pdf bib abs **Narrate Dialogues for Better Summarization**
Ruochen Xu | Chenguang Zhu | Michael Zeng

pdf bib abs **Towards Identifying Social Bias in Dialog Systems: Framework, Dataset, and Benchmark**
Jingyan Zhou | Jiawen Deng | Fei Mi | Yitong Li | Yasheng Wang | Minlie Huang | Xin Jiang | Qun Liu | Helen Meng

pdf bib abs **CrossRE: A Cross-Domain Dataset for Relation Extraction**
Elisa Bassignana | Barbara Plank

pdf bib abs **Probing Structural Knowledge from Pre-trained Language Model for Argumentation Relation Classification**
Yang Sun | Bin Liang | Jianzhu Bao | Min Yang | Ruifeng Xu

pdf bib abs **LogicNMR: Probing the Non-monotonic Reasoning Ability of Pre-trained Language Models**
Yeliang Xiu | Zhanhao Xiao | Yongmei Liu

pdf bib abs **Cheater's Bowl: Human vs. Computer Search Strategies for Open-Domain QA**
Wanrong He | Andrew Mao | Jordan Boyd-Graber

pdf bib abs **FRSUM: Towards Faithful Abstractive Summarization via Enhancing Factual Robustness**
Wenhao Wu | Wei Li | Jiachen Liu | Xinyan Xiao | Ziqiang Cao | Sujian Li | Hua Wu

pdf bib abs **PoeLM: A Meter- and Rhyme-Controllable Language Model for Unsupervised Poetry Generation**
Aitor Ormazabal | Mikel Artetxe | Manex Agirrezabal | Aitor Soroa | Eneko Agirre

pdf bib abs **ProGen: Progressive Zero-shot Dataset Generation via In-context Feedback**
Jiacheng Ye | Jiahui Gao | Zhiyong Wu | Jiangtao Feng | Tao Yu | Lingpeng Kong

pdf bib abs **Constructing Highly Inductive Contexts for Dialogue Safety through Controllable Reverse Generation**
Zhexin Zhang | Jiale Cheng | Hao Sun | Jiawen Deng | Fei Mi | Yasheng Wang | Lifeng Shang | Minlie Huang

pdf bib abs **Language Prior Is Not the Only Shortcut: A Benchmark for Shortcut Learning in VQA**
Qingyi Si | Fandong Meng | Mingyu Zheng | Zheng Lin | Yuanxin Liu | Peng Fu | Yanan Cao | Weiping Wang | Jie Zhou

pdf bib abs **Bridging the Training-Inference Gap for Dense Phrase Retrieval**
Gyuwan Kim | Jinhyuk Lee | Barlas Oguz | Wenhan Xiong | Yizhe Zhang | Yashar Mehdad | William Yang Wang

pdf bib abs **Beyond Counting Datasets: A Survey of Multilingual Dataset Construction and Necessary Resources**
Xinyan Yu | Trina Chatterjee | Akari Asai | Junjie Hu | Eunsol Choi

pdf bib abs **ERNIE-Layout: Layout Knowledge Enhanced Pre-training for Visually-rich Document Understanding**
Qiming Peng | Yinxu Pan | Wenjin Wang | Bin Luo | Zhenyu Zhang | Zhengjie Huang | Yuhui Cao | Weichong Yin | Yongfeng Chen | Yin Zhang | Shikun Feng | Yu Sun | Hao Tian | Hua Wu | Haifeng Wang

pdf bib abs **Do Charge Prediction Models Learn Legal Theory?**
Zhenwei An | Quzhe Huang | Cong Jiang | Yansong Feng | Dongyan Zhao

pdf bib abs **Keep Me Updated! Memory Management in Long-term Conversations**
Sanghwan Bae | Donghyun Kwak | Soyoung Kang | Min Young Lee | Sungdong Kim | Yuin Jeong | Hyeri Kim | Sang-Woo Lee | Woomyoung Park | Nako Sung

pdf bib abs **A Unified Dialogue User Simulator for Few-shot Data Augmentation**
Dazhen Wan | Zheng Zhang | Qi Zhu | Lizi Liao | Minlie Huang

pdf bib abs **An Error-Guided Correction Model for Chinese Spelling Error Correction**
Rui Sun | Xiuyu Wu | Yunfang Wu

pdf bib abs **Describing Sets of Images with Textual-PCA**
Oded Hupert | Idan Schwartz | Lior Wolf

pdf bib abs **Learning to Model Editing Processes**
Machel Reid | Graham Neubig

pdf bib abs **PALT: Parameter-Lite Transfer of Language Models for Knowledge Graph Completion**
Jianhao Shen | Chenguang Wang | Ye Yuan | Jiawei Han | Heng Ji | Koushik Sen | Ming Zhang | Dawn Song

pdf bib abs **Prompt-based Connective Prediction Method for Fine-grained Implicit Discourse Relation Recognition**
Hao Zhou | Man Lan | Yuanbin Wu | Yuefeng Chen | Meirong Ma

pdf bib abs **On Utilizing Constituent Language Resources to Improve Downstream Tasks in Hinglish**
Vishwajeet Kumar | Rudra Murthy | Tejas Dhamecha

pdf bib abs **SYGMA: A System for Generalizable and Modular Question Answering Over Knowledge Bases**
Sumit Neelam | Udit Sharma | Hima Karanam | Shajith Ikbal | Pavan Kapanipathi | Ibrahim Abdelaziz | Nandana Mihindukulasooriya | Young-Suk Lee | Santosh Srivastava | Cezar Pendus | Saswati Dana | Dinesh Garg | Achille Fokoue | G P Shrivatsa Bhargav | Dinesh Khandelwal |

Srinivas Ravishankar | Sairam Gurajada | Maria Chang | Rosario Uceda-Sosa | Salim Roukos | Alexander Gray | Guilherme Lima | Ryan Riegel | Francois Luus | L V Subramaniam

pdf bib abs **Instance-Guided Prompt Learning for Few-Shot Text Matching**
Jia Du | Xuanyu Zhang | Siyi Wang | Kai Wang | Yanquan Zhou | Lei Li | Qing Yang | Dongliang Xu

pdf bib abs **M3: Multi-level dataset for Multi-document summarisation of Medical studies**
Yulia Otmakhova | Karin Verspoor | Timothy Baldwin | Antonio Jimeno Yepes | Jey Han Lau

pdf bib abs **Adapters for Enhanced Modeling of Multilingual Knowledge and Text**
Yifan Hou | Wenxiang Jiao | Meizhen Liu | Carl Allen | Zhaopeng Tu | Mrinmaya Sachan

pdf bib abs **SepLL: Separating Latent Class Labels from Weak Supervision Noise**
Andreas Stephan | Vasiliki Kougia | Benjamin Roth

pdf bib abs **Probing Relational Knowledge in Language Models via Word Analogies**
Kiamehr Rezaee | Jose Camacho-Collados

pdf bib abs **Semi-Supervised Lifelong Language Learning**
Yingxiu Zhao | Yinhe Zheng | Bowen Yu | Zhiliang Tian | Dongkyu Lee | Jian Sun | Yongbin Li | Nevin L. Zhang

pdf bib abs **Parameter-free Automatically Prompting: A Latent Pseudo Label Mapping Model for Prompt-based Learning**
Jirui Qi | Richong Zhang | Junfan Chen | Jaein Kim | Yongyi Mao

pdf bib abs **Exploring Logographic Image for Chinese Aspect-based Sentiment Classification**
Xiabing Zhou | Renjie Feng | Xiaotong Jiang | Zhongqing Wang

pdf bib abs **On the Role of Bidirectionality in Language Model Pre-Training**
Mikel Artetxe | Jingfei Du | Naman Goyal | Luke Zettlemoyer | Veselin Stoyanov

pdf bib abs **You Are What You Talk About: Inducing Evaluative Topics for Personality Analysis**
Josip Jukić | Iva Vukojević | Jan Snajder

pdf bib abs **CAT-probing: A Metric-based Approach to Interpret How Pre-trained Models for Programming Language Attend Code Structure**
Nuo Chen | Qiushi Sun | Renyu Zhu | Xiang Li | Xuesong Lu | Ming Gao

pdf bib abs **Learning to Revise References for Faithful Summarization**
Griffin Adams | Han-Chin Shing | Qing Sun | Christopher Winestock | Kathleen McKeown | Noémie Elhadad

pdf bib abs **Towards Intention Understanding in Suicidal Risk Assessment with Natural Language Processing**
Shaoxiong Ji

pdf bib abs **On the Impact of Temporal Concept Drift on Model Explanations**
Zhixue Zhao | George Chrysostomou | Kalina Bontcheva | Nikolaos Aletras

pdf bib abs **Text-Only Training for Image Captioning using Noise-Injected CLIP**
David Nukrai | Ron Mokady | Amir Globerson

pdf bib abs **Improving Sharpness-Aware Minimization with Fisher Mask for Better Generalization on Language Models**
Qihuang Zhong | Liang Ding | Li Shen | Peng Mi | Juhua Liu | Bo Du | Dacheng Tao

pdf bib abs **TINA: Textual Inference with Negation Augmentation**
Chadi Helwe | Simon Coumes | Chloé Clavel | Fabian Suchanek

pdf bib abs **Improving Bilingual Lexicon Induction with Cross-Encoder Reranking**
Yaoyiran Li | Fangyu Liu | Ivan Vulić | Anna Korhonen

pdf bib abs **Mixed-modality Representation Learning and Pre-training for Joint Table-and-Text Retrieval in OpenQA**
Junjie Huang | Wanjun Zhong | Qian Liu | Ming Gong | Daxin Jiang | Nan Duan

pdf bib abs **The Effects of Corpus Choice and Morphosyntax on Multilingual Space Induction**
Vinit Ravishankar | Joakim Nivre

pdf bib abs **Modeling Complex Dialogue Mappings via Sentence Semantic Segmentation Guided Conditional Variational Auto-Encoder**
Bin Sun | Shaoxiong Feng | Yiwei Li | Weichao Wang | Fei Mi | Yitong Li | Kan Li

pdf bib abs **Graph Embeddings for Argumentation Quality Assessment**
Santiago Marro | Elena Cabrio | Serena Villata

pdf bib abs **SMiLE: Schema-augmented Multi-level Contrastive Learning for Knowledge Graph Link Prediction**
Miao Peng | Ben Liu | Qianqian Xie | Wenjie Xu | Hua Wang | Min Peng

pdf bib abs **Multilingual Multimodal Learning with Machine Translated Text**
Chen Qiu | Dan Oneață | Emanuele Bugliarello | Stella Frank | Desmond Elliott

pdf bib abs **Learning From the Source Document: Unsupervised Abstractive Summarization**
Haojie Zhuang | Wei Emma Zhang | Jian Yang | Congbo Ma | Yutong Qu | Quan Z. Sheng

pdf bib abs **How to Do Things without Words: Modeling Semantic Drift of Emoji**
Eyal Arviv | Oren Tsur

pdf bib abs **Mind Your Bias: A Critical Review of Bias Detection Methods for Contextual Language Models**
Silke Husse | Andreas Spitz

pdf bib abs **ZeroPrompt: Scaling Prompt-Based Pretraining to 1,000 Tasks Improves Zero-Shot Generalization**
Hanwei Xu | Yujun Chen | Yulun Du | Nan Shao | Wang Yanggang | Haiyu Li | Zhilin Yang

pdf bib abs **Semantic Role Labeling Meets Definition Modeling: Using Natural Language to Describe Predicate-Argument Structures**
Simone Conia | Edoardo Barba | Alessandro Scirè | Roberto Navigli

pdf bib abs **Is anisotropy really the cause of BERT embeddings not being semantic?**
Alejandro Fuster Baggetto | Victor Fresno

pdf bib abs **m^4 Adapter: Multilingual Multi-Domain Adaptation for Machine Translation with a Meta-Adapter**
Wen Lai | Alexandra Chronopoulou | Alexander Fraser

pdf bib abs **Textual Enhanced Contrastive Learning for Solving Math Word Problems**
Yibin Shen | Qianying Liu | Zhuoyuan Mao | Fei Cheng | Sadao Kurohashi

pdf bib abs **What Do Compressed Multilingual Machine Translation Models Forget?**
Alireza Mohammadshahi | Vassilina Nikoulina | Alexandre Berard | Caroline Brun | James Henderson | Laurent Besacier

pdf bib abs **Controllable Dialogue Simulation with In-context Learning**
Zekun Li | Wenhu Chen | Shiyang Li | Hong Wang | Jing Qian | Xifeng Yan

pdf bib abs **Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards**
Jean-Benoit Delbrouck | Pierre Chambon | Christian Bluethgen | Emily Tsai | Omar Almusa | Curtis Langlotz

pdf bib abs **Recursive Neural Networks with Bottlenecks Diagnose (Non-)Compositionality**
Verna Dankers | Ivan Titov

pdf bib abs **HumSet: Dataset of Multilingual Information Extraction and Classification for Humanitarian Crises Response**
Selim Fekih | Nicolo' Tamagnone | Benjamin Minixhofer | Ranjan Shrestha | Ximena Contla | Ewan Oglethorpe | Navid Rekabsaz

pdf bib abs **Viterbi Decoding of Directed Acyclic Transformer for Non-Autoregressive Machine Translation**
Chenze Shao | Zhengrui Ma | Yang Feng

pdf bib abs **Lexical Generalization Improves with Larger Models and Longer Training**
Elron Bandel | Yoav Goldberg | Yanai Elazar

pdf bib abs **Realistic Data Augmentation Framework for Enhancing Tabular Reasoning**
Dibyakanti Kumar | Vivek Gupta | Soumya Sharma | Shuo Zhang

pdf bib abs **Inducing Generalizable and Interpretable Lexica**
Yilin Geng | Zetian Wu | Roshan Santhosh | Tejas Srivastava | Lyle Ungar | João Sedoc

pdf bib abs **The Curious Case of Absolute Position Embeddings**
Koustuv Sinha | Amirhossein Kazemnejad | Siva Reddy | Joelle Pineau | Dieuwke Hupkes | Adina Williams

pdf bib abs **Goal-oriented Vision-and-Dialog Navigation via Reinforcement Learning**
Yan Cao | Keting Lu | David DeFazio | Shiqi Zhang

pdf bib abs **Leveraging Data Recasting to Enhance Tabular Reasoning**
Aashna Jena | Vivek Gupta | Manish Shrivastava | Julian Eisenschlos

pdf bib abs **Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again**
Bernal Jimenez Gutierrez | Nikolas McNeal | Clayton Washington | You Chen | Lang Li | Huan Sun | Yu Su

pdf bib abs **Attention weights accurately predict language representations in the brain**
Mathis Lamarre | Catherine Chen | Fatma Deniz

pdf bib abs **Improving HowNet-Based Chinese Word Sense Disambiguation with Translations**
Xiang Zhang | Bradley Hauer | Grzegorz Kondrak

pdf bib abs **Mask-then-Fill: A Flexible and Effective Data Augmentation Framework for Event Extraction**
Jun Gao | Changlong Yu | Wei Wang | Huan Zhao | Ruifeng Xu

pdf bib abs **MOBA-E2C: Generating MOBA Game Commentaries via Capturing Highlight Events from the Meta-Data**

Dawei Zhang | Sixing Wu | Yao Guo | Xiangqun Chen

pdf bib abs **Enhancing Automatic Readability Assessment with Pre-training and Soft Labels for Ordinal Regression**
Jinshan Zeng | Yudong Xie | Xianglong Yu | John Lee | Ding-Xuan Zhou

pdf bib abs **Opening up Minds with Argumentative Dialogues**
Youmna Farag | Charlotte Brand | Jacopo Amidei | Paul Piwek | Tom Stafford | Svetlana Stoyanchev | Andreas Vlachos

pdf bib abs **You Are My Type! Type Embeddings for Pre-trained Language Models**
Mohammed Saeed | Paolo Papotti

pdf bib abs **Generating Textual Adversaries with Minimal Perturbation**
Xingyi Zhao | Lu Zhang | Depeng Xu | Shuhan Yuan

pdf bib abs **SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings**
Jan Engler | Sandipan Sikdar | Marlene Lutz | Markus Strohmaier

pdf bib abs **Contextualizing Language Models for Norms Diverging from Social Majority**
Niklas Kiehne | Hermann Kroll | Wolf-Tilo Balke

pdf bib abs **Empathetic Dialogue Generation via Sensitive Emotion Recognition and Sensible Knowledge Selection**
Lanrui Wang | Jiangnan Li | Zheng Lin | Fandong Meng | Chenxu Yang | Weiping Wang | Jie Zhou

pdf bib abs **Joint Multilingual Knowledge Graph Completion and Alignment**
Vinh Tong | Dat Quoc Nguyen | Trung Thanh Huynh | Tam Thanh Nguyen | Quoc Viet Hung Nguyen | Mathias Niepert

pdf bib abs **A Framework for Automatic Generation of Spoken Question-Answering Data**
Merve Ünlü Menevşe | Yusufcan Manav | Ebru Arisoy | Arzucan Özgür

pdf bib abs **Readability Controllable Biomedical Document Summarization**
Zheheng Luo | Qianqian Xie | Sophia Ananiadou

pdf bib abs **Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions**
Torsten Wörtwein | Lisa Sheeber | Nicholas Allen | Jeffrey Cohn | Louis-Philippe Morency

pdf bib abs **Generalization Differences between End-to-End and Neuro-Symbolic Vision-Language Reasoning Systems**
Wang Zhu | Jesse Thomason | Robin Jia

pdf bib abs **Learning to Model Multimodal Semantic Alignment for Story Visualization**
Bowen Li | Thomas Lukasiewicz

pdf bib abs **SciFact-Open: Towards open-domain scientific claim verification**
David Wadden | Kyle Lo | Bailey Kuehl | Arman Cohan | Iz Beltagy | Lucy Lu Wang | Hannaneh Hajishirzi

pdf bib abs **COMET-QE and Active Learning for Low-Resource Machine Translation**
Everlyn Chimoto | Bruce Bassett

pdf bib abs **MedicalSum: A Guided Clinical Abstractive Summarization Model for Generating Medical Reports from Patient-Doctor Conversations**
George Michalopoulos | Kyle Williams | Gagandeep Singh | Thomas Lin

pdf bib abs **Leveraging Training Dynamics and Self-Training for Text Classification**
Tiberiu Sosea | Cornelia Caragea

pdf bib abs **Learning to Infer from Unlabeled Data: A Semi-supervised Learning Approach for Robust Natural Language Inference**
Mobashir Sadat | Cornelia Caragea

pdf bib abs **Unsupervised Text Deidentification**
John Morris | Justin Chiu | Ramin Zabih | Alexander Rush

pdf bib abs **Federated Continual Learning for Text Classification via Selective Inter-client Transfer**
Yatin Chaudhary | Pranav Rai | Matthias Schubert | Hinrich Schütze | Pankaj Gupta

pdf bib abs **DOROTHIE: Spoken Dialogue for Handling Unexpected Situations in Interactive Autonomous Driving Agents**
Ziqiao Ma | Benjamin VanDerPloeg | Cristian-Paul Bara | Yidong Huang | Eui-In Kim | Felix Gervits | Matthew Marge | Joyce Chai

pdf bib abs **He Said, She Said: Style Transfer for Shifting the Perspective of Dialogues**
Amanda Bertsch | Graham Neubig | Matthew R. Gormley

pdf bib abs **Dynamic Augmentation Data Selection for Few-shot Text Classification**
Guangliang Liu | Lifeng Jin | Owen Yuan | Jiayu Zhou

pdf bib abs **KPDROP: Improving Absent Keyphrase Generation**

Jishnu Ray Chowdhury | Seo Yeon Park | Tuhin Kundu | Cornelia Caragea

pdf bib abs **Natural Language Deduction through Search over Statement Compositions**
Kaj Bostrom | Zayne Sprague | Swarat Chaudhuri | Greg Durrett

pdf bib abs **EnDex: Evaluation of Dialogue Engagingness at Scale**
Guangxuan Xu | Ruibo Liu | Fabrice Harel-Canada | Nischal Reddy Chandra | Nanyun Peng

pdf bib abs **LOPS: Learning Order Inspired Pseudo-Label Selection for Weakly Supervised Text Classification**
Dheeraj Mekala | Chengyu Dong | Jingbo Shang

pdf bib abs **Train Flat, Then Compress: Sharpness-Aware Minimization Learns More Compressible Models**
Clara Na | Sanket Vaibhav Mehta | Emma Strubell

pdf bib abs **Structural Contrastive Representation Learning for Zero-shot Multi-label Text Classification**
Tianyi Zhang | Zhaozhuo Xu | Tharun Medini | Anshumali Shrivastava

pdf bib abs **Improving Generalization of Pre-trained Language Models via Stochastic Weight Averaging**
Peng Lu | Ivan Kobyzev | Mehdi Rezagholizadeh | Ahmad Rashid | Ali Ghodsi | Phillippe Langlais

pdf bib abs **Learn What Is Possible, Then Choose What Is Best: Disentangling One-To-Many Relations in Language Through Text-based Games**
Benjamin Towle | Ke Zhou

pdf bib abs **Structurally Diverse Sampling for Sample-Efficient Training and Comprehensive Evaluation**
Shivanshu Gupta | Sameer Singh | Matt Gardner

pdf bib abs **Unsupervised Multi-Granularity Summarization**
Ming Zhong | Yang Liu | Suyu Ge | Yuning Mao | Yizhu Jiao | Xingxing Zhang | Yichong Xu | Chenguang Zhu | Michael Zeng | Jiawei Han

pdf bib abs **HeLo: Learning-Free Lookahead Decoding for Conversation Infilling**
Ivan Lee | Taylor Berg-Kirkpatrick

pdf bib abs **Invernet: An Inversion Attack Framework to Infer Fine-Tuning Datasets through Word Embeddings**
Ishrak Hayet | Zijun Yao | Bo Luo

pdf bib abs **LawngNLI: A Long-Premise Benchmark for In-Domain Generalization from Short to Long Contexts and for Implication-Based Retrieval**
William Bruno | Dan Roth

pdf bib abs **Distillation-Resistant Watermarking for Model Protection in NLP**
Xuandong Zhao | Lei Li | Yu-Xiang Wang

pdf bib abs **NeuroCounterfactuals: Beyond Minimal-Edit Counterfactuals for Richer Data Augmentation**
Phillip Howard | Gadi Singer | Vasudev Lal | Yejin Choi | Swabha Swayamdipta

pdf bib abs **Don't Just Clean It, Proxy Clean It: Mitigating Bias by Proxy in Pre-Trained Models**
Swetasudha Panda | Ari Kobren | Michael Wick | Qinlan Shen

pdf bib abs **The Undesirable Dependence on Frequency of Gender Bias Metrics Based on Word Embeddings**
Francisco Valentini | Germán Rosati | Diego Fernandez Slezak | Edgar Altszyler

pdf bib abs **BioNLI: Generating a Biomedical NLI Dataset Using Lexico-semantic Constraints for Adversarial Examples**
Mohaddeseh Bastan | Mihai Surdeanu | Niranjan Balasubramanian

pdf bib abs **Self-supervised Cross-modal Pretraining for Speech Emotion Recognition and Sentiment Analysis**
Iek-Heng Chu | Ziyi Chen | Xinlu Yu | Mei Han | Jing Xiao | Peng Chang

pdf bib abs **Multimodal Conversation Modelling for Topic Derailment Detection**
Zhenhao Li | Marek Rei | Lucia Specia

pdf bib abs **Active Learning for Abstractive Text Summarization**
Akim Tsvigun | Ivan Lysenko | Danila Sedashov | Ivan Lazichny | Eldar Damirov | Vladimir Karlov | Artemy Belousov | Leonid Sanochkin | Maxim Panov | Alexander Panchenko | Mikhail Burtsev | Artem Shelmanov

pdf bib abs **Finding Memo: Extractive Memorization in Constrained Sequence Generation Tasks**
Vikas Raunak | Arul Menezes

pdf bib abs **SALTED: A Framework for SAlient Long-tail Translation Error Detection**
Vikas Raunak | Matt Post | Arul Menezes

pdf bib abs **Discord Questions: A Computational Approach To Diversity Analysis in News Coverage**
Philippe Laban | Chien-Sheng Wu | Lidiya Murakhovs'ka | Xiang | Caiming Xiong

pdf bib abs **FocusQA: Open-Domain Question Answering with a Context in Focus**
Gianni Barlacchi | Ivano Lauriola | Alessandro Moschitti | Marco Del Tredici | Xiaoyu Shen | Thuy Vu | Bill Byrne | Adrià de Gispert

pdf bib abs **Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media**
Chan Young Park | Julia Mendelsohn | Anjalie Field | Yulia Tsvetkov

pdf bib abs **Disentangling Task Relations for Few-shot Text Classification via Self-Supervised Hierarchical Task Clustering**
Juan Zha | Zheng Li | Ying Wei | Yu Zhang

pdf bib abs **XRICL: Cross-lingual Retrieval-Augmented In-Context Learning for Cross-lingual Text-to-SQL Semantic Parsing**
Peng Shi | Rui Zhang | He Bai | Jimmy Lin

pdf bib abs **Continuation KD: Improved Knowledge Distillation through the Lens of Continuation Optimization**
Aref Jafari | Ivan Kobyzev | Mehdi Rezagholizadeh | Pascal Poupart | Ali Ghodsi

pdf bib abs **Detecting Dementia from Long Neuropsychological Interviews**
Nauman Dawalatabad | Yuan Gong | Sameer Khurana | Rhoda Au | James Glass

pdf bib abs **Sarcasm Detection is Way Too Easy! An Empirical Comparison of Human and Machine Sarcasm Detection**
Ibrahim Abu Farha | Steven Wilson | Silviu Oprea | Walid Magdy

pdf bib abs **Cross-lingual Text-to-SQL Semantic Parsing with Representation Mixup**
Peng Shi | Linfeng Song | Lifeng Jin | Haitao Mi | He Bai | Jimmy Lin | Dong Yu

pdf bib abs **JamPatoisNLI: A Jamaican Patois Natural Language Inference Dataset**
Ruth-Ann Armstrong | John Hewitt | Christopher Manning

pdf bib abs **Are Neural Topic Models Broken?**
Alexander Miserlis Hoyle | Rupak Sarkar | Pranav Goel | Philip Resnik

pdf bib abs **Know Thy Strengths: Comprehensive Dialogue State Tracking Diagnostics**
Hyundong Cho | Chinnadhurai Sankar | Christopher Lin | Kaushik Sadagopan | Shahin Shayandeh | Asli Celikyilmaz | Jonathan May | Ahmad Beirami

pdf bib abs **Open-domain Question Answering via Chain of Reasoning over Heterogeneous Knowledge**
Kaixin Ma | Hao Cheng | Xiaodong Liu | Eric Nyberg | Jianfeng Gao

pdf bib abs **Detecting Languages Unintelligible to Multilingual Models through Local Structure Probes**
Louis Clouatre | Prasanna Parthasarathi | Amal Zouaq | Sarath Chandar

pdf bib abs **Cards Against AI: Predicting Humor in a Fill-in-the-blank Party Game**
Dan Ofer | Dafna Shahaf

pdf bib abs **Open-Vocabulary Argument Role Prediction For Event Extraction**
Yizhu Jiao | Sha Li | Yiqing Xie | Ming Zhong | Heng Ji | Jiawei Han

pdf bib abs **Token-level Sequence Labeling for Spoken Language Understanding using Compositional End-to-End Models**
Siddhant Arora | Siddharth Dalmia | Brian Yan | Florian Metze | Alan W Black | Shinji Watanabe

pdf bib abs **Baked-in State Probing**
Shubham Toshniwal | Sam Wiseman | Karen Livescu | Kevin Gimpel

pdf bib abs **ClinicalT5: A Generative Language Model for Clinical Text**
Qiuhao Lu | Dejing Dou | Thien Nguyen

pdf bib abs **Find Someone Who: Visual Commonsense Understanding in Human-Centric Grounding**
Haoxuan You | Rui Sun | Zhecan Wang | Kai-Wei Chang | Shih-Fu Chang

pdf bib abs **CrisisLTLSum: A Benchmark for Local Crisis Event Timeline Extraction and Summarization**
Hossein Rajaby Faghihi | Bashar Alhafni | Ke Zhang | Shihao Ran | Joel Tetreault | Alejandro Jaimes

pdf bib abs **Prompt-Tuning Can Be Much Better Than Fine-Tuning on Cross-lingual Understanding With Multilingual Language Models**
Lifu Tu | Caiming Xiong | Yingbo Zhou

pdf bib abs **BERT Meets CTC: New Formulation of End-to-End Speech Recognition with Pre-trained Masked Language Model**
Yosuke Higuchi | Brian Yan | Siddhant Arora | Tetsuji Ogawa | Tetsunori Kobayashi | Shinji Watanabe

pdf bib abs **EtriCA: Event-Triggered Context-Aware Story Generation Augmented by Cross Attention**
Chen Tang | Chenghua Lin | Henglin Huang | Frank Guerin | Zhihao Zhang

pdf bib abs **LADIS: Language Disentanglement for 3D Shape Editing**
Ian Huang | Panos Achlioptas | Tianyi Zhang | Sergei Tulyakov | Minhyuk Sung | Leonidas Guibas

pdf bib abs **Effective Pretraining Objectives for Transformer-based Autoencoders**

Luca Di Liello | Matteo Gabburo | Alessandro Moschitti

pdf bib abs **Language Model Detoxification in Dialogue with Contextualized Stance Control**
Jing Qian | Xifeng Yan

pdf bib abs **Multilingual SubEvent Relation Extraction: A Novel Dataset and Structure Induction Method**
Viet Lai | Hieu Man | Linh Ngo | Franck Dernoncourt | Thien Nguyen

pdf bib abs **A Two-Stage Approach towards Generalization in Knowledge Base Question Answering**
Srinivas Ravishankar | Dung Thai | Ibrahim Abdelaziz | Nandana Mihindukulasooriya | Tahira Naseem | Pavan Kapanipathi | Gaetano Rossiello | Achille Fokoue

pdf bib abs **Few-Shot (Dis)Agreement Identification in Online Discussions with Regularized and Augmented Meta-Learning**
Yuanyuan Lei | Ruihong Huang

pdf bib abs **Data Cartography for Low-Resource Neural Machine Translation**
Aquia Richburg | Marine Carpuat

pdf bib abs **Augmenting Multi-Turn Text-to-SQL Datasets with Self-Play**
Qi Liu | Zihuiwen Ye | Tao Yu | Linfeng Song | Phil Blunsom

pdf bib abs **Prompt Compression and Contrastive Conditioning for Controllability and Toxicity Reduction in Language Models**
David Wingate | Mohammad Shoeybi | Taylor Sorensen

pdf bib abs **NaturalAdversaries: Can Naturalistic Adversaries Be as Effective as Artificial Adversaries?**
Saadia Gabriel | Hamid Palangi | Yejin Choi

pdf bib abs **Multi-Path Transformer is Better: A Case Study on Neural Machine Translation**
Ye Lin | Shuhan Zhou | Yanyang Li | Anxiang Ma | Tong Xiao | Jingbo Zhu

pdf bib abs **Unsupervised Learning of Hierarchical Conversation Structure**
Bo-Ru Lu | Yushi Hu | Hao Cheng | Noah A. Smith | Mari Ostendorf

pdf bib abs **Task Compass: Scaling Multi-task Pre-training with Task Prefix**
Zhuosheng Zhang | Shuohang Wang | Yichong Xu | Yuwei Fang | Wenhao Yu | Yang Liu | Hai Zhao | Chenguang Zhu | Michael Zeng

pdf bib abs **Sharpness-Aware Minimization with Dynamic Reweighting**
Wenxuan Zhou | Fangyu Liu | Huan Zhang | Muhao Chen

pdf bib abs **Predicting Long-Term Citations from Short-Term Linguistic Influence**
Sandeep Soni | David Bamman | Jacob Eisenstein

pdf bib abs **Joint Audio/Text Training for Transformer Rescorer of Streaming Speech Recognition**
Suyoun Kim | Ke Li | Lucas Kabela | Ron Huang | Jiedan Zhu | Ozlem Kalinli | Duc Le

pdf bib abs **TyDiP: A Dataset for Politeness Classification in Nine Typologically Diverse Languages**
Anirudh Srinivasan | Eunsol Choi

pdf bib abs **Probing Cross-modal Semantics Alignment Capability from the Textual Perspective**
Zheng Ma | Shi Zong | Mianzhi Pan | Jianbing Zhang | Shujian Huang | Xinyu Dai | Jiajun Chen

pdf bib abs **Hidden State Variability of Pretrained Language Models Can Guide Computation Reduction for Transfer Learning**
Shuo Xie | Jiahao Qiu | Ankita Pasad | Li Du | Qing Qu | Hongyuan Mei

pdf bib abs **Language Models as Agent Models**
Jacob Andreas

pdf bib abs **Combinatory Grammar Tells Underlying Relevance among Entities**
Yuanhe Tian | Yan Song

pdf bib abs **Leveraging Open Data and Task Augmentation to Automated Behavioral Coding of Psychotherapy Conversations in Low-Resource Scenarios**
Zhuohao Chen | Nikolaos Flemotomos | Zac Imel | David Atkins | Shrikanth Narayanan

pdf bib abs **Learning to Detect Noisy Labels Using Model-Based Features**
Zhihao Wang | Zongyu Lin | Junjie Wen | Xianxin Chen | Peiqi Liu | Guidong Zheng | Yujun Chen | Zhilin Yang

pdf bib abs **Keyphrase Generation Beyond the Boundaries of Title and Abstract**
Krishna Garg | Jishnu Ray Chowdhury | Cornelia Caragea

pdf bib abs **Composition, Attention, or Both?**
Ryo Yoshida | Yohei Oseki

pdf bib abs **CDGP: Automatic Cloze Distractor Generation based on Pre-trained Language Model**
Shang-Hsuan Chiang | Ssu-Cheng Wang | Yao-Chung Fan

pdf bib abs **G3: Geolocation via Guidebook Grounding**
Grace Luo | Giscard Biamby | Trevor Darrell | Daniel Fried | Anna Rohrbach

pdf bib abs **Controlling Bias Exposure for Fair Interpretable Predictions**
Zexue He | Yu Wang | Julian McAuley | Bodhisattwa Prasad Majumder

pdf bib abs **Investigating the Benefits of Free-Form Rationales**
Jiao Sun | Swabha Swayamdipta | Jonathan May | Xuezhe Ma

pdf bib abs **Data-Efficient Concept Extraction from Pre-trained Language Models for Commonsense Explanation Generation**
Yanbo Fang | Yongfeng Zhang

pdf bib abs **Unsupervised Domain Adaptation for Joint Information Extraction**
Nghia Ngo | Bonan Min | Thien Nguyen

pdf bib abs **Foiling Training-Time Attacks on Neural Machine Translation Systems**
Jun Wang | Xuanli He | Benjamin Rubinstein | Trevor Cohn

pdf bib abs **Learning Action-Effect Dynamics for Hypothetical Vision-Language Reasoning Task**
Shailaja Keyur Sampat | Pratyay Banerjee | Yezhou Yang | Chitta Baral

pdf bib abs **Named Entity and Relation Extraction with Multi-Modal Retrieval**
Xinyu Wang | Jiong Cai | Yong Jiang | Pengjun Xie | Kewei Tu | Wei Lu

pdf bib abs **Calibrating Factual Knowledge in Pretrained Language Models**
Qingxiu Dong | Damai Dai | Yifan Song | Jingjing Xu | Zhifang Sui | Lei Li

pdf bib abs **MCPG: A Flexible Multi-Level Controllable Framework for Unsupervised Paraphrase Generation**
Yi Chen | Haiyun Jiang | Lemao Liu | Rui Wang | Shuming Shi | Ruifeng Xu

pdf bib abs **WordTies: Measuring Word Associations in Language Models via Constrained Sampling**
Peiran Yao | Tobias Renwick | Denilson Barbosa

pdf bib abs **Exploring The Landscape of Distributional Robustness for Question Answering Models**
Anas Awadalla | Mitchell Wortsman | Gabriel Ilharco | Sewon Min | Ian Magnusson | Hannaneh Hajishirzi | Ludwig Schmidt

pdf bib abs **Collaborative Reasoning on Multi-Modal Semantic Graphs for Video-Grounded Dialogue Generation**
Xueliang Zhao | Yuxuan Wang | Chongyang Tao | Chenshuo Wang | Dongyan Zhao

pdf bib abs **Partitioned Gradient Matching-based Data Subset Selection for Compute-Efficient Robust ASR Training**
Ashish Mittal | Durga Sivasubramanian | Rishabh Iyer | Preethi Jyothi | Ganesh Ramakrishnan

pdf bib abs **Adaptive Graph Convolutional Network for Knowledge Graph Entity Alignment**
Renbo Zhu | Xukun Luo | Meng Ma | Ping Wang

pdf bib abs **Towards Robust NLG Bias Evaluation with Syntactically-diverse Prompts**
Arshiya Aggarwal | Jiao Sun | Nanyun Peng

pdf bib abs **PcMSP: A Dataset for Scientific Action Graphs Extraction from Polycrystalline Materials Synthesis Procedure Text**
Xianjun Yang | Ya Zhuo | Julia Zuo | Xinlu Zhang | Stephen Wilson | Linda Petzold

pdf bib abs **Validity Assessment of Legal Will Statements as Natural Language Inference**
Alice Kwak | Jacob Israelsen | Clayton Morrison | Derek Bambauer | Mihai Surdeanu

pdf bib abs **AdaPrompt: Adaptive Model Training for Prompt-based NLP**
Yulong Chen | Yang Liu | Li Dong | Shuohang Wang | Chenguang Zhu | Michael Zeng | Yue Zhang

pdf bib abs **Code Generation From Flowcharts with Texts: A Benchmark Dataset and An Approach**
Zejie Liu | Xiaoyu Hu | Deyu Zhou | Lin Li | Xu Zhang | Yanzheng Xiang

pdf bib abs **Focus! Relevant and Sufficient Context Selection for News Image Captioning**
Mingyang Zhou | Grace Luo | Anna Rohrbach | Zhou Yu

pdf bib abs **Generative Aspect-Based Sentiment Analysis with Contrastive Learning and Expressive Structure**
Joseph Peper | Lu Wang

pdf bib abs **Semantic Dependency Parsing with Edge GNNs**
Songlin Yang | Kewei Tu

pdf bib abs **Explore Unsupervised Structures in Pretrained Models for Relation Extraction**
Xi Yang | Tao Ji | Yuanbin Wu

pdf bib abs **Identifying Human Strategies for Generating Word-Level Adversarial Examples**
Maximilian Mozes | Bennett Kleinberg | Lewis Griffin

pdf bib abs **Refinement Matters: Textual Description Needs to be Refined for Zero-shot Learning**
Chandan Gautam | Sethupathy Parameswaran | Vinay Verma | Suresh Sundaram | Savitha Ramasamy

pdf bib abs **SAT: Improving Semi-Supervised Text Classification with Simple Instance-Adaptive Self-Training**
Hui Chen | Wei Han | Soujanya Poria

pdf bib abs **Answer Quality Aware Aggregation for Extractive QA Crowdsourcing**
Peide Zhu | Zhen Wang | Claudia Hauff | Jie Yang | Avishek Anand

pdf bib abs **Search to Pass Messages for Temporal Knowledge Graph Completion**
Zhen Wang | Haotong Du | Quanming Yao | Xuelong Li

pdf bib abs **Code Vulnerability Detection via Nearest Neighbor Mechanism**
Qianjin Du | Xiaohui Kuang | Gang Zhao

pdf bib abs **Robust Question Answering against Distribution Shifts with Test-Time Adaption: An Empirical Study**
Hai Ye | Yuyang Ding | Juntao Li | Hwee Tou Ng

pdf bib abs **ParaMac: A General Unsupervised Paraphrase Generation Framework Leveraging Semantic Constraints and Diversifying Mechanisms**
Jinxin Liu | Jiaxin Shi | Ji Qi | Lei Hou | Juanzi Li | Qi Tian

pdf bib abs **Semi-supervised New Slot Discovery with Incremental Clustering**
Yuxia Wu | Lizi Liao | Xueming Qian | Tat-Seng Chua

pdf bib abs **Con-NAT: Contrastive Non-autoregressive Neural Machine Translation**
Hao Cheng | Zhihua Zhang

pdf bib abs **Improved Knowledge Distillation for Pre-trained Language Models via Knowledge Selection**
Chenglong Wang | Yi Lu | Yongyu Mu | Yimin Hu | Tong Xiao | Jingbo Zhu

pdf bib abs **Syntactically Robust Training on Partially-Observed Data for Open Information Extraction**
Ji Qi | Yuxiang Chen | Lei Hou | Juanzi Li | Bin Xu

pdf bib abs **A Benchmark and Dataset for Post-OCR text correction in Sanskrit**
Ayush Maheshwari | Nikhil Singh | Amrith Krishna | Ganesh Ramakrishnan

pdf bib abs **Knowledge-Enhanced Self-Supervised Prototypical Network for Few-Shot Event Detection**
Kailin Zhao | Xiaolong Jin | Long Bai | Jiafeng Guo | Xueqi Cheng

pdf bib abs **VarMAE: Pre-training of Variational Masked Autoencoder for Domain-adaptive Language Understanding**
Dou Hu | Xiaolong Hou | Xiyang Du | Mengyuan Zhou | Lianxin Jiang | Yang Mo | Xiaofeng Shi

pdf bib abs **Exploring Methods for Building Dialects-Mandarin Code-Mixing Corpora: A Case Study in Taiwanese Hokkien**
Sin-En Lu | Bo-Han Lu | Chao-Yi Lu | Richard Tzong-Han Tsai

pdf bib abs **Recurrence Boosts Diversity! Revisiting Recurrent Latent Variable in Transformer-Based Variational AutoEncoder for Diverse Text Generation**
Jinyi Hu | Xiaoyuan Yi | Wenhao Li | Maosong Sun | Xing Xie

pdf bib abs **Tweet Based Reach Aware Temporal Attention Network for NFT Valuation**
Ramit Sawhney | Megh Thakkar | Ritesh Soun | Atula Neerkaje | Vasu Sharma | Dipanwita Guhathakurta | Sudheer Chava

pdf bib abs **Entity Embedding Completion for Wide-Coverage Entity Disambiguation**
Daisuke Oba | Ikuya Yamada | Naoki Yoshinaga | Masashi Toyoda

pdf bib abs **Entity-level Interaction via Heterogeneous Graph for Multimodal Named Entity Recognition**
Gang Zhao | Guanting Dong | Yidong Shi | Haolong Yan | Weiran Xu | Si Li

pdf bib abs **Status Biases in Deliberation Online: Evidence from a Randomized Experiment on ChangeMyView**
Emaad Manzoor | Yohan Jo | Alan Montgomery

pdf bib abs **Empathetic and Emotionally Positive Conversation Systems with an Emotion-specific Query-Response Memory**
Zhiliang Tian | Yinliang Wang | Yiping Song | Chi Zhang | Dongkyu Lee | Yingxiu Zhao | Dongsheng Li | Nevin L. Zhang

pdf bib abs **Trial2Vec: Zero-Shot Clinical Trial Document Similarity Search using Self-Supervision**
Zifeng Wang | Jimeng Sun

pdf bib abs **From Mimicking to Integrating: Knowledge Integration for Pre-trained Language Models**
Lei Li | Yankai Lin | Xuancheng Ren | Guangxiang Zhao | Peng Li | Jie Zhou | Xu Sun

