Novel representations and methods in text classification

Manuel Montes, Hugo Jair Escalante

Instituto Nacional de Astrofísica, Óptica y Electrónica, México.

http://ccc.inaoep.mx/~mmontesg/

http://ccc.inaoep.mx/~hugojair/

{mmontesg, hugojair}@inaoep.mx

7th Russian Summer School in Information Retrieval
Kazan, Russia, September 2013

Novel represensations and methods in text classification

ENHANCING THE BOW WITH SEQUENTIAL INFORMATION

Outline

- Bag of words
- Extensions to incorporate sequential information
  - Maximal frequent sequences
  - Sequential patterns
  - The LOWBOW framework
- Text categorization under LOWBOW
- Authorship attribution with LOWBOW

Bag of words

- Under the bag-of-words framework a document is represented by the **set of terms** that appear in it

- By definition, BOW is an orderless representation

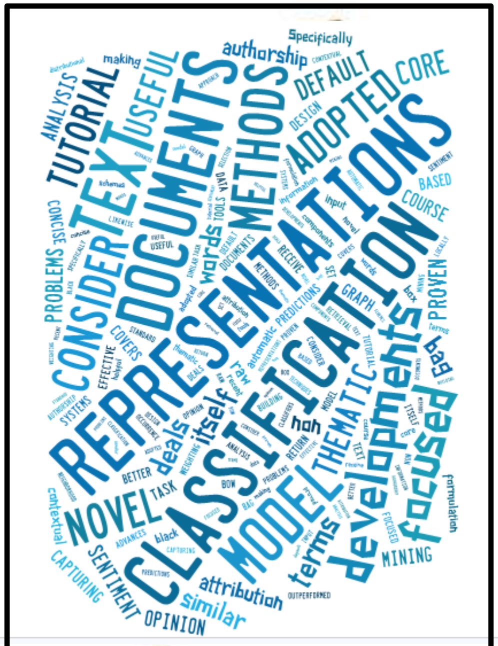
Novel representations and methods in text classification

Manuel Montes-y-Gómez & Hugo Jair Escalante

Two core components of any classification system are the adopted representation for documents and the classification model itself. This tutorial deals with recent advances and developments on both components. The default representation for documents in text classification is the bag-of-words(BOW), where weighting schemes similar to those used in information retrieval are adopted.Whereas this representation has proven to be very helpful for thematic text classification, in novel, non-thematic text classification problems (e.g., authorship attribution, sentiment analysis and opinion mining, etc.), the standard BOW can be outperformed by other advanced representations.

This course is focused on three document representations that have proved to be useful for capturing more information than the raw occurrence of terms in documents as in BOW. The considered representations are: locally weighted BOW, distributional term representations,concise representations and graph-based representations. Likewise, the tutorial covers recent developments in the task of building classification models. Specifically, we consider contextual classification techniques and full model selection methods. The former approach is focused in the design of classifiers that consider the neighborhood of a document for making better predictions.

The latter formulation focuses in the development of automatic methods for building classification systems, that is, black box tools that receive as input a data set and return a very effective classification model.



Bag of words

- Under the bag-of-words framework a document is represented by the **set of terms** that appear in it

- By definition, BOW is an orderless representation

Yo me rio en el baño
(I am laughing at the bathroom)

ahora	…	baño	…	en	…	me	…	río	…	zorro
0	0	1	0	1	0	1	0	1	0	0

Yo me baño en el río
(I am taking a shower at the river)

ahora	…	baño	…	en	…	me	…	río	…	zorro
0	0	1	0	1	0	1	0	1	0	0

Same BoW representation different meaning

Bag of words

- There have been several efforts trying to incorporate sequential information into BoW-based representations
  - **Ngrams:** Terms are defined as sequences of characters or words
  - **Maximal frequent sequences:** Frequent sequences of words are discovered (with/without gaps)
  - **Phrase patterns:** Sequential data mining is applied to detect sequential patterns (with gaps)
  - **Methods based on linguistic analyses:** POS tagging, syntactic trees, etc.

Bag of Ngrams

- An Ngram is a sequence of N-terms (e.g., words / characters):
  - Russian-federation / bag-of-words / in-god-we-trust ...
  - the / mex / lol / wtf ...

- A sliding window is applied to the documents, all Ngrams found in the corpus form the vocabulary

- Documents are represented by the bag of Ngrams that they contain

Document: Russian Summer School in Information Retrieval

Unigrams	Bigrams	Tri-grams
Russian, Summer School, in, Information, Retrieval	**Russian-summer,** summer-school, school-in, in-information, **information-retrieval**	**Russian-summer-school**, **Summer-School-in,** School-in-Information, **in-Information-Retrieval**

Bag of Ngrams

- An Ngram is a sequence of N-terms (e.g., words / characters):

✓ Ngrams capture low-range sequential information

✗ Fixed length patterns (usually n≤5);

✓ Satisfactory results have been reported in non-thematic tasks

✗ The size of the vocabulary increases dramatically

✗ No significant improvements over standard BOW

✓ When using characters, they can capture *style* aspects

Bag of Ngrams

- An Ngram is a sequence of N-terms (e.g., words / characters):

✓ Ngrams capture low-range sequential information

✗ Fixed length patterns (usually n≤5);

Skyp-grams: Extension to Ngrams that allows us to consider gaps between terms to build Ngrams. Example:

Russian Summer School in Information Retrieval

Bigrams	2-skyp-bigrams
Russian-summer, summer-school, school-in, in-information, information-retrieval	Russian-school, Russian-in, Summer-in, Summer-Information, School-information, in-retrieval

Increases the range of sequential information, but augments the vocabulary size

Maximal frequent sequences

- Each document is seen a sequence of words (items)

- The goal is to identify *interesting* sequences of words that can be used to characterize documents, e.g.:

Russian-School-Information-Retrieval

✓ No fixed-length constraints are imposed (as in n-grams)

✓ Reduce overlapping information in the representation

✓ Gaps are allowed in sequences

Maximal frequent sequences

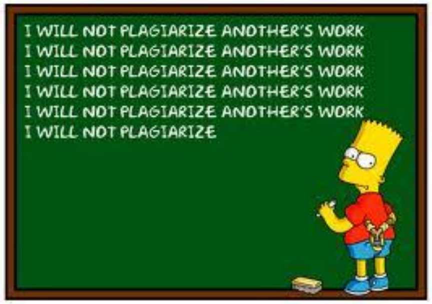
- Definitions:
  - A sequence $p = p_1,\ldots,p_k$ is a subsequence of another sequence $q = q_1,\ldots,q_m$ if all of the items p_i $1 \le i \le k$, occur in q and they occur in the same order as in p
  - A sequence p is frequent in document collection D if p is a subsequence of at least σ documents in D
  - A sequence p is a maximal frequent sequence in D if there does not exist any sequence p' in D such that p is a subsequence of p' and p' is frequent in D

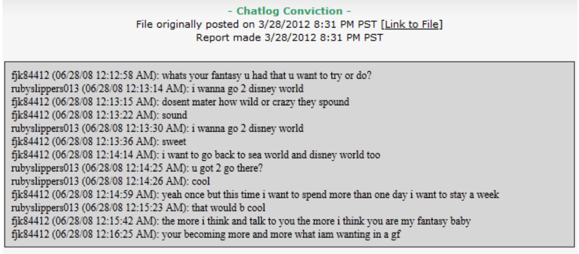
- There are *efficient algorithms* to identify all of the MFS

MFS for authorship attribution

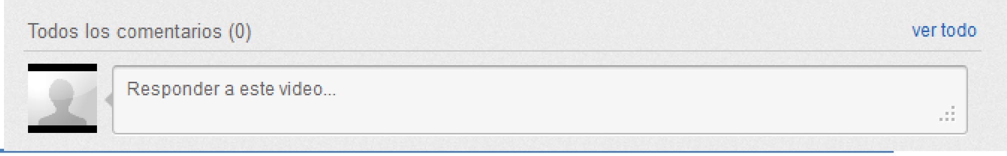
- **Authorship attribution:** Given texts of uncertain authorship and texts written by a set of candidate authors, the task is to map the uncertain texts onto their true authors among the candidates.

