GPT – 4

Title: Advancements in Efficiency and Robustness of Neural Ranking Models

Advancements in neural ranking models have led to significant improvements in efficiency and robustness. One such advancement is the use of greedy synonym replacement strategies, which have been shown to enhance rank promotion in models like PRADA [REF0]. Another development is the application of graph neural networks (GNNs) in self-supervised settings, which can be used to construct content-based retrieval systems [REF1].

Incorporating pseudo-relevance feedback in multiple representation dense retrieval settings has also been explored. ColBERT-PRF, for example, applies clustering to embeddings in the pseudo-relevant set and identifies the most discriminative embeddings among cluster centroids [REF2]. Additionally, noise data augmentation strategies have been employed to improve the robustness of ranking models, ensuring that the noise of perturbed documents has minimal impact on the ranking [REF3].

MetaAdaptRank, a method that synthesizes ranking-aware weak supervision, has been developed to train robust neural information retrieval models. This approach involves training an NLG model to capture the difference between contrastive document pairs and generate discriminative queries [REF4]. Another advancement is the use of query salience conditioned functions, which have been shown to outperform fixed log functions used in previous kernel-pooling approaches [REF5].

Graph neural network architectures have also been improved by incorporating common building blocks such as neighborhood aggregation, local pooling, and global pooling [REF6]. MetaAdaptRank further addresses the few-shot challenge in neural information retrieval by transferring source-domain supervision signals to target-domain weak supervision signals and meta-learning to reweight the synthetic weak supervision [REF7].

Preprocessing techniques, such as cleaning text and representing documents in vector form, have been essential in improving the efficiency of neural ranking models. Methods like Latent Semantic Analysis, Bag-of-Words, and Word2Vec have been employed to extract document features [REF8]. These advancements in neural ranking models have contributed to the development of more efficient and robust systems, paving the way for future research and improvements in the field [REF9].

References given to GPT:

[REF0] - paperID: 21ee4b66ce53de6b7b23c23cae0885bf5c96ad78

Title: PRADA: Practical Black-Box Adversarial Attacks against Neural Ranking Models

Chunk of text: Based on Table 5, we observe that: (1) By removing important word replacement, the performance of PRADA−IW R in terms of SR has a significant drop as compared with PRADA. The results indicate that the greedy synonym replacement strategy does help the rank promotion. PRADA−ESP has a similar performance with PARDA, which again demonstrates the effectiveness of the word replacement with synonyms. (2) PRADA−ESP−IW R performs much worse than the PRADA−IW R. Without the limitation given by the word replacement, the embedding space perturbation has an obvious influence on the results. (3) By including all the components, PRADA achieves the best performance among the variants in terms of all evaluation metrics. 6.4 Analysis between Black-box Setting vs. White-box Setting As mentioned in Section 3.2, there are different adversarial settings for the WSRA task in terms of the information that attackers rely on. In this work, we focus on the decision-based black-box attack setting because it is close to real-world search engine scenario.

[REF1] - paperID: 73f266f2f495f86331609d9d3e24cbc8a762510c

Title: Self-Supervised Learning with Graph Neural Networks for Region of Interest Retrieval in Histopathology

Chunk of text: We propose learning a graph neural network (GNN) that encodes ROI graphs into representations using a contrastive loss function in a self-supervised setting. Then, a content-based retrieval system is constructed using the trained GNN to extract representations from ROIs and Euclidean distance between the extracted representations to measure the similarity of ROIs. 6330 Authorized licensed use limited to: ULAKBIM UASL - Bilkent University. Downloaded on February 08,2022 at 10:10:24 UTC from IEEE Xplore. Restrictions apply. TABLE II CLASS DISTRIBUTION OF SLIDES AND ROIS IN TRAINING, VALIDATION, AND TEST SETS.

[REF2] - paperID: 44772b24ae2f68b77476c814b0607370f7195ddb

Title: Pseudo-Relevance Feedback for Multiple Representation Dense Retrieval

Chunk of text: In this work, we are concerned with applying pseudo-relevance feedback in a multiple representation dense retrieval setting. Indeed, as retrieval uses multiple representations, this allows additional useful embeddings to be appended to the query representation. Furthermore, the exact scoring stage provides the document embeddings in response to the original query, which can be used as pseudo-relevance information. Thus, in this work, we propose a pseudo-relevance feedback mechanism called ColBERT-PRF for dense retrieval. In particular, as embeddings cannot be counted, ColBERT-PRF applies clustering to the embeddings occurring inICTIR ’21, July 11, 2021, Virtual Event, Canada X. Wang et al. the pseudo-relevant set, and then identifies the most discriminative embeddings among the cluster centroids. These centroids are then appended to the embeddings of the original query.

