Information Extraction using Markov Models

Martin Labský
labsky@vse.cz
Dept. of Knowledge Engineering
VŠE Praha

1

Agenda

1. **The IE problem and its motivation**
   – introduction, applications [4]
2. **Methods used for IE**
   – symbolic vs. probabilistic, LP2 [6]
3. **Markov models for IE**
   – principles, variations [5]
4. **Markov models for Bike product IE**
   – training data, models, results [10]

2

I. The IE task

- **Identification of words or phrases of interest in texts**
- **Kind of semantic (pragmatic) annotation with:**
  – labels
  – database / ontology instances
- **Good for:**
  – database / knowledge base population
  – ... and then for querying over the texts

3

IE Applications

- **Local semantic search**
  – job offerings, advertisements
  – FAQ for judges (~10K questions & answers)
  – news servers, e.g. PlanetOnto:
    - ontology (people, projects, events etc. + inference rules)
    - automatic annotation of new texts (sent by email)
    - instances from text fed into ontology's knowledge base
    - structured querying using OCML or forms
- **Global semantic search**
  – Internet-scale annotation
  – annotation "bureaus"
    - structured querying
    - augmented search using the TAP ontology (→)

4

Augmented Google search

* R.V.Guha, Rob McCool: TAP (WWW Conference 2003)

5

IE types

Is the city **Washington** named after George Washington?

instance-of

Město

(named entity recognition, NER)

```
<p align="left" class="prodbold">MODEL NAME</p>
<p class="prodmain">TREK 850X</p>
<span class="grey">£249.99</span>
<a href="offers.php">ON OFFER FOR</a> £225.00
<span class=tiny><br>*€306 approx</span>
```

Produkt	
název	String
cena	Float

(template IE)

6

Agenda

1. **The IE problem and its motivation**
   – introduction, applications [4]
2. **Methods used for IE**
   – symbolic vs. probabilistic, LP2 [6]
3. **Markov models for IE**
   – principles, variations [5]
4. **Markov models for Bike product IE**
   – training data, models, results [10]

7

II. Methods used for IE

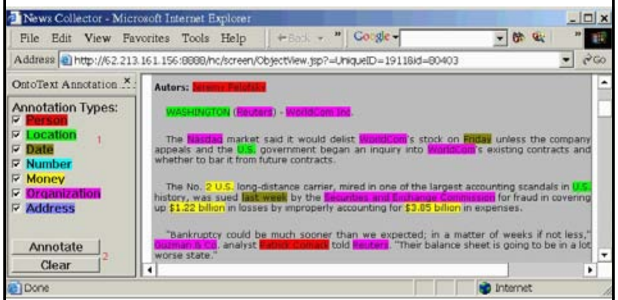
- **Symbolic**
  – assigning the most common sense (Washington=city)
  – induction of context-based rules (LP^2, Rapier, Stalker)
- **Probabilistic**
  – hidden markov models (HMMs)
  – maximum entropy models (MEMs)

8

KIM plugin

http://www.ontotext.com

- **Semantic annotation of named entities**



9

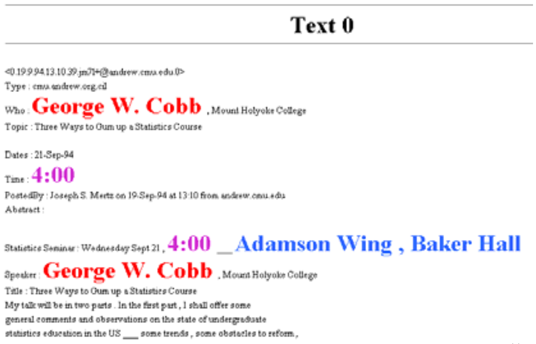
LP^2

- **Bottom-up induction of context-based rules by sequential covering of training examples**
  – positive examples = annotated instances in text
  – negative examples = the rest of the text
  – rules are generalized using lemmatization, upper/lower case letters, POS tags, and other categories (p.m. -> time etc.)
- **Types of the induced rules**
  – tagging (context trigger => "insert tag")
  – correction (context trigger => "move tag")
- **Sequential covering of positive examples**
  – positive examples covered by a newly induced rule are removed
  – induction continues untill all positive examples are covered

