Determining unintelligible words from their textual context

Balázs Pintér, Gyula Vörös, Zsolt Palotai, Zoltán Szabó, András Lőrincz

Sept. 3, 2012

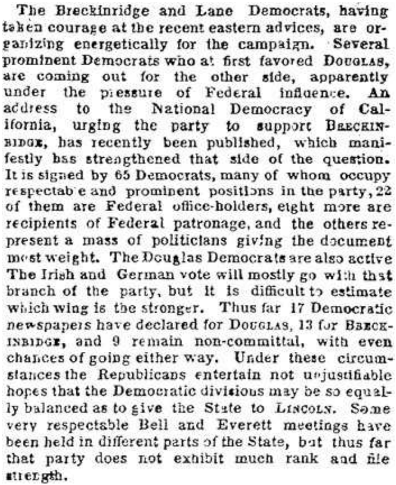
Contents

1 The problem

2 The method

3 Results

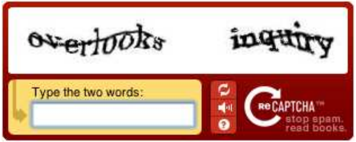
Motivation – reCAPTCHA



The Breckinridge and Lane Democrats, having taken courage at the recent eastern advices, are organizing energetically for the campaign. Several prominent Democrats who at first favored DOUGLAS, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support BRECKINRIDGE, has recently been published, which manifestly bss strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party, 22 of them are Federal office-holders, eight more are recipients of Federal patronage, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for DOUGLAS, 13 for BRECKINRIDGE, and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to LINCOLN. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.

The Hreckinridge' and Lane Democrats, having taken courage at the recent eastern advises, are xxxxxxxxxx energetically for the campaign: Several prominent Democrats who at first favored DonoLea, are coming out. for the other aide, apparently under the xxxxxxxx of Federal xxxxxxxxx. An address to the National Democracy of ,1ifornia, urging the party to support HaeeslipsIDas, has recently been published, which manifestly bss strengthened that aide of the xxxxxxxxx: It is signed by 65 Democrats,

Motivation



- reCAPTCHA determines unintelligible words by means of human effort
- by solving CAPTHAs, users on the Web also help digitize content

Our goal is to determine unintelligible words *automatically*: selecting the right word from a list of *candidate words*, using

- their *context*
- the *distributional hypothesis*
- and *structured sparse coding*

The distributional hypothesis and spurious similarities

- Words that occur in the same contexts tend to have similar meaning
  - Context: words preceding and following the target word
  - Example: Democrats who at first favored DonoLea, are coming out
- There are exceptions to the hypothesis
  - spuriously similar contexts: when two contexts are similar but belong to different words
- many candidate words → many spuriously similar contexts
- A mechanism is needed to deal with spurious similarities → structured sparse coding

Contents

1 The problem

2 The method

3 Results

Two steps

- *First step*: solve an inverse problem $\mathbf{D}\boldsymbol{\alpha} \approx \mathbf{x}$, where
  - $\mathbf{x}$ is the context of the unintelligible word
  - $\mathbf{D}$ is the *word-context matrix* of *dictionary*
  - $\boldsymbol{\alpha}$ is the representation vector
- *Second step*: Obtain a single candidate word from the representation vector $\boldsymbol{\alpha}$

