MSR KMG at KBA 2014 Vital Filtering Task

Jingtian Jiang, Chin-Yew Lin, and Yong Rui

Knowledge Mining Group, Microsoft Research

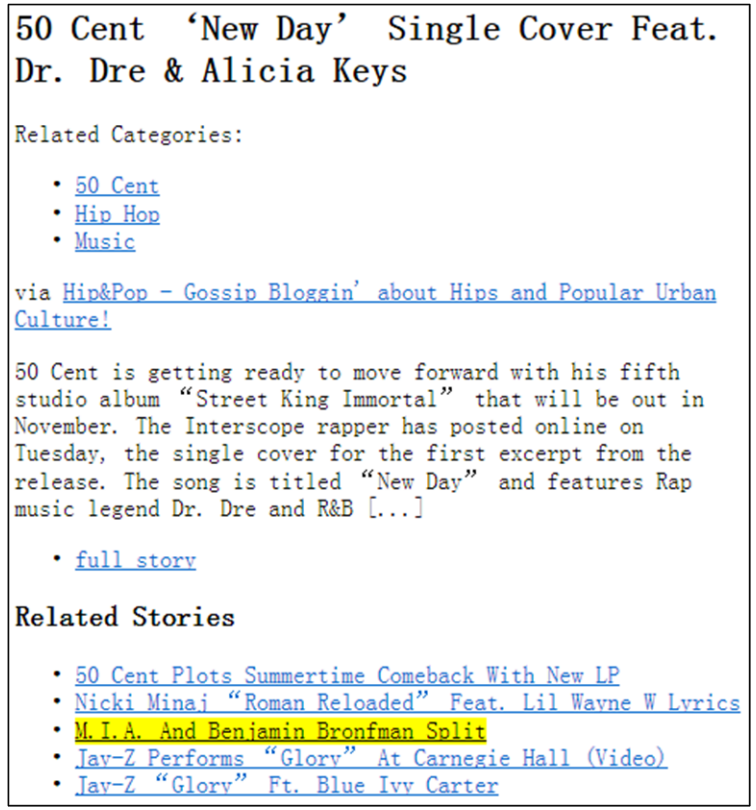
11/20/2014 @ NIST

Overview

- Observations
- Preprocessing
- Features
- Models
- Global Adjustment
- Results
- Error Analysis & Future work

Observations

- A vital event could be at sentence level



50 Cent ‘New Day’ Single Cover Feat. Dr. Dre & Alicia Keys

Related Categories:

- 50 Cent
- Hip Hop
- Music

via Hip&Pop - Gossip Bloggin’ about Hips and Popular Urban Culture!

50 Cent is getting ready to move forward with his fifth studio album “Street King Immortal” that will be out in November. The Interscope rapper has posted online on Tuesday, the single cover for the first excerpt from the release. The song is titled “New Day” and features Rap music legend Dr. Dre and R&B [...]

- full story

Related Stories

- 50 Cent Plots Summertime Comeback With New LP
- Nicki Minaj “Roman Reloaded” Feat. Lil Wayne W Lyrics
- M.I.A. And Benjamin Bronfman Split
- Jay-Z Performs “Glory” At Carnegie Hall (Video)
- Jay-Z “Glory” Ft. Blue Ivy Carter

Most Popular

- Viewed
- Commented

- 1. Who benefits when Mark Carney takes a hit over a whitewashed Asian face on a $100 bill? Co
- 2. Young woman strangles old man with her bra
- 3. New duty-free limits for Canadians: what you can bring back, alcohol and all Comments (13)
- 4. Choreographer Crystal Pite nabs national award Comments (2)
- 5. Beloit College's 2016 Mindset List reveals mind-wobbling generation gap
- 6. George Galloway comes under fire for saying Julian Assange was guilty of bad sexual etiquette
- 7. David Suzuki: Climate change deniers are almost extinct Comments (3)
- 8. Glowbal Group's Sanafir to close; new restaurant scheduled for fall
- 9. Chad Kroeger to marry Avril Lavigne. Yes, really.
- 10. Kelly Ripa has reportedly chosen a football player, Michael Strahan, as cohost

