Distributed Representations of Geographically Situated Language

David Bamman Chris Dyer Noah A. Smith

Presenter: Konstantinos Pappas

“Who?”

David Bamman
Assistant Professor

Chris Dyer
Assistant Professor

Noah Smith
Associate Professor

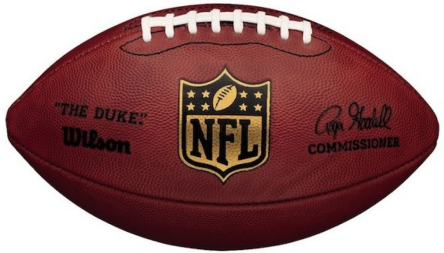
“What?”

1. Incorporate contextual information (geography) in learning vector-space representations of *situated* language.
   - contextual variable: 51 US states
   - dataset: 93M tweets - 1B words
   - representation: embeddings (Mikolov et al. 2013)
   - evaluation: qualitatively (manual inspection) and quantitatively (semantic similarity)

"WHY?" #1

“Why?” #2

In particular, how is a word’s meaning shaped by its geography?

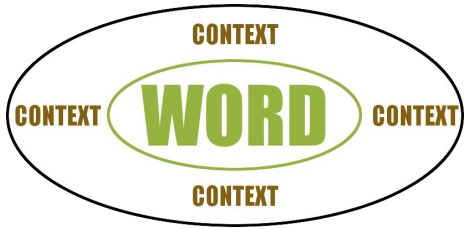


Distributional hypothesis

Words that occur in similar contexts tend to have similar meanings (Harris, 1954; Firth, 1957; Deerwester et al., 1990). - If words have similar row vectors in a word-context matrix, then they tend to have similar meanings.

1-modal learning:

- textual context



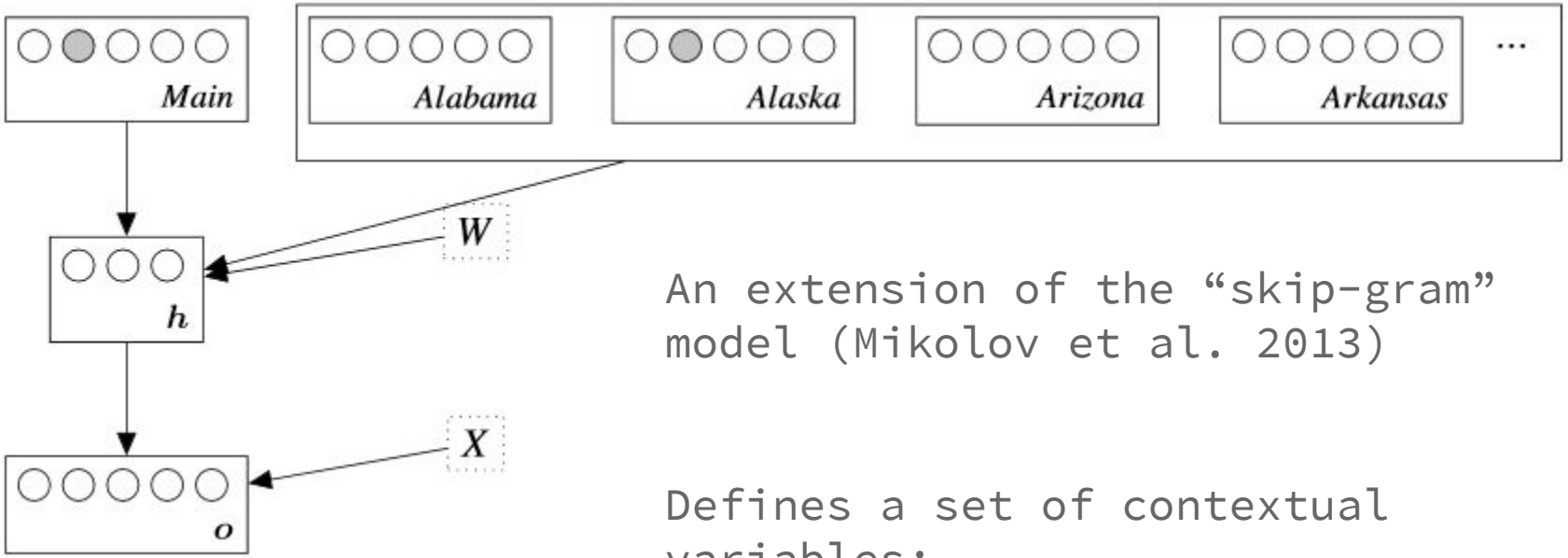
Language is situated

You shall know a word by the company it keeps . . .
Firth (1957)

Multimodal learning:

- object info e.g. visual cues
- speaker info e.g. location

MODEL



An extension of the “skip-gram” model (Mikolov et al. 2013)

Defines a set of contextual variables:
Cstate = (AK, AL, …, WY)

Model details

One global embedding matrix W.

Another 51 matrices which capture the effect that each variable value has on each word in the vocabulary.

Each deviation indicates how that common representation should shift in the k-dimensional space when used in each state.

Backpropagation, L2 regularization.

IMPLEMENTATION

github: https://github.com/dbamman/geoSGLM

GeoSGLM: Code for learning geographically-informed word embeddings.

To run, adjust the input/output parameters in run.sh and execute it.

word2vec

DATA=DATA/DATA.TEST.TXT

id	location	message
480326347508969000	PA	There is a great research question in how long a sequence of blog comments can go before it descends into madness http://t.co/NFqKgaZRuO
472023364908118000	PA	So much easier than hunting through individual websites : using Google Scholar to get BibTeX citations http://t.co/H2inkMGMom
105039889808109000	PA	Just discovered Conflict Kitchen in Pittsburgh - brilliant idea that needs to catch on in other cities . http://t.co/FkSLGD9

VOCABFILE=data/vocab.txt & MAXVOCAB=100000

The vocab file contains the maximal set of words to learn representations for.