Lei Li | Yankai Lin | Xuancheng Ren | Guangxiang Zhao | Peng Li | Jie Zhou | Xu Sun

pdf bib abs **Model and Data Transfer for Cross-Lingual Sequence Labelling in Zero-Resource Settings**
Iker García-Ferrero | Rodrigo Agerri | German Rigau

pdf bib abs **Early Guessing for Dialect Identification**
Vani Kanjirangat | Tanja Samardzic | Fabio Rinaldi | Ljiljana Dolamic

pdf bib abs **R-AT: Regularized Adversarial Training for Natural Language Understanding**
Shiwen Ni | Jiawen Li | Hung-Yu Kao

pdf bib abs **Multi-View Active Learning for Short Text Classification in User-Generated Data**
Payam Karisani | Negin Karisani | Li Xiong

pdf bib abs **Forging Multiple Training Objectives for Pre-trained Language Models via Meta-Learning**
Hongqiu Wu | Ruixue Ding | Hai Zhao | Boli Chen | Pengjun Xie | Fei Huang | Min Zhang

pdf bib abs **ConGen: Unsupervised Control and Generalization Distillation For Sentence Representation**
Peerat Limkonchotiwat | Wuttikorn Ponwitayarat | Lalita Lowphansirikul | Can Udomcharoenchaikit | Ekapol Chuangsuwanich | Sarana Nutanong

pdf bib abs **Large-Scale Differentially Private BERT**
Rohan Anil | Badih Ghazi | Vineet Gupta | Ravi Kumar | Pasin Manurangsi

pdf bib abs **Improving Zero-Shot Multilingual Translation with Universal Representations and Cross-Mapping**
Shuhao Gu | Yang Feng

pdf bib abs **Controllable Fake Document Infilling for Cyber Deception**
Yibo Hu | Yu Lin | Erick Skorupa Parolin | Latifur Khan | Kevin Hamlen

pdf bib abs **Weakly Supervised Headline Dependency Parsing**
Adrian Benton | Tianze Shi | Ozan İrsoy | Igor Malioutov

pdf bib abs **BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization**
Wojciech Kryscinski | Nazneen Rajani | Divyansh Agarwal | Caiming Xiong | Dragomir Radev

pdf bib abs **Not All Errors are Equal: Learning Text Generation Metrics using Stratified Error Synthesis**
Wenda Xu | Yi-Lin Tuan | Yujie Lu | Michael Saxon | Lei Li | William Yang Wang

pdf bib abs **Summarization as Indirect Supervision for Relation Extraction**
Keming Lu | I-Hung Hsu | Wenxuan Zhou | Mingyu Derek Ma | Muhao Chen

pdf bib abs **DIGAT: Modeling News Recommendation with Dual-Graph Interaction**
Zhiming Mao | Jian Li | Hongru Wang | Xingshan Zeng | Kam-Fai Wong

pdf bib abs **SMASH: Improving SMAll Language Models' Few-SHot Ability with Prompt-Based Distillation**
Yueqian Wang | Chang Liu | Kai Chen | Xi Wang | Dongyan Zhao

pdf bib abs **Consecutive Question Generation via Dynamic Multitask Learning**
Yunji Li | Sujian Li | Xing Shi

pdf bib abs **Subword Segmental Language Modelling for Nguni Languages**
Francois Meyer | Jan Buys

pdf bib abs **Towards Robust Visual Question Answering: Making the Most of Biased Samples via Contrastive Learning**
Qingyi Si | Yuanxin Liu | Fandong Meng | Zheng Lin | Peng Fu | Yanan Cao | Weiping Wang | Jie Zhou

pdf bib abs **P3LM: Probabilistically Permuted Prophet Language Modeling for Generative Pre-Training**
Junwei Bao | Yifan Wang | Ying Jiangyong | Yeyun Gong | Jing Zhao | Youzheng Wu | Xiaodong He

pdf bib abs **Holistic Sentence Embeddings for Better Out-of-Distribution Detection**
Sishuo Chen | Xiaohan Bi | Rundong Gao | Xu Sun

pdf bib abs **MuGER2: Multi-Granularity Evidence Retrieval and Reasoning for Hybrid Question Answering**
Yingyao Wang | Junwei Bao | Chaoqun Duan | Youzheng Wu | Xiaodong He | Tiejun Zhao

pdf bib abs **EntityCS: Improving Zero-Shot Cross-lingual Transfer with Entity-Centric Code Switching**
Chenxi Whitehouse | Fenia Christopoulou | Ignacio Iacobacci

pdf bib abs **MBTI Personality Prediction for Fictional Characters Using Movie Scripts**
Yisi Sang | Xiangyang Mou | Mo Yu | Dakuo Wang | Jing Li | Jeffrey Stanton

pdf bib abs **A Simple and Strong Baseline for End-to-End Neural RST-style Discourse Parsing**
Naoki Kobayashi | Tsutomu Hirao | Hidetaka Kamigaito | Manabu Okumura | Masaaki Nagata

pdf bib abs **Probing for Constituency Structure in Neural Language Models**
David Arps | Younes Samih | Laura Kallmeyer | Hassan Sajjad

pdf bib abs **Table-To-Text generation and pre-training with TabT5**
Ewa Andrejczuk | Julian Eisenschlos | Francesco Piccinno | Syrine Krichene | Yasemin Altun

pdf bib abs **A POMDP Dialogue Policy with 3-way Grounding and Adaptive Sensing for Learning through Communication**
Maryam Zare | Alan Wagner | Rebecca Passonneau

pdf bib abs **PaCo: Preconditions Attributed to Commonsense Knowledge**
Ehsan Qasemi | Filip Ilievski | Muhao Chen | Pedro Szekely

pdf bib abs **Improving Few-Shot Domain Transfer for Named Entity Disambiguation with Pattern Exploitation**
Philip Blair | Kfir Bar

pdf bib abs **Capturing Topic Framing via Masked Language Modeling**
Xiaobo Guo | Weicheng Ma | Soroush Vosoughi

pdf bib abs **WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation**
Alisa Liu | Swabha Swayamdipta | Noah A. Smith | Yejin Choi

pdf bib abs **Sequentially Controlled Text Generation**
Alexander Spangher | Yao Ming | Xinyu Hua | Nanyun Peng

pdf bib abs **Revisiting the Roles of "Text" in Text Games**
Yi Gu | Shunyu Yao | Chuang Gan | Josh Tenenbaum | Mo Yu

pdf bib abs **FPT: Improving Prompt Tuning Efficiency via Progressive Training**
Yufei Huang | Yujia Qin | Huadong Wang | Yichun Yin | Maosong Sun | Zhiyuan Liu | Qun Liu

pdf bib abs **Prompt-learning for Fine-grained Entity Typing**
Ning Ding | Yulin Chen | Xu Han | Guangwei Xu | Xiaobin Wang | Pengjun Xie | Haitao Zheng | Zhiyuan Liu | Juanzi Li | Hong-Gee Kim

pdf bib abs **TransLIST: A Transformer-Based Linguistically Informed Sanskrit Tokenizer**
Jivnesh Sandhan | Rathin Singha | Narein Rao | Suvendu Samanta | Laxmidhar Behera | Pawan Goyal

pdf bib abs **Fair NLP Models with Differentially Private Text Encoders**
Gaurav Maheshwari | Pascal Denis | Mikaela Keller | Aurélien Bellet

pdf bib abs **Modeling Context With Linear Attention for Scalable Document-Level Translation**
Zhaofeng Wu | Hao Peng | Nikolaos Pappas | Noah A. Smith

pdf bib abs **What do Large Language Models Learn beyond Language?**
Avinash Madasu | Shashank Srivastava

pdf bib abs **CONSISTENT: Open-Ended Question Generation From News Articles**
Tuhin Chakrabarty | Justin Lewis | Smaranda Muresan

pdf bib abs **Efficient (Soft) Q-Learning for Text Generation with Limited Good Data**
Han Guo | Bowen Tan | Zhengzhong Liu | Eric Xing | Zhiting Hu

pdf bib abs **Lexi: Self-Supervised Learning of the UI Language**
Pratyay Banerjee | Shweti Mahajan | Kushal Arora | Chitta Baral | Oriana Riva

pdf bib abs **Inferring the Reader: Guiding Automated Story Generation with Commonsense Reasoning**
Xiangyu Peng | Siyan Li | Sarah Wiegreffe | Mark Riedl

pdf bib abs **How to Stop an Avalanche? JoDeM: Joint Decision Making through Compare and Contrast for Dialog State Tracking**
Haoming Wang | Wang Xin

pdf bib abs **Contrastive Learning with Prompt-derived Virtual Semantic Prototypes for Unsupervised Sentence Embedding**
Jiali Zeng | Yongjing Yin | Yufan Jiang | Shuangzhi Wu | Yunbo Cao

pdf bib abs **Weight Perturbation as Defense against Adversarial Word Substitutions**
Jianhan Xu | Linyang Li | Jiping Zhang | Xiaoqing Zheng | Kai-Wei Chang | Cho-Jui Hsieh | Xuanjing Huang

pdf bib abs **CORT: A New Baseline for Comparative Opinion Classification by Dual Prompts**
Yequan Wang | Hengran Zhang | Aixin Sun | Xuying Meng

pdf bib abs **APEACH: Attacking Pejorative Expressions with Analysis on Crowd-Generated Hate Speech Evaluation Datasets**
Kichang Yang | Wonjun Jang | Won Ik Cho

pdf bib abs **Guiding Neural Story Generation with Reader Models**

Xiangyu Peng | Kaige Xie | Amal Alabdulkarim | Harshith Kayam | Samihan Dani | Mark Riedl

pdf bib abs **Reason first, then respond: Modular Generation for Knowledge-infused Dialogue**
Leonard Adolphs | Kurt Shuster | Jack Urbanek | Arthur Szlam | Jason Weston

pdf bib abs **Adapting Multilingual Models for Code-Mixed Translation**
Aditya Vavre | Abhirut Gupta | Sunita Sarawagi

pdf bib abs **LPC: A Logits and Parameter Calibration Framework for Continual Learning**
Xiaodi Li | Zhuoyi Wang | Dingcheng Li | Latifur Khan | Bhavani Thuraisingham

pdf bib abs **SlovakBERT: Slovak Masked Language Model**
Matúš Pikuliak | Štefan Grivalský | Martin Konôpka | Miroslav Blšták | Martin Tamajka | Viktor Bachratý | Marian Simko | Pavol Balážik | Michal Trnka | Filip Uhlárik

pdf bib abs **Efficient Zero-shot Event Extraction with Context-Definition Alignment**
Hongming Zhang | Wenlin Yao | Dong Yu

pdf bib abs **Logical Fallacy Detection**
Zhijing Jin | Abhinav Lalwani | Tejas Vaidhya | Xiaoyu Shen | Yiwen Ding | Zhiheng Lyu | Mrinmaya Sachan | Rada Mihalcea | Bernhard Schoelkopf

pdf bib abs **Topic-Aware Response Generation in Task-Oriented Dialogue with Unstructured Knowledge Access**
Yue Feng | Gerasimos Lampouras | Ignacio Iacobacci

pdf bib abs **Revisiting Transformer-based Models for Long Document Classification**
Xiang Dai | Ilias Chalkidis | Sune Darkner | Desmond Elliott

pdf bib abs **Time-aware Prompting for Text Generation**
Shuyang Cao | Lu Wang

pdf bib abs **Improving Scheduled Sampling with Elastic Weight Consolidation for Neural Machine Translation**
Michalis Korakakis | Andreas Vlachos

pdf bib abs **Ensemble Transformer for Efficient and Accurate Ranking Tasks: an Application to Question Answering Systems**
Yoshitomo Matsubara | Luca Soldaini | Eric Lind | Alessandro Moschitti

pdf bib abs **Uncertainty Quantification with Pre-trained Language Models: A Large-Scale Empirical Analysis**
Yuxin Xiao | Paul Pu Liang | Umang Bhatt | Willie Neiswanger | Ruslan Salakhutdinov | Louis-Philippe Morency

pdf bib abs **How to Represent Context Better? An Empirical Study on Context Modeling for Multi-turn Response Selection**
Jiazhan Feng | Chongyang Tao | Chang Liu | Rui Yan | Dongyan Zhao

pdf bib abs **CHIA: CHoosing Instances to Annotate for Machine Translation**
Rajat Bhatnagar | Ananya Ganesh | Katharina Kann

pdf bib abs **Guiding Neural Machine Translation with Semantic Kernels**
Ping Guo | Yue Hu | Xiangpeng Wei | Yubing Ren | Yunpeng Li | Luxi Xing | Yuqiang Xie

pdf bib abs **HiSMatch: Historical Structure Matching based Temporal Knowledge Graph Reasoning**
Zixuan Li | Zhongni Hou | Saiping Guan | Xiaolong Jin | Weihua Peng | Long Bai | Yajuan Lyu | Wei Li | Jiafeng Guo | Xueqi Cheng

pdf bib abs **Dependency Parsing via Sequence Generation**
Boda Lin | Zijun Yao | Jiaxin Shi | Shulin Cao | Binghao Tang | Si Li | Yong Luo | Juanzi Li | Lei Hou

pdf bib abs **Scaling Laws Under the Microscope: Predicting Transformer Performance from Small Scale Experiments**
Maor Ivgi | Yair Carmon | Jonathan Berant

pdf bib abs **Analyzing the Limits of Self-Supervision in Handling Bias in Language**
Lisa Bauer | Karthik Gopalakrishnan | Spandana Gella | Yang Liu | Mohit Bansal | Dilek Hakkani-Tur

pdf bib abs **Multiple Instance Learning for Offensive Language Detection**
Jiexi Liu | Dehan Kong | Longtao Huang | Dinghui Mao | Hui Xue

pdf bib abs **Grounded Keys-to-Text Generation: Towards Factual Open-Ended Generation**
Faeze Brahman | Baolin Peng | Michel Galley | Sudha Rao | Bill Dolan | Snigdha Chaturvedi | Jianfeng Gao

↑ up

pdf (full) bib (full) Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP

pdf bib **Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP**

Jasmijn Bastings | Yonatan Belinkov | Yanai Elazar | Dieuwke Hupkes | Naomi Saphra | Sarah Wiegreffe

pdf bib abs **A Minimal Model for Compositional Generalization on gSCAN**
Alice Hein | Klaus Diepold

pdf bib abs **Sparse Interventions in Language Models with Differentiable Masking**
Nicola De Cao | Leon Schmid | Dieuwke Hupkes | Ivan Titov

pdf bib abs **Where's the Learning in Representation Learning for Compositional Semantics and the Case of Thematic Fit**
Mughilan Muthupari | Samrat Halder | Asad Sayeed | Yuval Marton

pdf bib abs **Sentence Ambiguity, Grammaticality and Complexity Probes**
Sunit Bhattacharya | Vilém Zouhar | Ondrej Bojar

pdf bib abs **Post-Hoc Interpretation of Transformer Hyperparameters with Explainable Boosting Machines**
Kiron Deb | Xuan Zhang | Kevin Duh

pdf bib abs **Revisit Systematic Generalization via Meaningful Learning**
Ning Shi | Boxin Wang | Wei Wang | Xiangyu Liu | Zhouhan Lin

pdf bib abs **Is It Smaller Than a Tennis Ball? Language Models Play the Game of Twenty Questions**
Maxime De Bruyn | Ehsan Lotfi | Jeska Buhmann | Walter Daelemans

pdf bib abs **Post-hoc analysis of Arabic transformer models**
Ahmed Abdelali | Nadir Durrani | Fahim Dalvi | Hassan Sajjad

pdf bib abs **Universal Evasion Attacks on Summarization Scoring**
Wenchuan Mu | Kwan Hui Lim

pdf bib abs **How (Un)Faithful is Attention?**
Hessam Amini | Leila Kosseim

pdf bib abs **Are Multilingual Sentiment Models Equally Right for the Right Reasons?**
Rasmus Jørgensen | Fiammetta Caccavale | Christian Igel | Anders Søgaard

pdf bib abs **Probing for Understanding of English Verb Classes and Alternations in Large Pre-trained Language Models**
David Yi | James Bruno | Jiayu Han | Peter Zukerman | Shane Steinert-Threlkeld

pdf bib abs **Analyzing Gender Translation Errors to Identify Information Flows between the Encoder and Decoder of a NMT System**
Guillaume Wisniewski | Lichao Zhu | Nicolas Ballier | François Yvon

pdf bib abs **Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions**
Jenny Kunz | Martin Jirenius | Oskar Holmström | Marco Kuhlmann

pdf bib abs **Analyzing the Representational Geometry of Acoustic Word Embeddings**
Badr Abdullah | Dietrich Klakow

pdf bib abs **Understanding Domain Learning in Language Models Through Subpopulation Analysis**
Zheng Zhao | Yftah Ziser | Shay Cohen

pdf bib abs **Intermediate Entity-based Sparse Interpretable Representation Learning**
Diego Garcia-Olano | Yasumasa Onoe | Joydeep Ghosh | Byron Wallace

pdf bib abs **Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information**
Isar Nejadgholi | Esma Balkir | Kathleen Fraser | Svetlana Kiritchenko

pdf bib abs **Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?**
Digvijay Ingle | Rishabh Tripathi | Ayush Kumar | Kevin Patel | Jithendra Vepa

pdf bib abs **It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark**
Teemu Vahtola | Mathias Creutz | Jörg Tiedemann

pdf bib abs **Controlling for Stereotypes in Multimodal Language Model Evaluation**
Manuj Malik | Richard Johansson

pdf bib abs **On the Compositional Generalization Gap of In-Context Learning**
Arian Hosseini | Ankit Vani | Dzmitry Bahdanau | Alessandro Sordoni | Aaron Courville

pdf bib abs **Explaining Translationese: why are Neural Classifiers Better and what do they Learn?**
Kwabena Amponsah-Kaakyire | Daria Pylypenko | Josef Genabith | Cristina España-Bonet

pdf bib abs **Probing GPT-3's Linguistic Knowledge on Semantic Tasks**
Lining Zhang | Mengchen Wang | Liben Chen | Wenxin Zhang

pdf bib abs **Garden Path Traversal in GPT-2**
William Jurayj | William Rudman | Carsten Eickhoff

pdf bib abs **Testing Pre-trained Language Models' Understanding of Distributivity via Causal Mediation Analysis**
Pangbo Ban | Yifan Jiang | Tianran Liu | Shane Steinert-Threlkeld

pdf bib abs **Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence**
Jingcheng Niu | Wenjie Lu | Eric Corlett | Gerald Penn

pdf bib abs **DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models**
Royi Rassin | Shauli Ravfogel | Yoav Goldberg

pdf bib abs **Practical Benefits of Feature Feedback Under Distribution Shift**
Anurag Katakkar | Clay H. Yoo | Weiqin Wang | Zachary Lipton | Divyansh Kaushik

pdf bib abs **Identifying the Source of Vulnerability in Explanation Discrepancy: A Case Study in Neural Text Classification**
Ruixuan Tang | Hanjie Chen | Yangfeng Ji

pdf bib abs **Probing Pretrained Models of Source Codes**
Sergey Troshin | Nadezhda Chirkova

pdf bib abs **Probing the representations of named entities in Transformer-based Language Models**
Stefan Schouten | Peter Bloem | Piek Vossen

pdf bib abs **Decomposing Natural Logic Inferences for Neural NLI**
Julia Rozanova | Deborah Ferreira | Mokanarangan Thayaparan | Marco Valentino | Andre Freitas

pdf bib abs **Probing with Noise: Unpicking the Warp and Weft of Embeddings**
Filip Klubicka | John Kelleher

pdf bib abs **Look to the Right: Mitigating Relative Position Bias in Extractive Question Answering**
Kazutoshi Shinoda | Saku Sugawara | Akiko Aizawa

pdf bib abs **A Continuum of Generation Tasks for Investigating Length Bias and Degenerate Repetition**
Darcey Riley | David Chiang

pdf bib abs **Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation**
Oleg Serikov | Vitaly Protasov | Ekaterina Voloshina | Viktoria Knyazkova | Tatiana Shavrina

↑ up

pdf (full) bib (full) Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)

pdf bib **Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)**
Ali Hürriyetoğlu | Hristo Tanev | Vanni Zavarella | Erdem Yörük

pdf bib abs **A Multi-Modal Dataset for Hate Speech Detection on Social Media: Case-study of Russia-Ukraine Conflict**
Surendrabikram Thapa | Aditya Shah | Farhan Jafri | Usman Naseem | Imran Razzak

pdf bib abs **EventGraph: Event Extraction as Semantic Graph Parsing**
Huiling You | David Samuel | Samia Touileb | Lilja Øvrelid

pdf bib abs **NLP4ITF @ Causal News Corpus 2022: Leveraging Linguistic Information for Event Causality Classification**
Theresa Krumbiegel | Sophie Decher

pdf bib abs **A Hybrid Knowledge and Transformer-Based Model for Event Detection with Automatic Self-Attention Threshold, Layer and Head Selection**
Thierry Desot | Orphee De Clercq | Veronique Hoste

pdf bib abs **Improving Zero-Shot Event Extraction via Sentence Simplification**
Sneha Mehta | Huzefa Rangwala | Naren Ramakrishnan

pdf bib abs **SNU-Causality Lab @ Causal News Corpus 2022: Detecting Causality by Data Augmentation via Part-of-Speech tagging**
Juhyeon Kim | Yesong Choe | Sanghack Lee

pdf bib abs **LTRC @ Causal News Corpus 2022: Extracting and Identifying Causal Elements using Adapters**
Hiranmai Sri Adibhatla | Manish Shrivastava

pdf bib abs **Cross-modal Transfer Between Vision and Language for Protest Detection**
Ria Raj | Kajsa Andreasson | Tobias Norlund | Richard Johansson | Aron Lagerberg

pdf bib abs **IDIAPers @ Causal News Corpus 2022: Efficient Causal Relation Identification Through a Prompt-based Few-shot Approach**
Sergio Burdisso | Juan Pablo Zuluaga-gomez | Esau Villatoro-tello | Martin Fajcik | Muskaan Singh | Pavel Smrz | Petr Motlicek

pdf bib abs **IDIAPers @ Causal News Corpus 2022: Extracting Cause-Effect-Signal Triplets via Pre-trained Autoregressive Language Model**
Martin Fajcik | Muskaan Singh | Juan Pablo Zuluaga-gomez | Esau Villatoro-tello | Sergio Burdisso | Petr Motlicek | Pavel Smrz

pdf bib abs **NoisyAnnot@ Causal News Corpus 2022: Causality Detection using Multiple Annotation Decisions**
Quynh Anh Nguyen | Arka Mitra

pdf bib abs **GGNN@Causal News Corpus 2022:Gated Graph Neural Networks for Causal Event Classification from Social-Political News Articles**
Paul Trust | Rosane Minghim | Evangelos Milos | Kadusabe Provia

pdf bib abs **1Cademy @ Causal News Corpus 2022: Leveraging Self-Training in Causality Classification of Socio-Political Event Data**
Adam Nik | Ge Zhang | Xingran Chen | Mingyu Li | Jie Fu

pdf bib abs **1Cademy @ Causal News Corpus 2022: Enhance Causal Span Detection via Beam-Search-based Position Selector**
Xingran Chen | Ge Zhang | Adam Nik | Mingyu Li | Jie Fu

pdf bib abs **Hybrid Knowledge Engineering Leveraging a Robust ML Framework to Produce an Assassination Dataset**
Abigail Sticha | Paul Brenner

pdf bib abs **Political Event Coding as Text-to-Text Sequence Generation**
Yaoyao Dai | Benjamin Radford | Andrew Halterman

pdf bib abs **Zero-Shot Ranking Socio-Political Texts with Transformer Language Models to Reduce Close Reading Time**
Kiymet Akdemir | Ali Hürriyetoğlu

pdf bib abs **SPOCK @ Causal News Corpus 2022: Cause-Effect-Signal Span Detection Using Span-Based and Sequence Tagging Models**
Anik Saha | Alex Gittens | Jian Ni | Oktie Hassanzadeh | Bulent Yener | Kavitha Srinivas

pdf bib abs **CSECU-DSG @ Causal News Corpus 2022: Fusion of RoBERTa Transformers Variants for Causal Event Classification**
Abdul Aziz | Md. Akram Hossain | Abu Nowshed Chy

pdf bib abs **ARGUABLY @ Causal News Corpus 2022: Contextually Augmented Language Models for Event Causality Identification**
Guneet Kohli | Prabsimran Kaur | Jatin Bedi

pdf bib abs **ClassBases at the CASE-2022 Multilingual Protest Event Detection Task: Multilingual Protest News Detection and Automatically Replicating Manually Created Event Datasets**
Peratham Wiriyathammabhum

pdf bib abs **EventGraph at CASE 2021 Task 1: A General Graph-based Approach to Protest Event Extraction**
Huiling You | David Samuel | Samia Touileb | Lilja Øvrelid

pdf bib abs **NSUT-NLP at CASE 2022 Task 1: Multilingual Protest Event Detection using Transformer-based Models**
Manan Suri | Krish Chopra | Adwita Arora

pdf bib abs **CamPros at CASE 2022 Task 1: Transformer-based Multilingual Protest News Detection**
Neha Kumari | Mrinal Anand | Tushar Mohan | Ponnurangam Kumaraguru | Arun Balaji Buduru

pdf bib abs **ARC-NLP at CASE 2022 Task 1: Ensemble Learning for Multilingual Protest Event Detection**
Umitcan Sahin | Oguzhan Ozcelik | Izzet Emre Kucukkaya | Cagri Toraman

pdf bib abs **CEIA-NLP at CASE 2022 Task 1: Protest News Detection for Portuguese**
Diogo Fernandes | Adalberto Junior | Gabriel Marques | Anderson Soares | Arlindo Galvao Filho

pdf bib abs **SPARTA at CASE 2021 Task 1: Evaluating Different Techniques to Improve Event Extraction**
Arthur Müller | Andreas Dafnos

pdf bib abs **Event Causality Identification with Causal News Corpus - Shared Task 3, CASE 2022**
Fiona Anting Tan | Hansi Hettiarachchi | Ali Hürriyetoğlu | Tommaso Caselli | Onur Uca | Farhana Ferdousi Liza | Nelleke Oostdijk

pdf bib abs **Tracking COVID-19 protest events in the United States. Shared Task 2: Event Database Replication, CASE 2022**
Vanni Zavarella | Hristo Tanev | Ali Hürriyetoğlu | Peratham Wiriyathammabhum | Bertrand De Longueville

pdf bib abs **Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022): Workshop and Shared Task Report**
Ali Hürriyetoğlu | Hristo Tanev | Vanni Zavarella | Reyyan Yeniterzi | Osman Mutlu | Erdem Yörük

pdf bib abs **Extended Multilingual Protest News Detection - Shared Task 1, CASE 2021 and 2022**
Ali Hürriyetoğlu | Osman Mutlu | Fırat Duruşan | Onur Uca | Alaeddin Gürel | Benjamin J. Radford | Yaoyao Dai | Hansi Hettiarachchi | Niklas Stoehr | Tadashi Nomoto | Milena Slavcheva | Francielle Vargas | Pagé David | Fatih Beyhan | Erdem Yörük

↑ up

pdf (full)
bib (full)

Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)

pdf bib **Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)**
Antske Fokkens | Vivek Srikumar

pdf bib abs **A Multilingual Bag-of-Entities Model for Zero-Shot Cross-Lingual Text Classification**
Sosuke Nishikawa | Ikuya Yamada | Yoshimasa Tsuruoka | Isao Echizen

pdf bib abs **Collateral facilitation in humans and language models**
James Michaelov | Benjamin Bergen

pdf bib abs **How Hate Speech Varies by Target Identity: A Computational Analysis**
Michael Yoder | Lynnette Ng | David West Brown | Kathleen Carley

pdf bib abs **Continual Learning for Natural Language Generations with Transformer Calibration**
Peng Yang | Dingcheng Li | Ping Li

pdf bib abs **That's so cute!: The CARE Dataset for Affective Response Detection**
Jane Yu | Alon Halevy

pdf bib abs **A Fine-grained Interpretability Evaluation Benchmark for Neural NLP**
Lijie Wang | Yaozong Shen | Shuyuan Peng | Shuai Zhang | Xinyan Xiao | Hao Liu | Hongxuan Tang | Ying Chen | Hua Wu | Haifeng Wang

pdf bib abs **Towards More Natural Artificial Languages**
Mark Hopkins

pdf bib abs **Causal Analysis of Syntactic Agreement Neurons in Multilingual Language Models**
Aaron Mueller | Yu Xia | Tal Linzen

pdf bib abs **Combining Noisy Semantic Signals with Orthographic Cues: Cognate Induction for the Indic Dialect Continuum**
Niyati Bafna | Josef van Genabith | Cristina España-Bonet | Zdeněk Žabokrtský

pdf bib abs **Detecting Unintended Social Bias in Toxic Language Datasets**
Nihar Sahoo | Himanshu Gupta | Pushpak Bhattacharyya

pdf bib abs **Incremental Processing of Principle B: Mismatches Between Neural Models and Humans**
Forrest Davis

pdf bib abs **Parsing as Deduction Revisited: Using an Automatic Theorem Prover to Solve an SMT Model of a Minimalist Parser**
Sagar Indurkhya

pdf bib abs **Entailment Semantics Can Be Extracted from an Ideal Language Model**
William Merrill | Alex Warstadt | Tal Linzen

pdf bib abs **On Neurons Invariant to Sentence Structural Changes in Neural Machine Translation**
Gal Patel | Leshem Choshen | Omri Abend

pdf bib abs **Shared knowledge in natural conversations: can entropy metrics shed light on information transfers?**
Eliot Maës | Philippe Blache | Leonor Becerra

pdf bib abs **Leveraging a New Spanish Corpus for Multilingual and Cross-lingual Metaphor Detection**
Elisa Sanchez-Bayona | Rodrigo Agerri

pdf bib abs **Cognitive Simplification Operations Improve Text Simplification**
Eytan Chamovitz | Omri Abend

pdf bib abs **On Language Spaces, Scales and Cross-Lingual Transfer of UD Parsers**
Tanja Samardžić | Ximena Gutierrez-Vasques | Rob van der Goot | Max Müller-Eberstein | Olga Pelloni | Barbara Plank

pdf bib abs **Visual Semantic Parsing: From Images to Abstract Meaning Representation**
Mohamed Ashraf Abdelsalam | Zhan Shi | Federico Fancellu | Kalliopi Basioti | Dhaivat Bhatt | Vladimir Pavlovic | Afsaneh Fazly

pdf bib abs **Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities**
Suhas Arehalli | Brian Dillon | Tal Linzen

pdf bib abs **OpenStance: Real-world Zero-shot Stance Detection**
Hanzi Xu | Slobodan Vucetic | Wenpeng Yin

pdf bib abs **Optimizing text representations to capture (dis)similarity between political parties**
Tanise Ceron | Nico Blokker | Sebastian Padó

pdf bib abs **Computational cognitive modeling of predictive sentence processing in a second language**
Umesh Patil | Sol Lago

pdf bib abs **PIE-QG: Paraphrased Information Extraction for Unsupervised Question Generation from Small Corpora**
Dinesh Nagumothu | Bahadorreza Ofoghi | Guangyan Huang | Peter Eklund

pdf bib abs **Probing for targeted syntactic knowledge through grammatical error detection**
Christopher Davis | Christopher Bryant | Andrew Caines | Marek Rei | Paula Buttery

pdf bib abs **An Alignment-based Approach to Text Segmentation Similarity Scoring**
Gerardo Ocampo Diaz | Jessica Ouyang

pdf bib abs **Enhancing the Transformer Decoder with Transition-based Syntax**
Leshem Choshen | Omri Abend

pdf bib abs **Characterizing Verbatim Short-Term Memory in Neural Language Models**
Kristijan Armeni | Christopher Honey | Tal Linzen

↑ up

pdf (full) bib (full) Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP)

pdf bib **Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP)**
Francesco Barbieri | Jose Camacho-Collados | Bhuwan Dhingra | Luis Espinosa-Anke | Elena Gribovskaya | Angeliki Lazaridou | Daniel Loureiro | Leonardo Neves

pdf bib abs **MLLabs-LIG at TempoWiC 2022: A Generative Approach for Examining Temporal Meaning Shift**
Chenyang Lyu | Yongxin Zhou | Tianbo Ji

pdf bib abs **Using Deep Mixture-of-Experts to Detect Word Meaning Shift for TempoWiC**
Ze Chen | Kangxu Wang | Zijian Cai | Jiewen Zheng | Jiarong He | Max Gao | Jason Zhang

pdf bib abs **Using Two Losses and Two Datasets Simultaneously to Improve TempoWiC Accuracy**
Mohammad Javad Pirhadi | Motahhare Mirzaei | Sauleh Eetemadi

pdf bib abs **Class Incremental Learning for Intent Classification with Limited or No Old Data**
Debjit Paul | Daniil Sorokin | Judith Gaspers

pdf bib abs **CC-Top: Constrained Clustering for Dynamic Topic Discovery**
Jann Goschenhofer | Pranav Ragupathy | Christian Heumann | Bernd Bischl | Matthias Aßenmacher

pdf bib abs **HSE at TempoWiC: Detecting Meaning Shift in Social Media with Diachronic Language Models**
Elizaveta Tukhtina | Kseniia Kashleva | Svetlana Vydrina

pdf bib abs **Leveraging time-dependent lexical features for offensive language detection**
Barbara McGillivray | Malithi Alahapperuma | Jonathan Cook | Chiara Di Bonaventura | Albert Meroño-Peñuela | Gareth Tyson | Steven Wilson

pdf bib abs **Temporal Word Meaning Disambiguation using TimeLMs**
Mihir Godbole | Parth Dandavate | Aditya Kane

