Learning Disentangled Representations of Texts with Application to Biomedical Abstracts

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Edward Banner
Northeastern University
ebanner@ccs.neu.edu

Jan-Willem van de Meent
Northeastern University
j.vandemeent@northeastern.edu

Iain J. Marshall
King's College London
iain.marshall@kcl.ac.uk

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

We propose a method for learning *disentangled* representations of texts that code for distinct and complementary aspects, with the aim of affording efficient model transfer and interpretability. To induce disentangled embeddings, we propose an adversarial objective based on the (dis)similarity between triplets of documents with respect to specific aspects. Our motivating application is embedding biomedical abstracts describing clinical trials in a manner that disentangles the *populations*, *interventions*, and *outcomes* in a given trial. We show that our method learns representations that encode these clinically salient aspects, and that these can be effectively used to perform aspect-specific retrieval. We demonstrate that the approach generalizes beyond our motivating application in experiments on two multi-aspect review corpora.

1 Introduction

A classic problem that arises in (distributed) representation learning is that it is difficult to determine what information individual dimensions in an embedding encode. When training a classifier to distinguish between images of people and landscapes, we do not know *a priori* whether the model is sensitive to differences in color, contrast, shapes or textures. Analogously, in the case of natural language, when we calculate similarities between document embeddings of user reviews, we cannot know if this similarity primarily reflects user sentiment, the product discussed, or syntactic patterns. This lack of interpretability makes it difficult to assess whether a learned representations is likely to generalize to a new task or domain, hindering model transferability. Disentangled representations with known semantics could allow more efficient training in settings in which supervision is expensive to obtain (e.g., biomedical NLP).

Thus far in NLP, learned distributed representations have, with few exceptions (Ruder et al., 2016; He et al., 2017; Zhang et al., 2017), been *entangled*: they indiscriminately encode all aspects of texts. Rather than representing text via a monolithic vector, we propose to estimate multiple embeddings that capture complementary aspects of texts, drawing inspiration from the ML in vision community (Whitney, 2016; Veit et al., 2017a).

As a motivating example we consider documents that describe clinical trials. Such publications constitute the evidence drawn upon to support *evidence-based medicine* (EBM), in which one formulates precise clinical questions with respect to the Populations, Interventions, Comparators and Outcomes (PICO elements) of interest (Sackett et al., 1996).[1] Ideally, learned representations of such articles would factorize into embeddings for the respective PICO elements. This would enable aspect-specific similarity measures, in turn facilitating retrieval of evidence concerning a given condition of interest (i.e., in a specific patient population), regardless of the interventions and outcomes considered. Better representations may reduce the amount of supervision needed, which is expensive in this domain.

Our work is one of the first efforts to induce disentangled representations of texts,[2] which we believe may be broadly useful in NLP. Concretely, our contributions in this paper are as follows:

- We formalize the problem of learning disentangled representations of texts, and develop a relatively general approach for learning these from aspect-specific similarity judgments expressed as triplets $(s, d, o)_a$, which indicate that document d is more similar to document s than to document o, with respect to aspect a.

[1] We collapse I and C because the distinction is arbitrary.

[2] We review the few recent related works that do exist in Section 5.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4683–4693
Brussels, Belgium, October 31 - November 4, 2018.

- We perform extensive experiments that provide evidence that our approach yields disentangled representations of texts, both for our motivating task of learning PICO-specific embeddings of biomedical abstracts, and, more generally, for multi-aspect sentiment corpora.

2 Framework and Models

Recent approaches in computer vision have emphasized unsupervised learning of disentangled representations by incorporating information-theoretic regularizers into the objective (Chen et al., 2016; Higgins et al., 2017). These approaches do not require explicit manual annotations, but consequently they require post-hoc manual assignment of meaningful interpretations to learned representations. We believe it is more natural to use weak supervision to induce meaningful aspect embeddings.

2.1 Learning from Aspect Triplets

As a general strategy for learning disentangled representations, we propose exploiting aspect-specific document triplets $(s, d, o)_a$: this signals that s and d are *more* similar than are d and o, with respect to aspect a (Karaletsos et al., 2015; Veit et al., 2017b), i.e., $\text{sim}_a(d, s) > \text{sim}_a(d, o)$, where sim_a quantifies similarity w.r.t. aspect a.

We associate with each aspect an encoder enc_a (encoders share low-level layer parameters; see Section 2.2 for architecture details). This is used to obtain text embeddings $(\mathbf{e}^a_s, \mathbf{e}^a_d, \mathbf{e}^a_o)$. To estimate the parameters of these encoders we adopt a simple objective that seeks to maximize the similarity between $(\mathbf{e}^a_d, \mathbf{e}^a_s)$ and minimize similarity between $(\mathbf{e}^a_d, \mathbf{e}^a_o)$, via the following maximum margin loss

$$\mathcal{L}(\mathbf{e}^a_s, \mathbf{e}^a_d, \mathbf{e}^a_o) = max\{0, 1 - sim(\mathbf{e}^a_d, \mathbf{e}^a_s) + sim(\mathbf{e}^a_d, \mathbf{e}^a_o)\} \quad (1)$$

Where similarity between documents i and j with respect to a particular aspect a, $sim_a(i, j)$, is simply the cosine similarity between the aspect embeddings $\mathbf{e}^a_i$ and $\mathbf{e}^a_j$. This allows for the same documents to be similar with respect to some aspects while dissimilar in terms of others.

