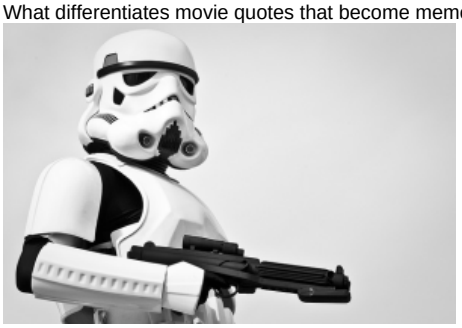
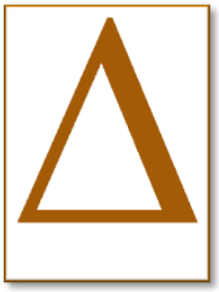
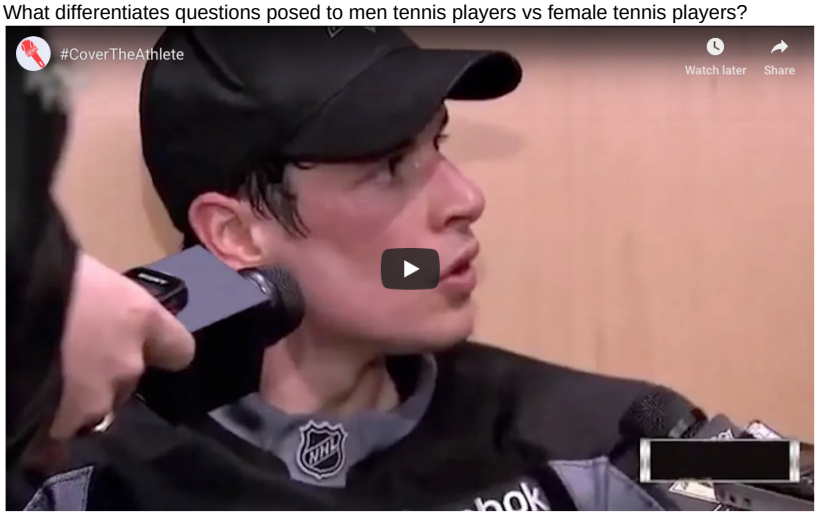
Choice2019

Lillian Lee, Choice 2019 Symposium "Wisdom from Words: Insight from Language and Text Analysis"
This URL: https://confluence.cornell.edu/display/~ljl2/Choice2019

Setting: what makes language type A different from type B?

For various reasons, including an eye towards deploying applications, we ultimately evaluate our hypothesis with *prediction* even though we are personally interested and invested in understanding what underlies the phenomenon being considered.

- What differentiates movie quotes that become memorable vs. those that don't?

- What differentiates tweets that will get many retweets vs. those that don't?

- What differentiates arguments that cause someone to change their mind vs unsuccessful arguments?

- What differentiates questions posed to men tennis players vs female tennis players?


- What differentiates social media posts that will attract controversy (lots of positive and lots of negative feedback) vs. those that won't?

Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. Proc. of the ACL.

> *Abstract:* Understanding the ways in which information achieves widespread public awareness is a research question of significant interest. We consider whether, and how, the way in which the information is phrased --- the choice of words and sentence structure --- can affect this process. To this end, we develop an analysis framework and build a corpus of movie quotes, annotated with memorability information, in which we are able to control for both the speaker and the setting of the quotes. We find that there are significant differences between memorable and non-memorable quotes in several key dimensions, even after controlling for situational and contextual factors. One is lexical distinctiveness: in aggregate, memorable quotes use less common word choices, but at the same time are built upon a scaffolding of common syntactic patterns. Another is that memorable quotes tend to be more general in ways that make them easy to apply in new contexts --- that is, more portable. We also show how the concept of "memorable language" can be extended across domains.

Tan, Chenhao, Lillian Lee and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. Proc. of the ACL.

