GPT – 3

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

Amharic is a complex language with unique characteristics such as morphology, syntax, and semantics. Therefore, it is essential to investigate learned text representation techniques that can handle these specificities. In recent studies, fine-tuned BERT models have shown promising results for Amharic text classification and information retrieval (IR) tasks [REF0]. BERT models represent each query and document word considering contextual information of a word in a sentence or a phrase, which makes them better at understanding and handling the specificity of Amharic. Additionally, fastText models have been shown to be effective in embedding rare words, misspelled words, and words that do not exist in corpora but are found in the topic set [REF1].

Word-based BERT models have been found to outperform both stem-based and root-based approaches in Amharic IR tasks [REF0]. The WordPiece tokenizer used in BERT splits a stem-based and root-based token into sub-units if it does not recognize a token, which degrades the performance of stem-based and root-based approaches compared to that of the word-based approach.

In the EVALITA 2020 shared task PRELEARN, methodologies based on gated recurrent units and classical word embeddings with ensemble methods were proposed to classify pairs of concepts according to whether one is a prerequisite for the other or not [REF6][REF7]. The second approach, which uses information from the rest of Wikipedia, was particularly useful in terms of extension to new entities close to the one in the corpus provided for the task but not actually present in it.

In conclusion, learned text representation techniques such as fine-tuned BERT models and fastText models have shown promising results for Amharic information retrieval and natural language processing tasks. Additionally, using information from external sources such as Wikipedia can improve the performance of classification tasks.

[REF0] - paperID: 7b8fe8c28a371120b4479540b2c8a0f7c5af25bf

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

Chunk of text: As depicted in Table 6, most fine-tuned BERT models show promising results for Amharic text classification and IR tasks. On Amharic IR, BERT significantly outperformed all embedding algorithms. For example, the F1-score and accuracy values of fine-tuned BERT model to classify documents as relevant and non-relevant for the IR system were higher than word embedding models, as the BERT model represents each query and document word considering contextual information of a word in a sentence or a phrase. From learned text representation techniques, fastText and BERT models are better at understanding and handling the specificity of Amharic, such as morphology, syntax, and semantics. As shown in Tables 4 and 5, word-based BERT models outperformed both stembased and root-based approaches. Even though stems and roots are the fundamental units for many text processing tasks, the WordPiece tokenizer splits a stem-based and rootbased token into sub-units if it does not recognize a token. As a result, the performances of stem-based and root-based approaches are degraded compared to that of the wordbased approach.

[REF1] - paperID: 7b8fe8c28a371120b4479540b2c8a0f7c5af25bf

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

Chunk of text: 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. However, fastText returns the words አገልግሎ /ʔəgəlɨgɨlo/, አገልግሎትም /ʔəgəlɨgɨlotɨmɨ/, አገልግሎትን /ʔəgəlɨgɨlotɨnɨ/, አገልግሎትና /ʔəgəlɨgɨlotɨna/, አገልገሎት /ʔəgəlɨgəlotɨ/, አገልግሎቱንና /ʔəgəlɨgɨlotunɨna/, አገልገሎቱን /ʔəgəlɨgəlotunɨ/, አገልግሎቱን /ʔəgəlɨgɨlotunɨ/, and አገልግሎቱም /ʔəgəlɨgɨlotumɨ/, which are variants referring to the concept “serve”. Furthermore, fastText based on skip-gram outperforms the baseline retrieval performance reported in . The effectiveness of the Amharic retrieval system without and with query expansion using fastText is presented in Table 12.

[REF2] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: UmBERTo has a high model capability with 125M trainable parameters and was trained on online crawled data, making it suitable for processing meme language. SentenceTransformers We use the SentenceTransformers framework (Reimers and Gurevych, 2019) to produce sentence embeddings by averaging all word embeddings produced by the original UmBERTo model since Miaschi and Dell’Orletta (2020) showed that those are usually much more informative than the default [CLS] sentence embedding. We fine-tune representations over the available meme textual data and use them as components of our end-to-end system. 1umberto-commoncrawl-cased-v1 in the HuggingFace’s model hub (Wolf et al., 2019)297 2.3 Visual input While we have so far discussed only using metadata to predict our results, it is essential to address the core of a meme: the image itself. We can internally distinguish a meme from a standard image through the aforementioned broken sentence structure, meme templates, and quick and messy edits, among other aspects. As previously mentioned, memes can be very difficult to individuate when they look like standard images but gain meme status through real-world knowledge grounding. Due to the inherently large variance in meme images’ styles and contents, it is impractical to expect a single framework to effectively describe each distinguishable feature and utilize it to classify an entry.