[REF3] - paperID: 7177d99f5a873ba8ad2772edbb02f85fcd281566

Title: Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models

Chunk of text: Similar to the process of obtaining 𝑇𝑤 from 𝑆𝑤, we achieve the collection perturbation dictionary 𝑇𝐶 by keeping the top 𝐽 nearest neighbors in 𝑆𝐶 for each word. The overall architecture is shown in Figure 2. The certified defense algorithm contains two key steps, i.e., noised data augmentation and Top-𝐾 Robustness Certification. We describe the two steps in the following. Noise Data Augmentation Strategy. The robustness certification holds regardless of how the original ranker 𝑓 is trained. However, to rank the document 𝑑 with respect to the 𝑞 correctly and robustly by ¯𝑓 , it is expected that 𝑓 properly ranks the perturbed document 𝑅 (recall that 𝑅 ∼ Π𝑑 ) such that it is close to the rank position of the original document 𝑑. That is, the noise of 𝑅 should have little effect on the ranking, making the ensembled ranking score ¯𝑓 (𝑞, 𝑑) close to the original ranking score 𝑓 (𝑞, 𝑑).

[REF4] - paperID: 65c2d2ffe45569101860a7defc7cccbd36b3602a

Title: Few-Shot Text Ranking with Meta Adapted Synthetic Weak Supervision

Chunk of text: and {sep} are special tokens. In inference, the trained query generator is directly used to generate new queries q ∗ for target-domain documents d ∗ , where d ∗ is regarded as the related (positive) document of q ∗ , while the unrelated (negative) document can be sampled from the target corpus. Despite some promising results, the vanilla training strategy may cause the NLG model to prefer to generate broad and general queries that are likely related to a crowd of documents in the target corpus. As a consequence, the synthetic relevance supervision does not have enough ranking awareness to train robust Neu-IR models. Source-domain NLG Training. To synthesize ranking-aware weak supervision, MetaAdaptRank trains the NLG model to capture the difference between the contrastive document pair (d +, d −) and generate a discriminative query q: q = T5-NLG([POS] ◦ d

[REF5] - paperID: de028eebe67b2bc74c471c9429914242fd5ed346

Title: Local Self-Attention over Long Text for Efficient Document Retrieval

Chunk of text: Furthermore, our novel query salience conditioned function outperforms the fixed log function used in previous kernel-pooling approaches, except for MAP on the TREC DL 2019 dataset. R5. How often does TKL attend to different parts of the document? We show the distribution of the top-3 relevant regions in Figure 3. While we can clearly see a focus on the beginning of the document, we observe a sizeable amount of relevant regions after 500 words. The most relevant region occur 24.5 %, the second 41.5 %, and the third 46.4 % of the time after the 500th word. Even though our baselines achieve acceptable results by only looking at the start of a document, this result in Figure 3 shows that TKL learns to detect relevant regions in every part of a document.

[REF6] - paperID: 73f266f2f495f86331609d9d3e24cbc8a762510c

Title: Self-Supervised Learning with Graph Neural Networks for Region of Interest Retrieval in Histopathology

Chunk of text: In our application, the data augmentation module removes a random subset of vertices from the ROI graph. The resulting two views represent the same ROI with different graph structures and vertex features. A GNN is employed as the encoder. As the projection head, we use a 2-layer multi-layer perceptron (MLP) with ReLU nonlinearity. This way, the GNN encoder is forced to learn high-level contextual features to maximize the agreement between the two views. The process is illustrated in Figure 1. C. Graph Neural Network Architectures Three common building blocks of GNN models are neighborhood aggregation, local pooling, and global pooling .

