Machine Learning for Information Extraction from XML marked-up text on the Semantic Web

Present: 吳佳厚

Outline

- Introduction & Background
- Corpora
- Named Entity Classes & Features
- Method
- Experiment & Results
- Analysis
  - Coordination, Apposition, Abbreviation
- Conclusion

Introduction

- How to find the information that meets users' requirements and present it in and understandable form.
- One scenario :

  A system will be an ability to learn to identify and classify terms based on examples of previously marked-up text.
- In two domains : news and molecular-biology

Background

- These systems, with previous IE (information extraction), can be classed as either predominantly dictionary-based or learning-based.
- But, the hand-built dictionary-based systems cannot be expected to be easily ported to new domains and they ignore a potentially valuable source of the domain expert's knowledge.

Background

- HMMs (Hidden Markov Models) are one of the most widely methods in ML for IE.
- It can be considered to be stochastic finite state machines and have enjoyed success in speech recognition and part-of-speech tagging.
- A system which uses HMMs is one of the most successful such sys. And trains on a corpus of marked-up text, using only character features in addition to word bigram.

Corpora

- In the experiments we used abstracts in the molecular biology domain available from PubMed's MEDLINE that were marked up in XML by a domain expert.
- As well as news texts used in the MUC-6 conference.

A graduate of <ENAMEX TYPE="ORGANIZATION">Harvard Law School</ENAMEX>, Ms. <ENAMEX TYPE="PERSON">Washington</ENAMEX> worked as a laywer for the corporate finance division of the <ENAMEX TYPE="ORGANIZATION">SEC</ENAMEX> in the late <TIMEX TYPE="DATE">1970s</TIMEX>. She has been a congressional staffer since <TIMEX TYPE= "DATE">1979</TIMEX>. Separately, <ENAMEX TYPE="PERSON">Clinton</ENAMEX> transition officials said that <ENAMEX TYPE="PERSON">Frank Newman</ENAMEX>, 50, vice chairman and chief financial officer of <ENAMEX TYPE="ORGANIZATION">BankAmerica Corp.</ENAMEX>, is expected to be nominated as assistant <ENAMEX TYPE="ORGANIZATION">Treasury</ENAMEX> secretary for domestic finance. Mr. <ENAMEX TYPE="PERSON">Newman</ENAMEX>, who would be giving up a job that pays <ENAMEX TYPE="MONEY">$1 million</ENAMEX> a year, would oversee the <ENAMEX TYPE="ORGANIZATION">Treasury</ENAMEX>'s auctions of government securities as well as banking issues. He would report directly to <ENAMEX TYPE="ORGANIZATION">Treasury</ENAMEX> Secretary-designate <ENAMEX TYPE="PERSON">Lloyd Bentsen</ENAMEX>. Mr. <ENAMEX TYPE="PERSON">Bentsen</ENAMEX>, who headed the <ENAMEX TYPE="ORGANIZATION">Senate Finance Committee</ENAMEX> for the past six years, also is expected to nominate <ENAMEX TYPE="PERSON"> Samuel Sessions</ENAMEX>, the committee's chief tax counsel, to one of the top tax jobs at <ENAMEX TYPE="ORGANIZATION">Treasury</ENAMEX>. As early as today, the <ENAMEX TYPE="PERSON">Clinton</ENAMEX> camp is expected to name five undersecretaries of state and several assistant secretaries.