[REF3] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: Evaluating n-gram candidates Once the methods have been selected and candidate n-grams are extracted from the model according to choice constraints, the outcome may be just one candidate and the evaluation stops or more than one candidate which is the rule. Now we have a list of candidate n-grams with the best ones at the top. The list may be created in a number of different manners. It has the KLD index inherited from the tweet and three other indices: one is the ratio of intersection words/lemmata, the higher this ratio the more relevant is the n-gram. Another index is the sum of the KLD indices associated to each of its word/lemma, the lower this sum the more relevant is the ngram (rare content words have a lower KLD index). Finally the third index is the one associated to the tweet in which the n-grams are contained. Choosing the best candidate in fact usually means selecting the best candidates from the list, because it almost never happens that there is only one candidate at the top with the best ratio or best index.

[REF4] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: Our approach is based on a convolutional neural network that exploits pretrained word embeddings. We also experimented a comparison among different architectures to understand the effectiveness of our method. The paper also described our submissions to both subtasks A and B to Automatic Misogyny Identification competition at Evalita 2020. 1 Introduction The paper describes our submission to the Automatic Misogyny Identification task at Evalita 2020 (Fersini et al., 2020; Basile et al., 2020). This competition is divided into two subtasks: • Subtask A Misogyny and Aggressive Behaviour Identification: identify if a text is misogynous or not, and, in case of misogyny, if it expresses an aggressive attitude. • Subtask B Unbiased Misogyny Identification: discriminate misogynistic contents from the non-misogynistic ones, while guaranteeing the fairness of the model (in terms of unintended bias) on a synthetic dataset (Nozza et al., 2019). We proposed a convolutional based approach to recognize misogynistic sentences.

[REF5] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: The task was organized in 4 subtasks: i) two of them concerned with the type of information that can be exploited by the submitted models, either solely textual or including metadata, e.g. Wikipedia hyperlinks; ii) the other two based on different classification scenarios, training and testing could happen either on the same domain or 1Copyright ­c 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). three domain could be used as training set and the fourth as testing. A more extensive description of the task together with all the results and more information is found in the report (Alzetta et al., 2020) which is part of the EVALITA 2020 (Basile et al., 2020). The concept of being a prerequisite is highly complex and can be misunderstood from humans as well. Indeed, this relation can be subtle and depending on the domain it may take a deep level of expertise to recognize. One of the reasons this challenge is very interesting, is the fact that several application can arise from this same setting.

[REF6] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: Abstract English. In this paper we describe the methodologies we proposed to tackle the EVALITA 2020 shared task PRELEARN. We propose both a methodology based on gated recurrent units as well as one using more classical word embeddings together with ensemble methods. Our goal in choosing these approaches, is twofold, on one side we wish to see how much of the prerequisite information is present within the pages themselves. On the other we would like to compare how much using the information from the rest of Wikipedia can help in identifying this type of relation. This second approach is particularly useful in terms of extension to new entities close to the one in the corpus provided for the task but not actually present in it. With this methodologies we reached second position in the challenge1 .