The above setup depends on the correlation between aspects in the training data. At one extreme, when triplets enforce identical similarities for all aspects, the model cannot distinguish between aspects at all. At the other extreme, triplets are present for only one aspect a, and absent for all other aspects a': In this case the model will use only the embeddings for aspect a to represent similarities. In general, we expect a compromise between these extremes, and propose using negative sampling to enable the model to learn targeted aspect-specific encodings.

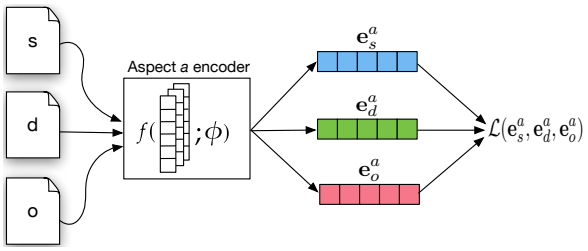

Figure 1: We propose associating aspects with encoders (low-level parameters are shared across aspects; this is not shown) and training these with triplets codifying aspect-wise relative similarities.

2.2 Encoder Architecture

Designing an aspect-based model requires specifying an encoder architecture. One consideration here is interpretability: a desirable property for aspect encoders is the ability to identify salient words for a given aspect. With this in mind, we propose using gated CNNs, which afford introspection via the token-wise gate activations.

Figure 2 schematizes our encoder architecture. The input is a sequence of word indices $d = (w_1, ..., w_N)$ which are mapped to m-dimensional word embeddings and stacked into a matrix $E = [\mathbf{e}_1, ..., \mathbf{e}_N]$. These are passed through sequential convolutional layers $C_1, ..., C_L$, which induce representations $H_l \in \mathbb{R}^{N \times k}$:

$$H_l = f_e(X * K_l + \mathbf{b}_l) \quad (2)$$

where $X \in \mathbb{R}^{N \times k}$ is the input to layer C_l (either a set of n-gram embeddings or H_{l-1}) and k is the number of feature maps. Kernel $K_l \in \mathbb{R}^{F \times k \times k}$ and $\mathbf{b}_l \in \mathbb{R}^k$ are parameters to be estimated, where F is the size of kernel window.[3] An activation function f_e is applied element-wise to the output of the convolution operations. We fix the size of $H_{l-1} \in \mathbb{R}^{N \times k}$ by zero-padding where necessary. Keeping the size of feature maps constant across layers allows us to introduce residual connections; the output of layer l is summed with the outputs of preceding layers before being passed forward.

We multiply the output of the last convolutional layer $H_L \in \mathbb{R}^{N \times k}$ with gates $\mathbf{g} \in \mathbb{R}^{N \times 1}$ to yield

[3] The input to C_1 is $E \in \mathbb{R}^{N \times m}$, thus $K_1 \in \mathbb{R}^{F \times m \times k}$.

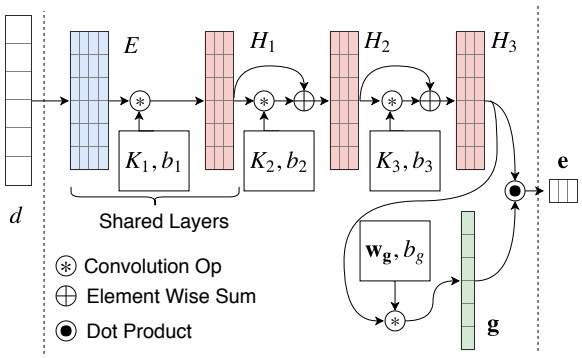

Figure 2: Schematic of our encoder architecture.

our final embedding $\mathbf{e}_d \in \mathbb{R}^{1\times k}$:

$$\begin{aligned} \mathbf{g} &= \sigma(H_L \cdot \mathbf{w}_g + b_g) \\ \mathbf{e}_d &= \mathbf{g}^T H_L \end{aligned} \tag{3}$$

where $\mathbf{w}_g \in \mathbb{R}^{k\times 1}$ and $b_g \in \mathbb{R}$ are learned parameters and σ is the sigmoid activation function. We impose a sparsity-inducing constraint on $\mathbf{g}$ via the ℓ_1 norm; this allows the gates to effectively serve as an attention mechanism over the input. Additionally, to capture potential cross-aspect correlation, weights in the embedding and first convolutional layers are shared between aspect encoders.

Alternative encoders. To assess the relative importance of the specific encoder model architecture used, we conduct experiments in which we fine-tune standard document representation models via triplet-based training. Specifically, we consider a single-layer MLP with BoW inputs, and a Neural Variational Document Model (NVDM) (Miao et al., 2016). For the NVDM we take a weighted sum of the original loss function and the triplet-loss over the learned embeddings, where the weight is a model hyperparameter.