Figure 1: Example sentences taken from the annotated MUC-6 NE corpus

TI - Activation of <PROTEIN> JAK kinases </PROTEIN> and <PROTEIN>STAT proteins </PROTEIN> by <PROTEIN> interleukin - 2 </PROTEIN> and <PROTEIN> interferon alpha </PROTEIN> , but not the <PROTEIN> T cell antigen receptor </PROTEIN> , in <SOURCE.ct> human T lymphocytes </SOURCE.ct> .
AB - The activation of <PROTEIN> Janus protein tyrosine kinases </PROTEIN> (<PROTEIN> JAKs </PROTEIN>) and <PROTEIN> signal transducer and activator of transcription </PROTEIN> (<PROTEIN> STAT </PROTEIN>) proteins by <PROTEIN> interleukin (IL) - 2 </PROTEIN> , the <PROTEIN> T cell antigen receptor </PROTEIN> (<PROTEIN> TCR </PROTEIN>) and <PROTEIN> interferon (IFN) alpha </PROTEIN> was explored in <SOURCE.ct> human peripheral blood - derived T cells </SOURCE.ct> and the <SOURCE.cl> leukemic T cell line Kit225 </SOURCE.cl> . An <PROTEIN>IL-2</PROTEIN>-induced increase in <PROTEIN>JAK1</PROTEIN> and <PROTEIN>JAK3</PROTEIN>, but not <PROTEIN>JAK2</PROTEIN> or <PROTEIN>Tyk2</PROTEIN>, tyrosine phosphorylation was observed. In contrast, no induction of tyrosine phosphorylation of <PROTEIN>JAKs</PROTEIN> was detected upon stimulation of the <PROTEIN>TCR</PROTEIN>. <PROTEIN>IFN alpha</PROTEIN> induced the tyrosine phosphorylation of <PROTEIN>JAK1</PROTEIN> and <PROTEIN>Tyk2</PROTEIN>, but not <PROTEIN>JAK2</PROTEIN> or <PROTEIN>JAK3</PROTEIN>. <PROTEIN>IFN alpha</PROTEIN> activated <PROTEIN>STAT1</PROTEIN>, <PROTEIN>STAT2</PROTEIN> and <PROTEIN>STAT3</PROTEIN> in <SOURCE.ct>T cells</SOURCE.ct>, but no detectable activation of these <PROTEIN>STATs</PROTEIN> was induced by <PROTEIN>IL-2</PROTEIN>.

Figure 2: Example MEDLINE sentence taken from the XML annotated molecular-biology NE corpus

Name Classes

Class	#	Example	Description
PROTEIN	2125	*JAK kinase*	proteins, protein groups, families, complexes and substructures.
DNA	358	*IL-2 promoter*	DNAs, DNA groups, regions and genes
RNA	30	*TAR*	RNAs, RNA groups, regions and genes
SOURCE.cl	93	*leukemic T cell line Kit225*	cell line
SOURCE.ct	417	*human T lymphocytes*	cell type
SOURCE.mo	21	*Schizosaccharomyces pombe*	mono-organism
SOURCE.mu	64	*mice*	multi-organism
SOURCE.vi	90	*HIV-1*	viruses
SOURCE.sl	77	*membrane*	sublocation
SOURCE.ti	37	*central nervous system*	tissue
UNK	-	*tyrosine phosphorylation*	background words

Table 1: Named entity classes for the molecular biology domain. # indicates the number of tagged terms in the corpus of 100 abstracts.

Class	#	Example	Description
ORGANISATION	1783	*Harvard Law School*	names of organisations
PERSON	838	*Washington*	names of people
LOCATION	390	*Houston*	names of places, countries etc.
DATE	542	*1970s*	date expressions
TIME	3	*midnight*	time expressions
MONEY	423	*$10 million*	money expressions
PERCENT	108	*2.5%*	percentage expressions
UNK	-	*start-up costs*	background words

Table 2: Named entity classes for the news domain. # indicates the number of tagged terms in the corpus of 100 abstracts.

Features

Table 3: Character features with examples. It should be noted that the examples do not show the full form of terms, but simply examples of 'words' that make up the term together with their semantic classification.