[REF7] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: This second approach is particularly useful in terms of extension to new entities close to the one in the corpus provided for the task but not actually present in it. With this methodologies we reached second position in the challenge1 . 1 Introduction The PRELEARN task consists in classifying pairs of concepts according to whether one is a prerequisite for the other or not. The concepts are presented as Wikipedia pages and they are divided into four different domains, physics, precalculus, data mining and geometry. The task was organized in 4 subtasks: i) two of them concerned with the type of information that can be exploited by the submitted models, either solely textual or including metadata, e.g. Wikipedia hyperlinks; ii) the other two based on different classification scenarios, training and testing could happen either on the same domain or 1Copyright ­c 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[REF8] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: Overall, the difference in the results between the first and the second model is considerable; therefore, we will show in the following table the F1 values obtained with the best run, for tasks A and B, respectively. TaskA Tweets TS News TS F-score NoHS HS 0.750 0.735 NoHS HS 0.835 0.615 M-F1 0.7432 0.7256 Table 3: Task A - Results for the Logistic Regression with Tfidf TaskB Tweets TS News TS F-score NoST ST 0.724 0.690 NoST ST 0.824 0.608 M-F1 0.7076 0.7166 Table 4: Task A - Results for the Logistic Regression with Tfidf Beyond the macro-F1 values obtained, it is interesting to note the behavior of the model with regard to the out-domain Test Set in both tasks. In particular, the F-scores show worse values in the classification of sentences that actually contain hate speech or stereotyping. This is actually due to low Recall values (about 0.51 for both tasks) which is probably due to the fact that the model is trained on a different type of data. 6 Discussion Observing the results on the in-domain Test Set, our best models obtained a ranking of 15/27 and 6/12 respectively for tasks A and B. Regarding the out-domain Test Set, they obtained the third-best score in both tasks. The result obtained with the first Test Set confirms that the proposed approach turned out to be too simplistic.

[REF9] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: Novel wordsense identification. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1624– 1635, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics. [Eger and Mehler2016] Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models.

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

Title: Efficient and Effective Text Representation for Information Retrieval

Efficient and Effective Text Representation for Information Retrieval is a crucial aspect of natural language processing. In recent years, deep learning models have shown significant improvements in various NLP tasks, including text classification, semantic similarity, and machine translation. Among the investigated architectures, ResNET and EfficientNet have shown impressive results in image processing tasks (REF0). However, for text representation, the use of noun phrases and function words has shown promising results in alleviating noise and achieving high accuracy in classification tasks (REF1).

Moreover, the proposed method of semantic sentence-image matching from a shared cross-modal space has shown effectiveness in natural language inference, semantic similarity, text classification, and machine translation tasks (REF2). This approach is universal and does not require manually annotated multimodal parallel corpora, making it cost-effective and efficient.

In identifying misogynistic and aggressive texts in Italian social media, simple classifiers with little feature engineering have shown a strong tendency to overfit and yield a strong bias on the test set (REF3).

In multilingual text representation, context-dependent models such as BERT have shown to be effective predictors, but relatively good results can also be obtained by using context-independent models and behavioral norms (REF5). However, for Italian, the range of available predictors is limited, and one effective solution is to translate the stimuli into English and use existing predictors for English.

In information retrieval tasks, the use of a target ranking model to re-rank the initial ranked list and the token importance ranking process has shown effectiveness in achieving high accuracy (REF6).

Finally, the incorporation of visual awareness into sentence modeling by retrieving a group of images for a given sentence has shown effectiveness in cross-modal retrieval tasks (REF7).

In summary, various approaches have been proposed for efficient and effective text representation for information retrieval, including the use of deep learning models, noun phrases, function words, semantic sentence-image matching, and visual awareness. These approaches have shown promising results in various NLP tasks, making them crucial for natural language processing.

[REF0] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: Among the investigated architectures, we first considered ResNET (He et al., 2016): this network is the first introducing Residual Learning to define very deep and effective CNNs. Several ResNET architectures are defined by stacking 50, 101, 152 up to 1001 layers of convolu-302 tion layers and skip connectors: as a result, deeper networks achieved significant improvements of previous state-of-art in a wide plethora of image processing tasks. Moreover, we investigated the recently proposed EfficientNet (Tan and Le, 2019): unlike ResNET, this is not a real architecture, but it provides an automatic methodology to improve the performance of an existing CNN (such as ResNET) by tuning its depth, width and resolution dimensions. The adoption of this methodology led to the definition of 8 CNNs (namely EfficientNET-B0, EfficientNET-B1 up to EfficientNET-B7), each characterized by an increasing depth and width. They achieve impressive results by efficiently balancing the number of the parameters of the network. The tuning process of (Tan and Le, 2019) demonstrated that a network such as EfficientNet-B3 achieves higher accuracy than ResNeXt101 (Xie et al., 2016) in using 18x fewer neural operations. Regardless of the adopted networks, these are already trained in a classification task involving the recognition of thousands of object types in several millions of images, i.e. in the ImageNet dataset (Deng et al., 2009).