- Applications include: fraud detection, spam filtering, computer forensics and plagiarism detection



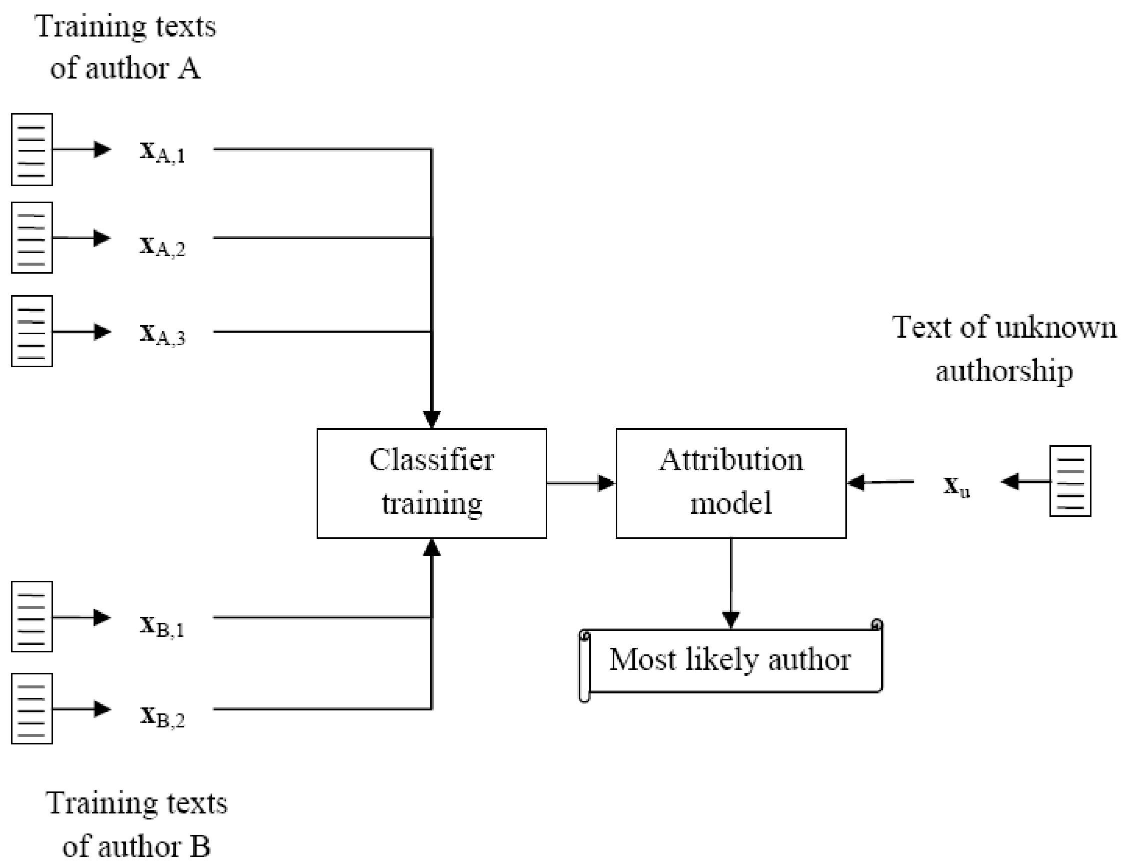

- Chatlog Conviction -
File originally posted on 3/28/2012 8:31 PM PST [Link to File]
Report made 3/28/2012 8:31 PM PST

fjk84412 (06/28/08 12:12:58 AM): whats your fantasy u had that u want to try or do?
rubyslippers013 (06/28/08 12:13:14 AM): i wanna go 2 disney world
fjk84412 (06/28/08 12:13:15 AM): dosent mater how wild or crazy they spound
fjk84412 (06/28/08 12:13:22 AM): sound
rubyslippers013 (06/28/08 12:13:30 AM): i wanna go 2 disney world
fjk84412 (06/28/08 12:13:36 AM): sweet
fjk84412 (06/28/08 12:14:14 AM): i want to go back to sea world and disney world too
rubyslippers013 (06/28/08 12:14:25 AM): u got 2 go there?
rubyslippers013 (06/28/08 12:14:26 AM): cool
fjk84412 (06/28/08 12:14:59 AM): yeah once but this time i want to spend more than one day i want to stay a week
rubyslippers013 (06/28/08 12:15:23 AM): that would b cool
fjk84412 (06/28/08 12:15:42 AM): the more i think and talk to you the more i think you are my fantasy baby
fjk84412 (06/28/08 12:16:25 AM): your becoming more and more what iam wanting in a gf



MFS for authorship attribution

- Instance-based approach to authorship attribution (~text categorization)



Maximal frequent sequences

- MSF for authorship attribution

Let D_T be the set of labeled documents that will be used for training
Let d be an anonymous document

TRAINING

1. Set the value of the frequency threshold $\sigma = 2$
2. Set the feature set $F_1 = \{\varnothing\}$
3. DO
   a. Enumerate all maximal frequent word sequences in D_T corresponding to the frequency threshold σ. Name the set of sequences S_σ
   b. Integrate new sequences to the feature set, i.e., $F_\sigma = F_{\sigma\text{-}1} \cup S_\sigma$
   c. Increment the frequency threshold; i.e., $\sigma = \sigma + 1$

   WHILE ($S_{\sigma\text{-}1}$ contain at least one sequence of two or more words not included in $F_{\sigma\text{-}2}$)
4. Build the training instances using the discovered Boolean features
5. Give the learning algorithm the training instances and perform training

CLASSIFICATION

1. Build the representation of d in accordance to the training feature space
2. Let the trained classifier label the new instance

Maximal frequent sequences

- Identify authors of poems written by different mexican poets

Poets	Number of documents	Size of Vocabulary	Number of Phrases	Average Words by Documents	Average Phrases by Documents
Efraín Huerta	48	3831	510	236.5	22.3
Jaime Sabines	80	3955	717	155.8	17.4
Octavio Paz	75	3335	448	162.6	27.2
Rosario Castellanos	80	4355	727	149.3	16.4
Rubén Bonifaz	70	4769	720	178.3	17.3

Features	Accuracy	Average Precision	Average Recall
Functional words	41.0%	0.42	0.39
Content words	73.0%	0.78	0.73
All kind of words	73.0%	0.78	0.74
n-grams (unigrams plus bigrams)	78.8%	0.84	0.79
n-grams (from unigrams to trigrams)	76.8%	0.84	0.77

- Baseline results

Poets	Precision	Recall
Efraín Huerta	1.00	0.75
Jaime Sabines	0.83	0.83
Octavio Paz	0.95	0.75
Rosario Castellanos	0.65	0.91
Ruben Bonifaz	0.94	0.87
Average Rates	0.87	0.82
Overall Accuracy	83%	

- Maximal frequent sequences approach

Maximal frequent sequences

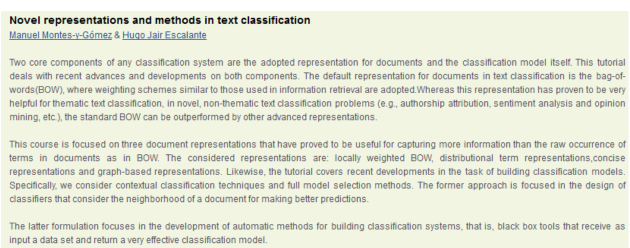
- MFS can discover interesting and useful patterns, however, extracting all of the MFS is a time consuming process

- MFS do not exploit information about the labels in training documents (it is an unsupervised method)

- *Informativeness* of patterns heavily depends on the frequency threshold σ

Phrase patterns

- A text is considered an ordered list of sentences, where each sentence is an unordered set of words
- The goal is to identify interesting *sequences of sets of words*. The order is at the sentence level
- Sequential patterns are extracted per each category


Novel representations and methods in text classification
Manuel Montes-y-Gómez & Hugo Jair Escalante

Two core components of any classification system are the adopted representation for documents and the classification model itself. This tutorial deals with recent advances and developments on both components. The default representation for documents in text classification is the bag-of-words(BOW), where weighting schemes similar to those used in information retrieval are adopted.Whereas this representation has proven to be very helpful for thematic text classification, in novel, non-thematic text classification problems (e.g., authorship attribution, sentiment analysis and opinion mining, etc.), the standard BOW can be outperformed by other advanced representations.

This course is focused on three document representations that have proved to be useful for capturing more information than the raw occurrence of terms in documents as in BOW. The considered representations are: locally weighted BOW, distributional term representations,concise representations and graph-based representations. Likewise, the tutorial covers recent developments in the task of building classification models. Specifically, we consider contextual classification techniques and full model selection methods. The former approach is focused in the design of classifiers that consider the neighborhood of a document for making better predictions.

The latter formulation focuses in the development of automatic methods for building classification systems, that is, black box tools that receive as input a data set and return a very effective classification model.

