Finding Convincing Arguments using Scalable Bayesian Preference Learning

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

We introduce a scalable Bayesian preference learning method for identifying convincing arguments in the absence of gold-standard ratings or rankings. In contrast to previous work, we avoid the need for separate approaches or pipelines to produce training data, predict rankings and perform pairwise classification. Although Bayesian methods are known to be effective when faced with sparse or noisy training data, they have not previously been used to identify convincing arguments. One deterrent is their perceived lack of scalability, which we address by developing a stochastic variational inference method for Gaussian process (GP) preference learning. We show how our method can be applied to predict argument convincingness from crowdsourced data, outperforming state-of-the-art methods, particularly when the data is sparse or noisy. We demonstrate how our Bayesian approach enables more effective active learning, thereby reducing the amount of data required to identify convincing arguments for new users and domains. While word embeddings are principally used with neural networks, our results show that word embeddings in combination with linguistic features also benefit GPs when predicting argument convincingness.

Two options for the paper's main claim (at the moment it is more toward number 1): (1) We provide a scalable Bayesian

1 Introduction

Argumentation is intended to persuade the reader of a particular point of view and is an important way for humans to reason about controversial topics (Mercier and Sperber, 2011). The amount of argumentative text on any given topic can, however, overwhelm a reader, particularly considering the scale of historical text archives and the prevalence of social media platforms with millions of authors. To gain an understanding of a topic, it is therefore useful to identify high-quality, persuasive arguments from different sides of a debate.

Theoretical approaches for identifying high-quality arguments have proved difficult to apply to everyday arguments (Boudry et al., 2015). However, empirical approaches using machine learning have recently shown success in identify convincing arguments in online discussions. Habernal and Gurevych (2016b) trained a model of convincingness on arguments taken from discussion forums, then used the model to predict convincingness of arguments from a new topic. To apply machine learning to predict which arguments are most convincing, we require our target audience to provide examples of arguments paired with judgements of their convincingness. Consider the arguments in Figure 1, which are taken from online discussion forums: how does a member of our audience assign a numerical convincingness score to each argument? If the audience considers each argument independently, it is difficult to ensure that scores remain consistent with their view of convincingness after they have judged multiple arguments. A solution to this problem is to compare argument 1 and 2 against one another. In this case we may judge that argument 1 is less convincing due to its writing style, whereas argument 2 presents evidence in the form of historical events. Pairwise comparisons are known to place less cognitive burden on human annotators than asking them

Topic: "William Farquhar ought to be honoured as the rightful founder of Singapore".
Stance: "No, it is Raffles!"
Argument 1: HE HAS A BOSS(RAFFLES) HE HAS TO FOLLOW HIM AND NOT GO ABOUT DOING ANYTHING ELSE...
Argument 2: Raffles conceived a town plan to remodel Singapore into a modern city. The plan consisted of separate areas for different...

Figure 1: Example of an argument pair.

to choose a numerical rating and allow fine-grained sorting of items that is not possible with categorical labels (Kendall, 1948; Kingsley, 2006). By using relative judgements instead of numerical scores, we can also avoid the problem that multiple annotators may have different biases toward high, low or middling values, making their sores hard to compare.

motivate use of word embeddings as additional input to GP. They have been shown to provided complementary information in applications such as... Mainly because they provided some semantic information not captured by the other features.

We propose the use of Gaussian process preference learning (GPPL) (Chu and Ghahramani, 2005) to model of argument convincingness as a function of textual features, including word embeddings, which can be inferred from noisy crowdsourced pairwise preferences. We address the poor scalability of GPPL by developing a stochastic variational inference (SVI) approach (Hoffman et al., 2013). Our evaluation using datasets provided by Habernal and Gurevych (2016b) shows that our method outperforms the previous state-of-the-art for ranking arguments by convincingness and identifying the most convincing argument in a pair. Further experiments with subsets of crowdsourced data show that our Bayesian approach is particularly advantageous with small, noisy datasets.

The rest of the paper is structured as follows. Section 2 reviews related work on argumentation, then Section 3 motivates the use of Bayesian methods by discussing their successful applications in NLP. In Section 4, we review preference learning methods and then in Section 5 we present our scalable Gaussian process-based approach. Section 6 then presents our evaluation: a comparison with the state-of-the art on predicting preferences in online debates; noisy datasets; active learning; and feature relevance determination. Finally, we present conclusions and avenues for future work.

2 Identifying Convincing Arguments

Habernal and Gurevych (2016b) used pairwise comparisons obtained using crowdsourcing to train models for predicting both pairwise labels for arguments and numerical scores of convincingness. However, this crowdsourced data required quality control techniques to account for errors. We may also wish to use other sources of noisy pairwise data to train a model of convincingness. For instance, when there is insufficient annotated data, we may try to learn which arguments a user finds convincing from their actions in a software application. A user's clicks can be interpreted as pairwise preferences (Joachims, 2002), for example if they select an argument from a list, however the resulting pairwise labels are very noisy. We may also be faced with very small amounts of data when we move to new domains and topics, which can present a problem to methods such as deep neural networks (Srivastava et al., 2014). The approach used by Habernal and Gurevych (2016b) to handle unreliable crowdsourced data involved first determining consensus labels using the MACE algorithm (Hovy et al., 2013); these consensus labels were then used to train a classifier and as input to PageRank; the resulting rankings were then used as training data for regression models. However, such pipeline approaches can be prone to error propagation (Chen and Ng, 2016) and consensus algorithms such as MACE require multiple crowdsourced labels for each argument pair, and so have higher annotation costs. Recently, Habernal and Gurevych (2016a) analysed reasons provided by annotators for why one argument is more convincing than another. In this paper we assume that explicit reasons are not provided. Investigations by Lukin et. al (2017) demonstrated the effect of personality and the audience's prior stance on persuasiveness, although their work does not extend to modelling persuasiveness using preference learning. The sequence of arguments in a dialogue is another important factor in their ability to change the audience's opinions (Tan et al., 2016). Reinforcement learning has been used to choose the best argument to present to a user (Rosenfeld and Kraus, 2016; Monteserin and Amandi, 2013), but such approaches do not model user preferences for arguments with certain qualities.

