Report Topics for Computational Linguistics & NLP
-Liviu P Dinu & Ana Uban-

Topics or projects:

1. Participate in shared tasks and competitions in the field of NLP (Kaggle is not accepted - if you need datasets start here): SemEval, CLEF, PAN, VarDial, any shared tasks associated with top ranking (A and A* according to core) NLP conferences (EMNLP, COLING, ACL, NAACL, EACL, LREC, etc.)
2. Automatic text summarization (abstractive or extractive)
    - Ideas: a news summarizer/generating news headlines (option: for Romanian - more challenging for abstractive), summarizer for scientific articles/generating the abstracts automatically), summarizer for stories/literature; + evaluate it using standard metrics and report results
    - Papers: Abstractive Summarization: A Survey of the State of the Art (2019), Recent automatic text summarization techniques: a survey (2016)
3. POS-tagging (part of speech tagging)
    - Ideas: implement a POS-tagging algorithm from scratch (optional: include graphical visualization) + evaluate and report results
    - Papers: POS Tagging for Arabic Tweets , Non-lexical neural architecture for fine-grained POS Tagging
    - Existing tools: The Stanford POS Tagger
    - Annotated datasets: Penn Treebank from NLTK
4. Named Entity Recognition and other Information Extraction tasks
    - Ideas: implement a NER algorithm from scratch, medical NER
    - Papers: A survey of named entity recognition and classification, Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition, Enhancing clinical concept extraction with contextual embeddings , OntoNERdIE – Mapping and Linking Ontologies to Named Entity Recognition and Information Extraction Resources
    - Applications: Lark needed it for food parsing, Bing needs it for search, processing medical text, populating knowledge bases
    - Existing tools: implemented in spacy, YODIE Named Entity Disambiguation (English)
    - Annotated datasets: juand-r/entity-recognition-datasets , Tweets here
5. Syntax Parsing; Text processing at the syntax level (dependency grammars/dependency parsing; constituent parsing)
    - Ideas: implement a syntax parser from scratch (optional: on Romanian/rare language), create a graphical visualization of parsed sentence
    - Papers: Accurate Unlexicalized Parsing, Proceedings of the... , Parsing as Sequence Labeling
    - Existing tools: Stanford parser
    - Applications: Grammarly, Grammarly Knock-off)

    - Annotated datasets: Penn Treebank in NTLK
6. Corpus/dataset creation (collection, cleaning, annotation, etc; e.g. Twitter/Reddit API, web scraper for news articles/political speeches/meeting transcripts/dictionaries, ...) (look at LREC, benchmark dataset type papers) - either create new kind of corpus OR complement corpus collection with some form of linguistic analysis
    - Ideas: corpus of Romanian product/movie reviews, annotate with sentiment; corpus of news, annotate with emotions expressed; corpus of non-English Tweets, annotate with optimism/pessimism/mental health (”I am diagnosed with depression”);
7. Text simplification
    - Ideas: implement a text simplification solution + evaluate and report results
    - Applications: Simple Wikipedia, language learning
    - Papers: Exploring neural text simplification methods
    - Annotated datasets: Simple Wikipedia,
8. Textual semantic similarity, text clustering; NLU (natural language inference, entailment)
    - Ideas: implement a textual entailment model + evaluate; train document/sentence embeddings for semantic similarity;
    - Papers: Distributed Representations of Sentences and Documents (doc2vec), Recognizing Textual Entailment in Twitter Using Word Embeddings , Siamese recurrent architectures for learning sentence similarity, Word n-gram attention models for sentence similarity and inference
    - Annotated datasets: SNLI (for entailment);
9. Distributional semantics, word embeddings, contextual embeddings
    - Ideas: implement, compare and evaluate various measures of similarity metrics on embeddings; visualization tool for embedding spaces; train embeddings on new domain and evaluate/discuss - needs large dataset (e.g. embeddings for medical data, embeddings for social media slang); evaluate and compare methods for embeddings compositionality
    - Papers: Distributed Representations of Words and Phrases and their Compositionality, Enriching Word Vectors with Subword Information, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, [1802.04302] Evaluating Compositionality in Sentence Embeddings, How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings
    - Existing tools: Embedding projector - visualization of high-dimensional data, GloVe: Global Vectors for Word Representation (pre-trained GloVe embeddings), https://github.com/facebookresearch/MUSE (multilingual FastText embeddings trained on Wikipedia)
    - Datasets of annotated word similarity: SimLex, WordSim
10. Fake news detection, rumor detection, propaganda detection
    - Examples: implement a fake news detection system + evaluate; implement automatic fact-checker (e.g. like https://www.factual.ro/); implement clickbait detector; detector of fake news in specific domain (political, medical)
    - Applications: pheme, Twitter/Instagram/FB integrated fake news detection

