Title: Investigating Learned Text Representation for Information Retrieval

Investigating Learned Text Representation for Information Retrieval

In recent years, there has been a growing interest in the use of deep learning techniques for natural language processing tasks, including information retrieval. One of the most promising approaches is the use of learned text representations, which can capture the semantic and syntactic properties of text in a more effective way than traditional bag-of-words models. In this section, we review some recent studies that investigate the use of learned text representations for information retrieval.

Giorgioni et al. (REF0) proposed a transformer-based architecture for stance detection in Italian tweets. The authors used transfer learning to improve the accuracy of their model, and investigated the contribution of three auxiliary tasks related to stance detection. The results showed that the proposed approach outperformed other methods in the competition, confirming the effectiveness of transformer-based architectures and transfer learning.

Ferracciolia et al. (REF1) developed a statistical model for predicting polarization in social networks. The authors used features extracted from both the texts and the social network of users, and trained the model using extreme gradient boosting. The results showed that the use of Multidimensional Scaling of the distance matrix on each network was a key factor for good performance.

In a study by Delmonte (REF9), a system for syntactic and semantic processing based on the parser for Italian was used to analyze each sentence. The author used both a fully and mixed statistically based approach and the semantic one used previously. The results showed that the use of n-grams and the usual tf-idf indices improved the performance of the model.

In a study by Xie et al. (REF4), a neural network classifier was developed for concept pair classification. The authors used a customizable set of features and a 20 neurons hidden layer trained on 15 epochs. The results showed that the input layer dynamically matching the size of the feature vector improved the performance of the model.

Finally, in a study by ArchiMeDe (REF7) a multimodal learning ensemble was developed for meme detection. The authors used sentence embeddings produced by the UmBERTo NLM concatenated to metadata and image embeddings produced by three popular pre-trained vision models. The results showed that the use of multiple views over an image improved the performance of the model.

In conclusion, the studies reviewed in this section demonstrate the effectiveness of learned text representations for information retrieval tasks. The use of deep learning techniques, transfer learning, and multimodal learning ensembles can improve the accuracy of models and provide complementary information about the content of text.

References given to GPT:

[REF0] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: Combining Transformer-based Architectures and Transfer Learning for Robust Stance Detection Simone Giorgioni, Marcello Politi, Samir Salman, Danilo Croce and Roberto Basili Department of Enterprise Engineering, University of Roma, Tor Vergata Via del Politecnico 1, 00133 Roma, Italy {[simone.giorgioni,marcello.politi,samir.salman}@alumni.uniroma2.eu](mailto:simone.giorgioni,marcello.politi,samir.salman%7D@alumni.uniroma2.eu) {[croce,basili}@info.uniroma2.it](mailto:croce,basili%7D@info.uniroma2.it) Abstract English. This paper describes the UNITOR system that participated to the Stance Detection in Italian tweets (Sardistance) task within the context of EVALITA 2020. UNITOR implements a transformer-based architecture whose accuracy is improved by adopting a Transfer Learning technique. In particular, this work investigates the possible contribution of three auxiliary tasks related to Stance Detection, i.e., Sentiment Detection, Hate Speech Detection and Irony Detection. Moreover, UNITOR relies on an additional dataset automatically downloaded and labeled through distant supervision. The UNITOR system ranked first in Task A within the competition. This confirms the effectiveness of Transformer-based architectures and the beneficial impact of the adopted strategies.

[REF1] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: A Smart Use of Social Network Data in Predicting Polarization Federico Ferracciolia , Andrea Sciandrab , Mattia Da Pontc , Paolo Girardia , Dario Solarid , Domenico Madonnaa , Livio Finosa a. Universita degli Studi di Padova ` b. Universita degli Studi di Modena e Reggio Emilia ` c. WMRI d. BeeViva [ferraccioli@stat.unipd.it](mailto:ferraccioli@stat.unipd.it), [andrea.sciandra@unimore.it](mailto:andrea.sciandra@unimore.it), [mattia.dapont@wmr.it](mailto:mattia.dapont@wmr.it), [paolo.girardi@unipd.it](mailto:paolo.girardi@unipd.it), [dario.solari@gmail.com](mailto:dario.solari@gmail.com), [domenico.madonna@studenti.unipd.it](mailto:domenico.madonna@studenti.unipd.it), [livio.finos@unipd.it](mailto:livio.finos@unipd.it) Abstract In this contribution we describe the system (i.e. a statistical model) used to participate in Evalita conference 2020, SardiStance (Tasks A and B) and Haspeede2 (Tasks A and B). We first developed a classifier by extracting features from the texts and the social network of users. Then, we fit the data through an extreme gradient boosting, with cross-validation tuning of the hyper-parameters. A key factor for a good performance in SardiStance Task B was the features extraction by using Multidimensional Scaling of the distance matrix (minimum path, undirected graph) applied on each network. The second system exploits the same features above, but it trains and performs predictions in twosteps. The performances proved to be lower than those of the single-step model.

