Semantic Comparisons for Natural Language Processing Applications

Lucy H. Lin

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2021

Reading Committee:
Noah A. Smith, Chair
Yejin Choi
Tim Althoff

Program Authorized to Offer Degree:
Computer Science and Engineering

© Copyright 2021

Lucy H. Lin

University of Washington

Abstract

Semantic Comparisons for Natural Language Processing Applications

Lucy H. Lin

Chair of the Supervisory Committee:
Professor Noah A. Smith
Computer Science and Engineering

For social scientists and other data practitioners, the abundance of available digital text data is a rich potential source for understanding social phenomena. As a result, practitioners have increasingly used text analysis methods on relevant corpora to help answer their substantive research questions; common abstractions for these analyses include text classification, topic modeling, and fixed keyword matching. While these tools are powerful, they impose strong assumptions about the structure of human language (e.g., documents as bags of words), and as a result limit the kinds of inferences that practitioners can draw from corpora. On the flip side, richer models trained on large corpora provided by the natural language processing community do not necessarily transfer to the needs of practitioners' applications.

In this work, we propose semantic comparison as another lens for study-

ing social phenomena in text data. We introduce two novel applications of semantic comparison methods for which standard abstractions are insufficient. First, we demonstrate the utility of finding semantic matches of a query sentence in a broader corpus through two case studies: community recovery after the 2010-2011 Christchurch, New Zealand earthquake sequence, as expressed in local news text; and policy attitudes in the United States Congress across 2000-2013, as expressed in archived websites from the `.gov` domain. We discuss model selection and end-user challenges involved, and introduce a procedure (nearest neighbor overlap) to compare sentence embedder behavior in the context of a corpus.

Second, we discuss sensationalism in medical journalism and the possible utility of NLP — particularly semantic comparison — in identifying sensationalized text. We survey past studies across communications, medicine, and psychology to illustrate the complexity of how and why sensationalism manifests in the health communications pipeline. In doing so, we critique the common NLP setup of attempting to label social phenomena in text with high accuracy and provide recommendations for developing user-facing NLP systems that seek to identify or reduce the occurrence of sensationalism.

Acknowledgements

I am incredibly grateful to many people for their support through the years:

My advisor, Noah Smith, for taking a chance on me, giving advice that I never regretted taking, and being the optimistic one in this relationship. This thesis would not exist but for his patience that never turned into apathy; this is especially true of Chapter 4, which was more than six years in the making. Bonus points for having great taste in music (because it is similar to mine) and cocktails (same reason).

The rest of my thesis committee — Yejin Choi, Tim Althoff, and Emma Spiro — for their invaluable feedback throughout this process.

My collaborators, mentees, and fellow members of Noah's ARK for thoughtful discussions and being such fun to work with; special thanks to: Tal August, Dallas Card, Ethan Chau, Elizabeth Clark, Jesse Dodge, Sarah Dreier, Saadia Gabriel, Emily Gade, Kelvin Luu, Scott Miles, Phoebe Mulcaire, Deric Pang, Maarten Sap, Sofia Serrano, and Swabha Swayamdipta.

All the friends who were wonderfully supportive throughout my PhD, including the espresso room crew — Emily Furst, Amrita Mazumdar, Kira Goldner, and Maarten Sap — who are responsible for any pop culture knowledge I have; Sofia Serrano, for our weekly shawarma lunches that always made my day better; and Jess Bawa & Mike Fitz, who never doubted that I could pull this off from the start.

And finally, my family — Mom, Dad, Patrick, and SoYoung — for whom words are not enough.

Funding sources. The research in this thesis was funded in part by a NSF Graduate Research Fellowship; Chapter 2 was also funded by NSF Award #1541025.

Land acknowledgment. The research in this thesis was performed on the ancestral lands of the Coast Salish people, including the Duwamish past and present, to whom I extend my gratitude.

DEDICATION

To the tides that lift my ship
and the stars that guide me home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For social scientists and other data practitioners, the abundance of available digital text data is a rich potential source for understanding social phenomena. Local news reports on-the-ground impacts to residents long after a disaster has occurred; legislators promote their policy positions through floor speeches, newsletters, and their official websites; and medical findings are pipelined from researchers to the broader public through peer-reviewed publications, press releases, and news. As a result, text is increasingly being considered as a data source for substantive research questions by practitioners.

When using text analysis methods, practitioners often fall back to known abstractions, such as text classification, topic modeling, and {keyword, n-gram, regular expression} matching [Grimmer and Stewart, 2013; Gentzkow et al., 2019]. These tools are powerful but require assumptions that limit the kinds of inferences that can be made. For example, standard text classification requires an up-front definition of what the classes are, as well as

data labeled with those classes to train a classification model; keyword or n-gram matching requires the creation of dictionaries and assumes that the matched segment is valid regardless of context.

On the other hand, models from the natural language processing (NLP) community are often trained on large datasets and capture a wider variety of linguistic behavior, but may not generalize well to a practitioner's corpus or specific application. Models may be trained on data from a different domain (e.g., news versus political speeches), capture unintentional artifacts from the training data instead of the desired property [Gururangan et al., 2018], or, if retraining the model on the desired corpus, make assumptions about the amount of training data available.

In this work, we propose *semantic comparison* as another lens for studying social phenomena in text data, and in doing so attempt to bridge the gap between NLP modeling and downstream practice. We introduce two novel applications of semantic comparison methods for which standard abstractions are insufficient.

In Chapter 2, we consider the scenario where a user seeks to identify occurrences of an idea in a text corpus. We introduce a framework based on semantic matching of a proposition query and sentences in the corpus, and then discuss considerations involved in selecting a matching function. We demonstrate the applicability of semantic matching through two case studies in collaboration with domain experts: community recovery after the 2010–2011 Christchurch, New Zealand earthquake sequence, as expressed in local news text in the following five years (§2.3); and policy positions in the United States Congress across 2000–2013, as expressed in archived

websites from the `.gov` domain (§2.4). One of the key challenges is model selection without the benefit of annotated training data in the application domain; to ameliorate this, in Chapter 3 we propose a method (nearest neighbor overlap) for comparing sentence embedder behavior in the context of a corpus.

In Chapter 4, we explore the possibility of using semantic comparisons to identify sensationalism in medical journalism. We survey past studies across communications, medicine, and psychology to illustrate where and how sensationalism manifests in the health communications pipeline, the incentives involved, and possible interventions; in doing so, we critique the common NLP setup of attempting to label social phenomena in text with high accuracy. We provide recommendations for development of end-user NLP systems that seek to identify or reduce the occurrence of sensationalism.

Chapter 2

Semantic Matching & Measurement

Consider the following possible end-users: (i) historians of science tracking expression of the idea that "vaccines cause autism" after the 1998 study in *The Lancet* making this claim; (ii) political scientists tracking stated policy positions by legislators, like "welfare programs help needy American families"; and (iii) public servants seeking to understand the challenges facing a community after a disaster by tracking claims like "dealing with authorities is causing stress and anxiety."

What all of these examples have in common is that a user specifies a natural language *proposition query* (an idea of interest likely to occur in their

The work in this chapter draws from three papers: Lin et al. [2018a] and Lin et al. [2018b] introduce semantic measurement as a viable framework for hypothesis generation (§2.1-2.2), with recovery from the 2010–2011 Christchurch earthquake sequence as a case study (§2.3); Dreier et al. [in prep] applies this abstraction towards identifying policy positions in Congressional websites (§2.4), as part of a broader effort to understand the role of religiosity in United States politics.

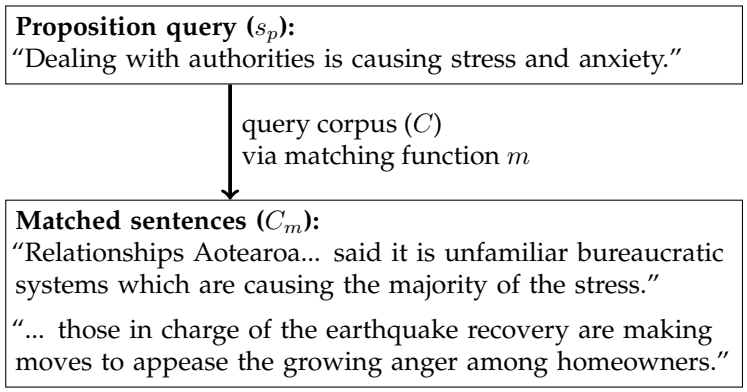


Figure 2.1: An example of semantic matching in the domain of natural disaster recovery.

text collection) and want to identify all expressions of this idea. See Fig. 2.1 for an example.

Tracking and idea measurement has long been considered in a more exploratory way using topic modeling and other unsupervised methods [e.g., Blei and Lafferty, 2006]; however, these do not allow specifying a specific proposition query. In contrast, other work allowing user-specified queries has often done so at the expense of idea complexity, such as through keywords or short phrases [Michel et al., 2011]. Given that natural language has many ways of expressing the above proposition query examples, exhaustive enumeration of exact matches is infeasible.

Instead, we introduce a framework based on semantic matching between a proposition query and sentences in the corpus of interest (§2.1). It is an open question what kinds of semantic matching methods will be required to fulfill the information needs of different kinds of users; we turn to work in proposition-level semantic comparisons (e.g., semantic similar-

ity, natural language inference) and sentence representations as a starting point (§2.2). We provide two case studies in collaboration with domain experts as example applications of semantic matching and demonstrate their potential for evidence gathering and hypothesis generation (§2.3–2.4). Finally, we discuss common findings between the two case studies and draw connections to related work (§2.5–2.6).

2.1 Problem formulation

We formalize the semantic matching problem as follows. Let C denote a corpus consisting of a collection of documents, each a list of sentences (individually denoted by s). s_p will be the proposition query, also a sentence.

The goal is to find sentences $s \in C$ such that s expresses the idea contained in s_p.[1] To do so, we assume that sentences $s \in C$ will be ranked by some function $m(s_p, s)$ and the top-n scoring sentences will be returned to the user as the set C_m. This setup is quite similar to sentence-level text retrieval [Balasubramanian et al., 2007], except that the user is assumed to be interested in the full set C_m, rather than answering a specific information need using *any* relevant match.

We note that our approach assumes segmentation at the sentence level, but alternative formulations (where the expression of an idea may span several sentences or only a clause or phrase in a sentence) could also be considered. We do not use document structure and metadata to help iden-

[1] What it means for a sentence to "express the idea" contained in the proposition query is dependent on the user and their goals; we will see this vary between the two case studies later in this chapter.

tify matches, though those could be an interesting source of information in future work.

2.2 Semantic matching models

In this section, we describe a number of considerations in selecting the semantic matching model $m(s_p, s)$. We will later see these come into play for the two case studies in §2.3 and §2.4, which have different desiderata and constraints.

2.2.1 Modeling considerations

User scenario. We envision that this setup will help a data practitioner or domain expert explore a corpus as follows. Given a text corpus, the user writes a set of proposition queries expressing ideas or concepts of interest. The system retrieves semantic matches using the procedure in §2.1, which the user can then examine manually for validity or to better understand aspects of the text corpus. They can iterate over the queries or update the semantic matching model to better identify sentences of interest. Finally, when the user is satisfied that the matched sentences sufficiently express the queries, they can perform an aggregate analysis over the matched sentences.

Semantic relationships. We would like the semantic matching function $m(s_p, s)$ to capture the desired semantic relationship between s_p and s. The precise nature of this relationship can vary by application; examples in-

clude semantic relatedness (to what extent are s and s_p describing associated concepts?), semantic similarity (how close are the meanings of s and s_p?), and entailment (if s is true, then is s_p also true?). Even within these categories, the details may differ in practice: for example, given the sentences (a) "*The cat walked upstairs*," (b) "*The dog walked upstairs*," and (c) "*The cat walked downstairs*," one might consider (a) and (b) to be more similar (as both sentences have pets going upstairs), and another might consider (a) and (c) to be more similar (as both sentences have cats walking on stairs).

Evaluating match correctness. While the semantic relationships above have benchmark datasets available to train and perform preliminary evaluations on, these may not necessarily translate to the semantic matching problem at hand — the target corpus could be of a different domain (e.g., news text vs. image captions in §2.3) or the desired semantic relationship is different. In that case, we may want to:

- Train a new model with domain-specific data: the domain expert specifies the semantic relationship through annotation guidelines or examples, which are used to build data for training; or:
- Use existing models trained on a close semantic comparison dataset, and then have the domain expert evaluate matched output from those models; iterate as needed.

Computational capacity. There are a couple of computational constraints to take into account:

- **Runtime**: how quickly does a system need to get matched sentences for a given query?
- **Storage**: how much space can be taken up by models and precomputed corpus information?

The impact of each of these will increase with the size of the corpus, since some form of scoring function needs to be executed between the query and every sentence.[2]

2.2.2 Architecture types

In this subsection, we describe two common architecture classes for semantic matching, dual encoders and cross encoders, as well as means for pipelining them.

Dual encoder. In the dual encoder[3] setup, we have a sentence embedder $\mathbf{e}(s)$ that takes a sentence as input and outputs a fixed-length embedding. The sentence embeddings for s_p and s are computed separately, and $m(s_p, s)$ is the cosine similarity between their embeddings:

$$m(s_p, s) = \frac{\mathbf{e}(s_p) \cdot \mathbf{e}(s)}{||\mathbf{e}(s_p)|| \cdot ||\mathbf{e}(s)||}$$

[2] In the dual encoder setup described in §2.2.2, one can bypass the linear search through approximate nearest neighbor methods, which involve additional preprocessing for sublinear query time. But even in that case, an embedding function still needs to be run over every sentence in the corpus during preprocessing — a constraint we will run up against in the second case study.

[3] Dual encoders are sometimes referred to as "bi-encoders" or the outdated term "Siamese models."

Algorithm 1 Pipelined dual encoder (filter) + cross encoder (reranker) setup. Hyperparameters include filter width k and output size n.

Input: corpus C, proposition query s_p
  for all $s \in C$ **do**
    obtain filter score $m_f(s_p, s)$
  end for
  take the top k scoring sentences to be C_f
  for all $s \in C_f$ **do**
    obtain reranker score $m_r(s_p, s)$
  end for
  return the top n scoring sentences (C_m)

From there, identifying the k most similar sentences to s_p is simply identifying the k nearest neighbors of $\mathbf{e}(s_p)$ in the set of embedded corpus sentences.

This procedure offers the benefit of being very fast at query time: given a fixed corpus, sentence embeddings need only be precomputed once, and computing cosine similarity on normalized vectors is simply matrix multiplication. Furthermore, this procedure does not require annotated pairs of sentences for training $m(s_p, s)$; the only training data necessary is for $\mathbf{e}(s)$.

Cross encoder. In a cross encoder setup,[4] s_p and s are joint inputs into m at runtime, rather than having representations computed separately. This offers the ability to capture interactions between the two sentences, but with a higher runtime cost; as a result, cross encoders are often impractical in information retrieval settings on their own.

[4]Cross encoders (or "cross-encoders") are also referred to as "cross-attention encoders," especially in the context of transformer-based models (e.g., BERT); to avoid adding yet another descriptor, we abuse the term here to more broadly mean joint modeling of the two inputs.

Pipelining. Given the performance-runtime tradeoff between the two setups, a common approach is to pipeline them: the dual encoder acts as a fast filtering step to quickly eliminate the majority of sentences unrelated to the query, and then the cross encoder reranks the top filtered sentences. The full procedure is outlined in Algorithm 1.

2.3 Case study 1: Earthquake recovery

Researchers and public servants are interested in understanding the challenges facing a community after a disaster. However, on-the-ground empirical studies can be expensive to conduct, especially across a multi-year recovery period and a wide variety of variables, and as a result recovery is one of the least understood disaster topics [Smith and Wenger, 2007]. While there have been efforts to characterize aspects of recovery through text data (e.g., news articles, government documents), most analyses have been through manual inspection [McDaniels et al., 2007; Chang et al., 2014b].

In Lin et al. [2018a] and Lin et al. [2018b], we propose that disaster recovery experts may be able to obtain additional data through semantic matching of ideas of interest. As a case study, we examine recovery of Christchurch, New Zealand after the M7.1 September 2010 and M6.3 February 2011 Canterbury earthquakes through semantic matches of proposition queries in local news text. We explore different dual and cross encoder options for semantic matching and evaluate model outputs through a user study of twenty professional emergency managers. We then illustrate how semantic matching output can be used in aggregate to help gain insight

about disaster recovery.

2.3.1 Background

Measuring disaster recovery

Disaster recovery can be quantified or measured in multiple ways [Chang et al., 2014a; duPont and Noy, 2015]. The most common way of measuring recovery is to compare post-event states to immediate pre-event states; less commonly, recovery can be measured by comparing post-event states to some modeled or assumed counterfactual state without a disaster [Kennedy et al., 2008]. However, disasters can manifest substantial changes, adaptation, and restructuring so that impacted communities do not fully return to either a pre-event state or a foreseeable counterfactual state [Miles, 2015]. More realistically, recovery can instead be measured as longitudinal changes to indicators of adaptation or community identity [O'Connor et al., 2011]. This most often takes the form of quantification approaches that emphasize quantity of supply and speed of recovery; however, such approaches do not represent qualitative characteristics of recovery, like negative or inequitable impacts of differential recovery across space and time [Chang et al., 2014a] — which are arguably more likely to be captured by text data.

The use of NLP to understand social, political, and economic processes — aspects of disaster recovery — has become popular with the increase in the volume of data about human communication, including text, audio, and video [O'Connor et al., 2011]. Example applications include automatic

extraction of international events from political context [O'Connor et al., 2013], public opinion measurement from social media posts [O'Connor et al., 2010], sense of place [Adams and Raubal, 2014], and community happiness [Ramírez-Esparza et al., 2012]. There are a growing number of uses of NLP methods to understand topics of disasters [Cohn et al., 2004; Lin and Margolin, 2014; Alam et al., 2018].

2010–2011 Christchurch earthquake sequence

A moderately damaging M7.1 earthquake struck the Canterbury region of New Zealand's South Island on September 4, 2010; the epicenter was located near the town of Darfield, approximately 35 kilometers west of the large city Christchurch. Six months later, on February 22, 2011, another earthquake struck but with the epicenter only 10 kilometers southeast of the central business district of Christchurch. The 2011 earthquake generated more than 7300 felt aftershocks in the first year alone. Measured ground acceleration from the earthquake was the highest ever recorded in New Zealand and one of the highest recorded worldwide at the time of the earthquake.

The February 2011 earthquake had significant impacts on community functioning and well-being in Christchurch, including the deaths of 185 people [Miles et al., 2014; Chang et al., 2014b; Morgan et al., 2015]. New Zealand Treasury estimated the capital cost of the Canterbury earthquakes to be around $40 billion. The high shaking intensity, the simultaneous vertical and horizontal ground movement, and the extreme liquefaction of the February 2011 earthquake caused significant damage: access to 45% of the

4,000 downtown buildings was banned for safety reasons, and 1,000 buildings were marked for demolition. Roughly 7,500 houses in Christchurch required demolition and zoning changes to restrict future construction, while almost 100,000 houses needed repairs.

The February 2011 earthquake damaged and disrupted the infrastructure of the city. Electric power was restored to 98% of occupied homes in less than two weeks of the earthquake; on the other hand, roads and bridges were extensively damaged, as were water and wastewater systems. The Christchurch City Council received over 36,000 water and wastewater service requests in the six months following the earthquake. After those six months, around 800 houses still remained without wastewater service.

In March 2011, the Canterbury Earthquake Recovery Authority (CERA) was established to lead economic, residential, social, ecological, and cultural recovery for the subsequent five years. The Stronger Christchurch Infrastructure Rebuild Team (SCIRT) was formed to rebuild the city's horizontal infrastructure and, similar to CERA, sunset after five years. Funds for the recovery were largely from a combination of sources: government and private insurance, central government, local government (including borrowing), and private savings or debt. New Zealand's Earthquake Commission (EQC), another government organization, provided earthquake insurance at a very low rate to residential policyholders; as a result, insurance played a larger role in recovery compared to other earthquake disasters. At the time of writing, more than a decade later, decisions about recovery and rebuilding in Christchurch still continue.[5]

[5] https://www.theguardian.com/world/2021/feb/22/before-and-after-

2.3.2 Modeling

In this case study, we adopt the pipeline setup described in §2.2.2. For the fast filtering step, we use averaged word embeddings, and for the reranker step we use a model based on syntactic differences between the input sentences.[6]

Dual encoder: Averaged word embeddings

There are many sentence embedders we could choose from;[7] we introduce a general procedure for comparing sentence embedding models in the context of a corpus in Chapter 3. In these case studies, we use a simple construction method inspired by work on paraphrase [Wieting et al., 2016] and averaging networks [Iyyer et al., 2015]; each sentence is simply represented as the average of its word embeddings. While this throws away word order, in practice, averaged word embeddings are a reasonable and fast baseline in semantic similarity tasks [Arora et al., 2017].

Of course, the choice of pretrained word embeddings could have a large effect on the quality of a semantic matching system, so we examine two options. We first consider 300-dimensional paraphrastic word embeddings generated by Wieting et al. [2016]; we select these because they were de-

how-the-2011-earthquake-changed-christchurch

[6] Although not directly relevant to the applications that are the focus of this paper, we note that sentence pairs in the Stanford Natural Language Inference (SNLI) corpus, which we use to train the reranker, tend to obtain higher scores from the dual encoder models than sentence pairs from the studies. The high similarity within SNLI sentence pairs is also supported by Gururangan et al. [2018]. We take this as encouraging evidence for performing this filtering step before applying the SNLI-based models described later in this section.

