Automatic Extraction of Geographic and Chronological References

Ying Wang   Xiangrui Meng

Institute for Computational and Mathematical Engineering
Stanford University

CADS 2009

Outline

Task Description by Example

Stephen C. Massett gave a concert of vocal music in the schoolhouse that stood at the northwest corner of the plaza. This was on Monday evening, June 22, 1849; and it was the first public entertainment ever given in San Francisco.

— from “California as I Saw It” (abridged)

Using Gazetteers

Gazetteers provide us a very simple way to identify location names from text through string matching.

Pros:

- Super fast and easy to implement

Cons:

- It may recognize other types of entities as locations, e.g., person and organization.
- The accuracy also depends on the gazetteers we use.

Improve the Accuracy

Using spatial prepositions as indicators

If we find an instance in a gazetteer following a spatial preposition such as “in” and “at”, we should mark the instance as a location name with confidence.

> Stephen C. Massett gave a concert of vocal music in the schoolhouse that stood at the northwest corner of the plaza. This was on Monday evening, June 22, 1849; and it was the first public entertainment ever given in San Francisco.

Counterexamples

Location names appear without spatial prepositions:

Victoria is the capital of British Columbia.

Which word indicates Victoria is a location, Victoria or capital?

Victoria was the queen of the United Kingdom.

Think the Opposite

Can we identify geographical entities without gazetteers?

What if we mask the proper nouns and numbers?

> Xxxxxxx X. Xxxxxxx gave a concert of vocal music in the schoolhouse that stood at the northwest corner of the plaza. This was on Monday evening, June ##, ####; and it was the first public entertainment ever given in Xxx Xxxxxxxxx.

— from "California as I Saw It" (abridged)

You can still mark the entity types easily.

Think the Opposite

Can we identify geographical entities without words?

What if we only show you the word classes?

(Personal pronoun) (Verb, past tense) (Determiner) (Adjective) (Adjective) (Noun, singular common) (Adverb) (Verb, past participle) (Preposition) (Noun, singular proper).

Certainly you cannot recover the original sentence:

It was the first public entertainment ever given in San Francisco.

But you can still guess that the last word is a location name.

The Semantic Approach

To take the semantic approach, we need:

- a set of features, e.g., word classes or shapes (Xxx, ####)
- a tool that converts words in a text to a feature sequence
- a set of rules that could identify geographic entities from a feature sequence

Part-of-Speech Tagging

POS tagging is to assign a part-of-speech or other lexical class marker to each word in a text.

It/PRP was/VBD the/DT first/JJ public/JJ entertainment/NN ever/RB given/VBN in/IN San/NNP Francisco./NNP

— annotated by Stanford POS Tagger

- PRP: Personal pronoun
- VB?: Verb
- NN: Noun
- IN: Preposition
- NNP: Proper noun

Making Rules

Is it easy to make rules that can identify geographic entities?

DT NN VBD NNP IN NNP NNP IN DT CD.

A steamer left Calcutta for Hong Kong on the 25th.

DT NN VBD NNP IN NNP NNP IN DT CD.

The pianist performed Waldstein at Carnegie Hall on the 21st.

DT NN VBD NNP IN NNP NNP IN DT CD.

The company released iPhone for 3G Network on the 11th.

Working rules are much more sophisticated!

The Semantic Approach: Pros and Cons

Pros:

- High accuracy with well crafted grammar rules

Cons:

- Need experienced linguists and months of work
- Language dependent

Automatically Learn Rules from Tagged Data

Observation

If we have a large amount of text with entity types tagged, computer can easily measure the accuracy of grammar rules.

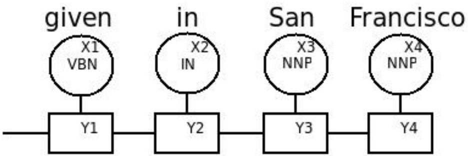
Question

Is it possible to let computer learn rules automatically?

Answer

Yes, if we replace the set of explicit rules by a probabilistic model (implicit rules).

Conditional Random Field Model



$(X1, X2, X3, X4)$ is the part-of-the-speech sequence.

$(Y1, Y2, Y3, Y4)$ is the type sequence to be determined.

