

# Identifying Authorships of very Short Texts using Flexible Patterns

Roy Schwartz<sup>+</sup>, Oren Tsur<sup>+</sup>,  
Ari Rappoport<sup>+</sup> and Moshe Koppel<sup>\*</sup>

<sup>+</sup>The Hebrew University, <sup>\*</sup>Bar Ilan University  
ICRI-CI Retreat, May 2014



# Agenda

- Our goal is to gain semantic knowledge about the world
  - The *sky* is **blue**
  - “to *kick* the bucket” does not involve kicking anything
  - “Although many people think iphone 5 is a *great* device, I wonder if it’s that *good*” is a **negative** review
- We have previously shown that ***flexible patterns*** are useful for extracting semantic information
- We apply this technology to a new task – identifying the author of a very short text

# Flexible Patterns

- A generalization of word n-grams
  - Capture potentially unseen word n-grams
- Computed automatically from plain text
  - Language and domain independent
- Shown to be useful in various NLP applications
  - Extraction of semantic relationships (Davidov, Rappoport and Koppel, ACL 2007)
  - Detection of sarcasm (Tsur, Davidov and Rappoport, ICWSM 2010)
  - Sentiment analysis (Davidov, Tsur and Rappoport, Coling 2010)

# Flexible Patterns Examples

- “X and Y” indicates semantic similarity between X and Y:
  - apples and oranges
  - France and Canada
- “as X as Y” indicates that Y is X:
  - John is as clever as Mary
  - Cheetahs run as fast as racing cars
- “X can’t Y these Z. great!” indicates a sarcastic review
  - The Sony eBook can’t read these formats. Great!

# Authorship Attribution



• “To be, or not to be: that is the question”

• “Romeo, Romeo! wherefore art thou Romeo”

• ...“Love all, trust a few, do wrong to none.”

• “Taking a new step, uttering a new word, is what people fear most”

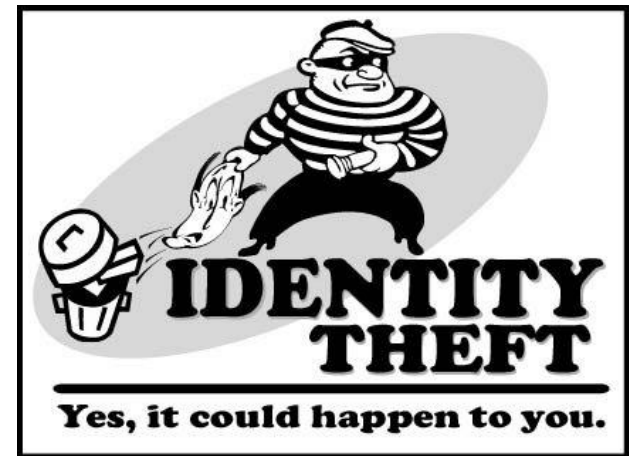
• “If they drive God from the earth, we shall shelter Him underground.”

• ...

• “Before all masters, necessity is the one most listened to, and who teaches the best.”

• “The Earth does not want new continents, but new men.”

# Authorship Attribution Applications



# History of Authorship Attribution

- Mendenhall, 1887
- Traditionally: long texts
- Recently: short texts
- Very recently: **very** short texts



# Tweets as Candidates for Short Text

- Tweets are limited to 140 characters
- Tweets are (relatively) self contained
- Compared to standard web data sentences
  - Tweets are shorter (14.2 words vs. 20.9)
  - Tweets have smaller sentence length variance (6.4 vs. 21.4)



# Experimental Setup

- Methodology
    - SVM with linear kernel, word n-gram, **flexible patterns** features
  - Experiments
    - Varying training set sizes, number of authors, recall-precision tradeoff
- Some Interesting Findings First**
- Results
    - 6.1% improvement over current state-of-the-art