Feature code	Examples
TwoDigitNumber	$[25]_{percent}$
FourDigitNumber	$[2000]_{date}$
DigitNumber	$[2]_{percent}$ $[3]_{DNA}$
SingleCap	$[I]_{protein}$ $[B]_{protein}$ $[T]_{source.ct}$
GreekLetter	$[alpha]_{protein}$
CapsAndDigits	$[2A]_{DNA}$ $[BW5147]_{source.cl}$ $[CD4+]_{source.ct}$
TwoCaps	$[RelB]_{protein}$ $[TAR]_{RNA}$ $[HMG]_{DNA}$ $[NF]_{DNA}$ $[NF]_{protein}$
LettersAndDigits	$[p50]_{protein}$ $[Kit225]_{source.cl}$
InitCap	$[Interleukin]_{protein}$ $[Washington]_{person}$
LowCaps	$[kappaB]_{protein}$ $[mRNA]_{RNA}$
Lowercase	$[cytoplasmic]_{source.sl}$ $[tax]_{protein}$
Determiner	the
Conjunction	and
FullStop	.
Comma	,
Hyphen	$[-]_{protein}$ $[-]_{DNA}$
Colon	:
SemiColon	;
OpenParen	(
CloseParen	)
CloseSquare	]
OpenSquare	[
Percent	%
Other	*+#
Backslash	$[/]_{protein}$

Method

- We considered that a HMM approach based on raw text strings of words and no deep linguistic analysis is well suited.
- The task is to maximize Pr(C|W) where C is the name classes and W is a given sequence of words.
- Words are ordered pairs consisting of a surface word, W, and a word feature, F, given as <W,F>

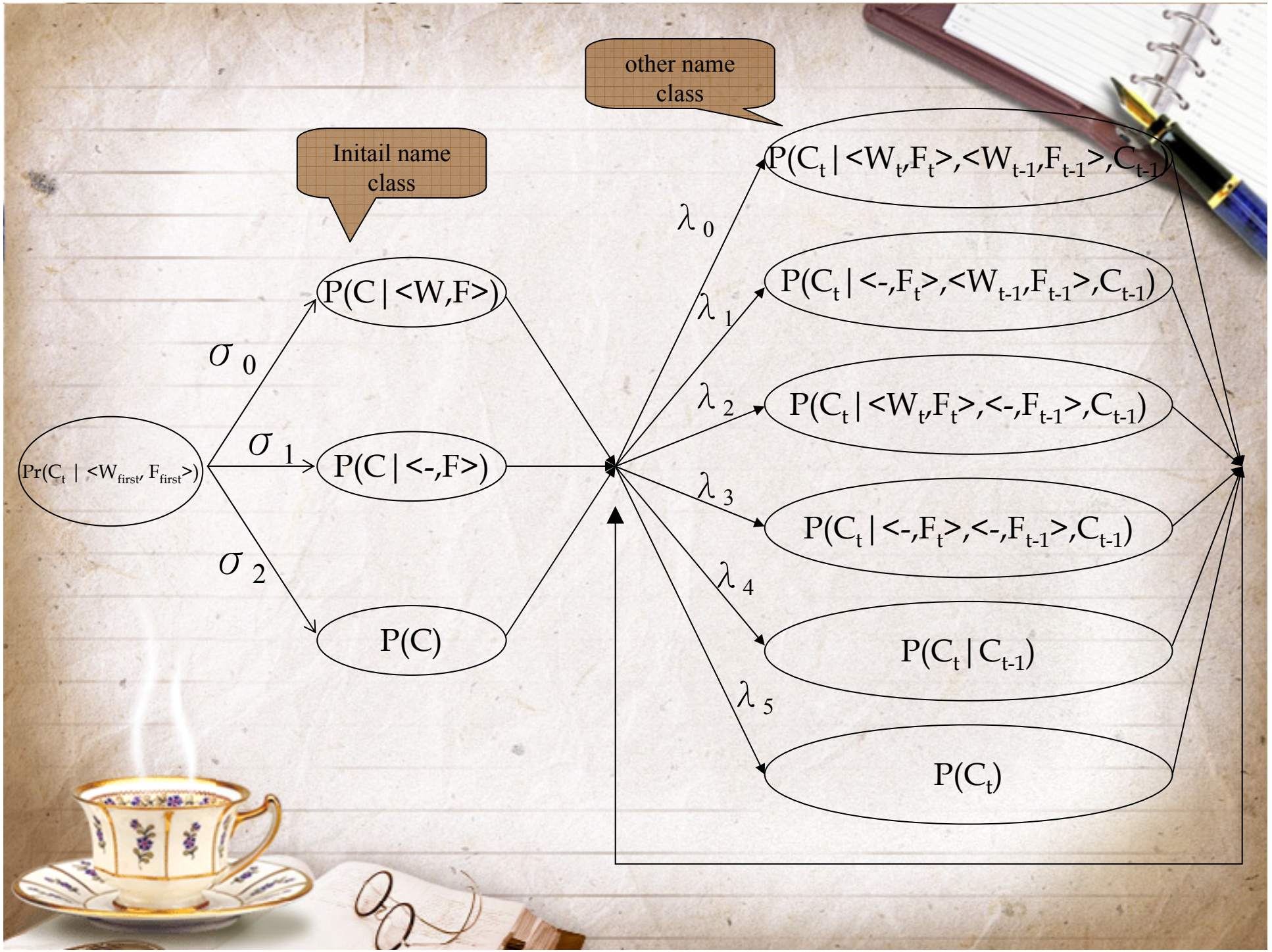
Use the following equation to calculate the initail name class probability