We first notice that $Y1 \neq$ LOC since $X2 =$ VBN. Mathematically, it is described by conditional probability: $\text{Prob}(Y1 = \text{LOC} \mid X1 = \text{VBN}) = 0$. Then we may ask:

$$\text{Prob}(Y3 = \text{LOC},\ Y4 = \text{LOC} \mid X1 = \text{VBN},\ X2 = \text{IN},\ X3 = \text{NNP},\ X4 = \text{NNP}) = ?$$

Finding the Best Tag Sequence

Let $\mathcal{X} = (X1, X2, \ldots)$ and $\mathcal{Y} = (Y1, Y2, \ldots)$. We want to find the optimal tag sequence $\mathcal{Y}^*$ that maximizes $P(\mathcal{Y}|\mathcal{X})$.

Under the CRF model, this function is a combination of some predefined functions, with some unknown parameters. We use tagged text to estimate those parameters, called training.

Training Data

The training data is from the Conference on Computational Natural Language Learning (CoNLL-2003). It is a collection of news wire articles, annotated by people.

```
U.N.        NNP   I-NP   I-ORG
official    NN    I-NP   O
Ekeus       NNP   I-NP   I-PER
heads       VBZ   I-VP   O
for         IN    I-PP   O
Baghdad     NNP   I-NP   I-LOC
.           .     O      O
```

Improve the Accuracy

The achieve high accuracy, the training data should be marked correctly and the type of the training data should be similar to our document type. We can do this in a stepwise manner:

$$\text{Training Data} \underset{\texttt{Revision}}{\overset{\texttt{NER Software}}{\rightleftharpoons}} \text{Tagged Text}$$

Additional Rules pt.1

We apply additional rules to make the disambiguation easier.

- Only output entities in our database.

... escaped Azkaban to seek revenge. $\Leftarrow$ Where is Azkaban?

- Suffix check

... turn left at Princeton street. $\Leftarrow$ not Princeton, NJ.

- Hierarchy check

... Orientalists call the Athens of India. $\Leftarrow$ not Athens, Greece.

Additional Rules pt.2

- If an geographical entity is also recognized as a person name for many times in a book, it is generally not a location name.

There they are, Sydney. Fire away! $\Leftarrow$ Sydney Carton.

Language Independence

- European Languages: Spanish, German, French, Italian ...
  The probabilistic model is almost language independent among European languages. We only need POS taggers for different languages and some training data to start.
- Chinese, Japanese and Korean (CJK)
  CJK NER is more difficult. These languages don't have capitalization. Moreover, the words are not separated by spaces in CJ. Word segmenters are needed for preprocessing.

Task Description by Example

Stephen C. Massett gave a concert of vocal music in the schoolhouse that stood at the northwest corner of the plaza. This was on Monday evening, June 22, 1849; and it was the first public entertainment ever given in San Francisco.

— from “California as I Saw It” (abridged)

You Find Something Similar Here

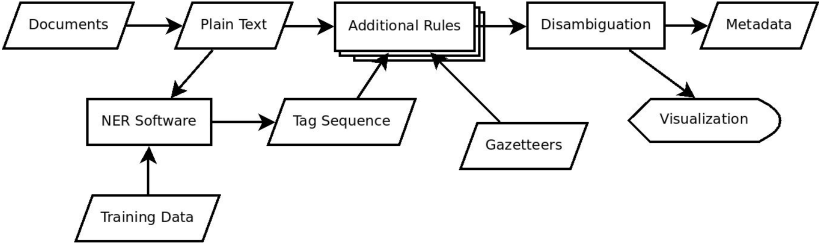
- Do string matching on “Monday”, ..., “Friday”’, and numbers between 1000 and 2050.
- Define a set of grammar rules.
- Use the probabilistic model.

Training Data

The training data is from the Message Understanding Conference Proceedings (MUC-7). It is a collection of newswire articles, annotated by people.

```
President <ENAMEX TYPE="PERSON">Jimmy Carter
</ENAMEX>, for instance, made consistent efforts
from <TIMEX TYPE="DATE">1977</TIMEX> on to
reduce tensions between the two countries.
```

Workflow


Documents
Plain Text
Additional Rules
Disambiguation
Metadata
NER Software
Tag Sequence
Gazetteers
Visualization
Training Data

Softwares

- The Stanford NER and POS Tagger, from the Stanford Natural Language Processing Group
- ANNIE, a GATE component with NER capabilities
- CRF++, C++ implementation of Conditional Random Fields model
- ...