# Interesting Finding

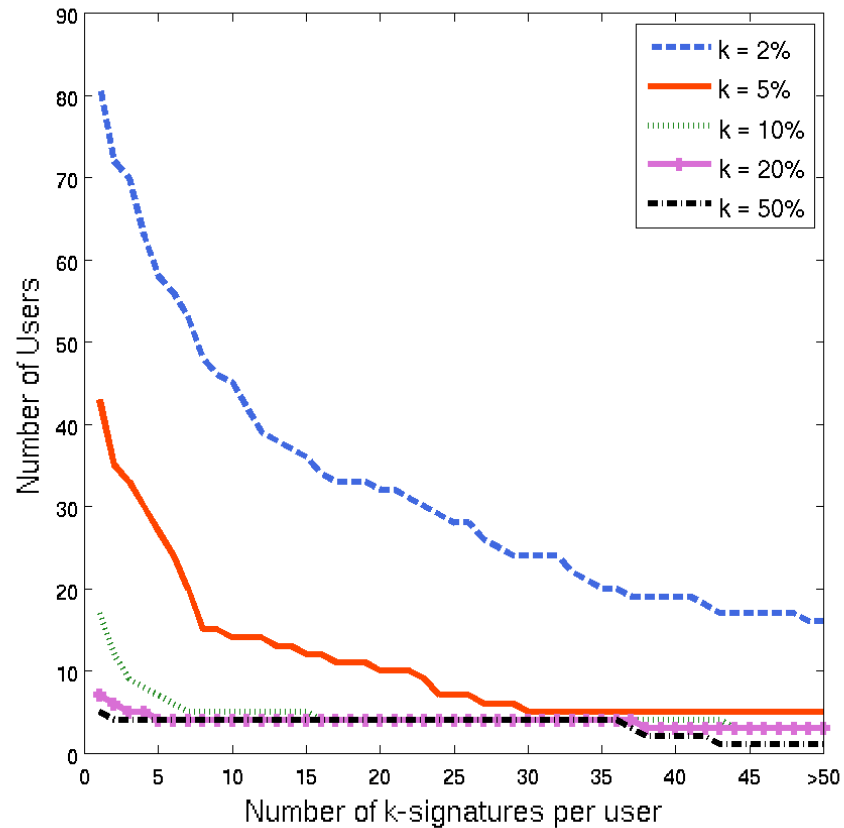
- Users tend to adopt a **unique style** when writing short texts
- K-signatures
  - A feature that is unique to a specific author  $A$
  - Appears in at least  $k\%$  of  $A$ 's training set, while not appearing in the more than  $0.5\%$  of the training set of any other user

# K-signatures Examples

Signature Type	10%-signature	Examples
Character n-grams	‘ ^ _ ^ ’	REF oh ok <u>^_</u> Glad you found it!
		Hope everyone is having a good afternoon <u>^_</u>
		REF Smirnoff lol keeping the goose in the freezer <u>^_</u>
	‘ <b>yew</b> ’	gurl <u>yew</u> serving me tea nooch
		REF about wen <u>yew</u> and ronnie see each other
		REF lol so <u>yew</u> goin to check out tini’s tonight huh???

# K-signatures per User

100 authors, 180 training tweets per author



# Structured Messages / Bots?

User	20%-signature	Examples
1	<b>I'm listening to :</b>	<b><u>I'm listening to:</u></b> Sigur R?s ? Intro: http://www.last.fm/music/Sigur+R%C3%B3s http://bit.ly/3XJHyb
		<b><u>I'm listening to:</u></b> Tina Arena ? In Command: http://www.last.fm/music/Tina+Arena http://bit.ly/7q9E25
		<b><u>I'm listening to:</u></b> Midnight Oil ? Under the Overpass: http://www.last.fm/music/Midnight+Oil http://bit.ly/7IH4cg
2	<b>news now ( str )</b>	#Hotel <b><u>News Now(STR)</u></b> 5 things to know: 27 May 2009: From the desks of the HotelNewsNow.com editor... http://bit.ly/aZTZOq #Tourism #Lodging
		#Hotel <b><u>News Now(STR)</u></b> Five sales renegotiating tactics: As bookings representatives press to renegot... http://bit.ly/bHPn2L
		#Hotel <b><u>News Now(STR)</u></b> Risk of hotel recession retreats: The Hotel Industry's Pulse Index increases... http://bit.ly/a8EKrm #Tourism #Lodging
3	<b>( NUM bids ) end date :</b>	NEW PINK NINTENDO DS LITE CONSOLE WITH 21 GIFTS + CASE: &#163;66.50 <b><u>(13 Bids) End Date:</u></b> Tuesday Dec-08-2009 17:.. http://bit.ly/7uPt6V
		Microsoft Xbox 360 Game System - Console Only - Working: US \$51.99 <b><u>(25 Bids) End Date:</u></b> Saturday Dec-12-2009 13:.. http://bit.ly/8VgdTv
		Microsoft Sony Playstation 3 (80 GB) Console 6 Months Old: &#163;190.00 <b><u>(25 Bids) End Date:</u></b> Sunday Dec-13-2009 21:21:39 G.. http://bit.ly/7kwtDS

