

Hierarchical classification for Multilingual Language Identification and Named Entity Recognition

Saatvik Shah Vaibhav Jain
Anshul Mittal Sarthak Jain
Shubham Tripathi Jatin Verma
Rajesh Kumar

Malaviya National Institute of Technology, Jaipur

December 04, 2015

Problem Statement

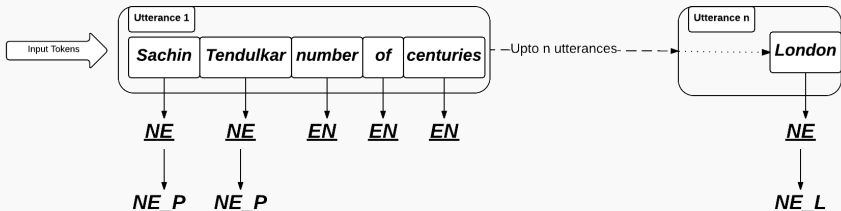
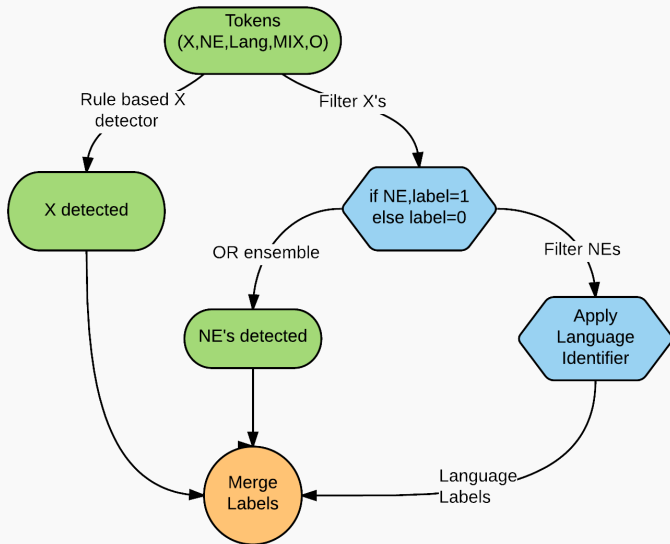


Figure: Task Description

Challenges

- ▶ Small dataset
- ▶ NE's not subclassified in training data
- ▶ 9 languages!!

Hierarchical Classification Workflow



Let's Begin!

Sample Utterance

kohli ki consistency !! IndiAn3 team lineup #wow

Punctuation Recognition

Sample Utterance

kohli ki consistency !! IndiAn3 team lineup #wow

1. Regex based cleanup to mark tokens such as
 - ▶ **http://** or **abc@def.xyz** for Web URLs
 - ▶ **:), >:(, :-P** for Emoticons
2. Gazetted List of **Acronyms**

Feature Set[1]

Sample Utterance - {X}

kohli ki consistency **IndiAn3** team lineup

- ▶ Word Context - (*ki, consistency, **IndiAn3**, team, lineup*)
- ▶ Character level n-gram, 2,3,4,5-gram
 - ▶ 3-gram - (*ind*),(*ndi*),(*diA*),(*iAn*),(*An3*)
- ▶ Relative Position - *BEGIN/END Tag*
- ▶ Word Normalization - *AaaaAa0*
- ▶ Composition Features - *WordDigit*
- ▶ POS Tags

and a few more...

Converted to a **Bag Of Words** representation

Named Entity Recognition[2, 3]

kohli ki consistency **IndiAn3** team lineup

Supervised NE classification

- ▶ Binary classification of words : **NE** or **Non-NE**.
- ▶ Linear SVM, Logistic Regression and Random forests
- ▶ **Logical OR** ensemble for final prediction.

Unsupervised NE subclassification

- ▶ Extract Wikipedia Page
- ▶ If **Disambiguation**: Search '*\$query wikipedia*' on Bing
- ▶ Extract most relevant non-Disambiguation
- ▶ Apply text keyword scoring

Unsupervised approach - I[6]

22,70,000 RESULTS Narrow by language Narrow by region

AAP - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/AAP

Aap or **Aa** may refer to: **Aa** (water), a Sanskrit word Argumentum ad populum, a logical fallacy Para Arara language, a Cariban language of Brazil Contexts: 1 Aviation and Aviation and aerospace · Business · Media

Aam Aadmi Party - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Aam_Aadmi_Party

Aam Aadmi Party (transliteration: common Man's Party; abbreviated as AAP) is an Indian political party, formally launched on 26 November 2012 and is currently the ruling party ...
 Background · Ideology and issues · Agenda · Support · Protests

AAP TV - Wikipedia, the free encyclopedia
https://en.m.wikipedia.org/wiki/AAP_TV

AAP TV. Type: cable television network: Branding: AAP TV. Country: UK: Availability: United Kingdom, Europe, Asia, Africa: Slogan: The Channel That Speaks Your Language

Sanjay Singh AAP - Official Site
www.aamaadmiparty.org/official-spokespersons

AAP Report on crimes against women in UP. Metro Link Campaign. Grievance Redressal Mechanism. Make a Donation. Contribute to fight Corruption! Rs.201 · Rs.501 · ...