3 Varieties of Supervision

Our approach entails learning from triplets that codify relative similarity judgments with respect to specific aspects. We consider two approaches to acquiring such triplets: the first exploits aspect-specific summaries written for texts, and the second assumes a more general scenario in which we solicit aspect-wise triplet judgments directly.

3.1 Deriving Triplets from Aspect Summaries

In the case of our motivating example – disentangled representations for articles describing clinical trials – we have obtained aspect-specific summaries from the *Cochrane Database of Systematic Reviews (CDSR)*. Cochrane is an international organization that creates and curates biomedical *systematic reviews*. Briefly, such reviews seek to formally synthesize all relevant articles to answer precise clinical questions, i.e., questions that specify a particular PICO frame. The CDSR consists of a set of reviews $\{R_i\}$. Reviews include multiple articles (studies) $\{S_{ij}\}$. Each study S consists of an abstract A and a set of free text summaries (s_P, s_I, s_O) written by reviewers describing the respective P, I and O elements in S.

Reviews implicitly specify PICO frames, and thus two studies in any given review may be viewed as equivalent with respect to their PICO aspects. We use this observation to derive document triplets. Recall that triplets for a given aspect include two comparatively similar texts (s, d) and one relatively dissimilar (o). Suppose the aspect of interest is the trial population. Here we match a given abstract (d) with its matched population summary from the CDSR (s); this encourages the encoder to yield similar embeddings for the abstract and the population description. The dissimilar o is constructed to distinguish the given abstract from (1) other aspect encodings (of interventions, outcomes), and, (2) abstracts for trials with different populations.

Concretely, to construct a triplet (s, d, o) for the PICO data, we draw two reviews R_1 and R_2 from the CDSR at random, and sample two studies from the first (s_1, s_1') and one from the second (s_2). Intuitively, s_2 will (very likely) comprise entirely different PICO elements than (s_1, s_1'), by virtue of belonging to a different review. To formalize the preceding description, our triplet is then: $(s = [s_1'^P], d = [s_1^{\text{abstract}}], o = [s_2^P | s_1'^I | s_1'^O])$, where s_1^{abstract} is the abstract for study s_1, and aspect summaries for studies are denoted by superscripts. We include a concrete example of triplet construction in the Appendix, Section D.

3.2 Learning Directly from Aspect-Wise Similarity Judgments

The preceding setup assumes a somewhat unique case in which we have access to aspect-specific summaries written for texts. As a more general setting, we also consider learning directly from triplet-wise supervision concerning relative similarity with respect to particular aspects (Amid and Ukkonen, 2015; Veit et al., 2017a; Wilber et al., 2014). The assumption is that such judgments can

be solicited directly from annotators, and thus the approach may be applied to arbitrary domains, so long as meaningful aspects can be defined implicitly via pairwise similarities regarding them.

We do not currently have corpora with such judgments in NLP, so we constructed two datasets using aspect-specific sentiment ratings. Note that this highlights the flexibility of exploiting aspect-wise triplet supervision as a means of learning disentangled representations: existing annotations can often be repurposed into such triplets.

4 Datasets and Experiments

We present a series of experiments on three corpora to assess the degree to which the learned representations are disentangled, and to evaluate the utility of these embeddings in simple downstream retrieval tasks. We are particularly interested in the ability to identify documents similar w.r.t. a target aspect. All parameter settings for baselines are reported in the Appendix (along with additional experimental results). The code is available at `https://github.com/successar/neural-nlp`.

4.1 PICO (EBM) Domain

We first evaluate embeddings quantitatively with respect to retrieval performance. In particular, we assess whether the induced representations afford improved retrieval of abstracts relevant to a particular systematic review (Cohen et al., 2006; Wallace et al., 2010). We then perform two evaluations that explicitly assess the degree of disentanglement realized by the learned embeddings.

The PICO dataset comprises 41K abstracts of articles describing clinical trials extracted from the CDSR. Each abstract is associated with a review and three summaries, one per aspect (P/I/O). We keep all words that occur in ≥ 5 documents, converting all others to `unk`. We truncate documents to a fixed length (set to the 95th percentile).

4.1.1 Quantitative Evaluation

Baselines. We compare the proposed **P**, **I** and **O** embeddings and their concatenation $[\mathbf{P}|\mathbf{I}|\mathbf{O}]$ to the following. **TF-IDF**: standard TF-IDF representation of abstracts. **RR-TF**: concatenated TF-IDF vectors of sentences predicted to describe the respective PICO elements, i.e., sentence predictions made using the pre-trained model from (Wallace et al., 2016) — this model was trained using distant supervision derived from the CDSR. **doc2vec**: standard (entangled) distributed representations of abstracts (Le and Mikolov, 2014). **LDA**: Latent Dirichlet Allocation. **NVDM**: A generative model of text where the representation is a vector of log-frequencies that encode a topic (Miao et al., 2016). **ABAE**: An autoencoder model that discovers latent aspects in sentences (He et al., 2017). We obtain document embeddings by summing over constituent sentence embeddings. **DSSM**: A CNN based encoder trained with triplet loss over abstracts (Shen et al., 2014).

