Text Analysis and Retrieval – Project Topics

UNIZG FER, Academic Year 2016/2017

Announced: 10 March 2017
Bidding deadline: 15 March 2017 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 2016/2017. 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 try to assign the projects to groups based on your preferences, but we may also have to take into account the time when the bid was received on a first-come-first-served basis. We'll assign each topic to at most three groups.

1 Author Clustering

Given a document collection, the task is to group documents into clusters that represent different authors. That is, documents that are written by the same author should be clustered into the same group. The task comprises two subtask: (1) a complete author clustering and (2) authorship-link ranking. The former is concerned with the detection of k different authors and their documents, whereas the latter one asks to predict the authorship similarity between the pairs of documents. Even though the data for this task is in English, Dutch, and Greek, you are only required to tackle this task for English (but you may as well do it for other languages if you want).

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.
- Douglas Bagnall. Authorship Clustering Using Multi-headed Recurrent Neural Networks.
- Yunita Sari and Mark Stevenson. Exploring Word Embeddings and Character N-Grams for Author Clustering.

2 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.

3 Semantic Textual Similarity

Semantic textual similarity (STS) measures how similar two text snippets are. The task is, given two sentences, to determine their similarity score on a continuous scale from 0 to 5. The score of 5 denotes a perfect semantic equivalence. This task was offered in multiple tracks, covering various cross-lingual and monolingual pairs. Your task, however, is the track concerned with English monolingual pairs. Of course, you are encouraged to give a shot on other tracks as well.

Competition website:
http://alt.qcri.org/semeval2017/task1/
Dataset:
http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools

Entry points

- Agirre, Eneko, et al. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation.
- Brychcın, Tomáš, and Lukáš Svoboda. Semantic textual similarity using lexical, syntactic, and semantic information.
- Šarić, Frane, et al. Takelab: Systems for measuring semantic text similarity.

4 Sentiment Analysis in Twitter

The goal of this task is to build a sentiment analysis model capable of determining whether a given tweet is of positive, negative, or neutral sentiment. Besides a standard task, where your model is given a tweet and needs to determine its sentiment, you also need to address the subtasks that ask for topic-specific sentiment. Specifically, you are given a tweet and a topic, and your system should be able to determine the sentiment expressed towards that topic (on a 2-point and a 5-point scale).

Competition website:
http://alt.qcri.org/semeval2017/task4/SemEval-2017 Task 4: Sentiment Analysis in Twitter
Dataset:
http://alt.qcri.org/semeval2017/task4/index.php?id=download-the-full-training-data-for-semeval-2017-task-4

Entry points

- Nakov, Preslav, et al. SemEval-2016 task 4: Sentiment analysis in Twitter.
- Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. Twitter sentiment analysis: The good the bad and the OMG!.

5 Author Profiling

It is interesting how people of different gender, age, or a personality type express themselves in writing – some might use more adjectives, some might be more direct and use less of active voice, while other might use extremely dated vernacular. In this task, you need to implement a model that is able to predict the age (18–24, 25–34, 50–64, or 65+) and gender (male or female) of the author, given text. Note that you are only required to tackle this task for English.

Competition website:
http://pan.webis.de/clef16/pan16-web/author-profiling.html
Dataset:
https://www.dropbox.com/s/x7zsudnnthccokq/pan16-author-profiling-training-dataset-english-2016-04-25.zip?dl=0

Entry points

- Rosso, Paolo et al. Overview of PAN'16 – New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation.
- Agrawal, Madhulika and Gonçalves, Teresa. Age and Gender Identification using Stacking for Classification.
- Ashraf, Shaina et al. Cross-Genre Author Profile Prediction Using Stylometry-Based Approach.

6 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.

7 Humor Detection and Ranking

This task is inspired by the Comedy Central's show @midnight, in which guests and audience propose funny tweets in a response to a topic (described by a hashtag). The machine learning setup is the same: given a hashtag and a list of tweets posted under that hashtag, rank the tweets by their humorousness. The task officially comprises two subtasks: (1) detecting which tweet of two is more humorous and (2) ranking the tweets by how humorous they are (three-partite ranking). These subtasks can obviously be tackled separately, but it is also possible to use the outputs of one to address the other one.

Competition website:
http://alt.qcri.org/semeval2017/task6/
Dataset:
http://alt.qcri.org/semeval2017/task6/index.php?id=data-and-tools

Entry points

- Potash et al. #HashtagWars: Learning a Sense of Humor.

