**David Copperfield** or *The Personal History, Adventures, Experience and Observation of David Copperfield the Younger of Blunderstone Rookery (which he never meant to be published on any account)* is a novel by Charles Dickens, first published in 1850.

**David Copperfield** (born **David Seth Kotkin**) is a multi Emmy Award winning, American magician and illusionist best known for his combination of illusions and storytelling. His most famous illusions include making the Statue of Liberty "disappear"; "flying"; "levitating" over the Grand Canyon; and "walking through" the Great Wall of China.
TWO TYPES OF ENTITY DISAMBIGUATION

- CROSS-DOCUMENT COREFERENCE
  - Extension of INTRA-DOCUMENT COREFERENCE
  - Cluster entity descriptions

- WEB ENTITY
  - One-entity-per-document assumption
  - Cluster documents
OUR APPROACH TO ENTITY DISAMBIGUATION

- Clustering of ENTITY DESCRIPTIONS containing
  - Distributional information
  - Information extracted through relation extraction techniques

- Building on the results of INTRA-DOCUMENT COREFERENCE (IDC)
On Friday, Datuk Daim added spice to an otherwise unremarkable address on Malaysia's proposed budget for 1990 by ordering the Kuala Lumpur Stock Exchange "to take appropriate action immediately" to cut its links with the Stock Exchange of Singapore.
[Petrie Stores Corporation, Secaucus, NJ,] said an uncertain economy and faltering sales probably will result in a second quarter loss and perhaps a deficit for the first six months of fiscal 1994.

[The women’s appareil specialty retailer] said sales at stores open more than one year, a key barometer of a retain concern strength, declined 2.5% in May, June and the first week of July.

[The company] operates 1714 stores.

In the first six months of fiscal 1993, [the company] had net income of $1.5 million ....
Petrie Stores Corporation, Secaucus, NJ, said an uncertain economy and faltering sales probably will result in a second quarter loss and perhaps a deficit for [the first six months of fiscal 1994].

The women’s apparel specialty retailer said sales at stores open more than one year, a key barometer of a retail concern strength, declined 2.5% in May, June and the first week of July.

The company operates 1714 stores.

In [the first six months of fiscal 1993], the company had net income of $1.5 million ....
[The FCC] took [three specific actions] regarding [AT&T]. By a 4-0 vote, it allowed AT&T to continue offering special discount packages to big customers, called Tariff 12, rejecting appeals by AT&T competitors that the discounts were illegal. …..

…..

[The agency] said that because MCI's offer had expired AT&T couldn't continue to offer its discount plan.
Why Wikipedia may help addressing the encyclopedic knowledge problem

http://en.wikipedia.org/wiki/FCC:

The overall picture

Documents (news/blogs/web)

INTRA-DOC COREF (IDC)

DOC_1

DOC_n

ENTITY DISAMB

LEXICAL & ENCYCLOPEDIC KNOWLEDGE

Web Wikipedia WordNet

Entry-423742:
Names: George Bush, G W Bush, ...
Descriptors: the president, US President, ...
Facts/Roles: lives-at(Washington), Place-of-birth(Maine), ...
Local Entity Network: Washington, Cheney, Putin, ...
Word context (collocations, word-vector, ...)
WHAT WE DID IN THE SUMMER

- Developed new corpora for evaluating both CDC and IDC
  - New ACE CDC
  - New ARRAU IDC
CDC annotation of ACE 2005

- Callisto / EDNA annotation tool
- ACE 05 CDC:
  - 257K
  - 18K entities
  - 55K mentions
ARRAU IDC CORPUS

- Includes texts from several genres
  - Penn Treebank II
  - Other text
  - Spoken dialogue
- All mentions
- A variety of features (agreement, semantic type)
- Bridging, discourse deixis, ambiguity
WHAT WE DID IN THE SUMMER

- Developed new corpora for evaluating both CDC and IDC
- Developed a variety of Web people and CDC systems evaluated
  - using Spock for Web People
  - ACE CDC05 for CDC
  - Including three different relation extraction systems
Entity disambiguation: Clustering methods

- Greedy agglomerative
- Metropolis-Hastings
- Gibbs sampling
Entity disambiguation: Features