Hyperparameters and Settings. We use three layers for our CNN-based encoder (with 200 filters in each layer; window size of 5) and the PReLU activation function (He et al., 2015) as f_e. We use 200d word embeddings, initialized via pretraining over a corpus of PubMed abstracts (Pyysalo et al., 2013). We used the Adam optimization function with default parameters (Kingma and Ba, 2014). We imposed ℓ_2 regularization over all parameters, the value of which was selected from the range ($1e$-2, $1e$-6) as $1e$-5. The ℓ_1 regularization parameter for gates was chosen from the range ($1e$-2, $1e$-8) as $1e$-6. All model hyperparameters for our models and baselines were chosen via line search over a 10% validation set.

Metric. For this evaluation, we used a held out set of 15 systematic reviews (comprising 2,223 studies) compiled by Cohen et al. (2006). The idea is that good representations should map abstracts in the same review (which describe studies with the same PICO frame) relatively near to one another. To compute AUCs over reviews, we first calculate all pairwise study similarities (i.e., over all studies in the Cohen corpus). We can then construct an ROC for a given abstract a from a particular review to calculate its AUC: this measures the probability that a study drawn from the same review will be nearer to a than a study from a different review. A summary AUC for a review is taken as the mean of the study AUCs in that review.

Results. Table 1 reports the mean AUCs over individual reviews in the Cohen et al. (2006) corpus, and grand means over these (bottom row). In brief: The proposed PICO embeddings (concatenated) obtain an equivalent or higher AUC than baseline strategies on 12/14 reviews, and strictly higher AUCs in 11/14. It is unsurprising that we outperform unsupervised approaches, but we also best RR-TF, which was trained with the same CDSR corpus (Wallace et al., 2016), and DSSM (Shen et al., 2014), which exploits the same triplet loss

Study	TF-IDF	Doc2Vec	LDA	NVDM	ABAE	RR-TF	DSSM	**P**	**I**	**O**	**[P|I|O]**
ACEInhib.	0.81	0.74	0.72	0.85	0.81	0.67	0.85	0.83	0.88	0.84	**0.92**
ADHD	0.90	0.82	0.83	**0.93**	0.77	0.85	0.83	0.86	0.75	0.91	0.89
Antihist.	0.81	0.73	0.67	0.79	0.84	0.72	**0.91**	0.88	0.84	0.89	**0.91**
Antipsych.	0.75	0.85	0.88	0.89	0.81	0.63	0.96	0.91	0.93	**0.97**	**0.97**
BetaBlockers	0.67	0.65	0.61	0.76	0.70	0.56	0.68	0.71	0.75	0.77	**0.81**
CCBlockers	0.67	0.60	0.67	0.70	0.69	0.58	0.76	0.73	0.69	0.74	**0.77**
Estrogens	0.87	0.85	0.60	0.94	0.85	0.82	0.96	**1.00**	0.98	0.83	**1.00**
NSAIDS	0.85	0.77	0.73	0.9	0.77	0.74	0.89	0.94	**0.95**	0.8	**0.95**
Opioids	0.81	0.75	0.80	0.83	0.77	0.76	0.86	0.80	0.83	**0.92**	**0.92**
OHG	0.79	0.80	0.70	0.89	0.90	0.72	0.90	0.90	0.95	0.95	**0.96**
PPI	0.81	0.79	0.74	0.85	0.82	0.68	0.94	0.94	0.87	0.87	**0.95**
MuscleRelax.	0.60	0.67	0.74	0.75	0.61	0.57	0.77	0.68	0.62	**0.78**	0.75
Statins	0.79	0.76	0.66	0.87	0.77	0.68	0.87	0.82	**0.94**	0.87	**0.94**
Triptans	0.92	0.82	0.83	0.92	0.75	0.81	**0.97**	0.93	0.79	**0.97**	**0.97**
Mean	0.79	0.76	0.73	0.85	0.78	0.70	0.87	0.85	0.84	0.87	**0.91**

Table 1: AUCs achieved using different representations on the Cohen et al. corpus. Models to the right of the | are supervised; those to the right of || constitute the proposed disentangled embeddings.

in a clinical trial mainly involving patients over qqq with coronary heart disease , ramipril reduced mortality while vitamin e had no preventive effect .
in a clinical trial mainly involving patients over qqq with coronary heart disease , ramipril reduced mortality while vitamin e had no preventive effect .
in a clinical trial mainly involving patients over qqq with coronary heart disease , ramipril reduced mortality while vitamin e had no preventive effect .

Table 2: Gate activations for each aspect in a PICO abstract. Note that because gates are calculated at the final convolution layer, activations are not in exact 1-1 correspondence with words.

as our model. We outperform the latter by an average performance gain of 4 points AUC (significant at 95% level using independent 2-sample t-test).

We now turn to the more important questions: are the learned representations actually disentangled, and do they encode the target aspects? Table 2 shows aspect-wise gate activations for PICO elements over a single abstract; this qualitatively suggests disentanglement, but we next investigate this in greater detail.

4.1.2 Qualitative Evaluation

To assess the degree to which our PICO embeddings are disentangled – i.e., capture complementary information relevant to the targeted aspects – we performed two qualitative studies.