# Methodology

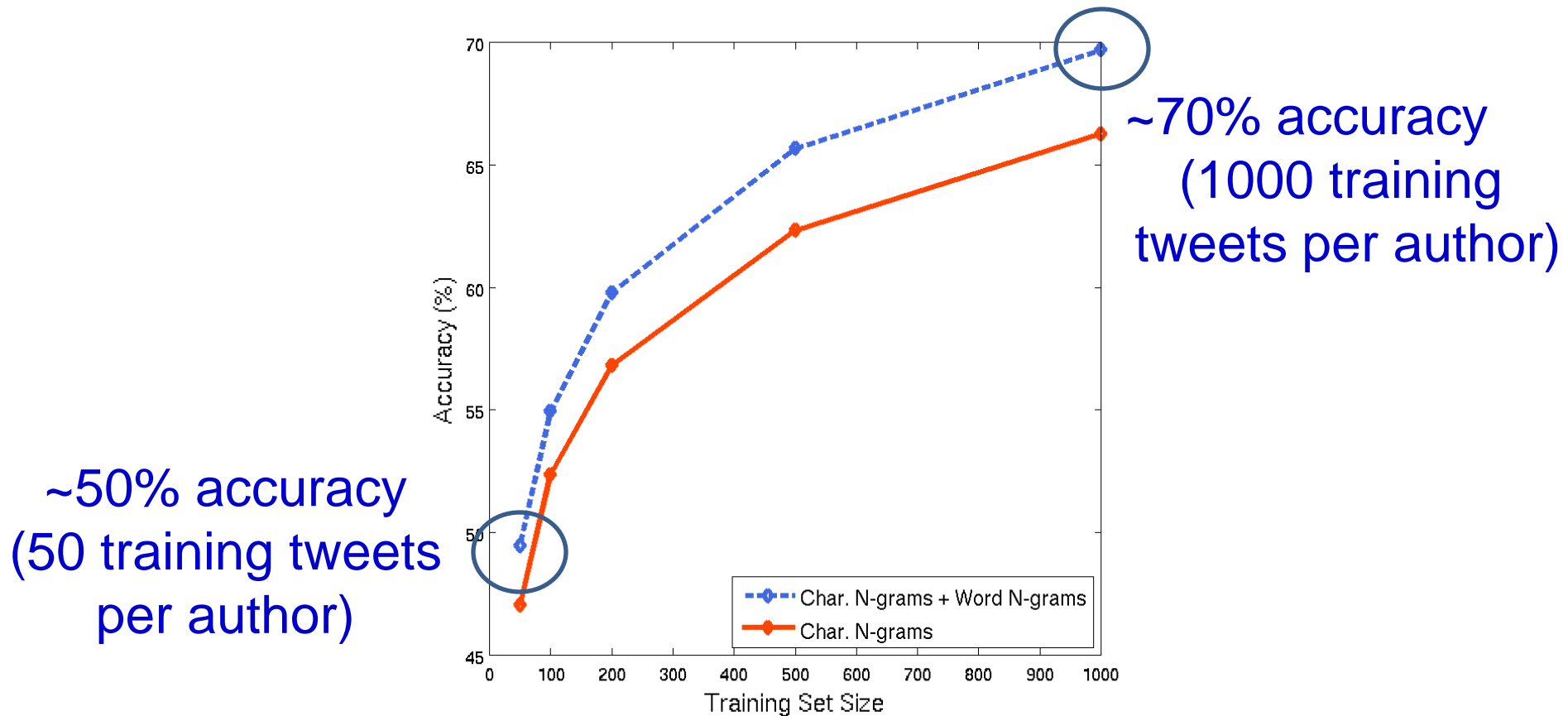
- Features
  - Character n-grams, word n-grams, **flexible patterns**
- Model
  - Multiclass SVM with a linear kernel

# Experiments

- Varying training set sizes
  - 10 groups of 50 authors each, 50-1000 training tweets per author
- Varying numbers of authors
  - 50-1000 authors, 200 training tweets per author
- Recall-precision tradeoff
  - “don’t know” option