- Papers: "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, We Built a Fake News / Click Bait Filter: What Happened Next Will Blow Your Mind!, SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours, SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, Can We Spot the" Fake News" Before It Was Even Written?,

11. Deception detection, style transfer
    - Ideas: system that automatically detects impersonation attempts in published long texts; generator of text in the style of a given author (challenging); identify impersonators on social media: discrepancies between declared age/sex and real one
    - Papers: Pastiche Detection Based on Stopword Rankings. Exposing Impersonators of a Romanian Writer, [2011.00416] Deep Learning for Text Style Transfer: A Survey
12. Authorship attribution / stylometry
    - Ideas: automatically identify author of literary texts (aim for similar authors) / classify characters in a play; automatically identify authors of social media posts (e.g. authorship attribution/verification on tweets); automatically classify authors of scientific papers; identify authors of song lyrics; + analyze which feature characterize the authors (explainability of the machine learning model)
    - Papers & authors: PAN, Liviu Dinu, Marius Popescu
13. Plagiarism detection
    - Ideas: implement a plagiarism detection system based on a collection of scientific articles (identify duplicate content and source); solve a shared task on plagiarism detection (PAN); plagiarism detection on political speeches
    - Papers & authors: PAN, Liviu Dinu, Marius Popescu
14. Author profiling (detecting the age/gender/personality/native language of an author)
    - Ideas: profile users on social media (automatically predict demographics like age/gender/location);
    - Papers & authors: PAN, Sulea& Dichiu, Sulea&Zampieri, Nisioi
15. Hate speech, offensive language identification, misogyny / stereotype detection
    - Ideas: annotate a corpus of Romanian tweets with hate speech labels; solve a shared task on hate speech/aggressive speech detection;
    - Sexism, Racism, Homophobia https://arxiv.org/pdf/2106.15896.pdf
    - Papers & authors: Marcos Zampieri, Paolo Rosso
16. Computational humor, sarcasm & irony detection
    - Ideas: implement an irony/sarcasm detector on social media, predict the #irony/#sarcasm hashtag; automatically identify sarcasm in movie or product reviews/news; measure correlation between sarcasm and optimism/pessimism/emotions expressed or author personality/profile on social media texts
    - Papers & authors: A multidimensional approach for detecting irony in Twitter , Cristian Danescu Mizil, Carlo Straparava, Paolo Rosso
17. Metaphor and figurative language detection