First, we assembled 87 articles (not seen in training) describing clinical trials from a review on the effectiveness of decision aids (Stacey et al., 2014) for: women with, at risk for, and genetically at risk for, breast cancer (BCt, BCs and BCg, respectively); type II diabetes (D); menopausal women (MW); pregnant women generally (PW) and those who have undergone a C-section previously (PWc); people at risk for colon cancer (CC); men with and at risk of prostate cancer (PCt and PCs, respectively) and individuals with atrial fibrillation (AF). This review is unusual in that it *studies a single intervention (decision aids) across different populations*. Thus, if the model is successful in learning disentangled representations, the corresponding P vectors should roughly cluster, while the I/C should not.

Figure 3 shows a TSNE-reduced plot of the P, I/C and O embeddings induced by our model for these studies. Abstracts are color-coded to indicate the populations enumerated above. As hypothesized, P embeddings realize the clearest separation with respect to the populations, while the I and O embeddings of studies do not co-localize to the same degree. This is reflected quantitatively in the AUC values achieved using each aspect embedding (listed on the Figure). This result implies disentanglement along the desired axes.

Next we assembled 50 abstracts describing trials involving *hip replacement arthroplasty* (HipRepl). We selected this topic because HipRepl will either describe the trial population (i.e., patients who have received hip replacements) or it will be the intervention, but not both. Thus, we would expect that abstracts describing trials in which HipRepl describes the population cluster in the corresponding embedding space, but not in the intervention space (and vice-versa). To test this, we first manually annotated the 50 abstracts, associating HipRepl with either P or I. We used these labels to calculate pairwise AUCs, reported in Table 3. The results imply that the population embeddings discriminate between studies that

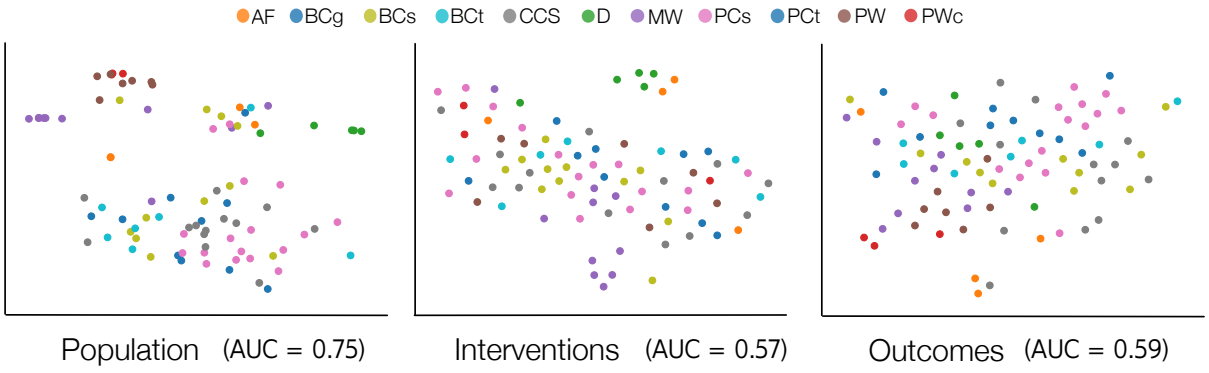

Figure 3: TSNE-reduced scatter of disentangled PICO embeddings of abstracts involving "decision aid" interventions. Abstracts are colored by known population group (see legend). Population embeddings for studies in the same group co-localize, more so than in the intervention and outcome space.

	HipRepl I	HipRepl P	Mean
Population	0.62	**0.68**	**0.66**
Intervention	**0.91**	0.46	0.57
Outcome	0.89	0.42	0.54

Table 3: AUCs realized over HipRepl studies using different embeddings. Column: Study label (HipRepl as P or I). Row: Aspect embedding used.

Population *american*, *area*, *breast*, *colorectal*, *diagnosis*, *inpatients*, *outpatients*, *stage*, *their*, *uk*
Intervention *adjunct*, *alone*, *an*, *discussion*, *intervention*, *methods*, *reduced*, *started*, *took*, *written*
Outcome *adults*, *area*, *either*, *eligible*, *importance*, *improve*, *mortality*, *pre*, *reduces*, *survival*

Table 4: Top ten most activated words, as determined by the gating mechanism.

enrolled patients with HipRepl and other studies. Likewise, studies in which HipRepl was the intervention are grouped in the interventions embedding space, but not in the populations space.

Aspect words. In Table 4, we report the most activated unigrams for each aspect embedding on the decision aids corpus. To derive these we use the outputs of the gating mechanism (Eq. 3), which is applied to all words in the input text. For each word, we average the activations across all abstracts and find the top ten words for each aspect. The words align nicely with the PICO aspects, providing further evidence that our model learns to focus on aspect-specific information.

4.2 Multi-Aspect Reviews

We now turn from the specialized domain of biomedical abstracts to more general applications. In particular, we consider learning disentangled representations of beer, hotel and restaurant reviews. Learned embeddings should capture different aspects, e.g., taste or look in the case of beer.