↑ up

pdf (full) bib (full) Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)

pdf bib **Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)**
Chung-Chi Chen | Hen-Hsen Huang | Hiroya Takamura | Hsin-Hsi Chen

pdf bib abs **Contextualizing Emerging Trends in Financial News Articles**
Nhu Khoa Nguyen | Thierry Delahaut | Emanuela Boros | Antoine Doucet | Gaël Lejeune

pdf bib abs **AstBERT: Enabling Language Model for Financial Code Understanding with Abstract Syntax Trees**
Rong Liang | Tiehua Zhang | Yujie Lu | Yuze Liu | Zhen Huang | Xin Chen

pdf bib abs **Disentangled Variational Topic Inference for Topic-Accurate Financial Report Generation**
Sixing Yan

pdf bib abs **Toward Privacy-preserving Text Embedding Similarity with Homomorphic Encryption**
Donggyu Kim | Garam Lee | Sungwoo Oh

pdf bib abs **TweetFinSent: A Dataset of Stock Sentiments on Twitter**
Yulong Pei | Amarachi Mbakwe | Akshat Gupta | Salwa Alamir | Hanxuan Lin | Xiaomo Liu | Sameena Shah

pdf bib abs **Stock Price Volatility Prediction: A Case Study with AutoML**
Hilal Pataci | Yunyao Li | Yannis Katsis | Yada Zhu | Lucian Popa

pdf bib abs **DigiCall: A Benchmark for Measuring the Maturity of Digital Strategy through Company Earning Calls**
Hilal Pataci | Kexuan Sun | T. Ravichandran

pdf bib abs **Learning Better Intent Representations for Financial Open Intent Classification**
Xianzhi Li | Will Aitken | Xiaodan Zhu | Stephen W. Thomas

pdf bib abs **Exploring Robustness of Prefix Tuning in Noisy Data: A Case Study in Financial Sentiment Analysis**
Sudhandar Balakrishnan | Yihao Fang | Xiaodan Zhu

pdf bib abs **A Taxonomical NLP Blueprint to Support Financial Decision Making through Information-Centred Interactions**
Siavash Kazemian | Cosmin Munteanu | Gerald Penn

pdf bib abs **Overview of the FinNLP-2022 ERAI Task: Evaluating the Rationales of Amateur Investors**
Chung-Chi Chen | Hen-Hsen Huang | Hiroya Takamura | Hsin-Hsi Chen

pdf bib abs **PromptShots at the FinNLP-2022 ERAI Task: Pairwise Comparison and Unsupervised Ranking**
Peratham Wiriyathammabhum

pdf bib abs **LIPI at the FinNLP-2022 ERAI Task: Ensembling Sentence Transformers for Assessing Maximum Possible Profit and Loss from Online Financial Posts**
Sohom Ghosh | Sudip Kumar Naskar

pdf bib abs **DCU-ML at the FinNLP-2022 ERAI Task: Investigating the Transferability of Sentiment Analysis Data for Evaluating Rationales of Investors**
Chenyang Lyu | Tianbo Ji | Liting Zhou

pdf bib abs **UOA at the FinNLP-2022 ERAI Task: Leveraging the Class Label Description for Financial Opinion Mining**
Jinan Zou | Haiyao Cao | Yanxi Liu | Lingqiao Liu | Ehsan Abbasnejad | Javen Qinfeng Shi

pdf bib abs **aiML at the FinNLP-2022 ERAI Task: Combining Classification and Regression Tasks for Financial Opinion Mining**
Zhaoxuan Qin | Jinan Zou | Qiaoyang Luo | Haiyao Cao | Yang Jiao

pdf bib abs **Yet at the FinNLP-2022 ERAI Task: Modified models for evaluating the Rationales of Amateur Investors**
Yan Zhuang | Fuji Ren

pdf bib abs **LDPP at the FinNLP-2022 ERAI Task: Determinantal Point Processes and Variational Auto-encoders for Identifying High-Quality Opinions from a pool of Social Media Posts**
Paul Trust | Rosane Minghim

pdf bib abs **Jetsons at the FinNLP-2022 ERAI Task: BERT-Chinese for mining high MPP posts**
Alolika Gon | Sihan Zha | Sai Krishna Rallabandi | Parag Pravin Dakle | Preethi Raghavan

pdf bib abs **No Stock is an Island: Learning Internal and Relational Attributes of Stocks with Contrastive Learning**
Shicheng Li | Wei Li | Zhiyuan Zhang | Ruihan Bao | Keiko Harimoto | Keiko Harimoto

pdf bib abs **Prospectus Language and IPO Performance**
Jared Sharpe | Keith Decker

pdf bib abs **It's Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset**
Alaa Alhamzeh | Romain Fonck | Erwan Versmée | Elöd Egyed-Zsigmond | Harald Kosch | Lionel Brunie

pdf bib abs **How Can a Teacher Make Learning From Sparse Data Softer? Application to Business Relation Extraction**
Hadjer Khaldi | Farah Benamara | Camille Pradel | Nathalie Aussenac-Gilles

pdf bib abs **Astock: A New Dataset and Automated Stock Trading based on Stock-specific News Analyzing Model**
Jinan Zou | Haiyao Cao | Lingqiao Liu | Yuhao Lin | Ehsan Abbasnejad | Javen Qinfeng Shi

pdf bib abs **Next-Year Bankruptcy Prediction from Textual Data: Benchmark and Baselines**
Henri Arno | Klaas Mulier | Joke Baeck | Thomas Demeester

pdf bib abs **AdaK-NER: An Adaptive Top-K Approach for Named Entity Recognition with Incomplete Annotations**
Hongtao Ruan | Liying Zheng | Peixian Hu

pdf bib abs **A Sentiment and Emotion Annotated Dataset for Bitcoin Price Forecasting Based on Reddit Posts**
Pavlo Seroyizhko | Zhanel Zhexenova | Muhammad Zohaib Shafiq | Fabio Merizzi | Andrea Galassi | Federico Ruggeri

pdf bib abs **FinSim4-ESG Shared Task: Learning Semantic Similarities for the Financial Domain. Extended edition to ESG insights**

Juyeon Kang | Ismail El Maarouf

pdf bib abs **Using Contextual Sentence Analysis Models to Recognize ESG Concepts**
Elvys Linhares Pontes | Mohamed Ben Jannet | Jose G. Moreno | Antoine Doucet

pdf bib abs **Automatic Term and Sentence Classification Via Augmented Term and Pre-trained language model in ESG Taxonomy texts**
Ke Tian | Zepeng Zhang | Hua Chen

pdf bib abs **Knowledge informed sustainability detection from short financial texts**
Boshko Koloski | Syrielle Montariol | Matthew Purver | Senja Pollak

pdf bib abs **TCS WITM 2022@FinSim4-ESG: Augmenting BERT with Linguistic and Semantic features for ESG data classification**
Tushar Goel | Vipul Chauhan | Suyash Sangwan | Ishan Verma | Tirthankar Dasgupta | Lipika Dey

pdf bib abs **Ranking Environment, Social And Governance Related Concepts And Assessing Sustainability Aspect of Financial Texts**
Sohom Ghosh | Sudip Kumar Naskar

pdf bib abs **Using Transformer-based Models for Taxonomy Enrichment and Sentence Classification**
Parag Pravin Dakle | Shrikumar Patil | Sai Krishna Rallabandi | Chaitra Hegde | Preethi Raghavan

↑ up

pdf (full) bib (full) Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)

pdf bib **Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)**
Debanjan Ghosh | Beata Beigman Klebanov | Smaranda Muresan | Anna Feldman | Soujanya Poria | Tuhin Chakrabarty

pdf bib abs **TEDB System Description to a Shared Task on Euphemism Detection 2022**
Peratham Wiriyathammabhum

pdf bib abs **A Prompt Based Approach for Euphemism Detection**
Abulimiti Maimaitituoheti | Yang Yong | Fan Xiaochao

pdf bib abs **Transfer Learning Parallel Metaphor using Bilingual Embeddings**
Maria Berger

pdf bib abs **Ring That Bell: A Corpus and Method for Multimodal Metaphor Detection in Videos**
Khalid Alnajjar | Mika Hämäläinen | Shuo Zhang

pdf bib abs **Picard understanding Darmok: A Dataset and Model for Metaphor-Rich Translation in a Constructed Language**
Peter A. Jansen | Jordan Boyd-Graber

pdf bib abs **The Secret of Metaphor on Expressing Stronger Emotion**
Yucheng Li | Frank Guerin | Chenghua Lin

pdf bib abs **Drum Up SUPPORT: Systematic Analysis of Image-Schematic Conceptual Metaphors**
Lennart Wachowiak | Dagmar Gromann | Chao Xu

pdf bib abs **Effective Cross-Task Transfer Learning for Explainable Natural Language Inference with T5**
Irina Bigoulaeva | Rachneet Singh Sachdeva | Harish Tayyar Madabushi | Aline Villavicencio | Iryna Gurevych

pdf bib abs **Detecting Euphemisms with Literal Descriptions and Visual Imagery**
Ilker Kesen | Aykut Erdem | Erkut Erdem | Iacer Calixto

pdf bib abs **Distribution-Based Measures of Surprise for Creative Language: Experiments with Humor and Metaphor**
Razvan C. Bunescu | Oseremen O. Uduehi

pdf bib abs **Euphemism Detection by Transformers and Relational Graph Attention Network**
Yuting Wang | Yiyi Liu | Ruqing Zhang | Yixing Fan | Jiafeng Guo

pdf bib abs **Just-DREAM-about-it: Figurative Language Understanding with DREAM-FLUTE**
Yuling Gu | Yao Fu | Valentina Pyatkin | Ian Magnusson | Bhavana Dalvi Mishra | Peter Clark

pdf bib abs **Bayes at FigLang 2022 Euphemism Detection shared task: Cost-Sensitive Bayesian Fine-tuning and Venn-Abers Predictors for Robust Training under Class Skewed Distributions**
Paul Trust | Kadusabe Provia | Kizito Omala

pdf bib abs **Food for Thought: How can we exploit contextual embeddings in the translation of idiomatic expressions?**
Lukas Santing | Ryan Jean-Luc Sijstermans | Giacomo Anerdi | Pedro Jeuris | Marijn ten Thij | Riza Batista-Navarro

pdf bib abs **EUREKA: EUphemism Recognition Enhanced through Knn-based methods and Augmentation**
Sedrick Scott Keh | Rohit Bharadwaj | Emmy Liu | [illegible] Gangal | Roberto Navigli

pdf bib abs **An insulin pump? Identifying figurative links in the construction of the drug lexicon**
Antonio Reyes | Rafael Saldivar

pdf bib abs **Can Yes-No Question-Answering Models be Useful for Few-Shot Metaphor Detection?**
Lena Dankin | Kfir Bar | Nachum Dershowitz

pdf bib abs **An Exploration of Linguistically-Driven and Transfer Learning Methods for Euphemism Detection**
Devika Tiwari | Natalie Parde

pdf bib abs **Back to the Roots: Predicting the Source Domain of Metaphors using Contrastive Learning**
Meghdut Sengupta | Milad Alshomary | Henning Wachsmuth

pdf bib abs **SBU Figures It Out: Models Explain Figurative Language**
Yash Kumar Lal | Mohaddeseh Bastan

pdf bib abs **NLP@UIT at FigLang-EMNLP 2022: A Divide-and-Conquer System For Shared Task On Understanding Figurative Language**
Khoa Thi-Kim Phan | Duc-Vu Nguyen | Ngan Luu-Thuy Nguyen

pdf bib abs **Adversarial Perturbations Augmented Language Models for Euphemism Identification**
Guneet Kohli | Prabsimran Kaur | Jatin Bedi

pdf bib abs **FigurativeQA: A Test Benchmark for Figurativeness Comprehension for Question Answering**
Geetanjali Rakshit | Jeffrey Flanigan

pdf bib abs **Exploring Euphemism Detection in Few-Shot and Zero-Shot Settings**
Sedrick Scott Keh

pdf bib abs **On the Cusp of Comprehensibility: Can Language Models Distinguish Between Metaphors and Nonsense?**
Bernadeta Griciūtė | Marc Tanti | Lucia Donatelli

pdf bib abs **A Report on the FigLang 2022 Shared Task on Understanding Figurative Language**
Arkadiy Saakyan | Tuhin Chakrabarty | Debanjan Ghosh | Smaranda Muresan

pdf bib abs **A Report on the Euphemisms Detection Shared Task**
Patrick Lee | Anna Feldman | Jing Peng

↑ up

pdf (full) bib (full) Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)

pdf bib **Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)**
Antoine Bosselut | Khyathi Chandu | Kaustubh Dhole | Varun Gangal | Sebastian Gehrmann | Yacine Jernite | Jekaterina Novikova | Laura Perez-Beltrachini

pdf bib abs **Improving abstractive summarization with energy-based re-ranking**
Diogo Pernes | Afonso Mendes | André F. T. Martins

pdf bib abs **Task-driven augmented data evaluation**
Olga Golovneva | Pan Wei | Khadige Abboud | Charith Peris | Lizhen Tan | Haiyang Yu

pdf bib abs **Generating Coherent Narratives with Subtopic Planning to Answer How-to Questions**
Pengshan Cai | Mo Yu | Fei Liu | Hong Yu

pdf bib abs **Weakly Supervised Context-based Interview Question Generation**
Samiran Pal | Kaamraan Khan | Avinash Kumar Singh | Subhasish Ghosh | Tapas Nayak | Girish Palshikar | Indrajit Bhattacharya

pdf bib abs **Analyzing Multi-Task Learning for Abstractive Text Summarization**
Frederic Thomas Kirstein | Jan Philip Wahle | Terry Ruas | Bela Gipp

pdf bib abs **CLSE: Corpus of Linguistically Significant Entities**
Aleksandr Chuklin | Justin Zhao | Mihir Kale

pdf bib abs **Revisiting text decomposition methods for NLI-based factuality scoring of summaries**
John Glover | Federico Fancellu | Vasudevan Jagannathan | Matthew R. Gormley | Thomas Schaaf

pdf bib abs **Semantic Similarity as a Window into Vector- and Graph-Based Metrics**
Wai Ching Leung | Shira Wein | Nathan Schneider

pdf bib abs **Towards In-Context Non-Expert Evaluation of Reflection Generation for Counselling Conversations**
Zixiu Wu | Simone Balloccu | Rim Helaoui | Diego Reforgiato Recupero | Daniele Riboni

pdf bib abs **WikiOmnia: filtration and evaluation of the generated QA corpus on the whole Russian Wikipedia**

Dina Pisarevskaya | Tatiana Shavrina

pdf bib abs **Evaluation of Response Generation Models: Shouldn't It Be Shareable and Replicable?**
Seyed Mahed Mousavi | Gabriel Roccabruna | Michela Lorandi | Simone Caldarella | Giuseppe Riccardi

pdf bib abs **Enhancing and Evaluating the Grammatical Framework Approach to Logic-to-Text Generation**
Eduardo Calò | Elze van der Werf | Albert Gatt | Kees van Deemter

pdf bib abs **Controllable Text Generation for All Ages: Evaluating a Plug-and-Play Approach to Age-Adapted Dialogue**
Lennert Jansen | Štěpán Lars Laichter | Arabella Sinclair | Margot van der Goot | Raquel Fernandez | Sandro Pezzelle

pdf bib abs **Template-based Contact Email Generation for Job Recommendation**
Qiuchi Li | Christina Lioma

pdf bib abs **Are Abstractive Summarization Models truly 'Abstractive'? An Empirical Study to Compare the two Forms of Summarization**
Vinayshekhar Bannihatti Kumar | Rashmi Gangadharaiah

pdf bib abs **Transfer learning for multilingual vacancy text generation**
Anna Lorincz | David Graus | Dor Lavi | Joao Lebre Magalhaes Pereira

pdf bib abs **Plug-and-Play Recipe Generation with Content Planning**
Yinhong Liu | Yixuan Su | Ehsan Shareghi | Nigel Collier

pdf bib abs **Towards Attribute-Entangled Controllable Text Generation: A Pilot Study of Blessing Generation**
Shulin Huang | Shirong Ma | Yinghui Li | Li Yangning | Shiyang Lin | Haitao Zheng | Ying Shen

pdf bib abs **Unsupervised Token-level Hallucination Detection from Summary Generation By-products**
Andreas Marfurt | James Henderson

pdf bib abs **A Corpus and Evaluation for Predicting Semi-Structured Human Annotations**
Andreas Marfurt | Ashley Thornton | David Sylvan | Lonneke van der Plas | James Henderson

pdf bib abs **T5QL: Taming language models for SQL generation**
Samuel David Arcadinho | David Aparicio | Hugo Veiga | Antonio Alegria

pdf bib abs **Human perceiving behavior modeling in evaluation of code generation models**
Sergey V. Kovalchuk | Vadim Lomshakov | Artem Aliev

pdf bib abs **Nearest Neighbor Language Models for Stylistic Controllable Generation**
Severino Trotta | Lucie Flek | Charles Welch

pdf bib abs **On reporting scores and agreement for error annotation tasks**
Maja Popović | Anya Belz

pdf bib abs **Answerability: A custom metric for evaluating chatbot performance**
Pranav Gupta | Anand A. Rajasekar | Amisha Patel | Mandar Kulkarni | Alexander Sunell | Kyung Kim | Krishnan Ganapathy | Anusua Trivedi

pdf bib abs **Improved Evaluation of Automatic Source Code Summarisation**
Jesse Phillips | David Bowes | Mahmoud El-Haj | Tracy Hall

pdf bib abs **Most NLG is Low-Resource: here's what we can do about it**
David M. Howcroft | Dimitra Gkatzia

pdf bib abs **GiCCS: A German in-Context Conversational Similarity Benchmark**
Shima Asaadi | Zahra Kolagar | Alina Liebel | Alessandra Zarcone

pdf bib abs **Control Prefixes for Parameter-Efficient Text Generation**
Jordan Clive | Kris Cao | Marek Rei

pdf bib abs **A Survey of Recent Error Annotation Schemes for Automatically Generated Text**
Rudali Huidrom | Anya Belz

pdf bib abs **What's in a (dataset's) name? The case of BigPatent**
Silvia Casola | Alberto Lavelli | Horacio Saggion

pdf bib abs **Measuring the Measuring Tools: An Automatic Evaluation of Semantic Metrics for Text Corpora**
George Kour | Samuel Ackerman | Eitan Daniel Farchi | Orna Raz | Boaz Carmeli | Ateret Anaby Tavor

pdf bib abs **Multilingual Social Media Text Generation and Evaluation with Few-Shot Prompting**
Mack Blackburn

pdf bib abs **Assessing Inter-metric Correlation for Multi-document Summarization Evaluation**
Michael Ridenour | Ameeta Agrawal | Olubusayo Olabisi

pdf bib abs **Factual Error Correction for Abstractive Summaries Using Entity Retrieval**
Hwanhee Lee | Cheoneum Park | Seunghyun Yoon | Trung Bui | Franck Dernoncourt | Juae Kim | Kyomin Jung

pdf bib abs **Coherent Long Text Generation by Contrastive Soft Prompt**
Guandan Chen | Jiashu Pu | Yadong Xi | Rongsheng Zhang

pdf bib abs **Error Analysis of ToTTo Table-to-Text Neural NLG Models**
Barkavi Sundararajan | Somayajulu Sripada | Ehud Reiter

pdf bib abs **Improving Dialogue Act Recognition with Augmented Data**
Khyati Mahajan | Soham Parikh | Quaizar Vohra | Mitul Tiwari | Samira Shaikh

pdf bib abs **Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks? A Case Study of Image Paragraph Generation**
Nikolai Ilinykh | Simon Dobnik

pdf bib abs **20Q: Overlap-Free World Knowledge Benchmark for Language Models**
Maxime De Bruyn | Ehsan Lotfi | Jeska Buhmann | Walter Daelemans

pdf bib abs **What Was Your Name Again? Interrogating Generative Conversational Models For Factual Consistency Evaluation**
Ehsan Lotfi | Maxime De Bruyn | Jeska Buhmann | Walter Daelemans

pdf bib abs **Narrative Why-Question Answering: A Review of Challenges and Datasets**
Emil Kalbaliyev | Kairit Sirts

pdf bib abs **Exploring a POS-based Two-stage Approach for Improving Low-Resource AMR-to-Text Generation**
Marco Antonio Sobrevilla Cabezudo | Thiago Pardo

pdf bib abs **What Makes Data-to-Text Generation Hard for Pretrained Language Models?**
Moniba Keymanesh | Adrian Benton | Mark Dredze

pdf bib abs **Don't Say What You Don't Know: Improving the Consistency of Abstractive Summarization by Constraining Beam Search**
Daniel King | Zejiang Shen | Nishant Subramani | Daniel S. Weld | Iz Beltagy | Doug Downey

↑ up

pdf (full) bib (full) Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)

pdf bib **Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)**
Alberto Lavelli | Eben Holderness | Antonio Jimeno Yepes | Anne-Lyse Minard | James Pustejovsky | Fabio Rinaldi

pdf bib abs **Exploring the Influence of Dialog Input Format for Unsupervised Clinical Questionnaire Filling**
Farnaz Ghassemi Toudeshki | Anna Liednikova | Philippe Jolivet | Claire Gardent

pdf bib abs **Assessing the Limits of Straightforward Models for Nested Named Entity Recognition in Spanish Clinical Narratives**
Matias Rojas | Casimiro Pio Carrino | Aitor Gonzalez-Agirre | Jocelyn Dunstan | Marta Villegas

pdf bib abs **Can Current Explainability Help Provide References in Clinical Notes to Support Humans Annotate Medical Codes?**
Byung-Hak Kim | Zhongfen Deng | Philip Yu | Varun Ganapathi

pdf bib abs **Distinguishing between focus and background entities in biomedical corpora using discourse structure and transformers**
Antonio Jimeno Yepes | Karin Verspoor

pdf bib abs **FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain**
Yanis Labrak | Adrien Bazoge | Richard Dufour | Beatrice Daille | Pierre-Antoine Gourraud | Emmanuel Morin | Mickael Rouvier

pdf bib abs **A Large-Scale Dataset for Biomedical Keyphrase Generation**
Maël Houbre | Florian Boudin | Beatrice Daille

pdf bib abs **Section Classification in Clinical Notes with Multi-task Transformers**
Fan Zhang | Itay Laish | Ayelet Benjamini | Amir Feder

pdf bib abs **Building a Clinically-Focused Problem List From Medical Notes**
Amir Feder | Itay Laish | Shashank Agarwal | Uri Lerner | Avel Atias | Cathy Cheung | Peter Clardy | Alon Peled-Cohen | Rachana Fellinger | Hengrui Liu | Lan Huong Nguyen | Birju Patel | Natan Potikha | Amir Taubenfeld | Liwen Xu | Seung Doo Yang | Ayelet Benjamini | Avinatan Hassidim

pdf bib abs **Specializing Static and Contextual Embeddings in the Medical Domain Using Knowledge Graphs: Let's Keep It Simple**
Hicham El Boukkouri | Olivier Ferret | Thomas Lavergne | Pierre Zweigenbaum

pdf bib abs **BioSimCSE: BioMedical Sentence Embeddings using Contrastive learning**
Kamal raj Kanakarajan | Bhuvana Kundumani | Abhijith Abraham | Malaikannan Sankarasubbu

pdf bib abs **Proxy-based Zero-Shot Entity Linking by Effective Candidate Retrieval**
Maciej Wiatrak | Eirini Arvaniti | Angus Brayne | Jonas Vetterle | Aaron Sim

pdf bib abs **BERT for Long Documents: A Case Study of Automated ICD Coding**
Arash Afkanpour | Shabir Adeel | Hansenclever Bassani | Arkady Epshteyn | Hongbo Fan | Isaac Jones | Mahan Malihi | Adrian Nauth | Raj Sinha | Sanjana Woonna | Shiva Zamani | Elli Kanal | Mikhail Fomitchev | Donny Cheung

pdf bib abs **Parameter Efficient Transfer Learning for Suicide Attempt and Ideation Detection**
Bhanu Pratap Singh Rawat | Hong Yu

pdf bib abs **Automatic Patient Note Assessment without Strong Supervision**
Jianing Zhou | Vyom Nayan Thakkar | Rachel Yudkowsky | Suma Bhat | William F. Bond

pdf bib abs **DDI-MuG: Multi-aspect Graphs for Drug-Drug Interaction Extraction**
Jie Yang | Yihao Ding | Siqu Long | Josiah Poon | Soyeon Caren Han

pdf bib abs **Divide and Conquer: An Extreme Multi-Label Classification Approach for Coding Diseases and Procedures in Spanish**
Jose Barros | Matias Rojas | Jocelyn Dunstan | Andres Abeliuk

pdf bib abs **Curriculum-guided Abstractive Summarization for Mental Health Online Posts**
Sajad Sotudeh | Nazli Goharian | Hanieh Deilamsalehy | Franck Dernoncourt

pdf bib abs **Improving information fusion on multimodal clinical data in classification settings**
Sneha Jha | Erik Mayer | Mauricio Barahona

pdf bib abs **How Long Is Enough? Exploring the Optimal Intervals of Long-Range Clinical Note Language Modeling**
Samuel Cahyawijaya | Bryan Wilie | Holy Lovenia | Huan Zhong | MingQian Zhong | Yuk-Yu Nancy Ip | Pascale Fung

pdf bib abs **A Quantitative and Qualitative Analysis of Schizophrenia Language**
Amal Alqahtani | Efsun Sarioglu Kayi | Sardar Hamidian | Michael Compton | Mona Diab

pdf bib abs **Exploring Hybrid and Ensemble Models for Multiclass Prediction of Mental Health Status on Social Media**
Sourabh Zanwar | Daniel Wiechmann | Yu Qiao | Elma Kerz

pdf bib abs **A Knowledge-Graph-Based Intrinsic Test for Benchmarking Medical Concept Embeddings and Pretrained Language Models**
Claudio Aracena | Fabián Villena | Matias Rojas | Jocelyn Dunstan

pdf bib abs **Enriching Deep Learning with Frame Semantics for Empathy Classification in Medical Narrative Essays**
Priyanka Dey | Roxana Girju

pdf bib abs **Condition-Treatment Relation Extraction on Disease-related Social Media Data**
Sichang Tu | Stephen Doogan | Jinho D. Choi

pdf bib abs **Integration of Heterogeneous Knowledge Sources for Biomedical Text Processing**
Parsa Bagherzadeh | Sabine Bergler

↑ up

pdf (full) bib (full) Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP)

pdf bib **Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP)**
Deborah Ferreira | Marco Valentino | Andre Freitas | Sean Welleck | Moritz Schubotz

pdf bib abs **Tracing and Manipulating intermediate values in Neural Math Problem Solvers**
Yuta Matsumoto | Benjamin Heinzerling | Masashi Yoshikawa | Kentaro Inui

pdf bib abs **Investigating Math Word Problems using Pretrained Multilingual Language Models**
Minghuan Tan | Lei Wang | Lingxiao Jiang | Jing Jiang

pdf bib abs **Induced Natural Language Rationales and Interleaved Markup Tokens Enable Extrapolation in Large Language Models**
Mirelle Candida Bueno | Carlos Gemmell | Jeff Dalton | Roberto Lotufo | Rodrigo Nogueira

pdf bib abs **Towards Autoformalization of Mathematics and Code Correctness: Experiments with Elementary Proofs**
Garett Cunningham | Razvan Bunescu | David Juedes

pdf bib abs **Numerical Correlation in Text**
Daniel Spokoyny | Chien-Sheng Wu | Caiming Xiong

pdf bib abs **Extracting Operator Trees from Model Embeddings**
Anja Reusch | Wolfgang Lehner

pdf bib abs **End-to-End Evaluation of a Spoken Dialogue System for Learning Basic Mathematics**
Eda Okur | Saurav Sahay | Roddy Fuentes Alba | Lama Nachman

Eda Okur | Saurav Sahay | Roddy Fuentes Alba | Lama Nachman

↑ up

pdf (full) bib (full) Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22)

pdf bib **Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22)**
Jack FitzGerald | Kay Rottmann | Julia Hirschberg | Mohit Bansal | Anna Rumshisky | Charith Peris | Christopher Hench

pdf bib abs **Robust Domain Adaptation for Pre-trained Multilingual Neural Machine Translation Models**
Mathieu Grosso | Alexis Mathey | Pirashanth Ratnamogan | William Vanhuffel | Michael Fotso

pdf bib abs **Fine-grained Multi-lingual Disentangled Autoencoder for Language-agnostic Representation Learning**
Zetian Wu | Zhongkai Sun | Zhengyang Zhao | Sixing Lu | Chengyuan Ma | Chenlei Guo

pdf bib abs **Byte-Level Massively Multilingual Semantic Parsing**
Massimo Nicosia | Francesco Piccinno

pdf bib abs **HIT-SCIR at MMNLU-22: Consistency Regularization for Multilingual Spoken Language Understanding**
Bo Zheng | Zhouyang Li | Fuxuan Wei | Qiguang Chen | Libo Qin | Wanxiang Che

pdf bib abs **Play música alegre: A Large-Scale Empirical Analysis of Cross-Lingual Phenomena in Voice Assistant Interactions**
Donato Crisostomi | Alessandro Manzotti | Enrico Palumbo | Davide Bernardi | Sarah Campbell | Shubham Garg

pdf bib abs **Zero-Shot Cross-Lingual Sequence Tagging as Seq2Seq Generation for Joint Intent Classification and Slot Filling**
Fei Wang | Kuan-hao Huang | Anoop Kumar | Aram Galstyan | Greg Ver steeg | Kai-wei Chang

pdf bib abs C5L7**: A Zero-Shot Algorithm for Intent and Slot Detection in Multilingual Task Oriented Languages**
Jiun-hao Jhan | Qingxiaoyang Zhu | Nehal Bengre | Tapas Kanungo

pdf bib abs **Machine Translation for Multilingual Intent Detection and Slots Filling**
Maxime De bruyn | Ehsan Lotfi | Jeska Buhmann | Walter Daelemans

pdf bib abs **Massively Multilingual Natural Language Understanding 2022 (MMNLU-22) Workshop and Competition**
Jack FitzGerald | Christopher Hench | Charith Peris | Kay Rottmann

↑ up

pdf (full) bib (full) Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL)

pdf bib **Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL)**
Duygu Ataman | Hila Gonen | Sebastian Ruder | Orhan Firat | Gözde Gül Sahin | Jamshidbek Mirzakhalov

pdf bib abs **Entity Retrieval from Multilingual Knowledge Graphs**
Saher Esmeir | Arthur Câmara | Edgar Meij

pdf bib abs **Few-Shot Cross-Lingual Learning for Event Detection**
Luis Guzman Nateras | Viet Lai | Franck Dernoncourt | Thien Nguyen

pdf bib abs **Zero-shot Cross-Lingual Counterfactual Detection via Automatic Extraction and Prediction of Clue Phrases**
Asahi Ushio | Danushka Bollegala

pdf bib abs **Zero-shot Cross-Language Transfer of Monolingual Entity Linking Models**
Elliot Schumacher | James Mayfield | Mark Dredze

pdf bib abs **Rule-Based Clause-Level Morphology for Multiple Languages**
Tillmann Dönicke

pdf bib abs **Comparative Analysis of Cross-lingual Contextualized Word Embeddings**
Hossain Shaikh Saadi | Viktor Hangya | Tobias Eder | Alexander Fraser

pdf bib abs **How Language-Dependent is Emotion Detection? Evidence from Multilingual BERT**
Luna De Bruyne | Pranaydeep Singh | Orphee De Clercq | Els Lefever | Veronique Hoste

pdf bib abs **MicroBERT: Effective Training of Low-resource Monolingual BERTs through Parameter Reduction and Multitask Learning**
Luke Gessler | Amir Zeldes

pdf bib abs **Transformers on Multilingual Clause-Level Morphology**
Emre Can Acikgoz | Tilek Chubakov | Muge Kural | Gözde Şahin | Deniz Yuret

pdf bib abs **Impact of Sequence Length and Copying on Cl**[illegible]
Badr Jaidi | Utkarsh Sahoo | Yihan Wu | Garrett Nicolai | Miikka Silfverberg

Badr Jaidi | Utkarsh Sabdo | Xinan Wu | Garrett Nicolai | Mikka Silfverberg

pdf bib abs **Towards Improved Distantly Supervised Multilingual Named-Entity Recognition for Tweets**
Ramy Eskander | Shubhanshu Mishra | Sneha Mehta | Sofia Samaniego | Aria Haghighi

pdf bib abs **Average Is Not Enough: Caveats of Multilingual Evaluation**
Matúš Pikuliak | Marian Simko

pdf bib abs **The MRL 2022 Shared Task on Multilingual Clause-level Morphology**
Omer Goldman | Francesco Tinner | Hila Gonen | Benjamin Muller | Victoria Basmov | Shadrack Kirimi | Lydia Nishimwe | Benoît Sagot | Djamé Seddah | Reut Tsarfaty | Duygu Ataman

↑ up

pdf (full) bib (full) **Proceedings of the Workshop on Novel Ideas in Learning-to-Learn through Interaction (NILLI 2022)**

pdf bib **Proceedings of the Workshop on Novel Ideas in Learning-to-Learn through Interaction (NILLI 2022)**
Prasanna Parthasarathi | Marc-Alexandre Côté

↑ up

pdf (full) bib (full) **Proceedings of the Natural Legal Language Processing Workshop 2022**

pdf bib **Proceedings of the Natural Legal Language Processing Workshop 2022**
Nikolaos Aletras | Ilias Chalkidis | Leslie Barrett | Cătălina Goanță | Daniel Preoțiuc-Pietro

pdf bib abs **On Breadth Alone: Improving the Precision of Terminology Extraction Systems on Patent Corpora**
Sean Nordquist | Adam Meyers

pdf bib abs **On What it Means to Pay Your Fair Share: Towards Automatically Mapping Different Conceptions of Tax Justice in Legal Research Literature**
Reto Gubelmann | Peter Hongler | Elina Margadant | Siegfried Handschuh

pdf bib abs **ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US**
Gil Semo | Dor Bernsohn | Ben Hagag | Gila Hayat | Joel Niklaus

pdf bib abs **Combining WordNet and Word Embeddings in Data Augmentation for Legal Texts**
Sezen Perçin | Andrea Galassi | Francesca Lagioia | Federico Ruggeri | Piera Santin | Giovanni Sartor | Paolo Torroni

pdf bib abs **A Legal Approach to Hate Speech – Operationalizing the EU's Legal Framework against the Expression of Hatred as an NLP Task**
Frederike Zufall | Marius Hamacher | Katharina Kloppenborg | Torsten Zesch

pdf bib abs **Privacy Pitfalls of Online Service Terms and Conditions: a Hybrid Approach for Classification and Summarization**
Emilia Lukose | Suparna De | Jon Johnson

pdf bib abs **Abstractive Summarization of Dutch Court Verdicts Using Sequence-to-sequence Models**
Marijn Schraagen | Floris Bex | Nick Van De Luijtgaarden | Daniël Prijs

pdf bib abs **Legal-Tech Open Diaries: Lesson learned on how to develop and deploy light-weight models in the era of humongous Language Models**
Stelios Maroudas | Sotiris Legkas | Prodromos Malakasiotis | Ilias Chalkidis

pdf bib abs **Towards Cross-Domain Transferability of Text Generation Models for Legal Text**
Vinayshekhar Bannihatti Kumar | Kasturi Bhattacharjee | Rashmi Gangadharaiah

pdf bib abs **Parameter-Efficient Legal Domain Adaptation**
Jonathan Li | Rohan Bhambhoria | Xiaodan Zhu

pdf bib abs **Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer**
Dimitris Mamakas | Petros Tsotsi | Ion Androutsopoulos | Ilias Chalkidis

pdf bib abs **Data-efficient end-to-end Information Extraction for Statistical Legal Analysis**
Wonseok Hwang | Saehee Eom | Hanuhl Lee | Hai Jin Park | Minjoon Seo

pdf bib abs **Semantic Segmentation of Legal Documents via Rhetorical Roles**
Vijit Malik | Rishabh Sanjay | Shouvik Kumar Guha | Angshuman Hazarika | Shubham Nigam | Arnab Bhattacharya | Ashutosh Modi

pdf bib abs **Privacy-Preserving Models for Legal Natural Language Processing**
Ying Yin | Ivan Habernal

pdf bib abs **Named Entity Recognition in Indian court judgments**

Prathamesh Kalamkar | Astha Agarwal | Aman Tiwari | Smita Gupta | Saurabh Karn | Vivek Raghavan

pdf bib abs **The Legal Argument Reasoning Task in Civil Procedure**
Leonard Bongard | Lena Held | Ivan Habernal

pdf bib abs **Efficient Deep Learning-based Sentence Boundary Detection in Legal Text**
Reshma Sheik | Gokul T | S Nirmala

pdf bib abs **Tracking Semantic Shifts in German Court Decisions with Diachronic Word Embeddings**
Daniel Braun

pdf bib abs **Should I disclose my dataset? Caveats between reproducibility and individual data rights**
Raysa M. Benatti | Camila M. L. Villarroel | Sandra Avila | Esther L. Colombini | Fabiana Severi

pdf bib abs **Attack on Unfair ToS Clause Detection: A Case Study using Universal Adversarial Triggers**
Shanshan Xu | Irina Broda | Rashid Haddad | Marco Negrini | Matthias Grabmair

pdf bib abs **E-NER — An Annotated Named Entity Recognition Corpus of Legal Text**
Ting Wai Terence Au | Vasileios Lampos | Ingemar Cox

pdf bib abs **Detecting Relevant Differences Between Similar Legal Texts**
Xiang Li | Jiaxun Gao | Diana Inkpen | Wolfgang Alschner

pdf bib abs **Legal and Political Stance Detection of SCOTUS Language**
Noah Bergam | Emily Allaway | Kathleen Mckeown

pdf bib abs **Graph-based Keyword Planning for Legal Clause Generation from Topics**
Sagar Joshi | Sumanth Balaji | Aparna Garimella | Vasudeva Varma

pdf bib abs **Automatic Classification of Legal Violations in Cookie Banner Texts**
Marieke Van Hofslot | Almila Akdag Salah | Albert Gatt | Cristiana Santos

pdf bib abs **Text Simplification for Legal Domain: {I}nsights and Challenges**
Aparna Garimella | Abhilasha Sancheti | Vinay Aggarwal | Ananya Ganesh | Niyati Chhaya | Nandakishore Kambhatla

pdf bib abs **Legal Named Entity Recognition with Multi-Task Domain Adaptation**
Răzvan-Alexandru Smădu | Ion-Robert Dinică | Andrei-Marius Avram | Dumitru-Clementin Cercel | Florin Pop | Mihaela-Claudia Cercel

pdf bib abs **Computing and Exploiting Document Structure to Improve Unsupervised Extractive Summarization of Legal Case Decisions**
Yang Zhong | Diane Litman

pdf bib abs **AraLegal-BERT: A pretrained language model for Arabic Legal text**
Muhammad Al-qurishi | Sarah Alqaseemi | Riad Souissi

pdf bib abs **An Efficient Active Learning Pipeline for Legal Text Classification**
Sepideh Mamooler | Rémi Lebret | Stephane Massonnet | Karl Aberer

