Comparison of Corpora through Narrative Structure

Dan Simonson

Georgetown University

des62@georgetown.edu — @thedansimonson — thedansimonson.com

2014 May 14

Overview

How this started?

[1]

[1] http://www.eurweb.com/wp-content/uploads/2013/10/capitol-hill-shooting.jpg

2013 Capitol Hill “Shooting”

News sources presented the events in two different ways:

- Fox News/MSNBC: Strong Moral Leaning, Provocative
- CNN/New York Times: Ambiguous Moral Leaning, Contemplative

Analysis by hand.

As a computational linguist, I can study 10^6—instead of $10^{0.6}$—documents.

Goals

- Compare police narratives *en masse* from different news sources.

Available data: New York Times Corpus [Sandhaus 2008].

Goals

- Compare police narratives *en masse* ~~from different news sources.~~
- ...in the New York Times, from January 1, 1987 and June 19, 2007.

What might have caused changes in the way police are talked about in that time frame?

Goals

- Compare police narratives *en masse* ~~from different news sources.~~
- ...in the New York Times, ~~from January 1, 1987 and June 19, 2007.~~
- ...before and after September 11th. [Balko 2012]

Were there significant changes in how police were discussed before and after September 11th?

What is a narrative schema?

- [Chambers and Jurafsky 2008], [Chambers and Jurafsky 2009]

Examples: Firing of Employee and Executive Resigns

Y	accused	X	W	joined	V
X	claimed	_	W	served	_
X	denied		W	oversaw	_
Y	dismissed	X	W	resigned	

X (employee), Y (supervisor), W (executive), and V (company) represent co-referent in the argument slots.

Previous Work

[Chambers and Jurafsky 2008] and [Chambers and Jurafsky 2009]

- Use coreference chains and dependency parses to extract narrative event chains/schema.
- [Webber, Egg, and Kordoni 2012] cites those as the first such work in a long time, and a few follow ups.

An Example

Our whole corpus:

"he barricaded himself in."

Coreference

"he barricaded himself in."

- $\{\text{he}_0, \text{himself}_2\}$

Dependencies

"he barricaded himself in."

- $\{(\textit{subj}, \text{barricade}, \text{he}), (\textit{obj}, \text{barricade}, \text{himself})\}$

Narrative Event Chain

"he barricaded himself in."

- $\{(\text{barricade}, \mathit{subj}), (\text{barricade}, \mathit{obj})\}$

Originally included a temporal ordering step.

- This was irrelevant here.

Pointwise Mutual Information

When two events are independent:

$$p(x)p(y) = p(x, y) \tag{1}$$

Therefore,

$$pmi(x, y) \equiv \frac{p(x, y)}{p(x)p(y)} \tag{2}$$

In our case, x and y are verb-dependency pairs.

- $p(x, y)$ indicates the probability that a coreference chain is an argument to both x and y.

Using *pmi*

The next candidate verb-dependency pair:

$$\max_{j:0<j<m} \sum_{i=0}^{n} pmi(e_i, f_j) \tag{3}$$

e_i is our chain as it stands. f_j that maximizes is our next candidate.

Coreference is Slow

"Honey, I shrunk the corpus!"

- 1.6 million documents was too many to process in a reasonable time.
- Shrinking the corpus in a systematic way may help us focus on our target data as well.

Corpus Shrinking

- Keyword Selection — is the word "police" in the document? (That's it.)
- Categorical Selection

Categorical Selection

Table: Number documents per category retained from the "police" subset. There were many more; they were not explicitly excluded.

Murders and Attempted Murders	22,640
Crime and Criminals	21,315
Terrorism	17,565
Demonstrations and Riots	9,443
Accidents and Safety	8,742
World Trade Center (NYC)	7,128
Blacks	6,786
Law and Legislation	6,695
Violence	6,059
Police Brutality and Misconduct	4,823
Attacks on Police	2,583
Frauds and Swindling	2,247
Cocaine and Crack Cocaine	1,487
Organized Crime	1,434
Serial Murders	1,136
Suburbs	296
Noise	294
Prison Escapes	270

Time Periods

Four time periods:

- Three years in size (1095 days, technically).
- Retained two before/two after 9/11.

