Text Analysis and Retrieval – Project Topics

UNIZG FER, Academic Year 2019/2020

Announced: 12 March 2020
Bidding deadline: 19 March 2020 at 23:59 CET

This document contains the descriptions of project topics offered to the students enrolled in the Text Analysis and Retrieval course in the Academic Year 2019/2020. Each project is to be carried out in groups of three students. Each group is allowed to bid for three topics and rank them by preference (the most preferred topic ranked first). We'll do our best to assign the projects to groups based on their preferences, subject to the constraint that we assign each topic to at most three groups. Receiving time of the bid will not affect the assignment, as long as the bid is received before the deadline.

1 Keyphrase Extraction and Classification

The number of scientific publications grows rapidly each day, which makes it hard to keep track of all the research being done. What is more, it is not that easy to confirm whether someone has addressed a specific task, studied some processes, or utilized certain materials, as currently available publication search engines are rather limited. The goal of this task is to build a system that can automatically identify all the keyphrases from the scientific publication (subtask A) and label them as PROCESS, TASK, or MATERIAL (subtask B).

Competition website:
https://scienceie.github.io/
Dataset:
https://scienceie.github.io/resources.html

Entry points

- Hasan, Kazi Saidul, and Vincent Ng. Automatic Keyphrase Extraction: A Survey of the State of the Art.
- Bhaskar, Pinaki et al. Keyphrase Extraction in Scientific Articles: A Supervised Approach.

2 Detection of Implicit Polarity of Events

As events play a key role in the understanding of text, there's obviously been a large body of research on event analysis in text. However, not much attention was paid to detecting the sentiment polarity of events. The goal of this task is to develop a system that is capable of recognizing the sentiment polarity of an event mentioned in a given sentence. Note that this is an interesting, yet challenging task, as polarity may not be expressed

directly using obvious polarity words (e.g., *good*, *hideous*), but may be implicit (e.g., "Last night I finally completed Dark Souls.").

Competition website:
http://alt.qcri.org/semeval2015/task9/
Dataset:
http://alt.qcri.org/semeval2015/task9/index.php?id=data-and-tools

Entry points

- Thelwall, Mike, Kevan Buckley, and Georgios Paltoglou. Sentiment in Twitter events.
- Nakov, Preslav, et al. SemEval-2016 task 4: Sentiment analysis in Twitter.

3 Early Depression Detection from Language Use

Posts of a person on social media often convey considerable information on their psychological state. This is useful for psychiatrists, since a full history of posts can contain valuable insights to help better assess the patients. Unfortunately, the amount of text to be analyzed makes it impractical to do this kind of analysis manually. The aim of this task is automatically determine whether a person is depressed based on his or her use of language in the social media posts.

Competition website:
http://tec.citius.usc.es/ir/code/dc.html
Dataset:
http://tec.citius.usc.es/ir/code/dc.html

Entry points

- Losada, David E., and Fabio Crestani. A Test Collection for Research on Depression and Language Use.
- Benton, Adrian, Margaret Mitchell, and Dirk Hovy. Multitask Learning for Mental Health Conditions with Limited Social Media Data.

4 Emoji Prediction

The task of this project is to make a system that would automatically fill the text with the appropriate emoticons. This can be done in two steps. First, for each position within the text a prediction is made whether an emoticon should be placed there. Second, an appropriate emoticon is chosen from a list of available emoticons. Both these tasks can be set up as supervised classification problems.

Competition website:
https://competitions.codalab.org/competitions/17344
Dataset:
https://competitions.codalab.org/competitions/17344

Entry points

- Barbieri, Francesco, Miguel Ballesteros, and Horacio Saggion. Are Emojis Predictable?

5 Character Identification in Multiparty Dialogues

Your task is a combination of coreference resolution and entity linking, which are both crucial for successful text understanding. You are given the textual scenarios of episodes from the popular TV show "Friends". The task is to resolve each mention of a speaker (possibly not part of the current conversation) to an entry in a knowledge-base of characters from the show. For example, if Chandler were to say "I gave it to my wife," then your system must map "my wife" to Monica in the knowledge-base.