> *Abstract:* Consider a person trying to spread an important message on a social network. He/she can spend hours trying to craft the message. Does it actually matter? While there has been extensive prior work looking into predicting popularity of social-media content, the effect of wording per se has rarely been studied since it is often confounded with the popularity of the author and the topic. To control for these confounding factors, we take advantage of the surprising fact that there are many pairs of tweets containing the same url and written by the same user but employing different wording. Given such pairs, we ask: which version attracts more retweets? This turns out to be a more difficult task than predicting popular topics. Still, humans can answer this question better than chance (but far from perfectly), and the computational methods we develop can do better than both an average human and a strong competing method trained on non-controlled data.

Tan, Chenhao, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, Lillian Lee. 2016. "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions." Proc. of WWW.

> *Abstract:* Changing someone's opinion is arguably one of the most important challenges of social interaction. The underlying process proves difficult to study: it is hard to know how someone's opinions are formed and whether and how someone's views shift. Fortunately, ChangeMyView, an active community on Reddit, provides a platform where users present their own opinions and reasoning, invite others to contest them, and acknowledge when the ensuing discussions change their original views. In this work, we study these interactions to understand the mechanisms behind persuasion.
>
> We find that persuasive arguments are characterized by interesting patterns of interaction dynamics, such as participant entry-order and degree of back-and-forth exchange. Furthermore, by comparing similar counterarguments to the same opinion, we show that language factors play an essential role. In particular, the interplay between the language of the opinion holder and that of the counterargument provides highly predictive cues of persuasiveness. Finally, since even in this favorable setting people may not be persuaded, we investigate the problem of determining whether someone's opinion is susceptible to being changed at all. For this more difficult task, we show that stylistic choices in how the opinion is expressed carry predictive power.

Fu, Liye, Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. IJCAI workshop on NLP Meets Journalism Best paper award.

> *Abstract:* Gender bias is an increasingly important issue in sports journalism. In this work, we propose a language-model-based approach to quantify differences in questions posed to female vs. male athletes, and apply it to tennis post-match interviews. We find that journalists ask male players questions that are generally more focused on the game when compared with the questions they ask their female counterparts. We also provide a fine-grained analysis of the extent to which the salience of this bias depends on various factors, such as question type, game outcome or player rank.

Hessel, Jack and Lillian Lee. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. Proc. of NAACL.

> *Abstract:* Controversial posts are those that split the preferences of a community, receiving both significant positive and significant negative feedback. Our inclusion of the word "community" here is deliberate: what is controversial to some audiences may not be so to others. Using data from several different communities on www.reddit.com, we predict the ultimate controversiality of posts, leveraging features drawn from both the textual content and the tree structure of the early comments that initiate the discussion. We find that even when only a handful of comments are

available, e.g., the first 5 comments made within 15 minutes of the original post, discussion features often add predictive capacity to strong content-and-rate only baselines. Additional experiments on domain transfer suggest that conversation-structure features often generalize to other communities better than conversation-content features do.

https://www.flickr.com/photos/hyku/3614261299/in/photostream/

http://pixabay.com/en/twitter-tweet-twitter-bird-312464/

http://commons.wikimedia.org/wiki/File:Greek_uc_delta.png, colorized

Screen shot from video at http://covertheathlete.com/

Licensed from Shutterstock

Some features/technologies I like

The Cornell Conversational Analysis Toolkit

Features for: linguistic coordination, politeness strategies, conversation motifs, conversation graphs

Datasets: Wikipedia talk page conversations that (do not) become derailed by personal attacks; dialogs from movie scripts; UK Parliamentary question-answer pairs; Supreme Court oral arguments; Wikipedia talk pages conversations; post-tennis-match press interviews; reddit conversations.

Chenhao Tan's list of hedging phrases, such as "I suspect", "raising the possibility":

This is in the long line of LIWC-like lexicons.