Move some of the discussion from the intro to here about the pipeline etc.

3 Bayesian Methods for NLP

When faced with a lack of annotated data or noisy labels, Bayesian approaches have a number of advantages. Bayesian inference provides a mathematical framework for combining multiple observations with prior information. Given a model, M, and observed data, D, we can apply Bayes' rule to obtain a posterior distribution over M:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}. \quad (1)$$

If the dataset is large, the likelihood $P(D|M)$ will dominate the posterior, but when D is small, the posterior will remain closer to the prior, $P(M)$, thereby reducing overfitting. In contrast, neural network methods typically select model parameters to maximise the likelihood, meaning that Bayesian methods can perform better with small datasets (Xiong et al., 2011). Bayesian methods naturally operate using unsupervised or semi-supervised learning, which can be an advantage when labelled training data is in short supply. A popular example in NLP is the use of Latent Dirichlet Allocation (LDA) for topic modelling (Blei et al., 2003), and its extension, the hierarchical Dirichlet process (HDP) (Teh et al., 2005), which learns the the number of topics that best describes the data rather than requiring it to be fixed a priori. More recently, semi-supervised Bayesian learning has been used to achieve state-of-the-art results for semantic role labelling (Titov and Klementiev, 2012). Bayesian inference can also be used to combine multiple pieces of evidence. For instance, it is possible to infer attack relations between arguments by combining votes for acceptable arguments from different people using a Bayesian network (Hiroyuki Kido, 2017). Errors in crowdsourced annotations can also be remedied using a Bayesian approach that simultaneously learns a sentiment classifier (Simpson et al., 2015; Felt et al., 2016). Many successful Bayesian approaches make use of Gaussian processes (GP), which are a particular form of prior distribution over functions of input features. For example, they have been used to analyse the relationship between text features of tweets and user impact on Twitter (Lampos et al., 2014), to predict the level of emotion in text (Beck et al., 2014), or to estimate the quality of a machine translation given the source and translated texts (Cohn and Specia, 2013).

4 Preference Learning

The goal is to learn convincingness as a function of argument features given a set of *pairwise preference labels*, where a label $x_i \succ x_j$ expresses that the user finds argument x_i more convincing than argument x_j. Pairwise labels can be predicted using a generic classifier without the need to learn a total ordering. To do this, pairs of items are transformed either by concatenating the feature vectors of two items (Habernal and Gurevych, 2016b), or by computing the difference of the two feature vectors, as in SVM-Rank (Joachims, 2002). The classifier is then trained using the transformed feature vectors as input data and the preference labels binary class labels. However, the ranking of items is useful for producing ordered lists in response to a query – consider a sorted list of the most convincing arguments in favour of topic X. Another approach is to learn the ordering directly using Mallows models (Mallows, 1957), which define distributions over permutations of a list. Mallows models have been extended to provide a generative model (Qin et al., 2010) and to be trained from pairwise preferences (Lu and Boutilier, 2011), but inference is typically costly since the number of possible permutations to be considered is $\mathcal{O}(N^2)$, where N is the number of items to be ranked. Modelling only the order of items means we are unable to quantify how closely rated items at similar ranks are to one another: how much better is the top ranked item from the second-rated?

To avoid the problems of classifier-based and permutation-based methods, another approach is to learn a set of real-valued scores from pairwise labels that can be used to predict rankings, pairwise labels, or as ratings for individual items. There are two established approaches for mapping discrete pairwise labels to real-valued scores: the Bradley-Terry-Plackett-Luce model (Bradley and Terry, 1952; Luce, 1959; Plackett, 1975) and the Thurstone-Mosteller model (Thurstone, 1927; Mosteller, 2006). More recently, Bayesian extensions of the Bradley-Terry-Plackett-Luce model were proposed by (Guiver and Snelson, 2009;

Volkovs and Zemel, 2014), while the Thurstone-Mosteller model was used by (Chu and Ghahramani, 2005). This latter piece of work assumes a Gaussian process (GP) prior over the scores, which enables us to predict scores for previously unseen items given their features using a Bayesian nonparametric approach. Nonparametric methods allow the function complexity to grow with the amount of data. Gaussian processes are a well established tool for extrapolating from training data in a principled manner, taking into account model uncertainty that may arise when data for new domains is limited (Rasmussen and Williams, 2006).

reduce this scalability discussion

The inference method used by Chu and Ghahramani (2005)has memory and computational costs that scale with $\mathcal{O}(N^3)$. Besides this limitation, there is also a computational and memory cost during training of $\mathcal{O}(N^2)$ due to the number of pairs in the training dataset. Recently, the stochastic variational inference (SVI) algorithm proposed by (Hoffman et al., 2013) has been used to address this problem in Gaussian process models (Hensman et al., 2013; Hensman et al., 2015) but has not previously been adapted for preference learning with GPs. The next section explains how we apply this technique to create a scalable preference learning method for argument convincingness.

5 Scalable Bayesian Preference Learning

Summarise this whole section in a paragraph or two. Move some background detail from previous section to here. OR just make it more argument-focused. Where to put in the linguistic features etc?