Novel-representations text-classification
Representation-for-documents
Authorship-attribution
Return-a-effective-classification model

Phrase patterns

- Similar to MFS: Sequential patterns aim at discovering temporal relations between items (words) in a database (corpus)

- Main idea: extending work on mining association rules to extract meaningful sequential patterns

Mining association rules	Text categorization
Client	Text
Item	Word
Items/transaction	Sentence (set of words)
Data	Position of the sentence in document

Phrase patterns

- Let $s = s_1, \ldots, s_k$ be a sequence, the support of *s* is defined as:

$$sp(s) = \frac{\#texts \quad matching \quad s}{\#texts}$$

- Sequences with a support higher than *minsup* are considered for the next step. Frequent patterns are used to generate rules of the form:

$$\gamma :< s_1, \ldots, s_k > \rightarrow C_i$$

- The confidence of a frequent pattern is defined as follows:

$$conf(\gamma) = \frac{\#text - from - C_i \quad matching \quad < s_1, \ldots, s_k >}{\#text \quad matching \quad < s_1, \ldots, s_k >}$$

- Classification is done with a KNN scheme over rules with highest confidence

Phrase patterns

- Interesting patterns can be obtained with this formulation
- Class-information is considered in obtaining sequential rules
- Similar results to BOW using SVMs
- A large number of rules can be obtained and (as with MFS) extracting sequential patterns is a time consuming process

The locally weighted bag-of-words framework

- **LOWBOW:** an attempt to enrich BoW representations with sequential information *without defining/generating new terms/patters*

- Each document is represented by a set of local histograms computed across the whole document but smoothed by kernels and centered at different document locations

- LOWBOW-based document representations can preserve sequential information in documents

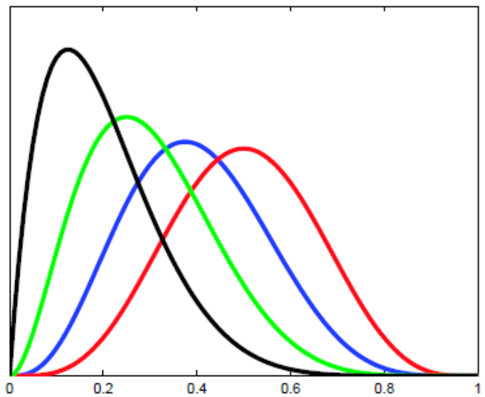
The locally weighted bag-of-words framework

- A document is a sequence of N words, it can be seen as a categorical time series:

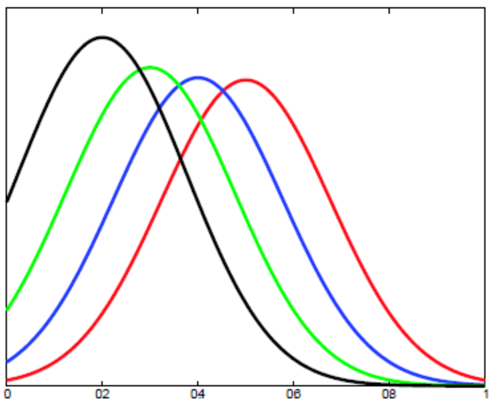
$$d_i = \langle d_{i,1}, \ldots, d_{i,N} \rangle \quad \text{with} \quad d_{i,j} \in V$$

- **Idea:** smooth temporarily this categorical times series with a Kernel: $K_{\mu,\sigma}(x)$

$$K_{\mu,\sigma}(x) = Beta(x; \beta\frac{\mu}{\sigma}, \beta\frac{1-\mu}{\sigma})$$



$$K_{\mu,\sigma}(x) = \begin{cases} \frac{N(x;\mu,\sigma)}{\Theta((1-\mu)/\sigma) - \Theta(\mu/\sigma)} & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$



The locally weighted bag-of-words framework

- Let: $$\delta_c(d_i)(k, j) = \begin{cases} \dfrac{c}{1+c\,|V\,|} & d_{i,k} \neq j \\ \dfrac{1+c}{1+c\,|V\,|} & d_{i,k} = j \end{cases}$$

  Denote the weight of term *j* at position *k* of document *i, for k a subset of locations at the documents*

- The LOWBOW representation of the word sequence d_i is:

  $$\mathrm{Y}(d_i) = \{\mathrm{Y}\mu(d_i) : \mu \in [0,1]\}$$

  where $\mathrm{Y}\mu(d_i)$ is the local word histogram at μ defined by

  $$[\mathrm{Y}\mu(d_i)] = \int_0^1 \varphi(\delta_c(d_i))(t, j) K\mu, \sigma(t)\,dt$$

The locally weighted bag-of-words framework

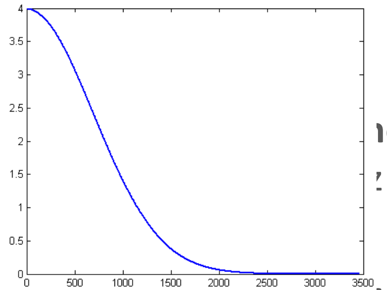
Identify locations in documents

Novel representations and methods in text classification
Manuel Montes-y-Gómez & Hugo Jair Escalante

Two core components of any classification system are the adopted representation for documents and the classification model itself. This tutorial deals with recent advances and developments on both components. The default representation for documents in text classification is the bag-of-words(BOW), where weighting schemes similar to those used in information retrieval are adopted. Whereas this representation has proven to be very helpful for thematic text classification, in novel, non-thematic text classification problems (e.g., authorship attribution, sentiment analysis and opinion mining, etc.), the standard BOW can be outperformed by other advanced representations.

This course is focused on three document representations that have proved to be useful for capturing more information than the raw occurrence of terms in documents as in BOW. The considered representations are: locally weighted BOW, distributional term representations, concise representations and graph-based representations. Likewise, the tutorial covers recent developments in the task of building classification models. Specifically, we consider contextual classification techniques and full model selection methods. The former approach is focused in the design of classifiers that consider the neighborhood of a document for making better predictions. The latter formulation focuses in the development of automatic methods for building classification systems, that is, black box tools that receive as input a data set and return a very effective classification model.

The locally weighted bag-of-words framework



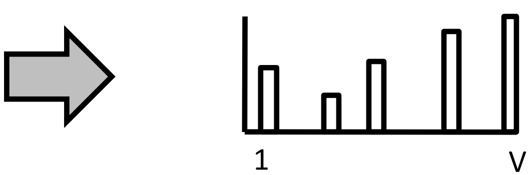
N... nd methods in text classification

M... z & Hugo Jair Escalante

Two core components of any classification system are the adopted representation for documents and the classification model itself. This tutorial deals with recent advances and developments on both components. The default representation for documents in text classification is the bag-of-words(BOW), where weighting schemes similar to those used in information retrieval are adopted. Whereas this representation has proven to be very helpful for thematic text classification, in novel, non-thematic text classification problems (e.g., authorship attribution, sentiment analysis and opinion mining, etc.), the standard BOW can be outperformed by other advanced representations.

This course is focused on three document representations that have proved to be useful for capturing more information than the raw occurrence of terms in documents as in BOW. The considered representations are: locally weighted BOW, distributional term representations, concise representations and graph-based representations. Likewise, the tutorial covers recent developments in the task of building classification models. Specifically, we consider contextual classification techniques and full model selection methods. The former approach is focused in the design of classifiers that consider the neighborhood of a document for making better predictions. The latter formulation focuses in the development of automatic methods for building classification systems, that is, black box tools that receive as input a data set and return a very effective classification model.

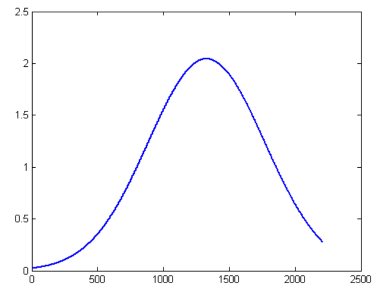
Weight the contribution of terms according to Gaussians at the different locations



The locally weighted bag-of-words framework

Novel representations and methods in text classification

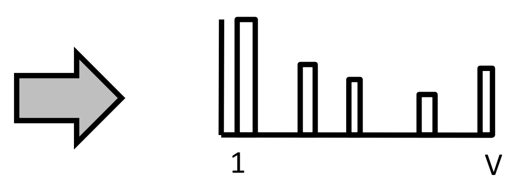
Manuel Montes-y-Gómez & Hugo Jair Escalante



Two core components of any cla the adopted representation for documents and the classification model itse vith recent advances and developments on both components. The default uments in text classification is the bag-of-words(BOW), where weighting s e used in information retrieval are adopted. Whereas this representation has ful for thematic text classification, in novel, non-thematic text classification orship attribution, sentiment analysis and opinion mining, etc.), the standa ormed by other advanced representations.

This course is focused on three document representations that have proved to be useful for capturing more information than the raw occurrence of terms in documents as in BOW. The considered representations are: locally weighted BOW, distributional term representations, concise representations and graph-based representations. Likewise, the tutorial covers recent developments in the task of building classification models. Specifically, we consider contextual classification techniques and full model selection methods. The former approach is focused in the design of classifiers that consider the neighborhood of a document for making better predictions. The latter formulation focuses in the development of automatic methods for building classification systems, that is, black box tools that receive as input a data set and return a very effective classification model.

Weight the contribution of terms according to Gaussians at the different locations

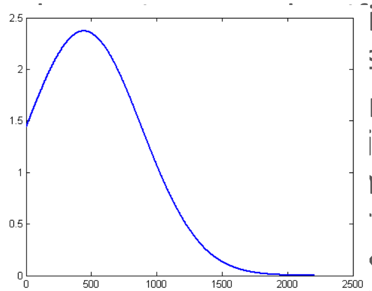


The locally weighted bag-of-words framework

Novel representations and methods in text classification

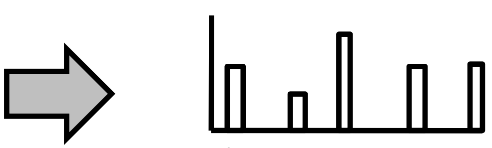
Manuel Montes-y-Gómez & Hugo Jair Escalante

Two core components of any classification system are the adopted representation for documents and the classification model itself. This tutorial deals with recent advances and developments on both components. The default representation for documents in text classification is the bag-of-words(BOW), where weighting schemes similar to those used in information retrieval are adopted. Whereas this representation has proven to be very helpful for thematic text classification, in novel, no ication problems (e.g., authorship attribution, sentiment analysis and op standard BOW can be outperformed by other advanced representations.