[REF7] - paperID: 65c2d2ffe45569101860a7defc7cccbd36b3602a

Title: Few-Shot Text Ranking with Meta Adapted Synthetic Weak Supervision

Chunk of text: Average Loss Optimal Weights Meta-forward Update Meta-backward Update Training ’ #’ $’ Source Domain Relevance Signals T5-NLG Inference Sampling Target Domain Documents … Contrast Doc Pair Meta-Reweighted Synthetic Signals ∗ # " ’ #’ Figure 1: The illustration of MetaAdaptRank, which first synthesizes massive weak supervision signals for target domains, and then meta-learns to reweight these synthetic data based on small target-domain relevance labels. where d + i and d − i denote the relevant (positive) and irrelevant (negative) documents of the query qi . In few-shot ranking scenarios, the number of relevance supervision signals (M) is limited, making it difficult to train an accurate Neu-IR model. To mitigate the few-shot challenge in Neu-IR, MetaAdaptRank first transfers source-domain supervision signals to target-domain weak supervision signals (Sec 3.2); then meta-learns to reweight the synthetic weak supervision (Sec 3.3) for selectively training Neu-IR models (Sec 3.4). 3.2 Contrastive Synthetic Supervision MetaAdaptRank transfers the relevance supervision signals from source domains to few-shot target domains in a zero-shot way.

[REF8] - paperID: cd52d4251de98217f32c3e556ea738ae97fc308d

Title: Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization

Chunk of text: First, the preprocessing step. This step usually requires cleaning the text (e.g., removing punctuation mark, stop word, numerals). Stemming is applied at this step. The second step involves representing the document in vector form to extract its features. Different techniques have been proposed, such as Latent Semantic Analysis , Bag-of-Words , and Word2Vec

[REF9] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

Title: UOBIT @ TAG-it: Exploring a Multi-faceted Representation for Profiling Age, Topic and Gender in Italian Texts

Chunk of text: Final Workshop (EVALITA 2020), Online. CEUR.org. Pedro Javier Ortiz Suarez, Laurent Romary, and Beno ´ ˆıt Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1703–1714, Online, July. Association for Computational Linguistics. Nils Reimers and Iryna Gurevych. 2019.

........................................................................................................................................................................................................

Title: Advancements in Neural Ranking Models and Attention Mechanisms for Information Retrieval

Advancements in neural ranking models and attention mechanisms have been driven by the need to improve information retrieval systems, particularly in the legal domain, where traditional natural language understanding systems struggle to transfer their success from non-legal domains [REF0]. One approach to address this issue is to focus on modeling semantic similarity at the paragraph level rather than whole case documents, which allows for a more limited number of important legal concepts to be considered [REF0].

Graph-based approaches have been employed to capture context and improve retrieval models. For example, the graph-of-word method uses an unweighted directed graph with nodes representing terms and edges linking terms within a sliding window of size N [REF1]. This approach, combined with term weighting and document length normalization, has shown promising results in information retrieval tasks [REF1]. Another graph-based method, the hypergraph-of-entity, has been used to explore different node and hyperedge types, leading to various extended versions of the base model [REF4].

Efficiency improvements have also been proposed, such as the HubRank method, which computes the personalization vector for a set of hub nodes selected based on query logs [REF3]. This approach uses a TypedWordGraph with word-to-entity relations, enabling mixed word and entity queries, and approximates vectors using precomputed fingerprints [REF3].

Recent advancements in neural ranking models, such as ColBERT-PRF, have shown promising results in document ranking tasks on datasets like MSMARCO and Robust04 [REF5]. Living Labs, a shared experimental environment for IR benchmarking, has also been introduced to facilitate the evaluation of retrieval models in real-world search engines, providing a unique degree of feedback that is often only available to commercial or institutional search engines [REF6].

Graph-based document retrieval has been explored by converting documents and queries into graphs consisting of words and POS tags as nodes and syntactic dependencies as edges [REF8]. Node weights are computed from TF-IDF, and this approach has contributed to the application of graphs as a tool for natural language understanding and question answering [REF8].

In summary, advancements in neural ranking models and attention mechanisms for information retrieval have been driven by the need to improve performance in challenging domains, such as the legal field. Graph-based approaches, efficiency improvements, and shared experimental environments have all contributed to the development of increasingly effective retrieval models.

