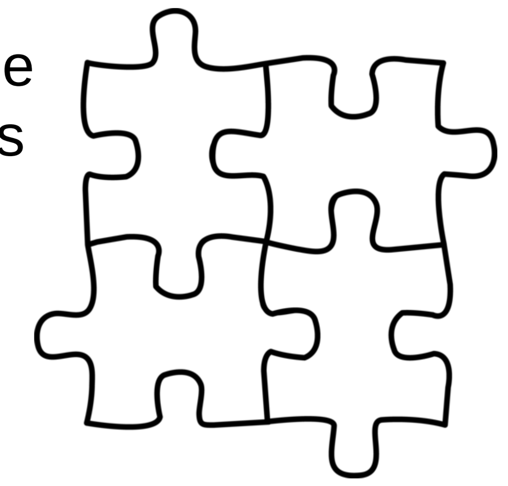
Word Games: Coherent Document Reconstruction

Mitchell Douglass (mrdoug95@stanford.edu) and Caelin Tran (cktt13@stanford.edu)
Stanford University | Department of Computer Science

CS 229

Reconstructing Documents

Making coherent text is a key characteristic of successful natural language generation. We tackled the subproblem of reconstructing large documents (traditionally constrained in size by other approaches[1]) from their unordered fragments. We achieved promising results.



Input: document fragments (varying length or varying number)
Model: determine maximum likelihood of sequence
Output: predicted fragment order

Dataset and Features

Wikipedia articles → Clean text and punctuation → Categorize by topic → Randomly fragment all articles → Unigrams and bigrams / GloVe word vector sums

Our documents were 3.3M Wikipedia articles. We focused on the categories of **film** (75K articles), **people** (120K), and **cities** (160K). Our ground truth is the correct order, which we store after the fragmentation step, and our two feature mappers produced from each fragment *n* grams and 100-dimensional embeddings (sums of GloVe[2] word vectors).

Approaches and Models

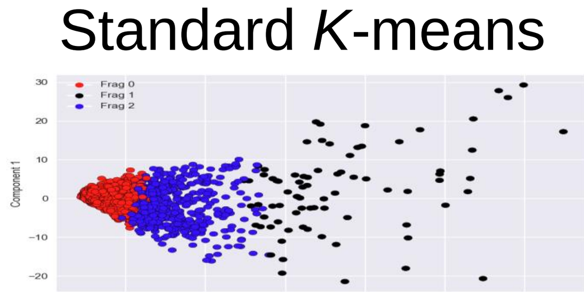
***k*-fragments:** make *k* fragments from each document; assume position carries semantics

Fragment 2, Fragment *k*, ⋮, Fragment 1 ≈ Background, Conclusion, ⋮, Introduction → Model → Introduction, Background, ⋮, Conclusion

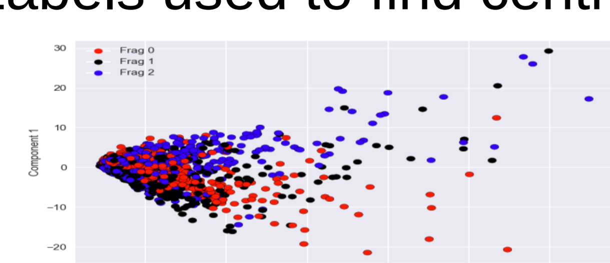
Naïve Bayes
Prob. of frag being in a pos.

$$\max_{p \in permutations} \prod_{i=1}^{k} P(pos = i \mid frag_{p_i})$$

Unsupervised *K*-means
Standard *K*-means



Supervised *K*-means
Labels used to find centroids



Transitional: make fragments of *m* sentences; assume adjacent fragments are related