↑ up

pdf (full) bib (full) Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)

pdf bib **Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)**
Laura Biester | Dorottya Demszky | Zhijing Jin | Mrinmaya Sachan | Joel Tetreault | Steven Wilson | Lu Xiao | Jieyu Zhao

pdf bib abs **A unified framework for cross-domain and cross-task learning of mental health conditions**
Huikai Chua | Andrew Caines | Helen Yannakoudakis

pdf bib abs **Critical Perspectives: A Benchmark Revealing Pitfalls in PerspectiveAPI**
Lucas Rosenblatt | Lorena Piedras | Julia Wilkins

pdf bib abs **Securely Capturing People's Interactions with Voice Assistants at Home: A Bespoke Tool for Ethical Data Collection**
Angus Addlesee

pdf bib abs **Leveraging World Knowledge in Implicit Hate Speech Detection**
Jessica Lin

pdf bib abs **A Dataset of Sustainable Diet Arguments on Twitter**
Marcus Hansen | Daniel Hershcovich

pdf bib abs **Impacts of Low Socio-economic Status on Educational Outcomes: A Narrative Based Analysis**
Motti Kelbessa | Ilyas Jamil | Labiba Jahan

pdf bib abs **Enhancing Crisis-Related Tweet Classification with Entity-Masked Language Modeling and Multi-Task Learning**

Philipp Seeberger | Korbinian Riedhammer

pdf bib abs **Misinformation Detection in the Wild: News Source Classification as a Proxy for Non-article Texts**
Matyas Bohacek

pdf bib abs **Modelling Persuasion through Misuse of Rhetorical Appeals**
Amalie Pauli | Leon Derczynski | Ira Assent

pdf bib abs **Breaking through Inequality of Information Acquisition among Social Classes: A Modest Effort on Measuring "Fun"**
Chenghao Xiao | Baicheng Sun | Jindi Wang | Mingyue Liu | Jiayi Feng

pdf bib abs **Using NLP to Support English Teaching in Rural Schools**
Luis Chiruzzo | Laura Musto | Santiago Gongora | Brian Carpenter | Juan Filevich | Aiala Rosa

pdf bib abs **"Am I Answering My Job Interview Questions Right?": A NLP Approach to Predict Degree of Explanation in Job Interview Responses**
Raghu Verrap | Ehsanul Nirjhar | Ani Nenkova | Theodora Chaspari

pdf bib abs **Identifying Condescending Language: A Tale of Two Distinct Phenomena?**
Carla Perez Almendros | Steven Schockaert

pdf bib abs **BELA: Bot for English Language Acquisition**
Muskan Mahajan

pdf bib abs **Applicability of Pretrained Language Models: Automatic Screening for Children's Language Development Level**
Byoung-doo Oh | Yoon-koung Lee | Yu-seop Kim

pdf bib abs **Transformers-Based Approach for a Sustainability Term-Based Sentiment Analysis (STBSA)**
Blaise Sandwidi | Suneer Pallitharammal Mukkolakal

pdf bib abs **Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features**
Gokul Karthik Kumar | Karthik Nandakumar

↑ up

pdf (full) bib (full) Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)

pdf bib **Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)**
David Bamman | Dirk Hovy | David Jurgens | Katherine Keith | Brendan O'Connor | Svitlana Volkova

pdf bib abs **Improving the Generalizability of Text-Based Emotion Detection by Leveraging Transformers with Psycholinguistic Features**
Sourabh Zanwar | Daniel Wiechmann | Yu Qiao | Elma Kerz

pdf bib abs **Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads**
Ann-sophie Gnehm | Eva Bühlmann | Helen Buchs | Simon Clematide

pdf bib abs **Experiencer-Specific Emotion and Appraisal Prediction**
Maximilian Wegge | Enrica Troiano | Laura Ana Maria Oberlaender | Roman Klinger

pdf bib abs **Understanding Narratives from Demographic Survey Data: a Comparative Study with Multiple Neural Topic Models**
Xiao Xu | Gert Stulp | Antal Van Den Bosch | Anne Gauthier

pdf bib abs **To Prefer or to Choose? Generating Agency and Power Counterfactuals Jointly for Gender Bias Mitigation**
Maja Stahl | Maximilian Spliethöver | Henning Wachsmuth

pdf bib abs **Conspiracy Narratives in the Protest Movement Against COVID-19 Restrictions in Germany. A Long-term Content Analysis of Telegram Chat Groups.**
Manuel Weigand | Maximilian Weber | Johannes Gruber

pdf bib abs **Conditional Language Models for Community-Level Linguistic Variation**
Bill Noble | Jean-philippe Bernardy

pdf bib abs **Understanding Interpersonal Conflict Types and their Impact on Perception Classification**
Charles Welch | Joan Plepi | Béla Neuendorf | Lucie Flek

pdf bib abs **Examining Political Rhetoric with Epistemic Stance Detection**
Ankita Gupta | Su Lin Blodgett | Justin Gross | Brendan O'connor

pdf bib abs **Linguistic Elements of Engaging Customer Service Discourse on Social Media**
Sonam Singh | Anthony Rios

pdf bib abs **Measuring Harmful Representations in Scandinavian Language Models**

Samia Touileb | Debora Nozza

pdf bib abs **Can Contextualizing User Embeddings Improve Sarcasm and Hate Speech Detection?**
Kim Breitwieser

pdf bib abs **Professional Presentation and Projected Power: A Case Study of Implicit Gender Information in English CVs**
Jinrui Yang | Sheilla Njoto | Marc Cheong | Leah Ruppanner | Lea Frermann

pdf bib abs **Detecting Dissonant Stance in Social Media: The Role of Topic Exposure**
Vasudha Varadarajan | Nikita Soni | Weixi Wang | Christian Luhmann | H. Andrew Schwartz | Naoya Inoue

pdf bib abs **An Analysis of Acknowledgments in NLP Conference Proceedings**
Winston Wu

pdf bib abs **Extracting Associations of Intersectional Identities with Discourse about Institution from Nigeria**
Pavan Kantharaju | Sonja Schmer-galunder

pdf bib abs **OLALA: Object-Level Active Learning for Efficient Document Layout Annotation**
Zejiang Shen | Weining Li | Jian Zhao | Yaoliang Yu | Melissa Dell

pdf bib abs **Towards Few-Shot Identification of Morality Frames using In-Context Learning**
Shamik Roy | Nishanth Sridhar Nakshatri | Dan Goldwasser

pdf bib abs **Utilizing Weak Supervision to Create S3D: A Sarcasm Annotated Dataset**
Jordan Painter | Helen Treharne | Diptesh Kanojia

pdf bib abs **A Robust Bias Mitigation Procedure Based on the Stereotype Content Model**
Eddie Ungless | Amy Rafferty | Hrichika Nag | Björn Ross

pdf bib abs **Who is GPT-3? An exploration of personality, values and demographics**
Marilù Miotto | Nicola Rossberg | Bennett Kleinberg

↑ up

pdf (full) bib (full) Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)

pdf bib **Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)**
Zhijian Ou | Junlan Feng | Juanzi Li

pdf bib abs **Oh My Mistake!: Toward Realistic Dialogue State Tracking including Turnback Utterances**
Takyoung Kim | Yukyung Lee | Hoonsang Yoon | Pilsung Kang | Junseong Bang | Misuk Kim

pdf bib abs **A GlobalPointer based Robust Approach for Information Extraction from Dialog Transcripts**
Yanbo J. Wang | Sheng Chen | Hengxing Cai | Wei Wei | Kuo Yan | Zhe Sun | Hui Qin | Yuming Li | Xiaochen Cai

pdf bib abs **A Token-pair Framework for Information Extraction from Dialog Transcripts in SereTOD Challenge**
Chenyue Wang | Xiangxing Kong | Mengzuo Huang | Feng Li | Jian Xing | Weidong Zhang | Wuhe Zou

pdf bib abs **Prompt Learning for Domain Adaptation in Task-Oriented Dialogue**
Makesh Narsimhan Sreedhar | Christopher Parisien

pdf bib abs **Disentangling Confidence Score Distribution for Out-of-Domain Intent Detection with Energy-Based Learning**
Yanan Wu | Zhiyuan Zeng | Keqing He | Yutao Mou | Pei Wang | Yuanmeng Yan | Weiran Xu

pdf bib abs **Semi-Supervised Knowledge-Grounded Pre-training for Task-Oriented Dialog Systems**
Weihao Zeng | Keqing He | Zechen Wang | Dayuan Fu | Guanting Dong | Ruotong Geng | Pei Wang | Jingang Wang | Chaobo Sun | Wei Wu | Weiran Xu

pdf bib abs **CMCC: A Comprehensive and Large-Scale Human-Human Dataset for Dialogue Systems**
Yi Huang | Xiaoting Wu | Si Chen | Wei Hu | Qing Zhu | Junlan Feng | Chao Deng | Zhijian Ou | Jiangjiang Zhao

pdf bib abs **State-Aware Adversarial Training for Utterance-Level Dialogue Generation**
Yi Huang | Xiaoting Wu | Wei Hu | Junlan Feng | Chao Deng

pdf bib abs **Information Extraction and Human-Robot Dialogue towards Real-life Tasks A Baseline Study with the MobileCS Dataset**
Hong Liu | Hao Peng | Zhijian Ou | Juanzi Li | Yi Huang | Junlan Feng

pdf bib abs **A Generative User Simulator with GPT-based Architecture and Goal State Tracking for Reinforced Multi-Domain Dialog Systems**
Hong Liu | Yucheng Cai | Zhijian Ou | Yi Huang | Junlan Feng

pdf bib abs **Offline-to-Online Co-Evolutional User Simulator and Dialogue System**
Dafeng Chi | Yuzheng Zhuang | Yao Mu | Bin Wang | Jianzhu Bao | Yasheng Wang | Yuhan Dong | Xin Jiang | Qun Liu | Jianye Hao

↑ up

pdf (full) bib (full)

Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)

pdf bib **Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)**
Angela Fan | Iryna Gurevych | Yufang Hou | Zornitsa Kozareva | Sasha Luccioni | Nafise Sadat Moosavi | Sujith Ravi | Gyuwan Kim | Roy Schwartz | Andreas Rücklé

pdf bib abs **Efficient Two-Stage Progressive Quantization of BERT**
Charles Le | Arash Ardakani | Amir Ardakani | Hang Zhang | Yuyan Chen | James Clark | Brett Meyer | Warren Gross

pdf bib abs **KGRefiner: Knowledge Graph Refinement for Improving Accuracy of Translational Link Prediction Methods**
Mohammad Javad Saeedizade | Najmeh Torabian | Behrouz Minaei-Bidgoli

pdf bib abs **Algorithmic Diversity and Tiny Models: Comparing Binary Networks and the Fruit Fly Algorithm on Document Representation Tasks**
Tanise Ceron | Nhut Truong | Aurelie Herbelot

pdf bib abs **Look Ma, Only 400 Samples! Revisiting the Effectiveness of Automatic N-Gram Rule Generation for Spelling Normalization in Filipino**
Lorenzo Jaime Flores | Dragomir Radev

pdf bib abs **Who Says Elephants Can't Run: Bringing Large Scale MoE Models into Cloud Scale Production**
Young Jin Kim | Rawn Henry | Raffy Fahim | Hany Hassan

pdf bib abs **Data-Efficient Auto-Regressive Document Retrieval for Fact Verification**
James Thorne

pdf bib abs **AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages**
Bonaventure F. P. Dossou | Atnafu Lambebo Tonja | Oreen Yousuf | Salomey Osei | Abigail Oppong | Iyanuoluwa Shode | Oluwabusayo Olufunke Awoyomi | Chris Emezue

pdf bib abs **Towards Fair Dataset Distillation for Text Classification**
Xudong Han | Aili Shen | Yitong Li | Lea Frermann | Timothy Baldwin | Trevor Cohn

↑ up

pdf (full) bib (full)

Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)

pdf bib **Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)**
Wenjuan Han | Zilong Zheng | Zhouhan Lin | Lifeng Jin | Yikang Shen | Yoon Kim | Kewei Tu

pdf bib abs **Named Entity Recognition as Structured Span Prediction**
Urchade Zaratiana | Nadi Tomeh | Pierre Holat | Thierry Charnois

pdf bib abs **Global Span Selection for Named Entity Recognition**
Urchade Zaratiana | Niama Elkhbir | Pierre Holat | Nadi Tomeh | Thierry Charnois

pdf bib abs **Visual Grounding of Inter-lingual Word-Embeddings**
Wafaa Mohammed | Hassan Shahmohammadi | Hendrik P. A. Lensch | R. Harald Baayen

pdf bib abs **A Subspace-Based Analysis of Structured and Unstructured Representations in Image-Text Retrieval**
Erica K. Shimomoto | Edison Marrese-Taylor | Hiroya Takamura | Ichiro Kobayashi | Yusuke Miyao

pdf bib abs **Discourse Relation Embeddings: Representing the Relations between Discourse Segments in Social Media**
Youngseo Son | Vasudha Varadarajan | H. Andrew Schwartz

pdf bib abs **Understanding Cross-modal Interactions in V&L Models that Generate Scene Descriptions**
Michele Cafagna | Kees van Deemter | Albert Gatt

pdf bib abs **DeepParliament: A Legal domain Benchmark & Dataset for Parliament Bills Prediction**
Ankit Pal

pdf bib abs **Punctuation and case restoration in code mixed Indian languages**
Subhashree Tripathy | Ashis Samal

pdf bib abs **Probing Script Knowledge from Pre-Trained Models**
Zijia Jin | Xingyu Zhang | Mo Yu | Lifu Huang

↑ up

pdf (full)
bib (full)

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)

pdf bib **Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)**
Houda Bouamor | Hend Al-Khalifa | Kareem Darwish | Owen Rambow | Fethi Bougares | Ahmed Abdelali | Nadi Tomeh | Salam Khalifa | Wajdi Zaghouani

pdf bib abs **CAraNER: The COVID-19 Arabic Named Entity Corpus**
Abdulmohsen Al-Thubaity | Sakhar Alkhereyf | Wejdan Alzahrani | Alia Bahanshal

pdf bib abs **Joint Coreference Resolution for Zeros and non-Zeros in Arabic**
Abdulrahman Aloraini | Sameer Pradhan | Massimo Poesio

pdf bib abs **SAIDS: A Novel Approach for Sentiment Analysis Informed of Dialect and Sarcasm**
Abdelrahman Kaseb | Mona Farouk

pdf bib abs **AraBART: a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization**
Moussa Kamal Eddine | Nadi Tomeh | Nizar Habash | Joseph Le Roux | Michalis Vazirgiannis

pdf bib abs **Towards Arabic Sentence Simplification via Classification and Generative Approaches**
Nouran Khallaf | Serge Sharoff | Rasha Soliman

pdf bib abs **Generating Classical Arabic Poetry using Pre-trained Models**
Nehal Elkaref | Mervat Abu-Elkheir | Maryam ElOraby | Mohamed Abdelgaber

pdf bib abs **A Benchmark Study of Contrastive Learning for Arabic Social Meaning**
Md Tawkat Islam Khondaker | El Moatez Billah Nagoudi | AbdelRahim Elmadany | Muhammad Abdul-Mageed | Laks Lakshmanan, V.S.

pdf bib abs **Adversarial Text-to-Speech for low-resource languages**
Ashraf Elneima | Mikołaj Bińkowski

pdf bib abs **NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task**
Muhammad Abdul-Mageed | Chiyu Zhang | AbdelRahim Elmadany | Houda Bouamor | Nizar Habash

pdf bib abs **The Shared Task on Gender Rewriting**
Bashar Alhafni | Nizar Habash | Houda Bouamor | Ossama Obeid | Sultan Alrowili | Daliyah AlZeer | Kawla Mohmad Shnqiti | Ahmed Elbakry | Muhammad ElNokrashy | Mohamed Gabr | Abderrahmane Issam | Abdelrahim Qaddoumi | Vijay Shanker | Mahmoud Zyate

pdf bib abs **Overview of the WANLP 2022 Shared Task on Propaganda Detection in Arabic**
Firoj Alam | Hamdy Mubarak | Wajdi Zaghouani | Giovanni Da San Martino | Preslav Nakov

pdf bib abs **ArzEn-ST: A Three-way Speech Translation Corpus for Code-Switched Egyptian Arabic-English**
Injy Hamed | Nizar Habash | Slim Abdennadher | Ngoc Thang Vu

pdf bib abs **Maknuune: A Large Open Palestinian Arabic Lexicon**
Shahd Salah Uddin Dibas | Christian Khairallah | Nizar Habash | Omar Fayez Sadi | Tariq Sairafy | Karmel Sarabta | Abrar Ardah

pdf bib abs **Developing a Tag-Set and Extracting the Morphological Lexicons to Build a Morphological Analyzer for Egyptian Arabic**
Amany Fashwan | Sameh Alansary

pdf bib abs **A Weak Supervised Transfer Learning Approach for Sentiment Analysis to the Kuwaiti Dialect**
Fatemah Husain | Hana Al-Ostad | Halima Omar

pdf bib abs **Mawqif: A Multi-label Arabic Dataset for Target-specific Stance Detection**
Nora Saleh Alturayeif | Hamzah Abdullah Luqman | Moataz Aly Kamaleldin Ahmed

pdf bib abs **Assessing the Linguistic Knowledge in Arabic Pre-trained Language Models Using Minimal Pairs**
Wafa Abdullah Alrajhi | Hend Al-Khalifa | Abdulmalik AlSalman

pdf bib abs **Identifying Code-switching in Arabizi**
Safaa Shehadi | Shuly Wintner

pdf bib abs **Authorship Verification for Arabic Short Texts Using Arabic Knowledge-Base Model (AraKB)**
Fatimah Alqahtani | Helen Yannakoudakis

pdf bib abs **A Semi-supervised Approach for a Better Translation of Sentiment in Dialectical Arabic UGT**
Hadeel Saadany | Constantin Orăsan | Emad Mohamed | Ashraf Tantawy

pdf bib abs **Cross-lingual transfer for low-resource Arabic language understanding**
Khadige Abboud | Olga Golovneva | Christopher DiPersio

pdf bib abs **Improving POS Tagging for Arabic Dialects on Out-of-Domain Texts**

Noor Abo Mokh | Daniel Dakota | Sandra Kübler

pdf bib abs **Domain Adaptation for Arabic Crisis Response**
Reem Alrashdi | Simon O'Keefe

pdf bib abs **Weakly and Semi-Supervised Learning for Arabic Text Classification using Monodialectal Language Models**
Reem AlYami | Rabah Al-Zaidy

pdf bib abs **Event-Based Knowledge MLM for Arabic Event Detection**
Asma Z Yamani | Amjad K Alsulami | Rabeah A Al-Zaidy

pdf bib abs **Establishing a Baseline for Arabic Patents Classification: A Comparison of Twelve Approaches**
Taif Omar Al-Omar | Hend Al-Khalifa | Rawan Al-Matham

pdf bib abs **Towards Learning Arabic Morphophonology**
Salam Khalifa | Jordan Kodner | Owen Rambow

pdf bib abs **AraDepSu: Detecting Depression and Suicidal Ideation in Arabic Tweets Using Transformers**
Mariam Hassib | Nancy Hossam | Jolie Sameh | Marwan Torki

pdf bib abs **End-to-End Speech Translation of Arabic to English Broadcast News**
Fethi Bougares | Salim Jouili

pdf bib abs **Arabic Keyphrase Extraction: Enhancing Deep Learning Models with Pre-trained Contextual Embedding and External Features**
Randah Alharbi | Husni Al-Muhtasab

pdf bib abs **ArabIE: Joint Entity, Relation and Event Extraction for Arabic**
Niama El Khbir | Nadi Tomeh | Thierry Charnois

pdf bib abs **Emoji Sentiment Roles for Sentiment Analysis: A Case Study in Arabic Texts**
Shatha Ali A. Hakami | Robert Hendley | Phillip Smith

pdf bib abs **Gulf Arabic Diacritization: Guidelines, Initial Dataset, and Results**
Nouf Alabbasi | Mohamed Al-Badrashiny | Maryam Aldahmani | Ahmed AlDhanhani | Abdullah Saleh Alhashmi | Fawaghy Ahmed Alhashmi | Khalid Al Hashemi | Rama Emad Alkhobbi | Shamma T Al Maazmi | Mohammed Ali Alyafeai | Mariam M Alzaabi | Mohamed Saqer Alzaabi | Fatma Khalid Badri | Kareem Darwish | Ehab Mansour Diab | Muhammad Morsy Elmallah | Amira Ayman Elnashar | Ashraf Hatim Elneima | MHD Tameem Kabbani | Nour Rabih | Ahmad Saad | Ammar Mamoun Sousou

pdf bib abs **Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions**
Saied Alshahrani | Esma Wali | Jeanna Matthews

pdf bib abs **A Pilot Study on the Collection and Computational Analysis of Linguistic Differences Amongst Men and Women in a Kuwaiti Arabic WhatsApp Dataset**
Hesah Aldihan | Robert Gaizauskas | Susan Fitzmaurice

pdf bib abs **Beyond Arabic: Software for Perso-Arabic Script Manipulation**
Alexander Gutkin | Cibu Johny | Raiomond Doctor | Brian Roark | Richard Sproat

pdf bib abs **Coreference Annotation of an Arabic Corpus using a Virtual World Game**
Wateen Abdullah Aliady | Abdulrahman Aloraini | Christopher Madge | Juntao Yu | Richard Bartle | Massimo Poesio

pdf bib abs **NatiQ: An End-to-end Text-to-Speech System for Arabic**
Ahmed Abdelali | Nadir Durrani | Cenk Demiroglu | Fahim Dalvi | Hamdy Mubarak | Kareem Darwish

pdf bib abs **The Effect of Arabic Dialect Familiarity on Data Annotation**
Ibrahim Abu Farha | Walid Magdy

pdf bib abs **Optimizing Naive Bayes for Arabic Dialect Identification**
Tommi Jauhiainen | Heidi Jauhiainen | Krister Lindén

pdf bib abs **iCompass Working Notes for the Nuanced Arabic Dialect Identification Shared task**
Abir Messaoudi | Chayma Fourati | Hatem Haddad | Moez BenHajhmida

pdf bib abs **TF-IDF or Transformers for Arabic Dialect Identification? ITFLOWS participation in the NADI 2022 Shared Task**
Fouad Shammary | Yiyi Chen | Zsolt T Kardkovacs | Mehwish Alam | Haithem Afli

pdf bib abs **Domain-Adapted BERT-based Models for Nuanced Arabic Dialect Identification and Tweet Sentiment Analysis**
Giyaseddin Bayrak | Abdul Majeed Issifu

pdf bib abs **Benchmarking transfer learning approaches for sentiment analysis of Arabic dialect**
Emna Fsih | Sameh Kchaou | Rahma Boujelbane | Lamia Hadrich-Belguith

pdf bib abs **SQU-CS @ NADI 2022: Dialectal Arabic Identification using One-vs-One Classification with TF-IDF Weights Computed on**

Character n-grams
Abdulrahman Khalifa AAlAbdulsalam

pdf bib abs **Ahmed and Khalil at NADI 2022: Transfer Learning and Addressing Class Imbalance for Arabic Dialect Identification and Sentiment Analysis**
Ahmed Oumar | Khalil Mrini

pdf bib abs **Arabic Sentiment Analysis by Pretrained Ensemble**
Abdelrahim Qaddoumi

pdf bib abs **Dialect & Sentiment Identification in Nuanced Arabic Tweets Using an Ensemble of Prompt-based, Fine-tuned, and Multitask BERT-Based Models**
Reem Abdel-Salam

pdf bib abs **On The Arabic Dialects' Identification: Overcoming Challenges of Geographical Similarities Between Arabic dialects and Imbalanced Datasets**
Salma Jamal | Aly M .Kassem | Omar Mohamed | Ali Ashraf

pdf bib abs **Arabic dialect identification using machine learning and transformer-based models: Submission to the NADI 2022 Shared Task**
Nouf AlShenaifi | Aqil Azmi

pdf bib abs **NLP DI at NADI Shared Task Subtask-1: Sub-word Level Convolutional Neural Models and Pre-trained Binary Classifiers for Dialect Identification**
Vani Kanjirangat | Tanja Samardzic | Ljiljana Dolamic | Fabio Rinaldi

pdf bib abs **Word Representation Models for Arabic Dialect Identification**
Mahmoud Sobhy | Ahmed H. Abu El-Atta | Ahmed A. El-Sawy | Hamada Nayel

pdf bib abs **Building an Ensemble of Transformer Models for Arabic Dialect Classification and Sentiment Analysis**
Abdullah Salem Khered | Ingy Yasser Hassan Abdou Abdelhalim | Riza Batista-Navarro

pdf bib abs **Arabic Dialect Identification and Sentiment Classification using Transformer-based Models**
Joseph Attieh | Fadi Hassan

pdf bib abs **Generative Approach for Gender-Rewriting Task with ArabicT5**
Sultan Alrowili | Vijay Shanker

pdf bib abs **AraProp at WANLP 2022 Shared Task: Leveraging Pre-Trained Language Models for Arabic Propaganda Detection**
Gaurav Singh

pdf bib abs **TUB at WANLP22 Shared Task: Using Semantic Similarity for Propaganda Detection in Arabic**
Salar Mohtaj | Sebastian Möller

pdf bib abs **SI2M & AIOX Labs at WANLP 2022 Shared Task: Propaganda Detection in Arabic, A Data Augmentation and Name Entity Recognition Approach**
Kamel Gaanoun | Imade Benelallam

pdf bib abs **iCompass at WANLP 2022 Shared Task: ARBERT and MARBERT for Multilabel Propaganda Classification of Arabic Tweets**
Bilel - Taboubi | Bechir Brahem | Hatem Haddad

pdf bib abs **ChavanKane at WANLP 2022 Shared Task: Large Language Models for Multi-label Propaganda Detection**
Tanmay Chavan | Aditya Manish Kane

pdf bib abs **AraBERT Model for Propaganda Detection**
Mohamad Sharara | Wissam Mohamad | Ralph Tawil | Ralph Chobok | Wolf Assi | Antonio Tannoury

pdf bib abs **AraBEM at WANLP 2022 Shared Task: Propaganda Detection in Arabic Tweets**
Eshrag Ali Refaee | Basem Ahmed | Motaz Saad

pdf bib abs **IITD at WANLP 2022 Shared Task: Multilingual Multi-Granularity Network for Propaganda Detection**
Shubham Mittal | Preslav Nakov

pdf bib abs **Pythoneers at WANLP 2022 Shared Task: Monolingual AraBERT for Arabic Propaganda Detection and Span Extraction**
Joseph Attieh | Fadi Hassan

pdf bib abs **CNLP-NITS-PP at WANLP 2022 Shared Task: Propaganda Detection in Arabic using Data Augmentation and AraBERT Pre-trained Model**
Sahinur Rahman Laskar | Rahul Singh | Abdullah Faiz Ur Rahman Khilji | Riyanka Manna | Partha Pakray | Sivaji Bandyopadhyay

pdf bib abs **NGU CNLP atWANLP 2022 Shared Task: Propaganda Detection in Arabic**
Ahmed Samir Hussein | Abu Bakr Soliman Mohammad | Mohamed Ibrahim | Laila Hesham Afify | Samhaa R. El-Beltagy

ACL materials are Copyright © 1963–2023 ACL; other materials are copyrighted by their respective copyright holders. Materials prior to 2016 here are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. Permission is granted to make copies for the purposes of teaching and research. Materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

The ACL Anthology is managed and built by the ACL Anthology team of volunteers.

Site last built on 16 March 2023 at 01:35 UTC with commit a06287ce.

https://web.archive.org/web/20230202011533/https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

INTERNET ARCHIVE
WayBackMachine
https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
Go
JAN FEB MAR
02
2022 2023 2024
2,610 capture
1 Feb 2023 - 6 Fe
About this capture

Exclusive news, data and analytics for financial market professionals REFINITIV

Technology

2 minute read · February 1, 2023 11:09 AM PST · Last Updated 6 hours ago

ChatGPT sets record for fastest-growing user base - analyst note

By Krystal Hu

Feb 1 (Reuters) - ChatGPT, the popular chatbot from OpenAI, is estimated to have reached 100 million monthly active users in January, just two months after launch, making it the fastest-growing consumer application in history, according to a UBS study on Wednesday.

The report, citing data from analytics firm Similar Web, said about 13 million unique visitors used ChatGPT per day in January, more than double the levels in December.

Advertisement · Scroll to continue

"In 20 years following the internet space, we cannot recall a faster ramp in a consumer internet app," UBS analysts wrote in the note.



Register for free to Reuters and know the full story
Register now
Exhibit 1027 Pa



Similar Web did not immediately respond to a request to confirm the figures. OpenAI, a private company backed by Microsoft Corp (MSFT.O), launched ChatGPT in late November.

It took TikTok about nine months after its global launch to add 100 million users and it took Instagram 2-1/2 years, according to data from Sensor Tower.

Latest Updates

- **Meta shares soar as company cuts spending, forecasts upbeat first-quarter sales**
- **U.S. investors have plowed billions into China's AI sector, report shows**
- **Apple set to post rare revenue drop as focus shifts to demand rebound**
- **As Tesla ignites an EV price war, suppliers brace for Musk seeking givebacks**
- **Musk did not need Tesla board to review buyout tweets, directors testify**

ChatGPT is currently free to use. The company on Thursday launched a $20 a month subscription that provides more stable and faster service as well as the opportunity to try new features first.

Analysts believe the viral launch of ChatGPT will give OpenAI a first-mover advantage against other AI companies. The growing usage, while adding up to a substantial amount of computing cost, has also provided valuable feedback to help train the chatbot's responses.

Advertisement · Scroll to continue

The tool has also raised questions about academic dishonesty and misinformation.

Last month, Microsoft announced another multi-billion investment in OpenAI in the form of cash and cloud credit.



Register for free to Reuters and know the full story

Register now

Reporting by Krystal Hu in Toronto; Editing by Cynthia Osterman

Our Standards: The Thomson Reuters Trust Principles.



Read Next

Technology
Intel slashes employee, exec pay amid PC market downturn

Technology
Meta shares soar as company cuts spending, forecasts upbeat first-quarter sales

Technology
Biden moves to slash U.S. credit card fees, app charges

Technology
Exclusive: ' Shanghai (price cuts s memo

Newsletter | Daily.

Technology Roundup

From startups to the FAANGs, get the latest news and trends in the global technology industry.

Sign up

More from Reuters

FEATURED VIDEOS FROM THE VERGE



Technology



Republican U.S. lawmaker meets with TikTok, but unpersuaded

Technology · February 1, 2023 · 4:28 PM PST · an hour ago

U.S. Republican Representative Mike Gallagher met with TikTok on



Future of Money
Britain's finance ministry sets out draft rules to regulate cryptoassets
4:30 AM PST

Technology
GoodRx pays $1.5 million to settle health privacy allegations
9:05 AM PST

Technology

Technology

U.S. Republican Representative Mike Gallagher met with TikTok on Wednesday about the Chinese-owned short video app's U.S. data security plans but still plans to seek to ban TikTok in the United States, a spokesperson for the lawmaker said.

PayPal's rocky road ahead warrants more cost cuts, Wall Street says

5:05 AM PST

Pinterest cuts about 150 jobs - Bloomberg News

4:56 PM PST . Updated 20 min ago

Information you can trust

Reuters, the news and media division of Thomson Reuters, is the world's largest multimedia news provider, reaching billions of people worldwide every day. Reuters provides business, financial, national and international news to professionals via desktop terminals, the world's media organizations, industry events and directly to consumers.

Follow Us

Thomson Reuters Products

Westlaw
Build the strongest argument relying on authoritative content, attorney-editor expertise, and industry defining technology.

Onesource
The most comprehensive solution to manage all your complex and ever-expanding tax and compliance needs.

Checkpoint
The industry leader for online information for tax, accounting and finance professionals.

Refinitiv Products

Refinitiv Workspace
Access unmatched financial data, news and content in a highly-customised workflow experience on desktop, web and mobile.

Refinitiv Data Catalogue
Browse an unrivalled portfolio of real-time and historical market data and insights from worldwide sources and experts.

Re
Scr
ind
glo
hid
rela
net

Advertise With Us **Advertising Guidelines** **Cookies** **Terms of Use** **Privacy** **Digital Accessibility** **Corrections** **Site Feedback** **FAQ**

All quotes delayed a minimum of 15 minutes. See here for a complete list of exchanges and delays.

© 2023 Reuters. All rights reserved

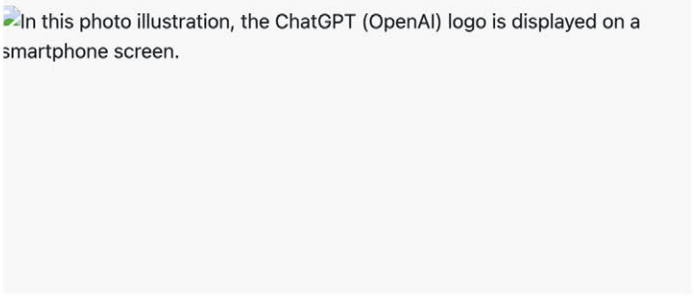
https://web.archive.org/web/20230203044225/https://www.businessinsider.com/chatgpt-may-be-fastest-growing-app-in-history-ubs-study-2023-2

INTERNET ARCHIVE
Wayback Machine
https://www.businessinsider.com/chatgpt-may-be-fastest-growing-app-in-history-ubs-study-2023-2
Go
JAN FEB MAR
03
2022 2023 2024
109 captures
2 Feb 2023 - 7 Fe
About this capture

US MARKETS CLOSED
In the news
Dow Jones -1.34%
Nasdaq -2.04%
S&P 500 -1.57%
AAPL -0.6%
NVDA -0.06%

ChatGPT may be the fastest-growing consumer app in internet history, reaching 100 million users in just over 2 months, UBS report says

Sawdah Bhaimiya Feb 2, 2023, 3:45 AM



In this photo illustration, the ChatGPT (OpenAI) logo is displayed on a smartphone screen. Sopa Images/Getty Images

- **ChatGPT may be the fastest-growing app in history with 100 million users after just over two months.**
- **A UBS report showed that it took TikTok over nine months to reach the same number of people.**
- **OpenAI is rolling out a $20-a-month subscription to ChatGPT for easy access and faster responses.**

Sign up for our newsletter for the latest tech news and scoops — delivered daily to your inbox.

Email address
SIGN UP

Email address

SIGN UP

By clicking 'Sign up', you agree to receive marketing emails from Insider as well as other partner offers and accept our Terms of Service and Privacy Policy.

ChatGPT may have become the fastest-growing app in history after it was estimated to have reached 100 million monthly active users, or MAUs, by the end of January, a report by Swiss banking giant UBS said.

The OpenAI-owned chatbot has surpassed 100 million MAUs just over two months after its launch — comfortably beating the growth speeds of major apps like TikTok and Instagram, according to the UBS note, which Insider has viewed.

RECOMMENDED VIDEO

It took TikTok nine months to reach the same level of users; Instagram took two and half years; and Spotify only amassed 100 million MAUs after four and a half years, the report said, citing data from data analytics firm Similar Web.

"In twenty years following the Internet space, we cannot recall a faster ramp in a consumer internet app," UBS analysts wrote in the report.

The number of MAUs on ChatGPT has almost doubled since December when it had around 57 million, per the report. Additionally the app was averaging 13 million daily unique visitors per day by the end of January, compared to

around six million daily unique views the prior month.

"We cannot remember an app scaling at this pace," UBS analysts added in the report.

OpenAI did not immediately respond to Insider's request for comment about the report, sent outside of normal US business hours.

The chatbot was launched in November and experienced a rapid surge in popularity as a result of its eerie human-like capabilities to code, write cover letters, and take exams. It amassed one million users in just five days.

Shortly afterwards, Microsoft extended its support of OpenAI through a multibillion-dollar investment in the AI firm, announced in January. The tech giant had already invested $1 billion into the firm in 2019.

INSIDER BUSINESS INSIDER

Download on the App Store GET IT ON Google Play

* Copyright © 2023 Insider Inc. All rights reserved. Registration on or use of this site constitutes acceptance of our Terms of Service, Privacy Policy and Cookies Policy.