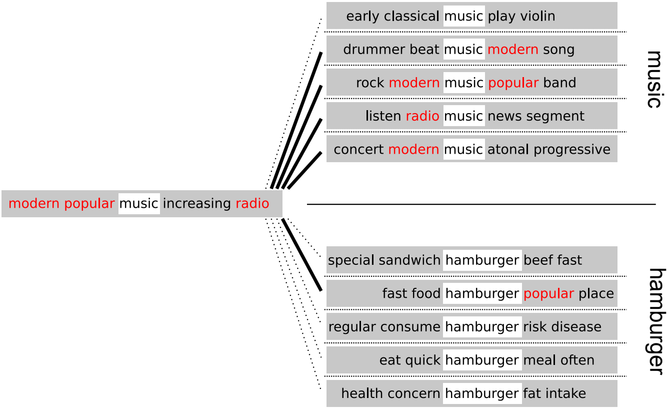
First step – solving an inverse problem

$$\begin{array}{c|cccc|cccc|c|cccc}
 & \multicolumn{4}{c}{\text{boot}} & \multicolumn{4}{c}{\text{root}} & \cdots & \multicolumn{4}{c}{\text{foot}} \\
\text{computer} & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 \\
\text{plant} & 0 & 0 & 0 & 0 & 0 & 1 & 3 & 0 & & 0 & 0 & 0 & 0 \\
\text{shoe} & 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & & 1 & 0 & 2 & 1 \\
\vdots & & & & & & & & & & & & & \\
\text{vegetable} & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & & 0 & 1 & 0 & 0
\end{array}$$

The representation vector $\boldsymbol{\alpha} = [\alpha_1; \alpha_2; \ldots; \alpha_n] \in \mathbb{R}^n$

$$\mathbf{x} = \alpha_1 \mathbf{d}_1 + \alpha_2 \mathbf{d}_2 + \ldots + \alpha_n \mathbf{d}_n$$

The structured sparsity inducing regularization



modern popular music increasing radio

early classical music play violin

drummer beat music modern song

rock modern music popular band

listen radio music news segment

concert modern music atonal progressive

music

special sandwich hamburger beef fast

fast food hamburger popular place

regular consume hamburger risk disease

eat quick hamburger meal often

health concern hamburger fat intake

hamburger

- Whole groups are selected → spurious similarities have less effect

Group Lasso

- The columns of **D** are organized into groups
  - $\mathcal{G} = \{G_l\}_{l \in L} \subseteq 2^{\{1,\ldots,n\}}$
  - G_l is a group labeled with $l \in L$, that contains indices of columns of **D**
- Our goal is to select only a few *groups*

Group Lasso

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{l \in L} w_l ||\boldsymbol{\alpha}_{G_l}||_2$$

Second step

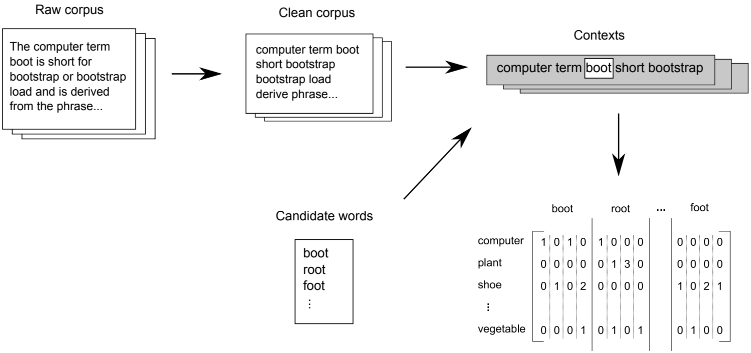
- Obtain a single candidate word from the representation vector $\boldsymbol{\alpha}$
  - sum the weights in each group
  - select the candidate word $l^* \in L$ whose group G_{l^*} contains the most weight

Selecting a single candidate word

$$l^* = \arg\max_{l \in L} \sum_i (\boldsymbol{\alpha}_{G_l})_i$$

Contents

Generating the datasets



Raw corpus

The computer term boot is short for bootstrap or bootstrap load and is derived from the phrase...

Clean corpus

computer term boot short bootstrap bootstrap load derive phrase...