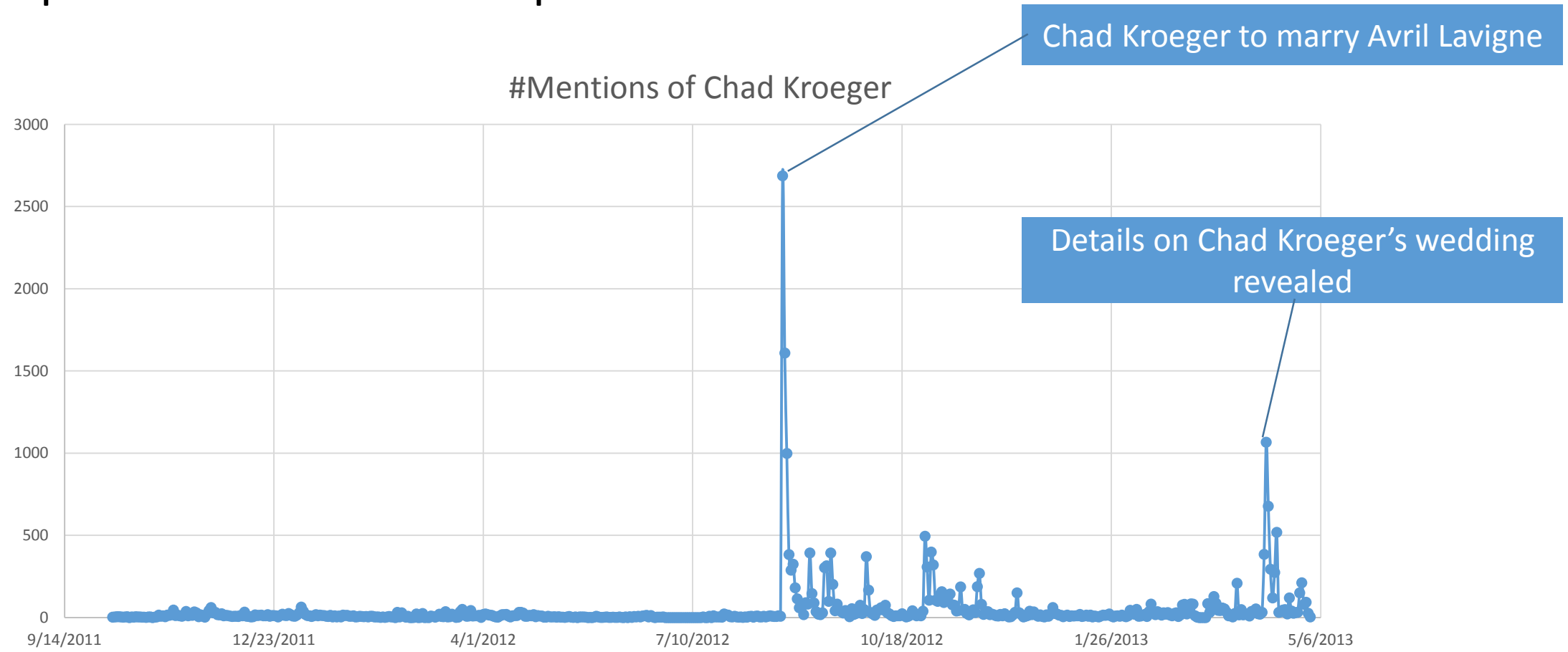
- 1. The Tiffin Project tackles takeout waste with Vancouver restaurants Comments (2)
- 2. A fake Jam pleases punters
- 3. Christy Clark clobbers Adrian Dix in online poll...for now Comments (23)
- 4. New Westminster mayor Wayne Wright wins reelection Comments (1)
- 5. COPE to hold masquerade ball at Museum of Vancouver Comments (2)
- 6. Gordon Campbell fan club president Kevin Krueger won't seek reelection Comments (11)
- 7. Protest targets private Vancouver clinic accused of extra billing patients Comments (10)
- 8. Slash slagged by Ear of Newt reader who calls Thin Lizzy and Led Zeppelin "pretty awesome"
- 9. Vancouver police: Don't call 911 if you see a "beggar" Comments (3)
- 10. Morning rush hour SkyTrain service extended for cyclists, TransLink says Comments (1)

