Text Analysis and Retrieval – Project Topics

UNIZG FER, Academic Year 2017/2018

Announced: 16 March 2018
Bidding deadline: 21 March 2018 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 2017/2018. Each project is to be carried out in groups of two or 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.

1 Intrinsic Plagiarism Detection

Intrinsic plagiarism detection (also called author diarization) attempts to detect plagiarized text fragments within a single document (in contrast with standard plagiarism detection across documents). Given a document, the task is to identify and group text fragments that were written by different authors. To make things a bit harder, the author change may occur at any position in the text, and not only at sentence or paragraph boundaries. The task is split into three subtasks of increasing difficulty, based on the number of (known) authors. First subtasks assumes two, second a fixed number, and third an unknown number of authors.

Competition website:
http://pan.webis.de/clef16/pan16-web/author-identification.html
Dataset:
http://pan.webis.de/clef16/pan16-web/author-identification.html

Entry points

- Paolo Rosso, Francisco Rangel, Martin Potthast, Efstathios Stamatatos, Michael Tschuggnall, and Benno Stein. Overview of PAN'16 – New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation.
- Mikhail Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, and Vadim Strijov. Methods for Intrinsic Plagiarism Detection and Author Diarization.
- Abdul Sittar, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. Author Diarization Using Cluster-Distance Approach.

2 Community Question Answering

Community question answering (CQA) Sites like StackOverflow, where people answer the community questions, but can also post their own questions, are rising in popularity.

Unfortunately, as such sites are quite leniently moderated (and therefore having content of varying quality), it is not always that easy to find the right answer for a given problem. The goal of this task is to implement a system that is capable of: (1) ranking the available answers by their relevance according to a given question, and (2) ranking the available questions by their similarity with a given question (subtasks A and B).

Competition website:
http://alt.qcri.org/semeval2017/task3/
Dataset:
http://alt.qcri.org/semeval2017/task3/index.php?id=data-and-tools

Entry points

- SemEval-2015 Task 3: Answer Selection in Community Question Answering.
- Tran, Quan Hung, et al. JAIST: Combining multiple features for answer selection in community question answering.

3 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.

4 Lexical Substitution

Lexical substitution is a task of finding meaning-preserving substitutes for a polysemous (multiple-sense) target word within its context. Given a word and its context, your system should be able to provide an appropriate set of substitutes, which will then be compared to the ones provided by the human annotators. You are free to assume you'll be dealing solely with single-word substitutes, even though the task proposes both single-word and multiword substitutes.

Competition website:
http://nlp.cs.swarthmore.edu/semeval/tasks/task10/summary.shtml
Dataset:
http://www.dianamccarthy.co.uk/task10index.html

Entry points

- McCarthy, Diana, and Roberto Navigli. SemEval-2007 Task 10: English lexical substitution task.
- Hassan, Samer, et al. UNT: Subfinder: Combining knowledge sources for automatic lexical substitution.
- Giuliano, Claudio, Alfio Gliozzo, and Carlo Strapparava. Fbk-irst: Lexical substitution task exploiting domain and syntagmatic coherence.
- Melamud, Oren, et al. A simple word embedding model for lexical substitution.

5 Extraction of Drug-Drug Interactions

The drug-drug interaction (DDI) describes changes that occur when two drugs are used at the same time. Most of these interactions are thoroughly explained in medical journals, and therefore it makes sense to develop an information extraction system that is able to automatically identify and extract relevant information on DDIs. Note that this information would greatly benefit the pharmaceutical industry. Your goal is to develop a system that can: (1) recognize and classify drug names from running text (DRUG, BRAND, GROUP, NO-HUMAN), and, given gold labels from the first subtask, (2) extract drug-drug interactions that occur (ADVICE, EFFECT, MECHANISM, INT).

Competition website:
https://www.cs.york.ac.uk/semeval-2013/task9/
Dataset:
https://www.cs.york.ac.uk/semeval-2013/task9/index.php%3Fid=data.html

Entry points

- Bedmar, Segura et al. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions From Biomedical Texts.

6 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.

7 Affect Detection in Tweets

The task is to build an emotion and sentiment analysis system for microblogs (tweets). In addition to identifying the presence of a particular emotion or sentiment, the system must also detect its *intensity*. The problem consists of five subtasks involving regression or ordinal regression applied to sentiment or emotion (a total of four tasks), and multiclass classfication of text into one of twelve classes (no-emotion or one of 11 possible emotion classes). You may choose a single task and explore several models, or solve more tasks with less experiments for each. In both cases a fair experimental evaluation of your system is mandatory.

Competition website:
https://competitions.codalab.org/competitions/17751
Dataset:
https://competitions.codalab.org/competitions/17751

Entry points

- Jabreel, Mohammed, and Antonio Moreno. EiTAKA at SemEval-2018 Task 1: An Ensemble of N-Channels ConvNet and XGboost Regressors for Emotion Analysis of Tweets.

8 Toxic Language Detection on Tweets

With the ever-increasing amount of (often anonimous) communication carried out over the internet, toxic comments are becoming a serious problem. Consequently, the task of automatic identification of such comments is becoming increasingly important. Your task is to build a model for detecting toxic comments on Twitter based on machine learning. You must experiment with several different models and provide a fair comparison of their performance.