Competition website:
https://competitions.codalab.org/competitions/17310
Dataset:
https://competitions.codalab.org/competitions/17310

Entry points

- Lee, Heeyoung, et al. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task.
- Shen, Wei, Jianyong Wang, and Jiawei Han. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions.

6 SMS Spam Detection

Given a SMS message, your task is to determine whether the message is a spam message or legitimate communication between people ("ham"). The corpus is crowdsourced from free research resources and contains 5,574 English SMS messages along with class labels. The domain of short text communication makes it difficult to use standard English NLP tools.

Dataset:
http://www.dt.fee.unicamp.br/%7Etiago/smsspamcollection/

Entry points

- Almeida, Tiago A., José María G. Hidalgo, and Akebo Yamakami. Contributions to the study of SMS spam filtering: new collection and results.
- Hidalgo, José María Gómez, Tiago A. Almeida, and Akebo Yamakami. On the validity of a new SMS spam collection.

7 Relation Extraction and Classification on Scientific Texts

Applying natural language processing techniques to scientific literature is an emerging trend. Due to the extremely high output of the scientific community, experts are overwhelmed by the amount of information being produced daily. This makes it difficult to keep

track with the state of the art in a given domain. Your task is to alleviate this problem by making a relation extraction and classification system. This can be viewed as two classification tasks. First, given two entities in text, the model must predict whether there exists a relation between them (binary classification). Second, given the information that two entities are in a relation, the model must predict the type of relation (multiclass classification). Alternatively, the two tasks can be tackled jointly, using joint learning or joint inference.

Competition website:
https://competitions.codalab.org/competitions/17422
Dataset:
https://competitions.codalab.org/competitions/17422

Entry points

- Dhyani, Dushyanta. OhioState at SemEval-2018 Task 7: Exploiting Data Augmentation for Relation Classification in Scientific Papers using Piecewise Convolutional Neural Networks.

8 Semantic Extraction from Cybersecurity Reports

With the widespread use of the internet, the danger of cyber-threats has also increased. A large repository of malware-related texts is available online, which contains detailed malware reports by various cybersecurity agencies or blog posts. Such texts are often used by cybersecurity researchers in the process of data collection. However, the volume and diversity of these texts make it very difficult for researchers to isolate useful information. There are four subtasks that you can tackle, ranging from simply classifying sentences as relevant or not relevant for malware to labeling tokens/relations/attributes with useful information.

Competition website:
https://competitions.codalab.org/competitions/17422
Dataset:
https://competitions.codalab.org/competitions/17422

Entry points

- Lim, Swee Kiat, et al. MalwareTextDB: A Database for Annotated Malware Articles

9 Bot Detection and Gender Profiling

Bots that pose as humans on social media can have considerable influence on users. The nature of such influence can be commercial (e.g., giving artificial positive reviews to a product), political (e.g., undermining the reputation of a political rival through fake negative user comments), or ideological (e.g., shifting public opinion towards an idea such as Brexit). It is also very common for bots to be related to spreading fake news. Consequently, developing methodology for identifying bots is very important. In this project your primary task is to build a system that, given a set of tweets of a Twitter user, determines

whether the user is a bot. A secondary task that you may also tackle using this data is to identify the gender of users that are not bots.

Dataset:
https://pan.webis.de/clef19/pan19-web/author-profiling.html

Entry points

- Andre Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on Facebook.
- Kai Shu, Suhang Wang, and Huan Liu. Understanding user profiles on social media for fake news detection.
- Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots sustain and inflate striking opposition in online social systems.

10 Celebrity Profiling

Public personas celebrities are most often heavy users of social media, which makes them ideal subjects for the study of author profiling. Namely, the question is how much can be learned about a person (in this case a celebrity) by having access only to texts they wrote. The task in this project is to predict traits of a famous person based solely on their social media communication represented as a limited set of their past tweets. Traits considered are degree of fame, occupation, age, and gender.

Dataset:
https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html

Entry points

- Matti Wiegmann, Benno Stein, Martin Potthast. The Celebrity Profiling Corpus.