[README] [list itself]

Chenhao Tan and Lillian Lee, "Talk it up or play it down? (Un)expected correlations between (de-)emphasis and recurrence of discussion points in consequential U.S. economic policy meetings", Text As Data 2016

> *Abstract:* In meetings where important decisions get made, what items receive more attention may influence the outcome. We examine how different types of rhetorical (de-)emphasis — including hedges, superlatives, and contrastive conjunctions — correlate with what gets revisited later, controlling for item frequency and speaker. Our data consists of transcripts of recurring meetings of the Federal Reserve's Open Market Committee (FOMC), where important aspects of U.S. monetary policy are decided on. Surprisingly, we find that words appearing in the context of hedging, which is usually considered a way to express uncertainty, are more likely to be repeated in subsequent meetings, while strong emphasis indicated by superlatives has a slightly negative effect on word recurrence in subsequent meetings. We also observe interesting patterns in how these effects vary depending on social factors such as status and gender of the speaker. For instance, the positive effects of hedging are more pronounced for female speakers than for male speakers.

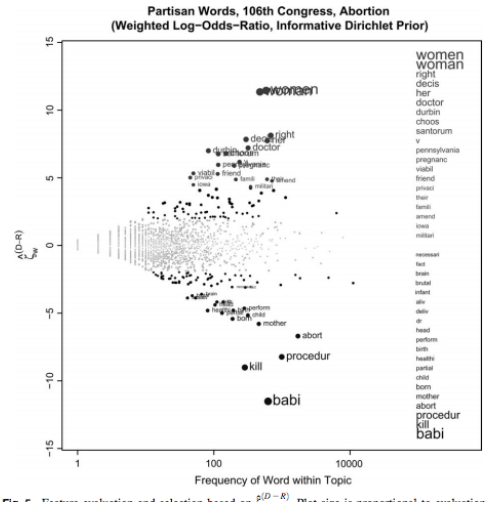
Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, Lillian Lee. "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions." Proc. of WWW 2016

Language models, which assign probabilities P(x) to words, sentences or text units after being trained on some language sample.

These are great for similarity, distinctiveness, visualization.

a. Monroe et al's "Fightin words": what makes two "languages" different?

   Slides and handout from Cristian Danescu-Niculescu-Mizil and my class "NLP and social interaction" : [pptx] [pdf] [handout]



Jurafsky, Dan, Victor Chahuneau, Bryan R. Routledge, Noah A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday* 19(4).

Mark Liberman on Language Log. The most Kasichoid, Cruzian, Trumpish, and Rubiositous words , 2016. The most Trumpish (and Bushish) words , 2015. Obama's favored (and disfavored) SOTU words , 2014. Draft words (descriptions of white vs black NFL prospects), 2014. Male and female word usage , 2014.
Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict . *Political Analysis* 16(4): 372-403. [alternate link]

> Abstract: *Entries in the burgeoning "text-as-data" movement are often accompanied by lists or visualizations of how word (or other lexical feature) usage differs across some pair or set of documents. These are intended either to establish some target semantic concept (like the content of partisan frames) to estimate word-specific measures that feed forward into another analysis (like locating parties in ideological space) or both. We discuss a variety of techniques for selecting words that capture partisan, or other, differences in political speech and for evaluating the relative importance of those words. We introduce and emphasize several new approaches based on Bayesian shrinkage and regularization. We illustrate the relative utility of these approaches with analyses of partisan, gender, and distributive speech in the U.S. Senate.*

The method is also described in Section 19.5.1, "Log odds ratio informative Dirichlet prior" of the 3rd edition of Jurafsky and Martin, *Spee ch and Language Processing.*

Slides adapted from slides 85-94 of Cristian Danescu-Niculescu-Mizil and Lillian Lee, Natural language processing for computational social science, Invited tutorial at NIPS 2016 [alternate link: tutorial announcement, slides] for lecture 16 of the class NLP and Social Interaction, Fall 2017.