10

LP^2 – seminar announcements

- **250 annotated announcements (F. Ciravegna, SSSW 2003)**



Text 0

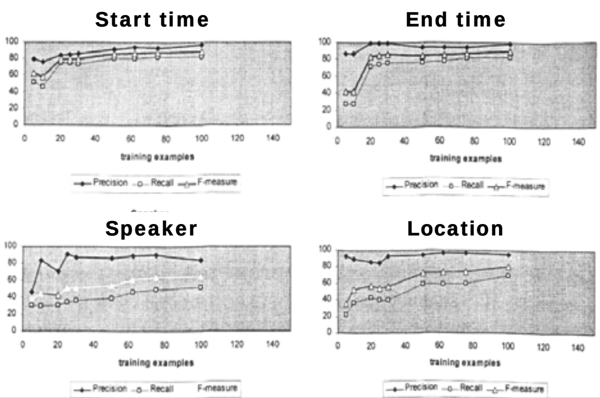
<0.19.9.94.13.10.39.jm7l+@andrew.cmu.edu.0>
Type : cmu.andrew.org.cl
Who : George W. Cobb , Mount Holyoke College
Topic : Three Ways to Gum up a Statistics Course

Dates : 21-Sep-94
Time : 4:00
PostedBy : Joseph S. Mertz on 19-Sep-94 at 13:10 from andrew.cmu.edu
Abstract :

Statistics Seminar : Wednesday Sept 21 , 4:00 __Adamson Wing , Baker Hall
Speaker : George W. Cobb , Mount Holyoke College
Title : Three Ways to Gum up a Statistics Course
My talk will be in two parts . In the first part , I shall offer some
general comments and observations on the state of undergraduate
statistics education in the US ___ some trends , some obstacles to reform ,

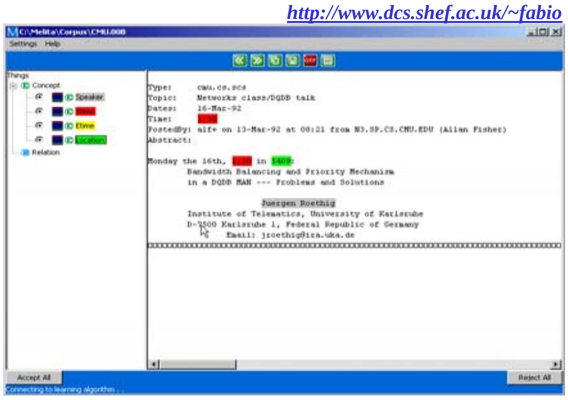
11

LP^2 – seminar results



Melita – interactive LP²

http://www.dcs.shef.ac.uk/~fabio



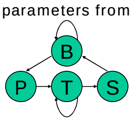
Agenda

1. **The IE problem and its motivation**
   – introduction, applications [4]
2. **Methods used for IE**
   – symbolic vs. probabilistic, LP2 [6]
3. **Markov models for IE**
   – principles, variations [5]
4. **Markov models for Bike product IE**
   – training data, models, results [10]

14

III. Using HMMs for IE

- **HMM - Probabilistic finite state machine**
  – generative model of text
  – model goes from state to state, each time generates 1 word
  – for each state we have:
    - **transition** distribution (probabilities of the next state)
    - **word generation** (emission) distribution (word probabilities)
  – probabilities of where to start
- **Efficient algorithms for determining:**
  – **P($w_1..w_n$|M)** = probability of text being generated by model M
  – **$S_1..S_n$** = the most probable state sequence generating that text
  – model parameters from training data



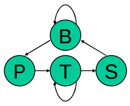
$w_1, w_2, w_3, w_4, w_5, w_6$

B B P T T S

15

Simple HMM structure for IE

- **4 state types:**
  – **B**ackground (generates words not of interest),
  – **T**arget (generates words to be extracted),
  – **P**refix (generates typical words preceding target)
  – **S**uffix (words typically following target)