Baseline	Look	Aroma	Palate	Taste
TF-IDF	0.63	0.62	0.62	0.61
LDA	0.73	0.73	0.73	0.73
Doc2Vec	0.61	0.61	0.61	0.61
NVDM	0.68	0.69	0.69	0.70
ABAE	0.50	0.50	0.50	0.50
BoW + Triplet	0.85	0.90	0.90	0.92
NVDM + Triplet	0.90	0.91	0.92	0.95
DSSM + Triplet	0.87	0.90	0.90	0.92
CNN + Triplet	**0.92**	**0.93**	**0.94**	**0.96**

Table 5: AUC results for different representations on the BeerAdvocate data. Models beneath the second line are supervised.

	Look	Aroma	Palate	Taste
Look	0.92	0.89	0.88	0.87
Aroma	0.90	0.93	0.91	0.92
Palate	0.89	0.92	0.94	0.95
Taste	0.90	0.94	0.95	0.96

Table 6: Cross AUC results for different representations on the BeerAdvocate data. Row: Embedding used. Column: Aspect evaluated against.

4.2.1 Beer Reviews (BeerAdvocate)

We conducted experiments on the BeerAdvocate dataset (McAuley et al., 2012), which contains 1.5M reviews of beers that address four aspects: *appearance*, *aroma*, *palate*, and *taste*. Free-text reviews are associated with aspect-specific numerical ratings for each of these, ranging from 1 to 5. We consider ratings < 3 as negative, and > 3 as positive, and use these to generate triplets of reviews. For each aspect a, we construct triplets $(s, d, o)_a$ by first randomly sampling a review d. We then select s to be a review with the same sen-

	Look		Aroma		Palate		Taste	
Look	-	-	0.42	0.60	0.40	0.63	0.38	0.65
Aroma	0.33	0.69	-	-	0.41	0.59	0.41	0.60
Palate	0.32	0.70	0.46	0.54	-	-	0.49	0.52
Taste	0.23	0.80	0.35	0.66	0.33	0.67	-	-

Table 7: 'Decorrelated' cross-AUC results on the BeerAdvocate data, which attempt to mitigate confounding due to overall sentiment being correlated. Each cell reports metrics over subsets of reviews in which the sentiment differs between the row and column aspects. The numbers in each cell are the AUCs w.r.t. sentiment regarding the column aspect achieved using the row and column aspect representations, respectively.

timent with respect to a as d, and o to be a review with the opposite sentiment regarding a. We selected 90K reviews for experiments, such that we had an equal number of positive and negative reviews for each aspect. We only keep words appearing in at least 5 documents, converting all others to `unk`. We truncated reviews to 95 percentile length. We split our data into 80/10/10 ratio for training, validation and testing, respectively.

Baselines. We used the same baselines as for the PICO domain, save for *RR-TF*, which was domain-specific. Here we also evaluate the result of replacing the CNN-based encoder with NVDM, BoW and DSSM based encoders, respectively, each trained using triplet loss.

Hyperparameters and Settings. For the CNN-based encoder, we used settings and hyperparameters as described for the PICO domain. For the BoW encoder, we used 800d output embeddings and a PReLU activation function with ℓ_2 regularization set to $1e$-5. For the NVDM based encoder, we used 200d embeddings.

Metrics. We again performed an IR-type evaluation to assess the utility of representations. For each aspect k, we constructed an affinity matrix A^k such that $A^k_{ij} = sim_k(r_i, r_j)$ for beer reviews r_i and r_j. We consider two reviews similar under a given aspect k if they have the same (dichotomized) sentiment value for said aspect. We compute AUCs for each review and aspect using the affinity matrix A_k. The AUC values are averaged over reviews in the test set to obtain a final AUC metric for each aspect. We also report cross AUC measures in which we use embeddings for aspect k to distinguish reviews under aspect k'.

Results We report the AUC measures for each aspect on our test set using different representations in Table 5. Our model consistently outperforms baseline strategies over all aspects. Unsurprisingly, the model outperforms unsupervised approaches.[4] We realize consistent though modest improvement over triplet-supervised approaches that use alternative encoders.

In Table 6 we present cross AUC evaluations. Rows correspond to the embedding used and columns to the aspect evaluated against. As expected, aspect-embeddings perform better w.r.t. the aspects for which they code, suggesting some disentanglement. However, the reduction in performance when using one aspect representation to discriminate w.r.t. others is not as pronounced as above. This is because aspect ratings are highly correlated: if taste is positive, aroma is very likely to be as well. Effectively, here sentiment entangles all of these aspects.[5]

In Table 7, we evaluate cross AUC performance for beer by first 'decorrelating' the aspects. Specifically, for each cell (k, k') in the table, we first retrieve the subset of reviews in which the sentiment w.r.t. k differs from the sentiment w.r.t. k'. Then we evaluate the AUC similarity of these reviews on the basis of sentiment concerning k' using both k and k' embeddings, yielding a pair of AUCs (listed respectively). We observe that the using k' embeddings to evaluate aspect k' similarity yields better results than using k embeddings.

