Profile of an Information Extraction Company

(or *How I Spent My Autumn Vacation*)

Presentation By: Steve Engels

Date: Jan. 18, 2001

- Information extraction overview
- Process involved in information extraction
- Technologies used
- Products & Areas of Research

Information Extraction Overview

- Example: CS Weekly Announcements

```
From uthoene@pythagoras.math.uwaterloo.ca Wed Jan 17 13:21:43 2001
Date: 12 Jan 2001 20:58:33 GMT
From: U.Thoene <uthoene@pythagoras.math.uwaterloo.ca>
Newsgroups: uw.cs.grad
Subject: CS Weekly Announcements  15-19 Jan 2001   #3

                COMPUTER SCIENCE WEEKLY ANNOUNCEMENTS

Announcements for the week of January 15 - 19, 2001               #3
---------------------------------------------------------------------

Monday, January 15, 2001
             DACA
             1:30 p.m.; DC3540

...

Wednesday, January 24, 2001
             The infraNET Project Smart Colmmunity Seminar Series
             "Adapting the Law for E-Commerce"
             C. Ian Kyer, Fasken Martineau DuMoulin LLP
             no charge, preregister: x4004 or http://infranet.uwaterloo.ca/
             2:30 - 4:00 p.m.; DC1302

.DE
```

Overview (cont'd)

- Want to find information about presentations
  - name of presenter
  - date & time
  - location
  - subject description
- Human readers skim, look for information patterns, find data for each category
- Information extraction performs the same task; difficulty is in specifying the method to recognize important data

Overview (cont'd)

- Why do information extraction?
  - Can automatically compile databases of information on a specified topic
  - Addresses the need for locating particular concepts within an growing pool of textual data
  - More reliable than simple keyword searching for when conducting a focussed search for information in a document

The IE Process

- The information extraction process involves:
  – Collecting “seed” web sites for search
  – Running a spider that recursively explores links found from seed web sites
  – Filtering a subset of relevant pages from the set of all retrieved web pages
  – Applying learned or rote extraction techniques to extract tuples of desired information

The IE Process (cont'd)

- "Seed" Web Sites
  - Mostly hand-collected
  - Grows incrementally after initial spidering runs reveal links to other relevant sites
- Web Spider
  - explores all links, filtering out links to CGI scripts, pictures, previously visited links, etc.
  - more advanced versions can fill out forms, invoke CGI scripts, access ftp and gopher sites, etc.
  - some spider applications stop after reaching a certain site (e.g. Continuing Education -- described later)

The IE Process (cont'd)

- Filtering Pages
  - filter out empty pages, non-relevant web sites, pages that match a certain criterion (e.g. no ads)
  - within a page, filter out non-relevant sections based on keywords or filtering cues
- Applying Extraction Patterns
  - Rote: hand-crafted rules/patterns, based on observations or established conventions (e.g. Dewey decimal system)
  - Learned: Markov models, decision trees, statistical parsing; training & test data

Technologies Used

- Rote Techniques
  - Feature detection
  - Regular expressions
  - Parsing
- Learned Techniques
  - Data collection
  - HMMs & MEMMs
  - Decision trees
  - Training enhancements

Rote Techniques

- Feature detection
  - Example: Name Extraction
    - “ **Jane Prime** *will be away this week* ”
  - Feature detection looks for various cues that indicate the likelihood of a word or phrase containing relevant data
  - Examples: Capitalized words, preceding/following words, etc.
  - Features can be made more elaborate, to include sentence fragments, HTML formatting and other features

Rote Techniques (cont'd)

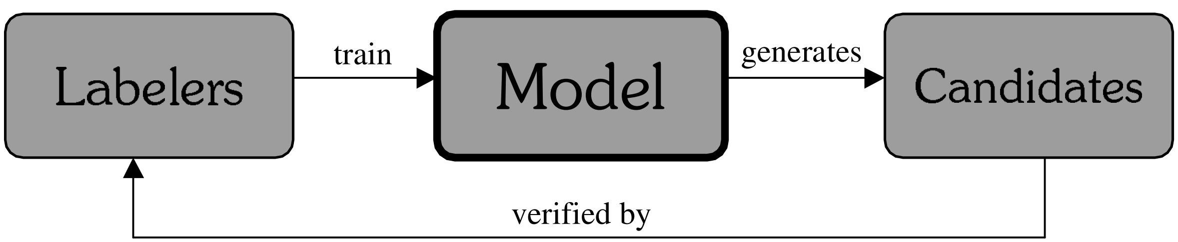
- Regular Expressions
  - Example: Name Extraction (again)
    - *" Prof.* **Bob Smith** *would bob for apples with the local smith "*
  - Uses prior knowledge, like rote lists of first & last names obtained through U.S. census reports
  - Regular expressions are created using some features and some syntactic knowledge to distinguish names (Bob Smith) from regular English words (bob, smith)
  - Not very reliable, usually used only to generate very basic candidates for creating training data