# Varying Training Set Sizes

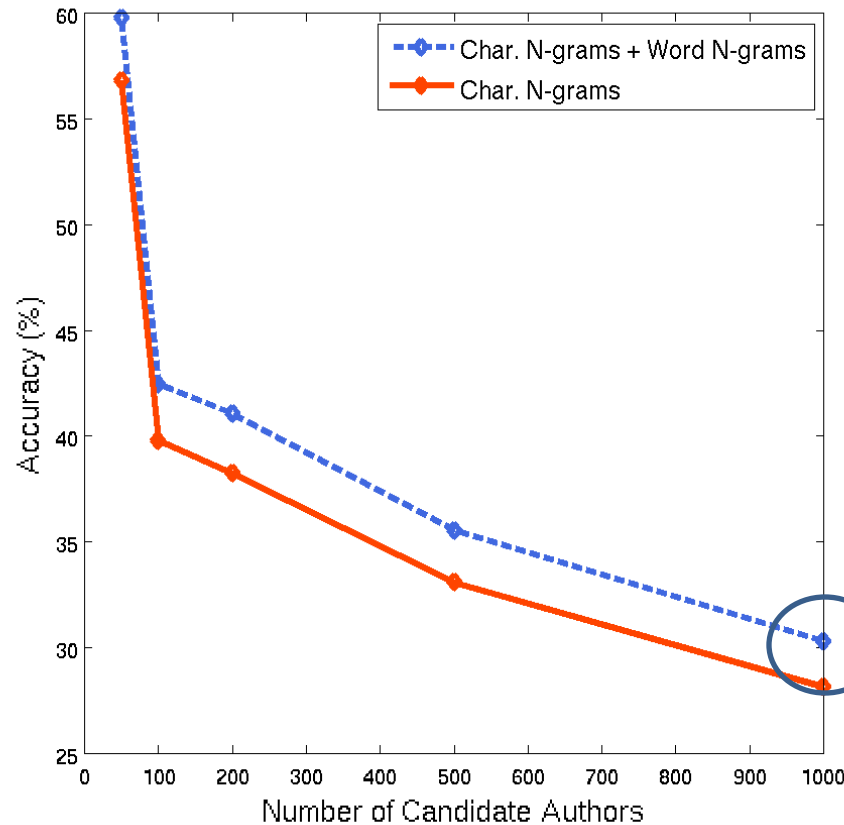
50 Authors (2% Random Baseline)