[REF2] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: For mBERT, AlBERTo, Italian BERT XXL and UmBERTo the best configuration was: maximum sequence length 256, batch 32, learning rate 5e-5, and 5 epochs. For GilBERTo we used the same values except the number of epochs, which was increased to 10. Finally, the best performing hyperparameters for XLM-RoBERTa was the following: maximum sequence length 256, batch 16, learning rate 2e-5, and 10 epochs. While the monolingual models clearly outperformed both mBERT and XLM-RoBERTa on the development data, we decided to submit the three best monolingual runs and the best multilingual one. Table 2 reports the official results obtained by each of the models and their position with respect to the ranking of constrained runs for Task A released by the task organizers. Our submission based on Italian BERT XXL was clearly the best of our four runs, although its performance was around 1.5 scores in F1 lower than the winner system for Task A. Furthermore, the ranking obtained in the test does not correspond with the results obtained during the development phase, where UmBERTo outperformed the other monolingual models by more than 3 points in F1 score.

[REF3] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: It would be quite expected for such semantic neighborhood to entitle heart and pain to higher concreteness values than what they receive in KonKretiKa. A more in-depth analysis into this contradiction revealed that it stems from the vulnerability in the semantic composition of the concrete paradigm which was used to compute the raw indexes (see Table 1). The words of this paradigm belong to the two major semantic classes – living organisms (animals and plants) and manmade artifacts. The class of words denoting human beings was intentionally excluded when the paradigm was compiled on the grounds that such nouns tend to indicate abstract social roles rather than physical humans. As a consequence, physical organic objects such as body parts and organs, or physical sensations and physiological conditions received non-uniform indexes in KonKretiKa: those that refer to humans as well as to animals (e.g. in veterinary or gastronomic discourse) ranked rather high in concreteness: e.g. liver (6.6), pancreas (6.4), foot (6.3), encephalitis (6.25), kidney (6.25), entrails (6.05), tummy (5.92), womb (5.6) – whereas those that tend to be primarily associated with humans received lower indexes, e.g. heart (2.63), heartburn (2.57), scar (2.53), nausea (2.5), headache (1.61), distress (1.5), pain (1.21), queasiness (1.12), etc. Thus, comparison of the KonKretiKa computational indexes with the psycholinguistic data of CONcreTEXT allowed us to detect a potential shortcoming in our approach to the design of the concrete paradigm. As was noted in previous study (Badryzlova, 2020), the class of concrete words seems to be more semantically heterogeneous than of abstract words; therefore, it may reasonable in future experiments to diversify the concrete paradigm and expand it in size by including words that denote human beings.

[REF4] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: The output layer consists of one neuron with sigmoid activation function. Some structural properties of the classifier can be customised by the user from a dedicated GUI. In particular, for what concerns the structure of the neural network the user can define the size of the hidden layer and the number of epochs, while for the evaluation the user can set the number of cross validation folds. Moreover, training can be performed on a customizable set of features (see Section 2.2 for the complete list) since the input layer is set to dynamically match the size of the feature vector. For the specific purposes of this work, we used in every scenario a model exploiting a 20 neurons hidden layer trained on 15 epochs. A 4- fold cross validation was used for the in-domain scenario. Training The official training set containing concept pairs and their binary labels was formatted as a pair of numpy arrays: one of them has variable length and contains the serialization of the features, which will be the model input, whilst the latter contains the binary labels of the pairs.