Candidate words

boot
root
foot
⋮

Contexts

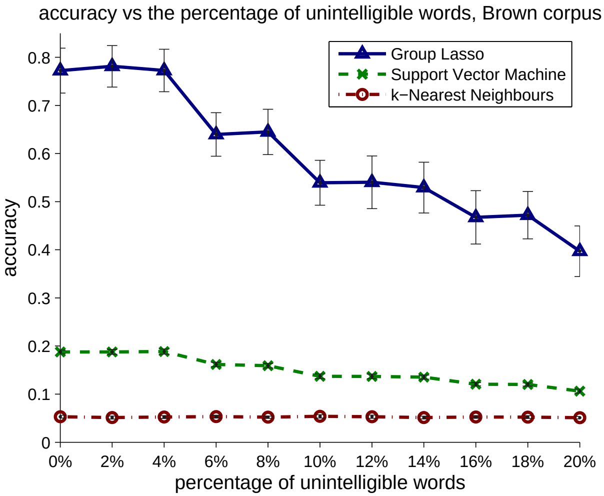
computer term boot short bootstrap

	boot				root				...	foot			
computer	1	0	1	0	1	0	0	0		0	0	0	0
plant	0	0	0	0	0	1	3	0		0	0	0	0
shoe	0	1	0	2	0	0	0	0		1	0	2	1
⋮													
vegetable	0	0	0	1	0	1	0	1		0	1	0	0

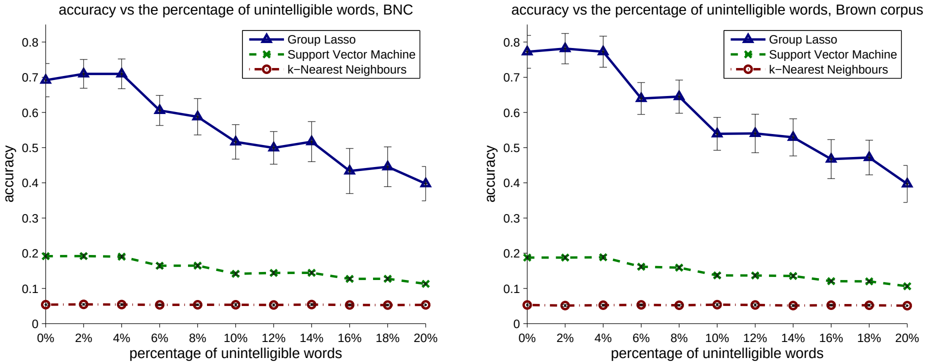
Evaluations

- Two goals:
  - Compare the method to baselines
    - Support Vector Machine
    - k-Nearest Neighbors
  - Examine the effect of the ratio of unintelligible words on the accuracy
    - Delete p percent of the words from the contexts in the test set ($\mathbf{x}$)
    - As p is increasing, measure the drop in accuracy
- Cross-validation: shuffle and split 30 times on each datapoint

Accuracy vs percentage of unintelligible words, Brown corpus



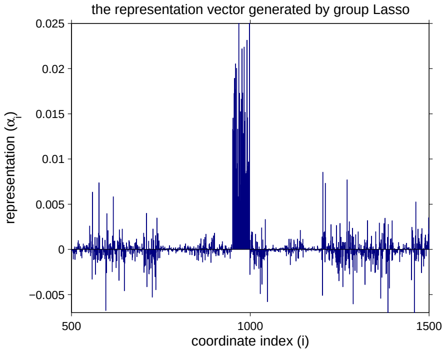
Accuracy vs percentage of unintelligible words, comparison



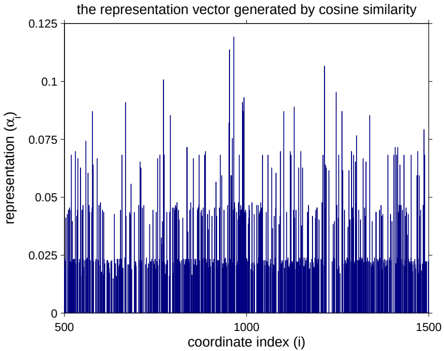
(a) BNC

(b) Brown corpus

Comparison of representation vectors



(a) group Lasso



(b) cosine similarity

Acknowledgments

The research has been supported by the ‘European Robotic Surgery’ EC FP7 grant (no.: 288233).

Nemzeti Fejlesztési Ügynökség
www.ujszechenyiterv.gov.hu
06 40 638 638

MAGYARORSZÁG MEGÚJUL

The Project is supported by the European Union and co-financed by the European Social Fund (grant agreement no. TAMOP 4.2.1./B-09/1/KMR-2010-0003).

The end

Thank you for your attention!