Contact Us | Sitemap | Disclaimer | Accessibility | Commerce Policy | Advertising Policies | CA Privacy Rights | Coupons | Made in NYC | Jobs @ Insider

Stock quotes by finanzen.net | Reprints & Permissions

International Editions: INTL | AS | AT | DE | ES | IN | JP | MX | NL | PL | ZA

Insider.com™ Insider Inc.

BUSINESS INSIDER TECH INSIDER

https://web.archive.org/web/20230930082124/ahdictionary.com/word/search.html?q=between

INTERNET ARCHIVE Wayback Machine
http://ahdictionary.com/word/search.html?q=between
Go
JUL SEP JAN
34 captures
19 Oct 2014 - 26
2022 30 2023 2024
About this capture

be·tween (bĭ-twēn′)

Share:

prep.

1. **a.** In or through the position or interval separating: *between the trees; between 11 o'clock and 12 o'clock.*
   b. Intermediate to, as in quantity, amount, or degree: *It costs between 15 and 20 dollars.*
2. Connecting spatially: *a railroad between the two cities.*
3. Associating or uniting in a reciprocal action or relationship: *an agreement between workers and management; a certain resemblance between the two stories.*
4. In confidence restricted to: *Between you and me, he is not qualified.*
5. **a.** By the combined effort or effect of: *Between them they succeeded.*
   b. In the combined ownership of: *They had only a few dollars between them.*
6. As measured against. Often used to express a reciprocal relationship: *choose between riding and walking.*

adv.

In an intermediate space, position, or time; in the interim.

Idioms:

in between
In an intermediate situation: *My roommates disagreed and I was caught in between.*

in between times
During an intervening period; in the meantime: *has written several books and teaches in between times.*

[Middle English *bitwene*, from Old English *betwēonum*; see **dwo-** in the Appendix of Indo-European roots.]

between′ness *n.*

Usage Note: The *–tween* in *between* comes from the same Indo-European root that gave us *two, twain,* and *duo,* and the *–mong* of *among* comes from an Old English word that meant "crowd" or "throng." It is thus unsurprising that a traditional rule requires *between* to be used only for sentences involving two items and *among* for sentences involving more than two. Indeed, in sentences involving two items, no rule is needed; native English speakers spontaneously use *between* (as in *the differences between* [not *among*] *karate and judo*). But when there are more than two items, practice is mixed. Many careful writers observe a more subtle distinction, using *among* when the sentence refers to the entities collectively or as a mass, as in *There were many outstanding players among the teams in the quarterfinal round* or *A thistle is growing among the roses,* but preferring *between* when the sentence refers to relationships involving particular pairs of entities from within the group, as in *We haven't yet assigned the matchups between teams in the quarterfinal round* or *I have sand between my toes.* In such sentences, the twoness of *between* has not, so to speak, been lost in the crowd—the pairings within the larger group are important to the meaning of the sentence and thus influence the writer's choice of preposition.

The American Heritage® Dictionary of the English Language, Fifth Edition copyright ©2022 by HarperCollins Publishers. All rights reserved.

HOW TO USE THE DICTIONARY

To look up an entry in *The American Heritage Dictionary of the English Language,* use the search window above. For best results, after typing in the word, click on the "Search" button instead of using the "enter" key.

Some compound words (like *bus rapid transit, dog whistle,* or *identity theft*) don't appear on the drop-down list when you type them in the search bar. For best results with compound words, place a quotation mark before the compound word in the search window.

GUIDE TO THE DICTIONARY

THE USAGE PANEL

The Usage Panel is a group of nearly 200 prominent scholars, creative writers, journalists, diplomats, and others in occupations requiring mastery of language. Annual surveys have gauged the acceptability of particular usages and grammatical constructions.

THE PANELISTS

AMERICAN HERITAGE DICTIONARY APP

The new American Heritage Dictionary app is now available for iOS and Android.

THE AMERICAN HERITAGE DICTIONARY BLOG

The articles in our blog examine new words, revised definitions, interesting images from the fifth edition, discussions of usage, and more.

THE 100 WORDS®

See word lists from the best-selling 100 Words Series!

FIND OUT MORE!

INTERESTED IN DICTIONARIES?

Check out the Dictionary Society of North America at http://www.dictionarysociety.com

Indo-European & Semitic Roots Appendices

Thousands of entries in the dictionary include etymologies that trace their origins back to reconstructed proto-languages. You can obtain more information about these forms in our online appendices:

Indo-European Roots

Semitic Roots

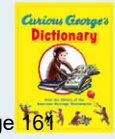
The Indo-European appendix covers nearly half of the Indo-European roots that have left their mark on English words. A more complete treatment of Indo-European roots and the English words derived from them is available in our **Dictionary of Indo-European Roots**.

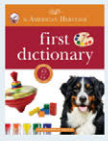
American Heritage Dictionary Products











The American Heritage... The American Heritage... The American Heritage... Curious... The American Heritage...

Heritage Dictionary, 5th Edition
Heritage Dictionary of Idioms
Heritage Roget's Thesaurus
George's Dictionary
Heritage Children's Dictionary

CONTACT US

Customer Service
Make Me An Author
Ebooks Help with Glose Reader

ABOUT US

Company Profile
Leadership Team
Corporate Social Responsibility
HarperCollins Careers
HarperCollins Imprints
HarperGreen
Social Media Directory
Accessibility

FOR READERS

Browse Reading Guides

FOR AUTHORS

Submit a Manuscript
Report Piracy
Agent Portal

MEDIA

Publicity Contacts
Press Room

SERVICES

HarperCollins Speakers Bureau
Library Services
Academic Services
Desk & Exam Copies
Review Copies
OpenBook API
Marketing Partnerships

COVID-19 RESOURCES & PERMISSIONS

Permissions for Adult Online Readings
Permissions for Kids Online Readings

SALES & RIGHTS

Booksellers & Retailer Ordering
HarperCollins Catalogs
Permissions
Subsidiary Rights
Media Rights and Content Development

GLOSE APP

iPhone
Android

GLOBAL DIVISIONS

HarperCollins US
HarperCollins Canada
HarperCollins Christian
HarperCollins Australia
HarperCollins India
HarperCollins UK

HarperCollins Publishers
News Corp

Terms of Use • Terms of Sale • Your Ad Choices • Privacy Policy • California Privacy Policy
Do Not Sell My Personal Information

Copyright 2022 HarperCollins Publishers All rights reserved.

*This website is best viewed in Chrome, Firefox, Microsoft Edge, or Safari. Some characters in pronunciations and etymologies cannot be displayed properly in Internet Explorer.

https://web.archive.org/web/20220322024512/https://dictionary.cambridge.org/dictionary/english/between

INTERNET ARCHIVE Wayback Machine https://dictionary.cambridge.org/dictionary/english/between Go JUN MAR JUN 22 2020 2022 2024 52 captures 24 Aug 2015 - 7 About this capture

*Meaning of **between** in English*

between

preposition**, **adverb

UK /bɪˈtwiːn/ US /bɪˈtwiːn/

between *preposition*, *adverb* (SPACE)

A1

in or into the space that separates two places, people, or objects:

- *The town lies halfway between Rome and Florence.*
- *Standing between the two adults was a small child.*
- *She squeezed between the parked cars and ran out into the road.*
- *There were two houses with a narrow path **in** between.*

− More examples

- *On the train I was sitting between two very large men.*
- *The plants have seeded themselves in the cracks between the paving stones.*
- *There was a revolting green slime in between the bathroom tiles.*
- *We watched a spider spin a web between three tall grass stems.*
- *Our neighbours have no claim to that strip of land between our houses.*

+ SMART Vocabulary: related words and phrases

between *preposition*, *adverb* (AMOUNT)

A2

If something is between two amounts, it is greater than the first amount but smaller than the second:

- *She weighs between 55 and 60 kilograms.*
- *The competition is open to children between six and twelve years of age.*
- *The room was either extremely cold or hot, never anything **in** between (= in the middle).*

+ SMART Vocabulary: related words and phrases

between *preposition*, *adverb* (TIME)

Test your vocabulary with our fun image quizzes

Image credits

Try a quiz now

More meanings of *between*

− All

go-between
in-between
length between
perpendiculars BETA
come between someone
between you and me *idiom*
(in) between times *idiom*
(just) between you and me *idiom*

See all meanings

+ Phrasal Verbs

+ Idioms

WORD OF THE DAY

A1

(also **in between**)

in the period of time that separates two different times or events:

- *You shouldn't eat between meals.*
- *There is a break of ten minutes between classes.*
- *The office is closed for lunch between 12.30 and 1.30.*
- *In between sobs, he managed to tell them what had happened.*
- *He visits his parents every month and sometimes **in** between.*

+ More examples

+ SMART Vocabulary: related words and phrases

Grammar

Between or among?

Between and among are prepositions. …

Between, among: meaning and use

We use between to refer to two things which are clearly separated. We use among to talk about things which are not clearly separated because they are part of a group or crowd or mass of objects: …

Between and prepositional phrases

Between is most often used to introduce a prepositional phrase which contains two singular or plural noun phrases: …

Nouns and verbs followed by between

The most common nouns often followed by between are: connection, difference, distinction, link, relationship: …

Among meaning 'surrounding, part of or included in'

We use among to suggest a sense of being a part of or surrounded by or included in something else. It is typically followed by a plural noun phrase: …

Among others and among other things

In the phrases among others and among other things, among means 'as well as': …

Idiom

(in) between times

between

preposition

UK /bɪˈtwiːn/ US /bɪˈtwiːn/

between *preposition* **(SHARED)**

B1

among two or more people or things:

- *The money was divided equally between several worthy causes.*
- *We drank two bottles of wine between the four of us.*
- *Trade between the two countries (= their trade with each other) has increased sharply in the past year.*
- *There is a great deal of similarity between Caroline and her mother (= they are very similar).*

tall oaks from little acorns grow

said about organizations or plans that start off very small or simple and become extremely large or successful

About this



BLOG

Rhyme, rhythm and verse: talking about poetry

March 16, 2022

Read More



NEW WORDS

Earthing

March 21, 2022

More new words

mother (= they are very similar).

+ More examples

+ SMART Vocabulary: related words and phrases

between *preposition* **(OPPOSING)**

A1

A discussion, argument, or game between two or more people or groups of people involves both people or groups:

- *The negotiations between the union and management have broken down.*
- *There has always been a fierce rivalry between the two schools.*
- *Tonight's game is between the New Orleans Saints and the St. Louis Rams.*

+ SMART Vocabulary: related words and phrases

between *preposition* **(CHOICE)**

If you choose between two things, you choose one thing or the other:

- *You'll have to **choose** between a holiday or a new washing machine.*
- *She was **torn** between loyalty to her father and love for her husband (= she could not decide which one to support).*

+ SMART Vocabulary: related words and phrases

between *preposition* **(CONNECTING)**

A2

connecting two or more places, things, or people:

- *There is a regular train service between Glasgow and Edinburgh.*
- *The survey shows a link between asthma and air pollution.*

from one place to another:

- *He commutes daily between Baltimore and Washington, D.C.*

+ More examples

+ SMART Vocabulary: related words and phrases

between *preposition* **(SEPARATING)**

A2

separating two places or things:

- *The wall between East and West Berlin came down in 1989.*
- *The report states that the gap between the rich and the poor has increased dramatically over the past decade.*

+ More examples

+ **SMART Vocabulary: related words and phrases**

Grammar

Between or among?

Between and among are prepositions. …

Between, among: meaning and use

We use between to refer to two things which are clearly separated. We use among to talk about things which are not clearly separated because they are part of a group or crowd or mass of objects: …

Between and prepositional phrases

Between is most often used to introduce a prepositional phrase which contains two singular or plural noun phrases: …

Nouns and verbs followed by between

The most common nouns often followed by between are: connection, difference, distinction, link, relationship: …

Among meaning 'surrounding, part of or included in'

We use among to suggest a sense of being a part of or surrounded by or included in something else. It is typically followed by a plural noun phrase: …

Among others and among other things

In the phrases among others and among other things, among means 'as well as': …

Idiom

(just) between you and me

*(Definition of **between** from the **Cambridge Advanced Learner's Dictionary & Thesaurus** © Cambridge University Press)*

between | AMERICAN DICTIONARY

between

preposition, ***adverb***

US /bɪˈtwin/

(also **in between**, US /ˌɪn·bɪˈtwin/)

between *preposition*, *adverb* (SPACE)

[not gradable]

in or into the space that separates two places, people, or objects:

- *We live halfway between Toronto and Montreal.*
- *She squeezed in between the parked cars.*

[not gradable]

If something is between two amounts, it is greater than the first amount but smaller than the second:

- *She weighs between 55 and 60 pounds.*

between *preposition*, *adverb* (TIME)

in the period of time that separates two different times or events:

- *There's a ten-minute break between classes.*
- *You should arrive between 8 and 8:30*

- *In between sobs, he managed to tell them what had happened.*

between

preposition

US /bɪˈtwin/

between *preposition* (AMONG)

shared by or involving two or more people or things:

- *The money was divided equally between her three children.*
- *Trade between the two countries has increased sharply.*
- *There is a great deal of similarity between Caroline and her mother.*
- *Between the three of us, we were able to afford a nice graduation gift.*
- *You'll have to choose between dinner and a movie.*
- *Next week's game will be between the two finalists.*

between *preposition* (CONNECTING)

connecting two or more places, things, or people:

- *There is regular train service between New York and Philadelphia.*
- *The survey shows a link between asthma and air pollution.*

between *preposition* (SEPARATING)

separating two places or things:

- *The wall between East and West Berlin came down in 1989.*
- *What's the difference between this $100 watch and the $500 one (= In what way are they different)?*

Idiom

between you and me

*(Definition of **between** from the **Cambridge Academic Content Dictionary** © Cambridge University Press)*

EXAMPLES of **between**

between

The existence and activities of ' go-betweens ' are intrinsic to any process of colonisation, however defined.

From the Cambridge English Corpus

They can sometimes defuse a situation and sometimes act as unofficial go-betweens with the commanders of the potential rivals.

From the Hansard archive

Instead, we have had to rely on journalists to act as go-betweens, and on many an occasion they have been rather inadequate go-betweens.

From the Hansard archive

The passengers are the unfortunate go-betweens in this usually politically-inspired action.

From the Hansard archive

They are the go-betweens of the officers and the men, and they are the finest lot of people we have in the ships.

From the Hansard archive

However, one does not want to see this legislation tip the scales against the interests of the in-betweens, or marginals.

From the Hansard archive

The provision is also likely to catch negotiators, mediators and go-betweens of one kind or another who probably genuinely seek to resolve a conflict or defuse a situation.

From the Hansard archive

The correlation *between* the two sets of country-level means is strong though not perfect (r 0.69, p 0.001).

From the Cambridge English Corpus

Convergence *between* the ideologies of elites should reduce (w z)and such convergence is known to be one factor promoting stability.

From the Cambridge English Corpus

First, it is not a dichotomous variable and it allows for a great degree of variation *between* countries.

From the Cambridge English Corpus

Differences *between* neighbourhoods are both significant and substantial.

From the Cambridge English Corpus

We also examine the interaction *between* electoral rules and presidentialism.

From the Cambridge English Corpus

We focus on the party leader because there is a well-attested relationship *between* electoral evaluation of leaders and election outcomes.

From the Cambridge English Corpus

One might suspect a curvilinear pattern *between* campaign funds and duration of candidacies.

From the Cambridge English Corpus

One difference *between* parliamentarism and presidentialism is obvious; indeed, definitional.

From the Cambridge English Corpus

See all examples of between

These examples are from corpora and from sources on the web. Any opinions in the examples do not represent the opinion of the Cambridge Dictionary editors or of Cambridge University Press or its licensors.

Translations of **between**

in Chinese (Traditional)

空間, (空間上) 在...中間，在...之間, 數量...

See more

in Chinese (Simplified)

空间, (空间上) 在...中间，在...之间, 数量...

See more

in Spanish

entre, entremedias, en medio...

See more

Need a translator?

Get a quick, free translation!

Translator tool

in Portuguese

entre, no meio, no intervalo…

See more

in more languages

What is the pronunciation of *between*?

Browse

betters

betting

betting shop

bettor BETA

between

between you and me *idiom*

between you, me, and the gatepost *idiom*

betwixt

bevel

© Cambridge University Press 2022

https://web.archive.org/web/20250817160037/https://techcrunch.com/2023/03/14/openai-releases-gpt-4-ai-that-it-claims-is-state-of-the-art/

INTERNET ARCHIVE WayBack Machine https://techcr 534 captures 14 Mar 2023 - 3 I
INTERNET ARCHIVE WayBack Machine https://techcrunch.com/2023/03/14/openai-releases-gpt-4-ai-that-it-claims-is-state-of-the-art/ Go 534 captures 14 Mar 2023 - 3 Feb 2026
APR AUG NOV 2024 17 2025 2026 About this capture

Advertisement



IMAGE CREDITS: DANI PADGETT WATSON / STRICTLYVC

AI

OpenAI releases GPT-4, a multimodal AI that it claims is state-of-the-art

Kyle Wiggers — 10:06 AM PDT · March 14, 2023

OpenAI has released a powerful new image- and text-understanding AI model, GPT-4, that the company calls "the latest milestone in its effort in scaling up deep learning."

GPT-4 is available today to OpenAI's paying users via ChatGPT Plus (with a usage cap), and developers can sign up on a waitlist to access the API.

Pricing is $0.03 per 1,000 "prompt" tokens (about 750 words) and $0.06 per 1,000 "completion" tokens (again, about 750 words). Tokens represent raw text; for example, the word "fantastic" would be split into the tokens "fan," "tas" and "tic." Prompt tokens are the parts of words fed into GPT-4 while completion tokens are the content *generated* by GPT-4.

GPT-4 has been hiding in plain sight, as it turns out. Microsoft confirmed today that Bing Chat, its chatbot tech co-developed with OpenAI, is running on GPT-4.

Other early adopters include Stripe, which is using GPT-4 to scan business websites and deliver a summary to customer support staff. Duolingo built GPT-4 into a new language learning subscription tier. Morgan Stanley is creating a GPT-4-powered system that'll retrieve info from company documents and serve it up to financial analysts. And Khan Academy is leveraging GPT-4 to build some sort of automated tutor.

GPT-4's new capabilities power a 'virtual volunteer' for the visually impaired

Tech and VC heavyweights join the Disrupt 2025 agenda

Netflix, ElevenLabs, Wayve, Sequoia Capital, Elad Gil — just a few of the heavy hitters joining the Disrupt 2025 agenda. They're here to deliver the insights that fuel startup growth and sharpen your edge. Don't miss the 20th anniversary of TechCrunch Disrupt, and a chance to learn from the top voices in tech — grab your ticket now and save up to $600+ before prices rise.

San Francisco | October 27-29, 2025

REGISTER NOW

Advertisement

DISRUPT
October 27-29, 2025 | San Francisco
Put your brand in front of 10,000+ tech and VC leaders across all three days of Disrupt 2025. Amplify your reach, spark real connections, and lead the innovation charge. Secure your exhibit space before your competitor does.
Book Your Table ›

Most Popular

- How your solar rooftop became a national security issue
- Co-founder of Elon Musk's xAI departs the company
- Google CEO adds a new calendar feature at Stripe co-founder's request
- NASA has sparked a race to develop the data pipeline to Mars
- iOS 26 beta 6 adds new

GPT-4 can generate text and accept image and text inputs — an improvement over GPT-3.5, its predecessor, which only accepted text — and performs at "human level" on various professional and academic benchmarks. For example, GPT-4 passes a simulated bar exam with a score around the top 10% of test takers; in contrast, GPT-3.5's score was around the bottom 10%.

OpenAI spent six months "iteratively aligning" GPT-4 using lessons from an internal adversarial testing program as well as ChatGPT, resulting in "best-ever results" on factuality, steerability and refusing to go outside of guardrails, according to the company. Like previous GPT models, GPT-4 was trained using publicly available data, including from public webpages, as well as data that OpenAI licensed.

OpenAI worked with Microsoft to develop a "supercomputer" from the ground up in the Azure cloud, which was used to train GPT-4.

"In a casual conversation, the distinction between GPT-3.5 and GPT-4 can be subtle," OpenAI wrote in a blog post announcing GPT-4. "The difference comes out when the complexity of the task reaches a sufficient threshold — GPT-4 is more reliable, creative and able to handle much more nuanced instructions than GPT-3.5."

Without a doubt, one of GPT-4's more interesting aspects is its ability to understand images as well as text. GPT-4 can caption — and even interpret — relatively complex images, for example identifying a Lightning Cable adapter from a picture of a plugged-in iPhone.

The image understanding capability isn't available to all OpenAI customers just yet — OpenAI's testing it with a single partner, Be My Eyes, to start with. Powered by GPT-4, Be My Eyes' new Virtual Volunteer feature can answer questions about images sent to it. The company&#65533;?explains how it works in a blog post:

"For example, if a user sends a picture of the inside of their refrigerator, the Virtual Volunteer will not only be able to correctly identify what's in it, but also extrapolate and analyze what can be prepared with those ingredients. The tool can also then offer a number of recipes for those ingredients and send a step-by-step guide on how to make them."

A more meaningful improvement in GPT-4, potentially, is the aforementioned steerability tooling. With GPT-4, OpenAI is introducing a new API capability, "system" messages, that allow developers to prescribe style and task by describing specific directions. System messages, which will also come to ChatGPT in the future, are essentially instructions that set the tone — and establish boundaries — for the AI's next interactions.

For example, a system message might read: "You are a tutor that always responds in the Socratic style. You *never* give the student the answer, but always try to ask just the right question to help them learn to think for themselves. You should always tune your question to the interest and knowledge of the student, breaking down the problem into simpler parts until it's at just the right level for them."

Even with system messages and the other upgrades, however, OpenAI acknowledges that GPT-4 is far from perfect. It still "hallucinates" facts and makes reasoning errors, sometimes with great confidence. In one example cited by OpenAI, GPT-4 described Elvis Presley as the "son of an actor" — an obvious misstep.

"GPT-4 generally lacks knowledge of events that have occurred after the vast majority of its data cuts off (September 2021), and does not learn from its experience," OpenAI wrote. "It can sometimes make simple reasoning errors which do not seem to comport with competence across so many domains, or be overly gullible in accepting obvious false statements from a user. And sometimes it can fail at hard problems the same way humans do, such as introducing security vulnerabilities into code it produces."

OpenAI does note, though, that it made improvements in particular areas; GPT-4 is less likely to refuse requests on how to synthesize dangerous chemicals, for one. The company says that GPT-4 is 82% less likely overall to respond to requests for "disallowed" content compared to GPT-3.5 and responds to sensitive requests — e.g. medical advice and anything pertaining to self-harm — in accordance with OpenAI's policies 29% more often.

OpenAI GPT-4

ringtones, snappy app launches, and more

- Ford throws out Henry Ford's assembly line to make low-cost EVs in America
- The computer science dream has become a nightmare

Advertisement

IMAGE CREDITS: OPENAI

There's clearly a lot to unpack with GPT-4. But OpenAI, for its part, is forging full steam ahead — evidently confident in the enhancements it's made.

"We look forward to GPT-4 becoming a valuable tool in improving people's lives by powering many applications," OpenAI wrote. "There's still a lot of work to do, and we look forward to improving this model through the collective efforts of the community building on top of, exploring, and contributing to the model."

Topics: AI AI Artificial Intelligence (AI) Enterprise gpt-4 OpenAI

Advertisement



Kyle Wiggers
AI Editor

Kyle Wiggers was TechCrunch's AI Editor until June 2025. His writing has appeared in VentureBeat and Digital Trends, as well as a range of gadget blogs including Android Police, Android Authority, Droid-Life, and XDA-Developers. He...

View Bio >

Loading the next article

Error loading the next article

Advertisement

© 2025 TechCrunch Media LLC.

Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.

EXHIBIT B

https://web.archive.org/web/20170807183515/www.aclweb.org/anthology/U10-1005

Corpus Based Classification of Text in Australian Contracts

Michael Curtotti
Australian National University
Canberra, Australia
michael.curtotti@anu.edu.au

Eric Mccreath
Australian National University
Canberra, Australia
eric.mccreath@anu.edu.au

Abstract

Written contracts are a fundamental framework for commercial and cooperative transactions and relationships. Limited research has been published on the application of machine learning and natural language processing (NLP) to contracts. In this paper we report the classification of components of contract texts using machine learning and hand-coded methods. Authors studying a range of domains have found that combining machine learning and rule based approaches increases accuracy of machine learning. We find similar results which suggest the utility of considering leveraging hand coded classification rules for machine learning. We attained an average accuracy of 83.48% on a multi-class labelling task on 20 contracts combining machine learning and rule based approaches, increasing performance over machine learning alone.

1 Introduction

Contracts govern economic transactions and commercial or organisational relationships from trivial purchases to major national infrastructure projects. Any large organisation (whether private or public) must unavoidably invest significant resources in developing and concluding contracts, as the contracts it enters define its legal relations with organisations and individuals with which it interacts. As one author has noted, contracts are an integral part of any business enterprise and it is "difficult to overstate their importance to the business world" (Khoury and Yamouni, 2007, p16). In addition, drafting contracts represents a major economic activity for the legal industry.

Ambiguity is an unresolved problem in natural language processing. It is also an important problem affecting the drafting and operation of contracts. The costs of ambiguity in a contract can be enormous. The presence of ambiguity increases the risk of litigation, project failure and loss of commercial relationships. Instances of litigated ambiguity are used in drafting manuals as instructional tales for the unwary drafter (Aitken and Butt, 2004, p37).

Our future research goal is to investigate approaches that will help identify such ambiguity, however, a precursor for addressing the problem of ambiguity is accurate labelling or classification of the constituent parts of contract documents.[1] In this paper we report work on classifying 'lines' in contracts into 32 classes. We compare three different classification approaches in order to determine which approach may provide the best performance with respect to this task: supervised machine learning; a hand-crafted rule based tagger; or combining hand-crafted rules based tagging with machine learning.

2 Background

Natural language processing involves applying a pipeline of transformations to text. A typical first step is to segment the text into its sentences, after which further processing is applied (Jurafsky and Martin, 2009, p 69). This works well in usual English prose, but is not necessarily well suited to contracts because of the heavy use of layout to embody structure and semantics. In a contract, a sentence may occur within a single paragraph of text, or may be spread over several line breaks due to the use of sub-paragraphing as an aid to comprehension. Such sub-paragraphs may embody disjunctive or conjunctive conditions associated with

[1] This task can also be thought of as a document segmentation task, although we here consider it as a classification exercise.

Michael Curtotti and Eric McCreath. 2010. Michael Curtotti and Eric McCreath. In *Proceedings of Australasian Language Technology Association Workshop*, pages 18−26
Exhibit 1027 Page 178

a rule. A 'clause'[2] is the typical structural unit which embodies a legal rule or set of rules. Sometimes clauses are organised into sub-clauses and clauses and sub-clauses may be numbered in hierarchical fashion (e.g. 1, 1.1, (a), (i), (A))(Aitken and Butt, 2004, pp 23, 27). Line breaks often occur immediately before data such as headings, clauses, sub-clauses, party names, dates or execution blocks. Being able to label a line as to its character, key content or function thus enriches available information for later NLP. The line is thus a valuable starting point for applying NLP.

3 Related Work

3.1 Classification and Segmentation of Contracts and Legislative Documents

There would appear to be little published work on the application of natural language processing (NLP) or machine learning specifically for the classification of text in legal contracts. An exception to this is (Indukuri and Krishna, 2010) who apply a support vector machine over feature vectors extracted from contract clauses. These feature vectors are comprised of n-grams up to size n = 4. They classify contract sentences as clauses or non-clauses and then sub-classify clauses as concerned with payment terms or not. Indukuri and Krishna note the lack of published work on classification of sentences in contract documents for workflow and monitoring purposes.

While also limited, there has been some work on segmentation or classification of legislative corpora using machine learning. (Mencía, 2009) reports 100% precision and recall in machine learning on classification of legislation into articles, sections and parts (although to a task for which he is able to craft an equally accurate regular expression based segmenter). (Bacci et al., 2009) automatically classify plain text legislative documents for the purpose of XML mark up. (Francesconi, 2006), on whose work Bacci et al. build, describes four separate modules for dealing with plain text legacy content which they are seeking to convert to XML tagged text. (Hasan et al., 2008) carry out segmentation of Spanish legislative bulletins using the table of contents as an aid to segmentation. This work, in the legislative field, provides a parallel application domain for the work undertaken here in respect of contracts, although it is generally true that legislation will tend to conform to far tighter stylistic rules, because of the central control of legislative drafting.

3.2 Combining Machine Learning and Hand Coded Approaches

A number of researchers have combined hand-coded approaches and machine learning to produce better overall results.

(Kipp, 2006) augments handcrafted rules with machine learnt rules for gesture generation.

(McCreath and Kay, 2003) used such a combined approach to improve performance on categorizing emails.

(Park and Zhang, 2003) address the problem of text chunking for the Korean language, combining hand-crafted rules with a memory based machine learning method. A rule based method is used to determine whether an instance may constitute an exceptional case, with machine learning used to correct classifications assigned by the rule based method. They report up to a 2.83% improvement in F-scores over memory-based learning alone.

(Rochery et al., 2002) report the use of hand-crafted rules to enhance the accuracy of machine learning utilising boosting. They apply the method to a multi-labelling classification problem where they are seeking to classify spoken utterances. They attribute increased accuracy to the fact that the rules they develop are not represented in their data at all or do not appear sufficiently to have statistical impact during the training process. They note that the combination has the greatest effect when there are less then 100 examples, with the benefit of the combination decreasing as the number of training examples increases. The method therefore has direct applicability to reducing the data requirement for machine learning (a costly aspect of supervised machine learning).

(Takahashi et al., 2005) utilise a combination of SVM based learning and hand-crafted rules for the task of classifying occupations from social surveys. They effect a combination by using rule derived labels as features for the machine learner. Again the work occurs in a multi-class labelling context (with around 200 occupation codes). They find that SVM approaches are superior to rule-based approaches alone but that combining rule-based and SVM learning produces the best results. They also examine the effect of the number of training examples, noting that the differential ben-

[2]The term 'clause' here refers to a legal rule or rules appearing in a contract, rather than the linguistic entity.

Exhibit 1027 Page 179

efits of combining methods reduces with increasing data.

In our study we use class labels assigned by a hand-coded tagger as additional features for machine learning to obtain an increase in accuracy in the domain of text classification.

The work reported above, and our own paper, come from diverse domains and illustrate the application of combined rule/ML approaches using a variety of machine learning algorithms.

This suggests the wide utility of using hand-coded rules to improve accuracy of machine learning. Symetrically we may say that machine learning may potentially be used to improve the accuracy of hand-coded rules. Here we essentially follow a hybrid model of development that examines each method separately as well as in combination.

4 Data

4.1 Australian Contract Corpus

The data which forms the basis of this study is a corpus of 256 contracts ('the Australian Contract Corpus') which has been compiled from the web using the Google Australia web search: `http://www.google.com.au`.[3] The corpus is constituted of 1043364 tokens (words and punctuation), 42910 sentences, and a vocabulary of 14217. The lexical diversity of this corpus is 73.39.[4] As far as we are aware there is no published work based on a corpus of Australian contract language, however space does not permit a fuller discussion of the design and analysis of the corpus.

4.2 Data Representation

For the purpose of this study we randomly selected 30 contracts from the Australian Contract Corpus to produce data sets for application of machine learning. The 30 contracts were divided into three sets: a training set (Set A)[5], a second set (Set B)[6] (primarily used as a test set, although sometimes combined with Set A to form a larger training set) and an additional test set (Set C)[7], each of 10 contracts. Set C was added to increase both available data and the number of contracts on which testing could be carried out.

To simplify processing all contracts were converted into text files as a preprocessing step.[8] Also we removed material that typically does not appear in contracts (primarily guideline notes for drafters).

For purposes of our study each 'line' or 'paragraph' in a contract constituted a data point. Sets A, B and C provided 1825, 2157 and 2231 lines/data points, respectively.

To classify our data for machine learning we first employed our hand-coded tagger to produce a labelled data set constituted of a line and primary and secondary labels according to our classification system (see below). The classifications applied by the hand-coded tagger were then manually corrected to remove any labelling errors. Both for the reason that hand tagging data from scratch is more error prone than correcting previously machine tagged data, and because of the high labour in manual tagging, machine assisted tagging is a standard method used in corpus creation (Biber et al., 1998, p 262), ((ed) Christopher S. Butler, 1992, p 131).

4.3 Feature Selection

Features were then extracted from our labelled lines using a hand-coded feature extractor to extract up to 40 features to be used for machine learning, with the primary class label earlier applied to each line serving as the target for machine learning.[9] Features for machine learning were selected based on assumed relevance to the intended classification.[10] While a bag of word representa-

[3]The search phrase we employed was: "clause party agreement" , limiting the search to "pages from Australia" and the filetype to ".doc". The collection of the corpus was undertaken in the period 6 - 24 December 2009. Each document was visually inspected by one of the authors to verify that it was an Australian contract and documents were added to the corpus in order of their appearance in Google search results. We have made available the list of URLs of the documents that make up the corpus at: `http://cs.anu.edu.au/people/Michael.Curtotti/`. We also make code and data referred to in this paper available at the same location.

[4]Lexical diversity refers to the ratio of the total number of tokens in the corpus to the total number of types of tokens i.e. tokens/vocabulary. The statistics and information provided here were extracted using the NLTK. (Bird et al., 2009).

[5]Contracts numbered in our corpus: 55, 74, 77, 91, 94, 144, 174, 185, 208 and 213

[6]Contracts 11, 12, 164, 193, 196, 199, 249, 59, 64, 9

[7]Contracts 200, 254, 175, 180, 120, 102, 207, 127, 75 and 79

[8]Notably, this step does have the downside of discarding valuable style and layout information found in the word document format.

[9]In some instances the assignment of class is not entirely disjunctive and implies a priority ordering of potential classifications. For instance a clause may contain address details, and be classified as such rather than as a clause. The feature extractor removed secondary labels.

[10]The full list of features is as follows: 1. relative position of a line within the contract; 2. line end punctuation; 3. num-

Exhibit 1027 Page 180

tion with automatic feature extraction was trialed, this was found to result in lower accuracy than a manually selected (but small) number of features. Traditionally, 'feature engineering' or 'selection' in text classification has focussed on 'bag of words' or proximate features such as n-grams or noun phrases (Scott and Matwin, 1999). Our selection of features might be regarded as 'hybrid' in that they implicitly encode domain knowledge and 'rules'. For example, we expect a definition line to contain the widely used word 'means' as in 'x means y'. We would expect clause headings to display a different pattern of parts of speech occurrence, as compared to clauses themselves. We also consider features such as the relationship between the particular line/datapoint and its neighbouring lines (e.g. is the preceding line long or short), or the document as a whole (e.g. relative placement of the line within the document).

Features used in the related work carried out by others on contracts and legislation include: n-grams (Indukuri), bag of word together with capitalisation and character patterns (Mencia), the use of sequential word and token data as a HMM to predict state where state is the classification of the portion of text (Bacci/Francesconi) and the use of word terms identifying titles and lists of indexed terms (Hasan).

4.4 Classes

We adopted 32 classes for our classification scheme. These classes represented significant structural elements (such as clause headings or content lines) or sometimes key contract meta data (for instance the parties, the date on which an agreement is made, email addresses, ABNs etc).[11] Of particular interest to us, was an ability to accurately identify clausematter (the primary location of legal rules of a contract) and clause headings (which in some contracts effectively mark boundaries between major rule sets).

5 Experimental Evaluation

Analysis of our data sets was carried out using two major methods: analysis using the training Set A measuring peformance on test Set B (i.e. testing for accuracy of classification of lines in Set B); and analysis using training Set A where each contract in test Sets B and C was used individually as a test set (testing for accuracy of classification of lines of each individual contract). The former testing approach is more widely used but the latter is more consistent with our intended end application: that is, developing an ability to accurately classify lines in a previously unseen contract as an individual document.

Set A consisted of 1825 lines of data, Set B consisted of 2157 lines of data and Set C consisted of 2231 lines of data.

5.1 Tools

The following tools were used to carry out machine learning tasks reported in this paper:

1. the *python* programming language was used to develop a hand coded line tagger, a feature extractor and code for evaluation of the performance of the hand coded tagger;

2. the Natural Language Toolkit (NLTK)(Bird et al., 2009) was used in corpus development and in the hand coded tagger for application of parts of speech tags; and

3. the Weka data mining software(Hall et al., 2009) was used to carry out machine learning.

bering at the beginning of the line; 4. line length in characters; 5. line length in tokens; 6. number of nouns; 7. number of adjectives; 8. number of verbs; 9. number of prepositions; 10. number of coordinating conjunctions and cardinal numbers; 11. number of modal verbs; 12. number of personal pronouns, possessives; 13. number of adjectives; 14. whether must occurs; 15. whether may occurs; 16. whether shall occurs; 17. whether means occurs;18. the position of the word means in the line; 19. whether the word 'include' occurs; 20. the position of the word 'includes' in the line; 21. whether the phase 'has the same meaning' occurs; 22. whether the line begins with a capital letter followed by a stop or a space; 23. do the letters 'abn' or 'acn' appear; 24. does the word address appear near the beginning of the line; 25. does the word 'contact' appear near the beginning of the line; 26. does the word 'email' appear near the beginning of the line; 27. does the word 'fax' appear near the beginning of the line; 28. does the word 'note' appear at the beginning of the line; 29. does the word 'phone' appear at the beginning of the line; 30. do the terms 'web' or 'www' appear near the beginning of the line; 31-36. token lengths of 3 lines before and after the data line; 37. the tag applied by the hand coded tagger to the data line; 38. whether the contract from which the line comes has clause headings; 39. whether the contract from which the line comes has clause sub-headings; 40. whether the contract has a schedule.