[REF5] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: 3.2 Italian BERT Finally, we use a pre-trained BERT4 to accomplish the calculation of a deep representation of the text. One of the most widely used autoencoding pre-trained Language Models (PLMs) is BERT (Devlin et al., 2018). BERT is trained using the masked language modeling task that randomly masks some tokens in a text sequence, and then independently recovers the masked tokens by conditioning on the encoding vectors obtained by a bidirectional Transformer. Inside BERT, the information is passed forward crosswise transformer layers. In this work, we used a specific output from one of those layers, this operation can be expressed by: h0 = Hl0(texttok) hi = Hli(hi−1)

[REF6] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: In this work, we used a specific output from one of those layers, this operation can be expressed by: h0 = Hl0(texttok) hi = Hli(hi−1) hn = Hln(hn−1) where texttok is the text after its tokenization5 , hi is the output of the i th transformer layer(Hli) called hidden state and n is the total transformer layers in BERT. Then, for an specific i, from the tensor of order 2 hi it is computed the vector fbert, as a deep representation of the initial text who will act as PLM feature. v = ­ k=0 hi [k, :] fbert = v ||v|| 3.3 Preprocessing In the preprocessing step, firstly stopwords were removed . Then, the hashtags composed of many words are split (e.g: #NessunDorma becomes # nessun dorma).

[REF7] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: Sentence embeddings produced by the UmBERTo NLM are concatenated to metadata and image embeddings produced by three popular pre-trained vision modals. The three resulting multimodal embeddings are fed separately to feedforward networks, and the final outcome is selected through majority voting. task of DANKMEMES, aimed at discriminating memes from standard images containing actors from the Italian political scene. Task organizers extracted a total of 1600 training images from the Instagram platform, and data available from each dataset entry – text, actors and user engagement, among others – were leveraged to train an ensemble of multimodal models performing meme detection through majority-vote. The following sections present our approach in detail, first showing our preliminary evaluation of multiple modeling approaches and then focusing on the final system’s main modules and the features we leverage from the dataset. Finally, results are presented, and we conclude by discussing the problems we faced with some inconsistencies in the data. Our code is made available at <https://github.com/> jinensetpal/ArchiMeDe 2 System Description ArchiMeDe is composed of a multimodal learning ensemble, with the final output being the result of a majority vote.

[REF8] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: These embeddings can be considered as different views over an image that may provide us with complementary information about its content. Then, each image embedding is concatenated with the sentence embedding and the raw image metadata and fed as input to an 8-layer feed-forward neural network to predict an image’s meme status. The feed-forward network also includes a single dropout layer to prevent overfitting and improve generalization. Lastly, the three predictions are weighted through majority voting to obtain the final prediction of the ensemble. Other simpler strategies using a single vision model to produce image embeddings were initially envisaged as potential candidates for our submission but were finally dismissed in light of the promising performances of the ArchiMeDe ensembling approach. We discuss those perspectives in Section 4. The remaining part of this section contains an296 in-depth description of our ensemble’s components, focusing on the input features that were used and how those were preprocessed to best suit learning.

[REF9] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0

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

Chunk of text: HaSpeeDe2 & SardiStance: Multilevel Deep Linguistically Based Supervised Approach to Classification Rodolfo Delmonte Dipartimento di Studi Linguistici e Culturali Comparati Ca’ Bembo – Dorsoduro 1075 – Università Ca’ Foscari – 30131 Venezia [delmont@unive.it](mailto:delmont@unive.it) Abstract In this paper1 we present the results obtained with ItVENSES a system for syntactic and semantic processing that is based on the parser for Italian called ItGetaruns to analyse each sentence. In previous EVALITA tasks we only used semantics to produce the results. In this year EVALITA, we used both a fully and mixed statistically based approach and the semantic one used previously. The statistic approaches are all characterized by the use of n-grams and the usual tf-idf indices.

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

Title: Efficient and Effective Text Representation for Information Retrieval

Efficient and Effective Text Representation for Information Retrieval

Text representation is a crucial step in information retrieval, as it determines how well a system can understand and retrieve relevant information from a large corpus of text. In recent years, various approaches have been proposed to improve the efficiency and effectiveness of text representation. In this section, we will discuss some of the recent developments in this area.

One approach to improving text representation is to incorporate metadata about the text, such as contextual and community-based features (REF1). This has been shown to be particularly effective in discriminating the stance of tweets, where metadata about the tweets was found to be more useful than the textual information of the tweets themselves (REF1). Another approach is to identify aspect terms in the text, which are product characteristics that are annotated based on the opinions expressed about them (REF2). This allows for a more fine-grained representation of the text, which can improve the accuracy of information retrieval.