Code

- Hessel, Jack: FightingWords.
- Lim, Kenneth: fightin-words 1.0.4. Compliant with sci-kit learn and distributed by PyPI; borrows (with acknowledgment) from Jack's version.
- Marzagão, Thiago: mcq.py

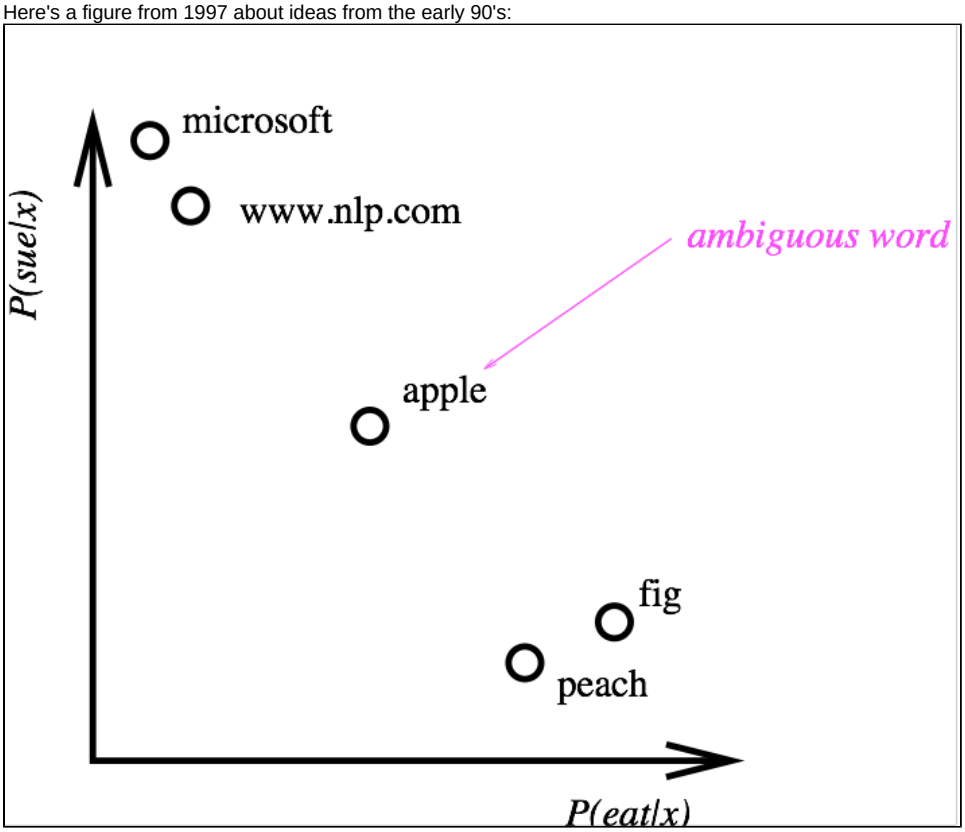
Visualizers

- Kessler, Jason. ScatterText , described Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. ACL System Demonstrations. 2017
- Schofield, Xanda. fightinwords.py (with acknowledgments to Jack Hessel)

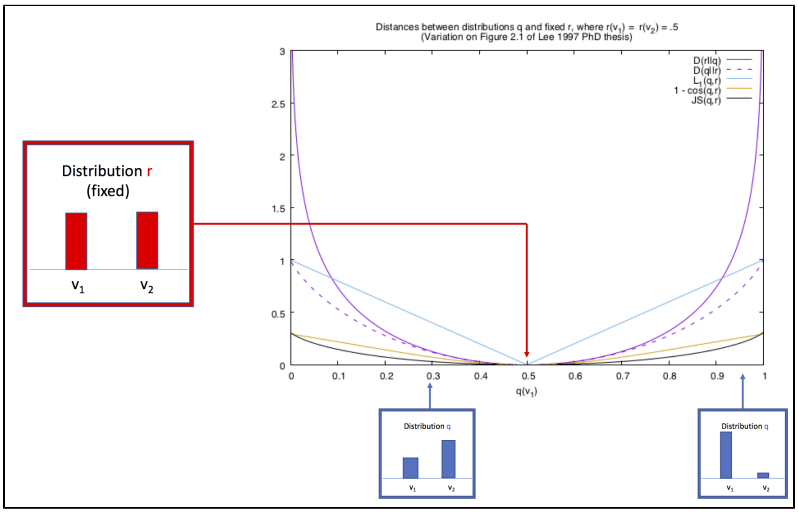
b. Similarity measured on the most frequent words ("stop words") only vs. on the content words
How similar are two language models? The standard measure is the cross-entropy: - p(x) log(q(x)) ; a variant is the KL divergence, p (x) log(p(x) / q(x)) = the cross entropy of p(x) and q(x) minus the entropy of p(x)
c. Similarity of each of A or B to a baseline of "regular" or "null hypothesis" language.