Thi n three document representations that have proved to be useful for ca ion than the raw occurrence of terms in documents as in BOW. The co s are: locally weighted BOW, distributional term representations, concise re aph-based representations. Likewise, the tutorial covers recent de k of building classification models. Specifically, we consider contextual cla and full model selection methods. The former approach is focused in the design of classifiers that consider the neighborhood of a document for making better predictions. The latter formulation focuses in the development of automatic methods for building classification systems, that is, black box tools that receive as input a data set and return a very effective classification model.

Weight the contribution of terms according to Gaussians at the different locations



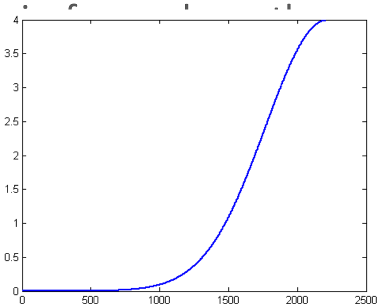
The locally weighted bag-of-words framework

Novel representations and methods in text classification

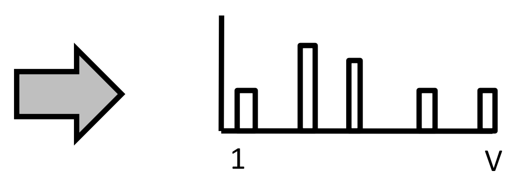
Manuel Montes-y-Gómez & Hugo Jair Escalante

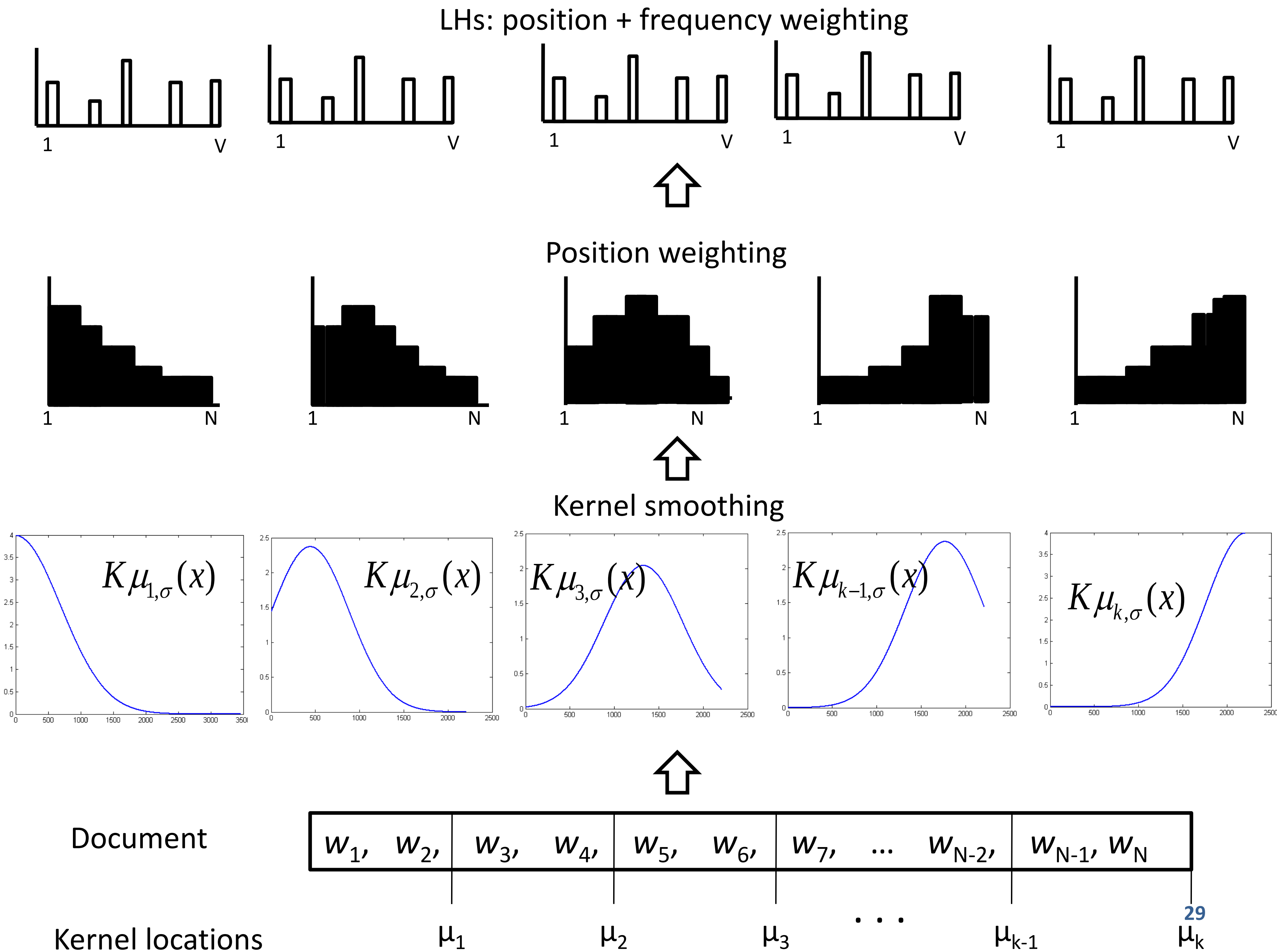
Two core components of any classification system are the adopted representation for documents and the classification model itself. This tutorial deals with recent advances and developments on both components. The default representation for documents in text classification is the bag-of-words(BOW), where weighting schemes similar to those used in information retrieval are adopted. Whereas this representation has proven to be very helpful for thematic text classification, in novel, non-thematic text classification problems (e.g., authorship attribution, sentiment analysis and opinion mining, etc.), the standard BOW can be outperformed by other advanced representations.

This course documen representations that have proved to be useful for capturing the raw occurrence of terms in documents as in BOW. The considered ally weighted BOW, distributional term representations, concise representat d representations. Likewise, the tutorial covers recent developme ling classification models. Specifically, we consider contextual classificatic odel selection methods. The former approach is focused in the design of c he neighborhood of a document for making better predictions. The latter f e development of automatic methods for building classification systems, that is, black box tools that receive as input a data set and return a very effective classification model.



Weight the contribution of terms according to Gaussians at the different locations

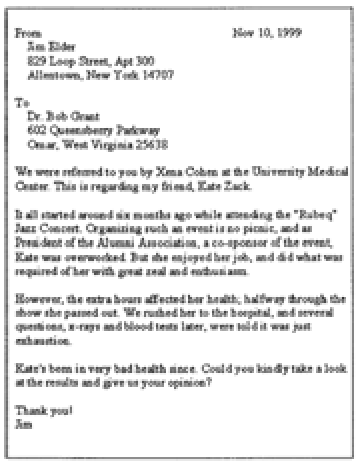




LHs: position + frequency weighting
1
V
Position weighting
1
N
Kernel smoothing
$K\mu_{1,\sigma}(x)$
$K\mu_{2,\sigma}(x)$
$K\mu_{3,\sigma}(x)$
$K\mu_{k-1,\sigma}(x)$
$K\mu_{k,\sigma}(x)$
Document
$w_1, \; w_2, \; w_3, \; w_4, \; w_5, \; w_6, \; w_7, \; \ldots \; w_{N-2}, \; w_{N-1}, w_N$
Kernel locations
μ_1
μ_2
μ_3
μ_{k-1}
μ_k

The locally weighted bag-of-words framework

- A set of histograms, each weighted according to selected positions in the document



From
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

Nov 10, 1999

To
Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

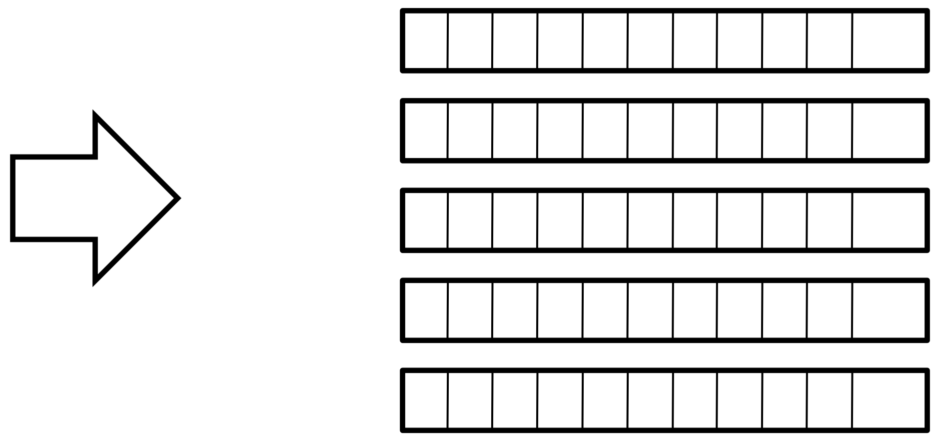
We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim



$$\mathbf{d}_i = \{\mathbf{dl}_i^1, \ldots, \mathbf{dl}_i^k\}$$

$$\mathbf{dl}_i^j = \mathbf{d}_i \times K_{\mu_j,\sigma}^s$$

The locally weighted bag-of-words framework

- Standard bag-of-words:

From
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

Nov 10, 1999

To
Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

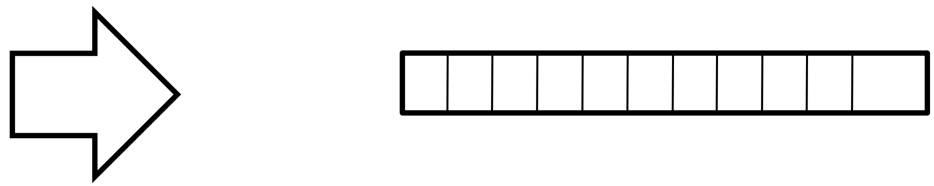
We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim

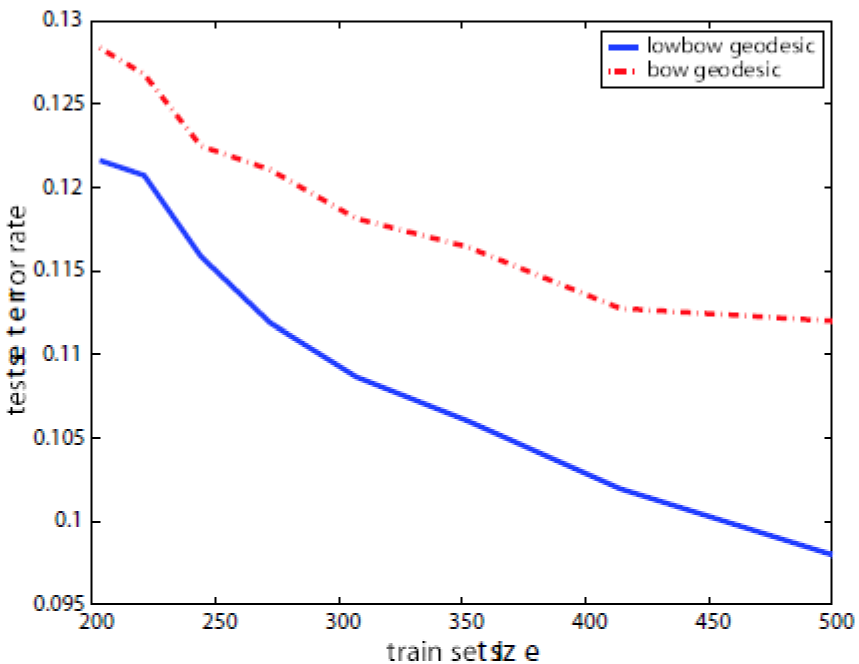


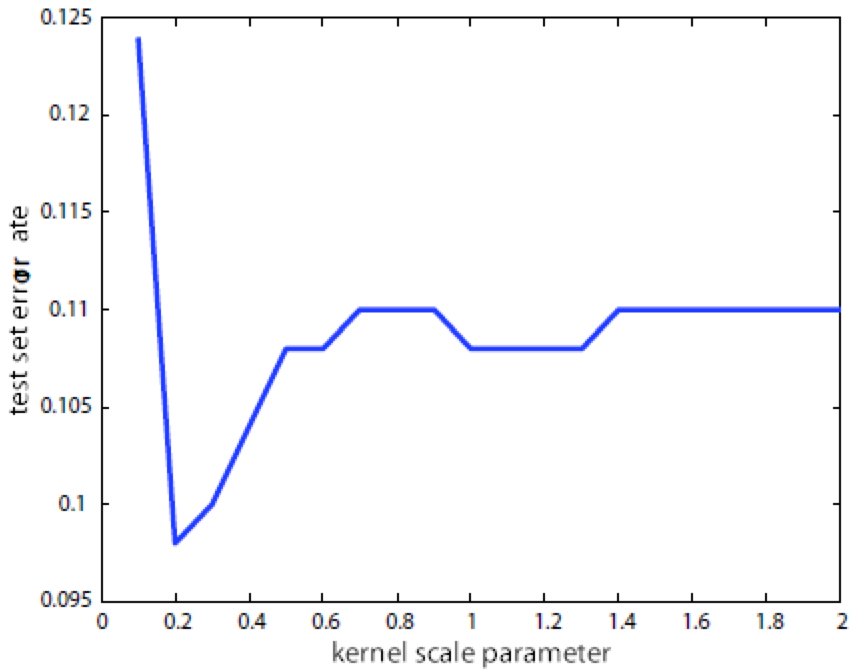
$$\mathbf{d}_i = \left[x_{i,1}, \ldots, x_{i,|V|} \right]$$

The locally weighted bag-of-words framework

- Documents represented under LOWBOW can be used for text categorization, using an appropriate distance measure (e.g.):

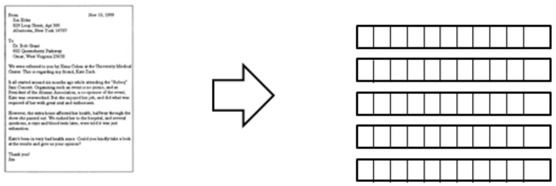
$$D(\theta,\eta) = \arccos\left(\sum_{i=1}^{m}\sqrt{\theta_i\eta_i}\right) \qquad \theta,\eta \in \mathbf{P}_{m+1}$$

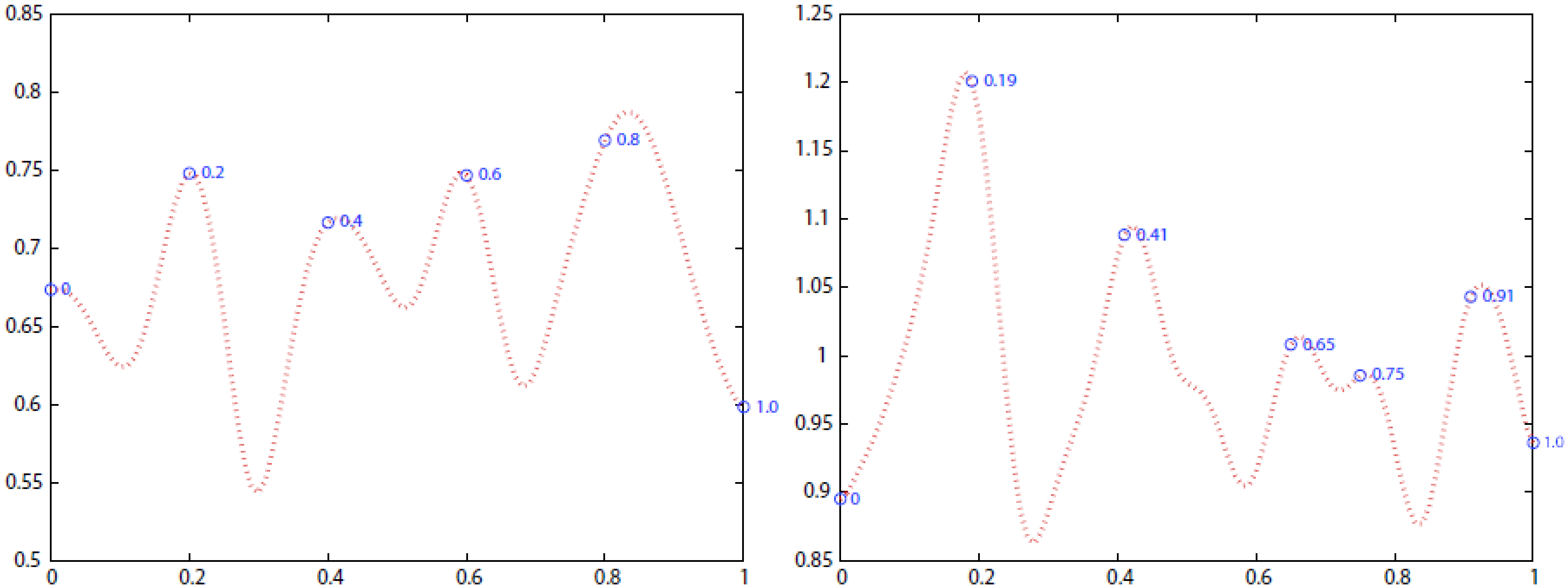




The locally weighted bag-of-words framework

- Text segmentation:

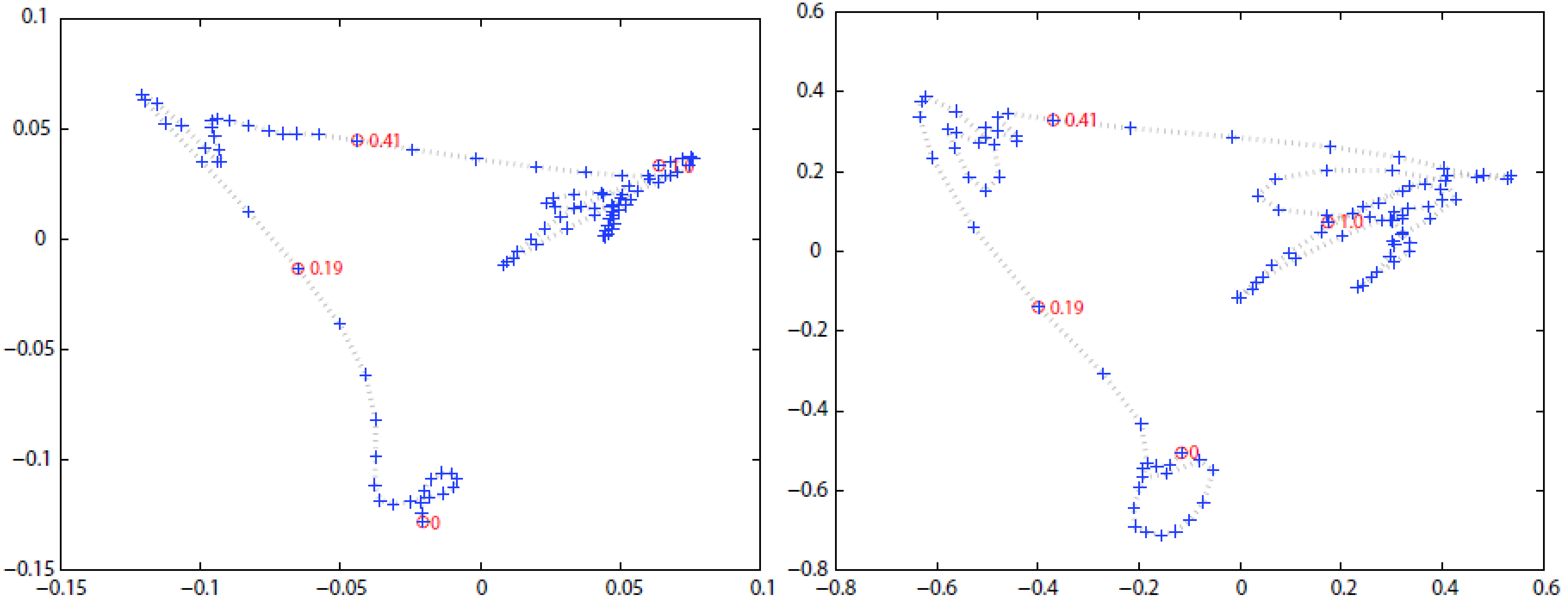




- Taking the gradient norm of the lowbow curve: $\| \dot{Y}_{\mu}(d_i) \|_2$

The locally weighted bag-of-words framework

- Text segmentation:

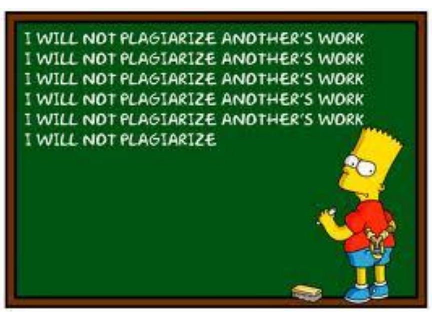


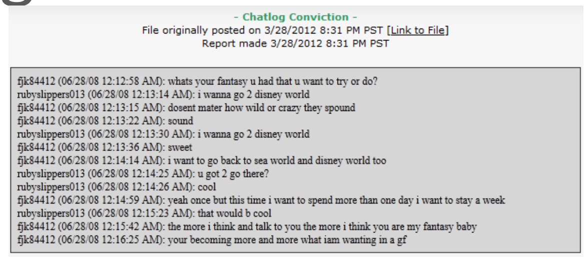
- PCA (left) and MDS (right) projections

LOWBOW for authorship attribution

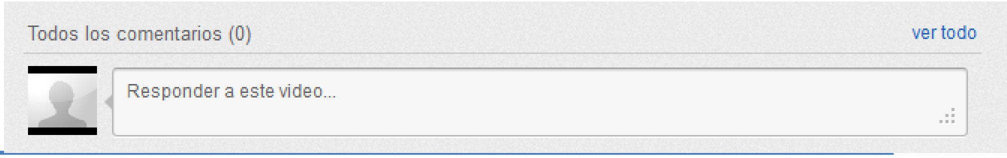
- **Authorship attribution:** Given texts of uncertain authorship and texts from a set of candidate authors, the task is to map the uncertain texts onto their true authors among the candidates.

- Applications include: fraud detection, spam filtering, computer forensics and plagiarism detection




- Chatlog Conviction -
File originally posted on 3/28/2012 8:31 PM PST [Link to File]
Report made 3/28/2012 8:31 PM PST

fjk84412 (06/28/08 12:12:58 AM): whats your fantasy u had that u want to try or do?
rubyslippers013 (06/28/08 12:13:14 AM): i wanna go 2 disney world
fjk84412 (06/28/08 12:13:15 AM): dosent mater how wild or crazy they spound
fjk84412 (06/28/08 12:13:22 AM): sound
rubyslippers013 (06/28/08 12:13:30 AM): i wanna go 2 disney world
fjk84412 (06/28/08 12:13:36 AM): sweet
fjk84412 (06/28/08 12:14:14 AM): i want to go back to sea world and disney world too
rubyslippers013 (06/28/08 12:14:25 AM): u got 2 go there?
rubyslippers013 (06/28/08 12:14:26 AM): cool
fjk84412 (06/28/08 12:14:59 AM): yeah once but this time i want to spend more than one day i want to stay a week
rubyslippers013 (06/28/08 12:15:23 AM): that would b cool
fjk84412 (06/28/08 12:15:42 AM): the more i think and talk to you the more i think you are my fantasy baby
fjk84412 (06/28/08 12:16:25 AM): your becoming more and more what iam wanting in a gf



LOWBOW for authorship attribution

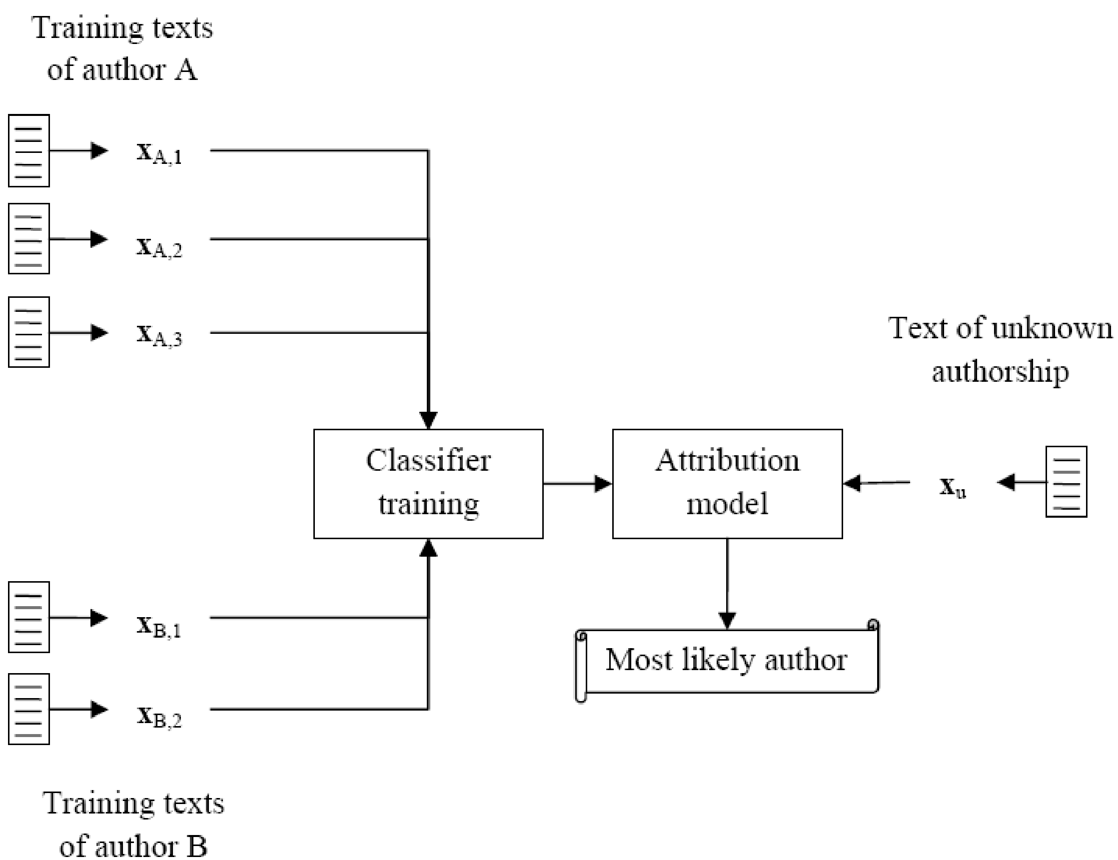
- LOWBOW acts as an expansion of the BOW approach that can be particularly suitable for AA

- Local histograms incorporate sequential information that reveal clues about the writing style of authors

- **Hypothesis:** Authors use similar distributions of certain words when writing documents