Observations (2)

- An acting target entity is a good vital indicator
  - Examples
    - Chad Kroeger to marry Avril Lavigne -> vital
    - Peter Goldmark won re-election -> vital
    - Lizette Graden appointed new chief curator -> vital
    - Ted Sturdevant was happy with the ruling -> non-vital
    - Dan Satterberg said in a statement -> non-vital

Observations (3)

- Temporal information is important



Preprocessing

- Index
  - Use ElasticSearch to index the corpus
- Retrieval
  - Retrieve the documents containing entity names or surface forms
    - The surface forms are from the redirect names in Wikipedia if the entity is from Wikipedia
- Sentence Extraction
  - Extract the sentences containing entity names or surface forms
- Stream Clustering
  - When a document arrives
    - Cluster it into an existing cluster
    - Or form a new cluster by itself

Features

- Time Range
- Temporal Feature
- Title/Profession Feature
- Action Patterns

Time Range

- An earlier document is likely more vital than a later document
- The documents after 3 days of the event are non-vital
- A document's feature value is defined as:

$$\mathrm{tr}(d_i) = 1.0 - (\,h_i - h_0)/72.0$$

  - d_i: the *ith* document in a cluster
  - h_i: the hour of the document d_i, e.g., 2012-11-12-09
  - h_0: the hour of the first document (the event beginning)
  - 72.0=24.0*3.0

Temporal Feature

- Given an entity e and a document d, its burst value is defined as below [Balog et al., OAIR'13]:

$$\text{burst\_value}(e, d) = \frac{m(e, h)}{\left(\frac{M(e)}{N}\right)}$$

  - h: the hour of document d, e.g., 2012-11-12-09
  - $m(e, h)$: number of mentions of entity e in hour h
  - $M(e)$: total mentions of entity e from the beginning of stream corpus to hour h
  - N: total hours from the beginning of stream corpus to hour h

Title/Profession Feature

- Some entity mentions are ambiguous
- Entity mentions are usually accompanied with entity title and profession
  - Lions coach Bill Templeton said
  - Bill Templeton, an organizer for the local chapter of Pennsylvania Association of Sustainable Agriculture (PASA)
- Use the similarity between the local words and the entity's title/profession as a feature

Title/Profession Feature (2)

- Title/Profession dictionaries from Freebase
  - 2,294 titles/2,440 professions
- Adopt word based n-gram (n=1,2,3,4,5) inside a [-5, 5] window around a mention
- Construct the title/profession vector using n-grams from Vital and Useful documents for each entity
- For each entity-document pair, compute the cosine similarity between title/profession vector and n-gram vector from the document

Action Patterns

- Employ ReVerb [Fader et al., EMNLP'11] to extract triples from sentences
- Use entity name + verb (relation) as binary features
  - Examples
    - {Entity} announces/appoints/fights/signs …
    - {Entity} told/said/asked…
    - {Entity} plans/expects/hopes…
    - {Entity} is/was
- 8K ReVerb patterns
  - Two annotators went through these patterns and select top 200 and 250 patterns in one week's efforts

