Named Entity Recognition and Tagging

An Overview

Aman Madaan

Indian Institute of Technology Bombay, Mumbai

January 23rd, 2014

Table of Contents

Recognition and Tagging : Two step problem

Michael Jordan is a Professor at Berkeley

Recognition and Tagging : Two step problem

Michael Jordan is a Professor at Berkeley

- Step 1 : **Identify** entities

  Michael Jordan_PERSON is a professor at Berkeley_INSTITUTION

Recognition and tagging : Two step problem

Michael Jordan is a Professor at Berkeley

- Step 1 : **Identify** entities

  Michael Jordan_PERSON is a professor at Berkeley_INSTITUTION

- Step 2 : **Link** entities to knowledge bases :

  Michael Jordan_ENTITY (`http://en.wikipedia.org/wiki/Michael_I._Jordan`) is a professor at Berkeley_ENTITY (`http://en.wikipedia.org/wiki/University_of_California,_Berkeley`)

NER : Problem Statement

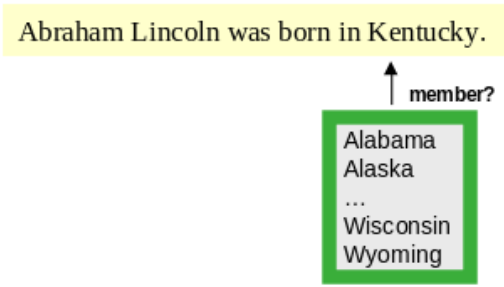
Definition (Named entity recognition[a])

[a] from 4

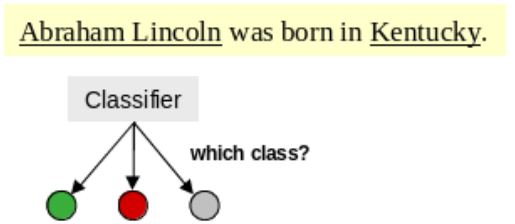
Named-entity recognition (NER) (also known as entity identification and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

NER : Solutions

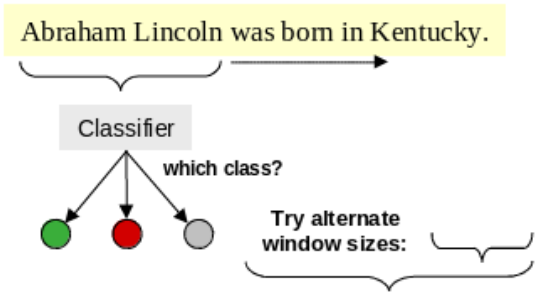
Lexicons



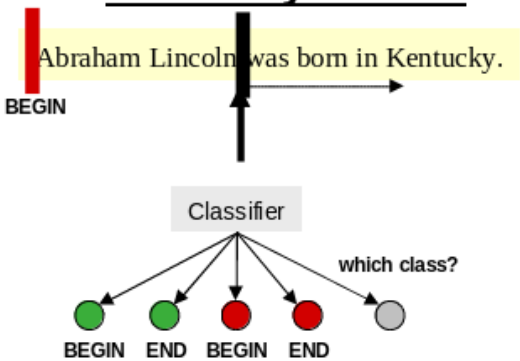
Classify Pre-segmented Candidates



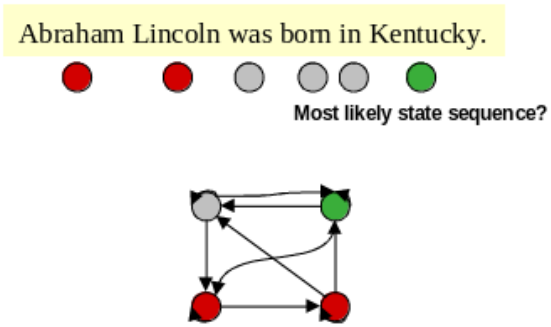
Sliding Window



Boundary Models



Finite State Machines



1

[1] from 3

NER as a sequence labeling problem

- Observation sequence : Text
- State sequence : Labeling of the sequence with elements in (PER, LOC, ORG) etc.
- Find $\text{argmax}_S P(S|O)$
- Candidates : HMM, MEMM, CRF

NER as a sequence labeling problem

- HMM
  - Generative
  - Makes strong independence assumption
  - Myopic (Refer to label bias problem in William Cohen's Survey)
- MEMM
  - Discriminative
  - No independence assumptions are made, by formulation
  - Allows the use of feature functions
  - Myopic
- CRF
  - Discriminative
  - MEMM + non myopic, avoids local normalization
  - Talks of "compatibility", not independence (CS 728)

