NAMED ENTITY RECOGNITION

Rajesh Arja
Vinit Atal
Sasidhar Kasturi
Balaji MMS

Under the guidance of

Prof. Pushpak Bhattacharya

The NER Problem

- The task is to classify named entities in a text into various name-classes such as:

✓ Entities (ENAMEX): People, organizations, locations

✓ Time (TIMEX): Date, time

✓ Numbers (NUMEX): Money, percentages

- A correct response implies a correct label (type and attribute) as well as the correct boundaries

Example

- E.g. John who is a student of Stanford University, Stanford, scored 95% in his seminar on the 11th of April.
- $ John^(ENAMEX, name) who is a student of $ Stanford University^(ENAMEX, org), $ Stanford ^(ENAMEX, location), scored $ 95% ^(NUMEX, percent) in his seminar on the $ 11th of April ^(TIMEX, date).

Motivation

- Because you NEED it and because you CAN do it.
- Applications:

QUESTION ANSWERING:

NER is extremely useful for systems that read text and answer queries.

e.g. Tasks such as “Name all the colleges in Bombay listed in the document”

INFORMATION EXTRACTION:

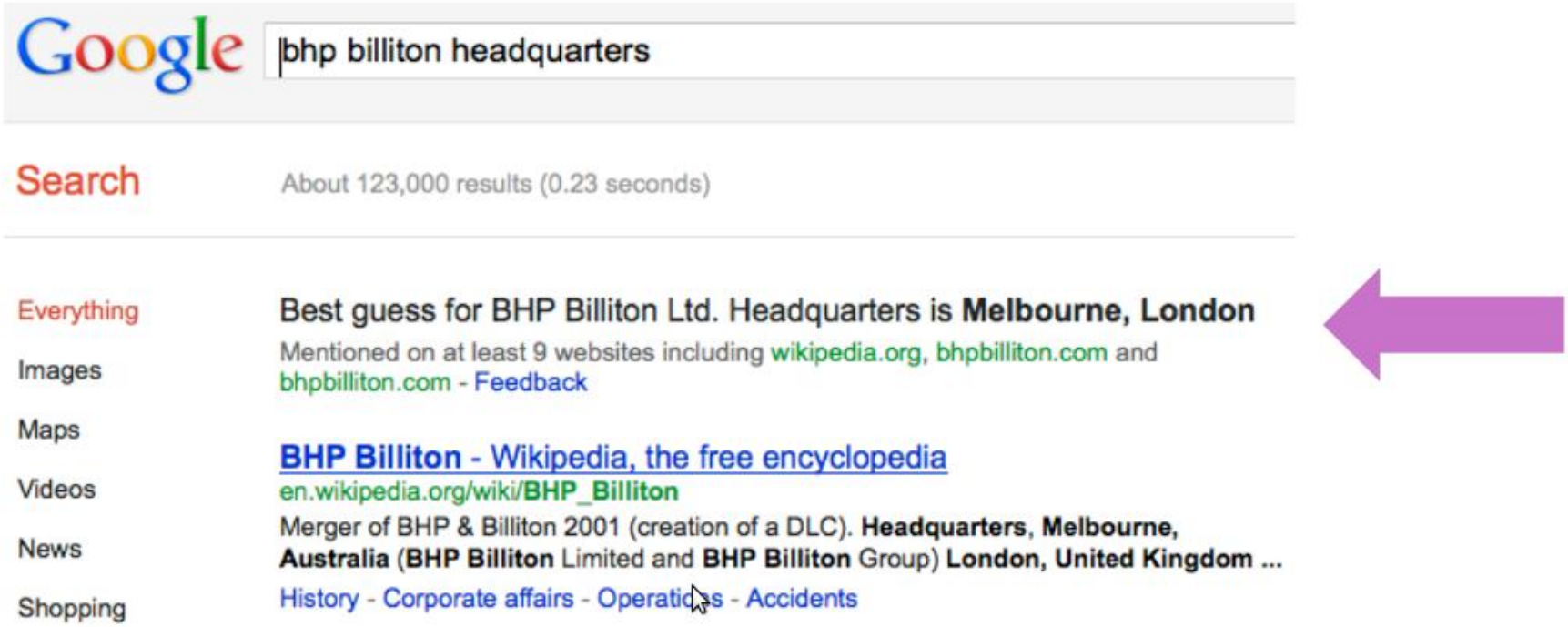
e.g. to find out and tag the subject of a web page