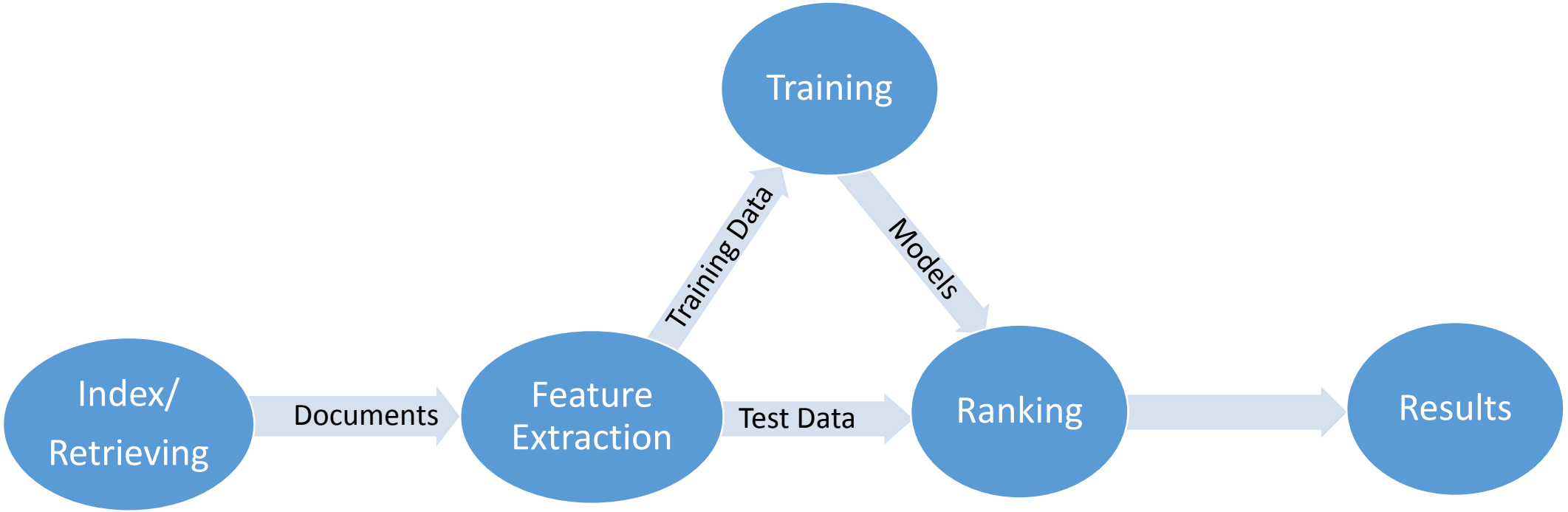
Models

- Ranking Method
  - Random forest (decision tree)
  - Use RandLib.jar from http://sourceforge.net/projects/lemur/
    - Number of bags (-bag): 800
    - Sub-sampling rate (-srate): 1.0
    - Feature-sampling rate (-frate): 0.3

Global Adjustment

- Each entity has its best threshold
- Different entities have different best thresholds
- There is only one global best threshold for all entities in evaluation
- So adjust the scores of each entity to align their best thresholds

System Overview



Results

Features	avg(P)	avg(R)	max(F(avg(P), avg(R)))	max(SU)
Baseline	0.288	0.953	0.442	0.267
+Time Range	0.342	0.774	0.474	0.349
+Temporal Feature	0.332	0.779	0.466	0.335
+Title/Profession Feature	0.305	0.899	0.456	0.312
+Action Patterns (200)	0.356	0.885	0.507	0.362
+Action Patterns (250)	0.381	0.780	0.512	0.397

Results (2)

Features	avg(P)	avg(R)	max(F(avg(P), avg(R)))	max(SU)
Baseline	0.288	0.953	0.442	0.267
+Time Range	0.342	0.774	0.474	0.349
+Time Range +Temporal Feature	0.367	0.743	0.491	0.367
+Time Range +Temporal Feature +Title/Profession Feature	0.378	0.744	0.501	0.377
+Time Range +Temporal Feature +Title/Profession Feature +Action Patterns	0.447	0.702	0.546	0.464
+Time Range +Temporal Feature +Title/Profession Feature +Action Patterns +Global Adjustment	0.472	0.706	0.566	0.510
Official Result	0.441	0.674	0.533	0.329

Results (3)

Features	avg(P)	avg(R)	max(F(avg(P), avg(R)))	max(SU)
Baseline	0.288	0.953	0.442	0.267
+Time Range	0.342	0.774	0.474	0.349