    - Ideas: automatically identify words used metaphorically in poems/song lyrics/social media; automatically identify meaning of a metaphor ("translate" the metaphor);
    - Papers: Brighter than Gold: Figurative Language in User Generated Comparisons, From humor recognition to irony detection: The figurative language of social media, A Computational Exploration of Exaggeration, Impact Analysis of Emotion in Figurative Language
18. Diachronic and historical linguistics: word formation, cognates identification, proto-word re-construction, borrowing, language similarity, etc
    - Ideas: collect lexicon of words and etymologies in a low resource language (Eastern European languages?); compare different measures of language/dialect similarity based on common vocabulary/similar syntax/similar phonetics;
    - Papers & authors: Liviu Dinu, Alina Maria Ciobanu
19. Semantic change – tracking the change in meanings of words
    - Ideas: identify semantic change in certain subset of words e.g. sentiment/emotion words (optional: for Romanian), in business terminology; compare metaphorical senses of words across languages; identify changes in slang terms and appearance of new senses from social media data (e.g. "lit"); solve shared task on semantic change (SemEval 2020)
    - Papers: Towards Computational Lexical Semantic Change Detection (LChange Workshop - check proceedings); Computational approaches to semantic change book
20. Temporal text classification, dating of texts
    - Ideas: predict period when text was written based on different features; identify which features are most useful for dating: news texts/scientific texts/
    - Papers: Temporal classification for historical Romanian texts
21. Law and NLP-AI
    - Ideas: predict outcome of court cases; automatically parse contracts
    - Papers: [1710.09306] Exploring the Use of Text Classification in the Legal Domain
22. NLP and ethics, biases in datasets and algorithms (explainability/interpretability)
    - Ideas: identify biases against immigrants in news texts using word embeddings (better: contextual embeddings) - biased sentiment/emotion; track changes in biases wrt certain minority over time; evaluate bias in Romanian embeddings (towards minorities in Romania?)
    - Papers: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,
23. Language and text generation; BERT, GPT-3 and other Transformer based language models
    - Ideas: build a language generator for Romanian based on pre-trained transformers (e.g. multilingual BERT); generate (fake?) news/scientific articles/food recipes
    - Examples: https://app.inferkit.com/demo, https://transformer.huggingface.co/, https://6b.eleuther.ai/

- Papers: [1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://aclanthology.org/2020.coling-main.581.pdf
- Resources: https://huggingface.co/transformers/pretrained_models.html, https://github.com/kingoflolz/mesh-transformer-jax/#gpt-j-6b

24. Social Media text processing (fb, ig, twitter, reddit)
    - Ideas: any task applied to social media text: hate speech, word embeddings (identify new terms), authorship, author profiling, opinion mining; tracking emotions expressed on social media (bonus: do it for Romanian - more rare); collect and annotate a corpus of social media data (ideally for a new language/task); meme analysis - multimodal text+visual (papers: Analysis of Facebook Meme Groups Used During the 2016 US Presidential Election, Findings of the WOAH 5 Shared Task on Fine Grained Hateful Memes Detection, Competition Memotion) (pretrained image analysis: VGG…)
25. Deep learning in NLP
    - Ideas: any NN architecture for any NLP task, explain the math
26. Machine learning in NLP (supervised, weakly supervised, zero shot learning, few shot learning; probabilistic models, variational inference)
    - Ideas: non-DL machine learning for an NLP task, compare models
27. Transfer learning – multi-stage machine learning where knowledge from one dataset/domain/task is leveraged to help with another
    - Papers & authors: Sebastian Ruder
28. (Neural) Machine Translation
    - Ideas: implement a machine translation model
    - Papers & tools: any Google Translate paper, Interlingua, WMT
29. Word sense disambiguation
    - Ideas: implement a word sense disambiguation model (bonus: for Romanian/a low-resource language); evaluate contextual embeddings (transformer-based) for disambiguation, compare with other models
    - Papers & authors: Florentina Hristea
30. Morphological re-inflection; Inflection generation
    - Ideas: reproduce models on inflection generation from existing papers on Romanian, apply for new parts of speech / use different architectures
    - Papers & authors: Maria Sulea, SIGMORPHON
31. Collocation detection; multi word expressions, phrase identification
    - Ideas: implement/compare collocation detection models; compare meanings of phrases in different languages; embeddings for collocations/phrases
    - Papers & authors: Mikolov, Katja Lipshikova
32. Anaphora resolution, coreference resolution
    - Ideas: implement model that identifies anaphora; implement a tool to visualize coreference in texts
    - Papers: Global Inference for Bridging Anaphora Resolution - Yufang Hou1, Katja Markert2, Michael Strube1 , An Algorithm for Pronominal Anaphora Resolution ,
33. Sentiment Analysis; optimism-pessimism identification

- Ideas: implement +evaluate an algorithm for sentiment analysis - predict sentiment expressed in tweet / news / review (optional: use Romanian data); optional: identify specifically sentiment for individual aspects of the object (aspect-based sentiment analysis)
- Papers & authors: Exploiting BERT for end-to-end aspect-based sentiment analysis , Rada Mihalcea