We present the most activated words for each aspect (as per the gating mechanism) in Table 8. And we present an illustrative review color-coded with aspect-wise gate activations in Table 9. For completeness, we reproduce the top words for aspects discovered using He et al. (2017) in the Appendix; these do not obviously align with the target aspects, which is unsurprising given that this is an unsupervised method.

4.2.2 Hotel & Restaurant Reviews

Finally, we attempt to learn embeddings that disentangle domain from sentiment in reviews. For this we use a combination of TripAdvisor and

[4] We are not sure why ABAE (He et al., 2017) performs so poorly on the review corpora. It may simply fail to prominently encode sentiment, which is important for these tasks. We note that this model performs reasonably well on the PICO data above, and qualitatively seems to recover reasonable aspects (though not specifically sentiment).

[5] Another view is that we are in fact inducing representations of <aspect, sentiment> pairs, and only the aspect varies across these; thus representations remain discriminative (w.r.t. sentiment) across aspects.

Look	*attractive*, *beautiful*, *fingers*, *pumpkin*, *quarter*, *received*, *retention*, *sheets*, *sipper*, *well-balanced*
Aroma	*beer*, *cardboard*, *cheap*, *down*, *follows*, *medium-light*, *rice*, *settled*, *skunked*, *skunky*
Palate	*bother*, *crafted*, *luscious*, *mellow*, *mint*, *range*, *recommended*, *roasted*, *tasting*, *weight*
Taste	*amazingly*, *down*, *highly*, *product*, *recommended*, *tasted*, *thoroughly*, *to*, *truly*, *wow*

Table 8: Most activated words for aspects on the beer corpus, as per the gating mechanism.

Yelp! ratings data. The former comprises reviews of hotels, the latter of restaurants; both use a scale of 1 to 5. We convert ratings into positive/negative labels as above. Here we consider aspects to be the domain (hotel or restaurant) and the sentiment (positive or negative). We aim to generate embeddings that capture information about only one of these aspects. We use 50K reviews from each dataset for training and 5K for testing.

Baselines. We use the same baselines as for the BeerAdvocate data, and similarly use different encoder models trained under triplet loss.

Evaluation Metrics. We perform AUC and cross-AUC evaluation as in the preceding section. For the domain aspect, we consider two reviews similar if they are from the same domain, irrespective of sentiment. Similarly, reviews are considered similar with respect to the sentiment aspect if they share a sentiment value, regardless of domain.

Results. In Table 10 we report the AUCs for each aspect on our test set using different representations. Baselines perform reasonably well on the domain aspect because reviews from different domains are quite dissimilar. Capturing sentiment information irrespective of domain is more difficult, and most unsupervised models fail in this respect. In Table 11, we observe that cross AUC results are much more pronounced than for the BeerAdvocate data, as the domain and sentiment are uncorrelated (i.e., sentiment is independent of domain).

5 Related Work

Work in representation learning for NLP has largely focused on improving word embeddings (Levy and Goldberg, 2014; Faruqui et al., 2015; Huang et al., 2012). But efforts have also been made to embed other textual units, e.g. characters (Kim et al., 2016), and lengthier texts including sentences, paragraphs, and documents (Le and Mikolov, 2014; Kiros et al., 2015).

Triplet-based judgments have been used in multiple domains, including vision and NLP, to estimate similarity information implicitly. For example, triplet-based similarity embeddings may be learned using 'crowdkernels' with applications to multi-view clustering (Amid and Ukkonen, 2015). Models combining similarity with neural networks mainly revolve around Siamese networks (Chopra et al., 2005) which use pairwise distances to learn embeddings (Schroff et al., 2015), a tactic we have followed here. Similarity judgments have also been used to generate document embeddings for IR tasks (Shen et al., 2014; Das et al., 2016).

Recently, He et al. (2017) introduced a neural model for aspect extraction that relies on an attention mechanism to identify aspect words. They proposed an autoencoder variant designed to tease apart aspects. In contrast to the method we propose, their approach is unsupervised; discovered aspects may thus not have a clear interpretation. Experiments reported here support this hypothesis, and we provide additional results using their model in the Appendix.

Other recent work has focused on text *generation* from factorized representations (Larsson et al., 2017). And Zhang et al. (2017) proposed a lightly supervised method for *domain adaptation* using aspect-augmented neural networks. They exploited *source* document labels to train a classifier for a *target* aspect. They leveraged sentence-level scores codifying sentence relevance w.r.t. individual aspects, which were derived from terms *a priori* associated with aspects. This supervision is used to construct a composite loss that captures both classification performance on the source task and a term that enforces *invariance* between source and target representations.

There is also a large body of work that uses probabilistic generative models to recover latent structure in texts. Many of these models derive from Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and some variants have explicitly represented topics and aspects jointly for sentiment tasks (Brody and Elhadad, 2010; Sauper et al., 2010, 2011; Mukherjee and Liu, 2012; Sauper and Barzilay, 2013; Kim et al., 2013).