- Examples
  - Non-vital documents are suppressed
    - State schools chief Randy Dorn has sent a letter to legislative leaders …
      - 10 days after the event
    - Maelle Ricker grabs another big gold medal.
      - 4 days after the event
  - Vital documents are suppressed
    - Schools Chief Randy Dorn Announces Re-Election Bid.
      - In fact, it is 8 days after the announcement (annotation error)

Results (4)

Features	avg(P)	avg(R)	max(F(avg(P), avg(R)))	max(SU)
Baseline	0.288	0.953	0.442	0.267
+Time Range +Temporal Feature	0.367	0.743	0.491	0.367

- Examples
  - Non-vital documents are suppressed
    - Incumbent Peter Goldmark will face Republican challenger Clint Didier …
    - Prosecutor Dan Satterberg characterized the judge's Thursday ruling as "simply wrong.“
  - Vital documents are suppressed
    - This included casting the majority of their ballots against Democrat Peter Goldmark for …

Results (5)

Features	avg(P)	avg(R)	max(F(avg(P), avg(R)))	max(SU)
Baseline	0.288	0.953	0.442	0.267
+Time Range +Temporal Feature +Title/Profession Feature	0.378	0.744	0.501	0.377

- Examples
  - Non-vital documents are suppressed
    - Bill Templeton, an organizer for the local chapter of Pennsylvania Association of Sustainable Agriculture (PASA)
      - Target entity: Lions coach Bill Templeton
    - Archaeologist Richard Hansen talks about a newly discovered Mayan panel March 7, 2009, in Guatemala's Peten jungle.
      - Target entity: Mayor Rick Hansen (Richard Hansen is one alias)
  - Vital documents are suppressed
    - Pierce County judges, lawyers aim to keep Mark Lindquist's clout in check Old bridge toll bills will be in mail soon
      - Target entity: Pierce County Prosecutor Mark Lindquist (no title around the mention)

Results (6)

Features	avg(P)	avg(R)	max(F(avg(P), avg(R)))	max(SU)
Baseline	0.288	0.953	0.442	0.267
+Time Range +Temporal Feature +Title/Profession Feature +Action Patterns	0.447	0.702	0.546	0.464

- Examples
  - Non-vital documents are suppressed
    - Ryan: Randy Dorn says a ton of stupid
    - Ted Sturdevant was happy with the ruling
    - Washington State Commissioner of Public Lands Peter Goldmark, member of the decision-making committee!
  - Vital documents are suppressed
    - Washington schools Superintendent Randy Dorn says the state is getting carried away with the exams high school students are required to pass before they graduate.