Following Chu and Ghahramani (2005), we model the relationship between a latent preference function, f, and each observed pairwise label, $v_k \succ u_k$, where k is an index into a list of P pairs, as follows:

$$p(v_k \succ u_k | f(v_k), f(u_k), \delta_{v_k}, \delta_{u_k}) = \begin{cases} 1 & \text{if } f(v_k) + \delta_{v_k} \geq f(u_k) + \delta_{u_k} \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where $\delta_i \sim \mathcal{N}(0, 1)$ is Gaussian-distributed noise. The noise term allows for variations in the observed preferences, which may occur if different annotators disagree or change their minds, or if the preferences are derived from noisy implicit data such as clicks streams. For the latent function , f, we assume a Gaussian process prior: $f \sim \mathcal{GP}(0, k_\theta/s)$, where k_θ is a kernel function with hyper-parameters θ, and $s \sim \mathcal{G}(a_0, b_0)$ is an inverse scale parameter drawn from a gamma prior with shape a_0 and scale b_0. The kernel function controls the smoothness of f over the feature space.

The inference goal is to learn the posterior distribution over the function values $f(i)$ for each item i. Chu and Ghahramani (2005) used a Laplace approximation, which finds a maximum a-posteriori (MAP) solution that has been shown to perform poorly in some cases (Nickisch and Rasmussen, 2008). Instead, we approximate a fully Bayesian approach adapting a variational inference method (Reece et al., 2011; Steinberg and Bonilla, 2014) to the preference likelihood given by Equation 2. Given a set of observed preference pairs, $\boldsymbol{y}$, the variational method assumes an approximation, $q(f, s)$, to the true posterior distribution, $p(f, s|\boldsymbol{y}, \theta, a_0, b_0)$. The algorithm iteratively maximises a lower bound on the log marginal likelihood, $\mathcal{L} \leq \log p(\boldsymbol{y}|\theta, a_0, b_0)$, and in doing so converges to an approximate posterior, $q(f, s)$, that minimises the Kullback-Leibler divergence of $p(f, s|\boldsymbol{y}, \theta, a_0, b_0)$ from $q(f, s)$.

To provide an inference algorithm that scales with the number of arguments and number of pairs, we adapt stochastic variational inference (SVI) (Hensman et al., 2013; Hensman et al., 2015) to our proposed variational method. For SVI, we assume M *inducing points*,which act as a substitute for the observed arguments. SVI further limits computational costs by considering only a subset of the data containing P_n pairs at each iteration. By choosing $M << N$ and $P_n << P$, we limit the computational complexity to $\mathcal{O}(M^3 + P_n)$ and the memory complexity $\mathcal{O}(M^2 + MP_n + P_n^2)$. To choose representative inducing points, we use K-means with $K = M$ to rapidly cluster the feature vectors, then take the cluster centres as inducing points.

We use a kernel function of the standard form $k_\theta(\boldsymbol{x}, \boldsymbol{x}') = \prod_{d=1}^{D} k_d(|x_d - x'_d|/l_d)$, where k_d is a function of the distance between the values of feature d for items x and x', and a length-scale hyper-parameter, l_d. The length-scale controls the smoothness of the function across the feature space, and can be optimised by by choosing the value of l_d that maximises the lower bound on the log marginal likelihood, $\mathcal{L}$. This process is known as maximum likelihood II and is often referred to as automatic relevance determination (ARD) (Rasmussen and Williams, 2006), since features with large length-

scales are less relevant because their values have less effect on $k_\theta(\boldsymbol{x}, \boldsymbol{x}')$ than features with short lengthscales. In this work we also test a variant of the median heuristic (Gretton et al., 2012), which has been shown to perform well in practice. This heuristic sets l_d according to: $l_{d,MH} = \frac{1}{D}\text{median}(\{|x_{i,d} - x_{j,d}|\forall i = 1, .., N, \forall j = 1, ..., N\})$.

6 Experiments

6.1 Datasets and Evaluation

Show examples from the dataset if not already done so in intro.

Make the point about using word embeddings with GPs less of an afterthought, more of a novelty.

We use crowdsourced preference datasets provided by Habernal and Gurevych (2016b), which contain pairwise preference labels for arguments taken from online discussion forums. Each pairwise label can have a value of 0 meaning the annotator found the second argument in the pair more convincing, 1 to express no preference, or 2 to indicate that the first argument was more convincing. All datasets contain 32 folds, which correspond to 16 controversial topics, and two stances for each topic. To test different scenarios, we apply different pre-processing steps to produce four variants of the data, which are shown in Table 1. We use the *UKPConvArgStrict* and *UKPConvArgRank* datasets to evaluate performance on 'clean' data for classification and ranking respectively. The *UKPConvArgCrowdSample* is used to evaluate both classification and ranking performance with noisy crowdsourced data including conflicts and no-preference labels.

We refer to our Bayesian preference learning method as *GPPL*. We compared our SVI method for GPPL against a less scalable variational method that did not use inducing points nor stochastic subsampling of data. Since the size of the datasets in Table 1 made it impractical to run the latter method given memory and time constraints, we tested on small subsets of the UKPConvArgStrict dataset. We found that GPPL converges to the same result using our SVI method as with variational inference. For the complete datasets listed in Table 1 we therefore report results only for our SVI approach. We compare GPPL against the *SVM* and *BLSTM* methods used in (Habernal and Gurevych, 2016b) on both pairwise classification task (predicting which of two arguments is more convincing) and a ranking task. For pairwise classifications, SVM and BLSTM concatenate the feature vectors of each pair of arguments. For ranking, PageRank scores for items in the training folds are used to train SVM and BLSTM for regression.