[REF1] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: Using these features yields comparable scores to the best of our methods, surpassing the classification attempts on uncleaned texts. This indicates that noun phrases alleviate the noise extracted by the tf-idf vectorizer. The model was less prone to overfitting and, therefore, more able to adapt to the unseen data. Function words are features with grammatical roles, consisting of conjunctions, prepositions, articles, etc. encompassing stylistic aspects of the texts. We tested the accuracy of a simple logistic regression using function words, and the results were higher than 50% by a non-trivial amount. This is a potential indicator that misogynistic and/or aggressive tweets have a slightly different syntax than those that do not fit in either of the two. Moreover, using the tf-idf vectorizer on plain function words achieved 0.628 F1 on the test set for misogyny identification, a result that is not at all negligible, given that these words do not encapsulate meaning.

[REF2] - paperID: 63483c9387d17e44eeb70c7321ad0dbb59b994fc

Title: Universal Multimodal Representation for Language Understanding

Chunk of text: Our approach can be easily applied to text-only tasks without manually annotated multimodal parallel corpora. Therefore, our method is universal in terms of the task requirements, in contrast to the recent vision-language models that require large-scale and expensive annotation datasets for each downstream task. The proposed method is evaluated on 14 NLP benchmark datasets involving natural language inference (NLI), semantic similarity, text classification, and machine translation. The experiments and analysis verify the effectiveness of the proposed method. To summarize, our contributions are primarily three-fold: (i) This work studies the universal visual representation for language representation in a broader view of the natural language processing scenario. Besides neural machine translation, this work leverages visual information as assistant signals for general NLP tasks, with the focus on investigating the global multimodality for general NLP, interpretability of effectiveness, and quality control of using universal visual representation. (ii) For the technical side, this work proposes new methods of semantic sentence-image matching from a shared cross-modal space to give more accurately paired images as topic information.

[REF3] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: arXiv preprint arXiv:1910.01108. Bin Wang and C.-C. Jay Kuo. 2020. SBERT-WK: A Sentence Embedding Method by Dissecting BERTbased Word Models. arXiv:2002.06652 [cs], June.55 MDD @ AMI: Vanilla Classifiers for Misogyny Identification Samer El Abassi Faculty of Mathematics and Computer Science University of Bucharest [samer.el-abassi@s.unibuc.ro](mailto:samer.el-abassi@s.unibuc.ro) Sergiu Nisioi Human Language Technologies Research Center, University of Bucharest [sergiu.nisioi@unibuc.ro](mailto:sergiu.nisioi@unibuc.ro) Abstract In this report1 , we present a set of vanilla classifiers that we used to identify misogynous and aggressive texts in Italian social media. Our analysis shows that simple classifiers with little feature engineering have a strong tendency to overfit and yield a strong bias on the test set.

[REF4] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: 13–23. Telmo Pires, Eva Schlinger, and Dan Garrette. “How multilingual is Multilingual BERT?” In: arXiv preprint arXiv:1906.01502 (2019). Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i-Nieto. “Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation”. In: arXiv preprint arXiv:1910.02334 (2019).

[REF5] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: The most effec-322 tive predictors are those derived from context-dependent models (e.g., BERT), but relatively good results can be obtained also by using context-independent models (e.g., Skip-gram) and behavioural norms (e.g., ratings of semantic diversity). Such an approach works very well for English, but less so for Italian, where the range of available predictors (i.e., pre-trained distributional models and large behavioural norms) is limited. One surprisingly effective solution to this problem is to simply translate the Italian stimuli into English, by relying on a neural machine translation system (e.g., MarianMT), and then make use of existing predictors for English. As an alternative to translating stimuli, it would be interesting to test whether comparable results can be obtained using multilingual versions of context-dependent models, such as BERT. Acknowledgements We would like to thank the anonymous reviewers, for their comments and suggestions, as well as the organizers of the competition, for their support. Table 1. Type and number of predictors obtained from behavioural norms and distributional models.

[REF6] - paperID: 21ee4b66ce53de6b7b23c23cae0885bf5c96ad78

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

Chunk of text: For the MS-MARCO-Pas, initial retrieval is performed using the Anserini toolkit with the BM25 model to obtain the top 100 ranked passages following . The ranked list Lc is obtained by utilizing the target ranking model to re-rank the above initial ranked list and the length N is set to 100. We set k = 1 in Equation (6) since every query in the MS-MARCO-Doc and most queries in the MS-MARCO-Pas have only one relevant document. In the token importance ranking process, the number of top important tokensm in PRADA is set to 50 and 20 for the MS-MARCO-Doc and MS-MARCO-pas, respectively. For fair comparison with 1https://tfhub.dev/google/universal-sentence-encoder/2. ACM Transactions on Information Systems, Vol. 41, No. 4, Article 89. Publication date: April 2023.89:16 C. Wu et al.