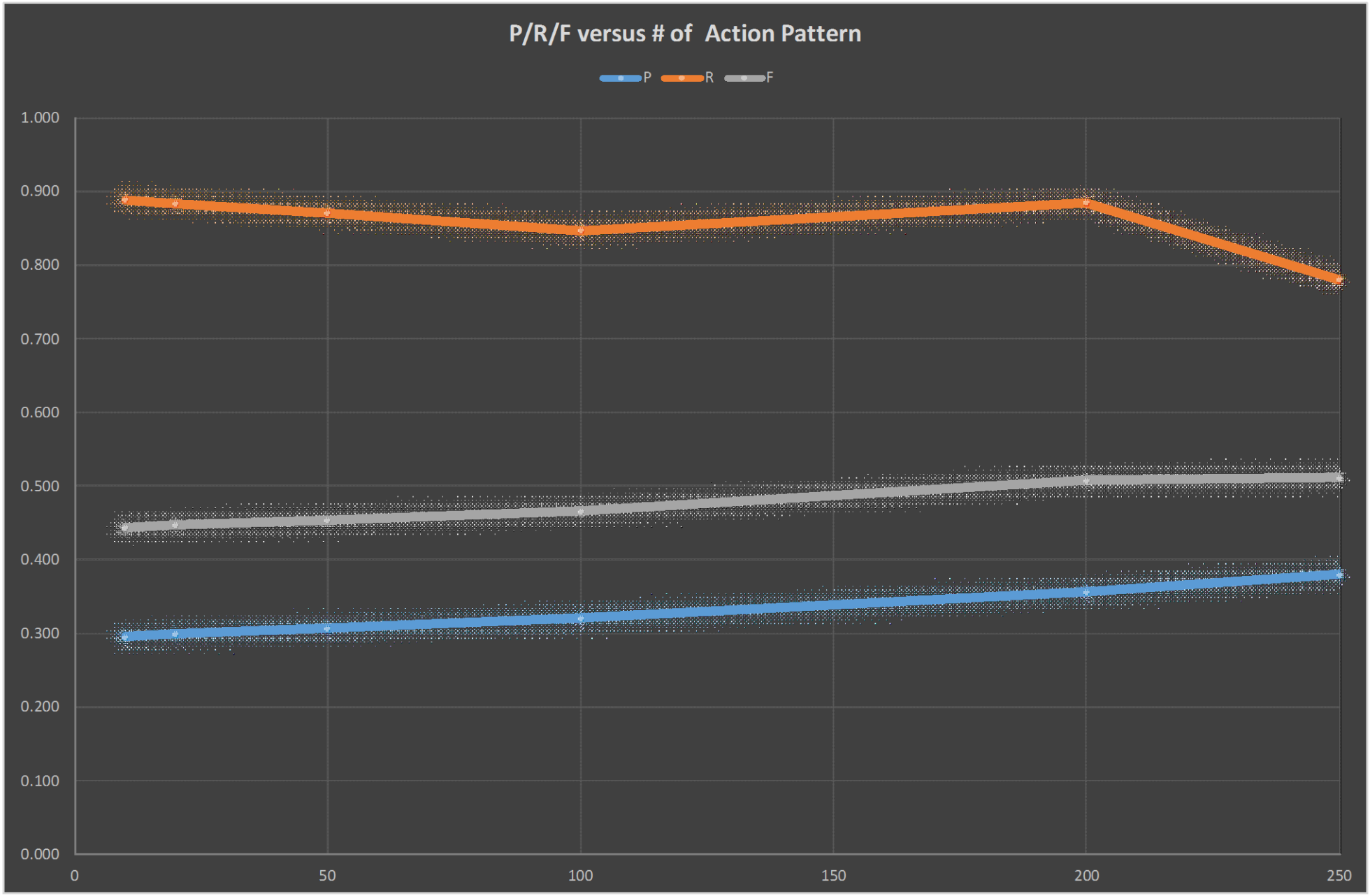
Error Analysis

- New patterns in test data (~70%)
  - These patterns do not exist in training data
    - Five candidates ran for the seat now held by state Rep. Andy Billig.
    - Jeff Mangum Extends Tour January 18, 2013 at 07:17 a.m.
- Inconsistent annotation (~20%)
  - Some documents' annotation violates the 1~3 days guideline
    - E.g., Schools Chief Randy Dorn Announces Re-Election Bid.
      - 8 days after the event, but it is also annotated as vital
    - E.g., Randy Dorn Issues Statement on Legalized Weed
      - Several hours after the event, but it is annotated as non-vital
  - Documents with the same content are annotated differently
    - E.g., one document has two same snapshots in the corpus, but one is vital while the other is non-vital

Error Analysis & Future Work

- Patterns' meaning changed (~10%)
  - Patterns related to vital in training data are related to non-vital in test data
    - E.g., Kennewick coach Bill Templeton said
      - It is related to all vital documents in training data, but most documents with this pattern are non-vital in test data
- Future Work
  - Revisit annotation guideline and annotations
  - Automatic mining and selection of action patterns (semantic feature helped!)
  - Scale to more entity types
  - New features
  - Great IR + Great NLP
  - Temporal summarization, dynamic domain, or KBP

The End


P/R/F versus # of Action Pattern
P
R
F
1.000
0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
0
50
100
150
200
250