6.2 Experiment 1: Toy Data

Our two basic tasks are to *score* arguments in terms of convincingness and to *classify* the preference label for a pair of arguments, i.e. predict which argument is preferred. We use simulated data to show how GPPL learns differently from the pairwise labels in comparison with SVM for the classification task and PageRank for the scoring task. We simulate four scenarios, each of which contains arguments labelled *arg0* to *arg4*. In each scenario, we generate a set of pairwise preference labels. These are depicted as convincingness graphs in Figure 2a. Each scenario is repeated 25 times: in each repeat, we select arguments at random from one fold of the UKPConvArgStrict, then associate these arguments with the labels arg0 to arg4. We then obtain feature vectors for each argument by computing mean Glove word embeddings, as in Habernal and Gurevych (2016b). We trained PageRank, GPPL and the SVM classifier on the preference pairs shown in each graph and used them to predict preferences for the arguments. Taking means over the 25 repeats, we plot the PageRank scores and GPPL latent function means in Figure 2b, and the GPPL and SVM classifications for pairs of arguments in Figures 2c and 2d.

In the "no cycle" scenario, arg0 is preferred to both arg1 and arg2, which is reflected in the PageRank and GPPL scores in Figure 2b. However, arg3 and arg4 are not connected to the rest of the graph and receive different scores with PageRank and GPPL. Unlike SVM, GPPL provides probabilistic classifications and is less confident for pairs that were not yet observed, e.g. arg2 $\succ$ arg4.

The "single cycle" scenario shows how each method handles a cycle in the preference graph. Both PageRank and GPPL produce equal values for the arguments in the cycle (arg0, arg1 and arg2). PageRank assigns lower scores to both arg3 and arg4 than the arguments in the cycle, while GPPL more intuitively gives a higher score to arg3, which was preferred to arg4. SVM predicts that arg0 and arg1 are preferred over arg3, although arg0 and arg1 are in a cycle so there is no reason to prefer arg0 and arg1. GPPL, in contrast, gives a weak prediction that

Dataset	Pairs	Arguments	No pref.	Dataset properties
UKPConvArgStrict	11642	1052	0	Combine crowdsourced labels with MACE and take $\geq$ 95% most confident labels; Discard arguments marked as equally convincing; Discard conflicting preferences.
UKPConvArgRank	16081	1052	3289	Combine crowdsourced labels with MACE and take $\geq$ 95% most confident labels; PageRank run on each topic to produce gold rankings.
UKPConvArg-CrowdSample	16927	1052	3698	No. pairs: ; no. arguments: ; no. don't knows: One original crowdsourced label per pair; PageRank run on each topic to produce gold rankings.

Table 1: Summary of the internet argument datasets produced using different processing steps.

arg3 is preferred.

In the "double cycle" scenario, PageRank and GPPL produce very different results. Here, the argument graph shows two paths from arg2 to arg0 via arg1 or arg3, and one conflicting preference arg2 $\succ$ arg0. GPPL scores the arguments as if the single conflicting preference, arg2 $\succ$ arg0, is less important than the two parallel paths from arg2 to arg0. In contrast, PageRank gives high scores to both arg0 and arg2. The classifications by GPPL and SVM are similar, but GPPL produces more uncertain predictions than in the first scenario due to the conflict.

Finally, "cycle with 9 undecided prefs" shows an exaggerated scenario in which we have added nine no-preference labels to the "no cycle" scenario, indicated by undirected edges, to simulate the case where multiple annotators labelled the pair and did not all agree. This does not affect the PageRank scores, but reduces the difference in GPPL scores between arg0 and the other arguments, since GPPL gives the edge from arg0 to arg0 less weight due to the undecided labels. This is reflected in the GPPL classifications, which are less confident than in the "no cycle" scenario. The SVM cannot be trained using the uncertain labels and therefore does not adapt to the undecided labels.

In conclusion, GPPL appears to resolve conflicts in the preference graphs in a more intuitive manner than PageRank, which was designed for ranking web pages by importance rather than preference. In contrast to SVM, GPPL is able to account for undecided labels to soften the latent convincingness function.

6.3 Experiment 2: Clean Data

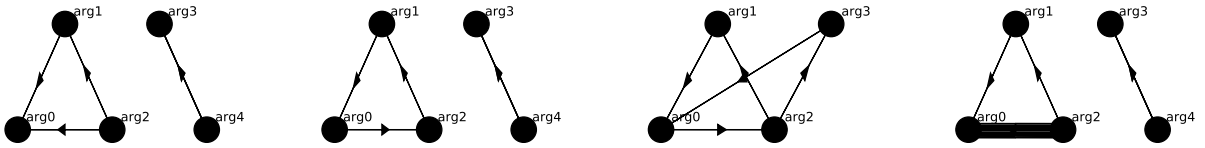
We comparemethods for classification on the UKPConvArgStrict dataset and ranking on the UKPConvArgRank dataset. Both datasets were cleaned to remove disagreements between annotators as stated in Table 1 To choose settings for the GPPL hyperparameters a_0 and b_0, which control the noise variance, we tested three different settings and found $a_0 = 2$, $b_0 = 200$ to be most effective. This is a weak prior favouring a moderate level of noise. To set the length-scale for GPPL, we compare the median heuristic (labelled "medi.") with the MLII optimisation method (labelled as "OptGPPL"). We also compare multiplicative and additive combinations for the kernel functions for each feature. We tested GPPL with different sets of input features: 32000 linguistic features labelled as *ling*, which we also use for SVM, as in Habernal and Gurevych (2016b)); *Glove* word embeddings with 300 dimensions, which we also use for BLSTM, also as in Habernal and Gurevych (2016b); and the combination of both lin and Glove embeddings (*ling+Glove*). To create a single embedding per argument as input for GPPL, we take the mean of the individual word embeddings for the tokens in the argument.