References given to GPT:

[REF0] - paperID: 05cd0c3998c6ca59d46669880f07b9134b08e8d9

Title: Searching for Legal Documents at Paragraph Level: Automating Label Generation and Use of an Extended Attention Mask for Boosting Neural Models of Semantic Similarity

Chunk of text: Such guesswork about the terms likely to be used in relevant documents, through the resulting errors, can prevent the discovery of the most relevant documents in the database, for example if those documents used slightly different terms that those that were selected by the user. As efforts to remedy such issues using linguistics and / or ontology-based approaches that explicitly model domain knowledge and language are rather expensive and difficult to keep up-to-date (Silveira et al. 2004), the research community is actively searching for new approaches to those problems, in particular the semantic similarity aspect of document relevance, as it enables improved document ranking. Recently, data-driven neural approaches have emerged as promising components of a new generation of increasingly automatic systems that may require less human interference, curation or updating effort (Tran et al. 2020, Zhong et al. 2020). While this trend is in its early stages, its maturation could help to deal with some of the above-mentioned limitations. However, it was observed by many authors (Alberts et al. 2020, Bhattacharya et al. 2020, Chalkidis et al., 2020, Draijer 2019, Raghav et al. 2016, Shao et al. 2020, Van Opijnen & Santos 2017, Xiao et al. 2019, Wang et al. 2019, Zhong et al. 2020) that current neural systems for natural language understanding that perform very well in non-legal domains do not transfer easily to tasks in the legal domain, for a variety of reasons that make this domain especially challenging (see Table 1). Such research efforts can then be further complicated by the need to adjust such systems to legal documents written in non-English languages, as much of the research in the public domain has been performed on English language texts. By performing the legal Information Retrieval task at the level of paragraphs rather than whole case documents, the modeling of semantic similarity is then focused on a more limited number of important legal concepts contained in such a paragraph, or query.

[REF1] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: They defined an unweighted directed graph (the graph-of-word), where nodes represented terms, and edges linked each term to its following terms within a sliding window of size N, in order to capture context. Based on information retrieval heuristics [117, 118] and the graph-based term weighting approach by Blanco and Lioma , they also defined a retrieval model over the graph-of-word, based on the indegree of the nodes (TW-IDF). The goal of the weighting model was to measure the number of contexts a given term appeared in. They also introduced a pivoted document length normalization component, tunable with parameter b (analogous to BM25’s b). The graph-of-word was generated per document, computing the TW metric and storing it within the inverted index, to be used as a replacement for TF. This meant that the document graphs could then be discarded without requiring persistence. They evaluated the TW-IDF ranking function with and without regularization over document length, as well as with and without parameter tuning for the pivoted document length normalization b parameter.

[REF2] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: Index Task Ranking Avg./query MAP GMAP P@10 NDCG@10 Lucene Doc. Index Ad hoc document retrieval TF-IDF 460ms 0.0228 0.0000 0.0692 0.0778 BM25 370ms 0.0324 0.0000 0.1173 0.1274 Ent. Index Ad hoc entity retrieval TF-IDF 1s 370ms 0.0373 0.0000 0.0636 0.0670 BM25 798ms 0.0668 0.0000 0.1182 0.1165 Entity list completion TF-IDF 1s 230ms 0.0558 0.0044 0.1000 0.1014 BM25 1s 221ms 0.0666 0.0067 0.1250 0.1212 Hypergraph-of-Entity Base Model Ad hoc document retrieval RWS 23s 405ms 0.0863 0.0278 0.2462 0.2662 Ad hoc entity retrieval 26s 330ms 0.1390 0.0002 0.2455 0.2425 Entity list completion 19s 162ms 0.0879 0.0376 0.0769 0.0594 Syns Ad hoc document retrieval RWS 55s 555ms 0.0937 0.0303 0.2615 0.2812 Ad hoc entity retrieval 30s 232ms 0.1337 0.0004 0.2473 0.2445 Entity list completion 19s 875ms 0.0857 0.0368 0.0635 0.0474 Context Ad hoc document retrieval RWS 24s 348ms 0.0869 0.0245 0.2654 0.2784 Ad hoc entity retrieval 27s 620ms 0.1304 0.0002 0.2364 0.2298 Entity list completion 19s 422ms 0.0875 0.0373 0.0692 0.0520 TF-Bins2 Ad hoc document retrieval RWS 2m 58s 0.0172 0.0033 0.0500 0.0508 Ad hoc entity retrieval 4m 41s 0.0300 0.0000 0.1145 0.1307 Entity list completion 1m 08s 0.0006 0.0000 0.0058 0.0053 Syns+Cont. Ad hoc document retrieval RWS 23s 265ms 0.0882 0.0246 0.2692 0.28830 Ad hoc entity retrieval 26s 877ms 0.1313 0.0002 0.2509 0.2422 Entity list completion 19s 824ms 0.0884 0.0369 0.0788 0.0594 terms according to word2vec embeddings; the TF-bins2 model, which added tf\_bin hyperedges for the most and least frequent terms per document; and the Syns+Cont. model, which included the hyperedges from the Syns and Context models. Indexing took between 33h05m (Syns) and 37h16m (Syns+Cont.)