People Selection

Still too big!

- People Metadata from NYT Corpus [Sandhaus 2008].
- Chose people who appeared 10 and 15 times per time period.
- Kept the documents they appeared in.

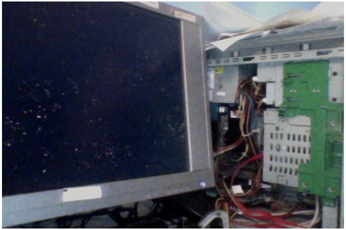
Most questionable reduction?

- Ensures we have coherent narratives from the handful of individuals selected.
- Worst case scenario, this is equivalent to random selection.

Coreference Chains, Dependency Parses, etc.

- Stanford CoreNLP[2]
- corenlp-python[3] serverlet — allowed for marginally graceful crashes.

Took two weeks on "Wisdom."



[2] http://nlp.stanford.edu/software/corenlp.shtml
[3] https://pypi.python.org/pypi/corenlp-python

Counter-Training

Inspired by [Yangarber 2003]:

- Document classification algorithm
- If multiple categories choose the same document, then that selection is a poor fit for the category.

This intuition is applied to our verb-dependency pairs and growing event chains.

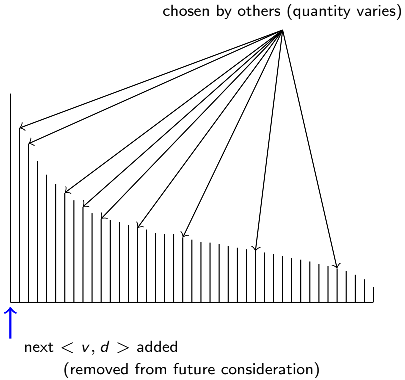
Probabilistic Rationale

How can this be justified?

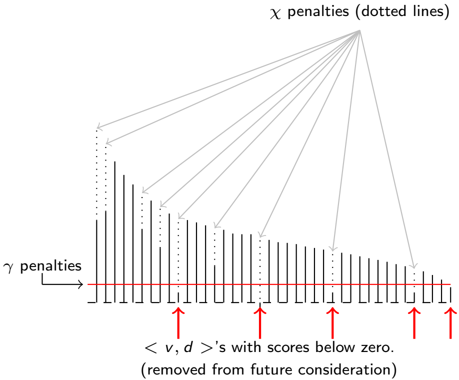
- We're producing a discrete representation of a continuous space.
- A $< v, d >$ choice may assign a disproportionate probability mass to our approximation.
- Penalizing prevents that disproportionality from becoming absurd.

Creates maximally distant schema.

Counter-Training



Counter-Training



Starting the Train

Something has to seed the counter-training process.

- Single $< v, d >$ Pairs
- Existing $< v, d >$ Pairs

[Chambers and Jurafsky 2008] is a bit vague about this aspect of the process:

$$\max_{j:0<j<m} \sum_{i=0}^{n} pmi(e_i, f_j) \tag{4}$$

Comparing Event Chains

There are many levels of comparison:

- verb-dependency pair $\times$ verb-dependency pair
- event chain $\times$ event chain
- corpus $\times$ corpus

Verb-Dependency Pair

- WordNet [Princeton 2010]
- Leacock-Chodorow Distance: $-\log \frac{p_{a,b}}{2D}$ [Leacock and Chodorow 1998]
- Penalize for mismatched dependencies—scale the path length.

Event Chains and Corpora

- Assume the closest possible match.
- corpora : event chains :: event chains : verb deps

$$\underline{\text{this level}}(A, B) = \text{norm} \sum_{b \in B} \max_{a \in A} \underline{\text{lower level}}(a, b) \tag{5}$$