[7] §3.2 surveys several classes of sentence embedders, many of which did not exist when this work was conducted.

signed specifically for semantic similarity between sequences. We also consider the widely used word2vec vectors [Mikolov et al., 2013], which are trained on Google News and contain 300-dimensional vectors for approximately 3 million words.[8] These are of interest because they are relatively fast to train on large amounts of data. Because they are derived from unstructured news text, they are more likely to contain proper nouns and entities of interest than the paraphrastic vectors, which are trained on the Paraphrase Database [Ganitkevitch et al., 2013].

Preliminary evaluation

Before investing in a user-focused evaluation, we exploit existing corpora labeled for related tasks (CNN/Daily Mail Reading Comprehension and Media Frames Corpus) to test the effectiveness of simple averaged word vector models. In both cases, our evaluation differs from the tasks originally introduced by the dataset, because our interest is in semantic matching applications.

CNN/Daily Mail. The CNN/Daily Mail Reading Comprehension dataset [Hermann et al., 2015] contains 93k articles from CNN and 220k articles from the Daily Mail. Each instance consists of an article, a query (constructed from bullet point summaries in the original articles), and an answer to the query. For each instance, we take the proposition query s_p to be its query and the "corpus" C to be the set of sentences in its article; the model is asked to find the sentence which contains the entity in the

[8] https://code.google.com/archive/p/word2vec/

answer.[9] This problem is simpler than our original problem: s_p is only being matched against sentences in one document (average 30 sentences), rather than an entire corpus. Nonetheless, this dataset provides an initial testbed.[10] We emphasize that we are interested only in identifying relevant sentences, and not in finding the answer-entity. We consider a sentence relevant if it contains the correct answer.

Media Frames Corpus. The Media Frames Corpus [Card et al., 2015] contains several thousand news articles related to three policy issues (immigration, tobacco, and same-sex marriage). These articles were annotated with fifteen "framing dimensions" according to a codebook developed by Boydstun et al. [2014].[11] The texts were annotated by a team of political science experts according to the framing dimensions; any span of text could be labeled with any frame, and overlapping is possible. An example span of text annotated with the *quality of life* frame is "we hear statistics rather than stories, stories of lives mired in human suffering."

Importantly, the codebook includes expert-designed **examples** for each framing dimension. We take proposition queries s_p to be these examples. The intuition is that a sentence in the corpus that matches a codebook example for frame F is also expected to evoke frame F. For instance, "immigra-

[9] The desired entity may appear in more than one sentence, but in general only is present in a small fraction of the total sentences in an article.

[10] Chen et al. [2016] found that in many of the "answerable" cases in their analysis of the CNN/Daily Mail dataset, identifying the most relevant single sentence goes a long way. While this property may work against advancing reading comprehension models, it is ideal for our evaluation.

[11] Readers interested in the details of the framing dimensions are referred to those works; examples of framing dimensions salient in the immigration data include *fairness and equality*, *crime and punishment*, and *cultural identity*.

tion rules have changed unfairly over time" and "allowing unauthorized immigration is unfair to those who apply and wait" are both examples of the *fairness and equality* frame.

In this work, we focus on the immigration-related articles, as the codebook for this subset of the corpus was most complete. From the codebook, we obtain 30 proposition queries across ten framing dimensions (not every framing dimension has examples provided for immigration). The full list of codebook examples used is provided in the appendix (Table A.1).

Because annotated spans can be any part of a sentence, we consider a sentence to be annotated with a frame if any part of it is annotated with that frame. In cases where the corpus annotators disagree on which framing dimension is evoked, we note agreement if any of the annotators has specified the frame of interest.[12] We will examine how well the output from the averaged word vector models aligns with existing frame annotations. We do not expect high recall on this task, since many annotations in the corpus evoke framing dimensions in ways semantically distant from the codebook's examples.

Results. We run each of the models across the train and test partitions of the CNN/Daily Mail corpus, and on the immigration section of the Media Frames Corpus. For the CNN/Daily Mail evaluation, we compute *recall* at different values of n (the number of top-scoring sentences to output) to see how well our models can identify the relevant sentence(s). In contrast, for the Media Frames Corpus, recall is not interesting since matches to frame

[12] As Card et al. [2015] note, some subjectivity in frame annotations is expected, as the same text can be interpreted differently depending on the reader.

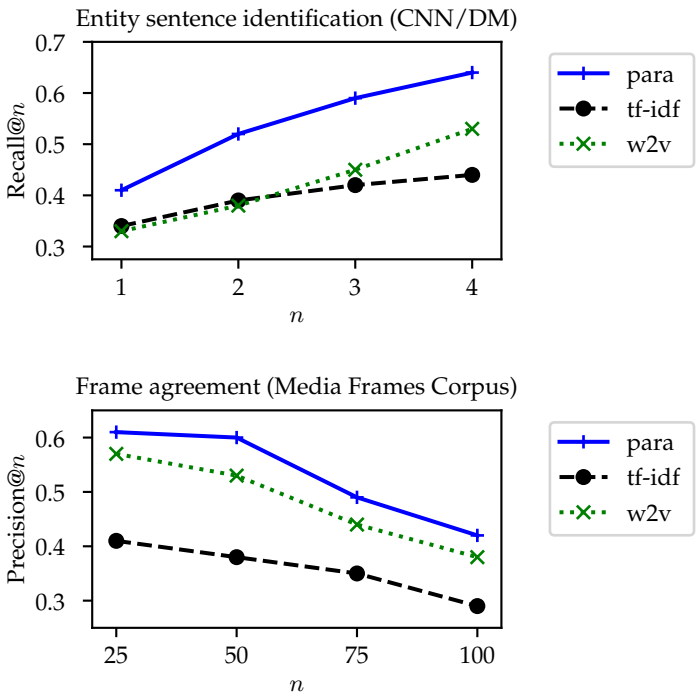


Figure 2.2: Averaged word vector model results for semantic matching tasks based on existing corpora.

annotations will certainly not cover all possible evocations of their frame, so we examine *precision* for varying values of n.

We plot the results in Fig. 2.2. In both tasks, we find that the word-vector-based variants result in improved performance over the tf-idf baseline. (In the CNN/Daily Mail task, the word2vec and tf-idf baselines behave similarly for $n = 1$ and $n = 2$; as n increases, word2vec becomes significantly better.) We also find that the paraphrastic vector model performs better than word2vec, which may be a result of the paraphrastic vectors being trained with semantic similarity tasks in mind. Of course, we expect

that this simple method can be improved with better sentence representations and/or application-specific supervision; we nonetheless consider these results encouraging for a fast filtering step.

Cross encoder: tree edit model

As a starting point for the reranker, we use the tree edit model introduced by Heilman and Smith [2010]. We select this model because it is simple and interpretable, and it was demonstrated to be suitable for a range of semantic similarity problems, including entailment, paraphrase, and answer ranking for question answering.

Base model. We summarize the base model from Heilman and Smith [2010] and refer the reader to the original paper for further details.

For the sentences s and s_p, we first obtain dependency parse trees[13] T and T_p, respectively. We then choose a tree edit sequence (i.e., a sequence of edit operations) that transforms T into T_p. Edit operations include adding nodes (words), deleting nodes, relabeling dependency relations, and so on; the full list is provided in Table 2.1. The edit sequence is found using beam search, with a heuristic function that depends on the lemmas, part of speech tags, arc labels, and whether a node is a left or right child of its parent.

A set of 33 integer-valued features are extracted from the edit sequence. These features include the sequence length and counts of different edit

[13] We use the Stanford CoreNLP pipeline [Manning et al., 2014] to obtain dependency parses, lemmas, and part of speech tags.

Operation	Arguments	Description
INSERT-CHILD	node n, new lemma l, POS p, edge label e, side $s \in \{\textit{left}, \textit{right}\}$	Insert a node with lemma l, POS p, and edge label e as the last child (i.e., farthest from parent) on side s of n.
INSERT-PARENT	non-root node n, new lemma l, new POS p, edge label e, side $s \in \{\textit{left}, \textit{right}\}$	Create a node with lemma l, POS p, and edge label e. Make n a child of the new node on side s. Insert the new node as a child of the former parent of n in the same position.
DELETE-LEAF	leaf node n	Remove the leaf node n.
DELETE-&-MERGE	node n (where n has exactly 1 child)	Remove n. Insert its child as a child of n's former parent in the same position.
RELABEL-NODE	node n, new lemma l, new POS p	Set the lemma of n to be l and its POS to be p.
RELABEL-EDGE	node n, new edge label e	Set the edge label of n to be e.
MOVE-SUBTREE	node n, node m (s.t. m is not a descendant of n), side $s \in \{\textit{left}, \textit{right}\}$	Move n to be the last child on the s side of m.
NEW-ROOT	non-root node n, side $s \in \{\textit{left}, \textit{right}\}$	Make n the new root node of the tree. Insert the former root as the last child on the s side of n.
MOVE-SIBLING	non-root node n, side $s \in \{\textit{left}, \textit{right}\}$, position $r \in \{\textit{first}, \textit{last}\}$	Move n to be the r child on the s side of its parent.

Table 2.1: Tree edit operations from Heilman and Smith [2010].

types; the full list is provided in the appendix (Table A.2). A logistic regression (LR) model is trained on these features.

Neural tree edit model. Given the many successes of non-linear models and the sequential nature of the tree edits, we introduce a neural network variant of the model. We select a tree edit sequence exactly as described above, and then use a LSTM [Hochreiter and Schmidhuber, 1997] that scores based on reading in the tree edits in sequence. Each element in the tree edit sequence is vectorized as the concatenation of:

- A one-hot encoding of the operation type.

- A word-embedding-like vector, in the same space as the word embeddings, that aims to capture the word-embedding-space "difference" between the sentences before and after the edit operation. For example, if a new node is added to the tree (`INSERT-CHILD`, `INSERT-PARENT`), then we use the word embedding for that word. If a node is relabeled (`RELABEL-NODE`) with a new lemma, then we use the difference between word embeddings for the replacement and original word. If a word is deleted (`DELETE-LEAF`, `DELETE-&-MERGE`), then we use the negated embedding of the deleted word. In other cases, we use a zero vector.

This approach allows the model to take lexical and sequential information into account rather than just counts of operations. Note that both approaches make use of syntactic context when representing edits to sentences.

Training. We use the Stanford Natural Language Inference corpus [SNLI; Bowman et al., 2015].[14] SNLI contains approximately 570,000 pairs of sentences (premise and hypothesis); each sentence pair is human-annotated with an *entailment*, *contradiction*, or *neutral* label of the relationship between the two sentences. (As is standard, we ignore examples marked as "unlabeled" due to annotator disagreement.)

For the purposes of our reranking function, we recast the SNLI examples into a binary framework as follows. We treat the premise sentence as analogous to the candidate s and the hypothesis as the proposition query s_p. Premise-hypothesis pairs labeled as entailment are considered positive matches, and those labeled as contradiction or neutral are considered negative matches.

We train three model variants: the original logistic regression (LR) version, and the LSTM using the two pre-trained word embeddings discussed and motivated previously. We use the standard SNLI train/development splits to tune hyperparameters; for the LSTM models, we optimize using Adam [Kingma and Ba, 2014].[15]

2.3.3 Data

We collected 982 earthquake-related articles from New Zealand news websites,[16] spanning 2011 through 2016; all articles are in English. We ob-

[14] This study began before the multi-domain version of the SNLI corpus, MultiNLI [Williams et al., 2018], was released; however, based on post-hoc experiments in §2.3.6, we suspect this would not have made a significant impact. Preliminary testing showed that paraphrase corpora (like the MSR Paraphrase Corpus; Dolan et al., 2004) were a poor fit.

[15] While performance on SNLI specifically is not the goal here, our models perform respectably well on the three-way task (best accuracy is 84.7%).

[16] `http://www.stuff.co.nz` and `http://www.nzherald.co.nz`

tained 20 proposition queries from our domain expert; the queries cover topics like community well-being, infrastructure restoration, decision making, and public opinion. Example queries include:

- "The council should have consulted residents before making decisions."
- "Confidence in Cera has been trending downwards." (*Cera* is short for the *Canterbury Earthquake Recovery Authority*.)
- "Some of the burden on mental health services is caused by lack of housing."

2.3.4 User study evaluation

To evaluate the viability of our approach, we conducted a user study with twenty emergency managers.[17] Emergency managers are state/local personnel responsible for planning, administration, operations, and logistics related to natural and man-made hazard events, and therefore might be interested in relevant ideas found in text.

We evaluated two hypotheses: (1) that adding the tree edit models on top of the averaged word vector ones yields better-quality matches; and (2) that using the LSTM-based tree edit model provides improved performance over the LR-based model.

Our preliminary investigation found that the tree edit models offered no consistent benefit on the existing-corpora tasks (§2.3.2). This is unsur-

[17] The emergency managers were solicited for this study through professional connections of our domain expert. Their judgments of our output were anonymized upon survey completion; their responses to a set of qualitative feedback questions were not. This study was IRB-approved.

prising; the semantic relationships in those tasks are much broader than entailment. We take the 250 top-scoring sentences from both averaged word vector models as "fast filter" output, and rerank them using the LR and LSTM tree edit models.

Study design. Ideally, we would have our users judge how well every candidate sentence matches every s_p. Since expert users and their time are finite, we instead sampled sentences from the following categories for each model: (i) C_m, the 25 highest-scoring sentences from the reranker (i.e., final output); (ii) top(C_f), the 25 highest-scoring sentences from the filter; (iii) rand(C_f), 25 sentences sampled randomly from those in ranks 26–250 from the filter; and (iv) rand($\neg C_f$), 25 sentences sampled randomly from those ranked at 251 or lower by the filter. (Since these are not necessarily semantic matches, we will refer to them as "candidate sentences.")

We gave each user the prompt, "Given an idea sentence, score each candidate sentence on a 1–5 scale based on how well it expresses the idea. The preceding and following sentences for each candidate are provided for context, but please score the quality of only the bolded candidate sentence."[18] We provided users with a sample idea sentence and candidate sentences scored by the same domain expert who supplied the idea sentences (Table 2.2). We also provided score descriptions from 1 through 5 (Table 2.3).

The candidate sentences to be scored were spread among all 20 partic-

[18]For our users' ease of understanding, we used the term "idea sentence" when referring to s_p instead of the more technical "proposition query." In the instruction sheet, we noted that an idea sentence "expresses a relationship between concepts," but did not provide a more formal definition to avoid overly constraining the idea sentences the participants created in the follow-up section.

Idea: *There is a shortage of construction workers.*

Score (1-5)	**Example candidate sentence (bold)** and preceding and following sentences for context.
1	The data was the latest demand and supply information on the Canterbury rebuild and wider recovery, MBIE said. **The quarterly report for Canterbury included analysis on Greater Christchurch Value of Work, Employment and Accommodation projections.** The forecasts were based on Canterbury Earthquake Recovery Authority projections of work to be done on the residential rebuild and repairs, infrastructure and commercial work.
3	Migrants were now filling most of the rising number of construction jobs but beneficiaries moving into work were also contributing, MBIE's quarterly "job-matching" report said. **The construction sector's workload was expected to peak in the December 2016 quarter at a value of about $1.6 billion.** The residential rebuild would run at "elevated levels" from 2015 until 2018 but commercial work would become increasingly important.
5	The additions to the current workforce of 30,000 will mostly work on commercial projects or infrastructure, the Ministry of Business, Innovation and Employment (MBIE) predicts. **Greater Christchurch's labour supply for the rebuild was tight and was likely to remain that way for the next three years.** Migrants were now filling most of the rising number of construction jobs but beneficiaries moving into work were also contributing, MBIEs quarterly "job-matching" report said.

Table 2.2: Example scored candidate sentences provided to user study participants.

Score	**Guidance**: *The candidate sentence...*
1	...is completely unrelated to the idea sentence.
2	...is tangentially related to the idea sentence.
3	...is related to but does not adequately express the idea sentence.
4	...almost expresses the idea sentence.
5	...expresses the idea sentence in its entirety.

Table 2.3: Scoring guidelines provided to user study participants.

ipants; users were not made aware of which model or sentence category the output came from. To allow calculation of inter-annotator agreement, half of the sentences received three judgments (rather than just one). We computed Krippendorf's α for interval data to be 0.784, which indicates

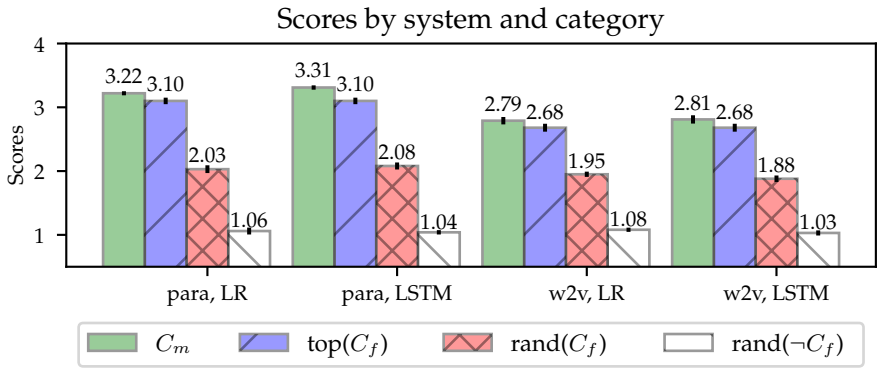


Figure 2.3: Average user study scores for each filter-reranker pair.

reasonable agreement when users rate the same sentence [Krippendorff, 2012].

Results. Our findings, shown in Fig. 2.3, confirm our expectations: users rated poorly-scoring sentences from the filtering step (rand($\neg C_f$)) very low; high-scoring sentences from both the filter and reranker (top(C_f) and C_m) highest, and top-250-ranked sentences from the filter (rand(C_f)) in between. The paraphrastic vectors lead to output receiving better ratings than word2-vec (3.1 vs. 2.7 on average), establishing a baseline that finds sentences "related to, but not (yet) adequately expressing" s_p.

We find that the tree edit model offers some benefit to sentence quality compared to using only the averaged word vector filters. This difference is significant with the paraphrastic filter but within the range of statistical chance with the word2vec-based filter. We also find that the LSTM on tree edit sequences offers slightly better matches than logistic regression; again, this difference is significant with the paraphrastic-based filter but not the

word2vec one. (In fact, the output from the word2vec filter with LR and LSTM tree edit models overlaps at about 85%.)

We note that the lower scores may be in part due to the propositions of interest only appearing sparsely or containing related content which does not fully express the proposition.

User feedback. To gauge interest in the utility of semantic matching systems, we also asked each user to answer an optional set of questions after providing judgments. (All users answered the questions.) We found that (i) 85% were interested in a way to measure ideas in news or other corpora, and (ii) half of the respondents were interested in a follow-up study evaluating semantic matches from idea sentences of their own choosing.

2.3.5 Follow-up study

Our follow-up study was executed similarly to the the original one described above, but with proposition queries solicited from users themselves. Instead of randomly distributing sentences among the follow-up study participants, we gave each user who participated in the follow-up the output for their own proposition queries. There were 18 idea sentences and seven participants in this study. (The full list of idea sentences is provided in Table A.4 in the appendix.) Each participant scored approximately 250 sentences, which were drawn from different parts of the output (as in the original study).

Results. We find that the follow-up study replicates the findings of the original study. The average scores for the top-ranked output (by the averaged word vector models, and reranked by the LR/LSTM models) are generally 0.1–0.2 lower than those in the original study. However, this decrease holds across different model variants, so the relative performance benefits of using paraphrastic word vectors in the averaging model, as well as using the tree edit LSTM model to rerank, still hold. We suspect that the decreased scores are partially a function of some of our users' queries being less applicable to the NZ earthquakes (resulting in fewer possible matches), as the emergency managers' expertise and interests are not centered around that particular disaster or region.

2.3.6 Other entailment models

Because of the limited availability of expert users, we were unable to include a wider range of entailment models in the user study. It is natural to ask whether alternatives to the tree edit model in §2.3.2 would have led to better results. We perform a post-hoc evaluation using the candidate sentences scored by our study participants. We consider two high-performing models: the decomposable attention model [DAM; Parikh et al., 2016] and the enhanced sequential inference model [ESIM; Chen et al., 2017b].[19]

To compare performance of these models in this domain, we take all candidate sentences from both the original and follow-up studies (paired

[19]DAM and ESIM were state of the art when these experiments were conducted; there are now more expressive classifiers based on large pretrained language models [e.g., BERT; Devlin et al., 2019]. An alternate and likely more successful approach would be to finetune such a model on the candidate sentences.

Model	Training data	F_1
Tree edit (LSTM/paraphrastic)	SNLI	55.6
	MultiNLI	51.3
DAM	SNLI	**56.5**
	MultiNLI	55.2
ESIM	SNLI	54.9
	MultiNLI	56.0

Table 2.4: Post-hoc evaluation results with other entailment models.

with their proposition query) and mark them as "entailment" if users scored them greater than or equal to a 4.[20] We split off a set of query-candidate sentence pairs to be a development set; we use these to tune the above models during training (rather than the development sets of SNLI or MultiNLI).

We train these in the two-class setting (entailment vs. combined contradiction and neutral) on SNLI; we use existing public implementations for DAM[21] and ESIM.[22] We also train these and the LSTM version of the tree-edit model on MultiNLI [Williams et al., 2018], a multi-domain version of SNLI.