$$\Pr(C_t \mid \langle W_{first}, F_{first} \rangle) = \sigma_0 f(C_{first} \mid \langle W_{first}, F_{first} \rangle) + \sigma_1 f(C_{first} \mid \langle -, F_{first} \rangle) + \sigma_2 f(C_{first}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

For all other words and their name classes :

$$\Pr(C_t \mid \langle W_t, F_t \rangle, \langle W_{t-1}, F_{t-1} \rangle, C_{t-1}) = \lambda_0 f(C_t \mid \langle W_t, F_t \rangle, \langle W_{t-1}, F_{t-1} \rangle, C_{t-1}) + \lambda_1 f(C_t \mid \langle -, F_t \rangle, \langle W_{t-1}, F_{t-1} \rangle, C_{t-1}) + \lambda_2 f(C_t \mid \langle W_t, F_t \rangle, \langle -, F_{t-1} \rangle, C_{t-1}) + \lambda_3 f(C_t \mid \langle -, F_t \rangle, \langle -, F_{t-1} \rangle, C_{t-1}) + \lambda_4 f(C_t \mid C_{t-1}) + \lambda_5 f(C_t) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

λ 、 σ : Constants


other name class
Initail name class
$P(C_t | <W_t,F_t>,<W_{t-1},F_{t-1}>,C_{t-1})$
$P(C | <W,F>)$
$P(C_t | <-,F_t>,<W_{t-1},F_{t-1}>,C_{t-1})$
λ_0
λ_1
σ_0
λ_2
$P(C_t | <W_t,F_t>,<-,F_{t-1}>,C_{t-1})$
$Pr(C_t | <W_{first}, F_{first}>)$
σ_1
$P(C | <-,F>)$
λ_3
$P(C_t | <-,F_t>,<-,F_{t-1}>,C_{t-1})$
λ_4
σ_2
$P(C_t | C_{t-1})$
$P(C)$
λ_5
$P(C_t)$