34. Opinion Mining
    - Ideas: implement an algorithm for understanding the opinion on a given product / service / public person from online reviews / social media (e.g. on iPhone, PNL, vaccine, …) (optional: use Romanian data); preliminary: annotate a social media dataset with sentiment scores / scrape reviews and annotate based on stars / use existing dataset; optional: identify specifically which aspects of the product are being referred to - see aspect-based sentiment analysis (e.g. camera is good, battery is bad)
35. Emotion analysis – detect the emotions in a text
    - Ideas: automatically extract emotion scores for individual emotions (see Plutchik's wheel of emotions) for news / blogs / social media texts; optional: use Romanian data; optional: analyze emotions with respect to a given label for a separate task, or use as feature for separate task: e.g. emotions in hate speech datasets, emotions in optimism/pessimism data, emotions in therapy sessions; tracking emotions on social media over time; annotate a dataset with emotion scores and learn to predict them (preferably for a low-resource language/new domain)
    - Resources: https://saifmohammad.com/WebPages/lexicons.html
36. NLP for clinical/ medical data
    - Ideas: NER model for clinical data; information retrieval in medical texts; build embeddings for medical terminology
    - Papers: NER for Medical Entities in Twitter using Sequence to Sequence Neural Networks, Adaptive Generation of Structured Medical Report Using NER Regarding Deep Learning, Proceedings of https://aclweb.org/aclwiki/BioNLP_Workshop , pubmed
37. https://naacl2018.wordpress.com/2018/01/14/test-of-time-paper-nominations-or-classic-computational-linguistics-papers/ (discutia unui articol din aceasta lista)
38. Recent research topics in NLP (articole recente relevante din Computational Linguistics, ACL, COLING, EMNLP, NAACL, EACL, PNAS, etc) see best papers proposals in the last 10 years (Disponibile on-line la https://aclweb.org/aclwiki/Best_paper_awards)
39. LREC 2020 papers for re-experimentation (https://lrec2020.lrec-conf.org/en/reprolang2020/selected-tasks/).
40. NLP applications
    - Ideas: resume analysis, automatic question tagging on StackOverflow/Quora etc, spam classification, automatic essay grading, bot detection, recommender system for products, movies etc
41. NLP & Art

- Ideas: lyrics generation (constrained to rhyme?); classification of literary texts/poems/song lyrics; generate text in the style of Shakespeare…
- Papers: https://rootroo.com/en/hucmac/, Creative GANs for generating poems, lyrics, and metaphors, Weird AI Yankovic: Generating Parody Lyrics

42. NLP for literary texts / digital humanities
    - Ideas: profiling literary characters, character networks, detecting events, profiling authors based on literary texts, OCR for historical texts; build a text processing tool to assist linguists/historians/etc…
    - Resources: https://www.gutenberg.org/
    - Papers: https://sighum.wordpress.com/ (look at proceedings), https://text2story22.inesctec.pt/
43. Search engine
    - Ideas: implement a search engine / information retrieval system on a corpus of data; implement application to allow users to perform searches
44. Text to Speech, Speech to Text
    - Ideas: speech2text system for a chatbot
45. Mental health, depression detection, etc
    - Ideas: solve an eRisk challenge/a CLPsych challenge (free datasets, you need to request the datasets from the organizers); collect and annotate a corpus on depression/another mental illness for a low-resource language
    - Papers: https://erisk.irlab.org/, https://clpsych.org/ (see proceedings)
    - Datasets:
        i. kharrigian/mental-health-datasets: An evolving list of electronic media data sets used to model mental-health status.
        ii. Anorexia. Data from Early risk prediction on the Internet | CLEF 2019 workshop
        iii. Self-harm. Data from CLEF eRisk: Early risk prediction on the Internet | CLEF 2021 workshop
        iv. PTSD Measuring Post Traumatic Stress Disorder in Twitter - Glen Coppersmith Craig Harman Mark Dredze
        v. Suicide Ideation (hard to get access datasets due to ethical concerns)
        vi. Bipolar disorder Not Just Depressed: Bipolar Disorder Prediction on Reddit
        vii. Stress Dreaddit: A Reddit Dataset for Stress Analysis in Social Media
        viii. Multiple Mental Health Conditions Classification SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions
46. Conversational agent / chatbot
    - Ideas: implement a chit-chat bot, customer support bot, Alexa-style bot, robot therapist… , optional: integrate with Google API for complete app including speech module
    - Papers: An Overview of Chatbot Technology, Chatbot for university related FAQs, Conversational agents in healthcare: a systematic review | Journal of the American Medical Informatics Association | Oxford Academic,