A bit more generally, *aspects* have also been interpreted as properties spanning entire texts, e.g., a perspective or theme which may then color the discussion of topics (Paul and Girju, 2010). This intuition led to the development of the *factorial LDA* family of topic models (Paul and Dredze, 2012; Wallace et al., 2014); these model individ-

Look : deep amber hue , this brew is topped with a finger of off white head . smell of dog unk , green unk , and slightly fruity . taste of belgian yeast , coriander , hard water and bready malt . light body , with little carbonation .	**Aroma** : deep amber hue , this brew is topped with a finger of off white head . smell of dog unk , green unk , and slightly fruity . taste of belgian yeast , coriander , hard water and bready malt . light body , with little carbonation .	**Palate** : deep amber hue , this brew is topped with a finger of off white head . smell of dog unk , green unk , and slightly fruity . taste of belgian yeast , coriander , hard water and bready malt . light body , with little carbonation .	**Taste** :deep amber hue , this brew is topped with a finger of off white head . smell of dog unk , green unk , and slightly fruity . taste of belgian yeast , coriander , hard water and bready malt . light body , with little carbonation .

Table 9: Gate activations for each aspect in an example beer review.

Baseline	Domain	Sentiment
TF-IDF	0.59	0.52
Doc2Vec	0.83	0.56
LDA	0.90	0.62
NVDM	0.79	0.63
ABAE	0.50	0.50
BoW + Triplet	**0.99**	0.91
NVDM + Triplet	**0.99**	0.91
DSSM + Triplet	**0.99**	0.90
CNN + Triplet	**0.99**	**0.92**

Table 10: AUC results for different representations on the Yelp!/TripAdvisor Data. Models beneath the second line are supervised.

Baseline	Domain	Sentiment
Domain	0.988	0.512
Sentiment	0.510	0.917

Table 11: Cross AUC results for different representations for Yelp!/TripAdvisor Dataset.

ual word probability as a product of multiple latent factors characterizing a text. This is similar to the Sparse Additive Generative (SAGE) model of text proposed by Eisenstein et al. (2011).

6 Conclusions

We have proposed an approach for inducing disentangled representations of text. To learn such representations we have relied on supervision codified in aspect-wise similarity judgments expressed as document triplets. This provides a general supervision framework and objective. We evaluated this approach on three datasets, each with different aspects. Our experimental results demonstrate that this approach indeed induces aspect-specific embeddings that are qualitatively interpretable and achieve superior performance on information retrieval tasks.

Going forward, disentangled representations may afford additional advantages in NLP, e.g., by facilitating transfer (Zhang et al., 2017), or supporting aspect-focused summarization models.

7 Acknowledgements

This work was supported in part by National Library of Medicine (NLM) of the National Institutes of Health (NIH) award R01LM012086, by the Army Research Office (ARO) award W911NF1810328, and research funds from Northeastern University. The content is solely the responsibility of the authors.

References

Ehsan Amid and Antti Ukkonen. 2015. Multiview triplet embedding: Learning attributes in multiple maps. In *International Conference on Machine Learning*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan).

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR 2005. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.*

Aaron M Cohen, William R Hersh, K Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2).

Arpita Das, Harish Yenala, Manoj Chinnakotla, and Manish Shrivastava. 2016. Together we stand: Siamese networks for similar question retrieval. In

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse Additive Generative Models of Text. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An Unsupervised Neural Attention Model for Aspect Extraction. In *Proceedings of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. Beta-Vae: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations*.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. 2015. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*.

Suin Kim, Jianwen Zhang, Zheng Chen, Alice H Oh, and Shixia Liu. 2013. A Hierarchical Aspect-Sentiment Model for Online Reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*.

Maria Larsson, Amanda Nilsson, and Mikael Kågebäck. 2017. Disentangled Representations for Manipulation of Sentiment in Text. *arXiv preprint arXiv:1712.10066*.

Quoc V Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In *Proceedings of the Association of Computational Linguistics (ACL)*.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining (ICDM)*.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International Conference on Machine Learning*.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics.

Michael Paul and Mark Dredze. 2012. Factorial LDA: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems*.

Michael Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *Bmj*, 312(7023).

Christina Sauper and Regina Barzilay. 2013. Automatic aggregation by joint modeling of aspects and values. *Journal of Artificial Intelligence Research*.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2010. Incorporating content structure into text analysis applications. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2011. Content models with attitude. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*.

D Stacey, F Légaré, N F Col, C L Bennett, M J Barry, K B Eden, H Thomas, A Lyddiatt, R Thomson, L Trevena, and J H C Wu. 2014. Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Reviews*, 1.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. 2017a. Conditional Similarity Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. 2017b. Conditional Similarity Networks. In *Computer Vision and Pattern Recognition (CVPR)*.

Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Brian Zhu, and Iain J Marshall. 2016. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research*, 17(132).

Byron C Wallace, Michael J Paul, Urmimala Sarkar, Thomas A Trikalinos, and Mark Dredze. 2014. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association*, 21(6).

Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1).

William Whitney. 2016. Disentangled Representations in Neural Models. *arXiv preprint arXiv:1602.02383*.

Michael J Wilber, Iljung S Kwak, and Serge J Belongie. 2014. Cost-effective hits for relative similarity comparisons. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2017. Aspect-augmented Adversarial Networks for Domain Adaptation. In *Transactions of the Association for Computational Linguistics*.