To extract the names of all the companies in a particular document

PRE PROCESSING FOR MACHINE TRANSLATION

WORD SENSE DISAMBIGUATION FOR PROPER NOUNS

Example given on next page


Google
bhp billiton headquarters
Search
About 123,000 results (0.23 seconds)
Everything
Images
Maps
Videos
News
Shopping
Best guess for BHP Billiton Ltd. Headquarters is Melbourne, London
Mentioned on at least 9 websites including wikipedia.org, bhpbilliton.com and bhpbilliton.com - Feedback
BHP Billiton - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/BHP_Billiton
Merger of BHP & Billiton 2001 (creation of a DLC). Headquarters, Melbourne, Australia (BHP Billiton Limited and BHP Billiton Group) London, United Kingdom ...
History - Corporate affairs - Operations - Accidents

Various Approaches

- Rule based approaches
  - Eg: Univ. of Scheffield's LaSIE-II
- Machine Learning based approaches
  - Hidden Markov Model based approach
  - Maximum Entropy Markov Model appraoch

I took the one less traveled by,
And that has made all the difference…

The HMM Model
IdentiFinder – D.M. Bikel, et al.

Naam mai rakha hi kya hai?

Our algorithm learns what's in a name!!

Bikel, et al.

Why HMM?

- Named entities are often preceded or followed by some markers which are give-aways to their class
- E.g. names are often preceded by titles such as “Mr.”, “President”, etc.
- Locations can be often recognized by the commas surrounding them e.g. “Kolkata, West Bengal”
- Companies also follow certain naming norms

  e.g. Matsushita Electrical Co., Touchmagix systems, Bremen Motor Werken, etc.
- This justifies using an HMM, which uses n-gram models

Word Features

- In Roman languages, capitalization => name
- Numeric symbols => NUMEX
- Special character sets used for transliterating names in Chinese and Japanese

(eg Scarlet O'Haara -> Si-a-ji-li O-haa-la

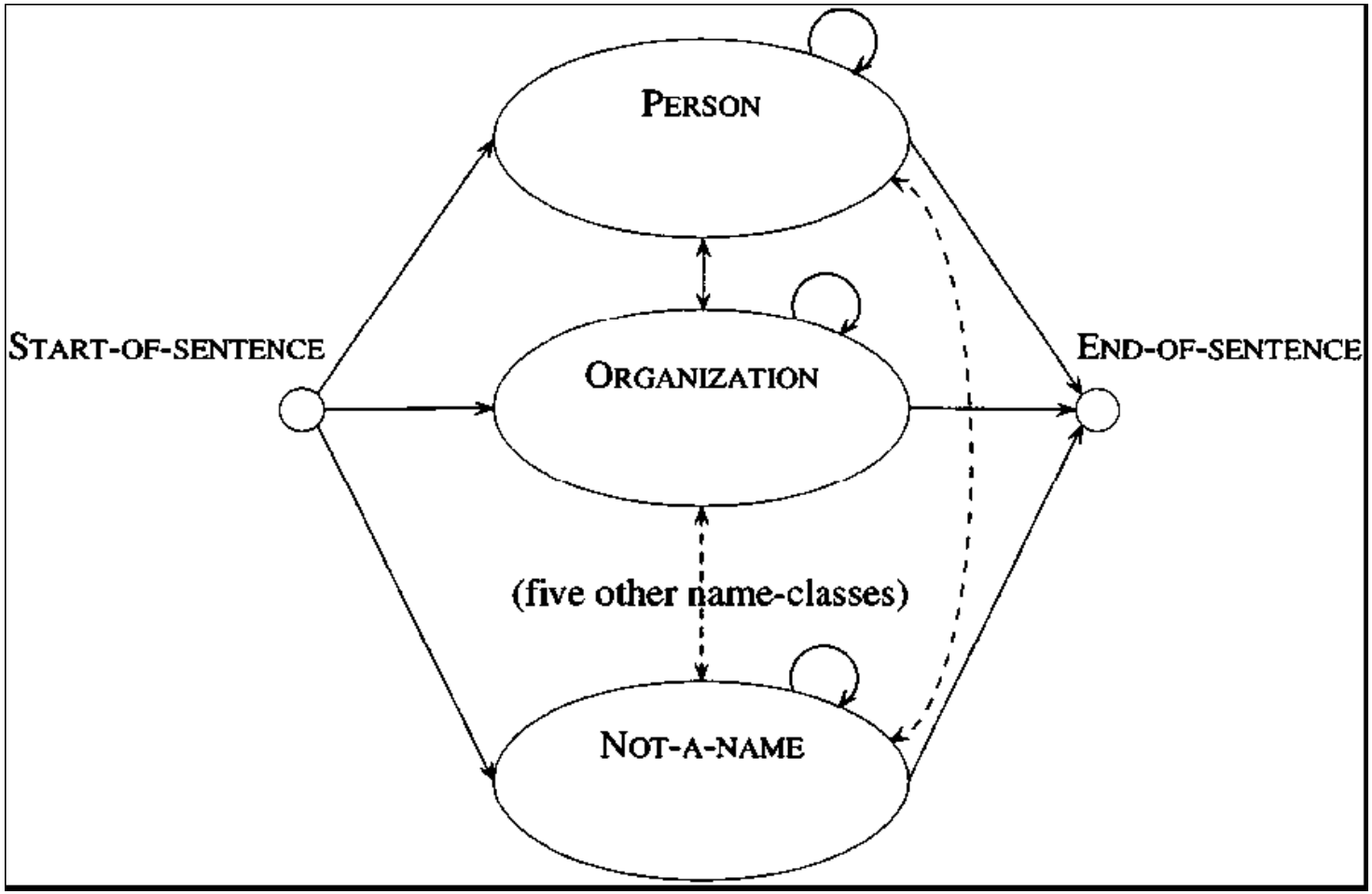
The "Si" is an archaic word usually used only in foreign names to imitate the 's' sound)

