Roger Bock
11/19/04

Assignment 4
SIMS 290-2: Applied Natural Language Processing

Overview

I decided to write a program that would take the Enron Email corpus and automatically generate an acronym dictionary for it. I used the Schwartz and Hearst algorithm and code, but modified it to tailor it to this specific corpus and to improve the precision and recall.

Method

In this section I will describe each of the modifications I made to the original algorithm, and I will provide examples of the resulting improvements when possible. The first change focused on improving the recall, and the rest of the changes were focused on improving precision.

1. Allow paragraphs to begin with a lowercase character
The first thing I noticed was that the algorithm was ignoring lines that were at the start of a new paragraph and had a lowercase character as their initial character. This restriction makes sense for working with the MEDLINE abstracts corpus, but for informal email communication it does not make sense. Removing this restriction increased the candidate pool of <"short form", "long form"> pairs from 167 pairs to 196 pairs.

2. Ignore email headers
The email headers containing information about the sender and the recipients were producing spurious definition candidates. For example, "yerskerbal@shb.com; Fritz Kolb (E-mail);" is clearly not an acronym definition. Ignoring the text in these headers removed five incorrect pairs from the dictionary.

3. Strip out junk characters
Various emails contained characters that seemed to be automatically introduced to signify when a line of text originally wrapped onto the next line. To account for this I removed the following strings from each line of text: "=20", "= ", and "=". I also removed quotation marks and replaced multiple spaces with a single space. This affected 47 entries in the dictionary and led to some entries being combined.

4. Prefer initial character matches to internal character matches
The paper by Schwartz and Hearst says "The algorithm is based on the observation that it is very rare for the first character of the short form to match an internal letter of the long form." I found that this situation actually occurred with some frequency in the Enron

Email corpus. To fix this I changed the algorithm so that whenever a short form character matched a character internal to a long form word, it checked the first character of that word to see if it could match that instead. To use an example from the paper, <_TTF-1, Thyroid transcription factor 1_> would now match as <_TTF-1, **T**hyroid **t**ranscription **f**actor **1**_> instead of <_TTF-1, Thyroid **t**ranscrip**t**ion **f**actor **1**_>. The following table shows the dictionary entries that were improved by this change.

Short Form	**Original Long Form**	**Improved Long Form**
AABE	Association of Blacks in Energy	American Association of Blacks in Energy
EEA	Espionage Act	Economic Espionage Act
EEF	Exchange Fellows	Eisenhower Exchange Fellows
EEF	Exchange Fellowship	Eisenhower Exchange Fellowship
EES	Energy Services	Enron Energy Services

5. Discard long forms that contain the short form

Parenthetical statements accounted for a number of false positives. For example, the line "SAVA Model Comparison (comparison of curves to model generated results)" could lead to the interpretation that the parenthetical statement is the long form definition for the short form "Comparison". These errors were fixed by discarding pairs where the long form contained the short form when the short form had at least four characters. The threshold was chosen because bigrams and trigrams of characters could reasonably be expected to be both the entire short form and part of the long form (<_SO, Southern Company_>, for example). This change eliminated eleven false positives from the dictionary.

6. Merge similar long forms

Some of the long forms matching a single short form were essentially the same concept but with slight morphological variation. These variations were due to the addition of suffixes to indicate pluralization or possession. To fix this I merged long forms whose edit distances were within (short form length - 1) of each other. The length-dependent threshold allowed greater variation for longer definitions. To calculate the edit distance I used an implementation I found on the web at http://www.merriampark.com/ld.htm.

The following table shows the dictionary entries that were improved by this change.

Short Form	**Original Long Forms**	**Merged Long Form**
ISO	Independent System Operator Independent System Operators	Independent System Operator
MSEB	Maharashtra State Electricity Board Maharashtra's State Electricity Board	Maharashtra State Electricity Board
REI	Reliant Energy Reliant Energy's	Reliant Energy

Results

Here I show the output of my system. The format is (short form) (long form) (frequency in corpus).

```
AABE       American Association of Blacks in Energy                    1
ABA        American Bioenergy Association                              7
AD's       Airworthiness Directives                                    1
AESP       Association of Energy Services Professionals International  2
AFC        Application for Certification                               7
AGA        American Gas Association                                    2
ALS        amyotrophic lateral sclerosis                               1
ANOPR      Advanced Notice of Proposed Rulemaking                      1
ANZ        and New Zealand                                             1
AOL        America Online                                              1
AOL        attribute it to technology                                  1
AP         Associated Press Writer WASHINGTON                          1
ATS        automated trading system                                    1
AWK        American Water Works                                        1
AYE        Allegheny Energy Inc.'s                                     1
B.ELE      Belgian utility Electrabel                                  1
BGS        Beta Gamma Sigma                                            1
BRKA       Berkshire-Hathaway                                          6
Body       being modelled on the Forestry Stewardship Council          1
CABC       Competitive Analysis and Business Controls                  1
CAEM       Center for the Advancement of Energy Markets                2
CAISO      California Independent System Operator                      2
CARB       California Air Resources Board                              2
CARE       Clean Air Responsibility Enterprise                         2
CCS        current cost of supplies                                    2
CEC        California Energy Commission                                7
CEM        California Energy Market                                    2
CERA       Cambridge Energy Research Associates                        1
CFMA       Commodities Futures Modernization Act                       1
CFTC       Commodities Futures Trading Commission                      1
CI         competitive intelligence                                    1
CIS        Cargill Investor Services                                   1
CIS        customer information system                                 1
CPN        http://www.calpine.com                                      12
CPUC       California Public Utilities Commission                      2
CRP        Conservation Reserve Program                                7
Cal-ISO    California Independent System Operator                      2
CalPERS    California Public Employees Retirement System               2
CalPX      California Power Exchange                                   1
Cut        cut reports and some photographs                            1
D-NM       Dan Albert in Senator Bingaman's office                     1
DENA       Duke Energy North America                                   4
DOE        Department of Energy                                        7
DPC        Dabhol Power Co                                             1
DPC        Dabhol Power Company                                        3
DSM        demand side management                                      2
DUK        Duke Energy                                                 2
DYN        Dynegy                                                      1
DYN        Dynegy Inc's                                                1
E.FEN      Electrica Fenosa SA                                         1
EBR        Executive Business Review                                   1
EBS        Enron Broadband Services                                    5
ECAR       East Central Area Reliability region                        1
ECS        Enron Center South                                          1
EEA        Economic Espionage Act                                      1
EEF        Eisenhower Exchange Fellows                                 2
EEF        Eisenhower Exchange Fellowship                              1
EES        Enron Energy Services                                       1
EFET       European Federation of Energy Traders                       2
EFS        Enron Facility Services                                     2
EGA        Enron Government Affairs                                    1
ENE        energy company Enron                                        1
ENE        energy major Enron Corp.'s                                  1
```