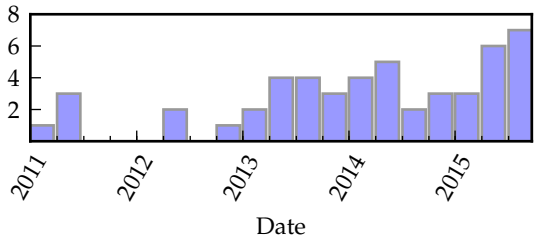
Results. Table 2.4 summarizes the F_1 scores. The relatively low performance from all models, despite high performance on SNLI,[23] indicates that this application is indeed challenging. We also find that training on MultiNLI instead of SNLI does not offer consistent improvement; that is, the multi-domain nature of that dataset does not seem to improve generaliza-

[20] When a candidate sentence was scored by multiple users, we average their scores.

[21] https://github.com/allenai/allennlp

[22] https://github.com/nyu-mll/multiNLI

[23] The SNLI website lists DAM and ESIM as having 85%+ three-way accuracy: https://nlp.stanford.edu/projects/snli/



- The initial trauma may be over but [...] Christchurch will endure at least six months of 'man-made' stressors as the system battles bureaucracy. (5)
- Add to this the growing frustration among the new, youthful leaders of the community who emerged in the wake of the quakes. (3)

Figure 2.4: Example of semantic measurement: frequency (3 month intervals) of the proposition "dealing with authorities is causing stress and anxiety" on the above. Two matched sentences from that histogram are below with user study scores in parentheses.

tion to our data. This suggests that our application requires more than modeling the kinds of sentential entailment captured in those datasets.

2.3.7 Semantic measurement

Here, we propose an application of obtaining semantic matches of ideas: measuring the frequency of an idea in a corpus across an independent variable (e.g., time). To demonstrate this, we return to the example query from Fig. 2.1: "Dealing with authorities is causing stress and anxiety." We select this example because it is not easily expressed through n-grams, and its output was one of the most highly scored in our user study.[24] We take the top 50 matched sentences from the paraphrastic-based model, determine

[24] Additional query-histogram sets are provided in Appendix A.3.

the publication dates of their source articles via metadata, and compute frequencies in bins of three months.

Our system detects an upward trend in expressions of this idea (Fig. 2.4). To our domain expert, this is an interesting yet explainable finding: in the short term after the earthquake, the focus is more on immediate response and relief. It takes time for frustration to set in among the population (e.g., due to dealing with bureaucracy and denied insurance claims). Furthermore, as recovery efforts stretch across years, the media may be more inclined to bring individual stories of continued distress to the forefront.

We note that the computed frequency range is not large, and the candidate sentences provided next to the plot illustrate a subjective range in how well each sentence matches the proposition. Arguably, topics within the proposition tend to be relatively well expressed in the candidate sentences, but less so the (characteristics of) relationships within the sentences. As a result, in this application, we believe that insight about a given recovery proposition is best gained through concurrent inspection of time-based visualization and sentence-level inspection of method outputs. Further, these insights should be put in context with attributes of the analyzed corpus.

As noted above, while a thousand articles is a large number for a researcher to process manually, it is possible that the propositions of interest only appear sparsely, such that only a subset of top-n outputs will ever be good semantic matches. In the next case study, our analysis is based on a corpus several orders of magnitude larger.

2.4 Case study 2: U.S. public policy

Politicians actively express their attitudes towards policy issues in text, such as floor speeches, constituent newsletters, and opinion pieces, and the amount of such text in digital format has rapidly increased. As a result, computational approaches for understanding the dynamics of politics through text data have become increasingly relevant [Grimmer and Stewart, 2013].

In Dreier et al. [in prep], we examine the use of religious rhetoric when legislators advocate for policy positions in the U.S. Congress, as well as whether this behavior varies by partisan affiliation. Part of this work requires being able to extract and quantify mentions of policy attitudes from a broader web corpus (§2.4.2); we describe this process here.

Like with the disaster recovery case study, enumerating and strictly matching against all possible phrasings of a policy attitude rapidly becomes infeasible. For example, we might be interested in expressions of the policy attitude "government should enforce immigration laws and secure U.S. borders." Semantic matches of this policy attitude found in our corpus include:

- "America needs secure borders and enforcement of our immigration laws."
- "Support Arizona and enforce the border laws and LEGAL immigration."
- "This legislation strengthens border security and limits illegal immigration."

A simple keyword- or phrase-based matching process would not necessarily capture these matches as reflecting comparable policy attitudes.

2.4.1 Background

As American partisan politics have become increasingly polarized [Layman, 1999], partisanship and political ideology largely mediate religion's political effects [Yamane and Oldmixon, 2006; Marietta, 2009; Norris and Inglehart, 2011]. For example, while Catholic Democrats tend to prioritize Catholic Social Teachings' commitments to alleviating public social injustices, Republican Catholics tend to focus on private morality issues like abortion [Oldmixon and Hudson, 2008]. However, among other policy areas, we might expect religiosity to intervene to disrupt partisan leanings. A highly religious conservative politician, for example, may deviate from conservative disavowals of poverty-focused development assistance to support foreign aid programs that target impoverished communities (e.g., U.S. President George W. Bush's Emergency Plan for AIDS Relief). Less religious conservatives, on the other hand, would likely maintain low levels of support for foreign aid.

In collaboration with domain experts, we focus on four policy areas which we expect U.S. legislators are likely to discuss in relative proximity to religious rhetoric in documents posted on their official congressional websites. We anticipate that partisanship and political ideology will shape legislators' application of religious rhetoric to social policies. We therefore select two policy areas which we expect Democrats or progressives will be more likely to frame in religious terms and two policy areas which we ex-

pect Republicans or conservatives will be more likely to frame in religious terms.

Policies expected to activate religious rhetoric among Democratic legislators:

- **Social welfare** spending: Legislators will justify domestic welfare spending as compliant with religious edicts to care for community-members in need.

- Distributing poverty-focused **foreign assistance**: Legislators will justify foreign assistance programs as compliant with religious edicts to care for those experiencing extreme poverty or instability around the world, regardless of nationality or identity.

Policies expected to activate religious rhetoric among Republican legislators:

- Regulating **sexual and reproductive behavior**: Legislators will justify regulatory policies to prohibit behavior they view to be incompatible with religious behavioral mandates, specifically with regard to abortion and same-sex marriage.

- Protecting **national security** against external actors: Legislators will justify supporting strong national security policies to protect the safety and stability of their God-fearing constituents and their communities religious ways of life against threats from immigrants and/or those framed as "terrorists."

2.4.2 Data

We use the Internet Archive (IA) .GOV collection of official congressional website data between 2000 and 2013 to understand these dynamics. The broader .GOV corpus contains 1.1 billion website captures from the `.gov` domain, and includes content published on the official congressional websites of U.S. senators, representatives, committees, and other congressional entities; we use a subset of this data from the `house.gov` and `senate.gov` domains (163 million documents).[25]

This data enables a more comprehensive collection of legislators' publicly available policy materials [Esterling et al., 2010], relative to previous political science analyses of a single type of document (e.g., legislators' press releases, newsletters, or floor speeches, like in Maltzman and Sigelman [1996] and Osborn and Mendez [2010]). Material available on a representative's official congressional website may include her floor-speech transcripts, constituent newsletters, opinion pieces, policy platforms, and legislative priorities.[26] We note that the IA data is particularly valuable for capturing content which has been subsequently amended or removed, which is the case for large amounts of Congressional data.[27]

[25] The corpus is primarily English-language text, though legislator websites sometimes have documents in other languages (e.g., Spanish) for the benefit of specific constituent groups.

[26] It excludes material she may have posted on her campaign website material and social media platforms, which would fall under other URL domains.

[27] For example, a Senator or Representative's government website is removed once the person leaves office. While the National Archives and Records Administration takes snapshots at the end of every Congress (i.e., every two years), content uploaded and removed within the two-year period may not be captured in that data, whereas the IA snapshots based on detected changes in website content.

2.4.3 Modeling

Given the size of the IA corpus, we focus on computational efficiency at both preprocessing time and runtime. In the previous case study, we found that the paraphrastic word vector approach had reasonable performance at identifying semantic matches that were at least related to the query; given that the goal here is to match broader policy attitudes (rather than more specific entities), we adopt that approach here.

Preprocessing. We use `spaCy`[28] to segment the entire corpus into sentences, and then pre-compute the sentence embedding for each sentence to minimize computation at query time. We label text as coming from a specific legislator or party website through URL matching.

2.4.4 Experimental procedure

Rather than perform a user study evaluation like in §2.3, with fixed proposition queries and matched output, we instead focus on an iterative approach towards answering the substantive questions of interest:

1. The domain expert supplies a set of proposition queries of interest.
2. The system identifies the top-n semantic matches of each query in the corpus.
3. The domain expert reviews the semantic matches for each query and either (a) accepts the output as sufficiently valid (i.e., suitable for later

[28] https://spacy.io

analysis or aggregation), or (b) revises the query sentence or the number of retrieved sentences (n) if she finds that the query is capturing irrelevant or tangential material.

While this procedure does not offer a quantitative evaluation of the user study as in the previous section, it offers several practical advantages. First, this process helps the domain expert become directly acquainted with the text data, without having to do a formal annotation. Second, it allows the domain expert to iteratively hone what the proposition queries (which act as exemplars for the concepts of interest) should look like in a data-driven manner. Third, it accommodates different views of what it means for a sentence to "express the idea" in a proposition query: a domain expert with a more exacting take on this may restrict the size of C_m or go through further iteration rounds. Finally, it yields a "transcript" of the expert's qualitative decision-making process in finalizing the proposition queries and matched output; given the corpus and series of proposition queries, another researcher can re-traverse the text in the same way.

We note that in this process, the domain expert does not look at aggregated data for a query until the matched output is finalized. This avoids the potential for selecting matched sentences to suit expected proportions or trends, rather than because the sentences are actually capturing the given idea.

Social welfare:

- Welfare builds a healthy America.
- America deserves a better health care policy.
- America faces a health care crisis.
- Government should provide economic benefits to American communities.
- Federal government should provide health care and economic support to help our communities thrive.
- Welfare helps American families thrive.
- Temporary assistance helps needy American families thrive.
- Child welfare programs help needy American families thrive.
- Poor Americans need a domestic hunger safety net.

Foreign aid:

- Disaster relief and lifesaving assistance amidst complex crises.
- United States should provide humanitarian relief and international assistance to global communities in need.
- Foreign aid supports global stability.
- Foreign aid reduces global poverty and supports sustainable development and security.
- Foreign aid promotes global health.
- Foreign aid helps fight HIV/AIDS and malaria abroad.
- Foreign aid empowers women and girls.
- Foreign aid promotes economic prosperity and resilience.
- Foreign aid boosts the economy of developing nations and alleviates poverty.
- Human rights norms are the cornerstone of U.S. foreign policy.
- We stand for human rights and democratic values abroad.

National security:

- Government should end legal loopholes and secure our borders.
- Government should safeguard the American people, our homeland, and our values.
- Terrorists attack the American people, our country, and our way of life.
- Government should identify potential terrorists and prevent attacks.
- Government should enforce immigration laws and secure U.S. borders.
- Government should disrupt cartels, smugglers, nefarious actors, illegal border crossers.
- Government should rebuild our military.

(Table 2.5 continued on next page)

(Table 2.5 continued from previous page)

Sexual and reproductive regulation:
• States should defend traditional marriage and oppose gay marriage.
• Americans overwhelmingly oppose same-sex marriage.
• Marriage is between a man and a woman.
• Gay marriage causes societal collapse.
• I do not support gay marriage.
• Gay marriage erodes/deteriorates traditional marriage and family.
• Traditional marriage and the family are the foundation of American society.
• Abortion kills unborn children.
• Partial birth abortion is murder.
• Partial birth abortion is a violent procedure that is truly traumatic for the mother and her unborn child.
• Partial birth abortion is cruel and inhumane.

Table 2.5: Proposition queries used in policy experiments; this does not include earlier iterations of queries that were discarded.

Proposition queries. Our domain expert focused on the four policy areas described in §2.4.1: social welfare, sexual and reproductive regulation, foreign aid, and increased national security. A fifth topic, constituent services (e.g., Capitol tours, internships), served as a "control group" in which we expect minimal partisan affiliation.

The domain expert drafted 7–10 proposition queries containing expressions of support for each policy area. These initial propositions were drafted based on text that appeared on relevant U.S. government agency websites (e.g., U.S. Department of Agriculture, U.S. Agency for International Development, U.S. Department of Health and Human Services, U.S. Department of Homeland Security) and other governmental and policy-relevant material.

Social welfare: *Federal government should provide health care and economic support to help our communities thrive.*

- This federal assistance would help increase access to quality health care and provide economic development in the community.
- False positive: Reality: While we need to ensure that people who need government assistance receive help, increasing unemployment and health care benefits doesn't help our ailing economy.
- I will continue to help local communities and work at the federal level to improve health care and foster economic development.
- It is important for the economic growth of our state that government and health care organizations in Rhode Island pursue this funding opportunity, and I look forward to supporting their efforts.
- Quality, affordable health care is critical to helping the South Bronx thrive.

Foreign aid: *United States should provide humanitarian relief and international assistance to global communities in need.*

- The international community is providing the people of Kosovo with needed humanitarian support.
- The Armed Forces will continue to execute the mission in support of USAID and the international community in providing humanitarian aid and disaster relief.
- USG RESPONSE EFFORTS It is the obligation of the international community to provide humanitarian assistance wherever it is needed.
- Top Ranked in Efficiency Since 1948, providing humanitarian aid to people in need worldwide.
- Additional humanitarian assistance is required as well.

National security: *Government should identify potential terrorists and prevent attacks.*

- We need the PATRIOT Act to prevent attacks and apprehend terrorists.
- We will do everything possible to deter and prevent terrorist attacks.
- FPS own Policy Handbook identifies patrolling as necessary to prevent and deter crime and terrorist attacks.
- These strikes were intended to prevent and deter additional attacks by a clearly identified terrorist threat.
- TSP reportedly helped to unveil and prevent terrorist attacks.

(Table 2.6 continued on next page)

(Table 2.6 continued from previous page)

Sexual and reproductive regulation: *Gay marriage causes societal collapse.*
• The public purpose of marriage is the reason why society creates laws around marriage.
• False positive: But the collapse of marriage is not inevitable.
• Harvard sociologist Pitirim Sorokin found that throughout history, societal collapse was always brought about following an advent of the deterioration of marriage and family.
• The collapse of marriage, rise of illegitimacy, and absence of fathers are the root cause behind most of the nation's social problems.
• The barriers between marriage and cohabitation collapse.

Table 2.6: Top five matched sentences from the corpus for selected proposition queries (in italics). False positives (matched sentences with opposite policy attitude) are marked accordingly.

We then iteratively honed these propositions to reduce false positive outputs and to accommodate sentiment negations. Table 2.5 contains the final proposition queries for each policy area; Tables A.5–A.8 contain all queries examined during the development process.

Using the procedure described above, we retrieved the top 500 highest-scoring sentences for each proposition query.[29] Table 2.6 shows example output, including instances of false-positive matches. While such matches will inevitably occur, the following analysis was performed with concurrent manual inspection of the matched output sentences to reduce our inclusion of false-positive, negated, or irrelevant sentences.

[29] While 500 sentences per query may seem small in the context of this corpus, we chose this number so that it would be feasible to manually examine output at this stage of our analysis. We plan to increase the output size in future work.

Duplication issues. The IA's web-scraping bots identified and captured changes to webpages only if some part of the document had changed. In theory, this method prevents multiple web captures of content that remain unchanged over time. However, in many cases, the specific matched sentences we retrieved did not change from one web capture to the next, even if other content on that page had been altered. This yields some duplication in our data.

We have not deduplicated the matched sentences for two reasons. First, policy sentences are often duplicated *across* different URLs for what is effectively the same document (e.g., the printable version of a page), which a meaningful deduplication strategy should also take into account. Second, policy sentences may be shared, replicated, or otherwise repeated from one legislator to another; for example, multiple legislators publish the sentence, "I do not support gay marriage." Simple approaches to deduplicating results would remove these common sentences. We plan to further investigate appropriate approaches to deduplication in future work.

2.4.5 Results

We present some basic results using the same semantic measurement idea from the previous case study. These aggregations are based on each proposition's 500 best-scoring matched sentences (see Table 2.6 for examples).[30]

Fig. 2.5 shows the distribution of matched sentence output for each policy area, split by partisan affiliation of the legislator whose site the output

[30]Unless otherwise mentioned, we do not include the negated queries in this analysis.

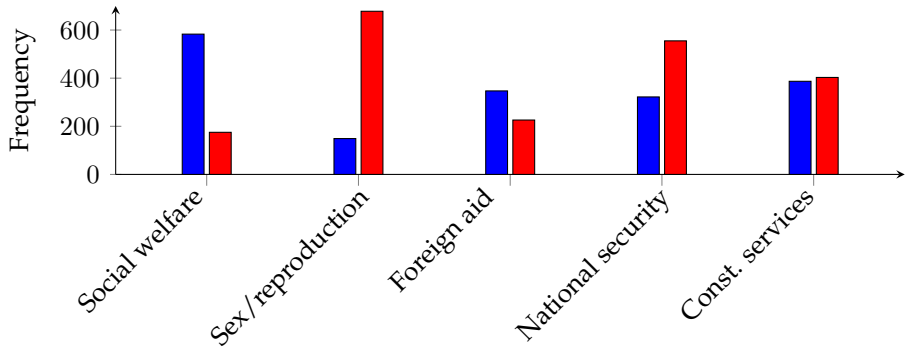


Figure 2.5: Frequency of semantic matches by political affiliation and policy area. Red indicates a Republican legislator or group, and blue a Democratic one. Neutral bodies (such as Congressional committees) are omitted here.

appeared on.[31] In general, we observe expected partisan leanings: Democrats express support for social welfare and foreign aid much more frequently than Republicans, and vice versa for sexual & reproductive regulation and national security. Furthermore, the difference in partisan support for constituent services is minimal.

As a further validity check, we would expect that in heavily polarized policy areas, a proposition capturing the opposing policy attitude would yield a flipped partisan divide. Fig. 2.6 shows an example for same-sex marriage, where we contrast the original proposition, "Americans overwhelmingly oppose same-sex marriage," with its inverse, "Americans support marriage equality." As expected, the former attitude (opposing same-sex marriage) appears more frequently on Republican legislators' websites, and the latter (for marriage equality) almost exclusively on Democratic leg-

[31] The frequency counts do not sum to 1250 (5 proposition queries $\times$ 250 matches/query) because the semantically matched sentences also came from committee (and other neutral) webpages, which we omit in this part of the analysis.

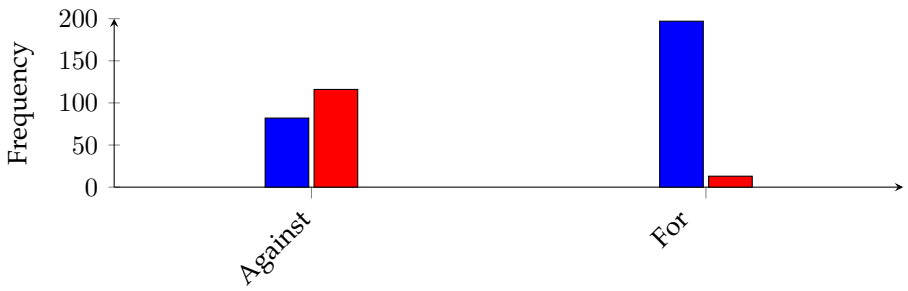


Figure 2.6: Frequency of semantic matches for proposition queries expressing opposite attitudes towards same-sex marriage. *Against*: "Americans overwhelmingly oppose same-sex marriage." *For*: "Americans support marriage equality." Red indicates a Republican legislator and blue a Democratic one.

islators' websites. We note that the smaller difference in expressions of opposition to same-sex marriage is not as surprising as it may seem: based on inspection of the matched sentences and documents they appear in, many of the statements by Democrats are in support of the Defense of Marriage Act in the early 2000s.

This analysis provides a proof of concept for the utility of semantic matching in large text corpora; in this case study, manually identifying and aggregating instances of policy attitudes would be infeasible.

2.5 Discussion

In this section , we discuss some broader findings from our two case studies as well as avenues for potential future work.

Semantic matches. The criteria for what constitutes a semantic match varied between the applications we presented. The disaster recovery study

demanded more specific semantic matches, as the domain expert was interested in relations between specific concepts or entities, and during the user study, participants were less sure about scoring sentences where the idea was only partially expressed. In contrast, when identifying semantic matches for policy attitude propositions, less semantically similar sentences were also permissible if they still supported the broader policy position.

In general, what should be labeled a semantic match is dependent on the practitioner and their application, and as demonstrated in §2.3.6, performance on outside or proxy tasks may not be a wholly reliable indicator; this echoes similar advice by Grimmer and Stewart [2013]. While we use models based on semantic similarity and entailment as a starting point, our setup is general enough that a practitioner could use a matching function that scores based on other semantic relationships. Our setup is also flexible; the practitioner can iteratively refine their queries and output set sizes based on continued exploration of the corpus, as seen in the policy attitudes study. Furthermore, the judgment of whether sentences are good semantic matches for a query is separate from labeling other attributes and performing downstream analysis, allowing for flexibility.

Entities. In the disaster recovery application, corpus-specific entities like government agencies and insurance companies were of interest. However, such entities may lack appropriate distributed representations (sometimes even in the Google News word2vec case) or presence in the embedding training corpus. (A frequent example in our earthquake news corpus is the

Canterbury Earthquake Recovery Authority, often written as "Cera" and conflated with the actor Michael Cera.) A possible solution is to retrain word embeddings on the target corpus if it is sufficiently large, or in the case of large pretrained language models, continue pretraining on the target corpus [Gururangan et al., 2020].