[REF3] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: hubrank Chakrabarti proposed an efficiency improvement over ObjectRank, where the personalization vector was only computed for a set of hub nodes selected based on query logs. They proposed a TypedWordGraph, where they introduced word-to-entity relations, thus enabling mixed word and entity queries. Each vector was approximated using precomputed fingerprints — i.e., the end nodes from random walks of various lengths, as sampled from a geometric distribution, and initiated from each node — as described by Fogaras et al. . In order to compute HubRank, a subgraph limited by boundary nodes was first prepared. The boundary was established by a subset of hub nodes called blockers, and by loser nodes that were too far to significantly influence the personalized PageRank of the word nodes. Personalized PageRank was then estimated for the remaining active nodes and iteratively computed using dynamic programming, while fixing the value of boundary nodes.

[REF4] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: ⊇ Vt} terms in the same TF range Eb = {v : v ∈ Vt ∧ v ∈ ed ∧ ed ∈ Ed} where V 0 t and V 00 t are extensions of Vt (i.e., supersets), which may contain additional terms from external synonyms or contextually similar neighbors, respectively. ED = Ec, where Ec is the only subset of directed hyperedges, such that Ec = {(t, h) : t ⊆ Vt ∧ h ⊂ Ve ∧ |h| = 1}. In the experiments we carry throughout this thesis, we control the structure of the index by enabling or disabling each of the described node and hyperedge types. For example, in Section 9.1, we explore the base model Hbm, where Es = ∅, Ex = ∅, and Eb = ∅. In the same section, we also explore several extended versions of the base model, namely by adding synonyms, which results in |Vt(Hsyns)| > |Vt(Hbm)| and |Es| > 0. Another example can be found in Section 9.2, where we explore a text-only version of the model, such that Vt 6= ∅, and Ve = ∅. 7.4.1.2 Base model The hypergraph-of-entity is, in many ways, a simplification of the graph-of-entity (Chapter 6).

[REF5] - paperID: 5537feedc97256e81c6f1af66664dbcd19621d11

Title: ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval

Chunk of text: 0.6158𝑔 0.2592𝑎𝑔ℎ𝑗 0.4624𝑎ℎ 0.6681 0.6289𝑔ℎ𝑗 In this section, we further investigate the effectiveness of our proposed ColBERT-PRF for document ranking task. Table 5 and Table 6 present the performance of ColBERT-PRF models as well as the baselines on the MSMARCO document dataset and the Robust04 dataset, respectively. 6.3.1 Results for RQ5.

[REF6] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: Living Labs Balog et al. presented the first practical methodology and implementation of the Living Labs2 for IR benchmarking, focusing on two use cases: local domain search on the website of the University of Amsterdam, and product search in the webshop of a toy retailer operating in Hungary. Living labs represents a central and shared experimental environment that replaces individually setup evaluation infrastructure. This infrastructure can be used by different research groups, avoiding the hassle of preparing it themselves, and enabling them to compare evaluation metrics within a similar platform. This is different from the classical approach of evaluating retrieval models through test collections and instead takes advantage of interleaving in real-world search engines. Therefore, it confers the approach a unique degree of feedback, which is frequently only available to commercial or institutional search engines, due to the challenges of finding a representative set of test subjects in a lab environment. Table 2.4 provides a summary of the most relevant TREC tracks, their datasets and mission, in the context of entity-oriented search.

[REF7] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: Track Datasets Mission Ad Hoc track 2006 English Wikipedia (XML), INEX 2009 Wikipedia collection (XML, annotated with YAGO classes) Explore the internal structure of documents to retrieve relevant information. Entity Ranking track 2006 English Wikipedia (XML), INEX 2009 Wikipedia collection (XML, annotated with YAGO classes) Explore the direct retrieval of entities as a way to solve information needs, either based on a keyword query or a multiple entity query. Linked Data track Wikipedia-LOD v1.1, Wikipedia-LOD v2.0 Explore retrieval techniques over a combination of text and linked data. with document-level relevance [248, Table 13], we found values of MAP ranging from 0.3177 to 0.4294.

