bananaNER

https://github.com/keighrim/bananaNER

Keigh Rim, Todd Curcuru, Yalin Liu

2/6/2015

Overview

- Classifier:

  CRF++

- Features:

  66 feature functions

  - including 13 Brown, 12 CTF, 13 CDF

  4 name dictionaries

  - geo, org, person, other

- Combination for features:

  feature templates

Methodology

Take text

Extract features

Combine features using template

Train using CRF++

Test and hope for the best

Classifier: CRF++

1. Combine variable features
2. Avoid label bias problem
3. Faster than Mallet

Feature functions

Various local features

Some global features

Functions either:

- Binary (T/F)
  - Initial capital, first word, includes digit
- Categorical
  - Brown cluster, zone, term freq

Used Chieu and Ng (2003) as a guide

Combination

Template

- unigram

  generate a set of features L*N
- bigram

  current and previous output token, features: L*L*N
- features' optimization is important in CRF++

Results on Dev data

Vary by machine!

Different features good/bad

K:

P: 77.6

R: 71.3

F: 74.3

T:

P: 77.7

R: 71.1

F: 74.3

Y:

P: 77.7

R: 71.0

F: 74.2

Useful Features

- Word Token
- POS tag
- Zone
- Bias
- First Word
- Initial Capitalization
- One Capital
- All Capitals
- Contains Digit
- 2 Digits
- 4 Digits
- Symbols ($, /, etc.)
- > Average Length
- Brown Clusters
- Name Dictionaries
- Sequence of Capitals

Useless Features

Hyphen

Zone (On its own)

CTF and CDF

Questionable

- Document Freq, Term Freq

Feature implementations

- Frequency counts
  - Document frequency (helped)
  - Term frequency (not helping)
  - Cluster ID freq counts (not helping)

Feature implementations

- Zones mentioned in XML
  - Not in our text
  - Made heuristics
- Name dictionaries
  - Names scraped from online sources
  - Messy/Not uniform
  - Implemented using N-gram matching

Building Dictionaries

Person names (PER)

- http://www2.census.gov/topics/genealogy/1990surnames/dist.all.last
- http://www.ssa.gov/oact/babynames/limits.html
- http://deron.meranda.us/data/census-derived-all-first.txt

Building Dictionaries

Organization names (ORG)

- Companies
  - http://www.sec.gov/rules/other/4-460list.htm
  - http://www.wordlab.com/archives/company-names-list/
  - http://www.nasdaq.com/screening/company-list.aspx
  - http://en.wikipedia.org/wiki/List_of_company_name_etymologies

Building Dictionaries

Organization names (ORG)

NGO/NPO

- https://www.charitywatch.org/charities
- http://earth-info.nga.mil/gns/html/gazetteers2.html
- http://topnonprofits.com/lists/best-nonprofits-on-the-web/
- http://charity.lovetoknow.com/List_of_Nonprofit_Organizations
- http://en.wikipedia.org/wiki/Category:Non-profit_organizations_based_in_the_United_States

Building Dictionaries

- Organization names (ORG)
  - News company
    - http://en.wikipedia.org/wiki/Lists_of_newspapers
    - http://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_States
    - http://en.wikipedia.org/wiki/List_of_newspapers_in_the_world_by_circulation
  - Militant Organizations
    - http://www.nctc.gov/site/other/fto.html
    - http://en.wikipedia.org/wiki/List_of_designated_terrorist_organization

Building Dictionaries

Geo-entities (GPE, LOC)

- http://www.geonames.org/
- http://www.unece.org/cefact/locode/welcome.html
- http://jordonmeyer.com/text-list-of-us-cities/
- http://jordonmeyer.com/50-states-list-text/
- https://gist.github.com/marijn/396531
- http://www.textfixer.com/resources/dropdowns/country-list-array.txt
- http://weather.rap.ucar.edu/surface/stations.txt

Building Dictionaries

- Others (FAC, VEH)
  - Extracted from train data

Using entities in the data

1) Using Only Collected Dictionaries
2) Combine with entities from training set
3) Combine with entities from train+dev+test

	P	R	F
1)	.77	.71	.74
2)	.79	.57	.66
3)	.87	.85	.86

Useful Combinations

Zone + First Word + Initial Capital

Zone + Initial Capital

Word Token + Initial Capital

Various N-grams of the same feature

Bad Combinations

First Word, POS Tag

First Word, Next Word Initial Capital

First Word, Next Word POS Tag

Anything with Hyphen

Useful N-grams

Good

- Trigram Init Caps
- Bigram Tokens
- Trigram POS
- Bigram POS

Bad

- Bigram Init Caps
- Trigram Tokens

Useful Windows

Unigram Word Tokens: 3

Unigram Tokens and Initial Caps: 3

Bigram Word Tokens: 3

Unigram POS Tag: 5

Bigram POS Tag: 3

Trigram POS Tag: 5

Final results

- On Test data
  - P: 75.8
  - R: 72.4
  - F: 74.1
- On Dev data (from prev slide)
  - P: 77.6
  - R: 71.3
  - F: 74.3