[REF7] - paperID: 63483c9387d17e44eeb70c7321ad0dbb59b994fc

Title: Universal Multimodal Representation for Language Understanding

Chunk of text: Prior studies have verified that representations of images and text can be jointly leveraged to build visual-semantic embeddings in a shared representation space [39, 40, 41, 44]. To this end, a popular approach is to connect both the monomodal text and image encoding paths using fully connected3 layers [45, 46]. The shared deep embedding can be used for cross-modal retrieval; thus, it can associate sentence text with associated images. Partly inspired by this line of research, we are motivated to incorporate visual awareness into sentence modeling by retrieving a group of images for a given sentence. One of the first techniques to align two views of heterogeneous data is the canonical correlation analysis method , in which linear projections defined on both sides are optimized to maximize the cross-correlation. Recent studies have followed the two-path architecture [45, 46], in which the encoder consists of a joint embedding of textual and image representations extracted from both the images and corresponding caption. Notably, Engilberge et al.

[REF8] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: Vocabulary alignment for collaborative agents: a study with real-world multilingual how-to instructions. In IJCAI, pages 159–165. D. Colla, E. Mensa, A. Porporato, and D.P. Radicioni. 2018. Conceptual Abstractness: From Nouns to Verbs. In Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), volume 2253. CEUR.

[REF9] - paperID: 21ee4b66ce53de6b7b23c23cae0885bf5c96ad78

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

Chunk of text: The SSdoc of TSrep has a larger drop than PRADA in the range of [30, 60], and the performance of PRADA in terms of SR and PP is always better than TSrep with different m. These results again illustrate the effectiveness of PRADA. 6.8 Case Study To obtain a better qualitative understanding of how different models perform, we show the adversarial examples from PRADA as well as that from TSrep , with the number of important tokens m set to 50. We take one query “government does do” from the dev set of the MS-MARCO-Doc as an ACM Transactions on Information Systems, Vol. 41, No. 4, Article 89. Publication date: April 2023.89:22 C. Wu et al. Table 9. Adversarial Samples Generated by TSrep and PRADA on the MS-MARCO-Doc Dataset Method Query: “government does do” Rank Position Original . . . what kind of government does japan have today?

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

Title: Incorporating Graph Information for Neural Information Retrieval

Graph-based models have been proposed as a way to better relate concepts and establish new connections that make documents more findable [REF0]. Hypergraphs, in particular, have been useful in music recommendation and compressing semantic graphs [REF4]. However, hypergraphs still present some limitations when applied to more complex representation needs [REF4]. To tackle entity-oriented search using graph-based approaches for representation and retrieval, a model called graph-of-entity has been proposed as a novel approach for indexing combined data, where terms, entities, and their relations are jointly represented [REF6].

In addition to graph-based models, deep generative ranking models have also been investigated for information retrieval [REF9]. The importance of the copy mechanism in the generative models in the context of retrieval has been shown, as provided by the PGN and T-PGN models [REF9].

To compute global statistics and plot distributions for the full hypergraph, the number of nodes, hyperedges, average degree, average clustering coefficient, average path length, diameter, and density are computed [REF1]. For each snapshot, the average node degree, hyperedge cardinality, diameter, path length, clustering coefficient, and density are computed and plotted to show its evolution as the number of documents increases [REF1].

Overall, incorporating graph information into neural information retrieval has shown promise in improving search results and establishing new connections between concepts. However, further research is needed to address the limitations of hypergraphs and to explore the potential of deep generative ranking models in information retrieval.