Another important aspect of text representation is the use of deep learning models, such as BERT, to extract features from the text (REF1). These models have been shown to be highly effective in various natural language processing tasks, including information retrieval. However, the use of deep learning models can also be computationally expensive, which can limit their efficiency in large-scale information retrieval systems.

To address this issue, some researchers have proposed the use of multimodal architectures that specialize in processing different types of information, such as text and images (REF4). This allows for a more efficient use of computational resources, as each component of the architecture can focus on a specific type of information. Additionally, some researchers have proposed the use of data augmentation and certified defenses to improve the robustness of text representation models against adversarial attacks (REF7).

In conclusion, text representation is a critical component of information retrieval systems, and recent developments in this area have focused on improving the efficiency and effectiveness of text representation through the use of metadata, aspect terms, deep learning models, multimodal architectures, and defenses against adversarial attacks. These developments have the potential to significantly improve the accuracy and scalability of information retrieval systems.

References given to GPT:

[REF0] - paperID: 7177d99f5a873ba8ad2772edbb02f85fcd281566

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

Chunk of text: 𝐾, for all 𝑑 ∈ 𝐿 𝑠 𝑞 [𝐾 + 1 :] and any 𝑑 ′ ∈ 𝑆𝑑 if Δ𝐿𝑞 def = ¯𝑓 (𝑞, 𝑑𝐾) − ¯𝑓 (𝑞, 𝑑𝐾+1) − max 𝑑 ∈𝐿 𝑠 𝑞 [𝐾+1:] 𝑜𝑑 > 0, (3) where Δ𝐿𝑞 can be estimated by Monte Carlo estimation as we show in the next section. We can certify whether the ranking model is top-𝐾 robust on the ranked list 𝐿 𝑠 𝑞 by simply checking Δ𝐿𝑞. If Δ𝐿𝑞 is positive, the model is certified top-𝐾 robust.

[REF1] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: Metadata about the tweets have served in discriminating the stance better than the textual information of the tweets themselves. 5 Conclusion In this paper, we presented the suitable models for stance detection in Italian tweets about Sardine movement. The three stances considered for this work are in favour of the movement, against and neutral. Multilayer perceptron is the classifier used for classification of stance of tweets. The deep learning pre-trained model BERT has been used to extract the features from the tweets along with several stylistic, contextual and community based features namely, The features are extracted Unigram , Char-grams , num-hashtag , Length, network quote community, network reply community, network retweet community, network friend community, user info bio, tweet info retweet, tweet info create at are few of the attributes that are extracted to detect the stance. The Models are trained using the dataset provided by SardiStance task for textual and contextual stance detections.

[REF2] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: In other words, if an opinion is expressed about the object itself, and it is then stated for which characteristic the judgment is applied, these characteristics are annotated as an aspect term (i.e., “questo telefono e ottimo, ` soprattutto per la durata della batteria”); 3. Deductible: the opinion is not expressed directly but it is inferable from the review or from the knowledge of the reference domain. The aspect term must represent the product characteristics, but it cannot represent a concept that is larger than the product itself. An aspect term does not identify opinions regarding elements external to the object, such as: (a) The shipment (it is not an intrinsic property of the object); (b) the company that produced them, the series to which the product belongs or other products with which the object is compared; (c) the elements that refer to the action of purchasing the item; (d) the elements that refer to the customer care. Moreover, in the case of aspect term composed of several words, all the words that make up the aspect term must be contiguous. In case they are separated by one or more words that are not part of the aspect term, the whole expression is discarded. More details and example of annotations are available on the task website2 . 5 Evaluation measures and baselines We evaluate the three subtasks (ATE, ABSA and SA) separately by comparing the results obtained by the participant systems on the gold standard annotations of the test set.

[REF3] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: UniBO 20 c 0.626 AMI the winner 21 c 0.490 NoPlaceForHateSpeech After the deadline the team UniBO submitted an amended run (\*\*) , that has not been ranked in the official results of the AMI shared task. However, we believe interesting to mention their achievement showing an Average Macro F1-score equal to 0.744. 5.2 Subtask B: Unbiased Misogyny Identification Table 6 reports the results for the Unbiased Misogyny Identification task, which received 11 submissions by 4 teams, among which 4 unconstrained and 7 constrained. The highest Average Macro F1 score has been achieved by jigsaw at 0.8825 with an unconstrained run and by PoliTeam at 0.8180 with a constrained submission. Similarly to the previous task, most of the systems have shown better performance compared to the AMI-BASELINE. By analizing the runs, we can highlight that the two best results achieved on Subtask B have been obtained by the unconstrained run submitted by jigsaw, where a simple debiasing technique based on data augumentation have been adopted, and by the constrained run provided by Politeam, where the problem of biased prediction Table 6: Results of Subtask B. Constrained runs are marked as “c”, while the unconstrained ones with “u”.