Collective Annotation of WikiPedia Entities in Web Text

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti

Key Intuition : Topical Coherence

- A document is usually about one topic
- Disambiguating each entity using the local clues misses out on a major piece of information : Topic of a page
- A page is usually has one topic, you can expect all the entities to be *related* to the topic *somehow*

Eg. : Michael Jackson : 30 Disambiguations http://en.wikipedia.org/wiki/Michael_Jackson_(disambiguation) John Paul : But if they are mentioned on the same page, the page is most likely about Christianity, A big hint towards disambiguating **both** of them

Challenges

- Capturing local compatibility
- Inculcating topical coherence in the overall objective

Challenges

- Capturing local compatibility
  - Create a scoring function to rank possible candidates
- Inculcating topical coherence in the overall objective

Challenges

- Capturing local compatibility
  - Create a scoring function to rank possible candidates
- Inculcating topical coherence in the overall objective
  - Define Topical coherence

Local compatibility

- s : Spot, an Entity to be disambiguated (Christian leader John Paul)
- γ : An entity label value (`http://en.wikipedia.org/wiki/Pope_John_Paul_II`)
- $f_s(\gamma)$: A feature function that creates a vector of features

Local compatibility : Feature design

- 1. Take
  - Text from the first descriptive paragraph of γ
  - Text from the whole page for γ
  - Anchor text within Wikipedia for γ.
  - Anchor text and 5 tokens around γ
- 2. Apply each of the following operation with one argument as Spot
  - Dot-product between word count vectors
  - Cosine similarity in TFIDF vector space
  - Jaccard similarity between word sets

Total 12 Features (3 operations, 4 argument pairs) + Sense Probability Prior[2]

[2] Obtained by counting intra wiki links

Compatibility Score

- Local compatibility score between a spot s and a candidate is given by $w^T f_s(\gamma)$
- Thus, candidate is picked by $argmax_{\gamma \in \Gamma} w^T f_s(\gamma)$
- w is trained using an SVM like training objective

$$w^T f_s(\gamma) - w^T f_s(\gamma) \geq 1 - \epsilon_s$$

Defining topic Relatedness

- We need some notion of capturing the fact that 2 topics are related to each other
- Given
  - $g(\gamma)$: Set of wikipedia pages that link to γ
  - c : Total number of Wikipedia pages
  - $r(\gamma, \gamma')$: Relatedness of topics γ and γ'
- Define $r(\gamma, \gamma') = \frac{log|g(\gamma) \bigcap g(\gamma')| - log(max\{|g(\gamma)|, |(\gamma')|\})}{logc - log(min\{|g(\gamma)|, |(\gamma')|\})}$

The Dominant Topic Model

- Need to define a collective score based on pairwise topical coherence of all γ_s used for labeling.
- The pairwise topical coherence, $r(\gamma_s, \gamma'_s)$ is as defined above.
- For a page, overall topical coherence :

$$\Sigma_{s \neq s' \in S_0} r(\gamma_s, \gamma'_s)$$

- Can be written as clique potential as in case of node potential

$$exp(\Sigma_{s \neq s' \in S_0} r(\gamma_s, \gamma'_s))$$

The Optimization objective

$$\frac{1}{\binom{|S_0|}{2}} \Sigma_{s \neq s' \in S_0} r(\gamma_s, \gamma'_s) + \frac{1}{|S_0|} \Sigma_{s \in S_0} w^T f_s(\gamma)$$