Tentative Results

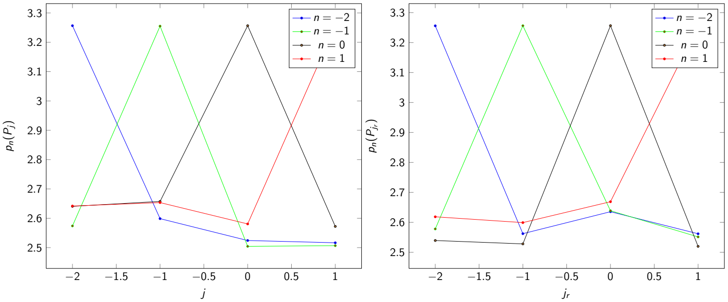


Figure: $p_n(P_j) = period(P_n, P_j)$, where n refers to period index of the series under consideration. Each series contains a common i value. The peaks are where $n = j$ or $n = j_r$—e.g. self-comparison. The left plot represents the `periodic` test; the second, the `randomized` test. The `randomized` i values are denoted with the subscript r.

Hypothesis? Nay!

The distributions are about the same.

- `periodic`: 2.750 ($\sigma = 0.306$)
- `randomize`: 2.752 ($\sigma = 0.304$)

On the bright side, this indicates that the narrative schema are quite stable across the New York Times!

Event Chains? Yay!

Table: Computation of mean of the difference set D for each test. $p_n(P_i)$ is short-hand here for $period(P_n, P_i)$. D_p and D_r refer to the difference set for the `periodic` and `randomized` tests, respectively. Each $d \in D$ is $p_n(P_i) - p_n(P_j)$ with respect to the row it is contained in.

periodic						randomized					
n	i	j	$p_n(P_i)$	$p_n(P_j)$	$d \in D_p$	n	i	j	$p_n(P_i)$	$p_n(P_j)$	$d \in D_r$
-1	-2	0	2.573	2.504	0.069	-1	-2	0	2.577	2.638	-0.060
-1	-2	1	2.573	2.506	0.067	-1	-2	1	2.577	2.551	0.026
0	-2	1	2.640	2.572	0.068	0	-2	1	2.539	2.519	0.019
0	-1	1	2.657	2.572	0.085	0	-1	1	2.527	2.519	0.008
				Mean of D_p =	0.072					Mean of D_r =	0.002

Continuing Work

Upgrading from event chains to the schema as given in [Chambers and Jurafsky 2009].

- Start with typed event chains.

Examples

Type `PERSON`: comes from the CoreNLP NamedEntityRecognizer/pronouns in the chain.

	arraign	X
	arrest	X
X	brag	
	charge	X
X	plead	

Examples

Type `SELF`: 1st person pronouns appear in the coreference chain.

X	believe
X	feel
X	hear
X	see
X	think
X	want

Examples

Type `judge`: a preferred type.

X	adopt
X	chide
X	come
X	reprimand

Examples

Type `diallo`: a preferred type.

X	crouch	
	fire	X*
	hit	X
	kill	X
	shoot	X
X	stand	

* = PREP, not OBJ

What's Next?

Finalize this project.

- The schema, and improved comparison, should give more interesting results.

Product reviews are stories.

- What are their schema like?
- Can schema similarity classify product reviews?

Infuse schema with other features.

- Modality — are the events described hypothetical?

Acknowledgements

- Dr. Tony Davis — without whose support, faith and guidance, this wouldn't be possible.
- Georgetown University Department of Linguistics
- DC NLP — for letting me speak today!

References

Balko, R. (2012). Rise of the warrior cop: The militarization of america's police forces. (1st ed.). PublicAffairs.

Chambers, N., & Jurafsky, D. (2008, June). Unsupervised Learning of Narrative Event Chains. In ACL (pp. 789-797).

Chambers, N., & Jurafsky, D. (2009, August). Unsupervised learning of narrative schemas and their participants. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 602-610). Association for Computational Linguistics. Chicago

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, 49(2), 265-283.

References

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.
- Princeton University. (2010). About WordNet. WordNet. Princeton University. Retrieved from `http://wordnet.princeton.edu`
- Sandhaus, E. (2008) The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.
- Webber, B., Egg, M., & Kordoni, V. (2012). Discourse structure and language technology. Natural Language Engineering, 18(4), 437-490.
- Yangarber, R. (2003, July). Counter-training in discovery of semantic patterns. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 343-350). Association for Computational Linguistics.