Rote Techniques (cont'd)

- Parsing (3 applications)
  - Use established English parser (e.g. Eugene Charniak's statistical parser) to identify named fields, actions, relations, etc. from grammatical data
  - Given entities, use parser to identify relations between entities
  - Given training data, train parser to include a new parse category for the entity to extract, then run parser on target data

Learning Techniques

- Data collection
  - Hand-tagging: Human operators manually search for desired entities in data
  - Feature extraction: Identify context and formatting cues in data (e.g. “Mr.”, “<TITLE>…</TITLE>”)
  - Candidate extraction: Use parsing, regular expressions, formatting cues to identify potentially relevant data, to be used or verifieded by data labelers



Learning Techniques (cont’d)

- Hidden Markov Model (HMM)
  - number & nature of states were set manually (e.g. name, pre-name, post-name, other)
  - unreliable sometimes, due to sparseness of training data
- Maximum Entropy Markov Models (MEMM)
  - maximum entropy bases its training on features of elements of the data (e.g. capitalization, word tense), instead of just the elements of the data alone (e.g. words) --> reduces sparseness problem
  - transition and emission weights are modified until ratio of features found in training set match those found in test set
  - results in more general model

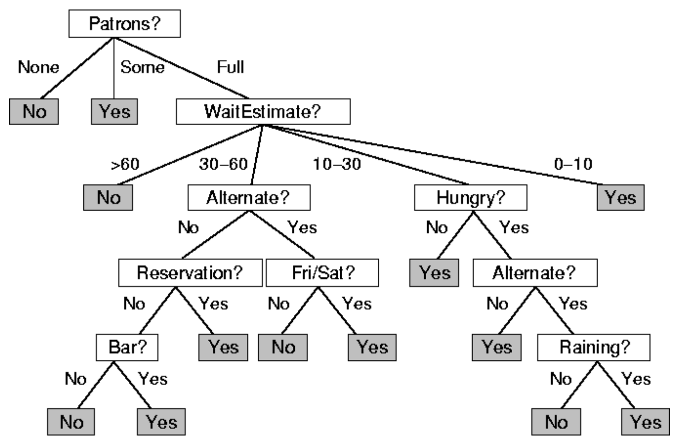
If you can read this, you’re sitting too close

Learning Techniques (cont'd)

- Uses of HMMs/MEMMs
  - candidate/entity identification (names, companies, etc.)
  - sentence segmenting
  - within an entity, identification of different component (title, first name, last name, etc.)

Learning Techniques (cont'd)

- Decision Trees
  - Instead of mapping feature sequences to Markov chains (MEMMs), decision trees evaluate a candidate's features one at a time until relative certainty about its category is achieved



  - Can be reliable, but very expensive to store and time-consuming to calculate

Learning Techniques (cont’d)

- Enhancements
  - Boosting
    - Using a weighted combination of weak hypotheses to produce a stronger combined hypothesis
    - When training the model, give higher weights to cases that were labeled incorrectly on the previous run
  - Co-training
    - When there is a natural split between the features of a document, training on each set of features individually and then combining the results has been shown to improve classification results
    - Example: Page classification
      - analyze words on a web page, hyperlinks to that page, combine results

Products & Areas of Research

- Job Databases
  - Finds all jobs on web
    - Company
    - Position
    - Salary
    - Etc.
  - First commercial venture
  - Spun off FlipDog.com as portal to job search database

PLEASE don’t ask me why it’s called “FlipDog”.