The results are shown in Table 2. When using

emphasise we beat state of the art

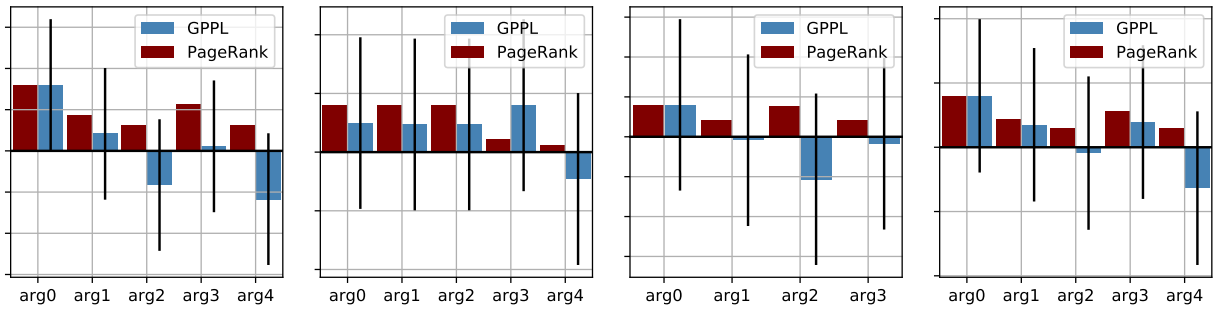
what kind of error analysis is required? Review other papers on similar topics (see comments in latex source). Where do the models differ?

Note that human upper bound is 93% so we could still improve (with better background knowl-

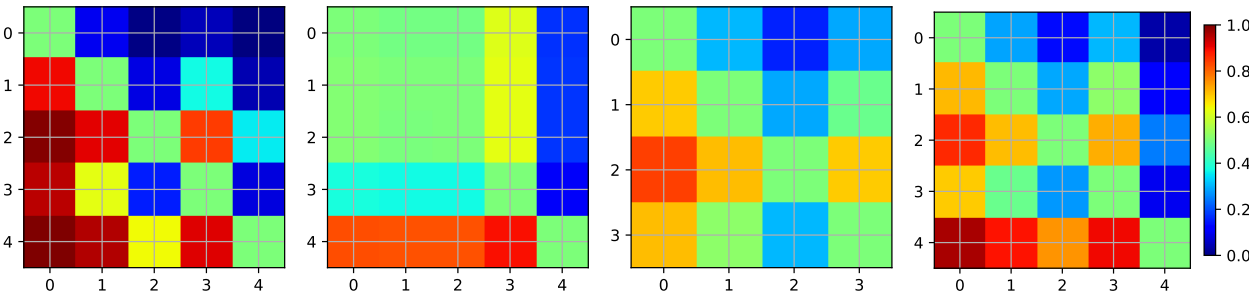


no cycle | single cycle | double cycle | cycle with 9 undecided prefs.

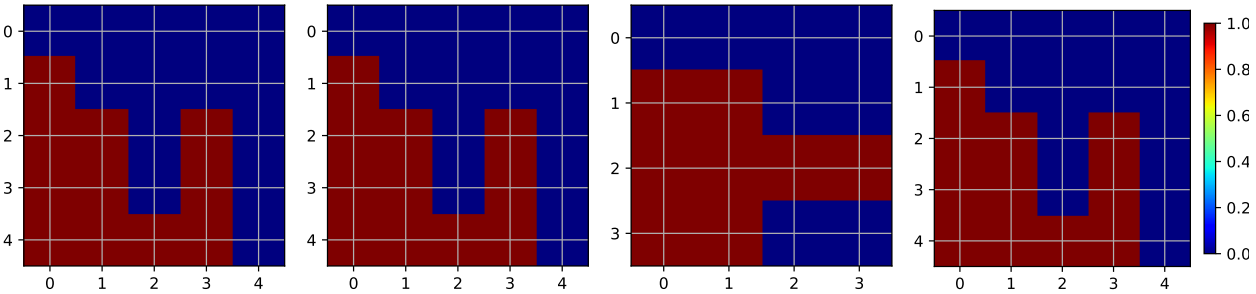
(a) Argument preference graphs for each scenario. Arrows point to the preferred argument.



(b) Ranking scores for arguments (bars for GPPL show standard deviation of latent preference function)



(c) GPPL predictions: probability that the argument on the horizontal axis is preferred to the argument on the vertical axis.



(d) SVM predictions: probability that the argument on the horizontal axis is preferred to the argument on the vertical axis.

Figure 2: Preference graphs and predictions for simulated arguments in different scenarios. The plots in each column correspond to a single scenario.

ling features, GPPL produces similar accuracy and improves the area under the ROC curve (AUC) by 2%, and cross entropy error by 0.01. Much larger improvements can be seen in the ranking metrics. When GPPL is run with mean Glove embeddings, it performs worse than BLSTM for classification but improves the ranking metrics. Using a combination of features, GPPL performs substantially better than the alternative methods for both classification and ranking, suggesting that embeddings and linguistic features contain complementary information. Optimising the length-scale using Bayesian model selection improves performance by 2% over the median heuristic. However, the cost of these improvements is that each fold required around 2 hours to compute instead of approximately 10 minutes on the same machine (an Intel i7 quad core desktop) using the median heuristic. While there is an improvement in then mean accuracy with length-scale optimisation, the accuracy does not improve for every fold. Since the optimisation step is performed using only the training folds and the model is tested on a different topic, there is a possibility of overfitting: features that are important in the test fold may not appear relevant in the training folds.