https://www.academia.edu/download/61052907/reportmedical_chatbot20191028-44031-hq2g2a.pdf
    - Examples: Amazon Alexa, Google Assistant, Eliza, a chatbot therapist, https://www.talktopoppy.com/
47. Topic modeling – extract topics discussed in a text (classical LDA / neural topic modeling)
    - Ideas: identify and track topics over time in news/scientific texts/social media; implement a topic model from scratch; dynamic topic modelling
    - Papers: Dynamic Topic Models, A Novel Approach of Neural Topic Modelling for Document Clustering, Discovering Discrete Latent Topics with Neural Variational Inference, Studying the Evolution of Scientific Topics and their Relationships
48. Image captioning (automatically generate a description of an image – involves both NLP and computer vision)
    - Ideas: implement+evaluate a model that generates image captions, ideally in sentence format (focus on text generation part)
    - Papers: Convolutional Image Captioning, A Comprehensive Survey of Deep Learning for Image Captioning
49. Language identification
    - Ideas: implement model that identifies language/dialect of given text, code switching detection (“romgleza” etc)
    - Papers: Automatic Language Identification in Texts: A Survey , VarDial 2019 - Evaluation Campaign , Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, Code-switching detection using multilingual DNNS | IEEE Conference Publication, Recurrent-neural-network for language detection on twitter code-switching corpus
50. Question answering
    - Ideas: implement question answering model
    - Papers: QuAC: Question Answering in Context , The Question Answering Systems: A Survey.
51. Stance detection, hyperpartisanship etc
    - Ideas: predict political views, stance on vaccine / social matters / etc; solve shared task on stance detection, bias in news
    - Papers: Multi-Task Stance Detection with Sentiment and Stance Lexicons , http://snap.stanford.edu/quotus/#about
    - Datasets: An Interactive Visualization of the SemEval-2016 Stance Dataset , Stance Detection
52. Explainable AI for NLP
    - Ideas: implement explainability methods for a NLP machine learning model (e.g. attention weights analysis, LIME, adversarial examples, gradients analysis…)
    - Papers: [2009.13295] A Diagnostic Study of Explainability Techniques for Text Classification
    - Resources: https://github.com/marcotcr/lime, https://tf-explain.readthedocs.io/en/latest/
53. Other NLP & CL topics (send an email for approval)

Requirements & guidelines:
Projects should consist of 3 parts:
- paper / technical report
- implementation (code)
- slide presentation
+ a short document explaining the contribution of each student in the team

A project can be focused either on:
- the paper (**survey)** – detailed presentation of existing methods; in this case the paper should be focused on describing the state of the art, comparing existing methods + should contain a proof of concept implementation of a solution to the chosen problem,
- or the **implementation** (in this case the paper will be focused on the methodology and describing technical and experimental details):

- implementing an end-to-end application to solve the problem
- implementing a solution described in a paper (projects 37, 38, 39)
- implementing a novel solution to a problem

All papers/technical reports will follow the classical structure of a research article (approx 4 pages):
- short summary (abstract)
- analysis of main idea
- related work: state of the art (SOTA) where it exists, short history, recent and/or related results
- In case you're presenting a survey: explain main methodologies and selection process (i.e. you are surveying either chronologically, or in order of SOTA achievements), discuss advantages and disadvantages to the methods used and introduced
- in the case of presenting specific applications: describe the method, compare it with other results in the field
- conclusions and future work, directions for further improvement
- references

Teams of 2-3, max 4 people. Any topic can be chosen by max 4 teams.
Add your name on the google sheets document next to the chosen project topic.
Additional details on useful resources here:
https://github.com/ananana/nlp-projects/blob/main/README.md