[11] The following class labels were applied: CLAUSEMATTER, CLAUSEHEAD, DEF, CONTENTSHEAD, BLANK, DROSS, AND, BETWEEN, DATEMADELINE, PARTIESHEAD, PARTYLINE, RECITALHEAD, RECITALLINE, HEAD, CONTENTLINE, PRELIM, OPERATIVEPROVISIONLINE, NUMBEREDLINE, SCHEDULEHEAD, SCHEDULEITEM, EXECUTIONBLOCK, NOTE, ABNLINE, ADDRESSLINE, EMAILLINE, WEBLINE, FAXLINE, PH-LINE, REF-LINE, SCHOTHERMATTER, CONTACTOFFICER, TITLELINE.

Exhibit 1027 Page 181

5.2 Hand-Coded Tagger

Our hand-coded tagger follows a pipeline methodology. A contract is separated into a list of lines. In the first phase, each line is passed through a series of "if-then-else" routines (essentially utilising regular expressions) to assign primary and secondary labels to the line. This produces a contract with enhanced information content, in the form of a tuple of labels and line text. This output is then fed through a series of methods, progressively bootstrapping improvements to labelling, based on the information available when the method is applied. For instance content lines will be identified, these are then used in later methods to seek to identify clause headings. In addition, the tagger extracts contract meta data such as: a list of definitions, a list of clause headings, where the beginning of operative provisions occurs, where the schedule begins. Such features may or may not occur in a contract. In addition the tagger also provided a method for hand correcting tagging which was used to create our data set for supervised machine learning purposes.

Development of the tagger relied heavily on domain knowledge to identify logical tests for label application. For instance the occurrence of numbering at the beginning a line was used to identify clause headings, or the presence of a common text pattern which occur in definitions i.e. (the pattern *["word[1-3]] means [...]"* was used as a criterion for identifying a line as a definition).

	$\overline{x}$	s
On Set B Contracts	86.27%	NA
Contract x Contract	82.84%	10.85%

Table 1: Accuracy of Line Tagger.

The first row of Table 1 reports accuracy of the hand coded tagger when applied to Set B as the test set (the % accuracy of tagging of individual lines in Set B is reported). The second row of Table 1 reports average accuracy with which the lines in each of 20 contracts in Sets B and C are tagged (standard deviation is also reported). The high variance of the results at a contract by contract level suggests a need for future work to identify if issues such as contract 'type' may account for the differences (the lowest level of accuracy was around 50% and the highest around 95%).

Table 2 provides measures of precision, recall and F-measure attained by the hand coded tagger on Set B for key data items: clause headings, clauses themselves and definitions.[12] These results on key measures are sufficient for our initial purposes and will allow us to focus subsequent work on addressing ambiguity in the substantive rules found in a contract.

	Precision	Recall	F-Measure
CL.HEAD	0.98	0.83	0.89
CL.MATTER	0.91	0.91	0.91
DEF	1.00	0.66	0.80

Table 2: Performance results for Hand-Coded Line Tagger

5.3 Supervised Machine Learning

Column 1 of Table 3 reports the accuracy of various supervised machine learning algorithms with Set A as training set and Set B as test set. Of the algorithms applied, Random Forest (implemented using 100 trees and 30 random features) performed best. No ML method performed better than the hand coded tagger tested on the same test set (see row 1 of Table 1).

5.4 Machine Learning vs Hand-coded methods

Given our ultimate focus on developing applied tools, we are primarily concerned to identify the best methods to maximise the accuracy of our classification task. Whether this proves to be machine learning methods or hand-coded methods or a combination of the two is immaterial from this viewpoint (other than in regard of development costs).

Also it is useful to consider whether methods and insights from one method may assist in the other. For instance, hand-coded rules that successfully labelled lines directly suggested 'features' to be extracted for machine learning purposes. Conversely the identification of 'features' for machine learning can be regarded as a programming abstraction where the algorithm is 'under the hood' but human intervention is required in the form of feature selection for 'learning' to occur (implicitly encoding of rules in the feature set). In deriving our results we used a confusion matrix as a tool to assist in identifying areas where the hand-

[12] Precision is $\frac{TP}{TP+FP}$, Recall is $\frac{TP}{TP+FN}$, and F-Measure is $\frac{2PR}{P+R}$, where TP is true positives, FP is false positives, FN is false negatives, P is precision, and R is recall.

Exhibit 1027 Page 182

coded tagger required improvement and as a common baseline for performance. In certain forms of machine learning, the resulting machine learning model can be represented as code in "if, then, else" form (as in the output of a decision list or decision tree learner). We may thus conceive of machine learning and hand coding of rules as different 'programming paradigms' - a perspective that encourages us to apply insights across the boundary between them.

Further, we wished to explore the relative accuracy of machine learning vs hand coding for this particular task. Such an exploration is valuable from a cost benefit point of view. Supervised machine learning (where we have focused our attention) requires the production of human annotated data (which is costly to produce) as well as feature testing and refinement. Also machine learning may be computationally intensive. Production of an accurate hand-coded tagger requires expert knowledge and considerable testing and code refinement in order to achieve a high level of accuracy (which again is resource intensive). We did not undertake quantitative data collection in respect of time employed in the two methods, particularly as we followed an iterative development model that switched between rule based and machine learning focussed work and as we were searching for optimal outcomes. Nonetheless we anecdotally report that each method presented significant demands in terms of development time.

In the results reported above we note that no machine learning approach alone exceeded the performance of the hand coded tagger when considering performance on test Set B as a whole. This is the result also found at contract by contract level (see below Table 4).

5.5 Leveraging Hand-Coded Labels for Machine Learning

As noted above we also wished to explore whether utilising the output from the hand coded tagger as input features for machine learning might improve the accuracy of machine learning alone. A converse question that could be framed is whether the rule based tagger's performance could be improved by combination with machine learning.

We carried out two trials that bear on these questions:

1. machine learning alone and machine learning enhanced with rule based input on an entire test set (Set B); and
2. machine learning alone, rule alone and machine learning enhanced with rule based input at contract by contract level.

ML Algorithm	ML Alone	ML + Rule
Naive Bayes	82.38%	84.75%
SMO (SVM)	82.80%	85.40%
Cl. via Regression	83.91%	87.07%
Decision Tree	82.01%	87.02%
Bagging	83.54%	87.76%
Majority Vote	84.42%	87.11%
Random Forest	85.12%	86.69%

Table 3: Comparative Performance ML alone and ML + Rule on train and test set

Column 2 of Table 3 shows the improvement in performance of various machine learning methods when tags applied by the hand coded tagger were provided as input features. All methods show improvement with addition of the rule derived feature on Set B as a whole.[13]

We report below a comparison of accuracy of machine learning, rule based tagging and a combination of both in respect of Decision Trees. Decision trees offer the advantage that the output is easily interpretable as a rule set for tagging.

Method	$\overline{x}$	s
ML Alone	79.12%	10.71%
Rule Alone	82.84%	10.85%
ML + Rule	83.48%	9.83%

Table 4: Tagging Accuracy Mean and Standard Deviation - Contract by Contract

Although the results reported in Table 4 are different (as different contracts are included in the tested contracts)[14] average accuracy continued to show a differential between machine learning alone, and machine learning leveraged with hand rules, as described above. On this combination ML + Rule also outperformed the rule based tagger alone, but by a very narrow margin. In 17 of 20 cases ML + Rule improved or did not worsen performance of ML alone. In 15 of 20 cases, ML +

[13]The training set for the results reported in Table 3 was Set A. The test set was Set B.

[14]The training set was Set A. Sets B and C provided 20 individual contracts (effectively 20 test sets) for testing.

Exhibit 1027 Page 183

rule was no worse or outperformed the hand coded tagger alone. Notably variation is quite high as compared to the mean. Testing for statistical significance we can reject a null hypothesis that the mean accuracy for ML + Rule is the same as for ML alone. We cannot reject a null hypothesis that ML + Rule is the same as Rule alone.[15] We note also that this investigation was only carried out in respect of the decision tree learner.

5.5.1 Is it more effective to increase data or use rule based features?

To further explore the question of how much effort might be saved by combining rule based approaches with machine learning alone we tested the effect of adding random noise to the feature set provided by the hand coded tagger. We found that we had to randomly flip over 45% of these tags before they ceased to impact positively on classification accuracy. This suggests that even poorly prepared hand coded rules can improve performance.[16]

We also tested the effect of incrementally increasing the amount of available training data to explore any decrease in differential in performance. Testing using Set C as the test set and taking training data from Sets A and B, we progressively increased training data by increments of 500 instances of data. Differential performance decreases progressively, starting around 5% and reducing to 2.5% as more training data is made available.[17] (See Table 5.) Extrapolating from these figures we would expect at a minimum that training data would have to be doubled to remove differential accuracy. Relatively little effort in developing rule based features may substitute for considerable work in creating additional supervised data.

[15]We gratefully acknowledge the assistance of Bob Forrester of the ANU Statistical Consulting Unit in deriving this result, undertaken utilising a generalised linear model to undertake model fitting adjusting for different file (contract) types. Predictions of estimated mean proportions were: ML Alone 0.7760 (s.e. 0.007181), Rule alone 0.8106 (s.e. 0.006775), ML+ Rule 0.8181 (s.e. 0.006677).

[16]Using a Decision Tree, ML + Rule accuracy on test Set B with training Set A, was 87.0% with zero noise, 86.9% with 5% noise, 84.7% with 10% noise, 85.2% with 20% noise, 84.4% with 30% noise, 83.9% with 40% noise, 82.8% with 45% noise, 82.4% with 50% noise. Decision Tree accuracy on ML alone on this training and test set was 82.0%.

[17]The data has a negative correlation of -0.70

Data	0.5	1K	1.5	2K	2.5	3K	3.5	4K
Diff.	5.1	3.4	5	4.2	3.3	3.1	3.9	2.6

Table 5: Difference in % accuracy of ML alone and ML with rule based feature with increasing training data.

5.5.2 Decision tree output

An examination of the decision tree produced after adding the output of the hand-coded tagger to machine learning is of interest.

```
LTlabel = #CLAUSEMATTER#
| tknLngth <= 4
|| nounNum <= 1: #CLAUSEMATTER# (6.0/2.0)
|| nounNum > 1
||| linePos. <= 0.383929: #PRELIM# (2.0)
||| linePos. > 0.383929: #CLAUSEHEAD# (2.0)
| tknLngth > 4: #CLAUSEMATTER# (508.0/7.0)
: :
: :
LTlabel = #PH-LINE#
| prepNum <= 1: #PH-LINE# (5.0)
| prepNum > 1: #CLAUSEMATTER# (2.0)
LTlabel = #EMAILLINE#: #EMAILLINE# (3.0)
LTlabel = #ABNLINE#
| preLine2 <= 0: #TITLELINE# (2.0/1.0)
| preLine2 > 0: #ABNLINE# (3.0)
```

Figure 1: Part of the decision tree learnt when the labels from the hand-coded are provided as an attribute.

Using only the machine learning feature set, the size of the pruned tree is 199 with 108 leaves. After addition of labels provided by our hand-coded tagger the size of the decision tree is reduced to 105 with 69 leaves. Figure 1 shows part of this decision tree and as illustrated, the decision tree generally begins with the label assigned by the hand-coded tagger 'LTlabel' and amends the label (where necessary) utilizing other features available to the machine learner. Some labels (e.g. EMAILLINE) were simply adopted without change. In the case of the CLAUSEMATTER label, the machine learner used the number of tokens in the line ('tknLngth'), number of nouns in the line ('nounNum') and the relative position of the line in the contract ('linePos.') to re-assign the class of the line. The significance of the use of labels assigned by the hand-coded tagger is that the decision-tree identifies that attribute as providing the most significant information gain (that is reduction in entropy) and therefore partitions on that feature (Callan, 2003, p245). This remains true in this example irrespective of the class to which

Exhibit 1027 Page 184

the data is being assigned.[18] It is obvious that the best feature for assigning the correct label for an item of data is the correct or probably correct label (if one can obtain it without undue endeavour). From a practical viewpoint some effort directing to hand-crafting code to extract appropriate class labels may thus assist machine learning.

6 Conclusions and Future Work

In terms of the overall performance of the methods described in this paper the closest point of comparison is the work reported by Indukuri and Krishna (Indukuri and Krishna, 2010). While comparisons using different methods, software and data are of dubious validity, we may note their broad similarity. With a training set of 10 contracts and a test set of 20 contracts, the contract-by-contract average F-measure for the clausematter class was 0.8632. Also in 14 contracts of the 20 tested the F-measure for the clausematter class was above 0.85. With a training set of 10 contracts (Set A) and a test set of 10 contracts (Set B) (as a single test set) the F-measure for the clausematter class was 0.921.[19] Using 4-grams as features, Indukuri and Krishna report an accuracy of 83.48% on the task of classification clauses from non-clauses, applied to 73 sentences to be so classified. Whereas the task in that case was a binary classification task over one contract, we report results on multi-class classification over test sets of up to 20 contracts.[20]

The work we have undertaken and related work reviewed in this paper (Section 3) suggests the potentially broad utility of combining rule based and machine learning methods. One potential hybrid classification method in its simplest formulation may look something like this:

1. determine the required or anticipated level of accuracy for the intended application;

2. develop a simple code rule set for the classification task (i.e. with minimal expenditure of resources);

3. if classification accuracy is sufficient machine learning would not be required, if not proceed as usual in development of machine learning using the output of hand coded rules as an input to machine learning.

Such a development model does not avoid the development of a supervised training set, as such a set is required both to assess the accuracy of the rule based tagger as well as the machine learner (should development proceed to that stage). However it may reduce the amount of data required to attain a desired level of accuracy. The method could of course be applied iteratively, if the accuracy level is insufficient after a first iteration.

Future work to be explored includes the effective "typing" of contracts (particularly in light of the high variance of performance on individual contracts). For instance, some contracts use clause headings, others do not. Some use schedules, others do not, etc. Such type information may further assist in the classification task (the accuracy of which varies considerably depending on the contract to which it is applied). In typical preprocessing, documents in corpora are converted to plain text. This results in the loss of layout and hierarchical information found in word documents. The preservation of such information for use in classification tasks may improve accuracy.

The preliminary work here is also a precursor to work on identification of ambiguity and the development of practical tools to assist in contract drafting. In separate work, we intend also to describe and analyse the contract corpus that is referred to here as the basis of this work.

[18] The Weka decision tree implementation, "J.48", is based on the Quinlan's C4.5(Quinlan, 1986).

[19] The results reported here are derived using a random forest algorithm and using the hand coded tag as an input feature for classification.

[20] Coincidentally the 83.48% figure is the same as our average result over all classes for 20 contracts.

References

J. K. Aitken and P. Butt. 2004. *Piesse The Elements of Drafting*. Lawbook Co, 10th edition.

L. Bacci, P. Spinosa, C. Marchetti, R. Battistoni, I. Florence, I. Senate, and I. Rome. 2009. Automatic mark-up of legislative documents and its application to parallel text generation. In *Proc. of LOAIT Workshop*, pages 45–54.

D. Biber, S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Rob Callan. 2003. *Artificial Intelligence*. Palgrave Macmillan.

(ed) Christopher S. Butler. 1992. *Computers and written texts*. Blackwell.

E. Francesconi. 2006. The Norme in Rete Project: Standards and Tools for Italian Legislation. *International Journal Legal Information*, 34:358.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software. *SIGKDD Explorations*, 11(1).

I. Hasan, J. Parapar, and R. Blanco. 2008. Segmentation of legislative documents using a domain-specific lexicon. In *Proceedings of the 19th International Conference on Database and Expert Systems Application*, pages 665–669.

Kishore Varma Indukuri and P. Radha Krishna. 2010. Mining e-contract documents to classify clauses. In *COMPUTE '10: Proceedings of the Third Annual ACM Bangalore Conference*, pages 1–5, New York, NY, USA. ACM.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2nd edition.

D. Khoury and Y. S. Yamouni. 2007. *Understanding Contract Law*. Lexis Nexis Butterworths, 7th edition.

M. Kipp. 2006. Creativity meets automation: Combining nonverbal action authoring with rules and machine learning. In *Intelligent Virtual Agents*, pages 230–242. Springer.

Eric McCreath and Judy Kay. 2003. Iems: Helping Users Manage Email. *User Modeling 2003*, pages 146–146.

Eneldo Loza Mencía. 2009. Segmentation of legal documents. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 88–97. ACM.

Seong-Bae Park and Byoung-Tak Zhang. 2003. Text chunking by combining hand-crafted rules and memory-based learning. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 497–504.

J.R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–106.

M. Rochery, R. Schapire, M. Rahim, N. Gupta, G. Riccardi, S. Bangalore, H. Alshawi, and S. Douglas. 2002. Combining prior knowledge and boosting for call classification in spoken language dialogue. In *Proceedings of the IEEE International Conference of Acoustics and Speech and Signal Processing*.

S. Scott and S. Matwin. 1999. Feature Engineering for Text Classification. In *Machine learning: proceedings of the Sixteenth International Conference (ICML'99), Bled, Slovenia, June 27-30, 1999*, page 379. Morgan Kaufmann Pub.

K. Takahashi, H. Takamura, and M. Okumura. 2005. Automatic occupation coding with combination of machine learning and hand-crafted rules. *Advances in Knowledge Discovery and Data Mining*, pages 269–279.

https://web.archive.org/web/20110708111015/http://aclweb.org/anthology/U/U10/U10-1005.pdf

Corpus Based Classification of Text in Australian Contracts

Michael Curtotti
Australian National University
Canberra, Australia
michael.curtotti@anu.edu.au

Eric Mccreath
Australian National University
Canberra, Australia
eric.mccreath@anu.edu.au

Abstract

Written contracts are a fundamental framework for commercial and cooperative transactions and relationships. Limited research has been published on the application of machine learning and natural language processing (NLP) to contracts. In this paper we report the classification of components of contract texts using machine learning and hand-coded methods. Authors studying a range of domains have found that combining machine learning and rule based approaches increases accuracy of machine learning. We find similar results which suggest the utility of considering leveraging hand coded classification rules for machine learning. We attained an average accuracy of 83.48% on a multi-class labelling task on 20 contracts combining machine learning and rule based approaches, increasing performance over machine learning alone.

1 Introduction

Contracts govern economic transactions and commercial or organisational relationships from trivial purchases to major national infrastructure projects. Any large organisation (whether private or public) must unavoidably invest significant resources in developing and concluding contracts, as the contracts it enters define its legal relations with organisations and individuals with which it interacts. As one author has noted, contracts are an integral part of any business enterprise and it is "difficult to overstate their importance to the business world" (Khoury and Yamouni, 2007, p16). In addition, drafting contracts represents a major economic activity for the legal industry.

Ambiguity is an unresolved problem in natural language processing. It is also an important problem affecting the drafting and operation of contracts. The costs of ambiguity in a contract can be enormous. The presence of ambiguity increases the risk of litigation, project failure and loss of commercial relationships. Instances of litigated ambiguity are used in drafting manuals as instructional tales for the unwary drafter (Aitken and Butt, 2004, p37).

Our future research goal is to investigate approaches that will help identify such ambiguity, however, a precursor for addressing the problem of ambiguity is accurate labelling or classification of the constituent parts of contract documents.[1] In this paper we report work on classifying 'lines' in contracts into 32 classes. We compare three different classification approaches in order to determine which approach may provide the best performance with respect to this task: supervised machine learning; a hand-crafted rule based tagger; or combining hand-crafted rules based tagging with machine learning.

2 Background

Natural language processing involves applying a pipeline of transformations to text. A typical first step is to segment the text into its sentences, after which further processing is applied (Jurafsky and Martin, 2009, p 69). This works well in usual English prose, but is not necessarily well suited to contracts because of the heavy use of layout to embody structure and semantics. In a contract, a sentence may occur within a single paragraph of text, or may be spread over several line breaks due to the use of sub-paragraphing as an aid to comprehension. Such sub-paragraphs may embody disjunctive or conjunctive conditions associated with

[1] This task can also be thought of as a document segmentation task, although we here consider it as a classification exercise.

Michael Curtotti and Eric McCreath. 2010. Michael Curtotti and Eric McCreath. In *Proceedings of Australasian Language Technology Association Workshop*, pages 18−26
Exhibit 1027 Page 188

a rule. A 'clause'[2] is the typical structural unit which embodies a legal rule or set of rules. Sometimes clauses are organised into sub-clauses and clauses and sub-clauses may be numbered in hierarchical fashion (e.g. 1, 1.1, (a), (i), (A))(Aitken and Butt, 2004, pp 23, 27). Line breaks often occur immediately before data such as headings, clauses, sub-clauses, party names, dates or execution blocks. Being able to label a line as to its character, key content or function thus enriches available information for later NLP. The line is thus a valuable starting point for applying NLP.

3 Related Work

3.1 Classification and Segmentation of Contracts and Legislative Documents

There would appear to be little published work on the application of natural language processing (NLP) or machine learning specifically for the classification of text in legal contracts. An exception to this is (Indukuri and Krishna, 2010) who apply a support vector machine over feature vectors extracted from contract clauses. These feature vectors are comprised of n-grams up to size n = 4. They classify contract sentences as clauses or non-clauses and then sub-classify clauses as concerned with payment terms or not. Indukuri and Krishna note the lack of published work on classification of sentences in contract documents for workflow and monitoring purposes.

While also limited, there has been some work on segmentation or classification of legislative corpora using machine learning. (Mencía, 2009) reports 100% precision and recall in machine learning on classification of legislation into articles, sections and parts (although to a task for which he is able to craft an equally accurate regular expression based segmenter). (Bacci et al., 2009) automatically classify plain text legislative documents for the purpose of XML mark up. (Francesconi, 2006), on whose work Bacci et al. build, describes four separate modules for dealing with plain text legacy content which they are seeking to convert to XML tagged text. (Hasan et al., 2008) carry out segmentation of Spanish legislative bulletins using the table of contents as an aid to segmentation. This work, in the legislative field, provides a parallel application domain for the work undertaken here in respect of contracts, although it is generally true that legislation will tend to conform to far tighter stylistic rules, because of the central control of legislative drafting.

3.2 Combining Machine Learning and Hand Coded Approaches

A number of researchers have combined hand-coded approaches and machine learning to produce better overall results.

(Kipp, 2006) augments handcrafted rules with machine learnt rules for gesture generation.

(McCreath and Kay, 2003) used such a combined approach to improve performance on categorizing emails.

(Park and Zhang, 2003) address the problem of text chunking for the Korean language, combining hand-crafted rules with a memory based machine learning method. A rule based method is used to determine whether an instance may constitute an exceptional case, with machine learning used to correct classifications assigned by the rule based method. They report up to a 2.83% improvement in F-scores over memory-based learning alone.

(Rochery et al., 2002) report the use of hand-crafted rules to enhance the accuracy of machine learning utilising boosting. They apply the method to a multi-labelling classification problem where they are seeking to classify spoken utterances. They attribute increased accuracy to the fact that the rules they develop are not represented in their data at all or do not appear sufficiently to have statistical impact during the training process. They note that the combination has the greatest effect when there are less then 100 examples, with the benefit of the combination decreasing as the number of training examples increases. The method therefore has direct applicability to reducing the data requirement for machine learning (a costly aspect of supervised machine learning).

(Takahashi et al., 2005) utilise a combination of SVM based learning and hand-crafted rules for the task of classifying occupations from social surveys. They effect a combination by using rule derived labels as features for the machine learner. Again the work occurs in a multi-class labelling context (with around 200 occupation codes). They find that SVM approaches are superior to rule-based approaches alone but that combining rule-based and SVM learning produces the best results. They also examine the effect of the number of training examples, noting that the differential ben-

[2]The term 'clause' here refers to a legal rule or rules appearing in a contract, rather than the linguistic entity.

Exhibit 1027 Page 189

efits of combining methods reduces with increasing data.

In our study we use class labels assigned by a hand-coded tagger as additional features for machine learning to obtain an increase in accuracy in the domain of text classification.

The work reported above, and our own paper, come from diverse domains and illustrate the application of combined rule/ML approaches using a variety of machine learning algorithms.

This suggests the wide utility of using hand-coded rules to improve accuracy of machine learning. Symetrically we may say that machine learning may potentially be used to improve the accuracy of hand-coded rules. Here we essentially follow a hybrid model of development that examines each method separately as well as in combination.

4 Data

4.1 Australian Contract Corpus

The data which forms the basis of this study is a corpus of 256 contracts ('the Australian Contract Corpus') which has been compiled from the web using the Google Australia web search: `http://www.google.com.au`.[3] The corpus is constituted of 1043364 tokens (words and punctuation), 42910 sentences, and a vocabulary of 14217. The lexical diversity of this corpus is 73.39.[4] As far as we are aware there is no published work based on a corpus of Australian contract language, however space does not permit a fuller discussion of the design and analysis of the corpus.

4.2 Data Representation

For the purpose of this study we randomly selected 30 contracts from the Australian Contract Corpus to produce data sets for application of machine learning. The 30 contracts were divided into three sets: a training set (Set A)[5], a second set (Set B)[6] (primarily used as a test set, although sometimes combined with Set A to form a larger training set) and an additional test set (Set C)[7], each of 10 contracts. Set C was added to increase both available data and the number of contracts on which testing could be carried out.

To simplify processing all contracts were converted into text files as a preprocessing step.[8] Also we removed material that typically does not appear in contracts (primarily guideline notes for drafters).

For purposes of our study each 'line' or 'paragraph' in a contract constituted a data point. Sets A, B and C provided 1825, 2157 and 2231 lines/data points, respectively.

To classify our data for machine learning we first employed our hand-coded tagger to produce a labelled data set constituted of a line and primary and secondary labels according to our classification system (see below). The classifications applied by the hand-coded tagger were then manually corrected to remove any labelling errors. Both for the reason that hand tagging data from scratch is more error prone than correcting previously machine tagged data, and because of the high labour in manual tagging, machine assisted tagging is a standard method used in corpus creation (Biber et al., 1998, p 262), ((ed) Christopher S. Butler, 1992, p 131).

4.3 Feature Selection

Features were then extracted from our labelled lines using a hand-coded feature extractor to extract up to 40 features to be used for machine learning, with the primary class label earlier applied to each line serving as the target for machine learning.[9] Features for machine learning were selected based on assumed relevance to the intended classification.[10] While a bag of word representa-

[3]The search phrase we employed was: "clause party agreement" , limiting the search to "pages from Australia" and the filetype to ".doc". The collection of the corpus was undertaken in the period 6 - 24 December 2009. Each document was visually inspected by one of the authors to verify that it was an Australian contract and documents were added to the corpus in order of their appearance in Google search results. We have made available the list of URLs of the documents that make up the corpus at: `http://cs.anu.edu.au/people/Michael.Curtotti/`. We also make code and data referred to in this paper available at the same location.

[4]Lexical diversity refers to the ratio of the total number of tokens in the corpus to the total number of types of tokens i.e. tokens/vocabulary. The statistics and information provided here were extracted using the NLTK. (Bird et al., 2009).

[5]Contracts numbered in our corpus: 55, 74, 77, 91, 94, 144, 174, 185, 208 and 213

[6]Contracts 11, 12, 164, 193, 196, 199, 249, 59, 64, 9

[7]Contracts 200, 254, 175, 180, 120, 102, 207, 127, 75 and 79

[8]Notably, this step does have the downside of discarding valuable style and layout information found in the word document format.

[9]In some instances the assignment of class is not entirely disjunctive and implies a priority ordering of potential classifications. For instance a clause may contain address details, and be classified as such rather than as a clause. The feature extractor removed secondary labels.

[10]The full list of features is as follows: 1. relative position of a line within the contract; 2. line end punctuation; 3. num-

tion with automatic feature extraction was trialed, this was found to result in lower accuracy than a manually selected (but small) number of features. Traditionally, 'feature engineering' or 'selection' in text classification has focussed on 'bag of words' or proximate features such as n-grams or noun phrases (Scott and Matwin, 1999). Our selection of features might be regarded as 'hybrid' in that they implicitly encode domain knowledge and 'rules'. For example, we expect a definition line to contain the widely used word 'means' as in 'x means y'. We would expect clause headings to display a different pattern of parts of speech occurrence, as compared to clauses themselves. We also consider features such as the relationship between the particular line/datapoint and its neighbouring lines (e.g. is the preceding line long or short), or the document as a whole (e.g. relative placement of the line within the document).

Features used in the related work carried out by others on contracts and legislation include: n-grams (Indukuri), bag of word together with capitalisation and character patterns (Mencia), the use of sequential word and token data as a HMM to predict state where state is the classification of the portion of text (Bacci/Francesconi) and the use of word terms identifying titles and lists of indexed terms (Hasan).

4.4 Classes

We adopted 32 classes for our classification scheme. These classes represented significant structural elements (such as clause headings or content lines) or sometimes key contract meta data (for instance the parties, the date on which an agreement is made, email addresses, ABNs etc).[11] Of particular interest to us, was an ability to accurately identify clausematter (the primary location of legal rules of a contract) and clause headings (which in some contracts effectively mark boundaries between major rule sets).

5 Experimental Evaluation

Analysis of our data sets was carried out using two major methods: analysis using the training Set A measuring peformance on test Set B (i.e. testing for accuracy of classification of lines in Set B); and analysis using training Set A where each contract in test Sets B and C was used individually as a test set (testing for accuracy of classification of lines of each individual contract). The former testing approach is more widely used but the latter is more consistent with our intended end application: that is, developing an ability to accurately classify lines in a previously unseen contract as an individual document.

Set A consisted of 1825 lines of data, Set B consisted of 2157 lines of data and Set C consisted of 2231 lines of data.

5.1 Tools

The following tools were used to carry out machine learning tasks reported in this paper:

1. the *python* programming language was used to develop a hand coded line tagger, a feature extractor and code for evaluation of the performance of the hand coded tagger;

2. the Natural Language Toolkit (NLTK)(Bird et al., 2009) was used in corpus development and in the hand coded tagger for application of parts of speech tags; and

3. the Weka data mining software(Hall et al., 2009) was used to carry out machine learning.

bering at the beginning of the line; 4. line length in characters; 5. line length in tokens; 6. number of nouns; 7. number of adjectives; 8. number of verbs; 9. number of prepositions; 10. number of coordinating conjunctions and cardinal numbers; 11. number of modal verbs; 12. number of personal pronouns, possessives; 13. number of adjectives; 14. whether must occurs; 15. whether may occurs; 16. whether shall occurs; 17. whether means occurs;18. the position of the word means in the line; 19. whether the word 'include' occurs; 20. the position of the word 'includes' in the line; 21. whether the phase 'has the same meaning' occurs; 22. whether the line begins with a capital letter followed by a stop or a space; 23. do the letters 'abn' or 'acn' appear; 24. does the word address appear near the beginning of the line; 25. does the word 'contact' appear near the beginning of the line; 26. does the word 'email' appear near the beginning of the line; 27. does the word 'fax' appear near the beginning of the line; 28. does the word 'note' appear at the beginning of the line; 29. does the word 'phone' appear at the beginning of the line; 30. do the terms 'web' or 'www' appear near the beginning of the line; 31-36. token lengths of 3 lines before and after the data line; 37. the tag applied by the hand coded tagger to the data line; 38. whether the contract from which the line comes has clause headings; 39. whether the contract from which the line comes has clause sub-headings; 40. whether the contract has a schedule.

[11] The following class labels were applied: CLAUSEMATTER, CLAUSEHEAD, DEF, CONTENTSHEAD, BLANK, DROSS, AND, BETWEEN, DATEMADELINE, PARTIESHEAD, PARTYLINE, RECITALHEAD, RECITALLINE, HEAD, CONTENTLINE, PRELIM, OPERATIVEPROVISIONLINE, NUMBEREDLINE, SCHEDULEHEAD, SCHEDULEITEM, EXECUTIONBLOCK, NOTE, ABNLINE, ADDRESSLINE, EMAILLINE, WEBLINE, FAXLINE, PH-LINE, REF-LINE, SCHOTHERMATTER, CONTACTOFFICER, TITLELINE.

Exhibit 1027 Page 191

5.2 Hand-Coded Tagger

Our hand-coded tagger follows a pipeline methodology. A contract is separated into a list of lines. In the first phase, each line is passed through a series of "if-then-else" routines (essentially utilising regular expressions) to assign primary and secondary labels to the line. This produces a contract with enhanced information content, in the form of a tuple of labels and line text. This output is then fed through a series of methods, progressively bootstrapping improvements to labelling, based on the information available when the method is applied. For instance content lines will be identified, these are then used in later methods to seek to identify clause headings. In addition, the tagger extracts contract meta data such as: a list of definitions, a list of clause headings, where the beginning of operative provisions occurs, where the schedule begins. Such features may or may not occur in a contract. In addition the tagger also provided a method for hand correcting tagging which was used to create our data set for supervised machine learning purposes.

Development of the tagger relied heavily on domain knowledge to identify logical tests for label application. For instance the occurrence of numbering at the beginning a line was used to identify clause headings, or the presence of a common text pattern which occur in definitions i.e. (the pattern *["word[1-3]] means [...]"* was used as a criterion for identifying a line as a definition).

	$\overline{x}$	s
On Set B Contracts	86.27%	NA
Contract x Contract	82.84%	10.85%

Table 1: Accuracy of Line Tagger.

The first row of Table 1 reports accuracy of the hand coded tagger when applied to Set B as the test set (the % accuracy of tagging of individual lines in Set B is reported). The second row of Table 1 reports average accuracy with which the lines in each of 20 contracts in Sets B and C are tagged (standard deviation is also reported). The high variance of the results at a contract by contract level suggests a need for future work to identify if issues such as contract 'type' may account for the differences (the lowest level of accuracy was around 50% and the highest around 95%).

Table 2 provides measures of precision, recall and F-measure attained by the hand coded tagger on Set B for key data items: clause headings, clauses themselves and definitions.[12] These results on key measures are sufficient for our initial purposes and will allow us to focus subsequent work on addressing ambiguity in the substantive rules found in a contract.

	Precision	Recall	F-Measure
CL.HEAD	0.98	0.83	0.89
CL.MATTER	0.91	0.91	0.91
DEF	1.00	0.66	0.80

Table 2: Performance results for Hand-Coded Line Tagger

5.3 Supervised Machine Learning

Column 1 of Table 3 reports the accuracy of various supervised machine learning algorithms with Set A as training set and Set B as test set. Of the algorithms applied, Random Forest (implemented using 100 trees and 30 random features) performed best. No ML method performed better than the hand coded tagger tested on the same test set (see row 1 of Table 1).

5.4 Machine Learning vs Hand-coded methods

Given our ultimate focus on developing applied tools, we are primarily concerned to identify the best methods to maximise the accuracy of our classification task. Whether this proves to be machine learning methods or hand-coded methods or a combination of the two is immaterial from this viewpoint (other than in regard of development costs).

Also it is useful to consider whether methods and insights from one method may assist in the other. For instance, hand-coded rules that successfully labelled lines directly suggested 'features' to be extracted for machine learning purposes. Conversely the identification of 'features' for machine learning can be regarded as a programming abstraction where the algorithm is 'under the hood' but human intervention is required in the form of feature selection for 'learning' to occur (implicitly encoding of rules in the feature set). In deriving our results we used a confusion matrix as a tool to assist in identifying areas where the hand-

[12] Precision is $\frac{TP}{TP+FP}$, Recall is $\frac{TP}{TP+FN}$, and F-Measure is $\frac{2PR}{P+R}$, where TP is true positives, FP is false positives, FN is false negatives, P is precision, and R is recall.

Exhibit 1027 Page 192

coded tagger required improvement and as a common baseline for performance. In certain forms of machine learning, the resulting machine learning model can be represented as code in "if, then, else" form (as in the output of a decision list or decision tree learner). We may thus conceive of machine learning and hand coding of rules as different 'programming paradigms' - a perspective that encourages us to apply insights across the boundary between them.

Further, we wished to explore the relative accuracy of machine learning vs hand coding for this particular task. Such an exploration is valuable from a cost benefit point of view. Supervised machine learning (where we have focused our attention) requires the production of human annotated data (which is costly to produce) as well as feature testing and refinement. Also machine learning may be computationally intensive. Production of an accurate hand-coded tagger requires expert knowledge and considerable testing and code refinement in order to achieve a high level of accuracy (which again is resource intensive). We did not undertake quantitative data collection in respect of time employed in the two methods, particularly as we followed an iterative development model that switched between rule based and machine learning focussed work and as we were searching for optimal outcomes. Nonetheless we anecdotally report that each method presented significant demands in terms of development time.

In the results reported above we note that no machine learning approach alone exceeded the performance of the hand coded tagger when considering performance on test Set B as a whole. This is the result also found at contract by contract level (see below Table 4).

5.5 Leveraging Hand-Coded Labels for Machine Learning

As noted above we also wished to explore whether utilising the output from the hand coded tagger as input features for machine learning might improve the accuracy of machine learning alone. A converse question that could be framed is whether the rule based tagger's performance could be improved by combination with machine learning.

We carried out two trials that bear on these questions:

1. machine learning alone and machine learning enhanced with rule based input on an entire test set (Set B); and
2. machine learning alone, rule alone and machine learning enhanced with rule based input at contract by contract level.

ML Algorithm	ML Alone	ML + Rule
Naive Bayes	82.38%	84.75%
SMO (SVM)	82.80%	85.40%
Cl. via Regression	83.91%	87.07%
Decision Tree	82.01%	87.02%
Bagging	83.54%	87.76%
Majority Vote	84.42%	87.11%
Random Forest	85.12%	86.69%

Table 3: Comparative Performance ML alone and ML + Rule on train and test set

Column 2 of Table 3 shows the improvement in performance of various machine learning methods when tags applied by the hand coded tagger were provided as input features. All methods show improvement with addition of the rule derived feature on Set B as a whole.[13]

We report below a comparison of accuracy of machine learning, rule based tagging and a combination of both in respect of Decision Trees. Decision trees offer the advantage that the output is easily interpretable as a rule set for tagging.

Method	$\overline{x}$	s
ML Alone	79.12%	10.71%
Rule Alone	82.84%	10.85%
ML + Rule	83.48%	9.83%

Table 4: Tagging Accuracy Mean and Standard Deviation - Contract by Contract

Although the results reported in Table 4 are different (as different contracts are included in the tested contracts)[14] average accuracy continued to show a differential between machine learning alone, and machine learning leveraged with hand rules, as described above. On this combination ML + Rule also outperformed the rule based tagger alone, but by a very narrow margin. In 17 of 20 cases ML + Rule improved or did not worsen performance of ML alone. In 15 of 20 cases, ML +

[13]The training set for the results reported in Table 3 was Set A. The test set was Set B.