Context and coreference. Currently, we do not take multiple sentences into account at once when determining sentence matches. (The user study in §2.3 provided context in the survey for the users alone.) In some cases, this leads to the system finding a match at the sentence level when it would otherwise be invalid from context; in others, a potential match is spread across a sentence boundary. In the `.gov` corpus case, there are documents which, due to quirks from converting PDFs or HTML to text, do not necessarily have clear sentence boundaries at all. Including larger and smaller passages (not only sentences) may be worthwhile, potentially coupled with more preprocessing (e.g., coreference resolution, entity linking) as well.

Corpus heterogeneity. A possible challenge that we did not fully explore in either case study was matching across a highly heterogeneous text corpus — e.g., documents spanning many styles, like scientific text and news, covering large time periods where word usage may change, or including many languages.[32] Regarding changing word usage, it is possible that large pretrained language models which create contextual word embed-

[32]The corpus in the policy study consisted of many types of documents, but the documents suffered consistently from messy web-to-text translations and in that sense were of a similar style. Likewise, although there was non-English text in the corpus, we focused on matching policy attitudes in English using monolingual embeddings.

dings could cover some of these nuances, as shifts in word usage may be accompanied with shifts in the contexts that those words appear in.[33] However, such ambiguities matched would have to be verified by the practitioner, rather than assumed to be captured automatically.

2.6 Related work

The semantic matching applications presented here are reminiscent of several lines of research in NLP.

Retrieval. As mentioned in §2.1, finding coarse semantic matches of a proposition in a corpus is closely related to past work in information retrieval (IR), particularly sentence retrieval [Balasubramanian et al., 2007]; more recent work also uses sentence embeddings in a dual encoder setup for retrieval [Gillick et al., 2018; Reimers and Gurevych, 2019; Luan et al., 2021]. Other relevant work in IR includes *passage* retrieval, which is a component in many web-scale question answering systems [Tellex et al., 2003]. The main difference is that, here, we seek more than a single answer to a question-query; we seek *all matches* to the query (formed as a proposition). Our approach also resembles work on question answering known as *machine reading* on already-retrieved passages [Chen et al., 2017a], as well as more recent work on scalable open-domain question answering [Seo et al.,

[33] Work on shifts in word usage tends to focus on identification of such words across distinct corpora (e.g., split by time period), rather than matching within the same corpus. For example, Gonen et al. [2020] present a method for identifying word usage changes by computing differences in a word's nearest neighbors; however, this procedure requires computing separate embedding spaces for each corpus.

2018].

Measurement or tracking of ideas. Tracking or measurement of ideas in corpora has often been considered in a more exploratory way, without a user-generated query. Such exploration has long been a motivation for topic models [e.g., Blei and Lafferty, 2006]. For example, Prabhakaran et al. [2016] use topics and their rhetorical roles in scientific journal abstracts to understand when topics are in growth or decline. Other work has allowed user specification of a particular query, though usually as an n-gram [Michel et al., 2011], keywords or topics [Starbird et al., 2016; Tan et al., 2017], or short meme phrases [Leskovec et al., 2009]. We define matches at a more fine-grained proposition level.

The closest work to ours is perhaps by Metzler et al. [2005], which introduces the RECAP tool for tracking information reuse. They use basic comparisons (e.g., word overlap, IBM translation model 1, tf-idf) to determine similarity of sentences to a query. The main difference is one of application, as they are primarily interested in precise factual content, rather than evocation of an idea, and therefore focus on much tighter notions of what constitutes a semantic match.

Other semantic comparisons. There are several relevant semantic comparison tasks which could provide suitable matching functions. For example, natural language inference determines whether a hypothesis is entailed given a premise, and there is a long line of entailment tasks and corpora: among others, the Recognizing Textual Entailment challenges [RTE;

beginning with Dagan et al., 2006]; the Sentences Involving Compositional Knowledge dataset [SICK; Marelli et al., 2014]; the large-scale SNLI and MultiNLI datasets used here [Bowman et al., 2015; Williams et al., 2018]; and the SciTail dataset [Khot et al., 2018]. The RTE-5 through RTE-7 shared tasks, starting with Bentivogli et al. [2009], contain a similar matching task to ours; however, these have a very different end goal (using entailment models to improve text summarization) and much smaller corpora (10 documents).

Other tasks include identifying semantic similarity between two sentences [Cer et al., 2017] and identifying paraphrase pairs [Dolan et al., 2004; Dolan and Brockett, 2005], which is akin to semantic similarity but on a binary scale of is-a-paraphrase or not.

2.7 Conclusion

In this chapter, we introduced a framework based on semantic matching between a proposition query expressing an idea and sentences in a target corpus. We performed two case studies in different domains (disaster recovery and U.S. policy) with different information needs and computational constraints, and demonstrated their potential for hypothesis generation and corpus exploration.

Chapter 3

Nearest Neighbor Overlap

Model selection is an important practical consideration when applying semantic matching models to new corpora or applications, especially in the absence of appropriate training data. As noted in §2.2, there are an abundance of sentence embedders one could choose from, with no clear best option. In the prior case studies, we performed small preliminary rounds of evaluation with our domain experts to compare model outputs, but performing a formal evaluation (like a user study) for every model we wanted to try was clearly infeasible. While there are a multitude of intrinsic [e.g., Conneau et al., 2018] and extrinsic [e.g., GLUE; Wang et al., 2018] evaluations for sentence embedders, none of these necessarily speak to how an embedder will perform on the target domain data, or what cosine similarity captures in terms of the meaning relationship between two sentences.

In this chapter, we introduce **nearest neighbor overlap** (N2O), which

The work in this chapter is based on Lin and Smith [2019]; addition of more recent sentence embedding models (RoBERTa, SBERT, and GPT-2) is new to this thesis.

compares a pair of embedders in a linguistics- and task-agnostic manner using only a large unannotated corpus. The central idea is that two embedders are more similar if, for a fixed query sentence, they tend to find nearest neighbor sets that overlap to a large degree. By drawing a random sample of queries from the corpus itself, we can estimate N2O using realistic data drawn from a domain of interest. N2O enables exploration of nearest neighbor behavior without domain-specific annotation, and therefore can help an end-user compare embedder options not only in the semantic measurement applications above, but also text clustering [Cutting et al., 1992], information retrieval [Salton and Buckley, 1988], and open-domain question answering [Seo et al., 2018], among other tasks.

3.1 N2O procedure

We first motivate and introduce our nearest neighbor overlap (N2O) procedure for comparing embedders (maps from objects to vectors). Although we experiment with sentence embedders here, we note that this comparison procedure can be applied to other types of embedders (e.g., phrase-level or document-level).[1]

Desiderata. We would like to quantify the extent to which sentence embedders vary in their treatment of "similarity." For example, given the sentence *Mary gave the book to John*, embedders based on bag-of-words will treat *John gave the book to Mary* as being maximally similar to the first sentence,

[1]We also note that nearest neighbor search has been frequently used on *word* embeddings (e.g., word analogy tasks).

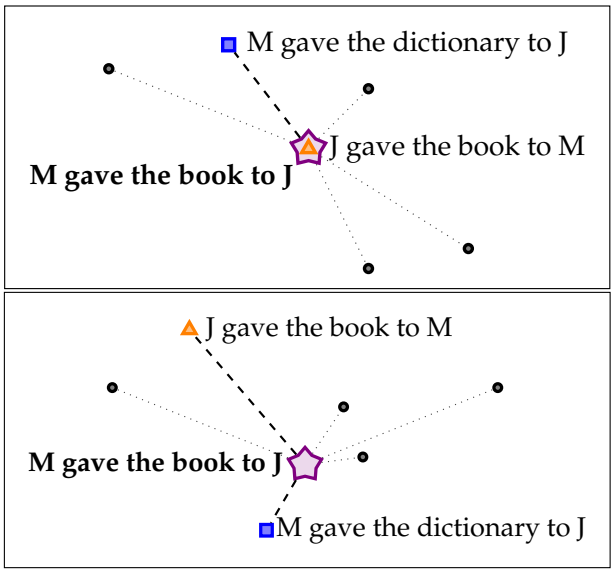


Figure 3.1: A toy example of two sentence embedders and how they might affect nearest neighbor sentences.

whereas different embedders may yield lower similarity for that compared to the sentence *Mary gave the dictionary to John*. We would like our comparison to reflect this intuition.

We would also like to focus on using naturally-occurring text for our comparison. Although there is merit in expert-constructed examples (see linguistic probing tasks referenced in §3.8), we have little understanding of how these models will generalize to text from real documents; many application settings involve computing similarity across texts in a corpus. Finally, we would like our evaluation to be task-agnostic, since we expect embeddings learned from large unannotated corpora in a self-supervised (and task-agnostic) manner to continue to play an important role in NLP.

As a result, we base our comparison on the property of *nearest neighbors*:

function $N2O(\mathbf{e}_A, \mathbf{e}_B, C, k)$
  for each query $\mathbf{q}_j \in \{\mathbf{q}_i\}_{i=1}^{n}$ **do**
    $neighbors_A \leftarrow nearest(\mathbf{e}_A, \mathbf{q}_j, C, k)$
    $neighbors_B \leftarrow nearest(\mathbf{e}_B, \mathbf{q}_j, C, k)$
    $o[j] \leftarrow |neighbors_A \cap neighbors_B|$
  end for
  return $\sum_j o[j]/(k \times n)$
end function

Figure 3.2: Computation of nearest neighbor overlap (N2O) for two embedders, $\mathbf{e}_A$ and $\mathbf{e}_B$, using a corpus C; the number of nearest neighbors is given by k. n is the number of queries ($\mathbf{q}_1 \ldots \mathbf{q}_n$), which are sampled uniformly from the corpus without replacement. The output is in [0, 1], where 0 indicates no overlap between nearest neighbors for all queries, and 1 indicates perfect overlap.

first, because similarity is often assumed as corresponding to nearness in embedding space (Fig. 3.1), which may not be true in practice; second, because nearest neighbor methods are used directly for clustering, retrieval, and other applications; and finally, because the nearest neighbors of a sentence can be computed for any embedder on any corpus without additional annotation.

Algorithm. Suppose we want to compare two sentence embedders, $\mathbf{e}_A(\cdot)$ and $\mathbf{e}_B(\cdot)$, where each embedding method takes as input a natural language sentence $\mathbf{s}$ and outputs a d-dimensional vector. For our purposes, we consider variants trained on different data or using different hyperparameters, even with the same parameter estimation *procedure*, to be different sentence embedders.

Take a corpus C, which is likely to have some semantic overlap in its sentences, and segment it into sentences $\mathbf{s}_1, \ldots, \mathbf{s}_{|C|}$. Randomly select a

small subset of the sentences in C as "queries" ($\mathbf{q}_1, \ldots, \mathbf{q}_n$). To see how similar $\mathbf{e}_A$ and $\mathbf{e}_B$ are, we compute the overlap in nearest neighbor sentences, averaged across multiple queries; the algorithm is in Figure 3.2. $nearest(\mathbf{e}_i, \mathbf{q}_j, C, k)$ returns the k nearest neighbor sentences in corpus C to the query sentence $\mathbf{q}_j$, where all sentences are embedded with $\mathbf{e}_i$.[2] There are different ways to define nearness and distance in embedding spaces (e.g., using cosine similarity or Euclidean distance); here we use cosine similarity.

We can think about this procedure as randomly probing the sentence vector space (through the n query sentences) from the larger space of the embedded corpus, under a sentence embedder $\mathbf{e}_i$; in some sense, k controls the depth of the probe.[3] The $N2O$ procedure then compares the sets of sentences recovered by the probes.

3.2 Sentence embedding methods

In the previous section, we noted that we consider a "sentence embedder" to encompass how it was trained, which data it was trained on, and any other hyperparameters involved in its creation. In this section, we first review the broader methods behind these embedders, turning to implementation decisions in §3.3.

[2] One of these will be the query sentence itself, since we sampled it from the corpus; we assume $nearest$ ignores it when computing the k-nearest neighbor lists.

[3] An alternate view is to consider each sentence as a node in a graph, where edges are drawn between nodes if they are within k-nearest neighbors of each other according to $\mathbf{e}$; N2O offers an estimate of the similarity between k-nearest neighbor graphs.

3.2.1 tf-idf

We consider **tf-idf**, which has been classically used in information retrieval settings. The tf-idf of a word token is based off two statistics: term frequency (how often a term appears in a document) and inverse document frequency (how rare the term is across all documents). The vector representation of the document is the idf-scaled term frequencies of its words; in this work we treat each sentence as a "document" and the vocabulary-length tf-idf vector as its embedding.

3.2.2 Word embeddings

Because sentence embeddings are often built from word embeddings (via initialization when training or other composition functions), we briefly review notable word embedding methods.

Static embeddings. We define "static embeddings" to be fixed representations of every word type in the vocabulary, regardless of its context. We consider three popular methods: **word2vec** [Mikolov et al., 2013] embeddings optimized to be predictive of a word given its context (continuous bag of words) or vice versa (skipgram); **GloVe** [Pennington et al., 2014] embeddings learned based on global cooccurrence counts; and **FastText** [Conneau et al., 2017], an extension of word2vec which includes character n-grams (for computing representations of out-of-vocabulary words).

Contextual embeddings. Contextual word embeddings, where a word token's representation is dependent on its context, have become popular

due to improvements over state-of-the-art on a wide variety of tasks. We consider:

- **ELMo** [Peters et al., 2018] embeddings are generated from a multi-layer, bidirectional recurrent language model that incorporates character-level information.
- **GPT** [Radford et al., 2018] and **GPT-2** [Radford et al., 2019] embeddings are generated from a unidirectional language model with multi-layer transformer decoder; subword information is included via byte-pair encoding [BPE; Sennrich et al., 2016].
- **BERT** [Devlin et al., 2019] embeddings are generated from a transformer model trained to predict (a) a word given both left and right context, and (b) whether a sentence is the "next sentence" given a previous sentence. Subword information is incorporated using the WordPiece model [Schuster and Nakajima, 2012]. Related models include **RoBERTa** [Liu et al., 2019], which removes the next sentence prediction objective and adjusts training decisions for better downstream performance; and Sentence-BERT [**SBERT**; Reimers and Gurevych, 2019], which is BERT finetuned with an objective function that maximizes cosine similarity between similar sentences.

Composition of word embeddings. The simplest way to obtain a sentence's embedding from its sequence of words is to average the word embeddings.[4] Despite the fact that averaging discards word order, it performs

[4] In the case of GPT- and BERT-based models, which yield subword embeddings, we treat those as we would standard word embeddings.

surprisingly well on sentence similarity, NLI, and other downstream tasks [Wieting et al., 2016; Arora et al., 2017].[5]

In the case of contextual embeddings, there may be other conventions for obtaining the sentence embedding, such as using the embedding for a special token or position in the sequence. With BERT-based models, the `[CLS]` token representation (normally used as input for classification) is also sometimes used as a sentence representation; similarly, the last token's representation may be used for GPT-based models.

3.2.3 Encoders

A more direct way to obtain sentence embeddings is to learn an encoding function that takes in a sequence of tokens and outputs a single embedding; often this is trained using a relevant supervised task. We consider two encoder-based methods:

- **InferSent** [Conneau et al., 2017]: supervised training on the Stanford Natural Language Inference [SNLI; Bowman et al., 2015] dataset; the sentence encoder provides representations for the premise and hypothesis sentences, which are then fed into a classifier.

- Universal Sentence Encoder [**USE**; Cer et al., 2018]: supervised, multi-task training on several semantic tasks (including semantic textual similarity); sentences are encoded either with a deep averaging network or a transformer.

[5] Arora et al. [2017] also suggest including a PCA-based projection with word embedding averaging to further improve downstream performance. However, because our focus is on behavior of the embeddings themselves, we do not apply this projection here.

3.3 Experimental details

Our main experiment is a broad comparison, using N2O, of the embedders discussed above and listed in Table 3.1. Despite the vast differences in methods, N2O allows us to situate each in terms of its functional similarity to the others.

N2O computation. We describe a N2O sample as, for a given random sample of n queries, the computation of $N2O(\mathbf{e}_A, \mathbf{e}_B, C, k)$ for every pair of sentence embedders through the procedure described in §3.1, using cosine similarity to determine nearest neighbors. The results in §3.4 are with k (the number of sentences retrieved) set to 50, averaged across five samples of $n = 100$ queries. We illustrate the effects of different k and N2O samples in §3.5.

Corpus. For our corpus, we draw from the English Gigaword [Parker et al., 2011], which contains newswire text from seven news sources. For computational feasibility, we use the articles from 2010, for a total of approximately 8 million unique sentences.[6] We note preprocessing details (segmentation, tokenization) in Appendix B.1.

Queries. For each N2O sample, we randomly select 100 ledes (opening sentences) from the news articles of our corpus, and use the same ones across all embedders. Because the Gigaword corpus contains text from multiple news sources covering events over the same time period, it is likely

[6] Because many news articles show up multiple times in the corpus, 23% of sentences in the English Gigaword are exact duplicates of one another; we remove these duplicates.

that the corpus will contain semantically similar sentences for a given lede. The average query length is 30.7 tokens (s.d. 10.2); an example query is: "Sandra Kiriasis and brakewoman Stephanie Schneider of Germany have won the World Cup bobsled race at Lake Placid."

Sentence embedders. Table 3.1 lists the sentence embedders we use in our experiments, their dimensions, and the manner in which their word embeddings were composed (if applicable). In general, we use popular pretrained versions of the methods described in §3.2. We also select pretrained variations of the same method (e.g., FastText embeddings trained from different corpora; pretrained ELMo models with different capacity) to permit more controlled comparisons.

In a couple of cases, we train/finetune models of our own. For tf-idf, we compute frequency statistics using our corpus, with each sentence as its own "document." For BERT, we use the Hugging Face implementation with default hyperparameter settings,[7] and finetune using the matched subset of the MultiNLI dataset [Williams et al., 2018] for three epochs (dev. accuracy 84.1%).

We note that additional embedders are easily situated among the ones tested in this chapter by first computing nearest neighbors of the same query sentences, and then computing overlap with the nearest neighbors obtained here. To enable this, the code, query sentences, and nearest neighbors are available at https://lucylin.github.io/projects/n2o.

[7] https://huggingface.co

Embed. method	Composition	Dim.	Model/data description
tf-idf	n/a	$\|V\|$	tf-idf statistics obtained on Gigaword corpus (2010 slice)
word2vec	`average`	300	Google News (3B tokens)
GloVe	`average`	100	Wikipedia 2014 + Gigaword 5 (6B tokens, uncased)
		300	Wikipedia 2014 + Gigaword 5 (6B tokens, uncased)
		300	Common Crawl (840B tokens, cased)
FastText	`average`	300	Wikipedia + UMBC + statmt.org (16B tokens)
		300	" + subword information
		300	Common Crawl (600B tokens)
		300	" + subword information
ELMo	`average`	256	pretrained small model (1 Billion Word Benchmark)
		1024	pretrained original model (1 Billion Word Benchmark)
		1024	pretrained original/5.5B model (Wikipedia/news)
BERT	`[CLS]`	768	pretrained cased/base model
	`average`	768	on Wikipedia + BooksCorpus
	`[CLS]`	768	" + finetuning on MultiNLI
	`average`	768	(matched subset)
RoBERTa	`[CLS]`	768	pretrained base model
	`average`	768	on Wikipedia, BooksCorpus, etc.
SBERT	`average`	768	BERT finetuned on MultiNLI with distance-based objective
GPT	`last`	512	pretrained model (110M params)
	`average`	512	trained on BooksCorpus
GPT-2	`last`	768	pretrained model (117M params)
	`average`	768	trained on WebText
InferSent	n/a	4096	V1 (GloVe-based) model, trained on SNLI
USE	n/a	512	deep averaging network (DAN) encoder; multitask training
		512	transformer encoder; multitask training

Table 3.1: Pretrained sentence embedder details. For methods which produce word embeddings, rather than a single sentence embedding, "composition" denotes how a single embedding was obtained from the sentence's word embeddings. ELMo embeddings are averaged across the three bi-LSTM layers; BERT* and GPT* embeddings come from the final hidden layer. All models besides tf-idf and the fine-tuned version of BERT are common pretrained versions.