We hypothesised that GPPL benefits from integrating the GP to learn the latent preference function directly from the discrete noisy preference labels. We compare GPPL against a two stage method shown in Table 2 as *PL+SVR*: first, we use use the GPPL preference likelihood method without any item features to infer convincingness scores for each argument from the pairwise labels; second, we perform SVM regression on the inferred scores with *ling+Glove* features. The results show that PL+SVR does not reach the same performance as GPPL. This suggests that GPPL benefits not just from its preference likelihood but also from the integration of the GP.

For the pairwise classification task, we also compare GPPL against a Gaussian process classifier (*GPC*) to investigate whether other GP-based approaches produce comparable performance. As shown in Table 2, GPC produces the best results on the classification task, although it cannot be used to rank the arguments. While the classification approach involves learning over twice as many features – the features of the first and second items in each pair are concatenated – the GPC may perform better on this dataset because it is trained directly on the classification task, rather than through a preference learning likelihood.

Remove results for different kernels and just show best

6.4 Experiment 3: Conflicts and Noisy Crowdsourced Data

In this experiment, we introduced noise to both the classification and the regression tasks by comparing on the UKPConvArgCrowdSample dataset. Our goal was to investigate whether a Bayesian approach is better able to handle noise and conflicts.

Can we crunch expt 2 and 3 together? Perhaps 3 can be dropped entirely so we make the same point with active learning.

The results are shown in Table 3, showing that all methods perform worse when there are noisy or conflicting preferences. GPPL and GPC produce the best results, but GPC no longer has a clear advantage over GPPL. GPPL now outperforms the other methods in all metrics except Spearman's ρ, where PL+SVR performs slightly better. It is possible that GPC and SVM have the largest changes in accuracy compared to the UKPConvArgStrict results because these classification-based methods have no mechanism to resolve conflicts in the preference graph. The performance of the BLSTM classifier also decreases by a smaller amount, but was already poorer than the other methods on UKPConvArgStrict so it is hard to compare this change directly. PL+SVR is again slightly poorer than GPPL and GPC. Metrics for ranking on UKPConvArgCrowdSample show that while GPPL and PL+SVR continue to perform well, the results for BLSTM and particularly for SVM are much poorer than with UKPConvArgRank.

6.5 Experiment 4: Active Learning

We hypothesised that a Bayesian approach would deal better with sparse data and provide more meaningful confidence estimates. To test this hypothesis, we simulated an active learning scenario, in which we simulate an agent that iteratively learns a model for each fold. Initially, $N_{inc} = 2$ pairs were chosen at random from the training set, then used to train the classifier. The agent then performs *uncertainty sampling* to select the $N_{inc} = 2$ pairs with the least confident classifications. The labels for these pairs are then taken from the training set and used to re-train the model. The result is plotted in Figure 3, showing that GPPL is able to reach accuracies above 65%

UKPConvArgStrict										
	SVM	BLSTM	GPPL*, medi.			GPPL*, opt	GPPL+, medi.	GPPL+, opt	PL +SVR	GPC
	ling	Glove	ling	Glove	ling+ Glove					
Acc.:	0.78	0.76	0.78	0.71	0.79	0.80	0.78	0.78	0.78	0.81
AUC:	0.83	0.84	0.85	0.77	0.87	0.87	0.86	0.86	0.85	0.89
CEE:	0.52	0.64	0.51	1.12	0.47	0.51	0.69	0.69	0.51	0.43
UKPConvArgRank										
Pears.:	0.36	0.32	0.38	0.33	0.45	0.44	0.40	0.40	0.39	-
Spear.:	0.47	0.37	0.62	0.44	0.65	0.67	0.64	0.64	0.63	-
Kend.:	0.34	0.27	0.47	0.31	0.49	0.50	0.49	0.49	0.47	-

Table 2: Performance comparison on clean datasets.

	SVM ling	B-LSTM Glove	GPPL ling+ Glove	PL+ SVR ling+ Glove	GPC ling+ Glove
UKPConvArgCrowdSample					
Acc:	0.70	0.73	0.77	0.75	0.73
AUC:	0.81	0.80	0.84	0.82	0.86
CEE:	0.58	0.54	0.50	0.55	0.53
Pears.:	0.06	0.26	0.35	0.31	-
Spear.:	0.04	0.20	0.54	0.55	-
Kend.:	0.04	0.13	0.40	0.40	-

Table 3: Performance comparison on datasets containing conflicts and noise.

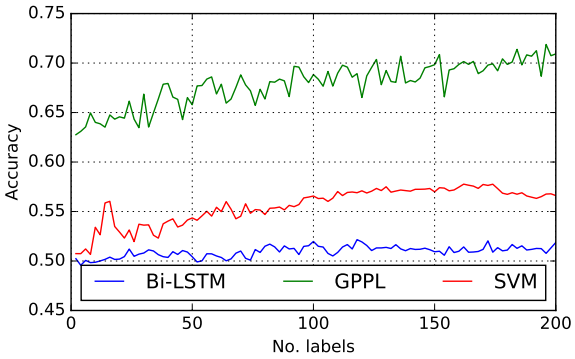


Figure 3: Active learning simulation for the three methods showing the mean accuracy of preference pair classifications over 32 runs.

with only 50 labels, while SVM and BLSTM do not reach the same performance given 200 labels. The accuracy of GPPL also increases by approximately 8% given 200 labels, while SVM increases approximately 6% and BLSTM only 2%. This suggest that GPPL may be a more suitable model to be used with uncertainty sampling in situations where obtaining labelled data is expensive.