[REF4] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: Considering L the maximum number of input tokens, the remainder of L − K tokens are being split between the text tokens associated with a meme and G VGCN reserved slots. Those slots are kept empty to be internally filled with VGCN embeddings during training. Alongside ordinary inputs required by ItalianBERT (i.e. input ids, input masks and segment ids ), we build a gcn ids vector similarly to input ids, by mapping each unique input token to the corresponding index in the task vocabulary Vtask; Vtask represents the set of tokens available in the task text corpus and in the ItalianBERT’s vocabulary. The second additional input is represented by a binary mask vector having the value of 1 for the VGCN reserved tokens, and 0 otherwise. During training, all ItalianBERT layers with the exception of the last 4 encoder blocks were frozen. 3.4 Multimodal Architecture The final solution consists of a multimodal architecture with two main components, each specialized on processing one informational channel, namely text or image-based. The dates are segmented and encoded by using complementary sine and cosine functions to preserve the cyclic characteristics of days (in a month) and months.

[REF5] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: In FIRE (Working Notes), pages 191–198.110 DH-FBK @ HaSpeeDe2: Italian Hate Speech Detection via Self-Training and Oversampling Elisa Leonardelli Fondazione Bruno Kessler Trento, Italy [eleonardelli@fbk.eu](mailto:eleonardelli@fbk.eu) Stefano Menini Fondazione Bruno Kessler Trento, Italy [menini@fbk.eu](mailto:menini@fbk.eu) Sara Tonelli Fondazione Bruno Kessler Trento, Italy [satonelli@fbk.eu](mailto:satonelli@fbk.eu) Abstract We describe in this paper the system submitted by the DH-FBK team to the HaSpeeDe evaluation task, and dealing with Italian hate speech detection (Task A). While we adopt a standard approach for fine-tuning AlBERTo, the Italian BERT model trained on tweets, we propose to improve the final classification performance by two additional steps, i.e. self-training and oversampling. Indeed, we extend the initial training data with additional silver data, carefully sampled from domain-specific tweets and obtained after first training our system only with the task training data. Then, we retrain the classifier by merging silver and task training data but oversampling the latter, so that the obtained model is more robust to possible inconsistencies in the silver data. With this configuration, we obtain a macro-averaged F1 of 0.753 on tweets, and 0.702 on news headlines. 1 Introduction Although hate speech detection may seem a solved task on English, with more than 60 systems participating in the last Offenseval edition reaching an F1 > 0.90

[REF6] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: However, the experiments imply a correlation between a text’s syntax and its misogynistic/aggressive value. This proposes the idea that text that falls into either categories, (or maybe even hate speech in general?) does have a slightly more recognisable grammatical pattern than text that isn’t. Whether it’s the POS n-grams, pronouns, or just function words, the wording matters and is worth looking into for more advanced feature engineering.59 References Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In 13th International Workshop on Semantic Evaluation, pages 54–63. Association for Computational Linguistics. Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro.

[REF7] - paperID: 7177d99f5a873ba8ad2772edbb02f85fcd281566

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

Chunk of text: Data augmentation [17, 42] is a representative empirical defense by augmenting adversarial examples with the original training data. Since empirical defenses are only effective for certain attacks rather than all attacks, a competition emerges between adversarial attacks and defense methods. To solve the attack-defense dilemma, researchers resort to certified defenses to make models provably robust to certain kinds of adversarial perturbations. Jia et al. and Huang et al. first proposed to certify the robustness to adversarial word substitutions by leveraging Interval Bound Propagation (IBP ) in NLP. These IBP-based methods are limited to continuous word embeddings and are not applicable to subword-level models like BERT.

[REF8] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: Computational Intelligence, 35(1):82–97, feb. Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2016. Exploring various linguistic features for stance detection. In Natural Language Understanding and Intelligent Applications, pages 840–847, Cham. Springer International Publishing. Michael Wojatzki, Torsten Zesch, Saif Mohammad, and Svetlana Kiritchenko. 2018. Agree or Disagree: Predicting Judgments on Nuanced Assertions.