Web Interface

Upload a text file:

Browse... | Generate Metadata

(TXT, TEI/SGML)

Metadata

Geographic references

Moscow, 5
Changsha, 4
Stalingrad, 4
Paris, 3
Berlin, 3

Chronological references

1939, 1944, 1945, 1943, 1941

Dataset

Collection	# Documents	Format
California as I Saw It	204	TEI/SGML
Winning the Vote for Women	153	TEI/SGML
Dance Instruction Manuals	175	TEI/SGML
A Century of Lawmaking for a New Nation	103	TEI/SGML
Pioneering the Upper Midwest	137	TEI/SGML
Early American Travel Narratives	282	TEI/SGML
Puerto Rico at the Dawn of the Modern Age	58	TEI/SGML
Chesapeake Bay Book Collection	141	TEI/SGML
The Foreign Affairs Oral History Collection	1303	TEI/SGML
Spalding Base Ball Guides	42	TEI/SGML
Newspapers (1918-1919)	418	PrimeOCR

How to Measure Accuracy

- Precision (P) measures the number of correct entities in the answer file over the total number of entities in the answer file.
- Recall (R) measures the number of correct entities in the answer file over the total number of entities in the key file.
- F-measure is the harmonic mean of precision and recall.

F-measure

$$F = \frac{RP}{R+P}$$

Precision Estimation pt.1

- We found 81132 location names, which contains 854 unique names, from the “California as I Saw It” collection.
- For each unique location name, we chose an instance at random and created a question on Amazon Mechanical Turk.

Sample Question

```
⋆ Entity: San Rafael
⋆ Context: I have just returned from a delightful
     drive to San Rafael and back.

 1. Is the entity a location name in the context?
    Yes, it is a location name!
    No, it is a person name.
    No, it is an organization name.
    No, it is a product name.
    No, it has other type.
```

Precision Estimation pt.2

- Each question is answered by three different person.
- For each entity, if more than one person says that it is not a location name, we treat it as a false negative.
- The precision is about 94.1%.

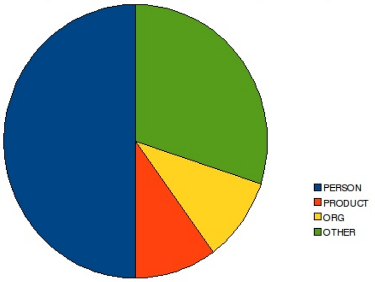


Figure: Distribution of false negatives

Precision Estimation pt.3

PERSON

Fremont was made Governor by Stockton at Los Angeles.

ORGANIZATION

... which was led by some of the graduates of Hampton or Carlisle,

PRODUCT

... minutes later she ran afoul of the big American ship Saint Paul,

OTHER

Texas was then the Mecca of adventurers and people who ...

Metadata of Proper Names and Named Events

Task

For proper names or named events, use external reference sources to generate appropriate geographic metadata.

Example

"Large Hadron Collider" $\Rightarrow$ Geneva, Switzerland, 2008

Wikipedia Extension

We first use Wikipedia to expand a proper name or a named event to an article. Then we can extract geographic and chronological entities from the Wikipedia text and use them as the metadata corresponding to the term given.

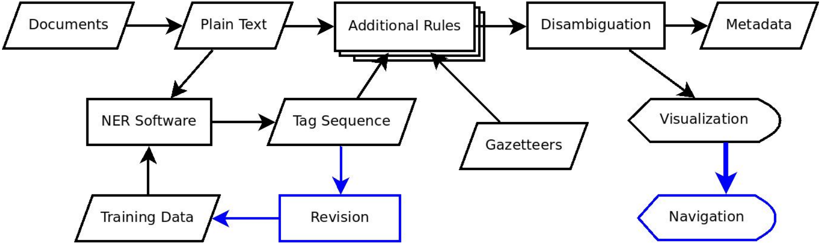
Large Hadron Collider — Wikipedia

The LHC ... lies underneath the Franco-Swiss border between the Jura Mountains and the Alps near Geneva, Switzerland ... were circulated through the collider on 10 September 2008.

Metadata of “Large Hadron Collider”

Locations	Years
Switzerland (4)	2008 (16)
Geneva (3)	2007 (6)
France (2)	2001 (4)
Vatican (1)	2005 (3)

Summary


Documents
Plain Text
Additional Rules
Disambiguation
Metadata
NER Software
Tag Sequence
Gazetteers
Visualization
Training Data
Revision
Navigation