11 Detecting Emotions from Text in Context

Automatically detecting emotions from text is a very difficult problem. The main reason for this are that computational models don't have access to the vast general knowledge and experience that humans take for granted, as well as the fact that text data lacks many non-verbal queues (such as facial expression) but also verbal cues that are lost in text (such as speech intonation). In this project your task is to tackle this problem. You must build a system that, given a user utterance (as text) and two additional utterances of context, can classify the emotion of the user utterance as *Happy*, *Sad*, *Angry*, or *Others*.

Dataset:
https://www.humanizing-ai.com/emocontext.html

Entry points

- Kashfia SailunazEmail authorManmeet DhaliwalJon RokneReda Alhajj. Emotion detection from text and speech: a survey

12 Unsupervised Lexical Semantic Change Detection

This task adresses unsupervised detection of lexical semantic change, i.e., word sense changes over time, in text corpora. There are two tasks: a classification task, and a ranking task. For each language, both tasks are based on the same two corpora, which span different periods. Corpora will be provided in four languages: German, English, Swedish, and Latin. Systems will be evaluated against a ground truth, as annotated by human native speakers (except for Latin, which was annotated by scholars of Latin).

Competition website:
https://competitions.codalab.org/competitions/20948

Entry points

- Kutuzov, A., Øvrelid, L., Szymanski, T. and Velldal, E., 2018. Diachronic word embeddings and semantic shifts: a survey. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Schlechtweg, D., Schulte im Walde, S. and Eckmann, S., 2018. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 169–174, New Orleans, LA, USA.
- Tahmasebi, N., Borin, L. and Jatowt, A., 2018. Survey of Computational Approaches to Lexical Semantic Change. arXiv preprint arXiv:1811.06278.

13 Predicting Multilingual and Cross-Lingual (Graded) Lexical Entailment

This shared task is about predicting binary and graded Lexical Entailment (i.e., is-a or hyponym-hypernym relation) for several different languages (multilingual component) and across languages for several language pairs (cross-lingual component). For Graded LE, the participants need to predict the degree (on a 0-6 scale) to which the LE relation holds between two given concepts. The two concepts in each pair come from the same language (e.g., medvjed (bear) is-a sisavac (mammal)) in multilingual subtasks and from different languages (e.g. medvjed (bear) is-a mammal) in the cross-lingual subtasks. For Binary LE the participants merely need to predict whether the LE relation holds between two concepts or not.

Competition website:
https://competitions.codalab.org/competitions/20865

Entry points

- Vulić, I., Ponzetto, S. P., and Glavaš, G. (2019, July). Multilingual and cross-lingual graded lexical entailment. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4963-4974).

14 Commonsense Validation and Explanation (ComVE)

The task is to directly test whether a system can differentiate natural language statements that make sense from those that do not make sense. There exist three subtasks. The first task is to choose from two natural language statements with similar wordings which one makes sense and which one does not make sense. The second task is to find the key reason from three options why a given statement does not make sense. The third task asks machine to generate the reasons and we use BLEU to evaluate them. Examples of all tasks can be found on the competition website.

Competition website:
https://competitions.codalab.org/competitions/21080

Entry points

- Wang, C., Liang, S., Zhang, Y., Li, X., and Gao, T. (2019). Does it make sense? and why? a pilot study for sense making and explanation. arXiv preprint arXiv:1906.00363.

15 Modelling Causal Reasoning in Language: Detecting Counterfactuals

The task has two substasks. First, you are asked to determine whether a given statement is counterfactual or not. Counterfactual statements describe events that did not actually happen or cannot happen, as well as the possible consequence if the events had happened (e.g., "if they had been more careful they would not have gotten the virus"). More specifically, counterfactuals describe events counter to facts and hence naturally involve common sense, knowledge, and reasoning. The second task involves a further step of detecting the knowledge conveyed by these statements by locating the antecedent and consequent.

Competition website:
https://competitions.codalab.org/competitions/21691

Entry points

- Qin, L., Bosselut, A., Holtzman, A., Bhagavatula, C., Clark, E., and Choi, Y. (2019). Counterfactual Story Reasoning and Generation. arXiv preprint arXiv:1909.04076.

16 Assessing Humor in Edited News Headlines

Nearly all existing humor datasets are annotated to study whether a chunk of text is funny. However, it is interesting to study how short edits applied to a text can turn it from non-funny to funny. Such a dataset helps us focus on the humorous effects of atomic changes and the tipping point between regular and humorous text. The goal of our task is to determine how machines can understand humor generated by such short edits. Your model will estimate the funniness of news headlines that have been modified by humans using a micro-edit to make them funny. The problem can be framed as either regression or classification, leading to two substasks.