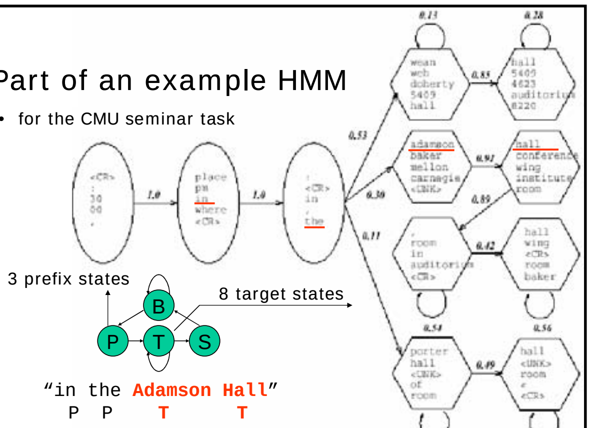


- **properties:**
  – extracts one type of target (e.g. target = bike name), need to build one model for each extracted type
  – one target state cannot model the inner structure of extracted phrases
  – model parameters can all be computed using counts from labeled training data

16

Part of an example HMM

- for the CMU seminar task



From: Freitag, D., McCallum, A.: IE with HMM Structures Learned by Stochastic Optimization

HMMs for IE - training

- **Counts from labeled training data**
  – paramaters computed directly from counts (e.g. how many times “hall” is marked as target)
  – only if there is always a unique way in the model that explains the labeling (e.g. cannot compute parameters for the 8 target states from the last slide)
- **Iterative reestimation (Baum-Welch)**
  – if there are multiple paths in the model explaining the labeling of training data
  – iterative improvement of parameters, maximizing the probability of training data

18

HMMs for IE - variations

- **emitting arcs instead of states**
- **null emissions**

- **using POS tags (certain states can emit only some POS tags) [5]**
- **emitting chunks of words instead of words [5]**

- **model structure learning [2]**

Agenda

1. **The IE problem and its motivation**
   – introduction, applications [4]
2. **Methods used for IE**
   – symbolic vs. probabilistic, LP2 [6]
3. **Markov models for IE**
   – principles, variations [5]
4. **Markov models for Bike product IE**
   – training data, models, results [10]

IV. Bike Product IE using HMMs

- **Goal**
  – semantic search application over English bikeshops in Google directory
  – e.g. “which Giant bikes are sold below 200 Euros?”, “where can I by the cheapest RockMachine Tsunami?”
- **Training data**
  – 100 labeled pages of HTML “product catalogues”
  – from English bike shops in Google directory
  – very diverse

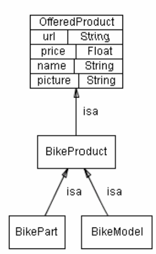
Training data

Preprocessing

- **HTML elements translated into generalized symbols using an element hierarchy (constructed ad-hoc)**
  – e.g. elements <b>, <i>, <em>, <span>, <tt>, <font>,<strong>... are grouped and treated as <styleChange>
- **Common HTML constructs translated into dedicated symbols**
  – “add to basket”, “submit form”, “choose amount”
- **Using only contents of block elements containing words or images**
- **Optionally unifying all numbers etc.**

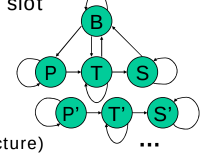
Extracted slots

- **Bike model**
  – ***name***
  – ***price***
  – ***picture***
  – category, make,
  – weight, size, color, year
- **Bike Part**
  – fork
  – frame
  – rear derailleur
  – front derailleur
  – brakes
  – ... [12]



HMM structures used

- **Single HMM for all extracted types:**
  - 1 **B**ackground state
  - 1 **T**arget, 1 **P**refix and 1 **S**uffix state type for each extracted slot
  - =1+3*N states
- **3 variants:**
  - A. simple model →
    (no internal target structure)
  - B. some target states are augmented with word ngram distributions
  - C. some target states are split into several states