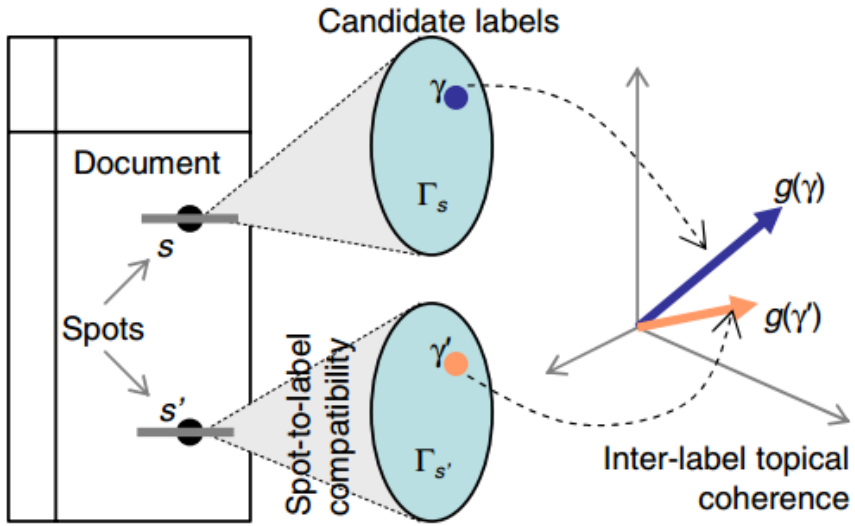


Figure 3: Labels $\gamma \in \Gamma_s, \gamma' \in \Gamma_{s'}$ have to be chosen for spots s, s' to maximize a combination of spot-to-label compatibility scores $\mathrm{NP}_s(\gamma), \mathrm{NP}_{s'}(\gamma')$ as well as topical similarity between γ and γ', say, $g(\gamma)^\top g(\gamma')$. [3]

[3] From 1

Solving the optimization objective

- LP rounding approach
- Hill climbing

1: initialize some assignment $y^{(0)}$
2: **for** $k = 1, 2, \ldots$ **do**
3: select a small spot set S_Δ
4: **for** each $s \in S_\Delta$ **do**
5: find new γ that improves objective
6: change $y_s^{(k-1)}$ to $y_s^{(k)} = \gamma$ greedily
7: **if** objective could not be improved **then**
8: **return** latest solution $y^{(k)}$

Experiments : Data preparation

- August 2008 version of WikiPedia used, 5.15 million entity IDs.
- Filter out IDs composed of verbs, adverbs, conjunctions etc.
- Create a trie from IDs.
- Identify spots (*NER*) by tokenizing the document and then matching spots with the trie.

Experiments : Preparing Ground Truth Collection

- Need data annotated with links to Wikipedia
- Done manually, pages obtained from popular links across various domains
- 19, 000 annotations marked, 40% marked NA, 3800 distinct entities used

Number of documents	107
Total number of spots	17,200
Spot per 100 tokens	30
Average ambiguity per Spot	5.3

Results : Only Local disambiguation

- Local approach performs well

$$\gamma_0 \leftarrow argmax_{\gamma \in \Gamma_s} w^T f_s(\gamma)$$

if $w^T f_s(\gamma_0) > \rho_{NA}$ then return γ_0 else return NA

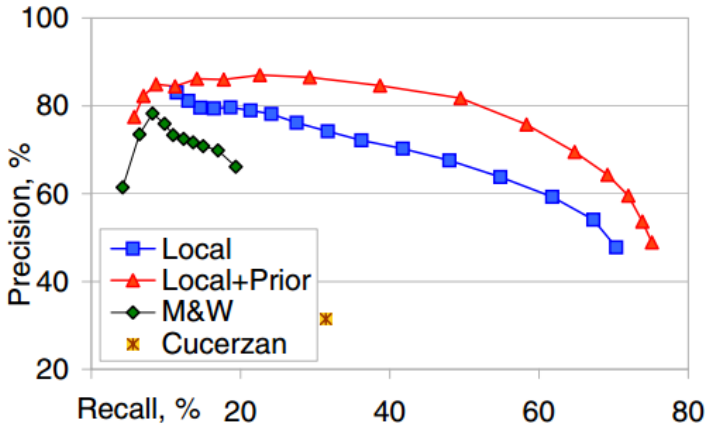


Figure 9: Even a non-collective Local approach that only uses trained node potential dominates both Cucerzan and M&W's algorithms wrt both recall and precision (IITB data).