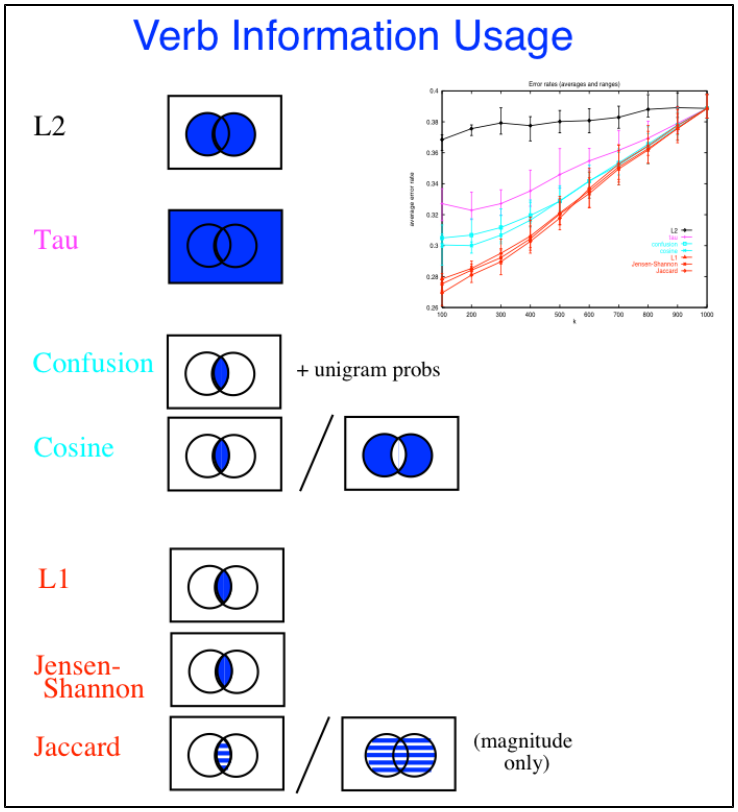
Distributional similarity (word embeddings are the modern version)

Here's a figure from 1997 about ideas from the early 90's:



For references, see the word embeddings section later in this document





Lee, Lillian. 1999. Measures of distributional similarity. Proc. of the ACL, 25--32

... and one feature that I both like and drives me crazy: length

It represents an intuitively slightly ridiculous null hypothesis that often works surprisingly well as a feature, most likely because it correlates with a lot of other features of interest.

Examples: (to be inserted)

A feature-effectiveness test that's caught my eye

Wang, Zhao and Aron Culotta, When do Words Matter? Understanding the Impact of Lexical Choice on Audience Perception using Individual Treatment Effect Estimation. AAAI 2019. [code]

How do we proceed during the age of deep learning, where, for prediction, we don't need to (aren't supposed to) worry about features anymore?