[REF8] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: This abstract graph query could then be instantiated into a graph query language available for the graph database, where ?author is a node of type #author. Despite the identified domain-dependent limitations of the model, this contributed to the application of graphs as a tool for natural language understanding and question answering. Zhang et al. explored graph-based document retrieval, by converting both documents and queries to graphs consisting of words and POS tags as nodes, and syntactic dependencies as edges. They segmented the documents into document semantic units (DSU), representing the atomic unit of parsing (e.g., a sentence, or a phrase within a sentence). They extracted graphs from each DSU, for each document. Node weights were computed from TF-IDF.

[REF9] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: “Markov logic networks”. In: Machine Learning 62.1-2 (2006), pp. 107–136. doi: 10.1007/s10994-006-5833-1 (cit. on p. 14). R. Biagioni, P. Vandenbussche, and V. Novácek. “Finding Explanations of Entity Relatedness in Graphs: A Survey”. In: CoRR abs/1809.07685 (2018).

........................................................................................................................................................................................................

Title: Advancements in Neural Models for Text Representation and Ranking

Advancements in neural models have significantly improved text representation and ranking in various applications. One such advancement is the use of BERT-based systems for multi-task learning, which has shown promising results in terms of f1-scores on validation sets [REF0]. Another development is the use of fastText models, which can generate embeddings for rare words, misspelled words, and words not found in the training corpora [REF1]. This capability addresses the limitations of word2vec and GloVe models, which cannot expand terms not found in their vocabularies.

In measuring semantic change degrees, distance-based methods such as Average Geometric Distance (AGD) have been employed [REF3]. These methods can be applied to the last-layer output of BERT for word representation. Furthermore, multilingual pretrained transformer language models like XLM-R have been adapted to initialize dense retrieval models for cross-language document ranking tasks [REF5]. This approach enables encoders to process multiple languages and identify appropriate resources for training the model.

In terms of ranking performance, methods like Joint Passage-Document (JPD) ranking have been shown to outperform init-LTR baselines [REF6]. JPDs, which use the features of the document's passage most highly ranked, have been found to be the most effective approach among the proposed methods. Another notable method is JPDm-max, which uses the maximum value of each passage-based feature over the document's passages [REF6].

Syntax-based neural network models have also been used to synthesize SemQL queries, which can then be transformed into SQL queries for information retrieval [REF8]. This approach allows for the extraction of table data from web pages and the organization of data in a formalized way using databases.

Overall, these advancements in neural models for text representation and ranking have led to improved performance in various applications, such as content moderation, evaluation of user-generated content [REF2], and cross-language document ranking [REF5]. As research in this area continues to progress, we can expect further improvements and innovations in the field.

References given to GPT:

[REF0] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper)

Chunk of text: On top of the model we added three ReLU classifiers, we applied the dropout method and we used the sum of the Cross-Entropy loss functions of the three classifiers as loss function. In Table 4 we report the system’s performances in terms of f1-score obtained on the validation set. f1-score Gender 0.86 Age 0.39 Topic 0.64 Table 4: Multi-Task Learning BERT-based system f1-scores on validation set 3 Results and Evaluation We run all our three systems on the test sets provided by the task organisers. The performances of our systems are reported in Table 5. For the Task 1 scoring, TAG-it considers two different rankings. The first ranking is obtained using a partial scoring scheme, giving 0 points if no correct predictions are provided for the three dimensions of the dataset, 1/3 points if one out of three correct answers is given, 2/3 points if two out of three correct answers are given and 1 point if all the answers given by the system are correct.

[REF1] - paperID: 7b8fe8c28a371120b4479540b2c8a0f7c5af25bf

Title: Learned Text Representation for Amharic Information Retrieval and Natural Language Processing

Chunk of text: Some query terms are not found in vocabulary/dictionary of the word2vec and GloVe models. As a result, such words are not expanded in the case of the word2vec and GloVe models. However, the fastText model returns expanded terms even though such words do not exist in its vocabulary. If a word is unseen during training, fastText segments a word into n-grams and generates its embedding. As a result, it helps to embed rare words, misspelled words, and words that do not exist in corpora but are found in the topic set. For example, the query term አገልግሎት /ʔəgəlɨgɨlotɨ ‘services’/ is not found in the corpora, and thus the word2vec and GloVe models do not return any expanded terms.

[REF2] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper)

Chunk of text: Figure 5: Subgroup AUC 7 Conclusions and Future Work Both of these challenges dealt with issues related to content moderation and evaluation of usergenerated content. While early research raised fears of censorship, the ongoing challenges platforms face have made it necessary to consider the potential of machine learning.