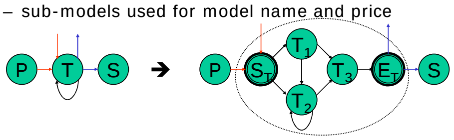
Use of word n-grams

- **Modification of the generative process**
  - if the process stays in a target state T for several time intervals, the next words generated at T are made dependent on the previously generated words at T
  - E.g. the present state is T:
    - then, if previous state was also T,
    - use $\mathbf{P(w_i|T,w_{i-1})}$ instead of $\mathbf{P(w_i|T)}$

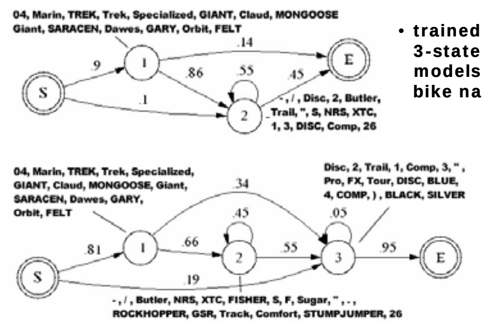
- **Word n-grams used**
  - smoothed word bigrams and trigrams were tried for chosen target states
  - linear interpolation smoothing used

Splitting target states (1)

- **Chosen target states were substituted with HMM sub-models modelling internal structure of the extracted type**
  - sub-models were iteratively re-estimated using the to-be-extracted word sequences from training data (via the Baum Welch algorithm)
  - number of sub-model states determined empirically during experiments
  - sub-models used for model name and price



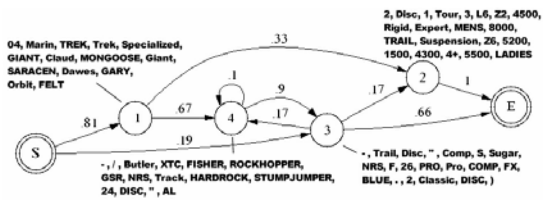
Splitting target states (2)



- **trained 2- and 3-state sub-models for bike name**

Splitting target states (3)

- **trained 4-state sub-model for bike name**



Results

- **Results were obtained using 10-fold cross-validation on the labeled set of 100 product catalogues**
- **Recall and precision are calculated on a per-word basis**
- **Bracketed numbers are with word trigram models enabled for that particular state**
- **Results for multiple target states will be available soon**

Tag	recall	precision	instances
name	77.9 (78.6)	63.5 (65.6)	927
price	98.9 (99.1)	89.5 (88.9)	971
picture	69.0	89.6	359
speed	86.8	93.6	186
size	83.2	93.7	173
year	98.1	70.0	160

References

1. Freitag D., McCallum A.: Information extraction with HMMs and shrinkage. Proceedings of the AAAI-99 Workshop on Machine Learning for IE, 1999
2. Freitag, D., McCallum, A.: Information Extraction with HMM Structures Learned by Stochastic Optimization *http://www.cs.umass.edu/~mccallum/papers/iehill-aaai2000s.ps.gz*
3. Borkar V., Deshmukh K., Sarawagi S.: Automatic segmentation of text into structured records. SIGMOD Conference, 2001
4. Schroeder I: A Case Study in POS Tagging Using the ICOPOST Toolkit. University of Hamburg, Computer Science Department, NATS, 2002
5. Ray S., Craven M.: Representing Sentence Structure in HMMs for IE. IJCAI 2001
6. Ciravegna, F.: LP2, an Adaptive Algorithm for Information Extraction from Web-related Texts *http://www.dcs.shef.ac.uk/%7Efabio/paperi/Atem01.pdf*
7. Dingli A., Ciravegna F., Guthrie D., Wilks Y.: Mining Web Sites Using Unsupervised Adaptive Information Extraction. EACL, 2003
8. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation, WWW2003 *http://www.almaden.ibm.com/webfountain/resources/semtag.pdf*
9. Semantic Web Summer school presentations *http://minsky.dia.fi.upm.es/summerschool/*
10. TAP ontology and KB *http://tap.stanford.edu*
11. Rainbow project at the Prague University of Economics *http://rainbow.vse.cz*