AAP - What does AAP stand for? The Free Dictionary
acronyms.thefreedictionary.com/AAP

See results for
Aam Aadmi Party
 aap

Related searches
 Aam Aadmi Party Wikipedia
 Raghav Chadha Aap Wiki
 Aap Ki Adalat Wiki
 En wikipedia Aap Ki Adalat
 What's Aap
 Aam Aadmi Party
 Raghav Chadha Aap Wikipedia
 Ashutosh Aap Wikipedia

Unsupervised approach - II

Your continued donations keep Wikipedia running!

article | discussion | edit this page | history

Ronn Torossian

From Wikipedia, the free encyclopedia

This article or section reads like a news release, or is otherwise written in an overly promotional tone.
 Please help rewrite this article from a neutral point of view to be less promotional.
 Where appropriate, blatant advertising may be marked for speedy deletion with {{db-ppan}}.

The creator of this article, or someone who has substantially contributed to it, may have a conflict of interest regarding its subject matter.
 It may require cleanup to comply with Wikipedia's content policies, particularly neutral point of view.
 Please discuss further on the talk page.

Ronn Torossian is the founder, president & CEO of New York City-based SW Public Relations.^[1] ^[2]

Torossian has been referred to by *New York Post* as a "publicity guru", by *Fox News* as a "high-powered PR CEO", and by *CNN* as "a leading PR expert". Torossian is regularly featured in and covered by the media including *CNN*, *Fox News Channel*, *MSNBC*, *ABC*, *NBC*, and *The New York Times*. He has also been named to *PR Week's* 40 under 40^[3] and *Advertising Week's* 40 under 40^[4]. CBS National News said "Ronn Torossian knows spin." and *Rolling Stone* magazine called him "the most powerful man in advertising, scrappy NY publicist."

Early Life

Torossian has been involved in Zionism since his childhood. He was a member of the *Beitar*, of which he became national president^[5], he became known and sought out among some of the hawkish Israeli public officials. He represented *Beitar* graduate Israeli Prime Minister *Yitzhak Olmert* while Olmert was still mayor of Jerusalem, the Israeli Ministry of Tourism and Knesset member *Benny Elon* among others.

He graduated from *Stuyvesant High School* in 1992. Torossian has represented conservative Christian groups and organizations and religious leaders such as *Pastor Benny Hinn*, *Regent University*, *Tinity Broadcasting Network*^[6]

References

- ↑ "Transcript: Corporate Terrorism ?, FOXNews.com, December 21, 2005, partial transcript from "Your World with Neil Cavuto," December 20, 2005.
- ↑ "Brash P.R. Guy Grabs Clients, link. [ⓘ]. *New York Times* (February 20, 2005). Retrieved on 2007-11-18. "At 30, Mr. Torossian is the founder and president of his own agency, whose name was inspired by the five journalistic Ws—who, where, when, etc. He may be particularly busy this month with his client L.F. Kim, the rapper who is headed to trial on Feb. 28 on perjury charges stemming from a shootout outside a Manhattan radio station, but who is just one of a large and seemingly contradictory stable of customers."
- ↑ http://www.presswire.com/40-under-40/industry/9999/ [ⓘ].
- ↑ http://adage.com/archive-date/State/2006-08-07/ [ⓘ]. SWPR, Ronn Torossian
- ↑ *Scheinman, Serge* (September 16, 1997), "Jews Cust Arab Tenants From House In Jerusalem" [ⓘ], *The New York Times*, "Inside the house, the settlers and their supporters celebrated their latest action, raising Israeli flags at the house and putting up slicters that read 'Jerusalem is Ours. Ronn Torossian, 23, who said he came from the Bronx, posed for television cameras with a large Israeli flag."
- ↑ Template error: argument title is required.
- ↑ *Sarayan, Strawberry* (April 16, 2006), "Christianity, the Brand" [ⓘ], *New York Times Magazine*.


Categories

Categories: Living people | Public relations people | Armerian-Americans

Summary

Section Heading

Ronn D. Torossian



Born	1974 United States
Nationality	Israel, United States
Occupation	Public relations
Employer	SW Public Relations
Title	CEO
Website	Torossian's Web Page [ⓘ]

Infobox Contents

Language Identification[1, 4, 5]

Sample Utterance - {X,NE}

ki consistency team lineup

- ▶ We implemented Linear-Kernel Support Vector Machines
- ▶ Important features, on cross validating,
 - ▶ Word N-gram, 2,3,4,5-gram
 - ▶ Local Knowledge
 - ▶ Part of Speech Tags
 - ▶ Composition Features

Merged Labels

Sample Utterance

kohli ki consistency !! IndiAn3 team lineup #wow

Labelled Utterance

NE_P hi en X NE_P en en X

Done!