If a word is not in this list, then don't learn a representation for it.

This list is further filtered in the code to only include words that are seen at least 5 times in the data, and a maximum of the $MAXVOCAB most frequent terms.

ETC

```
FEATUREFILE=data/states.txt

OUTFILE=data/out.embeddings

DIMENSIONALITY=100
Dimensionality specifies the size of the learned word
representations.

L2=0.0001
L2 regularization parameter.
```

Similarity

For a given query q, you can view the terms most similar to q in all 51 states using scripts/findNearest.py

```
python scripts/findNearest.py $OUTFILE
```

QUALITATIVE ANALYSIS #1

Kansas		Massachusetts	
term	cosine	term	cosine
wicked	1.000	wicked	1.000
evil	0.884	super	0.855
pure	0.841	ridiculously	0.851
gods	0.841	insanely	0.820
mystery	0.830	extremely	0.793
spirit	0.830	goddamn	0.781
king	0.828	surprisingly	0.774
above	0.825	kinda	0.772
righteous	0.823	#sarcasm	0.772
magic	0.822	soooooo	0.770

Table 1: Terms with the highest cosine similarity to *wicked* in Kansas and Massachusetts.

QUALITATIVE ANALYSIS #2

California		New York	
term	cosine	term	cosine
city	1.000	city	1.000
valley	0.880	suburbs	0.866
bay	0.874	town	0.855
downtown	0.873	hamptons	0.852
chinatown	0.854	big city	0.842
south bay	0.854	borough	0.837
area	0.851	neighborhood	0.835
east bay	0.845	downtown	0.827
neighborhood	0.843	upstate	0.826
peninsula	0.840	big apple	0.825

Table 2: Terms with the highest cosine similarity to *city* in California and New York.

QUANTITATIVE EVALUATION - SET UP

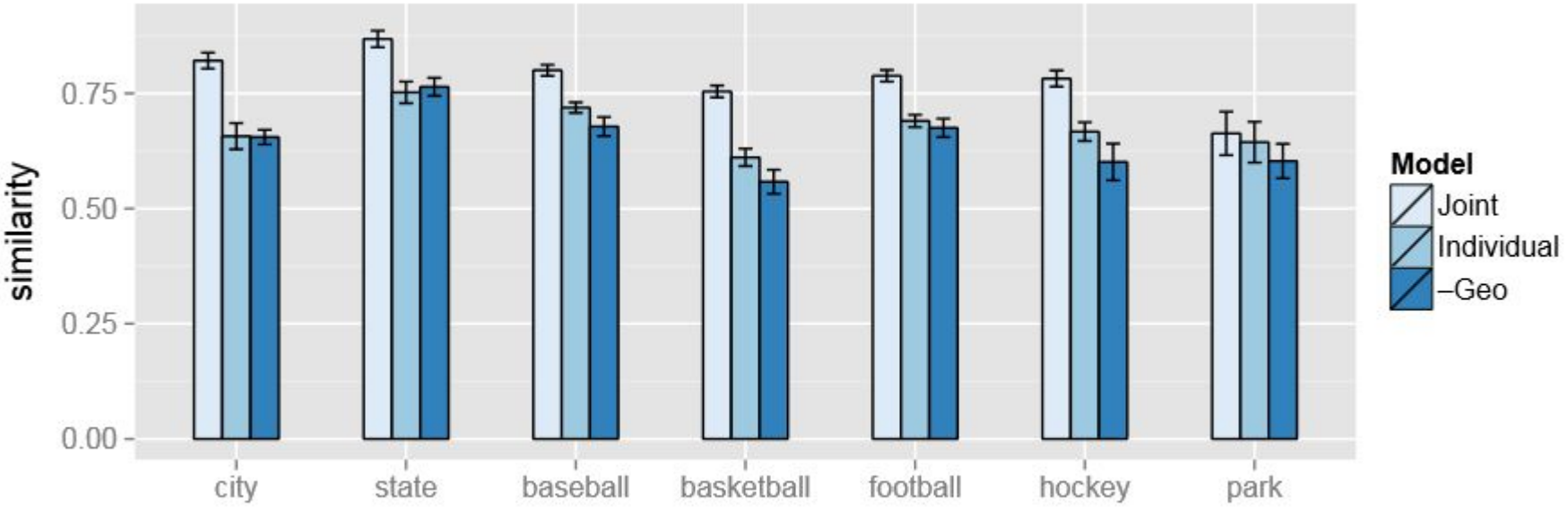
7 categories

1. city - most populous city/state
2. state - state name
3. football - NFL team names
4. basketball - NBA team names
5. baseball - MLB team names
6. hockey - NHL team names
7. park - US national parks

3 models

1. JOINT: global representation for each word + a deviation per state
2. INDIVIDUAL: each state one model
3. -GEO: one model from the whole US

SEMANTIC SIMILARITY



Average cosine similarity for all models across all categories, with 95% confidence intervals on the mean.

CONCLUSION

The paper provides an extension to vector-space representations that can take into account the context in which it is uttered.

Implements three models: joint, individual, normal.

Provides two different kinds of evaluation of the models.

Discusses possible extensions and applications of this tool.

NOTES

- ACL 2014 (+)
- Mentions that this tool for revealing periodic and historical influences on lexical semantics, but provides no evidence (-)
- Provides online implementation of the system (+)

Questions

Can we realistically find enough data for each contour that we are interested? E.g. a particular year?

How can these new embeddings be used for IR?

Would it make sense to create different embeddings per gender? Per age of author?