact Extraction and Verification with Neural Semantic Matching Networks*. 2018. arXiv: 1811.07039 [cs.CL].

[Tho+18a] James Thorne et al. *FEVER: a large-scale dataset for Fact Extraction and VERification*. 2018. arXiv: 1803.05355 [cs.CL].

[Tho+18b] James Thorne et al. "The Fact Extraction and VERification (FEVER) Shared Task". In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 1–9. DOI: 10.18653/v1/W18-5501. URL: https://aclanthology.org/W18-5501.

[Zho+19] Jie Zhou et al. "GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 892–901. DOI: 10.18653/v1/P19-1085. URL: https://aclanthology.org/P19-1085.

[Upa+20] Prajna Upadhyay et al. "Aspect-Based Academic Search Using Domain-Specific KB". In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose et al. Cham: Springer International Publishing, 2020, pp. 418–424. ISBN: 978-3-030-45442-5.