# Varying Numbers of Authors

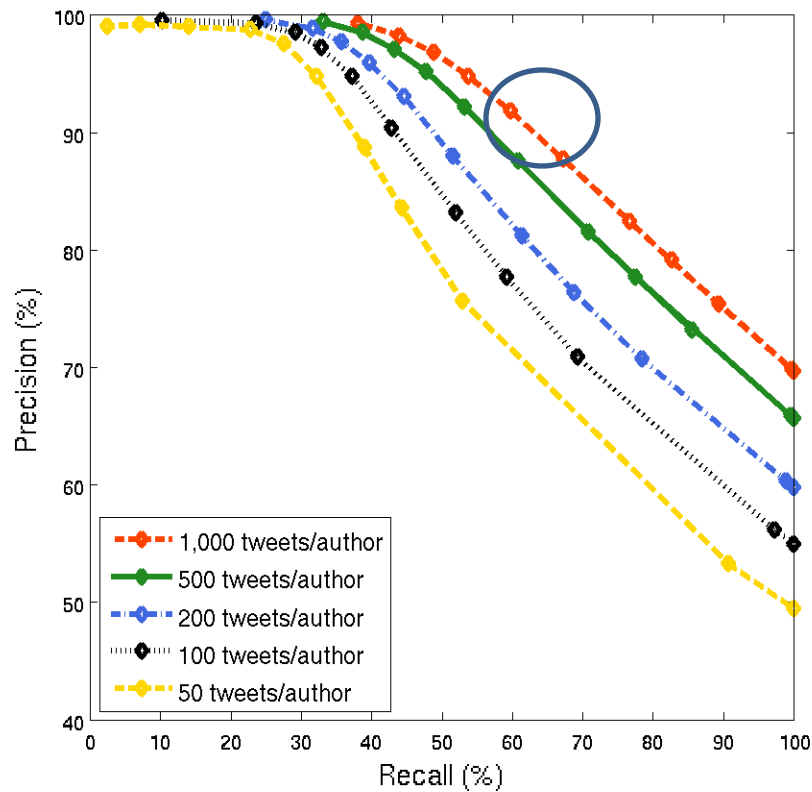
200 Training Tweets per Author



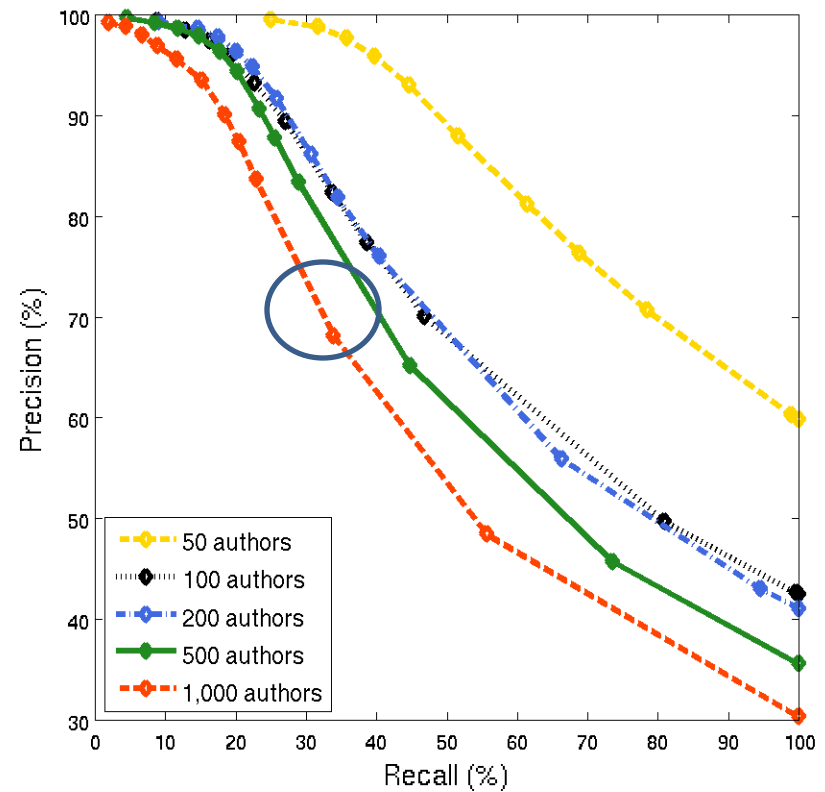
~30% accuracy  
(1000 authors,  
0.1% baseline)

# Recall-Precision Tradeoff

~90% precision,  
>~60% recall



~70% precision,  
~30% recall



# Flexible Patterns Features

- Examples of tweets written by the same author
  - “*the way I treated her*”
  - “*half of the things I’ve seen*”
  - “*the friends I have had for years*”
  - “*in the neighborhood I grew up in*”
- No word n-gram feature is able to capture this author’s style
- Author’s character n-grams (“the”, “ I ”) are unindicative

“the X I”

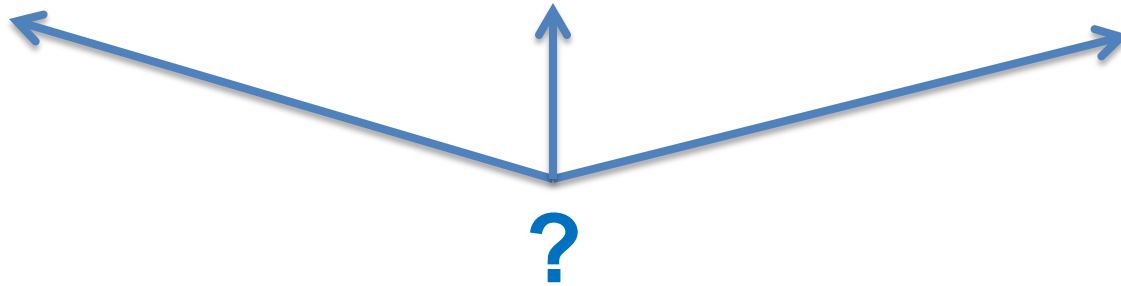
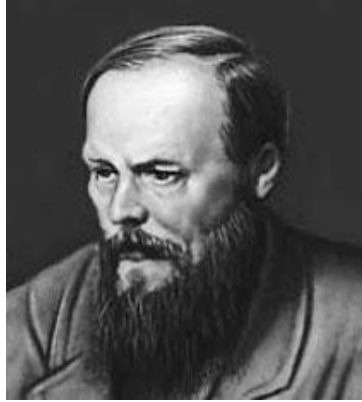
# Summary

- Accurate authorship attribution of very short texts
  - 6.1% improvement over current state-of-the-art
- Many authors use k-signatures in their writing of short texts
  - A partial explanation for our high-quality results
- Flexible patterns are useful authorship attribution features
  - Statistically significant improvement

# What's Next?

- Minimally supervised identification of semantic categories using flexible patterns
  - Animals, food, tools, ...
- Automatically obtain a complete semantic description of a concept
  - A **dog** is an *animal*, which *barks*, has a *tail*, is *faithful*, is related to *cats*, etc.

# Authorship Attribution



“Love all, trust a few, do wrong to none.”



*roys02@cs.huji.ac.il*

*<http://www.cs.huji.ac.il/~roys02/>*