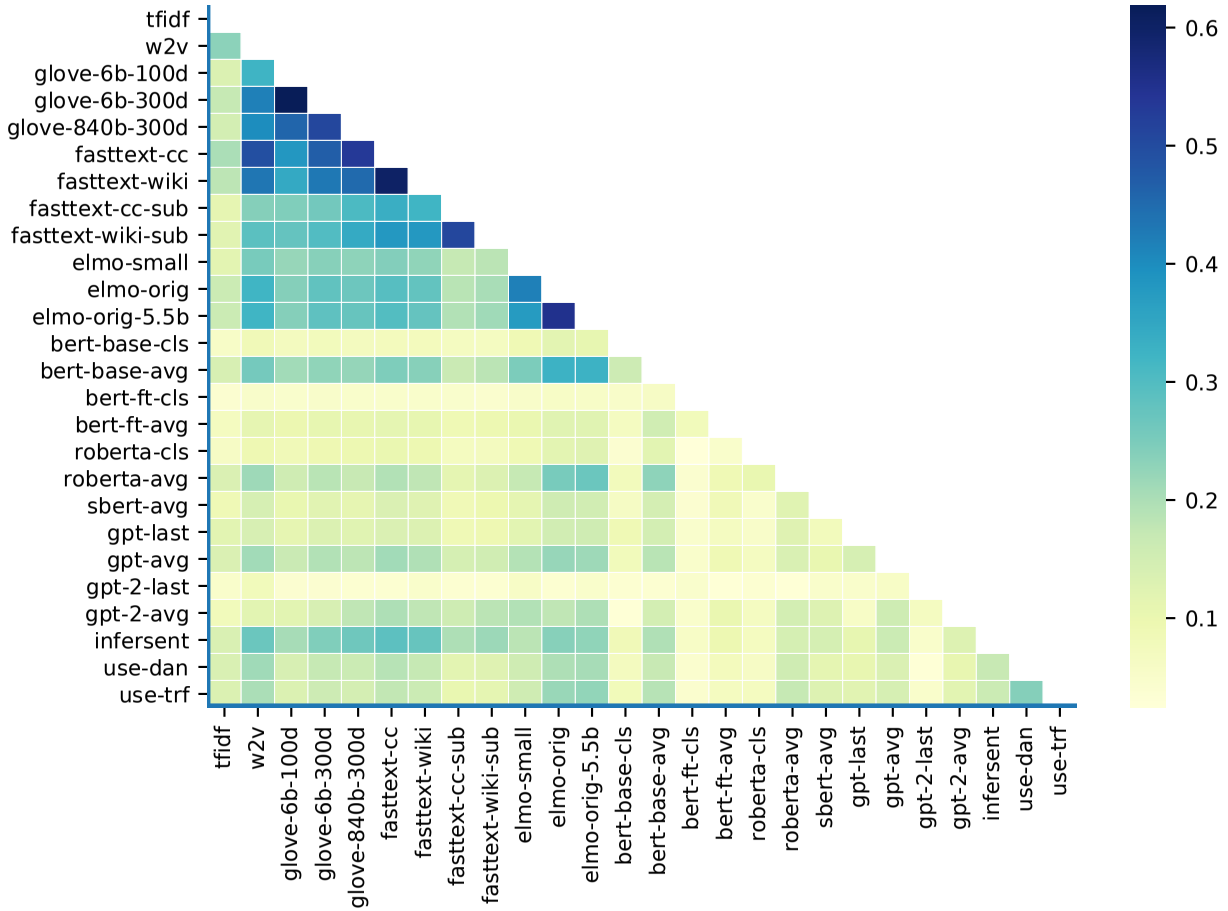


Figure 3.3: N2O values between all pairs of sentence embedders in Table 3.1. A version with all values annotated is in the appendix (Fig. B.1).

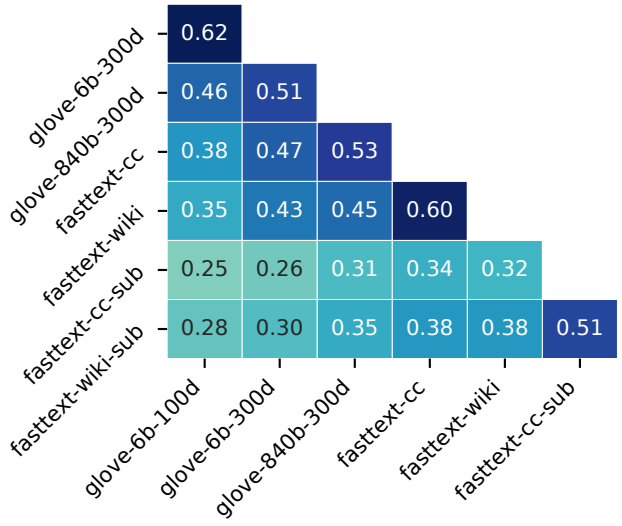


Figure 3.4: N2O values for a subset of embedders based on static word embeddings ($k = 50$).

3.4 Results

In this section, we present the results from the experiment described in §3.3. Fig. 3.3 shows N2O between each pair of sentence embedders listed in Table 3.1 over the 100 queries; the values range from 0.04 to 0.62. While even the maximum observed value may not seem large, we reiterate that overlap is computed over two draws of $k = 50$ sentences (nearest neighbors) from approximately 8 million sentences, and even an N2O of 0.04 is unlikely from random chance alone.

Averages of static word embeddings. We first observe that there is generally high N2O among this set of embedders in comparison to other categories (Fig. 3.4). Some cases where N2O is high for variations of the same

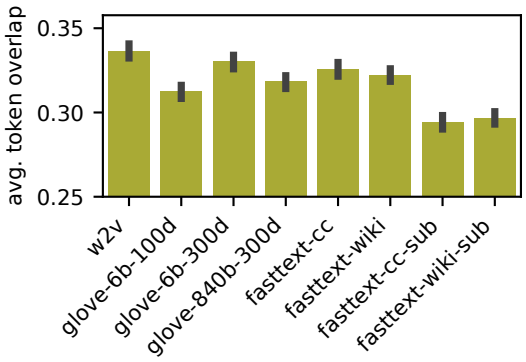


Figure 3.5: Average token overlap between a query and its nearest neighbors for tf-idf and static word embedding models ($k = 50$), averaged over all queries. Error bars represent 95% confidence intervals.

embedder: `glove-6b-100d` and `glove-6b-300d`, which have different dimensionality but are otherwise trained with the same method and corpus (and to a lesser extent `glove-840b-300d`, which retains casing and is trained on a different corpus); `fasttext-cc` and `fasttext-wiki`, which again are trained with the same method, but different corpora.

The use of subword information, unique to `fasttext-cc-sub` and `fasttext-wiki-sub`, has a large effect on N2O; there is a high (0.52) N2O value between these two and much lower N2O with other embedders, including their analogues without subword information. This effect is also illustrated by measuring, for a given embedder, the average token overlap between the query and its neighbors (Fig. 3.5). As we would expect, subword methods find near neighbors with lower token overlap, because they embed surface-similar strings near to each other.

Figure 3.6: N2O values for a subset of embedders based on contextual word embeddings ($k = 50$).

tf-idf. Unsurprisingly, tf-idf has low N2O with other embedders (even those based on static word embeddings). Like the subword case, we can also use token overlap to understand why this is the case: its nearest neighbors have by far the largest token overlap with the query (0.43).

Averages of ELMo embeddings. We test three ELMo pretrained models across different capacities (`elmo-small`, `elmo-orig`) but the same training data, and across different training data but the same model capacity (`elmo-orig`, `elmo-orig-5.5b`). These two embedder pairs have high N2O (0.42 and 0.55 respectively); the mismatched pair, with both different training data and capacities, has slightly lower N2O (0.38).

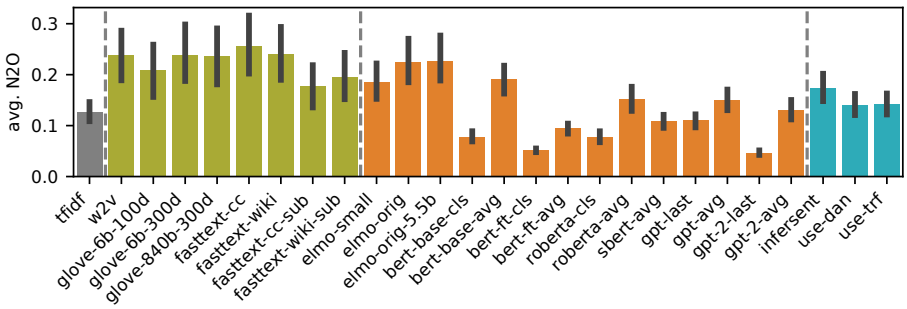


Figure 3.7: Comparison of N2O distribution between each embedder and all others.

Transformer-based models. We first find that specific-token representations for BERT, RoBERTa, GPT, and GPT-2 (`bert-base-cls`, `gpt-last`, etc.) are outliers compared to other embedders (i.e., low N2O). This itself is not unexpected, as the training objectives for both of the pretrained models (without finetuning) are not geared towards semantic similarity the way other embedders may be. This effect seems to hold even for the standard MultiNLI-finetuned version of BERT (`bert-ft-cls`).[8] To further confirm these findings, we plot the N2O values for each embedder (compared to all others) in Fig. 3.7.

We also find that taking *averaged* BERT* and GPT* embeddings yields higher N2O with other embedders, especially ELMo-based ones; we see this effect most strongly when comparing `bert-base-cls` to `bert-base-avg`, say. SBERT, which is finetuned on the same data and base model as `bert-ft-*` but with a semantic similarity objective, behaves more like the

[8] In preliminary experiments, we also saw similar results with BERT finetuned on the Microsoft Research Paraphrase Corpus [Dolan et al., 2004]; that is, the effect does not seem specific to MultiNLI.

averaged embeddings than its token-specific version as well.

Encoder-based embedders. We find that InferSent has highest N2O (∼0.2 –0.3) with the averaged word embeddings, despite InferSent being trained on a NLI task; that said, this is not wholly surprising as the model was initialized using GloVe vectors (`glove-840b-300d`) during training. The USE variants (DAN and Transformer) have fairly distinct nearest neighbors compared to other methods, with highest N2O between each other (0.24).

3.5 Robustness and runtime considerations

Varying k. One possible concern is how sensitive our procedure is to k (the number of nearest neighbors from which overlap is computed): we would not want conflicting judgments of how similar two sentence embedders are due to different k. To confirm that changing k does not significantly affect these judgments, we first compute the ranked lists of N2O output for each $k \in \{5, 10, \ldots, 45, 50\}$, where each list consists of all embedder pairs ordered by N2O for that k. We then compute Spearman's rank correlation coefficient (ρ) between each pair of ranked lists, where 1 indicates perfect positive correlation. We find that the average Spearman's ρ is very high (0.996; min. 0.986) — i.e., the rankings of embedder similarity by N2O are reasonably stable across different values of k, even as far as $k = 5$ and $k = 50$.

Query sampling. We also examine how the results may vary across different query samples; as noted previously, the presented results are averaged

across five samples of $n = 100$ queries each. Standard deviations for N2O values across the five samples range from 0.005 to 0.019 (avg. 0.011). That is, given the range of N2O values being compared, the differences due to different query samples is small. We compute Spearman's ρ across different N2O samples in the same manner as above ($k = 50$) and find an average ρ of 0.994 (min. 0.991).

Runtime. A theoretical concern with N2O is that, naively, its computation is linear in the size of the corpus, and to have reasonable semantic overlap within a diverse set of sentences, the corpus should be large. While our implementation of exact nearest neighbor search is sufficiently fast in practice,[9] we provide comments on use of approximate nearest neighbor methods in Appendix B.3.

3.6 Popularity of neighbors

Previously, we performed a basic comparison between sentence embedders using N2O. Here, we show one kind of analysis enabled by N2O: given a query, which sentences from the corpus C are consistently its neighbors across different embedders? We might expect, for example, that a nearly identical paraphrase of the query will be a "popular" neighbor chosen by most embedders. Table 3.2 shows an example query with a sentence that is in the 5-nearest neighborhood for all sentence embedders. We also show

[9] Given precomputed sentence embeddings, exact nearest neighbor search across the corpus takes 30 s.–1 min. (depending on dimensionality) for a batch of $n = 100$ queries and $k = 50$, across two 12-core Intel Xeon CPUs (E5-2960/2.60GHz).

sentences that are highly ranked for *some* embedder but not in the nearest neighbor sets for *any other* embedder (for larger $k = 50$).

Qualitatively, what we find with this example's outlier sentences is that they are often thematically similar in some way (such as fiscal matters in Table 3.2), but with different participants. We also observe that extremely "popular" neighbors tend to have high lexical overlap with the query.

Usage scenario. This kind of quantitative-qualitative analysis provides one means for exploring embedder behavior: suppose we are considering using a BERT-based embedder for a semantic matching application like the prior case studies, but on this Gigaword corpus, and are not sure if further analysis of both `bert-base-avg` and `bert-base-cls` is worthwhile. In our experiments on Gigaword, we find that the N2O between the `[CLS]` and averaged versions is 0.17, which is quite low relative to N2O computed between other embedders, especially given that the same base model is generating the sentence embeddings. (That is, the only difference is the composition of the output sentence embedding.) So the difference in matched sentence output between `bert-base-cls` and `bert-base-avg` is likely to be large across multiple queries, and further practitioner-guided comparison between the two is likely to have impact.

Query: Britain's biggest mortgage lender says that average house prices fell 3.6 percent in September, but analysts believe the market isn't that weak.

Embedder	Rank	Sentence
all embedders	≤ 5	Average house prices in Britain fell 3.6 percent in September from a month earlier, the country's biggest mortgage lender said Thursday, although analysts believe the market isn't that weak.
`bert-base-cls`	6	Some analysts say that the December data indicate that consumer spending remains weak, making it harder for the economy to keep a sustained rebound.
`bert-ft-cls`	2	Japanese consumer prices fell for 13th straight month in March, though the GDP data suggests that deflationary pressures are starting to ease.
`bert-ft-avg`	5	An industry group says German machinery orders were down 3 percent on the year in January but foreign demand is improving.
`fasttext-cc-sub`	6	It cautioned however that the economic situation abroad could still slow Sweden's recovery, and said the country's gross domestic product (GDP) would grow just 3.6 percent in 2011, down from its May estimate of 3.7 percent growth.
`glove-840b-300d`	12	Meanwhile, Australia's central bank left its key interest rate unchanged at 3.75 percent on Tuesday, surprising investors and analysts who had predicted the bank would continue raising the rate as the nation's economy rebounds.
`gpt-last`	8	The economy has since rebound and grew 8.9 percent year-on-year in the second quarter, the central bank said last month, with growth expected to exceed six percent in the full year.

Table 3.2: Popular and outlier near neighbors for the given query (top). The first sentence is in the 5-nearest neighborhood for all embedders; the remaining sentences are highly-ranked by the given embedder and outside the 50-nearest neighborhood for all other embedders.

3.7 Query paraphrasing

Attempts to derive sentence embeddings that capture semantic similarity are inspired by the phenomenon of paraphrase; in this section, we use nearest neighbors to probe how sentence embedders capture paraphrase. More specifically, we carry out a "needle-in-a-haystack" experiment using the Semantic Textual Similarity Benchmark [STS; Cer et al., 2017]. STS contains sentence pairs with human judgments of semantic similarity on a 1–5 continuous scale (least to most similar).

We take 75 sentence pairs in the 4–5 range from the STS development and test sets where the sentence pair has word-level overlap ratio < 0.6 — i.e., near paraphrases with moderately different surface semantics. We also constrain the sentence pairs to come from the newstext-based parts of the dataset. The first sentence in each sentence pair is the "query," and the second sentence is (temporarily) added to our Gigaword corpus. An example sentence pair, scored as 4.6, is: (A) *Arkansas Supreme Court strikes down execution law* and (B) *Arkansas justices strike down death penalty*. We then compute the *rank* of the sentence added to the corpus (i.e., the value of k such that the added sentence is part of the query's nearest neighbors). An embedder that "perfectly" correlates semantic similarity and distance should yield a rank of 1 for the sentence added to the corpus, since that sentence would be nearest to the query.

Results. Table 3.3 shows the performance of the sentence embedders; we compute mean reciprocal rank (MRR), the number of queries for which its

Embedder	MRR	# top	# top-5
elmo-orig-5.5b	0.910	67	70
elmo-orig	0.829	60	65
sbert	0.802	59	63
infersent-v1	0.799	55	64
roberta-avg	0.788	55	62
w2v	0.760	52	64
use-trf	0.759	54	60
fasttext-cc	0.756	52	62
use-dan	0.718	51	55
bert-base-avg	0.674	47	55
glove-6b-300d	0.673	48	52
tfidf	0.672	45	55
fasttext-wiki	0.662	45	54
elmo-small	0.638	44	51
gpt2-avg	0.627	42	50
glove-840b-300d	0.601	42	49
gpt-avg	0.600	41	50
fasttext-wiki-sub	0.552	37	47
glove-6b-100d	0.529	37	43
fasttext-cc-sub	0.515	35	41
roberta-cls	0.511	34	42
bert-ft-avg	0.493	31	44
bert-base-cls	0.450	27	42
gpt2-last	0.383	26	39
gpt-last	0.365	24	30
bert-ft-cls	0.302	19	27

Table 3.3: Results for the query-paraphrase experiment (§3.7), sorted by decreasing MRR. # *top* and # *top-5* are the number of queries for which the paraphrase was the nearest neighbor and in the 5-nearest neighborhood (max. 75), respectively.

paraphrase was its nearest neighbor, and the number of queries for which the paraphrase was in its 5-nearest neighborhood. We find that the larger ELMo models and SBERT do particularly well at placing paraphrase pairs near each other. We also can see that averaged BERT and GPT embeddings perform better than the `[CLS]`/final token ones[10]; this is consistent with our earlier observation (§3.4) that their training objectives may not yield specific-token embeddings that directly encode semantic similarity, hence why they are outliers by N2O.

3.8 Related work

Sentence embedder comparisons. Comparisons of sentence embedders have been primarily either (1) linguistic probing tasks or (2) downstream evaluations. Linguistic probing tasks test whether embeddings can distinguish surface level properties, like sentence length; syntactic properties, like tree depth; and semantic properties, like coordination inversion. See Ettinger et al. [2016], Adi et al. [2017], Conneau et al. [2018], and Zhu et al. [2018], among others. Downstream evaluations are often classification tasks for which good sentence representations are helpful (e.g., NLI). Evaluations like the RepEval 2017 shared task [Nangia et al., 2017], SentEval toolkit [Conneau and Kiela, 2018], and GLUE benchmark [Wang et al., 2018] seek to standardize comparisons across sentence embedding methods. N2O is complementary to the above, providing a task-agnostic way to compare embedders' functionality.

[10]The BERT results with STS are consistent with concurrent work by Reimers and Gurevych [2019].

Analysis of word embedding space/geometry. Analysis of nearest neighbors has been more often used for *word* vectors; e.g., Gonen et al. [2020] compute overlap between a word's nearest neighbors across two corpora to determine if a word's usage has changed, and Antoniak and Mimno [2018] examine the impact of minor changes in word embedding training corpora on nearest neighbor stability. Recent work on large language models has also examined the distribution of words in vector space; e.g., Ethayarajh [2019] find that ELMo, BERT, and GPT-2 word embeddings occupy a narrow cone (rather than uniformly distributed), meaning that even unrelated words may still have high cosine similarity.[11]

3.9 Conclusion

In this chapter, we introduced *nearest neighbor overlap* (N2O), a comparative approach to quantifying similarity between sentence embedders. Using N2O, we drew comparisons across a large number of commonly-used sentence embedders. We also provide additional analyses made possible with N2O, from which we found high variation in embedders' treatment of semantic similarity.

[11] This is consistent with our finding that specific token embeddings from BERT and GPT-2 tend to have low N2O with all other embedders. If an embedder tends to place all tokens near each other, then slight differences in token placement may result in substantially different nearest neighbors.

Chapter 4

Sensationalism in Medical News

Sometimes news articles covering medical advances misrepresent or sensationalize the studies they cite. For example, (a) is the lede of a news article linking a class of antidepressant to violent crime, and (b) is a discussion point from the study it cites:

1. "Use of selective serotonin reuptake inhibitors (SSRIs) increases the rate of violent crime among young adults..." [Torjesen, 2015]

2. "The reported association between SSRIs and violent crime in young people cannot be interpreted causally because of confounding by indication." [Molero et al., 2015]

In this case, the misrepresentation is clear: the news article presents the link as causation and not just correlation.

The work in this chapter is unpublished at the time of writing.

Misleading articles can undermine public trust in scientific findings and cause undue panic (e.g., vaccines purportedly causing autism).[1] To an extent, the potential for sensationalism in science and medical news has reached public awareness; for instance, the website "Kill or cure?"[2] lampoons the Daily Mail's tendency to assert that something causes or prevents cancer (or both, in the case of wine). Other resources, such as HealthNewsReview,[3] have sought to be educational by having experts manually review news articles covering medical advances. However, such manual efforts require expert annotation and can become cost-prohibitive. (HealthNewsReview ran out of funding in 2018.)

Given the cost of manual analysis, applying semantic comparison methods to identify such occurrences of sensationalized text (e.g., to alert readers that a text is misleading or journalists to moderate a claim) seems to be a natural fit. In this chapter, we survey past studies across communications, medicine, and psychology to illustrate properties of sensationalism, its occurrence in the health communications landscape, and why it might surface at different steps in the pipeline. We discuss possible user-facing roles that NLP systems could have in this ecosystem; in doing so, we critique the common NLP setup of attempting to label social phenomena in text with high accuracy. Finally, we provide suggestions for developing NLP systems that seek to identify or reduce the occurrence of sensationalism in medical journalism.

[1] While the original Wakefield study on vaccines/autism was itself found to be falsified, a key factor in the study's impact was the media overstating the risk involved [Jackson, 2003].

[2] http://kill-or-cure.herokuapp.com

[3] http://healthnewsreview.org

4.1 What is sensationalism?

Sensationalism, broadly speaking, is an editorial tactic meant to engage readers by providing information that is exciting or shocking. Sometimes this is due to a choice of topic or frame (e.g., violence, sex, scandal), and others due to choices in rhetoric or writing style. In this chapter, we primarily focus on the latter form of sensationalism — as a rhetorical strategy for presenting content, rather than selection of content that is meant to be shocking [Tannenbaum and Lynch, 1960; Molek-Kozakowska, 2013].[4]

Sensationalism in medical journalism. Within the context of medical journalism, sensationalism is more often associated with exaggeration of the discovery or treatment being presented. Although the canonical example of this is inflating correlation to causation (as seen in the example in the introduction), exaggeration of a claim can happen in more subtle ways, such as:

- Removal of uncertainty (e.g., omitting hedging words, like "can" or "might").
- Use of vivid metaphors and superlatives to gain reader attention; e.g., Doherty [2020] notes the overuse of zombie and other science fiction imagery when describing parasite-host interactions in both news and scientific literature, and Ottwell et al. [2021] find frequent use of exag-

[4]Note: the surveyed work in this chapter is based on English language news from the United States, United Kingdom, Australia, and Canada. Media norms, processes, and freedom can differ, and the conclusions drawn here may not apply to other countries.

gerated language like "game-changer" or "miracle" in news reporting on the COVID-19 pandemic.