8 Complex Word Identification

For non-native speakers, people with low literacy or intellectual disabilities, and language-impaired people (e.g., autistic, aphasic, congenitally deaf), regular texts (e.g., news) are often difficult to comprehend. Stylistically decorated sentences utilizing sophisticated vocabulary of infrequent words pose a particular difficulty for these groups of people. The goal of this project is to build a system that can identify which words in a given sentence can challenge such readers.

Competition website:
http://alt.qcri.org/semeval2016/task11/
Dataset:
http://alt.qcri.org/semeval2016/task11/index.php?id=participate

Entry points

- Paetzold, Gustavo. Reliable Lexical Simplification for Non-Native Speakers.
- Glavaš, Goran. Simplifying Lexical Simplification: Do We Need Simplified Corpora?
- Specia, Lucia et al. SemEval-2012 Task 1: English Lexical Simplification.

9 Sexual Predator Identification

Given chat logs involving two (or more) people, your task is to determine who is the one trying to convince the other(s) to provide some sexual favor. You also need to identify the conversation parts where the predator exhibits their bad behavior. This amounts to the implementation of two supervised classifiers: one for the identification of the predators (among all the users), and the other one for the detection of chat log lines that are the most distinctive of the predator bad behavior.
NB: The chat logs may be offensive or disturbing to some people. Please refrain from bidding for this topic if that bothers you.

Competition website:
http://pan.webis.de/clef12/pan12-web/author-identification.html
Dataset:
http://pan.webis.de/clef12/pan12-web/author-identification.html

Entry points

- Inches, Giacomo and Crestani, Fabio. Overview of the International Sexual Predator Identification Competition
- Villatoro-Tello, Esaú et al. Two-step Approach for Effective Detection of Misbehaving Users in Chats
- Parapar, Javier et al. A Learning-based Approach for the Identification of Sexual Predators in Chat Logs

10 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.

11 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.

12 Diachronic Text Evaluation

It's interesting to see how languages change over time. The goal of this task is to build a system that is able to automatically determine the time period when a piece of news was written (the so-called text epoch disambiguation), covering the time interval from 1700 to 2014. This is to be framed as a classification task, where a model should predict the correct time interval (out of all given ones). The task comprises three subtasks of increasing difficulty: the first one deals with texts with clear reference to time anchors, the second one with texts with specific time language use, and the last one with the recognition of time-specific phrases.

Competition website:
http://alt.qcri.org/semeval2015/task7/
Dataset:
http://alt.qcri.org/semeval2015/task7/index.php?id=data-and-tools

Entry points

- Popescu, Octavian and Strapparava, Carlo. SemEval-2015 Task 7: Diachronic Text Evaluation.
- Mihalcea, Rada and Nastase, Vivi. Word Epoch Disambiguation: Finding How Words Change Over Time.
- Popescu, Octavian, and Carlo Strapparava. Behind the Times: Detecting Epoch Changes using Large Corpora.

13 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.

14 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.

15 Factoid Question Answering

The task is to build a question answering system for short, factoid questions, e.g., "Who was the first man to walk in space?". Your system has to find the answer that is buried in a set of (single-sentence) documents, derived from the TREC-9 QA test collection. The solution has two steps. The first step is answer retrieval: retrieving and ranking the sentences such that those containing the answer are ranked at the top. The second step is answer extraction: singling out the phrase that constitutes the precise answer to the question. For the latter, you are to consider two setups: (1) answer extraction on sentences that we know contain the answer (an idealized setup) and (2) answer extraction on sentences retrieved as relevant in the first step (a realistic setup). You are free to use both unsupervised or supervised methods, as well as decide on which of the subtasks to put more focus. For at least one of the steps your system must outperform a sensible baseline. Answer retrieval shall be evaluated using mean reciprocal rank (MRR), while answer extraction shall be evaluated by means of token-based F1-measure averaged over all questions.

Competition website:
http://trec.nist.gov/data/qa.html
Dataset:
http://trec.nist.gov/data/qa/t9_qadata.html

Entry points

- Voorhees, Ellen M., and D. M. Tice. Overview of the TREC-9 Question Answering Track.
- Wang, Mengqiu. A survey of answer extraction techniques in factoid question answering.
- Kolomiyets, Oleksandr, and Marie-Francine Moens. A survey on question answering technology from an information retrieval perspective.