Competition website:
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
Dataset:
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

Entry points

- Sax, Sasha. Flame Wars: Automatic Insult Detection.
- Kennedy, George, et al. Technology Solutions to Combat Online Harassment.

9 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.

10 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?

11 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.

12 Hate Speech Identification

Given a short text, the task is to classify the text into one of the three following classes: (1) the text contains hate speech, (2) the text is offensive, but does not contain hate speech, and (3) the text is not offensive at all. The dataset consists of 25000 tweets which may contain offensive language. The language used on Twitter differs from the standard English, which makes the problem more difficult.

Dataset:
https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data

Entry points

- Davidson, T., Warmsley, D., Macy, M., and Weber, I. Automated hate speech detection and the problem of offensive language.
- Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. Understanding Abuse: A Typology of Abusive Language Detection Subtasks.

13 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.

14 Disasters on Social Media

Given a collection of tweets containing words such as *ablaze*, *quarantine*, and *pandemonium*, your task is to determine whether the tweet refers to (1) an actual disaster effect or (2) the catastrophic meaning is used in a non-literal manner, such as part of a movie review.

Dataset:
https://www.crowdflower.com/wp-content/uploads/2016/03/socialmedia-disaster-tweets-DFE.csv

Entry points

- Verma, Sudha, et al. Natural Language Processing to the Rescue? Extracting Situational Awareness Tweets During Mass Emergency.

15 U.S. Economic Performance Based on News Articles

Given a news article headline, a target sentence from a news article and two context sentences, your task is to determine if the target sentences provides an indication towards the health of the U.S. economy, and rate the indication on a scale from 1–9, where 9 signifies most positive and 1 signifies negative.

Dataset:
https://www.crowdflower.com/wp-content/uploads/2016/03/us-economic-newspaper.csv

Entry points

- Levenberg, Abby, et al. Predicting Economic Indicators from Web Text Using Sentiment Composition.

16 Relation Extraction and Classification on Scientific Texts

Applying natural language processing technologies 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 wit 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 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.

17 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

18 Personalized Medicine using NLP

Recent advances in cancer treatement rely on gene sequencing of the cancer cells. Once sequenced, a cancer tumor can have thousands of genetic mutations. The challenge lies in determining which of these mutations contribute to tumor growth ("drivers") and which are neutral ("passengers"). This information can lead to effective personalized treatment. Currently, this job is performed by a clinical pathologist, who classifies each mutation based on reviewing available text-based clinical literature. The aim of this task is to develop a machine learning based system that will automatically classify a mutation as a driver or passenger (binary text classification), making this process much faster.

Competition website:
`https://www.kaggle.com/c/msk-redefining-cancer-treatment`
Dataset:
`https://www.kaggle.com/c/msk-redefining-cancer-treatment`

Entry points

- Pestian, John P., et al. A Shared Task Involving Multi-label Classification of Clinical Free Text.

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval.

19 Analysis of Exaggerated Claims in Science News

The discrepancy between science and media has been affecting the effectiveness of science communication. Original findings from science publications may be distorted with altered claim strength when reported to the public, causing misinformation spread. Construction of prediction models for identifying claim strength levels in science reporting is one solution to avoid the problems followed by such exaggeration. Your task is to build such a prediction model.

Dataset:
https://figshare.com/articles/InSciOut/903704

Entry points

- Li, Yingya, Jieke Zhang, and Bei Yu. An NLP Analysis of Exaggerated Claims in Science News.

20 Patent Retrieval

With the ever-increasing number of filed patent applications every year, the need for effective and efficient systems for managing such tremendous amounts of data becomes inevitably important. Patent Retrieval (PR) is considered the pillar of almost all patent analysis tasks. Your task is to build an information retrieval system that will, given input claims, retrieve patents relevant for these claims and identify the most relevant paragraphs within these documents.

Competition website:
www.ifs.tuwien.ac.at/~clef-ip/2012/claims-to-passage.shtml
Dataset:
http://www.ifs.tuwien.ac.at/~clef-ip/download/2012/index.shtml

Entry points

- Gobeill, Julien, and Patrick Ruch. BiTeM Site Report for the Claims to Passage Task in CLEF-IP 2012.

21 Metaphor Detection

Metaphor is a linguistic phenomenon which implies reasoning about one thing in terms of another – it is a type of conceptual mapping, where words or phrases applied to objects and actions do not permit a literal interpretation. Another aspect of metaphors is that they violate selection restrictions of verbs. This means that metaphors allow verbs to combine with different syntactic arguments, accepting different semantic concepts in these positions. Given a text fragment, the goal is to detect, at a word level, all content-word metaphors.

Dataset:
https://github.com/EducationalTestingService/metaphor/tree/master/NAACL-FLP-shared-task

Entry points

- Schulder, Marc, and Eduard Hovy. Metaphor Detection through Term Relevance.
- Tsvetkov, Yulia, et al. Metaphor Detection with Cross-lingual Model Transfer.