[REF9] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3

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

Chunk of text: MIT Press, Cambridge, MA, USA. K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778. L. Jiao and J. Zhao. 2019. A survey on the new generation of deep learning in image processing. IEEE Access, 7:172231–172263. Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019

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

Title: Incorporating Graph Information for Neural Information Retrieval

Graphs have been widely used in various fields, such as image retrieval, bibliometrics, brain representation, protein-protein interactions, social networks, and counter-terrorism and intelligence [REF0]. In the field of information retrieval, graphs have been used to represent entities and their relationships, which can be used to improve retrieval performance. One approach to incorporating graph information for retrieval is to use hypergraphs, which can simplify and generalize the representation of entities and their relationships [REF3]. Another approach is to use random walks to sample the structure of the hypergraph, which can be used to learn a representation model for retrieval [REF5].

In the context of neural information retrieval, deep learning has been used to learn joint embeddings of text and image features for cross-modal retrieval [REF4]. The current approach is to learn modality-specific deep features in a common space for both modalities, using a two-branch framework that extracts visual and textual features and learns joint embeddings using an optimized objective [REF4]. However, the cross-modal retrieval performance is still far from satisfactory due to the heterogeneity gap between different modalities [REF4].

To improve retrieval performance, context representation can be used to compute the common or unified representation that exists between the code graph and query graph [REF9]. This can be achieved by computing the attention between the nodes of the code graph and the query graph, and using the attention to compute the context representation [REF9]. Additionally, BERT-based rankers can be used to classify the relevance of candidate entities, using Lucene as a candidate generator during training and prediction [REF8].

In summary, incorporating graph information for neural information retrieval can be achieved using hypergraphs, random walks, deep learning, context representation, and BERT-based rankers. These approaches can improve retrieval performance by representing entities and their relationships more accurately and learning joint embeddings of text and image features.

References given to GPT:

[REF0] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: , image retrieval has used them to establish similarity networks for reranking based on PageRank , and bibliometrics has relied on coauthorship network analysis to understand collaboration patterns and measure author importance . Graphs have been used to pose representations of the brain , to study protein-protein interactions , to analyze social networks 131.2 the importance of consolidating models and the semantic web , or even for counter-terrorism and intelligence . Graphs have also supported the development of unified frameworks across different areas. Take for instance Moro et al.

[REF1] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: This is an exercise to demonstrate how the core concepts of a search engine’s ranking function, defined in a basic TF-IDF (i.e., without normalizing term count), can be extended to reach a different ranking function, in this case the probabilistic approach given by BM25. Let us now identify the core components in BM25 , starting from Equation 9.1 and building on the analysis work already carried by Rousseau and Vazirgiannis [16, §5], as well as based on some observations by Turnbull : BM25(t, d) = (k1 + 1) × tf(t, d) k1 × 1 − b + b × |d| avdl + tf(t, d) × log N + 1 df(t) (9.1) Here, tf(t, d) is the frequency of term t in document d, N is the number of documents in the collection, and df(t) is the document frequency of term t. Before describing parameters |d|, avdl, k1 and b, let us first identify idf(t, D) = log N+1 df(t) as the inverse document frequency of term t in collection D, so that we can build up BM25 from a simple TF-IDF, as described in Equation 9.2. TF-IDF(t, d) = tf(t, d) × log N + 1 df(t) (9.2) By taking a probabilistic view, and departing from the three pillars of information retrieval, our goal should be to approximate the probability of retrieval and the probability of relevance, not only over document length, but also over term frequency and inverse document frequency. Were the probability of retrieval to perfectly match the probability of relevance and perfect a ranking could be obtained.