Results

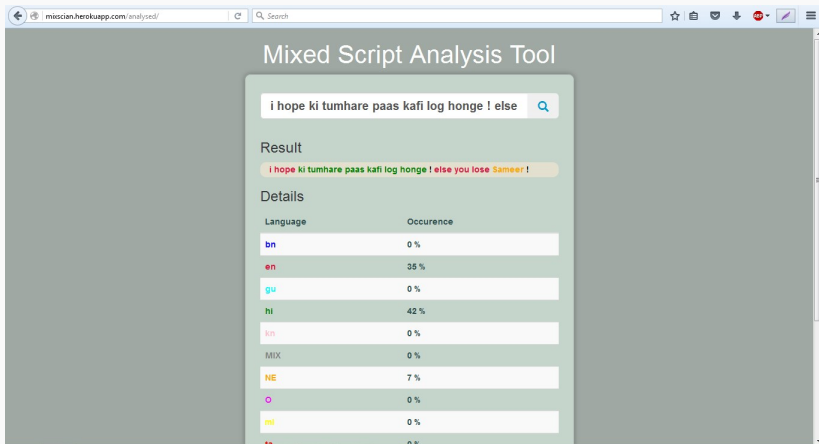
P, R, F-S	Bn	En	Gu	Hi	Kn	MI	Mr	Ta	Te	X	NE	NE_L	NE_P
P	0.878	0.958	0.097	0.817	0.575	0.394	0.705	0.937	0.431	0.961	0.368	0.722	0.2121
R	0.966	0.848	0.5	0.74	0.829	0.752	0.79	0.708	0.687	0.966	0.528	0.124	0.25
F-S	0.838	0.9	0.163	0.776	0.679	0.517	0.745	0.806	0.529	0.964	0.433	0.214	0.229

Table: Strict Precision (P), Recall (R), F-scores (F-S) for languages and NEs.

COMMENTS:

1. Weighted F-measure = 0.808
2. Gujarati and Hindi are morphologically quite similar.
3. Indian names commonly confused with Hindi

Implementation



Mixed Script Analysis Tool

i hope ki tumhare paas kafi log honge ! else

Result

i hope ki tumhare paas kafi log honge ! else you lose Sameer !

Details

Language	Occurrence
bn	0 %
en	35 %
gu	0 %
hi	42 %
kn	0 %
MIX	0 %
NE	7 %
o	0 %
ml	0 %
ta	0 %

- ▶ Web Application at <https://mixscian.herokuapp.com>
- ▶ Source code at <https://github.com/saatvikshah1994/hline>

Conclusion

1. What worked:
 - ▶ Splitting the classification problem into a hierarchy
 - ▶ Bing + Wikipedia for highly accurate NE subclassification
 - ▶ LinearSVM
2. What didn't work:
 - ▶ IRSLTM
 - ▶ LIGA
 - ▶ Home built Spell Corrector
 - ▶ Stanford NER Tagger
3. What will be done:
 - ▶ More Data!!
 - ▶ More Gazetted Lists
 - ▶ Word Embeddings as a feature for NER
 - ▶ Adding a Language Group level to Hierarchy

Acknowledgments

The authors would like to thank the reviewers and organizers.
We also like to thank **ACM SIGIR** for the travel grant provided.



Questions???

- [1] Gupta, D. K., Kumar, S., & Ekbal A. *Machine Learning Approach for Language Identification & Transliteration: Shared Task Report of IITP-TS..* Centre for Digital Music, 2012.
- [2] Abhinaya N., Neethu John, Dr. M. Anand Kumar, Dr. K.P. Soman Amrita @ FIRE-2014: *Named Entity Recognition for Indian Languages.* 2014.
- [3] Dubey, S., Goel, B., Prabhakar, D. K., & Pal S.. *ISM@ FIRE-2014: Named Entity Recognition Indian Languages.*
- [4] King, B., & Abney, S. P. *Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods.* In HLT-NAACL (pp. 1110-1119). 2013.
- [5] A. Das and B. Gamback. *Code-Mixing in Social Media Text: The Last Language Identification Frontier? Traitement Automatique des Langues (TAL): Special Issue on Social Networks and NLP , TAL Volume 54 no 3/2013, Pages 41-64.*
- [6] Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. *Learning multilingual named entity recognition from Wikipedia.* Artificial Intelligence, 194, 151-175. 2013.