Comparison of hand-crafted features against deep learning on predicting controversial social-media posts

	AskMen	(2)	(3)	(4)	(5)	(6)
HAND-crafted						
Word2Vec						
W2V-LSTM						
BERT-LSTM	☆	☆	☆	○	☆	○
BERT-meanpool-then-linear	○	○	○	☆	○	○
HAND+W2V			○	○		○
HAND+BERT-meanpool-512 then linear	○	○	○	○	○	☆

star = best in column; circle = performance within 1% of the best in column. Columns: different sub-reddits.

Image adapted from Table 2 of Hessel, Jack and Lillian Lee. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. Proc. of NAACL.

HAND = "for the title and text body separately, length, type-token ratio, rate of first-person pronouns, rate of second-person pronouns, rate of question-marks, rate of capitalization, and Vader sentiment. Combining the post title and post body: number of links, number of Reddit links, number of imgur links, number of sentences, Flesch-Kincaid readability score, rate of italics, rate of boldface, presence of a list, and the rate of word use from 25 Empath wordlists.

Word embeddings - now contextual/polysemy-aware!

Question/proposal : where is the word embedding version of LIWC? ("Can we BERT LIWC?").

Fast, Ethan, Binbin Chen, Michael S Bernstein. Lexicons on demand: Neural word embeddings for large-scale text analysis. IJCAI 2017. *Abstract*: Human language is colored by a broad range of topics, but existing text analysis tools only focus on a small number of them. We present Empath, a tool that can generate and validate new lexical categories on demand from a small set of seed terms (like "bleed" and "punch" to generate the category violence). Empath draws connotations between words and phrases by learning a neural embedding across billions of words on the web. Given a small set of seed words that characterize a category, Empath uses its neural embedding to discover new related terms, then validates the category with a crowd-powered filter. Empath also analyzes text across 200 built-in, pre-validated categories we have generated such as neglect, government, and social media. We show that Empath's data-driven, human validated categories are highly correlated (r=0.906) with similar categories in LIWC.
Smith, Noah A. 2019. Contextual word representations: A contextual introduction. arxiv:1092.06006, version 2, dated Feb 19. 2019.
Twitter commentary regarding the history as recounted in the above (Naftali Tishby and yours truly are among the "& co." referred to by Robert Munro): [1] [2] [3]Goldberg, Yoav. 2017. Neural network methods for natural language processing. Morgan Claypool. Earlier, shorter, open-access journal version: A primer on neural network models for natural language processing: JAIR 57:345--420, 2016.

Language modeling = the bridge?

Note that the basic units might be characters or unicode code points ("names of character") instead of words.

Thanks to Jack Hessel and Yoav Artzi for the below. Paraphrasing errors are my own.

The best off-the-shelf language model right now (caveat: this is a very fast-moving field) is the 12-or-so layer GPT-2, where GPT stands for Generative Pre-Training. [code] [(infamous) announcement] [hugging face's reimplementation of pre-trained GPT-2]

But a single-layer LSTM trained from scratch, with carefully chosen hyperparameters, is still often a very strong baseline, especially with small data (around 10K samples).

Both BERT and GPT seems to transfer well via fine-tuning to small new datasets, at least in expert hands. [code] [Colab] [hugging face's reimplementation of pre-trained BERT] [announcement]

The Giant Language model Test Room (GLTR) can be used for analyzing what a neural LM is doing, although its stated purpose is to enable "detect automatically generated text".

Devlin, Jacob, Ming-wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proc. of NAACL. [arXiv version]

Rush, Sasha, with VIncent Nguyen and Guillaume Klein. April 3, 2018. The annotated transformer — interpolates code line-by-line for Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, ukasz Kaiser, and Illia Polosukhin, 2017. Attention is all you need. Proc. of NIPS. [arxiv version]

Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, Sutskever, Ilya. 2019. Language models are unsupervised multitask learners. Manuscript. (The GPT-2 paper)
Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. April 21, 2019. BERTScore: Evaluating Text Generation with BERT. arxiv version 1. [code]
Belinkov, Yonatan and James Glass. 2019. Analysis methods in neural NLP. TACL 7:49–72. [supplementary materials]