- Basic features:
  - bags of words
  - nominals

- Topic models

- Relations
  - supervised
  - unsupervised
Relation extraction

- **ACE:**
  - Supervised: Su & Yong
  - Supervised: Giuliano

- **Spock:**
  - Unsupervised: Mann
Summary

- Web people: achieved improvements both through the improved clustering methods & the additional features
- CDC: very high baseline, but obtained improvements nevertheless
- Relation extraction: significant difference with / without IDC
WHAT WE DID IN THE SUMMER

- Developed resources for evaluating both CDC and IDC
- Developed a variety of Web people and CDC systems

**IDC:**
- Developed a platform for exploring IDC methods
- Implemented a variety of techniques for extracting knowledge from the Web and Wikipedia
- Tested better ML methods
AN ARCHITECTURE FOR IDC
(Working name: ELKFED / BART)

Can handle
- Different preprocessing methods (e.g. chunkers vs parsers)
- Different methods for generating training instances
- Different decoding methods
- Different types of output (including MUC, APF)

- Easy to customize
- Support for error analysis through MMAX2
Using lexical & encyclopedic knowledge

- Tested around 20 features
- Developed both
  - New methods for extracting knowledge
  - New methods for using this knowledge
- Improved mention detection crucial
Extracting lexical and commonsense knowledge

- From WordNet
  - A variety of similarity measures (Ponzetto & Strube, 2006)

- From the Web
  - Hyponymy (Markert & Nissim, 2005; Versley, 2007)

- From Wikipedia
  - From the categories (Ponzetto & Strube, 2006)
  - From a Wikipedia-extracted taxonomy
  - From the relatedness links
Using lexical & encyclopedic knowledge

- To detect SIMILARITY
  - GOP – the Republican Party

- To detect INCOMPATIBILITY
  - the first six months of fiscal 1994
  - the first six months of fiscal 1993
ML models

- Support Vector Machines
  - To detect ‘structured’ similarity / dissimilarity
- ‘Split’ models
  - Pronouns / definite descriptions
- Ranked models
- Global models
Results: quantitative
(ACE02 bnews)
Summary of contributions

Conclusions

- Developed two new corpora for evaluating CDC and IDC
- Demonstrated that improvements in Web People can be obtained using
  - Topic models
  - Metropolis-Hastings, Gibbs sampling
- Developed a new platform for IDC
- Achieved improvements in IDC using
  - Lexical and Encyclopedic knowledge
  - Support vector machines
  - (And the contributions are additive)
Program

- Resources & evaluation
- ED: web people, CDC, and relation extraction
- IDC development tool
- Extraction of lexical and commonsense knowledge
- SVMs
Conclusions
- Developed two new corpora for evaluating CDC and IDC
- Demonstrated that improvements in Web People can be obtained using
  - Topic models
  - Metropolis-Hastings, Gibbs sampling
- Developed a new platform for IDC
- Achieved improvements in IDC using
  - Lexical and Encyclopedic knowledge
  - Support vector machines
  - (And the contributions are additive)
Members of the team

- Senior staff on site
  - Artstein Day Duncan Hitzeman Mann Moschitti Poesio Strube Su Yang

- PhDs
  - Hall Ponzetto Smith Versley Wick

- Undergrads
  - Eidelman Jern

- Externals
  - Giuliano Hoste Jemison Pradhan Yong
  - Daelemans Hinrichs
Thanks

- The sponsors
- EML Research (MMAX2, the initial code for BART)
- UMass Amherst (Rob Hall)
- MITRE (David Day, Janet Hitzeman)
- I^2R
- EPSRC Project ARRAU (corpus, Gideon Mann)
- DoD (Jason Duncan)
- JHU Center of Excellence (Paul McNamee)
Immediate future: some ideas

- Delivering the corpora (LDC) and BART (SourceForge)
- More experiments with the new corpora (ARRAU, OntoNotes)
- Improved mention detectors
- ‘Backoff’ model of Wiki / Web / WN knowledge use
- Global models
- Semantic trees & incompatibility
  - With global models / with SVMs
- Relation extraction with ‘real’ IDC