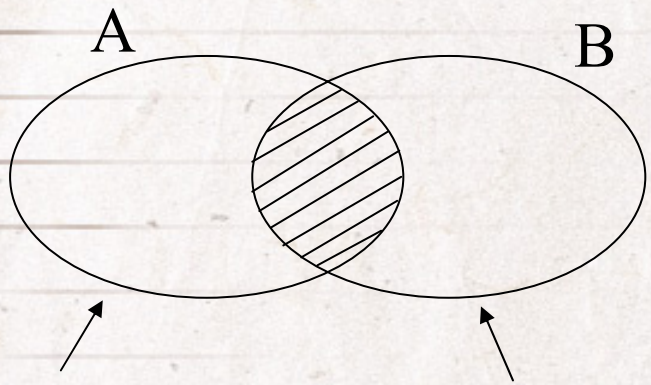
$$f(C_t \mid \langle W_t, F_t \rangle, \langle W_{t-1}, F_{t-1} \rangle, C_{t-1}) = \frac{T(\langle W_t, F_t \rangle, C_t, \langle W_{t-1}, F_{t-1} \rangle, C_{t-1})}{T(\langle W_t, F_t \rangle, \langle W_{t-1}, F_{t-1} \rangle, C_{t-1})}$$

$P(A|B) = P(AB)/P(B)$

Where T() has been found from counting the events in the traning corpus.

Experiment

- 60 news texts (50 training, 10 testing)
- 100 molecular biology (80 training, 20 testing)
- The results are given as F-scores, a common measurement for accuracy in the MUC conferences that is the harmonic mean of recall and precision.



標記的標準答案

程式產生

$$\text{Precision} = \frac{A \cap B}{B}$$

$$\text{Recall} = \frac{A \cap B}{A}$$

$$\text{F-score} = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

$$\Rightarrow 0 \le F - score \le 1$$

Analysis

- Broadly speaking the major ones can be divided into coordination, apposition and abbreviation …etc, there are many issues that the method cannot cover.
- Ex: “non-T-cells”

Coordination

Ex. 2. *...regulated by members of the [rel/NF-kappa B family]$_{protein}$*

Ex. 3. *...involves phosphorylation of several members of the [NF-kappa B]$_{protein}$/[I kappa B protein]$_{protein}$ families.*

Ex. 4. *...regulated by members of the rel/[NF-kappa B family]$_{protein}$*

Ex. 5. *...involves phosphorylation of several members of the [NF-kappa B/I kappa B] protein families.*

For example "< ENAMEX TYPE= "LOCATION"> North</ENAMEX> and <ENAMEX TYPE-="LOCATION">South America</ENAMEX>.

Apposition

Ex. 8. *However, similarly to the other* $[Rel]_{protein}$-$[NF\text{-}kappa\ B]_{protein}$ *complexes,* $[RelB]_{protein}$-$[p52]_{protein}$ *can up-regulate the synthesis of* $[I\ kappa\ B\ alpha]_{protein}$.

Ex. 9. *The transcription factor* $[NF\text{-}Kappa\ B]_{protein}$ *is stored in the* $[cytoplasm]_{source.sl}$...

Ex. 10. *However, similarly to the other* $[Rel]_{protein}$-$[NF\text{-}kappa\ B]_{protein}$ *complexes,* $[RelB\text{-}p52]_{protein}$ *can upregulate the synthesis of* $[I\ kappa\ B\ alpha]_{protein}$.

Ex. 11. *The transcription factor* $[NF\text{-}Kappa\ B]_{protein}$ *is stored in the* $[cytoplasm]_{source.sl}$...

Abbreviation

Ex. 12. *The [interleukin-2 (IL-2) promoter]$_{DNA}$ consists of several independent [T cell receptor (TcR) responsive elements]$_{DNA}$.*

Ex. 13. *The [interleukin-2]$_{protein}$ ([IL-2]$_{protein}$) promoter consists of several independent [T cell receptor]$_{protein}$ ([TcR]$_{protein}$) responsive elements.*

Conclusion

- In this paper, a scenario has been presented for the exploitation of terminology contained within XML marked-up documents that are produced by online domain-based communities for automatically annotateing untagged texts.
- A limitation of the HMM approach is that it cannot easily model large feature sets due to fragmentation of the probability distribution.

Future work

- Looking at combining orthographic knowledge with other types of lexcial info. as well as contecxtual clues from grammatical dependency analysis.