Fragmented documents → Model for transition probability P → Predict optimal permutation (beam search)

$$\max_{p} \prod_{i>1} P(frag_{p_i} \mid frag_{p_{i-1}})$$

Logistic regression

- Concatenated GloVe sums of adjacent fragments used as input
- Trained to predict whether fragment pairs were ordered properly

$$\sigma(x) = \frac{1}{1+\exp(\theta^T x)}$$

Results

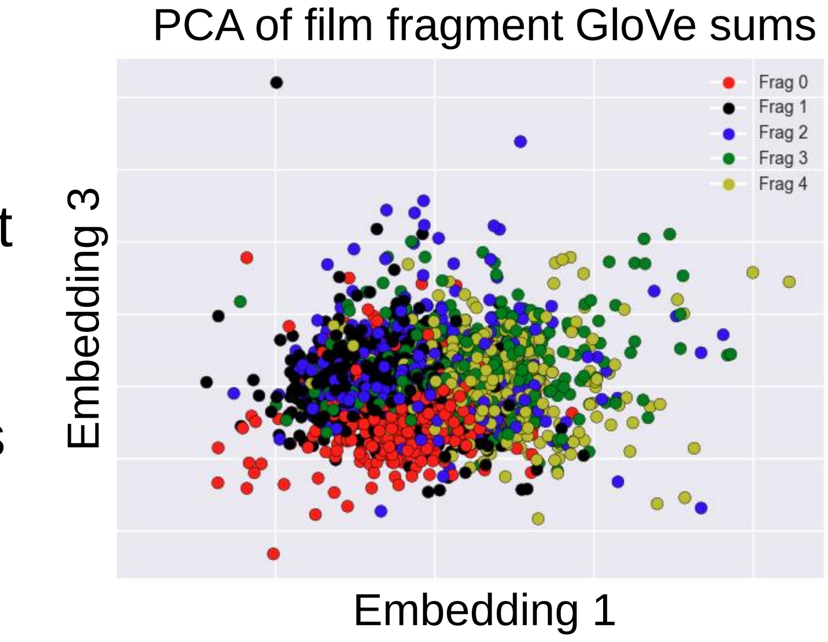
The *accuracy* is the rate at which fragments are matched to the correct document position.
τ is the fraction of correct pairwise fragment orderings: $1 - (\#\text{ pairwise inversions}) \div \binom{k}{2}$

80/20 train/test split		***k*-fragments approach** (*k* = 5)			**Transitional** (5-sent. frags)
		Naïve Bayes *n*-grams	Uns. *K*-means GloVe sums	Sup. *K*-means GloVe sums	Logistic Regression
Film	Train acc. %	39	15	**45**	31
	Test acc. %	39	15	**45**	31
	Train 100τ	20	6	**23**	20
	Test 100τ	21	6	**24**	21
People	Train acc. %	**47**	17	46	33
	Test acc. %	45	17	**45**	31
	Train 100τ	**26**	7	**26**	23
	Test 100τ	25	7	**26**	21
Cities	Train acc. %	**37**	22	33	25
	Test acc. %	**39**	9	36	20
	Train 100τ	**17**	23	18	16
	Test 100τ	**19**	10	18	12
Any	Train acc. %	33	20	**35**	32
	Test acc. %	32	20	**35**	28
	Train 100τ	17	0	**17**	20
	Test 100τ	14	0	**16**	19

Discussion

Making coherent text is a difficult, but we made promising headway in reconstructing documents. Training and testing on articles with similar structure, e.g., people, yielded the best results: test τ = 0.26. We were surprised by the performance of *K*-means using GloVe sums, but after conducting PCA on the sums we observed clustering (see right). The *k*-fragments position semantics assumption is correct to some extent. By comparison, the transitional approach is strictly harder as the number of fragments exceeds *k*, so the results were worse.

PCA of film fragment GloVe sums



Future Work

Advanced transitional probability: use recurrent neural networks to find P($frag_i$ | $frag_{i-1}$, …, $frag_1$)
Improved fragment embeddings: unsupervised representation learning for fragment embeddings
Variable fragment length and number: relax constraints on fragmentation

1. L. Logeswaran, H.Lee, & D. Radev. "Sentence Ordering using Recurrent Neural Networks." *arXiv*:1611.02654. Nov. 2016.
2. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation
3. Wikimedia Dumps. "English Wikipedia Dataset." Meta.Wikimedia.org/wiki/data_dump_torrents