References:

[REF0] As opposed to the back-of-the-book index, an inverted index contains most of the terms in the collection, usually discarding frequent words (stopwords) and sometimes storing a reduced form of the word (obtained from stemming or lemmatization). Automatization means that a larger volume of data can be processed efficiently, and stored statistics can be used as a way to measure relevance. However, one thing that is lost with the inverted index is the ability to relate concepts. In the back-of-the-book index, a domain expert might 402.2 graph-based models provide associations between concepts (e.g., using ‘see also’) or use keywords that are not explicitly mentioned in the page despite being more adequate for search. The inverted index is usually focused on representing the document as is, however we can use techniques like query expansion or latent semantic indexing to establish new connections that make documents more findable. With query expansion we can, for instance, also consider the synonyms of the query keywords to increase recall. With latent semantic indexing we can establish new relations based on contextual similarity, or we can use approaches like word2vec or explicit semantic analysis for a similar purpose.

[REF1] For the full hypergraph of each of the four models, we compute the following global statistics: • Number of nodes, in total and per type; • Number of hyperedges, in total, per direction, and per type; • Average degree; • Average clustering coefficient; • Average path length; • Diameter; • Density. We also plot the following distributions for the full hypergraph: • Node degree distributions per node type: – Node-based node degree; – Hyperedge-based node degree. • Hyperedge cardinality distributions per hyperedge type. Then, we define a temporal analysis framework based on an increasing number of documents (i.e., time passes as documents are added to the hypergraph-of-entity index). We prepare several snapshots, with a different number of documents each, for each of the four models. We then compute and plot the following statistics for each snapshot, showing its evolution as the number of documents increases: • Average node degree over time; • Average hyperedge cardinality over time; • Average diameter and average path length over time; • Average clustering coefficient over time; • Average density over time. • Size over time: – Number of nodes; – Number of hyperedges; – Space in disk; – Space in memory.

[REF4] Had noticed the inability of anthropologists and sociologists to study social networks based only on dyadic relationships, proposing hypergraphs as a way of better modeling non-dyadic relationships, such as group membership. Moreover, hypergraphs have already been particularly useful in music recommendation [83, 209–211] through unified approaches for modeling heterogeneous data or through the use of random walks. Given their ability to represent polyadic relations that group multiple nodes, hypergraphs have also been used as a way to compress semantic graphs . Assuming that we would be able to effectively represent text and entities using a hypergraph, then we might be able to take advantage of both set theory, using metrics like the Jaccard index to measure similarities, or random walks in hypergraphs , where we might rely on hyperedge weights, but also on node weights to control the traversal. While hypergraphs are a flexible data structure, they still present some limitations, when applied to more complex representation needs. For instance, weights associated with nodes and hyperedges might not be enough to represent all types of bias — e.g., we can define node weights, but not node weights per hyperedge. There are, however, extensions of hypergraphs, like fuzzy hypergraphs , intuitionistic fuzzy hypergraphs or hypergraphs with edge-dependent vertex weights , that provide increased flexibility in establishing bias.

[REF6] As search becomes increasingly dependent on the integration of text and knowledge, novel approaches for a joint representation of corpora and knowledge bases present the opportunity to unlock new ranking strategies. We tackle entity-oriented search using graph-based approaches for representation and retrieval. In particular, we begin by proposing a model called graph-of-entity, as a novel approach for indexing combined data, where terms, entities and their relations are jointly represented. We compare the graph-of-entity with the graph-of-word, a text-only model, verifying that, overall, it does not yet achieve a better performance, despite obtaining a higher precision. Our assessment was based on the INEX 2009 10T-NL subset, described in Section 4.1.1.2, which was created from a sample of 10 topics and respectively judged documents. The offline evaluation we do here is complemented by its counterpart from TREC 2017 OpenSearch track, where, during our participation, we assessed the graph-of-entity in an online setting, through team-draft interleaving. The online evaluation at TREC 2017 OpenSearch was carried over the SSOAR site (Social Science Open Access Repository), using the title and the abstract as a text block and the remaining metadata as a knowledge block.

[REF9] Proposing a novel deep generative ranking model, T-PGN, we investigate the performance of several deep generative IR models on two passage retrieval collections. Our evaluation results show the importance of the copy mechanism in the generative models in the context

[REF0] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: As opposed to the back-of-the-book index, an inverted index contains most of the terms in the collection, usually discarding frequent words (stopwords) and sometimes storing a reduced form of the word (obtained from stemming or lemmatization). Automatization means that a larger volume of data can be processed efficiently, and stored statistics can be used as a way to measure relevance. However, one thing that is lost with the inverted index is the ability to relate concepts. In the back-of-the-book index, a domain expert might 402.2 graph-based models provide associations between concepts (e.g., using ‘see also’) or use keywords that are not explicitly mentioned in the page despite being more adequate for search. The inverted index is usually focused on representing the document as is, however we can use techniques like query expansion or latent semantic indexing to establish new connections that make documents more findable. With query expansion we can, for instance, also consider the synonyms of the query keywords to increase recall. With latent semantic indexing we can establish new relations based on contextual similarity, or we can use approaches like word2vec or explicit semantic analysis for a similar purpose.

