ClaimCompare:
A Data Pipeline for Evaluation of Novelty Destroying Patent Pairs

Arav Parikh, **Shiri Dori-Hacohen**

PatentSemTech Workshop @ SIGIR
July 2024

UCONN
REDUCING INFORMATION ECOSYSTEM THREATS (RIET) LAB
RUTGERS

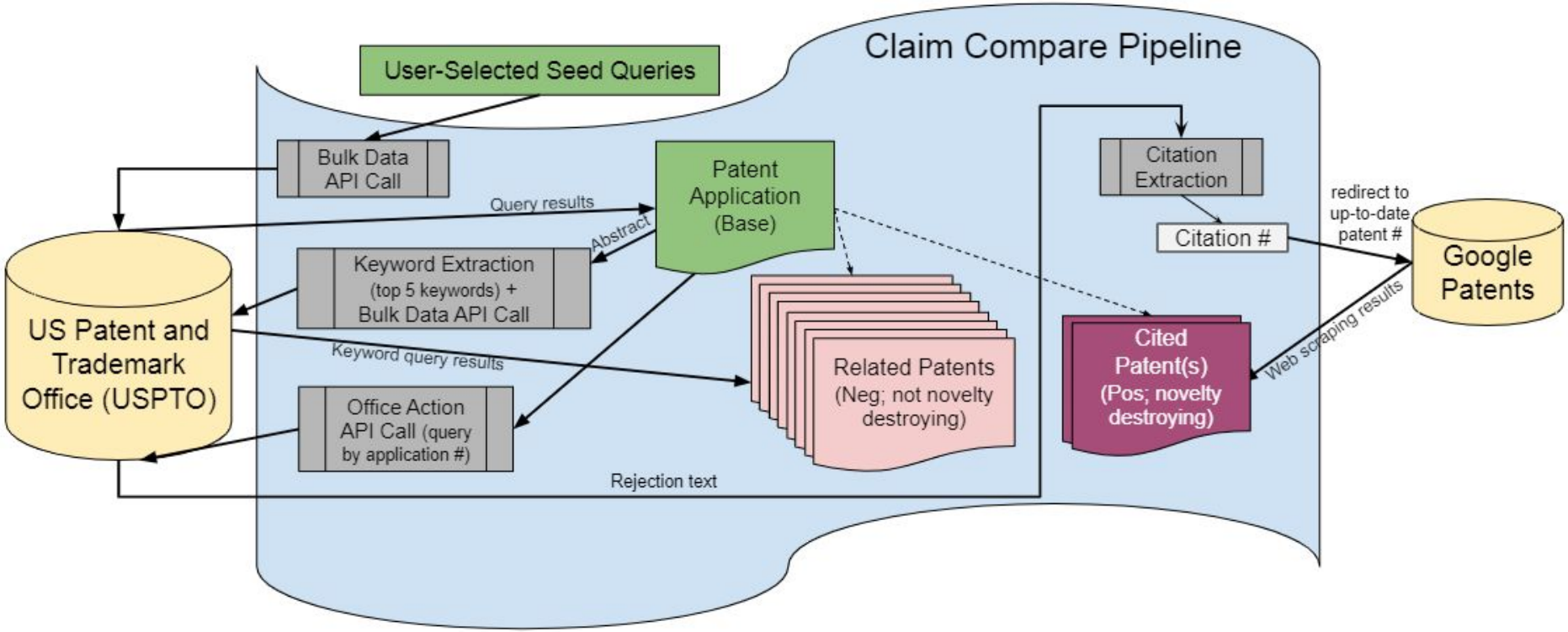
Background & Motivation

- Searching for novelty destroying documents is a critical yet **time and labor-intensive** part of the patent application process
- Automated approaches to this challenge are **lacking**
- Access to quality patent data needs to be **improved** in order to train viable AI/ML/IR models for this task

Contributions

1. The *ClaimCompare* pipeline, focused on **novelty destruction**, based on USPTO APIs and Google Patents, which can be used to generate **any number of datasets**
2. A sample dataset of 27K patents using *ClaimCompare*
3. Preliminary experimentation with LLMs and sample dataset, showcasing improvement over two baselines

The ClaimCompare Pipeline



Sample ClaimCompare Dataset

Dataset	Size	Data Source	Positive Samples	Negative Samples	Matching Strategy	Balanced?
PatentMatch v2	25K	EPO	Search report "X" citations	Search report "A" citations	Specific excerpts/lines	Yes
CC Sample	27K	USPTO	Office action 102 rejections	Similar keyword patents	Entire claim sets	No

Table 1 PatentMatch vs. a sample ClaimCompare (CC) dataset. Note that the ClaimCompare pipeline can be used as-is in order to generate many other datasets, including significantly larger ones.

Experimental Setup

- Dataset: 1K+ base patents in the electrochemical domain each matched with 25 related patents split into pairs (labeled with respect to novelty destruction, 0/1)
- Tasks: Pairwise classification (base vs. related) for novelty destruction & ranking
- Baseline: DistilRoBERTa (general) & BERT for Patents (enhanced, domain-specific) in zero-shot setting
- Experiments: DistilRoBERTa fine-tuned for 3 epochs on modified training datasets; tested downsampling of negative examples (k) to reduce class imbalance

Results

Model	AUROC	AP	MRR	P@1
General Baseline	0.473	0.350	0.697	0.651
Domain Baseline	0.589	0.464	0.703	0.651
Fine-Tuned (k=25)	0.999	0.999	0.989	0.978
Fine-Tuned (k=10)	0.999	0.998	0.987	0.975
Fine-Tuned (k=5)	0.982	0.975	0.967	0.974

Analysis & Future Work

- Dramatic improvement from the baseline in all cases
- Surprisingly robust generalization to unseen data
- No clear signs of overfitting or data leakage
- High results on test indicates our negative samples may be too “easy” to differentiate

Future Work

- Collecting harder, more semantically similar negatives
- Better leverage inter-patent relationships
- More complex querying to improve data quality

UCONN
REDUCING INFORMATION ECOSYSTEM THREATS (RIET) LAB

ClaimCompare:
A Data Pipeline for Evaluation of Novelty Destroying Patent Pairs

Thank you!
Questions?

@ShirKi
https://linkedin.com/in/shiri
shiridh@uconn.edu

Paper

Data & Code