- We explore the use of LOWBOW for AA using character n-grams

LOWBOW for authorship attribution



LOWBOW for authorship attribution

- How to take advantage of the multiple vectors associated to each document:
  - Combining the vectors (LOWBOW histogram)

$$\mathbf{L}_i = \sum_{j=1}^{k} \mathbf{dl}_i^j$$

  - Use the set of vectors to represent the document (BOLH)

$$\mathbf{L}_i = \{\mathbf{dl}_i^1, \ldots, \mathbf{dl}_i^k\}$$

- Classifier: Support vector machine

$$f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) - b$$

LOWBOW for authorship attribution

- Kernels for BOLHs

$$K(P,Q) = e^{-\frac{1}{\gamma} D(P,Q)^2}$$

Kernel	**Distance**
Diffusion	$D(P,Q) = \sum_{l=1}^{k} \arccos\left(\left\langle \sqrt{\mathbf{p}_l} \cdot \sqrt{\mathbf{q}_l} \right\rangle\right)$
Earth mover's distance	*EMD(P,Q)*
Euclidean	$D(P,Q) = \sum_{l=1}^{k} \sum_{i=1}^{\lvert V \rvert} \sqrt{(\mathbf{p}_l^i - \mathbf{p}_l^i)^2}$
Chi-squared	$D(P,Q) = \sum_{l=1}^{k} \sum_{i=1}^{\lvert V \rvert} \frac{(\mathbf{p}_l^i - \mathbf{q}_l^i)^2}{(\mathbf{p}_l^i + \mathbf{q}_l^i)}$

39

Experimental settings

- We consider a subset of RCV-I, documents written by 10 authors (about the same subject); 50 documents are available for training and 50 for testing for each author

- Experiments using words and 3-grams at the character level were performed, different number of locations and scale parameters were evaluated, we report the settings that showed better performance

- The 2500 most frequent terms were used to obtain the representations

Experimental settings

- Three settings were considered:
  - **Balanced data set (BC):** 50 documents for training per author
  - **Reduced data set (RBC):** 4 subsets using 1, 3, 5 and 10 training documents per author
  - **Imbalanced data set (IRBC):** 3 subsets generated with a Gaussian distribution over authors using at least 2, 5, 10 and at most 10, 10, and 20 documents, respectively.

Challenging conditions

Balanced data set (BC)

BOW a strong baseline

Method	Parameters	Words	Char. N-grams
BOW	-	78.2%	75.0%
LOWBOW	k = 2; σ = 0.2	75.8%	72.0%
LOWBOW	k = 5; σ = 0.2	77.4%	75.2%
LOWBOW	k = 20; σ = 0.2	77.4%	75.0%

LOWBOW histograms

k	Euc.	Diff.	EMD	Chi^2
Words				
2	78.6%	81.0%	75.0%	75.4%
5	77.6%	82.0%	72.0%	77.2%
20	79.2%	80.8%	75.2%	79.0%
Character N-grams				
2	83.4%	82.8%	84.4%	83.8%
5	83.4%	84.2%	82.2%	84.6%
20	84.6%	86.4%	81.0%	85.2%

BOLH

BOLHs obtained better performance

Balanced data set (BC)

BOW a strong baseline

Method	Parameters	Words	Char. N-grams
BOW	-	78.2%	75.0%
LOWBOW	k = 2; σ = 0.2	75.8%	72.0%
LOWBOW	k = 5; σ = 0.2	77.4%	75.2%
LOWBOW	k = 20; σ = 0.2	77.4%	75.0%

LOWBOW histograms

k	Euc.	Diff.	EMD	Chi^2
		Words		
2	78.6%	81.0%	75.0%	75.4%
5	77.6%	82.0%	72.0%	77.2%
20	79.2%	80.8%	75.2%	79.0%
		Character N-grams		
2	83.4%	82.8%	84.4%	83.8%
5	83.4%	84.2%	82.2%	84.6%
20	84.6%	86.4%	81.0%	85.2%

BOLH

BOLHs obtained better performance

Balanced data set (BC)

BOW a strong baseline

Method	Parameters	Words	Char. N-grams
BOW	-	78.2%	75.0%
LOWBOW	k = 2; σ = 0.2	75.8%	72.0%
LOWBOW	k = 5; σ = 0.2	77.4%	75.2%
LOWBOW	k = 20; σ = 0.2	77.4%	75.0%

LOWBOW histograms

k	Euc.	Diff.	EMD	Chi^2
Words				
2	78.6%	81.0%	75.0%	75.4%
5	77.6%	82.0%	72.0%	77.2%
20	79.2%	80.8%	75.2%	79.0%
Character N-grams				
2	83.4%	82.8%	84.4%	83.8%
5	83.4%	84.2%	82.2%	84.6%
20	84.6%	86.4%	81.0%	85.2%

BOLH

BOLHs obtained better performance

Reduced balanced data sets

- Using words as terms

Method \ dataset	1-doc	3-docs	5-docs	10-docs	50-docs
BOW	36.8%	57.1%	62.4%	69.9%	78.2%
LOWBOW	37.9%	55.6%	60.5%	69.3%	77.4%
Diff. Kernel	52.4%	63.3%	69.2%	72.8%	82.0%
Reference	-	-	53.4%	67.8%	80.8%

- Using character n-grams as terms

Method \ dataset	1-doc	3-docs	5-docs	10-docs	50-docs
BOW	65.3%	71.9%	74.2%	76.2%	75.0%
LOWBOW	61.9%	71.6%	74.5%	73.8%	75.0%
Diff. Kernel	70.7%	78.3%	80.6%	82.2%	86.4%
Reference	-	-	53.4%	67.8%	80.8%

Imbalanced data sets

- Using words as terms

Method \ dataset	2-10	5-10	10-20
BOW	62.3%	67.2%	71.2%
LOWBOW	61.1%	67.4%	71.5%
Diff. Kernel	66.6%	70.7%	74.1%
Reference	49.2%	59.8%	63.0%

- Using character n-grams as terms

Method \ dataset	2-10	5-10	10-20
BOW	70.1%	73.4%	73.1%
LOWBOW	70.8%	72.8%	72.1%
Diff. Kernel	77.8%	80.5%	82.2%
Reference	49.2%	59.8%	63.0%

LOWBOW for authorship attribution

- Conclusions:
  - Sequential information encoded in local histograms is useful for AA. Character-level representations, which have proved to be very effective for AA can be further improved by adopting a local histogram formulation
  - Our results are superior to state of the art approaches, with improvements ranging from 2%-6% in balanced data sets and from 14%-30% in imbalanced data sets (larger improvements were observed in challenging conditions)
  - In preliminary experiments with short texts we have found that LOWBOW does not work very well

Research opportunities with LOWBOW

- Automatically-dynamically setting the number of local histograms for documents according to their length

- Studying the performance of local histograms in terms of length of documents, training set size, sparseness, narrowness of domain, etc.

- Profile-based authorship attribution using local histograms

- Learning the appropriate smoothing function from data

Discussion

- One of the main limitations of the BOW formulation is its inability to incorporate sequential information
- Several extensions/alternatives to BOW have been proposed so far, each of which has limitations and advantages with respect to each other
- Too much work to do in this topic = research opportunities

References

- R. Bekkerman, J. Allan. **Using Bigrams in Text Categorization.** CIIR Technical Report IR-408 2004.
- H. Ahonen-Myka. **Finding All Maximal Frequent Sequences in Text.** Proceedings of the 16th International Conference on Machine Learning ICML-99 Workshop on Machine Learning in Text Data Analysis, eds. D. Mladenic and M. Grobelnik, p. 11-17, J. Stefan Institute, Ljubljana 1999.
- R. M. Coyotl-Morales, L. Villaseñor-Pineda, M. Montes-y-Gómez, P. Rosso. **Authorship Attribution using Word Sequences.** 11th Iberoamerican Congress on Pattern Recognition, CIARP 2006. Cancun, Mexico, November 2006.
- S. Jaillet , A. Laurent, M. Teisseire. **Sequential patterns for text categorization.** Intelligent Data Analysis, Vol. 10(3):199--214, 2006
- B. Zhang. Learning Features for Text Classification. PhD Thesis, University of Washington, Electrical Engineering dept. 2013.
- D. Guthrie, B. Allison, W. Liu, L. Guthrie, Y. Wilks. **A Closer Look at Skip-gram Modelling.** *Proceedings of the Fifth international Conference on Language Resources and Evaluation LREC-2006, Genoa, Italy,* (*2006*)
- G. Lebanon, Y. Mao, M. Dillon. **The Locally Weighted Bag of Words Framework for Document Representation.** Journal of Machine Learning Research. Vol. 8, pp. 2405—2441, 2007.
- H. J. Escalante, T. Solorio, M. Montes-y-Gómez. **Local Histograms of Character Ngrams for Authorship Attribution.** Proc. of ACL-HTL 2011, pp. 288—298, Portland, OR, 2011.