- Inaccurate or non-quantitative reporting of benefits, which can mislead readers into assuming treatment effectiveness [Moynihan et al., 2000; Hicks-Courant et al., 2021].
- Overly-strong advice to readers (e.g., "don't drink wine") not supported by the evidence [Sumner et al., 2014].
- Omission of caveats and other study limitations, such as small sample size, non-human subjects, and the study being observational rather than having controlled or blinded experiments [Bubela and Caulfield, 2004; Woloshin et al., 2009; Sumner et al., 2014, 2016].
- Unreported conflicts of interest (e.g., funding from a drug manufacturer) by researchers and other interviewed experts [Moynihan et al., 2000; Wang et al., 2017].

Relationship to misinformation, disinformation, and other forms of information disorder. First Draft, a non-profit which works to reduce societal impacts from misinformation, defines *dis*information as "content that is intentionally false and designed to cause harm," *mis*information as false content without the intent to deceive (e.g., reshared content on social media), and *mal*information as content based on truth that is framed to cause harm (e.g., with omitted details) [Wardle, 2019].

Among these forms of content, sensationalism can be characterized as an editorial *strategy* for increasing the proliferation of such information, re-

gardless of truth or authorial intent. In the rest of this chapter, we generally focus on the setting where presented medical findings are based on a source of some sort (e.g., research article), rather than active disinformation campaigns, though we note connections to research and interventions for misinformation and disinformation where relevant.

4.2 Sensationalism in the publishing pipeline

Although journalists get most of the critique for sensationalizing medical claims, there are a number of places in the research finding → news pipeline where exaggeration can occur.

Scientific writing. Research writing is not immune to overclaiming or other spin; for example, Jellison et al. [2019] identified use of writing strategies that "distracted the reader from statistically non-significant results" in the abstracts of >50% of published psychology/psychiatry trials (2012–2017).

The rise of unreviewed papers uploaded to preprint services (e.g., bioRxiv and medRxiv) has led to concerns about rushed presentation of research findings without adequate vetting [Kaiser, 2017] and news media and the general public not distinguishing between preprints and published articles [Hoy, 2020; Kharasch et al., 2021]. Misuse of preprints toward legitimizing misinformation has also become a concern during the COVID-19 pandemic, though it is an open debate whether those negative impacts outweigh the benefits of faster dissemination of findings during such a time-

sensitive event [Vlasschaert et al., 2020], and issues with fast release of results is not limited to preprints (e.g., high-profile retraction of a COVID-19 study in the *Lancet*; see Lipworth et al., 2020).

Outside of individual papers, systemic factors in scientific publishing can also contribute to misconceptions of the scientific process, such as a strong bias towards presenting only positive results, the role of funding sources in driving research, and mismatches between how scientists and journalists view the (un)certainty of research outcomes [Dumas-Mallet and Gonon, 2020].

Press release. A number of studies have found that exaggerations and omission of caveats arise in press releases provided by a university or medical journal [Woloshin and Schwartz, 2002; Brechman et al., 2009]. Furthermore, press releases often act as an intermediary between research and news, and such inaccuracies are associated with increased exaggeration and missing caveats in downstream news articles [Woloshin et al., 2009; Brechman et al., 2009; Sumner et al., 2014; Bratton et al., 2019].

The primary incentive for producing a press release is to get media coverage, although the extent to which stronger claims or omitting caveats has an impact in practice is inconclusive. That said, it is possible that moderating claims made in press releases could be a positive intervention without sacrificing media coverage; Bott et al. [2019] found that inclusion of caveats did not have an effect on journalism students' judgment of newsworthiness, and Adams et al. [2019] found that adjusting causal statements to an appropriate strength was correlated with more accurate news and no cost

to news uptake.

News articles. While guides from professional organizations like the Association of Health Care Journalists[5] exist, their recommendations can be challenging to implement in practice due to constraints such as: balancing newsworthiness or immediacy over reporting depth, gaining access to relevant experts, attempting to gain readers through interesting headlines, insufficient domain-relevant training, and tight deadlines [Leask et al., 2010; O'Keeffe et al., 2021]. From an accountability standpoint, these constraints can make it challenging to self-monitor one's reporting; third-party groups that critiqued health news, like Health News Review [Walsh-Childers et al., 2018] and Media Doctor AU [Wilson et al., 2009], have since run out of funding.

Separate from accurate reporting, the ways in which lay readers interpret scientific or medical claims are still being studied; Adams et al. [2017] found that study participants could distinguish between direct cause (e.g., "makes"), "can" cause, and moderate cause ("might cause," "associated with"), but could not necessarily distinguish between cause and association in the last category.

Social media. Social media platforms (e.g., Twitter, Facebook) often act as intermediaries between readers and news reporting, with the added barriers of information overload and competition with other misinformation (which need not come from reputable sources); Sharma et al. [2016] found

[5] https://healthjournalism.org

that, on Facebook, misleading and false posts about the Zika virus were significantly more popular than accurate ones from trustworthy sources (CDC, newsrooms).

Interventions on social media are often tied to broader efforts of combating misinformation and disinformation. Platform-level interventions include content restrictions, warnings to users that content is misleading, and other nudges to check content validity before posting [Jahanbakhsh et al., 2021].

Social media is not limited to the lay public; Merchant and Asch [2018] discuss possible impacts of *scientists* sharing their work on social media directly, with a potential pitfall being a desire to spin findings more positively to get attention quickly. On the other hand, engaging with social media could also allow scientists to more directly counter misinformation and disinformation narratives [Iyengar and Massey, 2019].

Social costs of sensationalism. Misleading health reporting can have direct, damaging effects on health-related behaviors; for example, Shuchman and Wilkes [1997] describe the impact of exaggerated news reporting regarding calcium channel blockers for hypertension, which was stated to severely increase the likelihood of a heart attack, causing patients to feel distress or discontinue taking the medication altogether.[6]

Beyond immediate impact to individual patients, continued sensationalized reporting can also contribute to decreased trust in science and medical research. Scheufele and Krause [2019] note that while past studies find

[6]They note that the briefing and press release given by the American Heart Association likely had a role in the finding being exaggerated, suggesting multiple points of failure.

that public trust in science in the United States is still relatively high compared to other agencies, other surveys demonstrate public concern about exaggeration of scientific findings and associated news coverage.

4.3 NLP system interventions and guidelines

A few themes emerge from the surveyed work above:

1. Informing the public accurately is a shared responsibility across the publication pipeline — researchers, doctors, press offices, journalists, and other communications officials.

2. Interventions must take into account the underlying reasons for why claims might be exaggerated, caveats omitted, etc. by a user group.

3. The ways in which different user groups interpret the same health news finding may not be the same; e.g., a researcher may interpret the phrase "associated with" as a weaker claim than a lay reader would.

We could set up a task where, given two documents — a source document, like a press release, and a document derived from that source — identify whether (or to what extent) an article or claim is sensationalized.[7,8] Given the above principles, however, even a system successful at such a task is ill-defined as an intervention without:

(a) A target user group to be informed.

[7]For the purposes of this argument, we focus on the identification of sensationalism, but these principles can also extend to tools for summarizing or rewriting sensationalized text to be less so.

[8]This is not an entirely hypothetical task setup; see e.g., Li et al. [2017] and Wright and Augenstein [2021] for identifying exaggerated claims.

(b) A desired effect/action that those users take with the new information.

Without (a), the labeling criteria may not be well-suited for its actual use case. While a communications researcher may be interested in understanding/measuring specific aspects of sensationalism across a corpus, a journalist is focused on writing a single article, and a reader is just trying to figure out whether it's worth clicking on the article at all. The costs of misclassification also differ between measurement and intervention: for the former, such measurement errors can be accounted for, but a journalist or reader trusting that the article has been independently judged to not be misleading could lead to the social costs described in the previous section.

Without (b), then the system serves no applicable purpose from an intervention standpoint — the intervention's success cannot be measured. For example, from a PR or news perspective, stating that the headline exaggerates the finding may not be useful without more specific guidance on what to change and assurances that the alternative can still draw a reader's attention. The intervention's long-term success would be dependent on both whether the writer moderated headline claims when identified and whether those changes impacted readership; a secondary effect would be whether moderating claim strength has an impact on readers' interpretations of the claim.

This is not to suggest that NLP systems cannot aid end users in some way, just that such development should be taken in collaboration with a target user population and with a clearly specified intervention goal.

Costs of underdeveloped tasks. There are a couple of costs of detecting social phenomena without proper scoping of the usage scenario: (1) researchers may attempt to develop and optimize models on a task that, as described above, may not exactly fulfill a particular scenario; and (2) if such a model is deployed anyways, the public may either overly trust or, given evidence of consistent mislabeling, distrust the model's output.

A recent case within the NLP community is hate speech (or abusive & toxic language) detection, in which the broader goal is to identify or remove such speech from social media posts. Early work focused on dataset collection and model development without clear definitions of the kinds of speech being flagged; a cost incurred by the research community has been playing catch-up in terms of verification of possible biases in models (e.g., flagging African American English tweets as toxic: Sap et al., 2019; Davidson et al., 2019), understanding different community norms of what speech is considered harmful [Park et al., 2021], and accounting for possible biases introduced by annotators themselves [Sap et al., 2021]. Concerns about model biases extend beyond just models iterated on by the research community; the initial release of the Perspective API, a tool for identifying toxic language in an effort to promote healthier conversations online, was swiftly followed by criticism that certain identity-related terms (e.g., "black", "gay") would result in higher toxicity scores regardless of context.[9,10] Even though research and API development was done in the hope

[9] https://medium.com/jigsaw/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23

[10] https://algorithmwatch.org/en/automated-moderation-perspective-bias/

of benefiting populations targeted by toxic language, unintended harms towards those same populations were accrued as well.

Possible steps. The cost of developing multiple NLP systems to address the myriad scenarios above would likely be high; currently, semantic comparisons such as identifying entailment or similarity primarily operate at the sentence level and assume that the text being compared is within the same domain.

One avenue forward for model development without being tied to specific end-user scenarios is to decouple the determination of *how* two documents differ in meaning from the *judgment* of what that semantic difference means in a particular user-focused setting. For example, aligning similar statements between two documents is a step towards understanding what information has been carried over, added, or omitted, with the downstream application determining which additions or omissions are important. Similarly, for comparing parallel claims, instead of a flat label of exaggeration (or entailment), perhaps a system can characterize more transparently how the semantic roles involved differ; the application can then dictate whether certain generalizations of terms or strengthening of predicates is misleading.

This approach parallels our earlier work on semantic matching: the underlying models for semantic comparison at the sentence level have been iterated on using standard semantic comparison datasets (e.g., MultiNLI) and become more powerful over time. However, we recognize that those datasets may not directly transfer to the end use case, and the ultimate

decision of what a semantic match means is still left to the practitioner to decide.

4.4 Related work

Exaggerated claims & claim strength. There are a number of studies that seek to identify exaggerated claims, often from the standpoint of measuring or characterizing textual aspects, including: Tan and Lee [2014], which compares statement strength across arXiv paper revisions; Yu et al. [2019] and Yu et al. [2020], which examine the use of causal language in PubMed abstracts and press releases respectively; and Wright and Augenstein [2021], which uses a multi-task approach for predicting whether one claim exaggerates the other as well as the strength of each claim. On the flip side, Starbird et al. [2016] and Pei and Jurgens [2021] focus on measuring uncertainty in scientific communications, and Farkas et al. [2010] provides a shared task to identify hedging in scientific text.

Other related problems. There are a number of other related problems, largely around identifying veracity or attempts to mislead. Deception detection seeks to identify whether a text is intentionally trying to mislead others [Rubin and Vaschilko, 2012]. Related work on truth discovery seeks to identify the truthful claims among a larger set of claims given a knowledge base; see Berti-Équille and Borge-Holthoefer [2015] for a survey.

Fake news detection has been formalized as several different tasks, including classification on the news article alone [Pérez-Rosas et al., 2018;

Zellers et al., 2019]; understanding stylistic aspects of fake news [Rashkin et al., 2017]; and subtasks like stance detection, used by the Fake News Challenge,[11] where a text is judged as for, against, or observing a claim [Ferreira and Vlachos, 2016]. Propaganda detection is closely related as well [Da San Martino et al., 2019]. However, these settings assume an intention by the writer to lie or mislead, which may not be the case in sensationalized text.

4.5 Conclusion

In this chapter, we considered the possibility of using semantic comparison methods to identify sensationalism in medical journalism. We surveyed past studies across communications, medicine, and psychology to illustrate properties of sensationalism, its occurrence in the health communications landscape, and why it might surface at different steps in the pipeline. In doing so, we critiqued the common NLP setup of attempting to label social phenomena in text with high accuracy and provided suggestions for developing user-facing NLP systems that seek to identify or reduce the occurrence of sensationalism in medical journalism.

[11] http://www.fakenewschallenge.org

Chapter 5

Conclusion

In this thesis, we proposed *semantic comparison* as another lens for studying social phenomena in text data (beyond text classification, say), in an effort to bridge the gap between NLP modeling and downstream practice. We explored two novel applications of semantic comparison methods for which standard abstractions are insufficient.

In Chapter 2, we tackled the scenario where a user seeks to identify occurrences of an idea in a text corpus. We introduced a framework based on semantic matching of a proposition query and sentences in the corpus, and then discuss considerations involved in selecting a matching function. We then demonstrated the applicability of semantic matching through two case studies in collaboration with domain experts: community recovery after the 2010–2011 Christchurch, New Zealand earthquake sequence, as expressed in local news text in the five years afterward (§2.3); and policy positions in the United States Congress across 2000–2013, as expressed in archived websites from the `.gov` domain (§2.4).

We found that one of the key challenges in this was model selection without the benefit of annotated training data in the application domain. To ameliorate this, in Chapter 3 we introduced a method (nearest neighbor overlap) for comparing the similarity of sentence embedder behavior in the context of a target corpus.

In Chapter 4, we explored the possibility of using semantic comparisons to identify sensationalism in medical journalism. We surveyed past studies across communications, medicine, and psychology to illustrate where and how sensationalism manifests in the health communications pipeline, the incentives involved, and possible interventions. In doing so, we critiqued the common NLP setup of attempting to label social phenomena in text with high accuracy and provided recommendations for developing end-user NLP systems that seek to identify or reduce the occurrence of sensationalism.

Throughout this thesis, we envision the use of semantic comparisons in the context of specific users and scenarios. While we are reliant on the ability to model semantic similarity or differences between texts, particular semantic comparisons operationalized by labeled datasets may not generalize neatly to other applications. A final takeaway from this work is to encourage practitioners to apply such models with caution and validate whether their behavior aligns with the end goal.

Bibliography

Benjamin Adams and Martin Raubal. 2014. Identifying salient topics for personalized place similarity. *Research@Locate*.

Rachel C. Adams, Aimée Challenger, Luke Bratton, Jacky Boivin, Lewis Bott, Georgina Powell, Andy Williams, Christopher D. Chambers, and Petroc Sumner. 2019. Claims of causality in health news: a randomised trial. *BMC Medicine*, 17(91).

Rachel C. Adams, Petroc Sumner, Solveiga Vivian-Griffiths, Amy Barrington, Andrew Williams, Jacky Boivin, Christopher D. Chambers, and Lewis Bott. 2017. How readers understand causal and correlational expressions used in news headlines. *J. Experimental Psychology: Applied*, 23(1):1–14.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proc. of ICLR*.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter datasets from natural disasters. In *Proc. of ICWSM*.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *TACL*, 6:107–119.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proc. of ICLR*.

Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2019. ANN-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, in press.

Niranjan Balasubramanian, James Allan, and W. Bruce Croft. 2007. A comparison of sentence retrieval techniques. In *Proc. of SIGIR*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proc. of TAC*.

Laure Berti-Équille and Javier Borge-Holthoefer. 2015. Veracity of data: from truth discovery computation algorithms to models of misinformation dynamics. *Synthesis Lectures on Data Management*, 7(3):1–155.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proc. of NAACL-HLT*.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proc. of ICML*.

Lewis Bott, Luke Bratton, Bianca Diaconu, Rachel C. Adams, Aimeé Challenger, Jacky Boivin, Andrew Williams, and Petroc Sumner. 2019.

Caveats in science-based news stories communicate caution without lowering interest. *J. Experimental Psychology: Applied*, 25(4):517–542.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP*.

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues. In *APSA 2014 Annual Meeting*.

Luke Bratton, Rachel C. Adams, Aimée Challenger, Jackie Boivin, Lewis Bott, Christopher D. Chambers, and Petroc Sumner. 2019. The association between exaggeration in health-related science news and academic press releases: a replication study. *Wellcome Open Research*, 4(148).

Jean Brechman, Chul-joo Lee, and Joseph N. Cappella. 2009. Lost in translation? A comparison of cancer-genetics reporting in the press release and its subsequent coverage in the news. *Science Communication*, 30(4):453–474.

Tania M. Bubela and Timothy A. Caulfield. 2004. Do the print media "hype" genetic research? A comparison of newspaper stories and peer-reviewed research papers. *CMAJ*, 170(9):1399–1407.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of frames across issues. In *Proc. of ACL*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proc. of SemEval*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. arXiv:1803.11175 [cs.CL].

Stephanie E. Chang, Timothy McDaniels, Jana Fox, Rajan Dhariwal, and Holly Longstaff. 2014a. Toward disaster-resilient cities: Characterizing resilience of infrastructure systems with expert judgments. *Risk Analysis*, 34(3):416–434.

Stephanie E. Chang, Josh E. Taylor, Kenneth J. Elwood, Erica Seville, Dave Brunsdon, and Mikaël Gartner. 2014b. Urban disaster recovery in Christchurch: The Central Business District Cordon and other critical decisions. *Earthquake Spectra*, 30(1):513–532.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proc. of ACL*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. Reading Wikipedia to answer open-domain questions. In *Proc. of ACL*.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Enhanced LSTM for natural language inference. In *Proc. of ACL*.

Michael A. Cohn, Matthias R. Mehl, and James W. Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, 15(10):687–693.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proc. of LREC*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proc. of ACL*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proc. of ACL*.

Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proc. of SIGIR*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proc. of EMNLP*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, Lecture Notes in Computer Science, pages 177–190. Springer.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proc. of the Third Workshop on Abusive Language Online*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*.

Jean-François Doherty. 2020. When fiction becomees fact: Exaggerating host manipulation by parasites. *Proc. of the Royal Society B*, 287(1936).

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proc. of IWP*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proc. of COLING*.

Sarah K. Dreier, Lucy H. Lin, Sofia Serrano, Emily K. Gade, and Noah A. Smith. in prep. Qualitative and NLP approaches to measuring Congressional behavior: The partisan dimensions of religious rhetoric. Presented at *Text as Data 2018*.

Estelle Dumas-Mallet and Francois Gonon. 2020. Messaging in biological psychiatry: Misrepresentations, their causes, and potential consequences. *Harvard Review of Psychiatry*, 28(6):395–403.

William duPont and Ilan Noy. 2015. What happened to Kobe? A reassess-

ment of the impact of the 1995 earthquake in Japan. *Economic Development & Cultural Change*, 63(4):777–812.

Kevin M. Esterling, David M. J. Lazer, and Michael A. Neblo. 2010. *Improving Congressional Websites*. Center for Technology Innovation at Brookings.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proc. of EMNLP-IJCNLP*.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proc. of RepEval*.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proc. of CoNLL – Shared Task*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proc. of NAACL-HLT*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proc. of NAACL-HLT*.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019. Text as data. *J. Economic Literature*, 57(3):535–574.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. arXiv:1811.08008 [cs.IR].

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proc. of ACL*.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proc. of ACL*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proc. of NAACL-HLT*.

Michael Heilman and Noah A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Proc. of NAACL-HLT*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proc. of NeurIPS*.

Katherine Hicks-Courant, Jenny Shen, Angela Stroupe, Angel Cronin, Elizabeth F. Bair, Sam E. Wing, Ernesto Sosa, Rebekah H. Nagler, and Stacy W. Gray. 2021. Personalized cancer medicine in the media: Sensationalism or realistic reporting? *J. Personalized Medicine*, 11(8):741.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Matthew B. Hoy. 2020. Rise of the Rxivs: How preprint servers are changing the publishing process. *Medical Reference Services Quarterly*, 39(1):84–89.

Masajiro Iwasaki and Daisuke Miyazaki. 2018. Optimization of indexing based on k-nearest neighbor graph for proximity search in high-dimensional data. arXiv:1810.07355 [cs.DB].

Shanto Iyengar and Douglas S. Massey. 2019. Scientific communication in a post-truth society. *PNAS*, 116(16):7656–7661.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proc. of ACL*.

Trevor Jackson. 2003. MMR: More scrutiny, please. *BMJ: British Medical Journal*, 326(7401):1272.

Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. 2021. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. In *Proc. of CSCW*.

