W5: Who did What to Whom, Where and When

Graduate

Timo Klingler

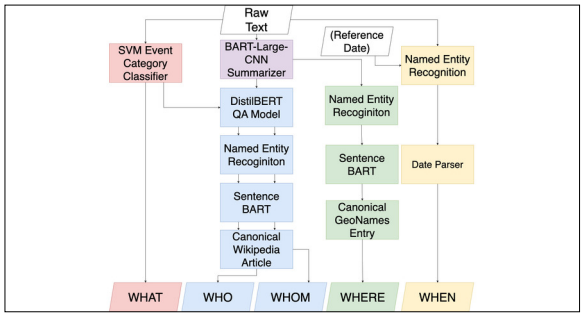
Davide Ferrara

Introduction: Vast quantities of unstructured text data are created daily through news articles and social media posts. To support decision-makers, the goal is to extract specific entities from this text to construct structured event data, answering the 'W5' questions: Who, What, Whom, Where and When. Traditional methods for creating these event data rely on manual processes or automated systems that parse the texts' syntactic structures. These approaches struggle with complex sentences, are language-specific and depend on dictionaries for ambiguity resolution, which are costly to maintain. This thesis presents a web-based application designed to convert unstructured textual data, such as news articles, into structured event records using the "W5" questions. The system leverages recent advancements in political event data extraction, particularly the shift from traditional Natural Language Processing (NLP) techniques to Large Language Models (LLMs), to identify and categorize event details with improved accuracy and flexibility.

Approach: Our approach processes input text through a multi-stage pipeline inspired by the Next Generation Political Event Coder (NGEC), a state-of-the-art framework developed by the political event data community for event extraction. The pipeline begins by assigning one of 16 distinct event types or categories (e.g., cooperation) to the input text through binary classification of each category independently, followed by selecting the one with the highest overall probability. The classified text is summarized by a Bidirectional and Auto-regressive Transformers (BART) model and then fed into a Bidirectional Encoder Representation from Transformers (BERT) question-answering (QA) model, which extracts answers from predefined questions (e.g., "Who is the actor?" "What is the action?"), tailored to the identified category. Named Entity Recognition (NER) is then applied on the answers of the QA model to extract "Who" and "Whom," as well as to identify temporal ("When") and geographical ("Where") information directly from the input text. Identified actors and recipients are linked to their canonical Wikipedia entries using pre-computed Sentence-BERT (SBERT) embeddings. Dates are parsed and evaluated relative to a given reference date. Lastly, extracted place names are matched to canonical entries in the GeoNames gazetteer using SBERT embeddings, allowing us to retrieve the corresponding geographic coordinates.

Result: We evaluated our system using two key resources: the Global Database of Events, Language, and Tone (GDELT), an open-source repository of global news events, and the Local-Global Lexicon (LGL), which maps place names to precise geographic coordinates. The evaluation was conducted under the constraint that the pipeline must run on limited consumer hardware—a laptop with 16GB of RAM, processing each text sample in approximately 30 to 40 seconds. Our implementation achieved the following F1-scores for the W5 pipeline components: What: 70.42%, Who: 0.06%, Whom: 0.05%, and Where: 0.17%. While the performance for event categorization (What) was strong, the results for entity and location resolution were significantly weaker. Nevertheless, the system constitutes a complete end-to-end solution, and its modular architecture provides a suitable foundation for future improvements.

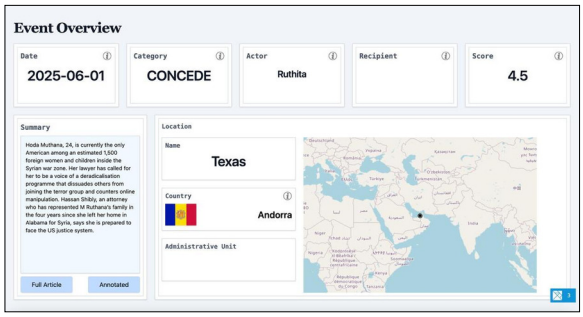
The W5 Pipeline: End-to-End Model Integration in the AI Workflow
Own presentment



Benchmarking Event Classification: Average Model Scores
Own presentment

Average Scores for Event Classification					
Algorithms	Test N	Accuracy	Precision	Recall	F1
Support Vector Machine	**905'807**	**0.7118**	**0.7247**	**0.6852**	**0.7042**
Bernoulli Naive Bayes	905'807	0.6704	0.6818	0.6424	0.6608
Decision Tree	905'807	0.6667	0.7605	0.4958	0.5942
Logistic Regression	905'807	0.7211	0.7620	0.6450	0.6981
Multinomial Naive Bayes	905'807	0.6802	0.6925	0.6599	0.6731
Random Forest	905'807	0.6507	0.6740	0.5946	0.6278

W5 Frontend View: Structured Results from News Analysis
Own presentment



Advisors
Prof. Dr. Mitra Purandare, Benjamin Plattner

Co-Examiner
Saskia Senn, Mettler Toledo GmbH

Subject Area
Artificial Intelligence