Novel representations and methods in text classification

SYNTACTIC INFORMATION IN TEXT CLASSIFICATION

Outline

- Complex linguistic features for text classification
- Use of syntactic features in authorship attribution
  - Brief review
  - Syntactic-based n-grams as features
  - AA using Probabilistic Context-Free Grammars
- Final remarks

Background

- Long history on the use of complex **linguistic features** in information retrieval (refer to TREC reports)
  - Have been used: lemmas, POS information, named entities, noun phrases, complex nominals, syntactic tuples such as subject-verb, verb-object, etc.
- General conclusion: the high computational cost of the adopted NLP algorithms, the small improvement produced over simple BoW representation, and the lack of accurate WSD tools are the reasons for the **failure of NLP in document retrieval**

Linguistic features in text classification

- Are they useful for text classification?
  - IR and text classification are similar tasks, both are rely on thematic similarities.
  - Strong evidence indicates that POS information, complex nominals, and word senses are **not adequate to improve TC accuracy**

Useful for other textual-based classification tasks?

Alessandro Moschitti, Roberto Basili. *Complex Linguistic Features for Text Classification: A Comprehensive Study*. Lecture Notes in Computer Science Volume 2997, 2004.

Features in authorship attribution

- AA deals with the definition of features that quantify the **writing style of authors**, and with the application of methods able to learn from that kind of features.
  - Lexical features → stylometric measures, words n-grams, function words
  - Character-based features → n-grams
  - **Syntactic features**
  - Semantic features → Use of synonyms and hyponyms, LSI
  - Domain specific features → Use/type of greetings, signatures, indentation, etc.

Efstathios Stamatatos. *A survey of modern authorship attribution methods*. Journal of the American Society for information Science and Technology 60(3): 538–556 (2009)

Syntactic features in AA

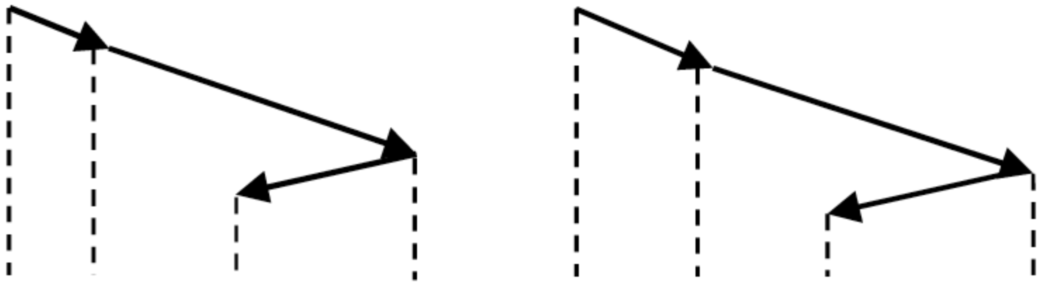
- The idea is that authors tend to use **similar syntactic patterns** unconsciously.
  – Strong authorial fingerprint
- Two basic approaches:
  – Use **POS tag frequencies** or POS n-gram frequencies as features
  – Apply a chunker, and use **phrase counts** as features
- Recent approaches:
  – Using syntactic-based n-grams as features
  – Using probabilistic context free grammars as language models for classification.

Syntactic n-grams

- Sn-grams are obtained based on the order in which the elements are presented in **syntactic trees**.
  - Constructed by **following a path in the tree**, rather than taking words as they appear in the text.
- Because sn-grams are based on syntactic relations of words, each word is bound to its **real neighbors**, ignoring the arbitrariness that is introduced by the surface structure

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, Liliana Chanona-Hernández. *Syntactic Dependency-Based N-grams as Classification Features*. Lecture Notes in Computer Science Volume 7630, 2013.

An example of sn-grams

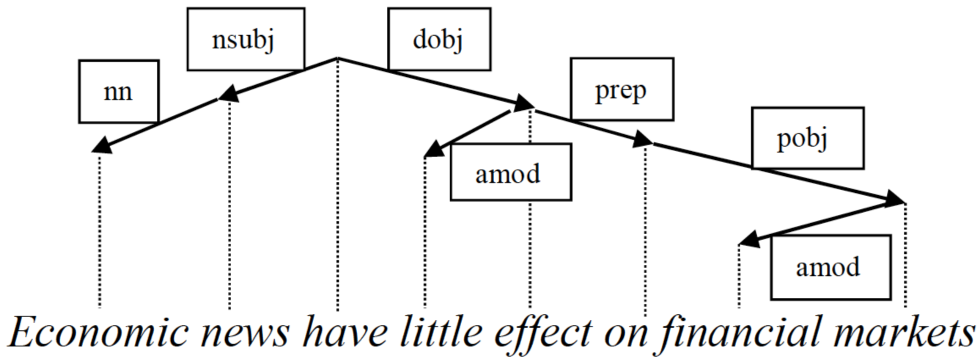


eat with wooden spoon *eat with metallic spoon*

- Common word n-grams:
  - eat with
- Common word sn-grams:
  - eat with, **with spoon**; **eat with spoon**
- Ignoring function words we would obtain:
  - **eat spoon**

Other variants of sn-grams

- In addition to word sn-grams, it is possible to build:
  - POS sn-grams
  - Sn-grams of syntactic relations tags (**SR tags**), where the elements are names of syntactic relations
  - Mixed sn-grams: composed by mixed elements like words (lexical units), POS tags and/or SR tags.



Sn-grams of SR tags

nsubj → *nn*
dobj → *amod*
dobj → *prep* → *pobj* → *amod*

Results

	Training		Classification	
Author	Novels	Size (MB)	Novels	Size (MB)
Booth Tarkington	8	3.6	5	1.8
George Vaizey	8	3.8	5	2.1
Louis Tracy	8	3.6	5	2.2
Total	**24**	**11**	**15**	**6.1**

Profile size	**sn-grams of SR tags**	**n-grams of POS tags**	**Character based n-grams**	**Word based n-grams**
400	100%	90%	76%	81%
1,000	100%	90%	86%	71%
4,000	100%	100%	95%	95%
7,000	100%	100%	90%	90%
11,000	100%	95%	100%	90%

- Profile size indicates the first most frequent n-grams/sngrams

AA using Probabilistic Context Free Grammars

- Idea: use of syntactic information by building complete models of each **author's syntax** to distinguish between authors.
- How: build a probabilistic context free grammar **(PCFG) for each author** and use this grammar as a **language model for classification**.
  - A PCFG is a probabilistic version of a CFG where each production has a probability
  - Probability of a sentence/derivation is the product of the probabilities of its productions

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. *Authorship attribution using probabilistic context-free grammars*. In Proceedings of the ACL 2010 Conference. Uppsala, Sweden, July 2010.

General procedure

Input – A training set of documents labeled with author names and a test set of documents with unknown authors.

1. Train a statistical parser on a generic corpus like the WSJ or Brown corpus.
2. Treebank each training document using the parser trained in Step 1.
3. Train a PCFG G_i for each author A_i using the treebanked documents for that author.
4. For each test document, compute its likelihood for each grammar G_i by multiplying the probability of the top PCFG parse for each sentence.
5. For each test document, find the author A_i whose grammar G_i results in the highest likelihood score.

Output – A label (author name) for each document in the test set.

- Generate a **parse tree** for each training document
- Estimate a **grammar** and its parameters from the assembled “tree-bank”
- Compute **probabilities** for each document, for each grammar
- **Select the author** (grammar) with the highest probability

Results

Dataset	# authors	# words/auth	# docs/auth	# sent/auth
Football	3	14374.67	17.3	786.3
Business	6	11215.5	14.16	543.6
Travel	4	23765.75	28	1086
Cricket	4	23357.25	24.5	1189.5
Poetry	6	7261.83	24.16	329

Words *Characters* *PCFG*

Dataset	MaxEnt	Naive Bayes	Bigram-*I*	PCFG	PCFG-*I*	PCFG-*E*
Football	84.45	86.67	86.67	**93.34**	80	91.11
Business	83.34	77.78	90.00	77.78	85.56	91.11
Travel	83.34	83.34	**91.67**	81.67	86.67	**91.67**
Cricket	91.67	**95.00**	91.67	86.67	91.67	**95.00**
Poetry	56.36	78.18	70.90	78.18	83.63	**87.27**

- **PCFG-I:** augments the training data with section of the Brown corpus; replicates the original data 3-4 times
- **PCFG-E:** an ensemble of MaxEnt, Bigram-I and PCFG-I

Final remarks

- Syntactic information is an important authorial **fingerprint**
- But, **both syntactic and lexical information** are useful in effectively capturing authors' overall writing style
  - Mixed sn-grams are a good compromise between these two sources of information
- Some disadvantages of using syntactic-based features:
  - Syntactic parsing is required!
    - Can take considerable **time**
    - Problem of **availability** of parsers for some languages
  - **Language-dependent** procedure

References

- Alessandro Moschitti, Roberto Basili. **Complex Linguistic Features for Text Classification: A Comprehensive Study.** Lecture Notes in Computer Science Volume 2997, 2004.
- Efstathios Stamatatos. **A survey of modern authorship attribution methods**. *Journal of the American Society for information Science and Technology* 60(3): 538–556 (2009)
- Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, Liliana Chanona-Hernández. **Syntactic Dependency-Based N-grams as Classification Features**. Lecture Notes in Computer Science Volume 7630, 2013.
- Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. **Authorship attribution using probabilistic context-free grammars**. *In Proceedings of the ACL 2010 Conference*. Uppsala, Sweden, July 2010.