Samuel Jellison, Will Roberts, Aaron Bowers, Tyler Combs, Jason Beaman, Cole Wayant, and Matt Vassar. 2019. Evaluation of spin in abstracts of pa-

pers in psychiatry and psychology journals. *BMJ Evidence Based Medicine*, 25:178–181.

Jocelyn Kaiser. 2017. The preprint dilemma. *Science*, 357(6358):1344–1349.

Jim Kennedy, Joseph Ashmore, Elizabeth Babister, and Ilan Kelman. 2008. The meaning of 'build back better': Evidence from post-tsunami Aceh and Sri Lanka. *J. Contingencies & Crisis Management*, 16(1):24–36.

Evan D. Kharasch, Michael J. Avram, J. David Clark, Andrew J. Davidson, Timothy T. Houle, Jerrold H. Levy, Martin J. London, Daniel I. Sessler, and Laszlo Vutskits. 2021. Peer review matters: Research quality and the public trust. *Anesthesiology*, 134(1):1–6.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SCITAIL: A textual entailment dataset from science question answering. In *Proc. of AAAI*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Klaus Krippendorff. 2012. *Content analysis: An introduction to its methodology*.

Geoffrey Layman. 1999. 'Culture wars' in the American party system: Religious and cultural change among partisan activists since 1972. *American Politics Quarterly*, 27(1):89–121.

Julie Leask, Claire Hooker, and Catherine King. 2010. Media coverage of health issues and how to work more effectively with journalists: A qualitative study. *BMC Public Health*, 10:535.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proc. of KDD*.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP analysis of exaggerated claims in science news. In *Proc. of EMNLP Workshop: Natural Language Processing Meets Journalism*.

Lucy H. Lin, Scott Miles, and Noah A. Smith. 2018a. Natural language processing for analyzing disaster recovery trends expressed in large text corpora. In *Proc. of IEEE Global Humanitarian Technology Conference*.

Lucy H. Lin, Scott Miles, and Noah A. Smith. 2018b. Semantic matching against a corpus: New applications and methods. arXiv:1808.09502 [cs.CL].

Lucy H. Lin and Noah A. Smith. 2019. Situating sentence embedders with nearest neighbor overlap. arXiv: 1909.10724 [cs.CL]; presented at *Text as Data 2019*.

Yu-Ru Lin and Drew Margolin. 2014. The ripple of fear, sympathy and solidarity during the Boston bombings. *EPJ Data Science*, 3(1):31.

Wendy Lipworth, Melanie Gentgall, Ian Kerridge, and Cameron Stewart. 2020. Science at warp speed: Medical research, publication, and translation during the COVID-19 pandemic. *J. Bioethical Inquiry*, 17(4):555–561.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.

2019. RoBERTa: A robustly optimized BERT training approach. arXiv: 1907.11692 [cs.CL].

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *TACL*, 9:329–345.

Forrest Maltzman and Lee Sigelman. 1996. The politics of talk: Unconstrained floor time in the U.S. House of Representatives. *The Journal of Politics*, 58(3):819–830.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of ACL (System Demonstrations)*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proc. of LREC*.

Morgan Marietta. 2009. The absolutist advantage: Sacred rhetoric in contemporary presidential debate. *Political Communication*, 26(4):388–411.

Timothy McDaniels, Stephanie E. Chang, Krista Peterson, Joey Mikawoz, and Dorothy Reed. 2007. Empirical framework for characterizing infrastructure failure interdependencies. *J. Infrastructure Systems*, 13(3):175–184.

Raina M. Merchant and David A. Asch. 2018. Protecting the value of

medical science in the age of social media and "fake news". *JAMA*, 320(23):2415–2416.

Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel. 2005. Similarity measures for tracking information flow. In *Proc. of CIKM*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffery Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of NeurIPS*.

Scott B. Miles. 2015. Foundations of community disaster resilience: Well-being, identity, services, and capitals. *Environmental Hazards*, 14(2):103–121.

Scott B. Miles, Dana Brechwald, Rachel Davidson, Katalin Demeter, David Johnston, Stefano Pampanin, and Suzanne Wilkinson. 2014. Building back better — case study of the 2010–2011 Canterbury, New Zealand earthquake sequence. Technical report, Earthquake Engineering Research Institute.

Katarzyna Molek-Kozakowska. 2013. Towards a pragma-linguistic frame-

work for the study of sensationalism in news headlines. *Discourse & Communication*, 7(2):173–197.

Yasmina Molero, Paul Lichtenstein, Johan Zetterqvist, Clara Hellner Gumpert, and Seena Fazel. 2015. Selective serotonin reuptake inhibitors and violent crime: A cohort study. *PLoS Med.*, 12(9):e1001875.

Jane Morgan, Annabel Begg, Sarah Beaven, Philip Schluter, Kath Jamieson, Sarb Johal, David Johnston, and Mary Sparrow. 2015. Monitoring well-being during recovery from the 2010–2011 Canterbury earthquakes: The CERA Wellbeing Survey. *International Journal of Disaster Risk Reduction*, 14:96–103.

Ray Moynihan, Lisa Bero, Dennis Ross-Degnan, David Henry, Kirby Lee, Judy Watkins, Connie Mah, and Stephen B. Soumerai. 2000. Coverage by the news media of the benefits and risks of medications. *New England Journal of Medicine*, 342(22):1645–1650.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R. Bowman. 2017. The RepEval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proc. of RepEval*.

Pippa Norris and Ronald Inglehart. 2011. *Sacred and Secular: Religion and Politics Worldwide*. Cambridge University Press.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. of the 4th International AAAI Conference on Weblogs and Social Media*.

Brendan O'Connor, David Bamman, and Noah A. Smith. 2011. Computational text analysis for social science: Model assumptions and complexity. In *Proc. of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. 2013. Learning to extract international relations from political context. In *Proc. of ACL*.

Mary O'Keeffe, Brooke Nickel, Thomas Dakin, Chris G. Maher, Loai Albarqouni, Kirsten McCaffery, Alexandra Barratt, and Ray Moynihan. 2021. Journalists' views on media coverage of medical tests and overdiagnosis: A qualitative study. *BMJ Open*, 11(6):e043991.

Elizabeth A. Oldmixon and William Hudson. 2008. When church teachings and policy commitments collide: Perspectives on Catholics in the U.S. House of Representatives. *Politics and Religion*, 1(1):113–136.

Tracy Osborn and Jeanette Morehouse Mendez. 2010. Speaking as women: Women and floor speeches in the Senate. *J. Women, Politics & Policy*, 31(1):1–21.

Ryan Ottwell, Madison Puckett, Taylor Rogers, Savannah Nicks, and Matt Vassar. 2021. Sensational media reporting is common when describing COVID-19 therapies, detection methods, and vaccines. *J. Investigative Medicine*, 69:1256–1257.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit.

2016. A decomposable attention model for natural language inference. In *Proc. of EMNLP*.

Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting community sensitive norm violations in online conversations. In *Findings of ACL: EMNLP 2021*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition, LDC2011T07. Linguistic Data Consortium.

Jiaxin Pei and David Jurgens. 2021. Measuring sentence-level and aspect-level (un)certainty in science communications. In *Proc. of EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of EMNLP*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proc. of COLING*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT*.

Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proc. of ACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Preprint.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Preprint.

Nairán Ramírez-Esparza, Cindy K. Chung, Gisela Sierra-Otero, and James W. Pennebaker. 2012. Cross-cultural constructions of self-schemas: Americans and Mexicans. *J. Cross-Cultural Psychology*, 43(2):233–250.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proc. of EMNLP*.

Niels Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. of EMNLP-IJCNLP*.

Victoria L. Rubin and Tatiana Vaschilko. 2012. Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. In *Proc. of the EACL 2012 Workshop on Computational Approaches to Deception Detection*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proc. of ACL*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. arXiv:2111.07997 [cs.CL].

Dietram A. Scheufele and Nicole M. Krause. 2019. Science audiences, misinformation, and fake news. *PNAS*, 116(16):7662–7669.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *Proc. of ICASSP*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL*.

Minjoon Seo, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Phrase-indexed question answering: A new challenge for scalable document comprehension. In *Proc. of EMNLP*.

Megha Sharma, Kapil Yadav, Nitika Yadav, and Keith C. Ferdinand. 2016. Zika virus pandemic — analysis of Facebook as a social media health information platform. *American Journal of Infection Control*, 45(3):301–302.

Miriam Shuchman and Michael S. Wilkes. 1997. Medical scientists and health news reporting: A case of miscommunication. *Annals of Internal Medicine*, 126(12):976–982.

Gavin P. Smith and Dennis Wenger. 2007. Sustainable disaster recovery: Operationalizing an existing agenda. *Handbook of Disaster Research*, pages 234–257.

Kate Starbird, Emma Spiro, Isabelle Edwards, Kaitlyn Zhou, Jim Maddock, and Sindu Narasimhan. 2016. Could this be true? I think so! Expressed uncertainty in online rumoring. In *Proc. of CHI*.

Kohei Sugawara, Hayato Kobayashi, and Masajiro Iwasaki. 2016. On approximately searching for similar word embeddings. In *Proc. of ACL*.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andrew Williams, Lewis Bott, Rachel Adams, Christos A. Venetis, Leanne Whelan, Bethan Hughes, and Christopher D. Chambers. 2016. Exaggerations and caveats in press releases and health-related science news. *PLoS One*, 11(12):e0168217.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A. Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D. Chambers. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, 349:g7015.

Chenhao Tan, Dallas Card, and Noah A. Smith. 2017. Friendships, rivalries, and trysts: Characterizing relations between ideas in texts. In *Proc. of ACL*.

Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proc. of ACL*.

Percy H. Tannenbaum and Mervin D. Lynch. 1960. Sensationalism: The concept and its measurement. *Journalism Quarterly*, 37(3):381–392.

Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proc. of SIGIR*.

Ingrid Torjesen. 2015. Study links SSRIs to violent crime in young adults. *BMJ*, 351.

Caitlyn Vlasschaert, Joel M. Topf, and Swapnil Hiremath. 2020. Proliferation of papers and preprints during the coronavirus disease 2019 pandemic: Progress or problems with peer review? *Advances in Chronic Kidney Disease*, 27(5):418–426.

Kim Walsh-Childers, Jennifer Braddock, Cristina Rabaza, and Gary Schwitzer. 2018. One step forward, one step back: Changes in news coverage of medical interventions. *Health Communications*, 33(2):174–187.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. of ICLR*.

Michael T. M. Wang, Andrew Grey, and Mark J. Bolland. 2017. Conflicts of interest and expertise of independent commenters in news stories about medical research. *CMAJ*, 189:E553–559.

Claire Wardle. 2019. Understanding information disorder. Technical report, First Draft.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *Proc. of ICLR*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL-HLT*.

Amanda Wilson, Billie Bonevski, Alison Jones, and David Henry. 2009. Media reporting of health interventions: Signs of improvement, but major problems persist. *PLoS One*.

Steven Woloshin and Lisa M. Schwartz. 2002. Press releases: Translating research into news. *JAMA*, 287(21):2856–2858.

Steven Woloshin, Lisa M. Schwartz, Casella Samuel L., Abigail T. Kennedy, and Robin J. Larson. 2009. Press releases by academic medical centers: not so academic? *Annals of Internal Medicine*, 150:613–618.

Dustin Wright and Isabelle Augenstein. 2021. Semi-supervised exaggeration detection of health science press releases. In *Proc. of EMNLP*.

David Yamane and Elizabeth A. Oldmixon. 2006. Religion in the legislative arena: Affiliation, salience, advocacy, and public policymaking. *Legislative Studies Quarterly*, 31(3):433–460.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. In *Proc. of EMNLP-IJCNLP*.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring correlation-to-causation exaggeration in press releases. In *Proc. of COLING*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Proc. of NeurIPS*.

Xunjie Zhu, Tingfeng Li, and Gerard de Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proc. of ACL*.

Appendix A

Chapter 2 Supplementary

A.1 Model & preliminary evaluation details

(See next page.)

Frame	Proposition query
Crime & punishment	Punishments should be softer on immigration. Immigrants are no more likely to engage in criminal activity. Immigrants are more likely to deal or transport drugs.
Cultural identity	Immigrants do assimilate and have similar values to us. Immigrants are taking over the country. Immigrants have conflicted loyalties and nationalistic sentiments.
Economic	Highly skilled workers are attracted to work in the United States. Immigrants work for less money, driving the wages down for domestic workers. Immigrants often pay into the system, but do not qualify to receive government benefits.
Fairness & equality	Immigration rules have changed unfairly over time. Immigrants cannot wait for the system in place because it is not fair. Allowing unauthorized immigration is unfair to those who apply and wait. Law enforcement officials use racial and ethnic stereotypes to unfairly discriminate. The penalties for illegal immigration should fall on the individuals breaking the law, not businesses.
Health & safety	Immigrants aid law enforcement by acting as witnesses. Immigrants who try to enter the country illegally are responsible for any safety hazards they incur.
Legality & constitutionality	For free trade to be successful, there should be a free movement of people. Right to work does not mean right to cross national borders. The regulation of immigration should be done through Congress.

(Table A.1 continued on next page.)

(Table A.1 continued from previous page.)

Morality	It would be immoral to turn our backs on those in need. We have no moral obligation to help those who break the law. No path to citizenship creates permanent second-class citizens. Supporting the poor does not mean supporting immigrants.
Politics	The immigration issue is a way for politicians to pander to the Hispanic community. Businesses have a legitimate interest in lobbying for immigration issues.
Public sentiment	The public supports immigration rights. Public support for immigration should not influence policy.
Quality of life	Immigrants drive up the cost of living. Immigrants have a positive impact on diversity in the United States. Immigrants deserve better quality of life than they can get in their home countries

Table A.1: Proposition queries used in the Media Frames Corpus evaluation (§2.3.2).

Feature category	Description
General counts	# of edits in the sequence; #s of X edits (where X is one of the operations in Table 2.1).
INSERT-CHILD, INSERT-PARENT	#s of these which: insert nouns or verbs, insert proper nouns.
DELETE-LEAF, DELETE-&-MERGE	#s of these which: remove nouns or verbs, remove proper nouns, remove nodes with subject edge labels, remove nodes with object edge labels, remove nodes with verb complement edge labels, remove nodes with root edge labels (which may occur after NEW-ROOT edits).
RELABEL-NODE	#s of these which: preserve POS, preserve lemmas, convert between nouns and pronouns, change proper nouns, change numeric values by more than 5% (to allow rounding).
RELABEL-EDGE	#s of these which: change to or from subject edge labels, change to or from object edge labels, change to or from verb complement edge labels, change to or from root edge labels.
Unedited node counts	In total, numeric values, verbs, nouns, proper nouns.
Other	If a tree edit sequence was found or not.

Table A.2: Tree edit features for logistic regression classification from Heilman and Smith [2010].

A.2 Disaster recovery proposition queries

Proposition query
• Residents are frustrated by the slow pace of recovery.
• The repair programme is on schedule to be completed.
• Money for repairs is running out.
• The council should have consulted residents before making decisions.
• Mental health rates have been rising.
• Dealing with authorities is causing stress and anxiety.
• Most eligible property owners have accepted insurance offers.
• Confidence in Cera has been trending downwards.
• Water quality declined after the earthquakes.
• The power system was fully restored quickly.
• Cera missed several recovery milestones.
• Prices levelled off as more homes were fixed or rebuilt.
• People are suffering because they've lost the intimacy of their relationships.
• Coordination between rebuild groups has been problematic.
• Few people said insurance companies had done a good job.
• Having the art gallery back makes the city feel more whole.
• Scirt has spent less money than predicted.
• Traffic congestion was severe due to road repairs.
• Some of the businesses forced out by the earthquake are returning.
• Some of the burden on mental health services is caused by lack of housing.

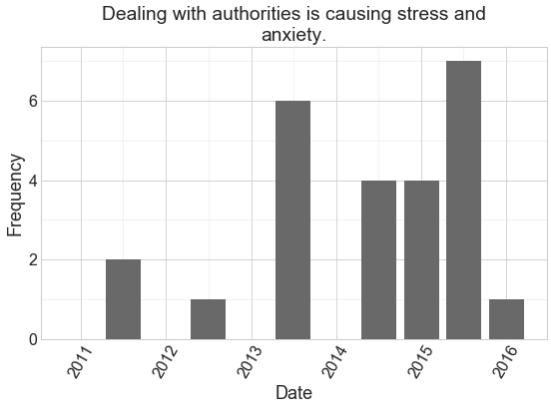
Table A.3: Proposition queries used in the initial disaster recovery user study (§2.3.4). ("Cera" is short for the "Canterbury Earthquake Recovery Authority", and "Scirt" is short for the "Stronger Christchurch Infrastructure Rebuild Team.")

Proposition query
• The cost of repairs is over budget.
• People are worried that rents will rise.
• There are many homeless people in need of shelter.
• People called on local corporations to provide additional aid.
• The city council could not agree on a plan forward.
• Economic inequality grew after the earthquake.
• There was a shortage of food and water.
• Public transit reroutes and delays caused frustration.
• Residents demanded accountability from government agencies.
• Small businesses are hit hard by rebuild costs and decreased sales.
• Access to electricity continues to be unreliable.
• People are struggling to get to their jobs.
• People feel less safe in the city.
• Hospitals have trouble accommodating all those who need health services.
• Residents note a greater sense of community within the neighborhood.
• People feel disconnected to the outside world due to unreliable internet access.
• Donations continue to flood in.
• The earthquake has exacerbated the housing shortage.

Table A.4: Proposition queries used in the disaster recovery follow-up study (§2.3.5).

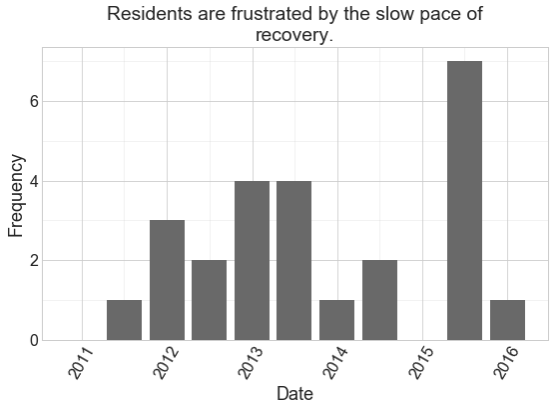
A.3 Disaster recovery histograms

(See next page.)



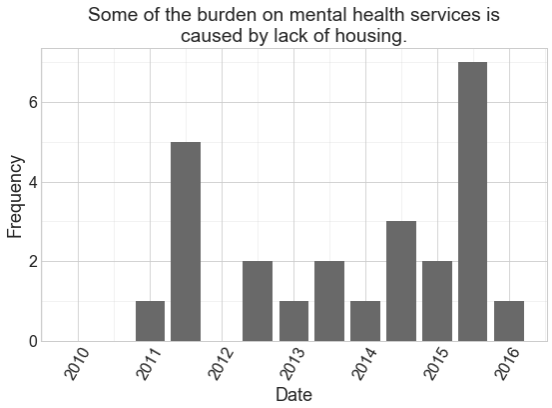
1. Stress and anxiety caused by dealing with authorities was “more debilitating” than the quakes, she said.
2. The battle to get back what they lost created huge financial stress.
3. $800m spent on east Christchurch not enough for frustrated residents
4. Most policies have clauses allowing replacement with “materials in common use”, frustrating owners of older villas and bungalows.
5. A Canterbury Earthquake Recovery Authority draft document on their psychosocial plan for the city says anxiety and stress will continue to dog the population due to ongoing battles with insurance, land issues, changes to schooling and problems rebuilding homes and businesses.
6. The key findings indicate that the secondary stressors of damaged homes, insurance wrangles, financial challenges and grief over the ‘lost Christchurch’ are taking its toll.
7. Add to this the growing frustration among the new, youthful leaders of the community who emerged in the wake of the quakes.
8. Mr Hodder accepted that the homeowners involved have been subjected to “great distress” but added that it was a “difficult” situation and it has taken time to get to this point.
9. ICNZ chief executive Tim Grafton said insurers could understand homeowners, whose claim had just reached their insurer, were frustrated.
10. Council anchor projects unit manager Liam Nolan said the council recognised retailers’ frustration that the carpark has not come down sooner.

Figure A.1: Histogram and randomly-selected subset of sentences found as expressing the proposition “Dealing with authorities is causing stress and anxiety” in our NZ news text corpus.



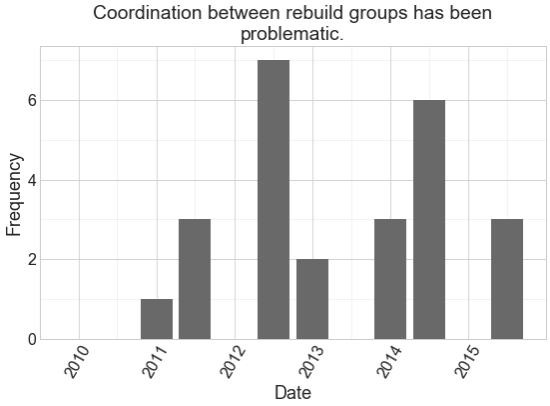
1. Residents are frustrated by the slow pace of recovery and the lack of proposed city council spending in New Brighton and surrounding suburbs.
2. She will also be asking them to put aside an extra $3m for the rejuvenation of New Brighton as part of a package of proposals aimed at addressing residents' concerns about the slow pace of recovery in the east.
3. It has slowed their recovery down hugely and made life for many much harder.
4. The submissions expressed frustration at the slow pace of recovery and the lack of proposed spending in New Brighton and the eastern suburbs.
5. Power has been restored to over 60 per cent of quake-ravaged Christchurch but progress is slow, lines company Orion says.
6. Read more: Christchurch anchor project delays cause frustration.
7. Many Christchurch residents have expressed anger at the slow pace of settling claims for damaged homes.
8. People are also learning to adjust to the pace of the recovery.
9. Worry, despair plague Christchurch residents.
10. He said progress in Christchurch is "far too slow".

