Spatial Denoising and Variational Bayesian Hierarchical Models for the Extraction of Commonsense Attributes from Text Corpora with Applications in NLP

Mauricio Tec

The University of Texas at Austin, Department of Statistics and Data Science

Introduction. The focus of this proposal is to develop novel strategies for extracting *commonsense knowledge* (CSK) of *quantitative* attributes from vast collections of text corpora. More precisely, for any given object, e.g. a *baseball bat*, we seek to extract the probability distribution of typical values associated with one or more of its attributes, e.g. *length* and *mass*. Knowing the distributions of CSK attributes can improve NLP systems in many ways; most obviously, they can support Question Answering (Tandon et al., 2014) and Textual Entailment (Dagan et al., 2013). More generally, such distributions can be regarded as the building blocks for solving complex tasks in a transparent and generalizable way. In this regard, Goodman et al. (2014) assert that human knowledge can be expressed as a taxonomy of simple concepts encoding probabilistic views, which allows reasoning to be extended to new situations. Thus leveraging CSK in algorithm design can lead to more robust and generalizable AI. We remark that existing collaborative databases with CSK such as Wikidata contain only a fraction of the potential information of interest on the web (Elazar et al., 2019), and they may lack data on many attributes of interest. Moreover, a pipeline for reliably extracting knowledge from task-specific corpora (e.g. technical manuals) can be very helpful in specialized language-assisted learning tasks (Luketina et al., 2019).

In what follows, we will discuss the three main challenges that we must confront in improving upon existing state-of-the-art methods in extracting CSK from text corpora. We will then outline a promising research direction, based on spatial denoising and scalable Bayesian modeling, that aims to address these challenges. A relevant and direct application of our proposed methods will be to enable a framework for the automatic construction of high-quality quantitative attributes in knowledge bases that is robust to noise in web text corpora.

Challenges. The first challenge arises from the inherent noisiness of the CSK data collection process. Typically, the process begins with identifying sentences with statements about the values of a given attribute for a given object; then, the values are standardized and aggregated to create an empirical distribution for the object–attribute tuple. This extraction process can be highly noisy, and has multiple sources of error, such as incorrect units, misplaced attribution, polysemy, and reporting bias. To illustrate, we examine the case of Distributions over Quantities (DoQ) (Elazar et al., 2019), a state-of-the-art resource containing 350k textual objects with information on 10 physical attributes extracted from web text corpora. Table 1a shows examples of extracted attribute statements, including both correct and incorrect statements (e.g. a peanut weighing 828,000 tons). The authors of DoQ partially address the erroneous collection issues via simple denoising filters. Yet inspection of the output shows the presence of extremely noisy values, as shown in table 1b, where DoQ infers that the average *dinner table* is of Valhalla proportions, at nearly 60 kilometers long. More sophisticated denoising filters than the ones used in DoQ have been attempted, e.g. (Bagherinezhad et al., 2016); but they are computationally prohibitive for large datasets. It is worth mentioning that the authors of DoQ did not include in their final data set any objects with fewer than 1000 observations due to privacy concerns. Still, we can hypothesize that including more terms would have resulted in more *coverage* of real-world objects, but less *quality*, since the inferred distributions are already noisy even with 1000 observations.

The second challenge in learning attributes of objects is how to pool or borrow information across semantically similar objects. For example, in Table 1b, the inferred length and area attributes for *coffee table* and *kitchen table* are sensible, while those for *dinner table* are highly noisy. Yet these three terms are all semantically similar; we expect that they would be close to each other in Euclidean distance in most vector-space representations, e.g. *word2vec* (Mikolov et al., 2013). The intuition is that semantically similar objects are more likely to have similar distributions over quantitative attributes than are two randomly chosen objects. Said concisely: quantitative attributes should be spatially correlated in a suitably chosen, semantically meaningful vector space. This intuition can be encoded in the form of a *spatial prior* or regularizer (Besag, 1974), and the challenge we address below is how to do so in a way that improves overall performance at CSK extraction.

The third challenge is how to take into account the correlation structure among different quantitative attributes. This correlation can potentially be a major source of information in CSK extraction. We expect, for example,

Dimension	*Correct measurement*	*Wrong measurement*
`time`	On 24 September drivers struck between *7.30 am* and *8.30 am*, the middle of the morning **rush hour**.	The largest **aftershock** measured 5.6 and struck *nine minutes* later at 1.51 pm. (*Aftershock wrongly labeled as 9 minutes long.*)
`area`	With over *36,000 acres*, **Smoky Hill** is the Air National Guard's largest weapons range.	The **plant** is spread over *25 acres* of land. (*Plant is a factory but wrongly identified as an organism.)*
`mass`	The **men**'s *73 kg* competition of the 2014 World Judo Championships was held on 27 August.	In recent years, the reported average annual **peanut** production lies around *828,000 tons* (95% for oil). (*Peanut labeled as weighing 828,000 tons.*)

(a) Examples of *correct* and *wrong* inferences from parsed sentences.