[REF1] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: For the full hypergraph of each of the four models, we compute the following global statistics: • Number of nodes, in total and per type; • Number of hyperedges, in total, per direction, and per type; • Average degree; • Average clustering coefficient; • Average path length; • Diameter; • Density. We also plot the following distributions for the full hypergraph: • Node degree distributions per node type: – Node-based node degree; – Hyperedge-based node degree. • Hyperedge cardinality distributions per hyperedge type. Then, we define a temporal analysis framework based on an increasing number of documents (i.e., time passes as documents are added to the hypergraph-of-entity index). We prepare several snapshots, with a different number of documents each, for each of the four models. We then compute and plot the following statistics for each snapshot, showing its evolution as the number of documents increases: • Average node degree over time; • Average hyperedge cardinality over time; • Average diameter and average path length over time; • Average clustering coefficient over time; • Average density over time. • Size over time: – Number of nodes; – Number of hyperedges; – Space in disk; – Space in memory.

[REF2] - paperID: 069c109507ee685dfe04534f0461b837c5cb224d

Title: Pre-trained Language Model based Ranking in Baidu Search

Chunk of text: Neural computation (2004). Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In CIKM’14. Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE:

[REF3] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: Callan has also more recently worked on with Chenyan Xiong, developing work on joint representation models for words and entities [257, 258], as well as for entity-oriented search using learning to rank , even contributing with a test collection based on DBpedia . Roi Blanco has contributed with the groundwork for graph-based information retrieval, proposing several approaches for modeling documents as graphs and computing term weights from this representation . David F. Gleich has done strong contributions in the area of PageRank, both providing an excellent survey on the subject , and proposing the Multilinear PageRank as a higher-dimension generalization of PageRank applicable to tensors. Renaud Delbru has also contributed with PageRank approaches, applied to entity-oriented search, or more specifically the web of data [72, 259]. He was also one of the creators, along with Balog and Tummarello (also on the top 10), of the Sindice Dataset , an historical test collection from entity-oriented search that is no longer available sensibly since the creation of the SindiceTech startup company.

[REF4] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: had noticed the inability of anthropologists and sociologists to study social networks based only on dyadic relationships, proposing hypergraphs as a way of better modeling nondyadic relationships, such as group membership. Moreover, hypergraphs have already been particularly useful in music recommendation [83, 209–211] through unified approaches for modeling heterogeneous data or through the use of random walks. Given their ability to represent polyadic relations that group multiple nodes, hypergraphs have also been used as a way to compress semantic graphs . Assuming that we would be able to effectively represent text and entities using a hypergraph, then we might be able to take advantage of both set theory, using metrics like the Jaccard index to measure similarities, or random walks in hypergraphs , where we might rely on hyperedge weights, but also on node weights to control the traversal. While hypergraphs are a flexible data structure, they still present some limitations, when applied to more complex representation needs. For instance, weights associated with nodes and hyperedges might not be enough to represent all types of bias — e.g., we can define node weights, but not node weights per hyperedge. There are, however, extensions of hypergraphs, like fuzzy hypergraphs , intuitionistic fuzzy hypergraphs or hypergraphs with edge-dependent vertex weights , that provide increased flexibility in establishing bias.

[REF5] - paperID: 06227bc74bcee55471fb37bde0149b317f8a2014

Title: Enhancing Semantic Code Search With Deep Graph Matching

Chunk of text: All of these studies follow a similar methodology, which involves using sequence encoders (such as RNN, LSTM, Attention, etc.) to first map both the code and the NL representation into vector format, and then compute vectors’ similarity (i.e., L2 and cosine). However, SeCoDeGrMa is different from earlier research in two crucial ways: (1) In the beginning, we propose a systematic structured-graph approach for the representation of query and code fragments. The graph representations are learned using GNN which is capable to retained structural and semantic information for a long term; (2) In addition, for improved vector representation, the SeCoDeGrMa model tries to exploit more semantic information and relationships of code fragments and NLqueries. D. DIFFERENT TASKS OF SOURCE CODE Other study fields that are currently active in this area include tasks like code summarization, code generation, and code generation which use machine learning , – . Fernandes et al. introduced a hybrid approach that enhances traditional sequence encoder schemes with GNNs.