6.6 Experiment 5: Embeddings

In our previous experiments, we found that including mean Glove word embeddings boosted performance above only using linguistic features. However, there are several alternative methods to mean word embeddings for representing longer pieces of text, notably skip-thoughts (Kiros et al., 2015) and Siamese-CBOW (Kenter et al., 2016). We compare mean Glove embeddings with skip-thoughts embeddings and Siamese-CBOW and show the results in Table 4. The best performance was obtained using mean Glove embeddings, despite the simplicity of this approach. Further work may be required to assess whether skip-thoughts and Siamese-CBOW can be improved if trained on different corpora.

cut this section and just state that we found no improvement with other embeddings?

remove results for different LS from table:embeddin

6.7 Experiment 6: Informative Features

Finally, we show how the length-scales learned by optimising GPPL can be used to identify informative sets of features. Since a larger length-scale causes greater smoothing, a very large length-scale implies that the value of that feature is irrelevant when predicting the function. In contrast, small length-scales indicate more informative features, since their precise value affects the latent preference function. Figure 4 shows the distribution of optimised length-scales on one fold of UKPConvArgStrict. The values shown are ratios of the optimised value to the

how much does this really show us? Could we skip this section to make more space for error analysis? In contrast

UKPConvArgStrict									
	Median heuristic						Optimised		
	Glove	Skip-thoughts	SCBOW	ling+ Glove	ling+ Skip-th.	ling+ SCBOW	ling+ Glove	ling+ Skip-th.	ling+ SCBOW
Acc.:	0.71	0.67	0.69	0.79	0.74	0.77	0.80	0.78	0.78
AUC:	0.77	0.72	0.75	0.87	0.81	0.85	0.87	0.85	0.85
CEE:	1.12	1.11	1.22	0.47	0.80	0.52	0.51	0.51	0.50
UKPConvArgRank									
Pears.:	0.33	0.30	0.29	0.45	0.34	0.39	0.44	0.34	0.40
Spear.:	0.44	0.49	0.40	0.65	0.59	0.63	0.67	0.52	0.63
Kend.:	0.31	0.36	0.28	0.49	0.43	0.47	0.50	0.37	0.47

Table 4: Comparison between different types of embeddings with GPPL

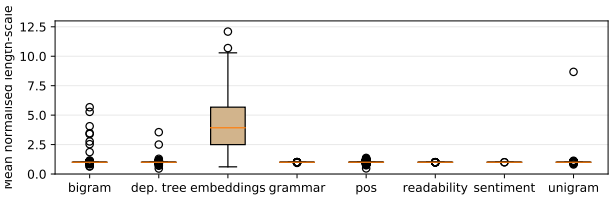


Figure 4: Distribution of length-scales for each type of feature after optimisation. Values are relative to the median heuristic value before optimisation, optimised on fold "should physical education be mandatory in schools – no", where optimisation increased accuracy from 75% to 80%.

Feature	Ratio
ProductionRule-S->ADVP,NP,VP,.,	0.466
Pos-ngram-PP-O-CARD	0.477
Unigram-"safer",	0.640
Bigram-"?"-"look"	5.672
Unigram-"safest"	8.673
Unigram-"safety"	271.190
Embedding-dimension-19	0.610
Embedding-dimension-241	12.093

Table 5: Ratios of optimised to median heuristic length-scales: largest and smallest ratios for linguistic features and word embeddings.

median heuristic. Due to the computation time required, our optimisation procedure was limited to 25 function evaluations. The large number of values close to 1 may be due to the L-BFGS-B algorithm not being able to optimise all features in the available time, suggesting that other features with larger gradients were prioritised for optimisation away from the median heuristic values. The length-scales for many dimensions of the mean word embeddings were increased, giving ratios close to 4 times the median heuristic, suggesting that these dimensions may be only very weakly informative. Table 5 shows the largest and smallest ratios for embeddings and linguistic features. The unigram "safety" has a very high length-scale, suggesting it is not informative and may be discarded.

6.8 Error Analysis

We compared the errors when using GPPL with mean Glove embeddings and with linguistic features. We manually inspected the twenty-five arguments most frequently mis-classified by GPPL *ling* and correctly classified by GPPL *Glove*. We found that GPPL *ling* mistakenly marked several arguments as less convincing where they contained grammar and spelling errors but otherwise made a logical point. In contrast, arguments that did not strongly take a side but did not contain language errors were often marked mistakenly as more convincing. We also examined the twenty-five arguments most frequently misclassified by GPPL *Glove* and correctly labelled by GPPL *ling*. GPPL *Glove* did not correctly mark arguments as less convincing even though they contained multiple exclamation marks and all-caps sentences. Other failures were very short arguments and underrating arguments containing the emotive term 'rape'. The analysis confirms that the different feature sets can identify different aspects of convincingness.

To investigate the differences between our best approach, GPPL opt. (ling+Glove), and the previous

best performer, SVM, we manually examined forty randomly chosen false classifications, where one of either GPPL (ling+Glove) or SVM (also ling) was correct and the other was incorrect. We found that both SVM and GPPL falsely classified arguments when they were either very short or long and complex, suggesting deeper semantic or structural understanding of the argument may be required. However, SVM also made mistakes where the arguments contained few verbs. We also compare the rankings produced by GPPL opt. (ling+Glove), and SVM on UKPConvArgRank, by examining the 20 largest deviations from the gold standard rank for each method. SVM underrated some arguments that GPPL did not where they contained exclamation marks, common spelling errors (likely due to unigram or bigram features). GPPL underrated short arguments with the ngrams "I think", "why?", and "don't know". In these cases, the phrases were used in a rhetorical question rather than to state that the author was uncertain or uninformed; these case may not be possible to distinguish given *ling* + *Glove* features.