[REF2] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: N YM S ACL Annual Meeting of the Association for Computational Linguistics 76 ACM Association for Computing Machinery 74 ANT Ad hoc search of eNtities and Text 24, 102, 125 AP Average Precision 250 API Application Programming Interface 63, 106 ARPA Advanced Research Projects Agency 4 bpref Binary Preference 44 BRAT BRAT Rapid Annotation Tool 109, 111, 112 CERN Conseil Européen pour la Recherche Nucléaire 23, 53, 252 CIKM Conference on Information and Knowledge Management 73, 76 CLEF Conference and Labs of the Evaluation Forum 30, 60 CoRR Computing Research Repository 73, 76 CRF Conditional Random Field 109, 111 DCG Discounted Cumulative Gain 53 DING Dataset rankING 10 DLN Document Length Normalization 7 DUL DOLCE+DnS Ultralite 114 ECIR European Conference on Information Retrieval 76 ELC Entity List Completion 63 ESA Explicit Semantic Analysis 38, 50 FEUP Faculdade de Engenharia da Universidade do Porto 98 FEUP InfoLab Laboratory of Information Systems from the Faculty of Engineering of the University of Porto 125 FTP File Transfer Protocol 12 GATE General Architecture for Text Engineering 109 GMAP Geometric Mean Average Precision 84, 140, 222, 231 GoE Graph-of-Entity 155, 220 GoW Graph-of-Word 154, 220 GPU Graphics Processing Unit 252 xviiiGSF Groupwise Scoring Function 38 HGoE HyperGraph-of-Entity 181, 206, 220 HITS Hyperlink-Induced Topic Search 6, 8 IDF Inverse Document Frequency 4, 5, 7, 31, 232 IE Information Extraction 115 INEX INitiative for the Evaluation of XML Retrieval 10, 24, 30, 60, 62, 65–67, 81, 122 IR Information Retrieval 3, 4, 9, 30, 62–64, 70 JASIST Journal of the Association for Information Science and Technology 76 JSON Javascript Object Notation 96, 104 JSON-LD Javascript Object Notation for Linked Data 9 KDD SIGKDD Conference on Knowledge Discovery and Data Mining 76 LODE Linked Open Descriptions Of Events 114 MAiP Mean Average interpolated Precision 65, 66, see also MAiP MAP Mean Average Precision 39, 63, 65, 66, 84, 89, 126, 140, 222, 228, 231, 315, 318 MART Multiple Additive Regression Trees 38 MMR Maximal Marginal Relevance 221 NDCG Normalized Discounted Cumulative Gain 37, 39, 45, 50, 63, 231 NDCG@p Normalized Discounted Cumulative Gain at a cutoff of p 84, 140, 222 NER Named Entity Recognition 111, 222, 247, 334 NERD Named Entity Recognition and Disambiguation 15 NIST National Institute of Standards and Technology 4, 63, 97, 221 NLTK Natural Language ToolKit 109 OWL Web Ontology Language 105 P@n Precision at a cutoff of n 43, 84, 140, 222, 231, 318 PCA Principal Component Analysis 58 PDLN Pivoted Document Length Normalization 4, 6, 31, 232 PhD doctor of philosophy 11 POS Part-Of-Speech 7, 53, 109 RDF Resource Description Framework 8, 30, 39, 41, 46, 60, 149 RDFa Resource Description Framework in attributes 9 xixREF Related Entity Finding 63 REMBRANDT Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto 109 REST REpresentational State Transfer 106 RWS Random Walk Score 131, 134, 182, 206, 219, 317 SIGIR Special Interest Group on Information Retrieval xx SIGIR (conf.)

[REF3] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: If we instead ranked entity nodes using the same strategy, we would have generalized the problem to ad hoc entity retrieval. Claude Berge stated that hypergraphs could be used to simplify as well as to generalize . We have shown that simplification can significantly reduce the order of complexity when compared to graph-based models like graph-of-entity. This is for instance the case with representing synonyms based on a single hyperedge rather than multiple edges in a graph. Moreover, it is not only a more natural, but also a more efficient modeling approach. On the other hand, the hypergraph is an adequate data structure to generalize, since it is able to represent clusters (n-ary groupings of elements), and other kinds of relations (e.g., directed relations between n-ary groupings of elements). It is also able to indirectly capture overlap through intersections, or hierarchies through subsumption, and it can factor uncertainty through node and hyperedge weights.

[REF4] - paperID: 758890bef9a1a85a25a1f6831a58f00a462476af

Title: SMAN: Stacked Multimodal Attention Network for Cross-Modal Image–Text Retrieval

Chunk of text: Conceptual illustration of general deep image-text retrieval framework. to search for images that are most relevant to the topic of a textual query, or captions that precisely describe the content of a visual query. However, solving the CMR problem is not easy, because data from different modalities separately reside in heterogeneous feature spaces, thus giving rise to difficulties in measuring the semantic relevances between cross-modal instances. Recently, there has been a surge of work , , , , , , , proposed to tackle the imagetext retrieval problem. Under the umbrella of deep learning, the current predominant schemes opt to learn modalityspecific deep features in a common space for both modalities. More concretely, they usually adopt a two-branch framework (as shown in Fig.1) carrying out two basic steps - visual branch (e.g., Convolutional Neural Network (CNN)) and textual branch (e.g., Long Short-Term Memory (LSTM)) extract visual and textural features respectively, followed by deploying an optimized objective (e.g., bidirectional triplet ranking loss) to learn the joint embeddings. Although thrilling progresses , , , have been achieved, due to the existence of heterogeneity gap, the cross-modal retrieval performance is still far from satisfactory.