[14]The training set was Set A. Sets B and C provided 20 individual contracts (effectively 20 test sets) for testing.

Exhibit 1027 Page 193

rule was no worse or outperformed the hand coded tagger alone. Notably variation is quite high as compared to the mean. Testing for statistical significance we can reject a null hypothesis that the mean accuracy for ML + Rule is the same as for ML alone. We cannot reject a null hypothesis that ML + Rule is the same as Rule alone.[15] We note also that this investigation was only carried out in respect of the decision tree learner.

5.5.1 Is it more effective to increase data or use rule based features?

To further explore the question of how much effort might be saved by combining rule based approaches with machine learning alone we tested the effect of adding random noise to the feature set provided by the hand coded tagger. We found that we had to randomly flip over 45% of these tags before they ceased to impact positively on classification accuracy. This suggests that even poorly prepared hand coded rules can improve performance.[16]

We also tested the effect of incrementally increasing the amount of available training data to explore any decrease in differential in performance. Testing using Set C as the test set and taking training data from Sets A and B, we progressively increased training data by increments of 500 instances of data. Differential performance decreases progressively, starting around 5% and reducing to 2.5% as more training data is made available.[17] (See Table 5.) Extrapolating from these figures we would expect at a minimum that training data would have to be doubled to remove differential accuracy. Relatively little effort in developing rule based features may substitute for considerable work in creating additional supervised data.

[15]We gratefully acknowledge the assistance of Bob Forrester of the ANU Statistical Consulting Unit in deriving this result, undertaken utilising a generalised linear model to undertake model fitting adjusting for different file (contract) types. Predictions of estimated mean proportions were: ML Alone 0.7760 (s.e. 0.007181), Rule alone 0.8106 (s.e. 0.006775), ML+ Rule 0.8181 (s.e. 0.006677).

[16]Using a Decision Tree, ML + Rule accuracy on test Set B with training Set A, was 87.0% with zero noise, 86.9% with 5% noise, 84.7% with 10% noise, 85.2% with 20% noise, 84.4% with 30% noise, 83.9% with 40% noise, 82.8% with 45% noise, 82.4% with 50% noise. Decision Tree accuracy on ML alone on this training and test set was 82.0%.

[17]The data has a negative correlation of -0.70

Data	0.5	1K	1.5	2K	2.5	3K	3.5	4K
Diff.	5.1	3.4	5	4.2	3.3	3.1	3.9	2.6

Table 5: Difference in % accuracy of ML alone and ML with rule based feature with increasing training data.

5.5.2 Decision tree output

An examination of the decision tree produced after adding the output of the hand-coded tagger to machine learning is of interest.

```
LTlabel = #CLAUSEMATTER#
| tknLngth <= 4
|| nounNum <= 1: #CLAUSEMATTER# (6.0/2.0)
|| nounNum > 1
||| linePos. <= 0.383929: #PRELIM# (2.0)
||| linePos. > 0.383929: #CLAUSEHEAD# (2.0)
| tknLngth > 4: #CLAUSEMATTER# (508.0/7.0)
: :
: :
LTlabel = #PH-LINE#
| prepNum <= 1: #PH-LINE# (5.0)
| prepNum > 1: #CLAUSEMATTER# (2.0)
LTlabel = #EMAILLINE#: #EMAILLINE# (3.0)
LTlabel = #ABNLINE#
| preLine2 <= 0: #TITLELINE# (2.0/1.0)
| preLine2 > 0: #ABNLINE# (3.0)
```

Figure 1: Part of the decision tree learnt when the labels from the hand-coded are provided as an attribute.

Using only the machine learning feature set, the size of the pruned tree is 199 with 108 leaves. After addition of labels provided by our hand-coded tagger the size of the decision tree is reduced to 105 with 69 leaves. Figure 1 shows part of this decision tree and as illustrated, the decision tree generally begins with the label assigned by the hand-coded tagger 'LTlabel' and amends the label (where necessary) utilizing other features available to the machine learner. Some labels (e.g. EMAILLINE) were simply adopted without change. In the case of the CLAUSEMATTER label, the machine learner used the number of tokens in the line ('tknLngth'), number of nouns in the line ('nounNum') and the relative position of the line in the contract ('linePos.') to re-assign the class of the line. The significance of the use of labels assigned by the hand-coded tagger is that the decision-tree identifies that attribute as providing the most significant information gain (that is reduction in entropy) and therefore partitions on that feature (Callan, 2003, p245). This remains true in this example irrespective of the class to which

Exhibit 1027 Page 194

the data is being assigned.[18] It is obvious that the best feature for assigning the correct label for an item of data is the correct or probably correct label (if one can obtain it without undue endeavour). From a practical viewpoint some effort directing to hand-crafting code to extract appropriate class labels may thus assist machine learning.

6 Conclusions and Future Work

In terms of the overall performance of the methods described in this paper the closest point of comparison is the work reported by Indukuri and Krishna (Indukuri and Krishna, 2010). While comparisons using different methods, software and data are of dubious validity, we may note their broad similarity. With a training set of 10 contracts and a test set of 20 contracts, the contract-by-contract average F-measure for the clausematter class was 0.8632. Also in 14 contracts of the 20 tested the F-measure for the clausematter class was above 0.85. With a training set of 10 contracts (Set A) and a test set of 10 contracts (Set B) (as a single test set) the F-measure for the clausematter class was 0.921.[19] Using 4-grams as features, Indukuri and Krishna report an accuracy of 83.48% on the task of classification clauses from non-clauses, applied to 73 sentences to be so classified. Whereas the task in that case was a binary classification task over one contract, we report results on multi-class classification over test sets of up to 20 contracts.[20]

The work we have undertaken and related work reviewed in this paper (Section 3) suggests the potentially broad utility of combining rule based and machine learning methods. One potential hybrid classification method in its simplest formulation may look something like this:

1. determine the required or anticipated level of accuracy for the intended application;

2. develop a simple code rule set for the classification task (i.e. with minimal expenditure of resources);

3. if classification accuracy is sufficient machine learning would not be required, if not proceed as usual in development of machine learning using the output of hand coded rules as an input to machine learning.

Such a development model does not avoid the development of a supervised training set, as such a set is required both to assess the accuracy of the rule based tagger as well as the machine learner (should development proceed to that stage). However it may reduce the amount of data required to attain a desired level of accuracy. The method could of course be applied iteratively, if the accuracy level is insufficient after a first iteration.

Future work to be explored includes the effective "typing" of contracts (particularly in light of the high variance of performance on individual contracts). For instance, some contracts use clause headings, others do not. Some use schedules, others do not, etc. Such type information may further assist in the classification task (the accuracy of which varies considerably depending on the contract to which it is applied). In typical preprocessing, documents in corpora are converted to plain text. This results in the loss of layout and hierarchical information found in word documents. The preservation of such information for use in classification tasks may improve accuracy.

The preliminary work here is also a precursor to work on identification of ambiguity and the development of practical tools to assist in contract drafting. In separate work, we intend also to describe and analyse the contract corpus that is referred to here as the basis of this work.

[18] The Weka decision tree implementation, "J.48", is based on the Quinlan's C4.5(Quinlan, 1986).

[19] The results reported here are derived using a random forest algorithm and using the hand coded tag as an input feature for classification.

[20] Coincidentally the 83.48% figure is the same as our average result over all classes for 20 contracts.

References

J. K. Aitken and P. Butt. 2004. *Piesse The Elements of Drafting*. Lawbook Co, 10th edition.

L. Bacci, P. Spinosa, C. Marchetti, R. Battistoni, I. Florence, I. Senate, and I. Rome. 2009. Automatic mark-up of legislative documents and its application to parallel text generation. In *Proc. of LOAIT Workshop*, pages 45–54.

D. Biber, S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Rob Callan. 2003. *Artificial Intelligence*. Palgrave Macmillan.

(ed) Christopher S. Butler. 1992. *Computers and written texts*. Blackwell.

E. Francesconi. 2006. The Norme in Rete Project: Standards and Tools for Italian Legislation. *International Journal Legal Information*, 34:358.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software. *SIGKDD Explorations*, 11(1).

I. Hasan, J. Parapar, and R. Blanco. 2008. Segmentation of legislative documents using a domain-specific lexicon. In *Proceedings of the 19th International Conference on Database and Expert Systems Application*, pages 665–669.

Kishore Varma Indukuri and P. Radha Krishna. 2010. Mining e-contract documents to classify clauses. In *COMPUTE '10: Proceedings of the Third Annual ACM Bangalore Conference*, pages 1–5, New York, NY, USA. ACM.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2nd edition.

D. Khoury and Y. S. Yamouni. 2007. *Understanding Contract Law*. Lexis Nexis Butterworths, 7th edition.

M. Kipp. 2006. Creativity meets automation: Combining nonverbal action authoring with rules and machine learning. In *Intelligent Virtual Agents*, pages 230–242. Springer.

Eric McCreath and Judy Kay. 2003. Iems: Helping Users Manage Email. *User Modeling 2003*, pages 146–146.

Eneldo Loza Mencía. 2009. Segmentation of legal documents. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 88–97. ACM.

Seong-Bae Park and Byoung-Tak Zhang. 2003. Text chunking by combining hand-crafted rules and memory-based learning. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 497–504.

J.R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–106.

M. Rochery, R. Schapire, M. Rahim, N. Gupta, G. Riccardi, S. Bangalore, H. Alshawi, and S. Douglas. 2002. Combining prior knowledge and boosting for call classification in spoken language dialogue. In *Proceedings of the IEEE International Conference of Acoustics and Speech and Signal Processing*.

S. Scott and S. Matwin. 1999. Feature Engineering for Text Classification. In *Machine learning: proceedings of the Sixteenth International Conference (ICML'99), Bled, Slovenia, June 27-30, 1999*, page 379. Morgan Kaufmann Pub.

K. Takahashi, H. Takamura, and M. Okumura. 2005. Automatic occupation coding with combination of machine learning and hand-crafted rules. *Advances in Knowledge Discovery and Data Mining*, pages 269–279.

https://web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu
The University of Arizona, Tucson, Arizona, USA
{alicekwak, jisraelsen, claytonm, derekbambauer, msurdeanu}@arizona.edu

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

1 Introduction

Natural language inference (NLI) in the legal domain has not been widely investigated, despite its importance and potential. One such important NLI application is validity assessment of legal documents such as wills. Legal procedures for creating and executing wills are evolving rapidly. Processing a will via probate is a costly, time-consuming process that can be exacerbated by errors or by challenges to validity. These problems will likely increase as people increasingly employ electronic wills. Developing natural language techniques that can determine a will's validity at creation and execution can increase validity, conserve legal resources, and effectuate the author's intent. To this end, this work explores how NLI models can be employed to evaluate the validity of will statements.

NLI in general deals with determining whether a premise entails, contradicts, or is neutral to the hypothesis given. Our project made two important changes to this approach to fit legal documents. First, our dataset contains three input types: statements from wills (as hypotheses), laws (as premises), and *conditions*, which are circumstances at the time of probate (e.g., eligibility of the testator or witnesses). This adaptation is necessary as will statements' validity cannot be evaluated without considering both circumstances at the time of probate ("conditions") and relevant legal rules ("laws"). Also, we switched the labels from entailment, contradict, neutral to support, refute, unrelated to better represent the relation between will statements and laws.
The major contributions of this work are:

- We create an open-access annotated dataset with 1,014 data points, generated from 23 publicly available wills. In addition to the three-tuple setting unique to the legal domain, this dataset also contains texts considerably longer than in other open-domain NLI datasets, such as SNLI (Bowman et al., 2015). The average length of our texts is 269 tokens, while the average lengths of premises and hypotheses in SNLI are 8 and 14 tokens, respectively.

- We demonstrate that validity assessment of legal will statements can be handled reasonably well with state-of-the-art NLI models when trained on our dataset. However, low scores in group accuracy, which is a stricter evaluation measure calculated with a group of positive and negative examples generated from the same statement as a unit, indicate that work remains before NLI models fully understand legal language.

- We explain how the trained models utilize our dataset via ablation tests that remove various

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6076 - 6085
December 7-11, 2022 ©2022 Association for Computational Linguistics
Exhibit 1027 Page 198

pieces of information at prediction time (laws, conditions, or both), and through post-hoc explainability analyses using LIME (Ribeiro et al., 2016). Our analyses indicate that in most cases the NLI classifiers use meaningful phrases from all three pieces of text, indicating the models do capture useful information. However, in some situations, the models use features that are not intuitive for humans, indicating that the task is incomplete.

2 Related Work

Natural language inference (NLI), also known as Recognizing Textual Entailment (RTE), determines entailment relations between a pair of sentences: a premise and a hypothesis. Their relationship is either: a) entailment, if a premise entails a hypothesis; b) contradiction, if a premise contradicts a hypothesis; or c) neutral, if a premise neither entails nor contradicts a hypothesis. NLI has been a key framework in natural language processing since Dagan et al. (2006) proposed the RTE challenge.

Numerous datasets exist for general NLI tasks. For example, the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) dataset contains about 570,000 pairs of sentences (premises and hypotheses) generated from Flickr image captions. More recently, datasets for domain-specific NLI tasks are being introduced. SciNLI (Sadat and Caragea, 2022) is drawn from scientific texts. It contains 107,412 sentence pairs extracted from papers in natural language processing and computational linguistics.

There have been a few NLI-related resources for the legal domain, including IFlyLegal (Wang et al., 2019), StAtutory Reasoning Assessment (SARA) dataset (Holzenberger et al., 2020), Graph-based Causal Inference (GCI) framework (Liu et al., 2021), and ContractNLI (Koreeda and Manning, 2021). One IFlyLegal module was developed for natural language article inference; it provides relevant legal articles when asked legal questions. SARA contains rules extracted from the US Internal Revenue Code and natural language questions. GCI generates causal graphs based on fact descriptions. Though the framework was not specifically designed for law, Liu et al. (2021) demonstrated that it can be utilized for legal text analysis. ContractNLI is a dataset consisting of 607 legal contracts designed to handle document-level natural language inference.

Our work differs from these existing NLI works in two ways: (a) it contains three types of input information; and (b) it operates in a pragmatic middle space with respect to text length: larger than datasets with sentence-level texts, which are insufficient to capture legal details, but shorter than document-level inference datasets.

3 Dataset

Our dataset includes three types of entries (will statements, applicable law, and facts representing external state) rather than the usual two types. Validity often depends upon such external circumstances. Thus, a will statement can be legally valid under some facts, but not others. Hence, a will can change from valid to invalid or vice versa; wills must, at minimum, be evaluated when executed (drafted and signed) and when probated (where a court determines whether to implement its provisions). For example, if a will contains a statement appointing a specific person as executor, that person's eligibility must be verified. This requires information about both the law and the person's circumstances at probate time. For example, according to Tennessee Code section 40-20-115, any person who has been imprisoned cannot be an executor. Thus, to assess eligibility of a Tennessee executor in Tennessee, one must know whether they have been in prison at probate time.

3.1 Data Collection

Our data was collected from the U.S. Wills and Probates datasets in Ancestry, which contains documents in the public domain.[1] We chose 23 wills based on three criteria: 1) whether the wills were handwritten; 2) the wills' execution and probation dates; and 3) where the wills were executed and probated. We excluded handwritten wills due to the difficulty of OCR text recognition. Execution and probation dates can affect interpretation of wills; we excluded wills executed before 1970 and probated before 2000. Lastly, we only collected wills from Tennessee, because including wills from multiple states would require analyzing each state's laws governing wills. Tennessee had the greatest number of wills meeting our criteria.

All collected wills were anonymized. Personal information was replaced with special tokens denoting the type of information (such as [Person-

[1]Court documents are in the public domain in the US.

Exhibit 1027 Page 199

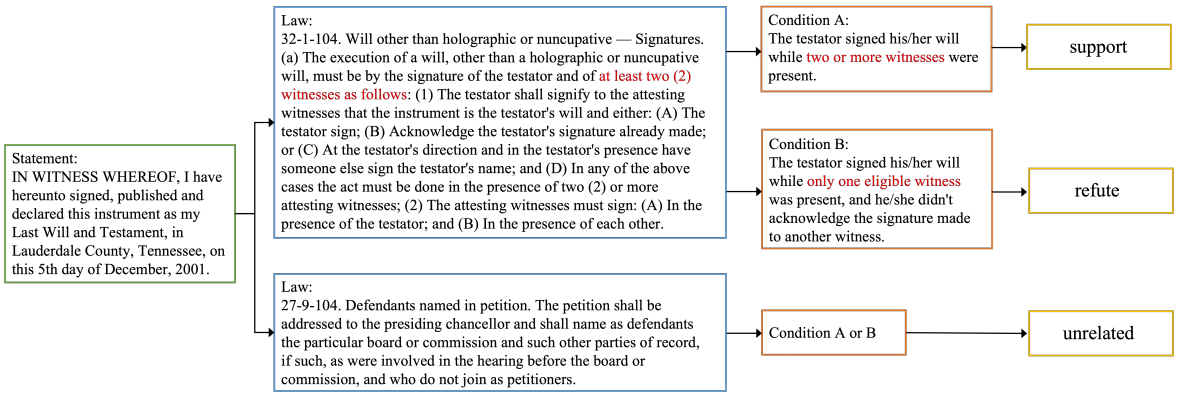


Figure 1: A demonstration of the annotation procedure. Given a law and condition, each statement is evaluated for validity. *Support*: it is supported by the law and condition provided; *Refute*: refuted by the law and condition. *Unrelated*: not relevant to the law. When a statement and a law are unrelated to each other, a condition is not required for classification. However, it is still assigned one of the conditions from either support or refute case. The goal is to prevent NLI models from depending on contextual difference (i.e., "texts without conditions are all unrelated cases") rather than features when making predictions.

n], [Address-n], [Number-n], where n identified each person or object in the will. (Suntwal et al. (2019) suggested this anonymization method). Anonymization was performed manually to prevent including personal information in the dataset.

3.2 Annotation

Each will statement was annotated as *support*, *refute*, or *unrelated* based on a given condition and law. (Our annotators set hypothetical conditions when necessary to categorize statements.)

Importantly, each statement was annotated multiple times. A statement was annotated once as support, once as refute, and thrice as unrelated.[2] This ensured that every statement was labeled with all three classes in the same ratio (support:refute:unrelated=1:1:3). Also, annotating statements three times more with "unrelated" than support and refute enabled including a greater range of laws in the dataset. Formally, our annotation procedure included the following steps:

(1) preprocessing: extracting texts from collected wills using OCR[3], and copying statements into the dataset;

(2) for each statement, identifying the Tennessee legal provision that supported, refuted, or was unrelated to the statement's validity;

(3) adding a condition specifying external circumstances relevant to validity;

(4) repeating these steps to generate five annotations per statement;

(5) anonymizing all statements by replacing personal information with special tokens.

Figure 1 demonstrates the annotation procedure with examples.

Two annotators participated in the task. One is a law student, and annotation was supervised by author Bambauer, a law professor. The other annotator does not have legal training, but ongoing discussions and reviews ensured the uniform quality. The annotators contributed equally. Annotators worked individually but shared annotation guidelines. After the annotation was complete, cross-annotation was conducted on 200 randomly selected items (100 items from each annotator's set; each annotator worked blinded on items drawn from the other annotator's set) to check inter-annotator agreement. We calculated Kappa agreement score based on the cross-annotation result, and the score is 0.91 (rounded to hundredth).

After the annotation was complete, the dataset was split into training, development, and test sets. There are 1014 data points in our dataset, split 50:25:25% (train: 504, development: 255, test: 255). When splitting, a group containing annotations from a single statement (one support, one refute, and three unrelated) was treated as a single

[2] A few exceptions occurred when:
a) it was impossible to create a case where the statement was refuted. In such cases, statements were not annotated as "refute."
b) a statement had more than one relevant legal rule. Such statements were annotated twice as support, twice as refute, and once as unrelated.

[3] The collected wills were scanned documents, so OCR was necessary to extract machine-readable text.

Exhibit 1027 Page 200

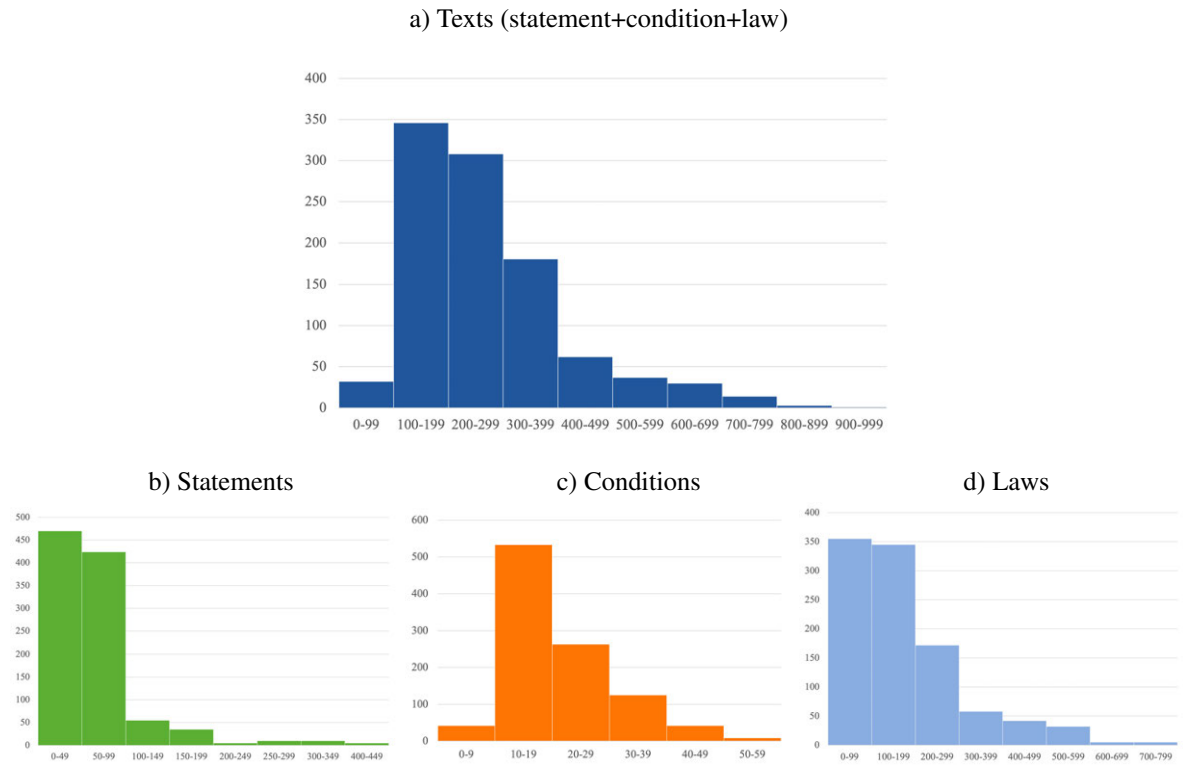


Figure 2: Histograms demonstrating the length distribution of a) full texts (i.e., statement+condition+law), b) statements, c) conditions, and d) laws in terms of token counts. Token counts are plotted on the X-axis; the Y-axis indicates the number of texts/statements/conditions/laws for each bin.

Datasets	Texts	Tokens
Train	504	135218
Development	255	69643
Test	255	68042
Total	**1014**	**272903**

Table 1: The number of texts and tokens contained in each dataset and in total

unit. Each annotator's work was equally represented in all three datasets. All 23 wills appear in more than one of the train, dev, and test partitions. However, given the nature of the will statements from the wills and how they were annotated, data leakage is unlikely. Will statements are independent and generally do not convey information about other statements from the same wills. Summary statistics are in Table 1 and Figure 2.

3.3 Characteristics of the Dataset

Our dataset has two characteristics distinguishing it from other datasets: a) texts are composed of three input types, and b) texts contain a large number of tokens.

First, our task requires a third input type: a condition, since validity often depends upon conditions.

Second, our texts tend to contain a large number of tokens. 44% of our texts contain 200 or more tokens. The average token number was 269.14; 79 texts contain more than 512 tokens (the threshold for most NLI models). Popular NLI datasets are shorter.

3.4 Implementation

We trained multiple NLI models with our dataset to assess their performance: four transformer models and four sentence-transformer models. The transformer models include bert-base-uncased (Devlin et al., 2018), distilbert-base-uncased (Sanh et al., 2019a), roberta-large-mnli (Liu et al., 2019), and longformer-base-4096 (Beltagy et al., 2020). Bert-base-uncased and dilstilbert-base-uncased were trained to set baselines on the task. In addition to the baseline models, we used roberta-large-mnli and longformer-base-4096. Among sentence-transformer models (Reimers and Gurevych, 2019), four pretrained models with top average performances (based on the Model Overview page[4]) were chosen: all-mpnet-base-v2[5], multi-qa-mpnet-base-

[4] https://www.sbert.net

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Exhibit 1027 Page 201

dot-v1[6], all-distilroberta-v1 (Sanh et al., 2019b)[7], and all-MiniLM-L12-v2 (Wang et al., 2020)[8].

To distinguish between statements, laws, and conditions in the concatenated texts, we prefixed each with a special token: `[STATE]`, `[LAW]`, and `[COND]`.

Both transformer models and sentence-transformer models were trained on PyTorch 1.11.0 with Cuda 11.3. All transformer models were trained using the Trainer class provided by HuggingFace. Learning rates and training epochs were tuned on the development partition. Sentence-transformer models were trained with Sentence Transformer Fine-Tuning (SetFit) proposed by Wasserblat (2021). It was slightly adapted to fit the multi-class classification task, but its fundamentals remain intact. SetFit utilizes sentence pairs generated within the same class as training data. It fits the model with the data to minimize the Softmax Loss. Once the model is fitted, it is used to encode the training and development (or test, when it is a testing phase) datasets. The encoded data is used to fit the Logistic Regression classifier (with 'liblinear' solver) which, finally, makes predictions.

All models except for longformer-base-4096 were trained and evaluated on the truncated datasets due to the models' token number limitations. The models can only process texts with 512 or fewer tokens without truncation (all-MiniLM-L12-v2: 256 tokens; all-mpnet-base-v2: 384 tokens; other models, except longformer-base-4096: 512 tokens). When truncating datasets, the ratio of average token numbers among three input types (statements, laws, and conditions) was considered to prevent losing excessive information from a single type of input.

4 Results and Analysis

4.1 Evaluation Measures

We report results using standard accuracy, precision, recall, and F1 scores. To ensure all labels are well represented, precision, recall, and F1 scores were computed with Macro average. Additionally, as suggested by Elazar et al. (2021), we introduce a new measure called group accuracy (GA). GA is calculated with a group of positive and negative examples from the same statement (rather than each text) as a unit. If a group has one or more incorrect predictions, the group is incorrect. If a model correctly understands a will statement, it should perform equally well on all examples from the same statement.

4.2 Results from Trained Models

Table 2 shows the models' performances on our test partition. Overall, the models demonstrated good performances. Each achieved more than 80% in all metrics but group accuracy. Roberta-large-mnli showed the best performance. It achieved over 96% in all metrics but group accuracy (84.31%), suggesting it can handle the task reasonably well. However, the large difference between accuracy and group accuracy suggests that the models' understanding of the task is superficial.

Accuracy for the *unrelated* label was higher than for *support* and *refute* in all models except all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1. This higher accuracy for *unrelated* label may be partially attributable to the gap between accuracy and group accuracy, as it would inflate overall accuracy. However, since the gap between accuracy and group accuracy is not significantly smaller for all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1 (i.e., the models where accuracy for *unrelated* was not higher than the other labels), it is more likely that the gap originated from the strictness of group accuracy and the models' superficial understanding of will statements.

Further, roberta-large-mnli (trained with truncated inputs) showed better performance than longformer-base-4096 (trained with full length inputs). This suggests that models with token number limitations can still perform well with long inputs when truncated properly.

4.3 Experiment with Different Input Lengths

To investigate whether the models' performance was affected by input length and/or input truncation, we varied input lengths. Inputs were classified into three categories by length: a) short (equal or less than 192 or 256 tokens); b) regular (larger than 192 or 256 tokens and equal or less than 384 or 512); and c) long (larger than 384 or 512)[9]. The

[6] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

[7] https://huggingface.co/sentence-transformers/all-distilroberta-v1

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[9] For all-mpnet-base-v2 model, 192 and 384 were used as dividing points instead of 256 and 512, as its maximum token length is 384.

Exhibit 1027 Page 202

Model	Precision	Recall	F1	Accuracy	GA
Transformers models					
bert-base-uncased	81.47	82.20	81.81	87.06	49.02
distilbert-base-uncased	82.37	82.59	82.48	87.06	50.98
longformer-base-4096	94.09	93.69	93.85	94.90	74.51
roberta-large-mnli	**96.67**	**96.25**	**96.42**	**96.86**	**84.31**
Sentence-Transformers models					
all-MiniLM-L12-v2	81.50	84.54	82.64	85.49	50.98
all-disilroberta-v1	83.43	85.99	84.22	87.06	49.02
multi-qa-mpnet-base-dot-v1	91.48	94.13	92.65	93.73	74.51
all-mpnet-base-v2	92.93	95.05	93.82	94.90	76.47

Table 2: Results of all classifiers trained on our dataset using five measures: precision, recall, F1, accuracy, and group accuracy (GA in table, see Section 4.1). Precision, recall, and F1 scores were computed with Macro average. Group accuracy is calculated with positive and negative examples generated from the same statement as a unit. If a group contains one or more incorrect predictions, the group is considered incorrect.

Model	Precision	Recall	F1	Accuracy
all-mpnet-base-v2 (sample n = 52)				
token <= 192	93.75	93.15	92.91	94.23
192 < token <= 384	90.48	89.10	89.63	94.23
384 < token	86.35	89.49	87.08	88.46
longformer-base-4096 (sample n = 20)				
token <= 256	95.24	88.89	90.77	95.00
256 < token <= 512	80.56	71.67	70.93	80.00
512 < token	73.81	66.67	66.93	75.00
roberta-large-mnli (sample n = 20)				
token <= 256	100	100	100	100
256 < token <= 512	94.87	85.00	88.76	90.00
512 < token	94.44	93.33	93.27	95.00

Table 3: Results with different input lengths. Models generally performed worse when token numbers in inputs were larger. Roberta-large-mnli performed better than longformer-base-4096 in predicting long inputs (token > 512) despite truncation. Roberta-large-mnli also showed better performance on long inputs than regular inputs, indicating input truncation did not worsen the model's performance.

number of inputs in each category varied. To control the impact of varying sample size, categories with larger sample sizes were reduced by random sampling to match the smallest sample size. For this experiment, we used the three models with best performance on the full dataset.

Table 3 shows the results. Models generally performed worse with larger input token numbers. All three models performed best with inputs with smaller token numbers. One interesting finding is that roberta-large-mnli performed better than longformer-base-4096 in predicting inputs with large token numbers (n > 512) despite truncation. Roberta-large-mnli showed better performance on long inputs than regular inputs, indicating truncation did not negatively affect the model's performance. This finding aligns with the observation from the overall results that models with token number limitation can still show good performances on long inputs when properly truncated. However, results differed with a smaller token number limitation. All-mpnet-base-v2 performed worst on long inputs (where truncation occurred), indicating truncation negatively impacted performance.

4.4 Results from Ablation Experiment

To determine whether the models correctly rely on all three types of texts (statements, conditions, and laws), we conducted an ablation experiment, training the models with datasets lacking one (law or condition) or two types (law+condition) of inputs. Poliak et al. (2018) suggested a similar experiment

Exhibit 1027 Page 203

Model	Precision	Recall	F1	Accuracy	GA
Statements and laws					
all-mpnet-base-v2	75.64	76.47	75.98	81.96	27.45
longformer-base-4096	75.72	75.98	75.79	82.75	37.25
roberta-large-mnli	77.93	77.51	77.55	84.31	39.22
Statements and conditions					
all-mpnet-base-v2	46.67	57.13	44.80	44.31	0.00
longformer-base-4096	58.00	49.44	51.30	58.82	0.00
roberta-large-mnli	37.63	41.72	38.08	56.86	0.00
Statements only					
all-mpnet-base-v2	36.76	35.94	34.04	36.47	0.00
longformer-base-4096	18.17	33.33	23.52	54.51	0.00
roberta-large-mnli	18.17	33.33	23.52	54.51	0.00

Table 4: Results from the ablation experiment where three models were trained with datasets lacking one (law or condition) or two types (law+condition) of inputs. Results demonstrate that the models' performances deteriorate if one or more input type(s) is omitted. Thus, both types of inputs affect the models' performances, though laws have larger impact than conditions.

to this one to set hypothesis-only baselines. We used the three models with the best performances on the full dataset for this experiment.

Table 4 shows the results. Performance deteriorated if any input type was omitted. Among the models trained on partial data, those trained with statements and laws performed better, with F1 scores over 75%. However, group accuracy dropped to the 30s or even 20%s, indicating actual understanding decreased considerably. Models trained with statements and conditions performed significantly worse. Results ranged between 38% and 52% for F1, and group accuracy dropped to 0. Models trained only with statements performed worst, with F1 scores dropping to 35% or lower.

This degradation from partial data shows including both conditions and laws positively affects the models' performance. Models use all three types of inputs when making predictions[10]. Laws have more impact on performance than conditions, since deterioration in results without laws was more significant than from models trained without conditions. This is realistic, since laws set the parameters within which conditions operate to make a given provision valid or invalid.

4.5 Understanding Results with LIME

To clarify model behavior, we used Local Interpretable Model-Agnostic Explanations (LIME) to explain our classifiers' predictions (Ribeiro et al., 2016). We implemented LIME on predictions made by the best performing model (roberta-large-mnli, fine-tuned on our data). The LIME results revealed that the model tends to correctly rely on all three texts (statements, laws, conditions) for a prediction, and uses meaningful features in many cases. However, sometimes the model utilized features that are not intuitive for humans.

4.5.1 A Correct Example

Figure 3 shows a case where the model made a correct prediction based on features sensible to humans. Figure 3's text is a statement saying two or more witnesses witnessed the testator signing the will and signed in each other's presence. It includes a condition stating that one of the two witnesses was ineligible to serve under Tennessee law, and the law specifying witness eligibility in Tennessee. The given condition and the law invalidate the will statement, rendering it as *refute*. The model correctly predicted the outcome based on tokens such as *ineligible*, *witnesses*, *One*, *testament*, and *2*. The top two tokens with greatest impact on the model's prediction were *ineligible* and *witnesses*, with importance score of 0.60 and 0.14, respectively. It is

[10] A reviewer suggested an experiment where laws and conditions are packed into premises and treat the task as normal NLI. A pilot experiment with the best performing model (roberta-large-mnli) found a slight deterioration in performance when laws and conditions were combined as premises (precision: 93.09, recall: 93.89, F1: 93.41, accuracy: 94.12, GA: 74.51; see Table 2 for comparison with main result). The biggest difference came from GA score (74.51 vs. 84.31). This lowered GA score implies that the model's understanding worsened when conditions and laws were combined into premises. Given this preliminary result, it is likely that three-way distinction in input types positively contributes to the model's performance. Nevertheless, as the performance difference is not significantly large, further investigation is needed.

Exhibit 1027 Page 204

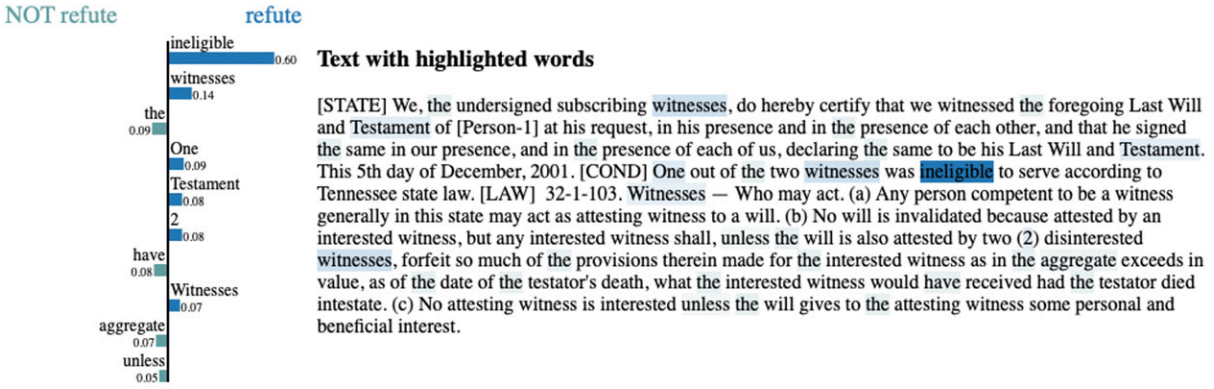


Figure 3: Examples of LIME explanations showing a case where the model (roberta-large-mnli) makes correct predictions based on features sensible to humans.

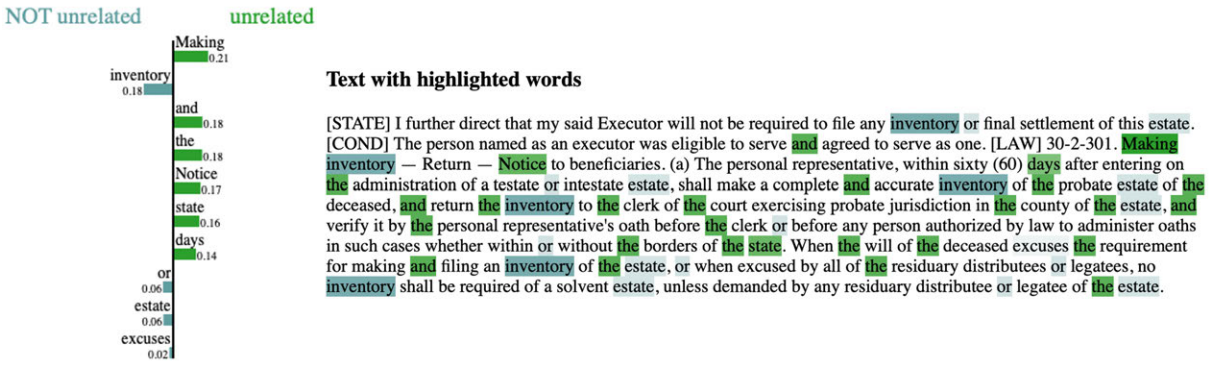


Figure 4: An example of a LIME explanation showing a case where the model (roberta-large-mnli) makes an incorrect prediction based on the tokens such as *Making*, *and*, and *the* which bear little semantic relevancy to the gist of the text.

understandable that these tokens have high scores, as they are the keywords ("**One** out of two (**2**) **witnesses** is **ineligible**") which provide grounds for revoking a will statement.