A proposed advantage of GPPL is that it provides more meaningful uncertainty estimates. We examined whether the erroneous classifications correspond to more uncertain predictions when using GPPL compared to SVM when both methods use the *ling* features. For UKPConvArgStrict, the mean Shannon entropy of the pairwise predictions from GPPL was 0.129 for correct predictions and 2.443 for errors, showing that on average, more confident predictions were less likely to lead to errors. For SVM, the mean Shannon entropy was 0.188 for correct predictions and 1.583 for incorrect. The more extreme values for the GPPL predictions suggest that the probabilities provided by GPPL were indeed more reflective of the probability of error than those given by the SVM classifier.

7 Conclusions and Future Work

We presented a novel, scalable approach to predicting argument convincingness using Bayesian preference learning, and demonstrated how our method outperforms the state-of-the-art. We showed particularly strong performance with sparse and noisy training data, as may be found in crowdsourcing or interactive learning scenarios. Future work will evaluate our approach on other NLP tasks such as the argument reasoning comprehension task (Habernal et al., 2017) where reliable classifications may be difficult to obtain. We also plan to investigate whether the GP preference function can be trained using a combination of classifications and absolute scores as well as pairwise labels.

Acknowledgments

References

Daniel Beck, Trevor Cohn, and Lucia Specia. 2014. Joint emotion analysis via multi-task gaussian processes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1798–1803. ACL.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. 2015. The fake, the flimsy, and the fallacious: demarcating arguments in real life. *Argumentation*, 29(4):431–456.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Chen Chen and Vincent Ng. 2016. Joint inference over a lightly supervised information extraction pipeline: Towards event coreference resolution for resource-scarce languages. In *AAAI*, pages 2913–2920.

Wei Chu and Zoubin Ghahramani. 2005. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144. ACM.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *ACL (1)*, pages 32–42.

Paul Felt, Eric K Ringger, and Kevin D Seppi. 2016. Semantic annotation aggregation with conditional crowdsourcing models and word embeddings. In *COLING*, pages 1787–1796.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. 2012. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213.

John Guiver and Edward Snelson. 2009. Bayesian inference for plackett-luce ranking models. In *proceedings of the 26th annual international conference on machine learning*, pages 377–384. ACM.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *EMNLP*, pages 1214–1223.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task. *arXiv preprint arXiv:1708.01425*.

James Hensman, Nicolò Fusi, and Neil D Lawrence. 2013. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 282–290. AUAI Press.

James Hensman, Alexander G. de G. Matthews, and Zoubin Ghahramani. 2015. Scalable Variational Gaussian Process Classification. In *AISTATS*.

Keishi Okamoto Hiroyuki Kido. 2017. A bayesian approach to argument-based reasoning for attack estimation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 249–255.

Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H Hovy. 2013. Learning whom to trust with mace. In *HLT-NAACL*, pages 1120–1130.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

Maurice George Kendall. 1948. *Rank correlation methods.* Griffin.

Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. In *Proceedings of the The 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

David C Kingsley. 2006. Preference uncertainty, preference refinement and paired comparison choice experiments. *Dept. of Economics. University of Colorado*.

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*, pages 3294–3302.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on twitter. In *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 405–413.

Tyler Lu and Craig Boutilier. 2011. Learning mallows models with pairwise preferences. In *Proceedings of the 28th international conference on machine learning (icml-11)*, pages 145–152.

R Duncan Luce. 1959. On the possible psychophysical laws. *Psychological review*, 66(2):81.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *15th European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Colin L Mallows. 1957. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74.

Ariel Monteserin and Analía Amandi. 2013. A reinforcement learning approach to improve the argument selection effectiveness in argumentation-based negotiation. *Expert Systems with Applications*, 40(6):2182–2188.

Frederick Mosteller. 2006. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. In *Selected Papers of Frederick Mosteller*, pages 157–162. Springer.

Hannes Nickisch and Carl Edward Rasmussen. 2008. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(Oct):2035–2078.

Robin L Plackett. 1975. The analysis of permutations. *Applied Statistics*, pages 193–202.

Tao Qin, Xiubo Geng, and Tie-Yan Liu. 2010. A new probabilistic model for rank aggregation. In *Advances in neural information processing systems*, pages 1948–1956.

C. E Rasmussen and C. K. I. Williams. 2006. Gaussian processes for machine learning. *The MIT Press, Cambridge, MA, USA*, 38:715–719.

Steven Reece, Stephen Roberts, David Nicholson, and Chris Lloyd. 2011. Determining intent using hard/soft data and gaussian process classifiers. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8. IEEE.

Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4):30.

Edwin D Simpson, Matteo Venanzi, Steven Reece, Pushmeet Kohli, John Guiver, Stephen J Roberts, and Nicholas R Jennings. 2015. Language understanding in the wild: Combining crowdsourcing and machine learning. In *Proceedings of the 24th International Conference on World Wide Web*, pages 992–1002. International World Wide Web Conferences Steering Committee.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Daniel M Steinberg and Edwin V Bonilla. 2014. Extended and unscented gaussian processes. In *Advances in Neural Information Processing Systems*, pages 1251–1259.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2005. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392.

Louis L Thurstone. 1927. A law of comparative judgment. *Psychological review*, 34(4):273.

Ivan Titov and Alexandre Klementiev. 2012. A bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 12–22. Association for Computational Linguistics.

Maksims Volkovs and Richard S. Zemel. 2014. New learning methods for supervised and unsupervised preference aggregation. *Journal of Machine Learning Research*, 15(1):1135–1176.

Hui Yuan Xiong, Yoseph Barash, and Brendan J Frey. 2011. Bayesian prediction of tissue-regulated splicing using rna sequence and cellular context. *Bioinformatics*, 27(18):2554–2562.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560.