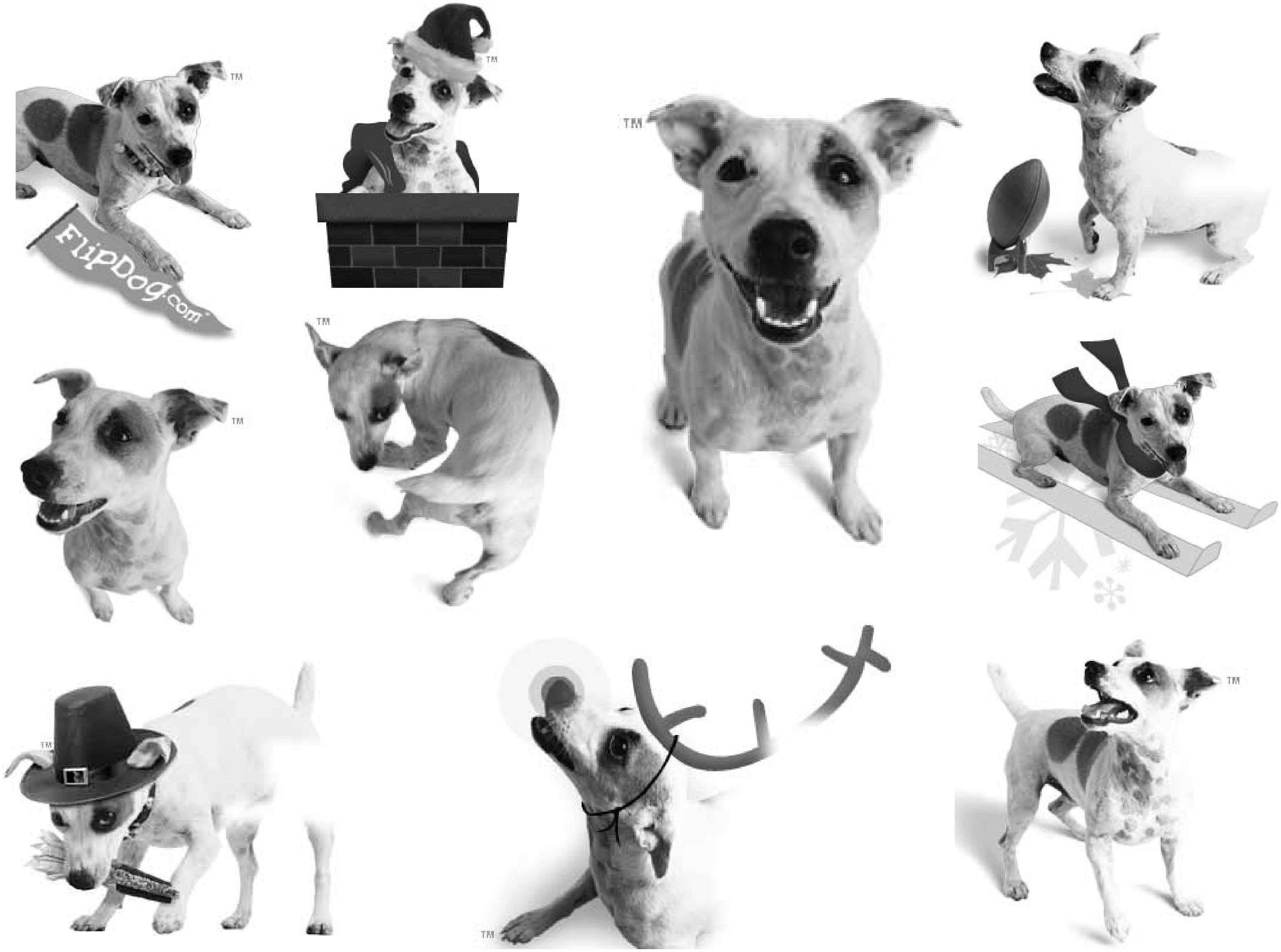
Products & Areas of Research (cont'd)

- Corporate Info
  - Use info extraction to gather data about companies on the web
    - corporate structure
    - contact information
    - affiliations/mergers/acquisitions
- Education
  - Compile list of continuing education courses offered by various institutions
  - Displays institution, course, time, date, instructor & description

All educators should be in an institution ;)

Products & Areas of Research (cont'd)

- News Clippings
  - Creates summaries of business or political events
  - Specifically, news about mergers, acquisitions, etc. most useful
- "diff" on WWW
  - Check pages of competitors or other important sites
  - Tool spiders every day, reports nature of all changes made to a site


FlipDog.com