Dimension	*Area (m2)*	*Length (m)*	*Mass (kg)*	*Volume (lt)*
coffee table	27.24	0.68	22.79	43.85
dinner table	74,114.27	59,188.71	14.70	0.97
kitchen table	170.42	1.64	10.76	2.49
wooden table	`NA`	1.1	11.48	0.62

(b) Examples of the means of the distributions of table-like objects. We observe extreme outliers in the *dinner table* row.

Table 1: (a) shows examples of the information extraction process in DoQ (Elazar et al., 2019, Supp.); (b) exemplifies the presence of outliers, the reported values are the distribution means taken from DoQ's output[1].

that *length*, *mass*, and *volume* of an object will be positively correlated; but that *time* and *speed* of a car journey will be negatively correlated. Having knowledge of this correlation at our disposal would allow us to reason, for example, that a 60-kilometer long, 15-kilogram dinner table is a huge *multivariate* outlier, since it has an extremely unlikely combination of these two physical attributes (in this case, 1% of the mass per unit length of tissue paper). The challenge is how to learn and leverage these kinds of correlation structures at scale. But existing state-of-the-art methods, including DoQ, simply parse each attribution independently, ignoring correlation.

Proposed research. To summarize, we identify three major limitations in current approaches for learning quantitative attributes of objects:

1. Their inferred distributions are inherently noisy, and they do not incorporate a statistical model for noise.
2. They do not exploit semantic or physical similarities between objects.
3. They ignore correlations among attributes.

We will now explain our proposed research direction for addressing these limitations.

Step 1: Robust Graph Smoothing. We first propose to address limitations 1 and 2 using denoising and spatial (semantic) regularization. There are four underlying assumptions of this approach, in which we treat space as a discrete for reasons of computational scalability:

- Every object (e.g. *dinner table*) is a node v in a graph.
- Every node is parameterized by some latent true probability distribution over a quantitative attribute, e.g. a randomly chosen dinner table has true mass distributed according to density f_v.
- Edges between nodes represent proximity in some suitably defined space (e.g. we might expect an edge between kitchen table and dinner table). We will explore both weighted and unweighted edges.
- At each node we have noisy data from the true distribution f, i.e. $y_i = x_i + e_i$, where $x_i \sim f$ and e_i is a noise term.

Having formulated the problem in this way, the central task becomes one of *denoising* the inferred probability distributions f_v based on the raw data $\{y_1, \dots, y_N\}$ at each node, while simultaneously *smoothing* those inferred distributions across the graph. This falls within a broad class of methods known as graph smoothing, which has proven effective in dealing with outliers and data sparsity at large scales. Graph smoothing techniques based on penalizing likelihood estimates along the edges of a graph can be given a Bayesian interpretation as a spatial prior from a conditional autoregressive model (CAR) (Besag, 1974). My group has extensive experience with these methods, which we have successfully scaled to graphs with millions of

[1] https://github.com/google-research-datasets/distribution-over-quantities

vertices and edges, and which we have used, for example, to analyze fMRI data, to discover spatiotemporal productivity effects for ride-sharing services, and to detect radiological spatial anomalies. Of particular relevant interest is (Madrid-Padilla et al., 2018), in which my group developed the first linear-time algorithm for graph smoothing that achieves a near-minimax error rate over a wide class of graph structures. Another key building block here is our work in (Tansey et al., 2017), where we specifically address the problem of smoothing (although not denoising) a nonparametric probability density over a large graph. Other examples of my group's work on graph smoothing include (Tansey and Scott, 2015; Zuniga-Garcia et al., 2019; Tec et al., 2019).

However, existing graph-smoothing techniques, e.g. total variation denoising/fused lasso, typically assume Gaussian or sub-Gaussian noise (which cannot, for example, account for an 828,000-ton peanut). Thus to deal with outliers, we will need to extend graph smoothing methods to work with robust estimation models—for example, using heavy-tailed and/or skewed distributions for the error terms e_i. For more complex contamination models, we can also use scalable deconvolution techniques, similar to the model we explored in (Madrid-Padilla et al., 2018).

