Motivation

• In scholarly works, the past contribution of fellow scientists is acknowledged through formal citations.

• Automated Citation string parser is a crucial part for most patent database websites such as US Patent Office (USPTO) and Patent Lens.

• Programmatically recovering links between referring and referred-to documents requires a machine to understand the structure of the strings in a citation.

• Each citation string can be viewed as a set of fields (e.g., author, title, year, journal, …).
A Real World Example


Melanie Mitchell "An Introduction to Genetic Algorithms" (C) 1996 Massachusetts Institute of Technology (MIT), chapter II, pp.
Problem Statement

• Citation parsing problem, a sequence labelling problem which is common to large scale of language-based tasks such as POS-tagging.
• A reference string should first be broken down into a sequence of tokens
• Each token is to be assigned the correct label from a set of classes, which are the metadata fields such as author, journal, year, title, etc.
Common Approach: HMMs

- Hidden Markov Models (HMMs) is a well-known and powerful probabilistic tool for segmenting sequential data.

- HMMs has two major drawbacks:
  1. For the sake of complexity reduction, strong independence assumptions between the observation variables are made in HMMs.
  2. HMMs assigning a joint probability which is difficult to model when using rich features.
Our Approach: CRFs

- We are going to use Conditional Random Fields (CRFs) to build probabilistic models to segment and label our sequence data.
- CRFs satisfy both crucial characteristics of sequential data:
  1. Statistical dependencies between input entities.
  2. Using a rich set of features of entities to aid classification.
My Tasks

• Implement a baseline CRF-based system to segment citations.

• Applying feature engineering techniques to improve the performance of CRF for citation parsing problem.

• Aim of project is to achieve the performance over 95%.

• Compare the performance measures of my system with current system for the same data.
System Architecture

Diagram:
- TSV file
- Citation Extractor
- Tokenizer & Features Encoder
- CRF++ Toolkit
- Annotate Manually
- Template file
- Train file
- Test file
- Model
- CRF output
- Segmented Labelled Citations
- Performance Calculator
- Over 95%
Thank you!

Questions?