[REF3] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper)

Chunk of text: Only last-layer output of BERT is used as word representation. 3.2 Measuring Semantic Change Degree 3.2.1 Distance-based methods In this section, we introduce various methods to calculate the semantic change degree. Average Geometric Distance. Average Geometric Distance (AGD) (also can be seen in (Kutuzov and Giulianelli, 2020; Giulianelli, 2019)) is defined as below: AGD(ΦC1 i , Φ C2 i ) = 1 mn ­ x∈Φ C1 i ,y∈Φ C2

[REF4] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper)

Chunk of text: Google-Books-ID: KGIbfiiP1i4C. Andreas Blank. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. Historical semantics and cognition, 61. Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018.

[REF5] - paperID: d1ccffb8eb1b7a99cd586723074b82fa5399bdd2

Title: Transfer Learning Approaches for Building Cross-Language Dense Retrieval Models

Chunk of text: This generalization poses two challenges: enabling the encoders to process multiple languages, and identifying appropriate resources with which to train the model. To address the former, we adapt XLM-R , a multilingual pretrained transformer language model, to initialize the dense retrieval model. For the latter challenge, we use translations of MS MARCO , a widely-used passage ranking collection for training monolingual neural retrieval models. We evaluate ColBERT-X on ad hoc document ranking tasks using English queries to retrieve documents in other languages, exploring two ways to cross the language barrier. In the zero-shot setting, where we lack cross-language training resources, we train the model only on English MS MARCO.

[REF6] - paperID: f6d69afebcebcbd3e511faf19375f71dd679cdcb

Title: A passage-based approach to learning to rank documents

Chunk of text: We see in Table 3 that all the proposed methods outperform the init-LTR baselines — often statistically significantly — in the vast majority of relevant comparisons and are never outperformed in a statistically significant manner by a baseline. JPDs is the most effective approach among those we proposed: its block in the table has the highest number of boldfaced numbers, it outperforms any other approach in most relevance comparisons, and it is never statistically significantly outperformed by other approaches while the reverse often holds. These findings attest to the merits of using the passage-query features of the document’s passage most highly ranked together with the document-query features to learn a document ranker. The JPDm-max approach is the second-best performing. This finding is not entirely surprising: JPDs, which is our best performing method, uses the features of the document’s passage most highly ranked while JPDm-max uses per each passage-based feature the maximum value over the document’s passages. As could be expected, both JPDm-max and JPDm-avg outperform JPDm-min. That is, using the average or the maximum of a feature value across the document’s passages yields better performance than using the minimal value. Table 3 also shows that RRF outperforms SMPD in most relevant comparisons when using SVM and the reverse holds when using LMart.

[REF7] - paperID: 01b6bf20e38818df0b1c9f5a55a5f013aadcef09

Title: Continual learning in cross-modal retrieval

Chunk of text: 3 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 6 James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka GrabskaBarwinska, et al. Overcoming catastrophic forgetting in neural networks.

[REF8] - paperID: 3355935d5e2d088f96effaa50f0f37fdfcea86c8

Title: Enhanced Natural Language Interface for Web-Based Information Retrieval

Chunk of text: A syntax-based neural network model is used to synthesize SemQL query. Finally, A SQL query is generated based on SemQL and domain knowledge . Transforming SemQL to SQL is done by traversing the SemQL tree from its root to leaf nodes. B. EXTRACTING TABLE DATA FROM WEB PAGES To obtain data on a web page and organize it in a formalized way using database, we analyze the tags corresponding to the 4234 VOLUME 9, 2021T. Bai et al.: Enhanced Natural Language Interface for Web-Based Information Retrieval FIGURE 1. Web of science’s query page, when using advanced search, users are required to create queries according to its grammar. FIGURE 2.

[REF9] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper)

Chunk of text: Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sergio Nunes, et al. 2019. A hierarchically-labeled ´ portuguese hate speech dataset. In Proceedings of the Third Workshop on Abusive Language Online, pages 94–104. Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Arantza Casillas, Arantza D´ıaz de Ilarraza, Nerea Ezeiza, Maite Oronoz, Alicia Perez, and Olatz ´ Perez-de Vinaspre. 2018. Automatic misogyny i- ˜ dentification using neural networks.

........................................................................................................................................................................................................