We will investigate the best methods to build the edges of the graph. One simple strategy would be to use a distance measure in a distributional embedding space such as *word2vec*. This approach can be partially justified by the work of Gupta et al. (2015), who show that distributional vectors encode hard facts—including quantitative attributes. We can also draw ideas from related previous research; for instance, Tandon et al. (2014) used *wordnet* (Miller, 1995) for creating constraints in a disambiguation model; and Fulda et al. (2017) used *word2vec* and *ConceptNet* (Speer et al., 2016) for pruning action spaces in text-based games.

Step 2: Variational Bayesian Hierarchical Modeling. We propose to address correlation among attributes using Bayesian modeling, a natural framework for making inferences about full probability distributions (as opposed to point estimates). Bayesian models can help us recover multivariate information because they allow us to define a probabilistic model where an object's observed values of an attribute depend on that object's *latent features*, which in turn can be modeled using a prior that encodes a complex correlation structure. Thus even with independently collected attributes, there is an underlying dependency induced by the latent variables. More precisely, we propose to explore models based on the following assumptions:

- Every textual object (e.g. *black bear*) is represented as a low-dimensional latent feature vector Z using a suitable embedding representation.
- We postulate a set of latent parameters λ_j for each marginal distribution of an attribute Y_j. In the simplest case, we might assume a Gaussian distribution for the marginals so that λ_j is the mean and variance (although we will go beyond such simplistic models). We assume that we observe samples of each attribute Y_j independently; for example, data on *mass* could have been extracted from a different document than the data on *length*.
- We define our probability model so that the parameters $\lambda = (\lambda_1, \dots, \lambda_n)$ have a joint probability distribution, conditional on Z, that captures correlations. This hierarchical step is what enables us to flexibly model correlations in the attributes Y: even when the attributes are independent given their latent parameters, integrating out the latent λ results in dependence among Y, because λ is dependent on Z.

The challenge here is to define an appropriate probability model for the latent parameters λ of the marginal distributions of the attributes in a way that extracts all the information available in Z, captures the multivariate structure of physical attributes, and that is trainable at scale. Traditional Bayesian inference methods such as MCMC are computationally prohibitive for the scale of our problem, but recent advances in variational inference have enabled Bayesian modeling at scale using stochastic gradient descent methods (Kingma and Welling, 2013). One possible approach is to use conditional distributions implicitly defined by complex transformations; for example, taking $\lambda \sim g_\theta(\lambda \mid Z)$ where g_θ is parametrized by a Bayesian neural network depending on Z (Yin and Zhou, 2018). Another promising approach is based upon decomposing the output space of the attributes as a Polya-tree structure, similar to a k-d tree, where the latent parameters λ are generated by a neural network and used to assign probability weights to each partition of the space. In (Tansey et al., 2018), we showed that this technique can produce rich and flexible conditional distributions.

We will also consider the problem of choosing the right framework of word embeddings. Gupta et al. (2015) show evidence that distributional vectors (e.g., *word2vec*) capture hard attribute information. In addition, Wang et al. (2017) show that distributional vectors work well for one-shot learning of definitional attributes. Thus, distributional embeddings are a sensible embedding framework. Another reasonable approach will be to use ConceptNet-based embeddings where the semantic distance is based on a graph distance between named categories (Speer et al., 2016). Finally, we will also try state-of-the-art language models such as BERT (Devlin et al., 2018), which have the advantage of being context-sensitive, and thus may help with polysemy and improve generalization. We remark that while we have treated the embedding representation of Z as given, embedding models can be fine-tuned to yield better results.

Another course of action to consider for this framework will be to investigate strategies for modeling the varying degrees of prior confidence on raw estimates. For example, there may be a subset of textual objects for which the estimated raw distributions are known to be correct with very high confidence. There are multiple ways in which prior confidence can be included in a Bayesian framework—for example, using informative priors or weighted/power likelihoods. This will be an area for experimentation and research.

To summarize, the proposed Bayesian modeling approach offers several advantages: first, we can obtain easy multivariate samples of the attributes of an object; second, it is based on embeddings, thus it considers the object's features and enables generalization to unseen objects; finally, it is compatible with deep learning technologies, which can be useful for transfer learning and are familiar in many existing NLP pipelines.

Concluding remarks. We have described a novel framework that can be used to improve the estimates of information extraction methods of CSK attributes from text corpora. Our proposed methods are based on spatial denoising and Bayesian modeling and are designed to explicitly deal with the three major limitations we have identified in current approaches. Our framework focuses on the extraction of quantitative attributes, but can be extended to deal with other types of extraction, such as *object-action affordances* (Fulda et al., 2017). Our methods leverage advancements in graph smoothing and variational inference to perform at the scale necessary for large resources. We intend to test our framework on DoQ (Elazar et al., 2019), the state-of-the-art resource containing information on 350k textual objects on 10 quantitative attributes. Our proposed approach takes as starting point a resource with noisy independent estimates by objects and attributes; and it generates a resource with denoised estimates that take into account the spatial (semantic) structure of the objects, multivariate information among attributes, and depends only on embedding representations, making the resource flexible, robust, semantic-aware and generalizable.