- Semantic classes according to lists of words having semantic features

Word features examples

- Word feature (Example text) = Intuition
- twoDigitNum (90) = Two-digit year
- fourDigitNum (1990) = Four-digit year
- containsDigitAndAlpha (A8956-67) = Product code
- containsDigitAndDash (09-96) = Date
- containsDigitAndSlash (11/9/89) =Date
- containsDigitAndComma (23,000.00) = Monetary amount
- containsDigitAndPeriod (1.00) = Monetary amount, percentage
- allCaps (BBN) = Organization
- capPeriod (M.) = Person name initial
- initCap (Sally) =Capitalized word
- lowerCase (can) = Uncapitalized word

The Algorithm



The Algorithm

- Each word in the text is assigned one of 8 classes, the 7 name classes mentioned earlier and a NOT-A-NAME class
- Further, each name class in the sentence has a start and end marker to mark its boundaries
- The bigram assumption is used
- We need to maximize Pr(NC|W) , i.e. Pr(W|NC)Pr(NC)

The Algorithm

- Probability for generating the first word of the name class has 2 factors:

  $Pr(NC|NC_{-1}, w_{-1}) * Pr(<w,f>_{first}|NC, NC_{-1})$

- Probability for generating all but the first word for a name class:

  $Pr(<w,f>|<w,f>_{-1}, NC)$

- Note: there is no transition probability within a name class. Hence, variations are possible.

Maximum Entropy Markov Model

- Maximum Entropy Markov Model:

$$P(f|h) = \frac{\prod_i \alpha_i^{g_i(h,f)}}{Z_\alpha(h)}$$
$$Z_\alpha(h) = \sum_f \prod_i \alpha_i^{g_i(h,f)}$$

h – History
f – Futures
Z – Normalization function
Alpha – Parameters
g – Feature function

MEMM – Features

- p(f | h_t) = p (f | information derivable from corpus relative to token t)

- g(h, t) = 1 if current_token_capitalized(h) and f = location_start

  = 0 other wise

MEMM - Formalization

$$
\begin{aligned}
Q &\equiv \left\{ \begin{array}{l} \text{An equivalence class over } \mathcal{H}. \text{ E.g. ``The} \\ \text{set of } h \text{ such that } h_{t+1} \text{ is `announced' ''} \end{array} \right\} \\
y &\equiv \text{the future ``organization\_unique''} \\
J &= \frac{|\{(h,f) \in C : h \in Q \wedge f = y\}|}{|\{(h,f) \in C : h \in Q|}
\end{aligned}
$$

- C - corpus

MEMM – Explanation

- p(y|h) = J not possible - Other characteristics of h
- Maximum Entropy – Condition on h
- Expected value over the equivalence class Q of p(y|h) is to be J

$$\sum_{(h,f)\in C:h\in Q\wedge f=y} P(h)P(f|h) = J \cdot P(h \in Q) = K$$

MEMM - Explanation

$$R \equiv \left\{ \begin{array}{l} \text{An equivalence class over } (\mathcal{H}, \mathcal{F}). \quad \text{I.e.} \\ \text{“}h_{t+1} = \text{‘announced’ and } f = \text{organiza-} \\ \text{tion\_unique} \end{array} \right\}$$

$$g_r(h, f) = \left\{ \begin{array}{lll} 1 & : & \text{if } (h, f) \in R \\ 0 & : & \text{else} \end{array} \right.$$