[REF6] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: As search becomes increasingly dependent on the integration of text and knowledge, novel approaches for a joint representation of corpora and knowledge bases present the opportunity to unlock new ranking strategies. We tackle entity-oriented search using graph-based approaches for representation and retrieval. In particular, we begin by proposing a model called graph-of-entity, as a novel approach for indexing combined data, where terms, entities and their relations are jointly represented. We compare the graph-of-entity with the graphof-word, a text-only model, verifying that, overall, it does not yet achieve a better performance, despite obtaining a higher precision. Our assessment was based on the INEX 2009 10T-NL subset, described in Section 4.1.1.2, which was created from a sample of 10 topics and respectively judged documents. The offline evaluation we do here is complemented by its counterpart from TREC 2017 OpenSearch track, where, during our participation, we assessed the graph-of-entity in an online setting, through team-draft interleaving. The online evaluation at TREC 2017 OpenSearch was carried over the SSOAR site (Social Science Open Access Repository), using the title and the abstract as a text block and the remaining metadata as a knowledge block.

[REF7] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: Listing 5.1 shows the SPARQL query used to calculate scoreclk(e, Ee) for all #lode:Event instances in the system. The scorepop(e, Ee) is calculated using a similar query, where we remove the statement in line 8, discarding the constraint for clicked events. As we can see, each entity score is calculated, per event, based on the total number of links to the entities that are associated with the event. We illustrate this by using lode:involvedAgent property for classes #dul:Person and #dul:Organization. After computing and storing the two entity scores for each event, we calculate the final score, for event ranking, as shown in Equation 5.1. Given event e, and entities Ee associated with event e, the final score is calculated based on a weighted average of three factors: days to event ∆Te, entity popularity score scorepop(e, Ee), and entity click score scoreclk(e, Ee). score(e, Ee) = w1 1 ∆Te + 1 + w2 scorepop(e, Ee) maxe scorepop(e, Ee) + w3 scoreclk(e, Ee) maxe scoreclk(e, Ee) (5.1)

[REF8] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: The original PageRank uses Jelinek-Mercer smoothing to combine the probability distribution of following a link with the probability distribution of randomly jumping to another page. Wang et al. [158, 385] explored another type of smoothing, called Dirichlet smoothing, as an alternative, thus proposing a Dirichlet PageRank (or DirichletRank). Their idea was that the outgoing links of hubs should have a higher probability of being followed and, as such, the probability of randomly jumping to a page should not be uniform, but dependent on the number of outgoing links — an indicator of a good hub. Another way to look at it is that the random walk should be biased towards visiting authorities (i.e., with strong links from hubs), as well as more likely to restart from hubs, since this more closely resembles the expected surfer behavior — it’s more probable for the surfer to restart from a good source node to increase the available paths to take. The authors also identified a zero-one gap problem with the original PagerRank that can be solved in an elegant manner using Dirichlet smoothing. In PageRank, when we reach a sink, the probability of randomly jumping to another page becomes one, but, as soon as there is a single outgoing link, this probability immediately drops to β = (1 − d), the complement of the damping factor1 , usually set to β = 0.15 (or d = 0.85).

[REF9] - paperID: d7d46a173fcb6808d1c78734b9d708078a20fc41

Title: A Modern Perspective on Query Likelihood with Deep Generative Retrieval Models

Chunk of text: Proposing a novel deep generative ranking model, T-PGN, we investigate the performance of several deep generative IR models on two passage retrieval collections. Our evaluation results show the importance of the copy mechanism in the generative models in the context of retrieval, as provided by the PGN and T-PGN models. We further explore the information provided by the uncertainty estimates, and showcase the value of such uncertainty information in a cut-off prediction task. ACKNOWLEDGEMENTS This research is supported in part by the NSF (IIS-1956221). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF or the U.S. Government. This research is also supported by Know-Center Graz, through project “Theory-inspired Recommender Systems”.