4

[4] from 1

LP vs Hill climbing approach

- Hill climbing and LP are equivalent

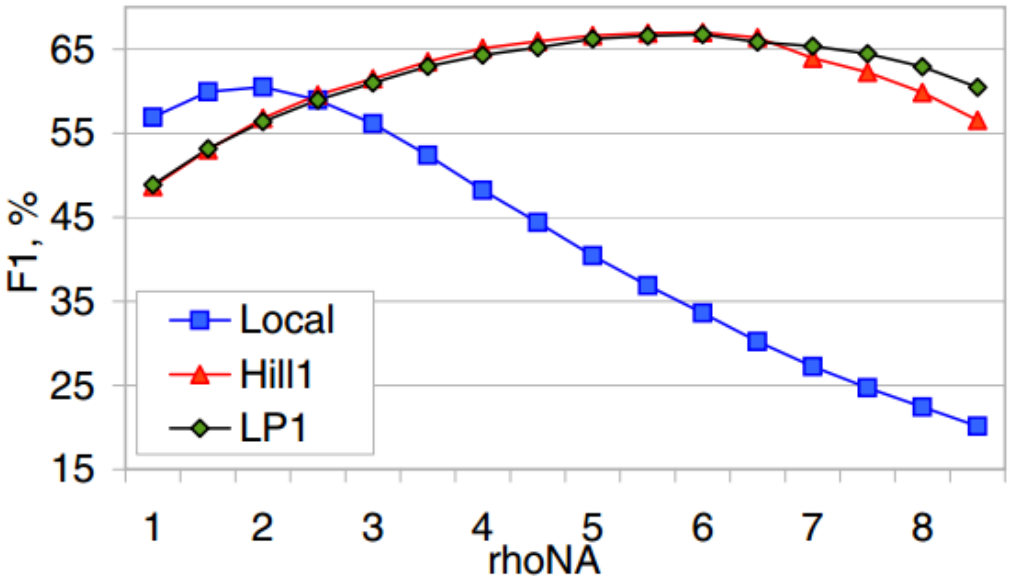


Figure 12: Hill1 attains almost the same F_1 score as LP1; both are better than Local (IITB data).

5

[5] from 1

Recall precision for various approaches

- Exploiting topical coherence improves precision by 9
- Adding topic prior also helps

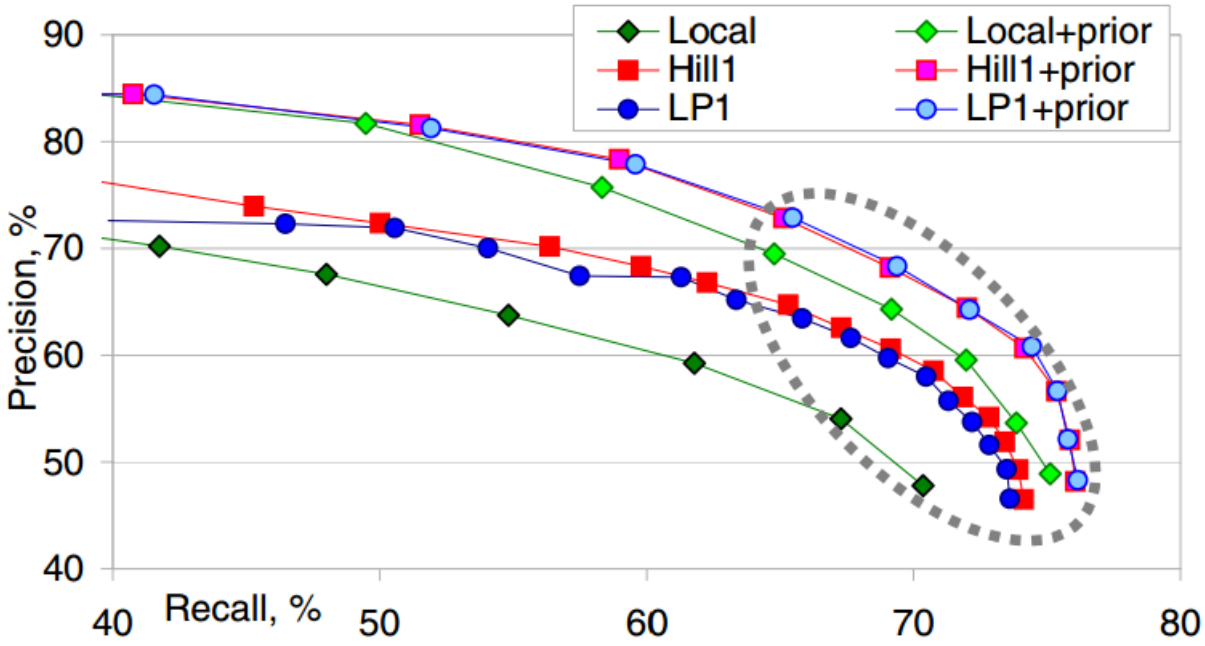


Figure 14: Recall/precision on IITB data.

[6]

[6] from 1

Summary

- Authors exploit the intuition that a document is usually about one topic
- Entities belonging to a document can thus have a common *background*
- Importance of good features design (local disambiguation)
- Global topical coherence was added to obtain the global optimization objective
- Extensive data preparation effort

References

[1] Collective Annotation of Wikipedia Entities in Web Text
Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti, IIT Bombay

[2] http://www.cse.iitb.ac.in/~soumen/OWI/Slides/

[3] William Cohen's Survey available at 2

[4] http://en.wikipedia.org/wiki/Named-entity_recognition