References

Bagherinezhad, Hessam, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi (2016). "Are Elephants Bigger than Butterflies? Reasoning about Sizes of Objects". In: *AAAI Conference on Artificial Intelligence*.

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems". In: *Journal of the Royal Statistical Society B* 24, pp. 192–236.

Dagan, I., D. Roth, F. Zanzotto, and M. Sammons (2013). *Recognizing Textual Entailment: Models and Applications*. Morgan & Claypool.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805.

Elazar, Yanai, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth (2019). "How Large Are Lions? Inducing Distributions over Quantitative Attributes". In: *CoRR* abs/1906.01327.

Fulda, Nancy, Daniel Ricks, Ben Murdoch, and David Wingate (2017). "What can you do with a rock? Affordance extraction via word embeddings". In: *CoRR* abs/1703.03429.

Goodman, Noah D, Joshua B Tenenbaum, and Tobias Gerstenberg (2014). *Concepts in a probabilistic language of thought*. Tech. rep. Center for Brains, Minds and Machines (CBMM).

Gupta, Abhijeet, Gemma Boleda, Marco Baroni, and Sebastian Pado (2015). "Distributional vectors encode referential attributes". In: *ACL*, pp. 12–21.

Kingma, Diederik P and Max Welling (2013). "Auto-Encoding Variational Bayes". In: *arXiv e-prints*. eprint: 1312.6114.

Luketina, Jelena, Nantas Nardelli, Gregory Farquhar, Jakob N. Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel (2019). "A Survey of Reinforcement Learning Informed by Natural Language". In: *CoRR* abs/1906.03926. arXiv: 1906.03926.

Madrid-Padilla, Oscar H., Nicholas G. Polson, and James G. Scott (2018). "A deconvolution path for mixtures". In: *Electron. J. Statist.* 12.1, pp. 1717–1751.

Madrid-Padilla, Oscar H., James G. Scott, James Sharpnack, and Ryan J. Tibshirani (2018). "The DFS Fused Lasso: Linear-Time Denoising over General Graphs". In: *Journal of Machine Learning Research* 18.176, pp. 1–36.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv e-prints*, arXiv:1301.3781, arXiv:1301.3781. arXiv: 1301.3781 [cs.CL].

Miller, George A. (Nov. 1995). "WordNet: A Lexical Database for English". In: *Commun. ACM* 38.11, pp. 39–41.

Speer, Robyn, Joshua Chin, and Catherine Havasi (2016). "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge". In: *CoRR* abs/1612.03975. arXiv: 1612.03975.

Tandon, Niket, Gerard Melo, and Gerhard Weikum (2014). "Acquiring Comparative Commonsense Knowledge from the Web". In: *AAAI Conference on Artificial Intelligence*.

Tansey, Wesley and James G. Scott (2015). "A Fast and Flexible Algorithm for the Graph-Fused Lasso". In: *arXiv e-prints*, arXiv:1505.06475, arXiv:1505.06475. arXiv: 1505.06475 [stat.ML].

Tansey, Wesley, Alex Athey, Alex Reinhart, and James G. Scott (2017). "Multiscale Spatial Density Smoothing: An Application to Large-Scale Radiological Survey and Anomaly Detection". In: *Journal of the American Statistical Association* 112.519, pp. 1047–1063.

Tansey, Wesley, Karl Pichotta, and James Scott (2018). "Leaf-Smoothed Hierarchical Softmax for Ordinal Prediction". In:

Tec, Mauricio, James G. Scott, and Natalia Zuniga-Garcia (2019). "Large-Scale Spatiotemporal Density Smoothing with the Graph-fused Elastic Net: Application to Ride-sourcing Driver Productivity Analysis". In: *Work in Progress*. eprint: https://github.com/mauriciogtec/RideAustinSpatioTemporalDensitySmoothing.

Wang, Su, Stephen Roller, and Katrin Erk (2017). "Distributional model on a diet: One-shot word learning from text only". In: *IJCNLP*.

Yin, Mingzhang and Mingyuan Zhou (2018). "Semi-implicit variational inference". In: *ICML*.

Zuniga-Garcia, Natalia, Mauricio Tec, James G. Scott, Natalia Ruiz-Juri, and Randy B. Machemehl (2019). "Evaluation of Ride-Sourcing Search Frictions and Driver Productivity: A Spatial Denoising Approach". In: *(to appear in) Journal of Transportation Research Part C*. eprint: arxiv:1809.10329.