Competition website:
https://competitions.codalab.org/competitions/20970

Entry points

- Nabil Hossain, John Krumm and Michael Gamon. "President Vows to Cut Taxes Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. 2019. In NAACL.

17 SentiMix Hindi-English

Mixing languages, also known as code- mixing, is a norm in multilingual societies. Multilingual people, who are non-native English speakers, tend to code-mix using English-based phonetic typing and the insertion of anglicisms in their main language. In addition to mixing languages at the sentence level, it is fairly common to find the code-mixing behavior at the word level. This poses challenges to NLP sytems. The objective of this task is to tackle the task of sentiment analysis in code-mixed social media text. Specifically, on the combination of English with Spanish (Spanglish) and Hindi (Hinglish), which are the 3rd and 4th most spoken languages in the world respectively.

Competition website:
https://competitions.codalab.org/competitions/20654

Entry points

- Barman, U., Das, A., Wagner, J., and Foster, J. (2014, October). Code mixing: A challenge for language identification in the language of social media. In Proceedings of the first workshop on computational approaches to code switching (pp. 13-23).
- Pratapa, A., Choudhury, M., and Sitaram, S. (2018). Word embeddings for code-mixed language processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3067-3072).

18 OffensEval 2020

Offensive language is pervasive in social media. Individuals frequently take advantage of the perceived anonymity of computer-mediated communication, using this to engage in behavior that many of them would not consider in real life. Online communities, social media platforms, and technology companies have been investing heavily in ways to cope with offensive language to prevent abusive behavior in social media. One of the most effective strategies for tackling this problem is to use computational methods to identify offense, aggression, and hate speech in user-generated content (e.g. posts, comments, microblogs, etc.). There are three substasks to choose from 1) detecting offensive language, 2) categorizing offense types, and 3) identifying the target of the offense.

Competition website:
https://sites.google.com/site/offensevalsharedtask/

Entry points

- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R. (2019) SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 75-86.

- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R. (2019) Predicting the Type and Target of Offensive Posts in Social Media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1415-1420.

19 Detection of Propaganda Techniques in News Articles

Propagandistic news articles use specific techniques to convey their message, such as whataboutism, red Herring, and name calling, among many others. The focus of this task is to develop automatic tools to detect such techniques! There are two substasks. First, given a text document your model must find fragments which contain at least one propaganda technique (a sequence labeling task). Second, given a fragment identified as propaganda and its document context, your model must identify the propaganda technique that was applied (multilabel text classification, as more than one technique could be applied in the same fragment). The data has been annotated with 18 different propaganda techniques!

Competition website:
`https://propaganda.qcri.org/semeval2020-task11/`

Entry points

- G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, "Fine-Grained Analysis of Propaganda in News Articles", in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), Hong Kong, China, November 3-7, 2019.

- A. Barrón-Cedeño, I. Jaradat, G. Da San Martino, P. Nakov, "Proppy: Organizing News Coverage on the Basis of Their Propagandistic Content", Information Processing and Management, 2019. DOI:10.1016/j.ipm.2019.03.005.

- A. Barrón-Cedeño, G. Da San Martino, I. Jaradat, and P. Nakov, "Proppy: A System to Unmask Propaganda in Online News", in Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI'19), Honolulu, HI, USA, 2019.

20 Mystery Topic

You might feel none of the above topics spark your interests in just the right way. Perhaps you feel you need something different, to stray off the beaten path, to explore what adventures or truths lay beyond the veil of what is currently popular. If this is the case

for you, we offer the opportunity to figure out your own topic. If we deem the topic to be sensible, you may pursue it for your project assignment.

Dataset:
https://www.google.com

Entry points

- Workshop on semantic evaluation – SEMEVAL
- PAN events.
- TREC evaluation campaigns.

  Note:

  – For each of the above links, be sure to check out editions from other years as well.
  – If you choose to go down this path, let us know your topic suggestion ASAP (by mail to mladen.karan@fer.hr), so we can give you feedback.