- Conditioning over history and features for computational ease

MEMM - Algorithm

$$\sum_{(h,f)} \tilde{P}(h) P_{ME}(f|h) g_r(h,f) = K$$

- Generalized iterative scaling

MEMM - Algorithm

- Randomly initialize alpha
- Compute K_ij s for each of the features

$$K_i^{(j)} \equiv \sum_h \tilde{P}(h) \cdot \sum_f P_j(f|h) \cdot g_i(h, f)$$

MEMM - Algorithm

- Update alphas

$$\alpha_i^{(j+1)} = \alpha_i^{(j)} \cdot \frac{K_i}{K_i^{(j)}}$$

MEMM - Algorithm

- Re-estimate conditional probabilities

$$P_{j+1}(f|h) \equiv \frac{\prod_i \alpha_i^{(j+1)g_i(h,f)}}{Z_\alpha(h)^{(j+1)}}$$

- Proved to converge
- Inference using Viterbi

The Unknown Word Conundrum

- Since we will typically deal with many proper nouns in NER, the occurrence of unknown words will be frequent, however large the training set
- It is imperative that we have a robust method to deal with unknown words
- The unknown word could be either the current or the previous word or both

The Unknown Word Model

- All unknown words are mapped to the token _UNK_
- We hold out 50% of the training data at a time and due to the generation of a lot of new unknown words, we train the unknown model on 50% of the training data and get statistics
- This is repeated for the other 50% and the statistics concatenated
- Now, whenever an unknown word is encountered, this model is invoked, else the regular one

Training data

- The training data for the mixed case English case included 650,000 words taken from the Wall Street Journal
- The Spanish dataset had 100,000 words (also, slightly inconsistent and slightly obscure domain)
- The accuracy did not drop substantially even for substantial decrease in size of training data
- Unicase and speech data made the NER task more difficult (even for humans)

Error analysis

- Eg: The Turkish company, $ Birgen Air ^ (location), was using the planes….
- Birgen = _UNK_; Air appears often in locations such as Sandhurst Air Base
- “Birgen Air” in between two commas, typically noticed for locations
- Getting rid of punctuations not a solution since they are useful
- Trigram would increase the computation

Performance of HMM v/s Rule-based

- The performance metric used is the F-measure:

  F = 2RP/(R+P) ; R = recall, P = precision

Language	Best rules	IdentiFinder
Mixed case English	96.4	94.9
Upper case English	89	93.6
Speech form English	74	90.7
Mixed case Spanish	93	90

Performance of HMM vs MEMM

- The performance metric used is the F-measure:

  F = 2RP/(R+P) ; R = recall, P = precision

Language	HMM	MEMM
English	92.5	94.02
Japanese	--	83.80

Multi Linguality

- Agglutinative Nature (Oorilo – ఊరిలో)
- Ambiguity

  person name Vs place name
  (Tirupathi - తిరుపతి)

  person first name Vs common noun
  (Bangaru - బంగారు)

  person last name Vs organization
  (TaTa – టాటా)
- Spelling Variation (B.J.P vs Ba.Ja.Pa)

Multi Linguality contd ..

- Frequent word list
- Useful unigrams (UNI)
- Useful bigrams (UBI)
  - Ex: In the village (Oorilo ఊరిలో (ooru + lo))
- Word suffixes (SUF)
  - Ex: Reddy, Naidu, Rao
- Name class suffixes (NCS)
  - Ex: party, samstha (పార్టీ, సంస్థ)

Conclusion

- NER
  - A very important task
- Can be solved with high accuracy
  - HMM
  - MEMM
- Challenges exist with various languages

References

- 1) Andrew Borthwick. 1999. A Maximum Entropy Approach to Named Entity Recognition. PhD thesis, New York University.

- 2) D. M. Bikel, R. L. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. Machine Learning, 34:211–231, 1999.

- 3) G.V.S.RAJU, B.SRINIVASU, Dr.S.VISWANADHA RAJU, K.S.M.V.KUMAR, Named Entity Recognition For Telugu Using Maximum Entropy Model, Journal of Theoretical and Applied Information Technology.

- 4) http://nlp.stanford.edu/