EON	E.On AG	3
EPA	Environmental Protection Agency	3
ERCOT	Electric Reliability Council of Texas	2
ESC	Energy Services Coalition	2
Elyse	Elizabeth Labanowski never returned my calls or messages	1
F.BNP	French banks BNP Paribas	1
FAP	Field Application Program	1
FCPA	Foreign Corrupt Practices Act	1
FERC	Federal Energy Regulatory Commission	26
FPA	Federal Power Act	6
GAO	General Accounting Office	5
GATS	General Agreement on Trade and Services	1
GSS	Global Strategic Sourcing	5
HPL	Houston Pipe Line Company	1
Hair?	have learned from watching Herr	2
IAM	Integrated Asset Management	1
IOGCC	Interstate Oil and Gas Compact Commission	2
IOS	Inventory of Skills	1
IPCC	Intergovernmental Panel for Climate Change	7
IPO	initial public offering	1
IPPs	independent power producers	5
ISAC	information sharing and analysis center	1
ISDA	International Swaps and Derivatives Association	1
ISO	Independent System Operator	20
ISO	it was riddled with errors	1
ISOs	Independent System Operators	1
KPTCL	Karnataka State Power Transmission Corp Ltd	1
Ken Peel	Ken_Peel@hagel.senate.gov	2
LDCs	Local distribution companies	2
LIHEAP	Low-Income Home Energy Assistance Program	1
LME	London Metal Exchange	1
MAA	Market Assessment Advisor	2
MAIN	Mid-America Interconnected Network	1
MIR	Mirant	1
MMP	mitigated market price	10
MOEA	Ministry of Economic Affairs	2
MPA	Market Participant Advisor	2
MPC	Mangalore Power Co	1
MSC	Market Surveillance Committee	1
MSEB	Maharashtra State Electricity Board	5
MW	megawatts	3
MWh	megawatt hours	1
NBMBAA	National Black MBA Association	1
NERC	North American Electric Reliability Council	2
NNG	Northern Natural Gas	2
OEC	Operational Energy Corporation	2
OOM	out-of-market	2
OPIC	Overseas Private Investment Corporation	2
OSP	optical-switched service provider	1
OSX	Oil Stock Index	2
PBR	performance-based rate	1
PJM	Pennsylvania-New Jersey-Maryland	3
PNW	Pacific Northwest Electric	1
POWER	Program on Workable Energy Regulation	2
PUC	Public Utilities Commission	5
PWC	PriceWaterhouseCoopers	1
PWPENS	POWER Working Paper Email Notification Service	1
PX	Power Exchange	3
RBI	Reserve Bank of India	1
RED Index	Retail Energy Deregulation Index	2
REI	Reliant Energy	7
REI	http://www.reliantenergy.com	7
RMR	reliability must-run	7
RTO	Regional Transmission Organization	1
SAFE	Securing Amerca's Future Energy	1
SCE	Southern California Edison	2
SEC	Securities and Exchange Commission	2
SERC	Southeastern Electric Reliability Council	1
SIEPR	Stanford Institute for Economic Policy Research	3
SO	Southern Company	1
SOL	shit out of luck	2

SPP	it's position as a bargaining ploy?	1
SRE	Sempra Energy	12
SRP	Sierra Pacific Resources	3
TEP	temporary extraordinary procedures	6
TM	Times; Source: World Reporter	1
TNRCC	Texas Natural Resources Conservation Commission	2
TNT	Turner Network	1
TURN	The Utility Reform Network	1
UDCs	utility distribution companies	1
UFE	Unaccounted For Energy	2
active	An OTC approved program that will provide eligible employees	1
bcf	billion cubic feet	2
drapes	drappies	2
e.g.	ed agreements	1
e.g.	entire coordinated legal strategy	1
grapes	grappies	2
i.e.	IOUs' core gas load	1
i.e.	Issues Developers of distributed	1
i.e.	invoked in CA	1
i.e.	issue	1
iii	ir contractual obligations	5
mw	megawatts	7
to date	to provide an HP contact on connectivity	3
we	where	1