Figure A.2: Histogram and randomly-selected subset of sentences found as expressing the proposition "Residents are frustrated by the slow pace of recovery" in our NZ news text corpus.



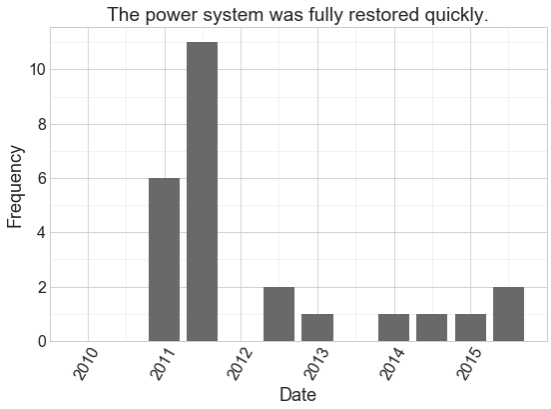
1. "We can see what it is that is causing the mental health issues and we know that if this person could get their housing issues sorted they would be transformed."
2. Four years on we are in the early days of recovery, and the anxiety about aftershocks has given way to stresses about insurance, repairs and relocating offices, schools and homes.
3. Priority has been given to essential service locations [medical facilities, etc] and main streets and thoroughfares."
4. Key urged anxious homeowners to be patient.
5. Some of the burden on mental health services is undoubtedly caused by the lack of housing in Christchurch.
6. Housing security was vital for recovery from mental health illness and for independent living, Duffy said.
7. Generally in Auckland, there will be more demand for residential building because of the lack of housing supply.
8. The pressure on mental health services continued to rise with more homeless mental health patients taking up beds.
9. For the majority of mental health patients cared for in the community, housing stress was a big element in their ill-health.
10. People outside Christchurch also find it hard to understand how mental health could be so affected by the insecurity of housing, or "the importance of your home to your sense of freedom and personal dignity".

Figure A.3: Histogram and randomly-selected subset of sentences found as expressing the proposition "Some of the burden on mental health services is caused by lack of housing" in our NZ news text corpus.



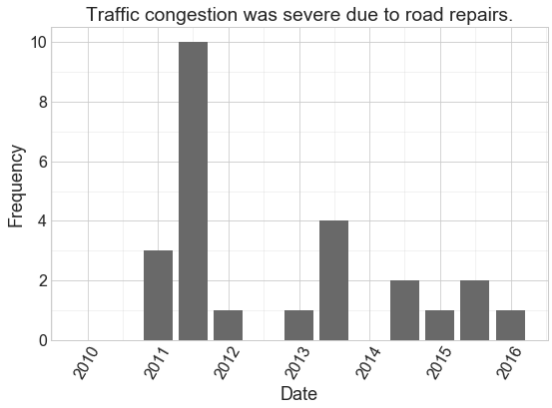
1. Householders often have different policies and various agencies have to be co-ordinated.
2. There will be lots of co-ordination needed between stakeholders, designers and industry.
3. After more than two years, and tired of fighting insurance companies and bureaucracy for what they believe is a fair deal, the group, described as the "Buggered All Stars Chorus", hope to use the song to raise awareness around New Zealand of their situation.
4. The challenge for the redevelopment of the city is to build demand for commercial, residential and retail space while planning for that redevelopment to occur in a coordinated way that lives up to the vision in the Central City Plan.
5. Appointing commissioners is not the answer to these difficult problems.
6. The Fletcher hubs oversee reconstruction and repair work throughout the rebuild area, where teams of building managers and advisers assess and co-ordinate work.
7. The new authority would pull together all of the resources of central government going into the city and co-ordinate where they went.
8. Calls are flooding into police communications centres around New Zealand, causing problems for staff already trying to deal with the massive earthquake and aftershocks which have hit Christchurch and Canterbury.
9. Another arm's length relationship was created that became a prime problem for the Blueprint [recovery plan].
10. He believed a coordinated response from the various organisations and agents involved with Christchurch's property market was key to dealing with the housing shortage.

Figure A.4: Histogram and randomly-selected subset of sentences found as expressing the proposition "Coordination between rebuild groups has been problematic" in our NZ news text corpus.



1. Canterbury electricity supplier Orion is confident that 90% of Christchurch city will have power restored by nightfall, says spokesman Roger Sutton.
2. Emergency services work frantically to restore water, power and sewerage systems and check for people who might still be trapped in their homes.
3. Then Orion's chief executive, his handling of the power lines company's staff working long hours to restore electricity to the city post-quake impressed many.
4. He had no water but power had been restored in his area.
5. The need to regain some sense of some control over one's life is central to the recovery process.
6. As in Christchurch, electricity, sewerage and chimneys were all down in post-quake Hawke's Bay.
7. Lines company Orion said the last of Christchurch's main power substations to be connected - Brighton substation - was confirmed as working yesterday.
8. City councillors have voted to fully restore the earthquake-damaged Christchurch Town Hall.
9. It had been unable to access the electricity network to restore power and the situation could remain for the next few days.
10. An update from Orion showed power had been restored to 87 per cent of Christchurch yesterday.

Figure A.5: Histogram and randomly-selected subset of sentences found as expressing the proposition "The power system was fully restored quickly" in our NZ news text corpus.



1. Karen Atkinson just up his road was on the verge of agreeing to a $280,000 repair job on her home. She called a rapid halt after her insurer told her that, if she wanted her house lifted, she would have to pay for it herself.
2. More than 1.2 million homes were without water, traffic was congested and fires started from gas mains rupturing could not be fought.
3. There was significant traffic congestion and drivers were told to avoid Ferry Road.
4. Congestion was making the road repairs more difficult and was holding up the delivery and servicing of portable toilets, the movement of trucks removing silt and numerous other important services.
5. Campbell-Reid says the council has developed strong criteria around the performance of new buildings, and it is looking at converting road space into green space so the central city streets become about people rather than rush-hour traffic.
6. Emergency service lines are working but remained heavily congested.
7. Civil Defence is asking people not to move road barriers and other signs on Christchurch city streets.
8. People may have been crushed or trapped by collapsing buildings, and brick facades.
9. Three years on from the February 2011 earthquake, the city's road network continues to be disrupted with lines of traffic cones, constant road closures and diversions.
10. People were being asked to stay off the roads to reduce congestion.

Figure A.6: Histogram and randomly-selected subset of sentences found as expressing the proposition "Traffic congestion was severe due to road repairs" in our NZ news text corpus.

A.4 Policy proposition queries

Social welfare spending
• Welfare builds a healthy America.
• America deserves a better health care policy.
• *The Democratic budget reflects what is important to America's families – a safe America with good jobs, better access to health care, the best possible education for our children, and a clean and healthy environment.* – attempt to see if longer sentences (pulled from the corpus as a seed query) would yield good matches; instead provided noisy output.
• *Helping minority and underprivileged families navigate the health system is a priority for America.* – yielded matches that were too general.
• America faces a health care crisis.
• *Families become dependent on welfare.* – attempted negation; in practice, this is generally not so flatly stated on Senate/House pages.
• Government should provide economic benefits to American communities.
• Federal government should provide health care and economic support to help our communities thrive.
• *Government should protect American families over corporate interests.* – mostly yielded false positives and partisan attacks.
• Welfare helps American families thrive.
• Temporary assistance helps needy American families thrive.
• Child welfare programs help needy American families thrive.
• Poor Americans need a domestic hunger safety net.

Table A.5: All examined proposition queries for social welfare policy attitudes. Italicized queries were cut before the analysis in §2.4.5.

Regulating sexual and reproductive behavior
• States should defend traditional marriage and oppose gay marriage.
• Americans overwhelmingly oppose same-sex marriage.
• *States should support/legalize gay marriage.* – attempted negation; we ended up needing to use the "marriage equality" framing to successfully match negations.
• *Americans overwhelmingly support same-sex marriage.* – attempted negation.
• Americans support marriage equality.
• Marriage is between a man and a woman.
• *Children need a mother and a father.* – matches more broadly about family/parenting.
• Gay marriage causes societal collapse.
• *I oppose same-sex marriage.* – matches were very similar to "Americans overwhelmingly oppose same-sex marriage," so just kept the former.
• I do not support gay marriage.
• *Weakening the legal status of marriage will only exacerbate these problems.* – generally unreliable matches.
• *I support same-sex marriage.* – attempted negation.
• Gay marriage erodes/deteriorates traditional marriage and family.
• Traditional marriage and the family are the foundation of American society.

(Table A.6 continued on next page)

(Table A.6 continued from previous page)

- *Government must protect the lives of unborn or pre-born children.* – matches more broadly about children and child welfare.
- Abortion kills unborn children.
- Partial birth abortion is murder.
- Partial birth abortion is a violent procedure that is truly traumatic for the mother and her unborn child.
- Partial birth abortion is cruel and inhumane.
- A woman has the right to have an abortion.
- *Women should have control/autonomy over their bodies.* – attempted negation; matches more broadly about autonomy (e.g., agency or state autonomy).
- *Providers can exercise religious freedom to refuse providing contraception and abortion.* – matches captured contraception/religious freedom content, but not necessarily the correct attitude.

Table A.6: All examined proposition queries for sexual and reproductive regulation policies. Italicized queries were cut before the analysis in §2.4.5.

Distributing foreign assistance
• *Public-private partnerships advance development.* – ended up capturing more matches on U.S.-based education and health sectors.
• Disaster relief and lifesaving assistance amidst complex crises.
• United States should provide humanitarian relief and international assistance to global communities in need.
• *United States should focus on domestic funding/interests.* – attempted negation.
• Foreign aid supports global stability.
• Foreign aid reduces global poverty and supports sustainable development and security.
• *Other nations should contribute more foreign aid.* – attempted negation.
• Foreign aid promotes global health.
• Foreign aid helps fight HIV/AIDS and malaria abroad.
• Foreign aid empowers women and girls.
• *Foreign aid empowers women and girls in developing nations.* – matches were not sufficiently restricted to foreign aid.
• Foreign aid promotes economic prosperity and resilience.
• Foreign aid boosts the economy of developing nations and alleviates poverty.
• *Foreign aid exacerbates poverty.* – attempted negation; in practice, this is unlikely to show up on Senate/House pages.
• *Foreign aid protects vulnerable populations.* – yielded too many false positives about vulnerable U.S. populations and other species

Table A.7: All examined proposition queries for foreign aid policy attitudes. Italicized queries were cut before the analysis in §2.4.5.

Protecting national security against terrorism
• Government should end legal loopholes and secure our borders.
• *Government should provide legal path to immigration.* – attempted negation; matched ideas that support legal immigration, but less so *expanding the path* to legal immigration.
• *Government should manage national cybersecurity risks.* – matches not particularly informative.
• *Government should target suspicious, malicious, or nefarious actors.* – matches not particularly informative.
• Government should safeguard the American people, our homeland, and our values.
• Government should provide asylum to immigrants in danger.
• *Terrorists threaten the U.S. borders and security.* – matches about border security, but not necessarily linked to terrorism.
• Terrorists attack the American people, our country, and our way of life.
• *Terrorists attack our way of life.* – captured other "way of life" phenomena.
• *Terrorists hate our American values.* – generally not great semantic matches.
• Government should identify potential terrorists and prevent attacks.
• Government should enforce immigration laws and secure U.S. borders.
• *Government should keep the American people safe.* – ambiguous; captures non-security based safety (e.g., food safety).
• Government should disrupt cartels, smugglers, nefarious actors, illegal border crossers.
• Government should rebuild our military.
• *Government already spends too much on the military.* – attempted negation.

Table A.8: All examined proposition queries for national security policies. Italicized queries were cut before the analysis in §2.4.5.

Appendix B

Chapter 3 Supplementary

B.1 N2O implementation details

In this section, we include additional implementation details for experiments performed in the paper. Generally, we use parameters consistent with the original work when possible.

Sentence segmentation. We use the `spacy`[1] library (2.0.16) to perform sentence segmentation; for word tokenization, we defer to preferences for the original embedder implementations if specified (see below), or use the `spacy` tokenizer otherwise.

Tf-idf. We use the `gensim` library (3.7.3) implementation of tf-idf,[2] with frequency statistics learned on the 2010 section of the Gigaword corpus (i.e., the same corpus used to find nearest neighbors). For tokenization, we use

[1] http://spacy.io
[2] https://radimrehurek.com/gensim/

the Gensim tokenizer and lowercase all word tokens.

Word2vec. We use pretrained 300D Google News embeddings available from Google.[3] We use `spacy` to perform word tokenization and embedding lookup.

GloVe. We use three sets of standard pretrained GloVe embeddings: 100D and 300D embeddings trained on Wikipedia and Gigaword (6B tokens), and 300D embeddings trained on Common Crawl (840B tokens).[4] We handle tokenization and embedding lookup identically to word2vec; for the Wikipedia/Gigaword embeddings, which are uncased, we lower case all tokens as well.

FastText. We use four sets of pretrained FastText embeddings: two trained on Wikipedia and other news corpora, and two trained on Common Crawl (each with an original version and one trained on subword information).[5] We use the Python port of the FastText implementation to handle tokenization, embedding lookup, and OOV embedding computation.[6]

ELMo. We use three pretrained models made available by AllenNLP: *small*, *original*, and *original (5.5B)*.[7] We use `spacy` to perform word tokenization, consistent with the `allennlp` library; we also use `allennlp` (0.7.2) to

[3] https://code.google.com/archive/p/word2vec/
[4] https://nlp.stanford.edu/projects/glove/
[5] https://fasttext.cc/docs/en/english-vectors.html
[6] https://github.com/facebookresearch/fastText/tree/master/python
[7] https://allennlp.org/elmo

compute the ELMo embeddings. We average the embeddings over all three bidirectional LSTM layers as suggested.

BERT. We use Hugging Face's `pytorch-transformers` (0.6.2) implementation and pretrained BERT base cased model.[8] To tokenize, we use the provided `BertTokenizer`, which handles WordPiece (subword) tokenization, and in general follow the library's recommendations for feature extraction.

For finetuning BERT on MultiNLI (matched subset), we use the default parameters provided in the library's `run_classifier.py` (batch size = 32, learning rate = 5e-5, etc.). We finetune for three epochs, and obtain 84.1% dev accuracy (reasonably consistent with the original work).

GPT. We use the same Hugging Face library and associated pretrained model for GPT; we use their BPE tokenizer and `spacy` for subword and word tokenization respectively.

InferSent. We use the authors' implementation of InferSent, as well as their pretrained V1 model based on GloVe.[9] (Unfortunately, the FastText-based V2 model was not available when performing these experiments; see issues #108 and #124 in the linked Github.) As per their README, we use the `nltk` tokenizer (3.2.5).

[8] Originally at `https://github.com/huggingface/pytorch-transformers`; now the `transformers` library (see GPT-2 notes).

[9] `https://github.com/facebookresearch/InferSent`

Universal Sentence Encoder. We use pretrained models available on TensorFlow Hub for both the DAN and Transformer variants.[10] The modules handle text preprocessing on their own.

GPT-2. For the newer models, we use an updated version of the Hugging Face `transformers` library (4.13.0)[11]; we use the pretrained GPT-2 model and associated tokenizer.[12]

RoBERTa. We use the same version of the `transformers` library as GPT-2 with the base RoBERTa model and tokenizer.[13]

SBERT. We use the authors' `sentence-transformers` library (2.1.0)[14] and pretrained `bert-base-nli-mean-tokens` model[15]; tokenization is handled by the library.

Computational details

Experiments for ELMo, BERT, GPT, and the Transformer version of USE were run on a NVIDIA Titan XP GPU with CUDA 9.2; later experiments with SBERT, RoBERTa, and GPT-2 were performed with the same GPU and CUDA 11.2. All other experiments were performed on CPUs.

[10] DAN: https://tfhub.dev/google/universal-sentence-encoder/2 Transformer: https://tfhub.dev/google/universal-sentence-encoder-large/3

[11] https://huggingface.co

[12] https://huggingface.co/gpt2

[13] https://huggingface.co/roberta-base

[14] https://www.sbert.net

[15] https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

B.2 Full results

The figure on the next page is a larger version of Fig. 3.3 that includes the actual N2O values.

B.3 Approximate nearest neighbors

As noted in §3.5, N2O computation is linear in the size of the corpus, and to have reasonable semantic overlap within a diverse set of sentences, the corpus should be large. The upfront cost of computing sentence embeddings across the corpus is unavoidable (and, for many applications, necessary anyways); our implementation of exact search is fast enough that repeated queries given precomputed embeddings is not a concern.

However, we note that approximate nearest neighbor (ANN) methods are also a viable option, where computation of building an index of the corpus is front-loaded to ensure sub-linear search time. We recommend use of a small held-out set of queries to tune the ANN method parameters towards higher precision/recall (vs. speed).

All of the results in this paper were obtained using exact (linear) search. However, we also performed preliminary experiments using the neighborhood graph tree (NGT) library, which achieves good recall in high-dimensional settings [Iwasaki and Miyazaki, 2018; Aumüller et al., 2019].[16] We were able to obtain similar N2O-ranked results (query recall ∼0.96) relatively quickly: 0.25–5 s./query (depending on embedding dimension).

[16] https://github.com/yahoojapan/NGT

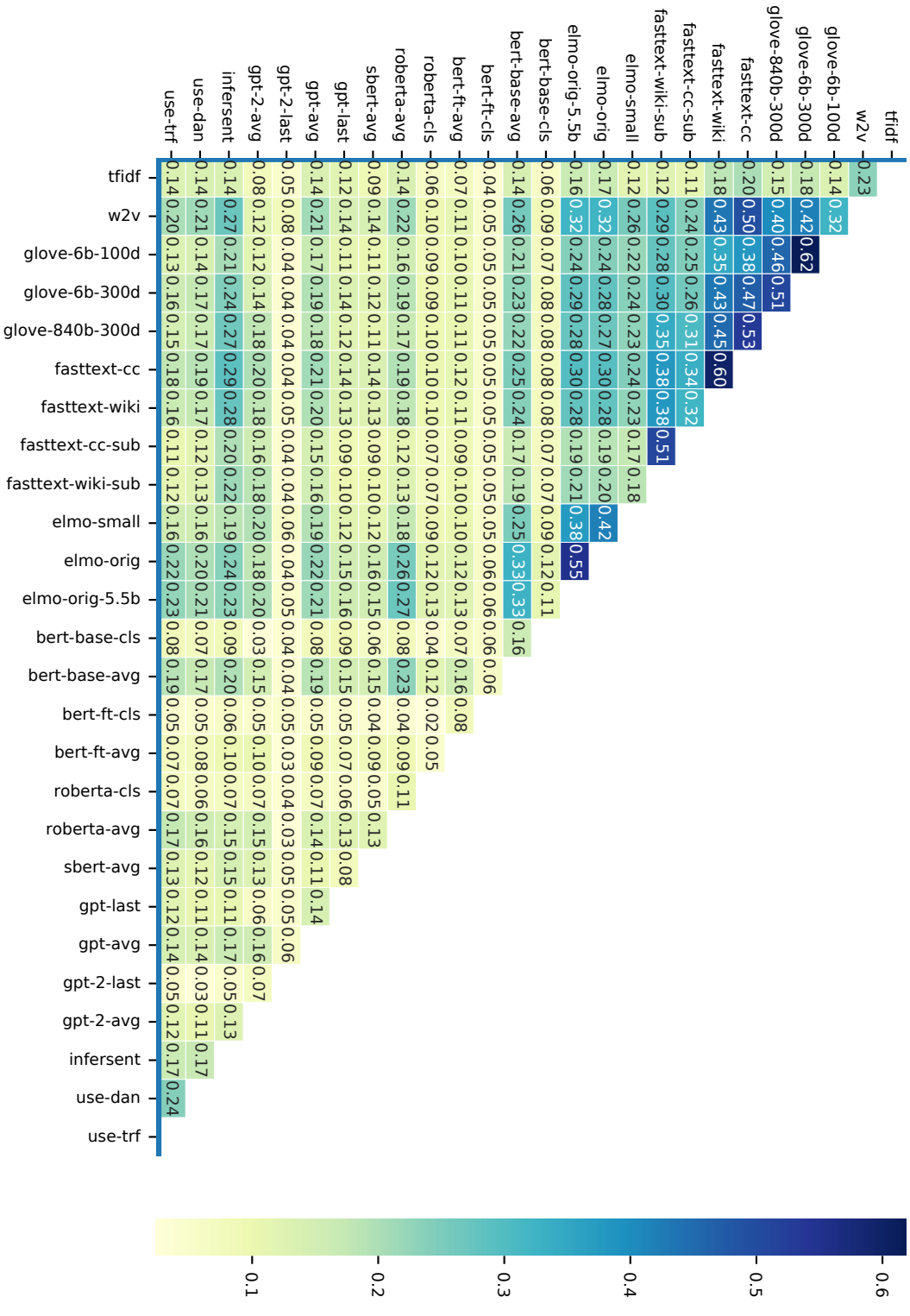


Figure B.1: N2O values between all pairs of sentence embedders.

We note that, in related work, ANN is commonly used in retrieval settings; e.g., Sugawara et al. [2016] test multiple ANN methods for similar *word* embedding search, and Bhagavatula et al. [2018] use an ANN method to index documents for citation recommendation. We believe that approximate methods can be of use for scalable N2O computation as well.