4.5.2 An Incorrect Example

Figure 4 shows a case where the model made an incorrect prediction based on irrelevant tokens. The intent of the will statement is to excuse the Executor from filing inventory. This is not contradicted by the condition and law, requiring the model to classify the text as *support*. Nevertheless, the model incorrectly classified it as *unrelated*. The reason for the incorrect prediction is that the model (probably) relied on irrelevant tokens such as *Making*, *and*, and *the*, which bear little semantic relevancy to the text, versus more relevant tokens such as *inventories* and *excuses* when making the prediction. This is likely due to overfitting induced by longer texts. This LIME explanation shows that even the best performing model still has room for improvement.

5 Future Work

This work can be expanded in several directions. First, it can be extended to cover multiple states by adapting the models or adding more data. Future work can investigate if models trained on a single state's data can be adapted to evaluate data from other states. Also, wills from multiple states could be added to the dataset. We expect the augmented dataset would enhance the models' performance on evaluating wills from other states.

This work can be expanded to other legal domains. Models trained on our data can be adapted to similar tasks in other legal domains.

Lastly, this work can be extended by investigating novel transformer models. Given the uniqueness of our dataset with regard to the number of inputs and text length, it is likely that further experiment and modification is needed to handle such characteristics.

Exhibit 1027 Page 205

6 Conclusion

This work presented an annotated dataset for natural language inference (NLI) in the legal domain, consisting of 1,014 data points generated from 23 publicly available wills. The dataset is novel for two reasons: it included texts with three input types (statement, law, and condition) rather than two (premise and hypothesis) in the traditional NLI, and included texts are longer than in general NLI datasets. The NLI models trained on our dataset showed reasonable performance in assessing the validity of will statements. Ablative experiments demonstrated that the models' performances worsen if any input type (condition, law, or both) is omitted. This suggests that the models utilize all three input types. The LIME implementation reveals that even the best performing model makes errors in some cases by using semantically irrelevant tokens. Our open-access dataset is publicly available at: https://github.com/ml4ai/nli4wills-corpus

Limitations

Our dataset consists of a relatively small number of data points (1,014 texts). Annotating will statements with relevant laws and conditions is a highly demanding, time intensive task. A larger size of dataset would likely further improve the models' performances.

Our dataset only includes wills executed and probated in Tennessee, with execution after 1970 and probate after 2000. Due to these restrictions, our framework might produce incorrect results on inputs from wills from different settings. Supplementing the dataset with wills from more diverse settings would address this limitation. Even though the scope is limited to a single state (Tennessee), our study demonstrates that transformer models trained on the dataset can evaluate the validity of statements from wills with reasonable accuracy. Future additions to our dataset will be available at the same URL: https://github.com/ml4ai/nli4wills-corpus

This work does not involve humans in the loop. Considering how crucial accuracy is for the task (i.e., legal validity evaluation), the work would have benefited much from involving domain experts. Even though our study discovered that the state-of-the-art transformer models can show good performances (over 85% accuracy in all 8 models) without human interaction, it was also found that the models' understanding on the task is rather superficial (GAs ranging from 49-85% in Table 2). Including humans-in-the-loop could be a solution for enhancing the models' understanding on the task.

Ethics Statement

We collected legal wills from Ancestry as a part of our dataset creation process. The wills probated in court are in the public domain in the US, and we did not violate Ancestry's Terms and Conditions when collecting wills. We also anonymized the wills by replacing any personal identifiable information contained in the documents with special tokens.

Our datasets and codes are released to public. We believe our released datasets and codes would contribute to society by promoting further NLI endeavors in the legal domain. The resources could potentially assist with people reviewing wills, but they should not be considered as legal advice. To avoid any confusion, we placed a disclaimer that the users must not rely on any information provided from our resources when making legal decisions and should instead consult with an attorney.

Acknowledgements

We thank the reviewers for their thoughtful comments and suggestions. This work was partially supported by the National Science Foundation (NSF) under grant 2217215, and by University of Arizona's Provost Investment Fund. Mihai Surdeanu and Clayton Morrison declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification,*

and Recognising Tectual Entailment, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *CoRR*, abs/2005.05257.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *CoRR*, abs/2110.01799.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. *CoRR*, abs/2104.09420.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3413–3418, Hong Kong, China. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Ziyue Wang, Baoxin Wang, Xingyi Duan, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2019. IFlyLegal: A Chinese legal system for consultation, law searching, and document analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 97–102, Hong Kong, China. Association for Computational Linguistics.

Moshe Wasserblat. 2021. Sentence transformer fine-tuning (setfit): Outperforming gpt-3 on few-shot text-classification while being 1600 times smaller.

Exhibit 1027 Page 207

https://web.archive.org/web/20221101104607/https://arxiv.org/pdf/2210.16989v1.pdf

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu
The University of Arizona, Tucson, Arizona, USA
{alicekwak, jisraelsen, claytonm, derekbambauer, msurdeanu}@arizona.edu

arXiv:2210.16989v1 [cs.CL] 30 Oct 2022

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

1 Introduction

Natural language inference (NLI) in the legal domain has not been widely investigated, despite its importance and potential. One such important NLI application is validity assessment of legal documents such as wills. Legal procedures for creating and executing wills are evolving rapidly. Processing a will via probate is a costly, time-consuming process that can be exacerbated by errors or by challenges to validity. These problems will likely increase as people increasingly employ electronic wills. Developing natural language techniques that can determine a will's validity at creation and execution can increase validity, conserve legal resources, and effectuate the author's intent. To this end, this work explores how NLI models can be employed to evaluate the validity of will statements.

NLI in general deals with determining whether a premise entails, contradicts, or is neutral to the hypothesis given. Our project made two important changes to this approach to fit legal documents. First, our dataset contains three input types: statements from wills (as hypotheses), laws (as premises), and *conditions*, which are circumstances at the time of probate (e.g., eligibility of the testator or witnesses). This adaptation is necessary as will statements' validity cannot be evaluated without considering both circumstances at the time of probate ("conditions") and relevant legal rules ("laws"). Also, we switched the labels from entailment, contradict, neutral to support, refute, unrelated to better represent the relation between will statements and laws.
The major contributions of this work are:

- We create an open-access annotated dataset with 1,014 data points, generated from 23 publicly available wills. In addition to the three-tuple setting unique to the legal domain, this dataset also contains texts considerably longer than in other open-domain NLI datasets, such as SNLI (Bowman et al., 2015). The average length of our texts is 269 tokens, while the average lengths of premises and hypotheses in SNLI are 8 and 14 tokens, respectively.

- We demonstrate that validity assessment of legal will statements can be handled reasonably well with state-of-the-art NLI models when trained on our dataset. However, low scores in group accuracy, which is a stricter evaluation measure calculated with a group of positive and negative examples generated from the same statement as a unit, indicate that work remains before NLI models fully understand legal language.

- We explain how the trained models utilize our dataset via ablation tests that remove various

pieces of information at prediction time (laws, conditions, or both), and through post-hoc explainability analyses using LIME (Ribeiro et al., 2016). Our analyses indicate that in most cases the NLI classifiers use meaningful phrases from all three pieces of text, indicating the models do capture useful information. However, in some situations, the models use features that are not intuitive for humans, indicating that the task is incomplete.

2 Related Work

Natural language inference (NLI), also known as Recognizing Textual Entailment (RTE), determines entailment relations between a pair of sentences: a premise and a hypothesis. Their relationship is either: a) entailment, if a premise entails a hypothesis; b) contradiction, if a premise contradicts a hypothesis; or c) neutral, if a premise neither entails nor contradicts a hypothesis. NLI has been a key framework in natural language processing since Dagan et al. (2006) proposed the RTE challenge.

Numerous datasets exist for general NLI tasks. For example, the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) dataset contains about 570,000 pairs of sentences (premises and hypotheses) generated from Flickr image captions. More recently, datasets for domain-specific NLI tasks are being introduced. SciNLI (Sadat and Caragea, 2022) is drawn from scientific texts. It contains 107,412 sentence pairs extracted from papers in natural language processing and computational linguistics.

There have been a few NLI-related resources for the legal domain, including IFlyLegal (Wang et al., 2019), StAtutory Reasoning Assessment (SARA) dataset (Holzenberger et al., 2020), Graph-based Causal Inference (GCI) framework (Liu et al., 2021), and ContractNLI (Koreeda and Manning, 2021). One IFlyLegal module was developed for natural language article inference; it provides relevant legal articles when asked legal questions. SARA contains rules extracted from the US Internal Revenue Code and natural language questions. GCI generates causal graphs based on fact descriptions. Though the framework was not specifically designed for law, Liu et al. (2021) demonstrated that it can be utilized for legal text analysis. ContractNLI is a dataset consisting of 607 legal contracts designed to handle document-level natural language inference.

Our work differs from these existing NLI works in two ways: (a) it contains three types of input information; and (b) it operates in a pragmatic middle space with respect to text length: larger than datasets with sentence-level texts, which are insufficient to capture legal details, but shorter than document-level inference datasets.

3 Dataset

Our dataset includes three types of entries (will statements, applicable law, and facts representing external state) rather than the usual two types. Validity often depends upon such external circumstances. Thus, a will statement can be legally valid under some facts, but not others. Hence, a will can change from valid to invalid or vice versa; wills must, at minimum, be evaluated when executed (drafted and signed) and when probated (where a court determines whether to implement its provisions). For example, if a will contains a statement appointing a specific person as executor, that person's eligibility must be verified. This requires information about both the law and the person's circumstances at probate time. For example, according to Tennessee Code section 40-20-115, any person who has been imprisoned cannot be an executor. Thus, to assess eligibility of a Tennessee executor in Tennessee, one must know whether they have been in prison at probate time.

3.1 Data Collection

Our data was collected from the U.S. Wills and Probates datasets in Ancestry, which contains documents in the public domain.[1] We chose 23 wills based on three criteria: 1) whether the wills were handwritten; 2) the wills' execution and probation dates; and 3) where the wills were executed and probated. We excluded handwritten wills due to the difficulty of OCR text recognition. Execution and probation dates can affect interpretation of wills; we excluded wills executed before 1970 and probated before 2000. Lastly, we only collected wills from Tennessee, because including wills from multiple states would require analyzing each state's laws governing wills. Tennessee had the greatest number of wills meeting our criteria.

All collected wills were anonymized. Personal information was replaced with special tokens denoting the type of information (such as [Person-

[1]Court documents are in the public domain in the US.

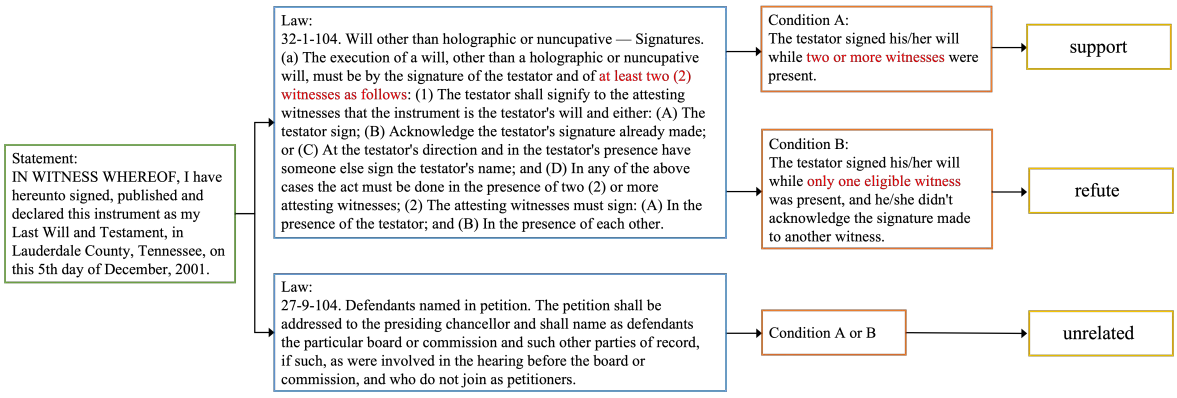


Figure 1: A demonstration of the annotation procedure. Given a law and condition, each statement is evaluated for validity. *Support*: it is supported by the law and condition provided; *Refute*: refuted by the law and condition. *Unrelated*: not relevant to the law. When a statement and a law are unrelated to each other, a condition is not required for classification. However, it is still assigned one of the conditions from either support or refute case. The goal is to prevent NLI models from depending on contextual difference (i.e., "texts without conditions are all unrelated cases") rather than features when making predictions.

n], [Address-n], [Number-n], where n identified each person or object in the will. (Suntwal et al. (2019) suggested this anonymization method). Anonymization was performed manually to prevent including personal information in the dataset.

3.2 Annotation

Each will statement was annotated as *support*, *refute*, or *unrelated* based on a given condition and law. (Our annotators set hypothetical conditions when necessary to categorize statements.)

Importantly, each statement was annotated multiple times. A statement was annotated once as support, once as refute, and thrice as unrelated.[2] This ensured that every statement was labeled with all three classes in the same ratio (support:refute:unrelated=1:1:3). Also, annotating statements three times more with "unrelated" than support and refute enabled including a greater range of laws in the dataset. Formally, our annotation procedure included the following steps:

(1) preprocessing: extracting texts from collected wills using OCR[3], and copying statements into the dataset;

(2) for each statement, identifying the Tennessee legal provision that supported, refuted, or was unrelated to the statement's validity;

(3) adding a condition specifying external circumstances relevant to validity;

(4) repeating these steps to generate five annotations per statement;

(5) anonymizing all statements by replacing personal information with special tokens.

Figure 1 demonstrates the annotation procedure with examples.

Two annotators participated in the task. One is a law student, and annotation was supervised by author Bambauer, a law professor. The other annotator does not have legal training, but ongoing discussions and reviews ensured the uniform quality. The annotators contributed equally. Annotators worked individually but shared annotation guidelines. After the annotation was complete, cross-annotation was conducted on 200 randomly selected items (100 items from each annotator's set; each annotator worked blinded on items drawn from the other annotator's set) to check inter-annotator agreement. We calculated Kappa agreement score based on the cross-annotation result, and the score is 0.91 (rounded to hundredth).

After the annotation was complete, the dataset was split into training, development, and test sets. There are 1014 data points in our dataset, split 50:25:25% (train: 504, development: 255, test: 255). When splitting, a group containing annotations from a single statement (one support, one refute, and three unrelated) was treated as a single

[2] A few exceptions occurred when:
a) it was impossible to create a case where the statement was refuted. In such cases, statements were not annotated as "refute."
b) a statement had more than one relevant legal rule. Such statements were annotated twice as support, twice as refute, and once as unrelated.

[3] The collected wills were scanned documents, so OCR was necessary to extract machine-readable text.

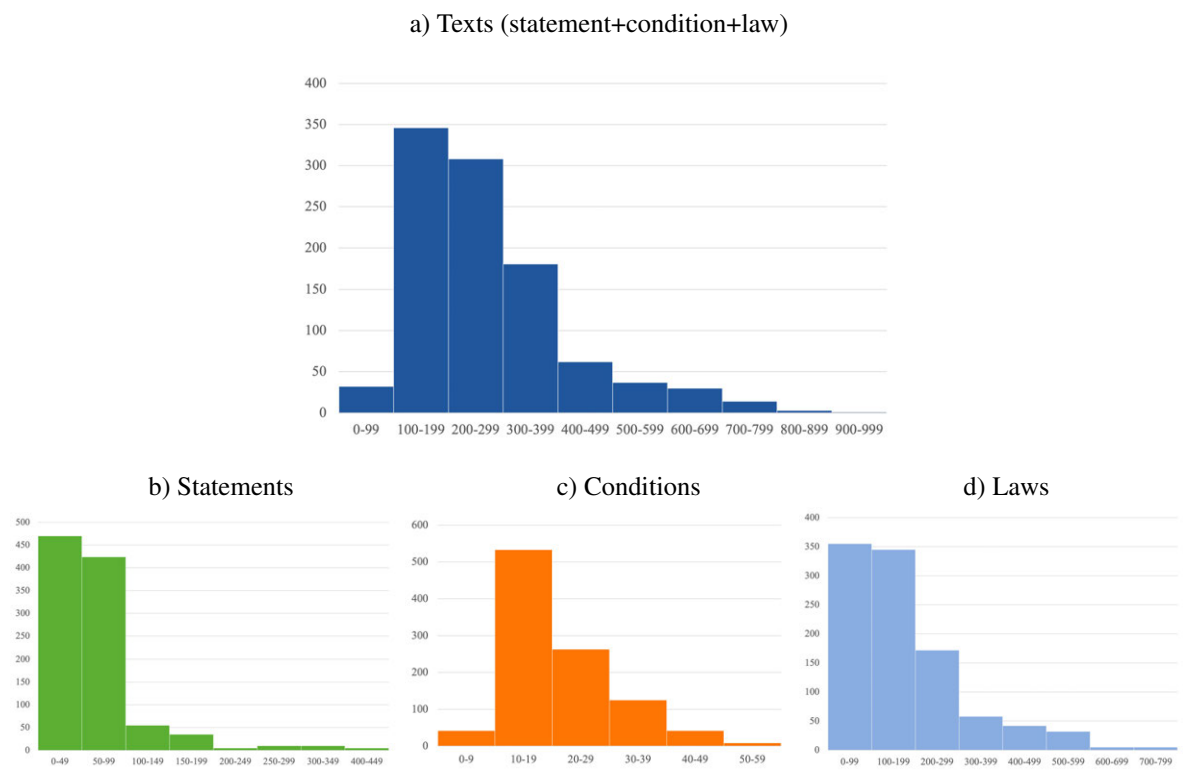


Figure 2: Histograms demonstrating the length distribution of a) full texts (i.e., statement+condition+law), b) statements, c) conditions, and d) laws in terms of token counts. Token counts are plotted on the X-axis; the Y-axis indicates the number of texts/statements/conditions/laws for each bin.

Datasets	Texts	Tokens
Train	504	135218
Development	255	69643
Test	255	68042
Total	**1014**	**272903**

Table 1: The number of texts and tokens contained in each dataset and in total

unit. Each annotator's work was equally represented in all three datasets. All 23 wills appear in more than one of the train, dev, and test partitions. However, given the nature of the will statements from the wills and how they were annotated, data leakage is unlikely. Will statements are independent and generally do not convey information about other statements from the same wills. Summary statistics are in Table 1 and Figure 2.

3.3 Characteristics of the Dataset

Our dataset has two characteristics distinguishing it from other datasets: a) texts are composed of three input types, and b) texts contain a large number of tokens.

First, our task requires a third input type: a condition, since validity often depends upon conditions.

Second, our texts tend to contain a large number of tokens. 44% of our texts contain 200 or more tokens. The average token number was 269.14; 79 texts contain more than 512 tokens (the threshold for most NLI models). Popular NLI datasets are shorter.

3.4 Implementation

We trained multiple NLI models with our dataset to assess their performance: four transformer models and four sentence-transformer models. The transformer models include bert-base-uncased (Devlin et al., 2018), distilbert-base-uncased (Sanh et al., 2019a), roberta-large-mnli (Liu et al., 2019), and longformer-base-4096 (Beltagy et al., 2020). Bert-base-uncased and dilstilbert-base-uncased were trained to set baselines on the task. In addition to the baseline models, we used roberta-large-mnli and longformer-base-4096. Among sentence-transformer models (Reimers and Gurevych, 2019), four pretrained models with top average performances (based on the Model Overview page[4]) were chosen: all-mpnet-base-v2[5], multi-qa-mpnet-base-

[4] https://www.sbert.net

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

dot-v1[6], all-distilroberta-v1 (Sanh et al., 2019b)[7], and all-MiniLM-L12-v2 (Wang et al., 2020)[8].

To distinguish between statements, laws, and conditions in the concatenated texts, we prefixed each with a special token: `[STATE]`, `[LAW]`, and `[COND]`.

Both transformer models and sentence-transformer models were trained on PyTorch 1.11.0 with Cuda 11.3. All transformer models were trained using the Trainer class provided by HuggingFace. Learning rates and training epochs were tuned on the development partition. Sentence-transformer models were trained with Sentence Transformer Fine-Tuning (SetFit) proposed by Wasserblat (2021). It was slightly adapted to fit the multi-class classification task, but its fundamentals remain intact. SetFit utilizes sentence pairs generated within the same class as training data. It fits the model with the data to minimize the Softmax Loss. Once the model is fitted, it is used to encode the training and development (or test, when it is a testing phase) datasets. The encoded data is used to fit the Logistic Regression classifier (with 'liblinear' solver) which, finally, makes predictions.

All models except for longformer-base-4096 were trained and evaluated on the truncated datasets due to the models' token number limitations. The models can only process texts with 512 or fewer tokens without truncation (all-MiniLM-L12-v2: 256 tokens; all-mpnet-base-v2: 384 tokens; other models, except longformer-base-4096: 512 tokens). When truncating datasets, the ratio of average token numbers among three input types (statements, laws, and conditions) was considered to prevent losing excessive information from a single type of input.

4 Results and Analysis

4.1 Evaluation Measures

We report results using standard accuracy, precision, recall, and F1 scores. To ensure all labels are well represented, precision, recall, and F1 scores were computed with Macro average. Additionally, as suggested by Elazar et al. (2021), we introduce a new measure called group accuracy (GA). GA is calculated with a group of positive and negative examples from the same statement (rather than each text) as a unit. If a group has one or more incorrect predictions, the group is incorrect. If a model correctly understands a will statement, it should perform equally well on all examples from the same statement.

4.2 Results from Trained Models

Table 2 shows the models' performances on our test partition. Overall, the models demonstrated good performances. Each achieved more than 80% in all metrics but group accuracy. Roberta-large-mnli showed the best performance. It achieved over 96% in all metrics but group accuracy (84.31%), suggesting it can handle the task reasonably well. However, the large difference between accuracy and group accuracy suggests that the models' understanding of the task is superficial.

Accuracy for the *unrelated* label was higher than for *support* and *refute* in all models except all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1. This higher accuracy for *unrelated* label may be partially attributable to the gap between accuracy and group accuracy, as it would inflate overall accuracy. However, since the gap between accuracy and group accuracy is not significantly smaller for all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1 (i.e., the models where accuracy for *unrelated* was not higher than the other labels), it is more likely that the gap originated from the strictness of group accuracy and the models' superficial understanding of will statements.

Further, roberta-large-mnli (trained with truncated inputs) showed better performance than longformer-base-4096 (trained with full length inputs). This suggests that models with token number limitations can still perform well with long inputs when truncated properly.

4.3 Experiment with Different Input Lengths

To investigate whether the models' performance was affected by input length and/or input truncation, we varied input lengths. Inputs were classified into three categories by length: a) short (equal or less than 192 or 256 tokens); b) regular (larger than 192 or 256 tokens and equal or less than 384 or 512); and c) long (larger than 384 or 512)[9]. The

[6] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

[7] https://huggingface.co/sentence-transformers/all-distilroberta-v1

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[9] For all-mpnet-base-v2 model, 192 and 384 were used as dividing points instead of 256 and 512, as its maximum token length is 384.

Model	Precision	Recall	F1	Accuracy	GA
Transformers models					
bert-base-uncased	81.47	82.20	81.81	87.06	49.02
distilbert-base-uncased	82.37	82.59	82.48	87.06	50.98
longformer-base-4096	94.09	93.69	93.85	94.90	74.51
roberta-large-mnli	**96.67**	**96.25**	**96.42**	**96.86**	**84.31**
Sentence-Transformers models					
all-MiniLM-L12-v2	81.50	84.54	82.64	85.49	50.98
all-disilroberta-v1	83.43	85.99	84.22	87.06	49.02
multi-qa-mpnet-base-dot-v1	91.48	94.13	92.65	93.73	74.51
all-mpnet-base-v2	92.93	95.05	93.82	94.90	76.47

Table 2: Results of all classifiers trained on our dataset using five measures: precision, recall, F1, accuracy, and group accuracy (GA in table, see Section 4.1). Precision, recall, and F1 scores were computed with Macro average. Group accuracy is calculated with positive and negative examples generated from the same statement as a unit. If a group contains one or more incorrect predictions, the group is considered incorrect.

Model	Precision	Recall	F1	Accuracy
all-mpnet-base-v2 (sample n = 52)				
token <= 192	93.75	93.15	92.91	94.23
192 < token <= 384	90.48	89.10	89.63	94.23
384 < token	86.35	89.49	87.08	88.46
longformer-base-4096 (sample n = 20)				
token <= 256	95.24	88.89	90.77	95.00
256 < token <= 512	80.56	71.67	70.93	80.00
512 < token	73.81	66.67	66.93	75.00
roberta-large-mnli (sample n = 20)				
token <= 256	100	100	100	100
256 < token <= 512	94.87	85.00	88.76	90.00
512 < token	94.44	93.33	93.27	95.00

Table 3: Results with different input lengths. Models generally performed worse when token numbers in inputs were larger. Roberta-large-mnli performed better than longformer-base-4096 in predicting long inputs (token > 512) despite truncation. Roberta-large-mnli also showed better performance on long inputs than regular inputs, indicating input truncation did not worsen the model's performance.

number of inputs in each category varied. To control the impact of varying sample size, categories with larger sample sizes were reduced by random sampling to match the smallest sample size. For this experiment, we used the three models with best performance on the full dataset.

Table 3 shows the results. Models generally performed worse with larger input token numbers. All three models performed best with inputs with smaller token numbers. One interesting finding is that roberta-large-mnli performed better than longformer-base-4096 in predicting inputs with large token numbers (n > 512) despite truncation. Roberta-large-mnli showed better performance on long inputs than regular inputs, indicating truncation did not negatively affect the model's performance. This finding aligns with the observation from the overall results that models with token number limitation can still show good performances on long inputs when properly truncated. However, results differed with a smaller token number limitation. All-mpnet-base-v2 performed worst on long inputs (where truncation occurred), indicating truncation negatively impacted performance.

4.4 Results from Ablation Experiment

To determine whether the models correctly rely on all three types of texts (statements, conditions, and laws), we conducted an ablation experiment, training the models with datasets lacking one (law or condition) or two types (law+condition) of inputs. Poliak et al. (2018) suggested a similar experiment

Model	Precision	Recall	F1	Accuracy	GA
Statements and laws					
all-mpnet-base-v2	75.64	76.47	75.98	81.96	27.45
longformer-base-4096	75.72	75.98	75.79	82.75	37.25
roberta-large-mnli	77.93	77.51	77.55	84.31	39.22
Statements and conditions					
all-mpnet-base-v2	46.67	57.13	44.80	44.31	0.00
longformer-base-4096	58.00	49.44	51.30	58.82	0.00
roberta-large-mnli	37.63	41.72	38.08	56.86	0.00
Statements only					
all-mpnet-base-v2	36.76	35.94	34.04	36.47	0.00
longformer-base-4096	18.17	33.33	23.52	54.51	0.00
roberta-large-mnli	18.17	33.33	23.52	54.51	0.00

Table 4: Results from the ablation experiment where three models were trained with datasets lacking one (law or condition) or two types (law+condition) of inputs. Results demonstrate that the models' performances deteriorate if one or more input type(s) is omitted. Thus, both types of inputs affect the models' performances, though laws have larger impact than conditions.

to this one to set hypothesis-only baselines. We used the three models with the best performances on the full dataset for this experiment.

Table 4 shows the results. Performance deteriorated if any input type was omitted. Among the models trained on partial data, those trained with statements and laws performed better, with F1 scores over 75%. However, group accuracy dropped to the 30s or even 20%s, indicating actual understanding decreased considerably. Models trained with statements and conditions performed significantly worse. Results ranged between 38% and 52% for F1, and group accuracy dropped to 0. Models trained only with statements performed worst, with F1 scores dropping to 35% or lower.

This degradation from partial data shows including both conditions and laws positively affects the models' performance. Models use all three types of inputs when making predictions[10]. Laws have more impact on performance than conditions, since deterioration in results without laws was more significant than from models trained without conditions. This is realistic, since laws set the parameters within which conditions operate to make a given provision valid or invalid.

4.5 Understanding Results with LIME

To clarify model behavior, we used Local Interpretable Model-Agnostic Explanations (LIME) to explain our classifiers' predictions (Ribeiro et al., 2016). We implemented LIME on predictions made by the best performing model (roberta-large-mnli, fine-tuned on our data). The LIME results revealed that the model tends to correctly rely on all three texts (statements, laws, conditions) for a prediction, and uses meaningful features in many cases. However, sometimes the model utilized features that are not intuitive for humans.

4.5.1 A Correct Example

Figure 3 shows a case where the model made a correct prediction based on features sensible to humans. Figure 3's text is a statement saying two or more witnesses witnessed the testator signing the will and signed in each other's presence. It includes a condition stating that one of the two witnesses was ineligible to serve under Tennessee law, and the law specifying witness eligibility in Tennessee. The given condition and the law invalidate the will statement, rendering it as *refute*. The model correctly predicted the outcome based on tokens such as *ineligible*, *witnesses*, *One*, *testament*, and *2*. The top two tokens with greatest impact on the model's prediction were *ineligible* and *witnesses*, with importance score of 0.60 and 0.14, respectively. It is

[10] A reviewer suggested an experiment where laws and conditions are packed into premises and treat the task as normal NLI. A pilot experiment with the best performing model (roberta-large-mnli) found a slight deterioration in performance when laws and conditions were combined as premises (precision: 93.09, recall: 93.89, F1: 93.41, accuracy: 94.12, GA: 74.51; see Table 2 for comparison with main result). The biggest difference came from GA score (74.51 vs. 84.31). This lowered GA score implies that the model's understanding worsened when conditions and laws were combined into premises. Given this preliminary result, it is likely that three-way distinction in input types positively contributes to the model's performance. Nevertheless, as the performance difference is not significantly large, further investigation is needed.

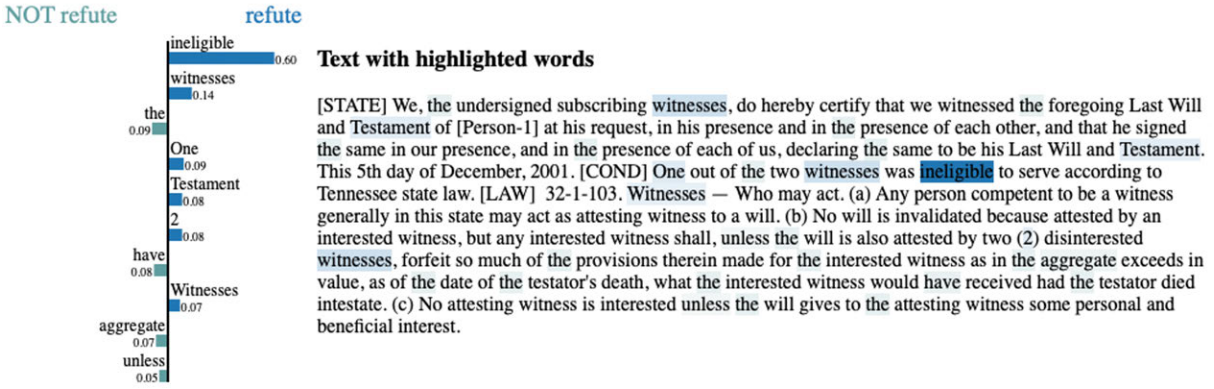


Figure 3: Examples of LIME explanations showing a case where the model (roberta-large-mnli) makes correct predictions based on features sensible to humans.

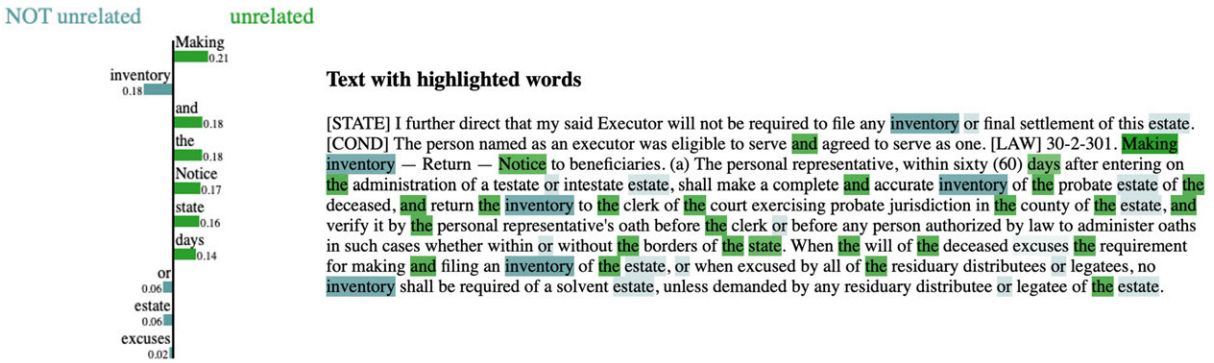


Figure 4: An example of a LIME explanation showing a case where the model (roberta-large-mnli) makes an incorrect prediction based on the tokens such as *Making*, *and*, and *the* which bear little semantic relevancy to the gist of the text.

understandable that these tokens have high scores, as they are the keywords ("**One** out of two (**2**) **witnesses** is **ineligible**") which provide grounds for revoking a will statement.

4.5.2 An Incorrect Example

Figure 4 shows a case where the model made an incorrect prediction based on irrelevant tokens. The intent of the will statement is to excuse the Executor from filing inventory. This is not contradicted by the condition and law, requiring the model to classify the text as *support*. Nevertheless, the model incorrectly classified it as *unrelated*. The reason for the incorrect prediction is that the model (probably) relied on irrelevant tokens such as *Making*, *and*, and *the*, which bear little semantic relevancy to the text, versus more relevant tokens such as *inventories* and *excuses* when making the prediction. This is likely due to overfitting induced by longer texts. This LIME explanation shows that even the best performing model still has room for improvement.

5 Future Work

This work can be expanded in several directions. First, it can be extended to cover multiple states by adapting the models or adding more data. Future work can investigate if models trained on a single state's data can be adapted to evaluate data from other states. Also, wills from multiple states could be added to the dataset. We expect the augmented dataset would enhance the models' performance on evaluating wills from other states.

This work can be expanded to other legal domains. Models trained on our data can be adapted to similar tasks in other legal domains.

Lastly, this work can be extended by investigating novel transformer models. Given the uniqueness of our dataset with regard to the number of inputs and text length, it is likely that further experiment and modification is needed to handle such characteristics.

6 Conclusion

This work presented an annotated dataset for natural language inference (NLI) in the legal domain, consisting of 1,014 data points generated from 23 publicly available wills. The dataset is novel for two reasons: it included texts with three input types (statement, law, and condition) rather than two (premise and hypothesis) in the traditional NLI, and included texts are longer than in general NLI datasets. The NLI models trained on our dataset showed reasonable performance in assessing the validity of will statements. Ablative experiments demonstrated that the models' performances worsen if any input type (condition, law, or both) is omitted. This suggests that the models utilize all three input types. The LIME implementation reveals that even the best performing model makes errors in some cases by using semantically irrelevant tokens. Our open-access dataset is publicly available at: https://github.com/ml4ai/nli4wills-corpus

Limitations

Our dataset consists of a relatively small number of data points (1,014 texts). Annotating will statements with relevant laws and conditions is a highly demanding, time intensive task. A larger size of dataset would likely further improve the models' performances.

Our dataset only includes wills executed and probated in Tennessee, with execution after 1970 and probate after 2000. Due to these restrictions, our framework might produce incorrect results on inputs from wills from different settings. Supplementing the dataset with wills from more diverse settings would address this limitation. Even though the scope is limited to a single state (Tennessee), our study demonstrates that transformer models trained on the dataset can evaluate the validity of statements from wills with reasonable accuracy. Future additions to our dataset will be available at the same URL: https://github.com/ml4ai/nli4wills-corpus

This work does not involve humans in the loop. Considering how crucial accuracy is for the task (i.e., legal validity evaluation), the work would have benefited much from involving domain experts. Even though our study discovered that the state-of-the-art transformer models can show good performances (over 85% accuracy in all 8 models) without human interaction, it was also found that the models' understanding on the task is rather superficial (GAs ranging from 49-85% in Table 2). Including humans-in-the-loop could be a solution for enhancing the models' understanding on the task.

Ethics Statement

We collected legal wills from Ancestry as a part of our dataset creation process. The wills probated in court are in the public domain in the US, and we did not violate Ancestry's Terms and Conditions when collecting wills. We also anonymized the wills by replacing any personal identifiable information contained in the documents with special tokens.

Our datasets and codes are released to public. We believe our released datasets and codes would contribute to society by promoting further NLI endeavors in the legal domain. The resources could potentially assist with people reviewing wills, but they should not be considered as legal advice. To avoid any confusion, we placed a disclaimer that the users must not rely on any information provided from our resources when making legal decisions and should instead consult with an attorney.

Acknowledgements

We thank the reviewers for their thoughtful comments and suggestions. This work was partially supported by the National Science Foundation (NSF) under grant #2217215, and by University of Arizona's Provost Investment Fund. Mihai Surdeanu and Clayton Morrison declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages

177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *CoRR*, abs/2005.05257.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *CoRR*, abs/2110.01799.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. *CoRR*, abs/2104.09420.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3413–3418, Hong Kong, China. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Ziyue Wang, Baoxin Wang, Xingyi Duan, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2019. IFlyLegal: A Chinese legal system for consultation, law searching, and document analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 97–102, Hong Kong, China. Association for Computational Linguistics.

Moshe Wasserblat. 2021. Sentence transformer fine-tuning (setfit): Outperforming gpt-3 on few-shot text-classification while being 1600 times smaller.