[REF5] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: can be seen as a form of randomized sampling of the structure of the hypergraph. The longer the random walk or the higher the number of repeats, the better the ability to capture hypergraph structure. This means that, assuming random walks do their job of correctly sampling structure, the representation model will then be the fundamental indicator of retrieval effectiveness, hence representation-driven retrieval. The current ranking approach is based on simulating individual steps of the random walk, but ideally this would be based on a Markov process over a matrix or tensor representation of the general mixed hypergraph that is the foundation for our model. We could then take advantage of the GPU for improving efficiency. Nevertheless, there are several challenges in this front. Only recently has CERN been studying algebraic approaches for representing general hypergraphs, using adjacency tensors [115, 201].

[REF6] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: It was particularly interesting to find that both synonyms and contextually similar terms were complementary. Moreover, the order by which we added synonyms and context to the index makes a difference. For the Syns+Cont. version, synonyms are added over the term vocabulary of the original collection, and then contextually similar terms are added for both the original terms and their synonyms. In a similar way, the opposite also happens for the Cont.+Syns version. We created an index for both combinations. However, in this particular case, since the word2vec model was trained based on the same INEX 2009 52T-NL subset, they are in fact equivalent — i.e., even after adding new terms from synonyms, these won’t be present in the word2vec model, and, conversely, after adding context, no new terms will be added, since the word2vec model was trained on the same collection, and therefore no new terms from synonyms will be added.

[REF7] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed

Title: Graph-based entity-oriented search

Chunk of text: And d0 k-hop elements, for an ontology with diameter d0 . website1 , a repository of biomedical ontologies, they found that, instead of teleporting to random ontology nodes, users showed a bias toward jumping to nodes at a particular distance k. They called this a k-hop, naming the probabilities of teleporting to k-hops as HopPortation. Given the diameter d 0 of the ontology (ignoring direction), consider the HopPortation vector ~d of size d 0 + 1, where dk ∈ ~d represents the probability of a k-hop happening. The authors computed dk based on the clickstream transitions in the BioPortal website, using add-one smoothing to ensure each available k-hop was considered. Also consider d 0 matrices Mk containing the transition probabilities for the corresponding k-hops, based on the undirected ontology links.

[REF8] - paperID: 72eee80d08f619cf845390a5ca484903b65e522c

Title: Unified Medical Language System resources improve sieve-based generation and Bidirectional Encoder Representations from Transformers (BERT)–based ranking for concept normalization

Chunk of text: 2. BERT-based ranker(f): the listwise classifier, which always requires a candidate generator. During training, we experiment with Lucene(e) or Lucene(c þ d þ e) to generate training instances for the BERT-based classifier, with these candidate generators being run over the training set, and any mentions that have multiple matched candidate concepts becoming training instances. During prediction, we experiment with Lucene(e), Lucene(c þ d þe) and Lucene(a þ b þ c þ d þ e) to generate candidates. For all experiments, we use BioBERT-base,38 which further pre-trains BERT on PubMed abstracts (PubMed) and PubMed Central full-text articles. In our preliminary experiments, we also explored the Clinical-BERT,39 but this resulted in worse performance than BioBERT.

[REF9] - paperID: 06227bc74bcee55471fb37bde0149b317f8a2014

Title: Enhancing Semantic Code Search With Deep Graph Matching

Chunk of text: • Context representation: To compute the context representation, the computed attention in Equation (4) is used in Equation (5). Equation (5) calculates the context representation (e i G) by computing weighted mean of cosine attention (αi,j ) with embedding nodes of code graph (Ge ) denoted by ej . The whole process achieves the common or unified representation that exists between the code graph (Ge) and query graph (Gq). e i G = 1 N X N j αi,j ej (5) • Comparison: Motivated by previous studies on textual implication in NLP .

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