Discovering Better AAAI Keywords via Clustering with Community-sourced Constraints

Kelly Moran
Department of Computer Science
Tufts University
khmoran@google.com

Byron C. Wallace
Health Services Policy and Practice
Brown University
byron_wallace@brown.edu

Carla E. Brodley
Department of Computer Science
Tufts University
brodley@cs.tufts.edu

Abstract

Selecting good conference keywords is important because they often determine the composition of review committees and hence which papers are reviewed by whom. But presently conference keywords are generated in an ad-hoc manner by a small set of conference organizers. This approach is plainly not ideal. There is no guarantee, for example, that the generated keyword set aligns with what the community is actually working on and submitting to the conference in a given year. This is especially true in fast moving fields such as AI. The problem is exacerbated by the tendency of organizers to draw heavily on preceding years' keyword lists when generating a new set. Rather than a select few ordaining a keyword set that that represents AI at large, it would be preferable to generate these keywords more directly from the data, with input from research community members. To this end, we solicited feedback from seven AAAI PC members regarding a previously existing keyword set and used these 'community-sourced constraints' to inform a clustering over the abstracts of all submissions to AAAI 2013. We show that the keywords discovered via this data-driven, human-in-the-loop method are at least as preferred (by AAAI PC members) as 2013's manually generated set, and that they include categories previously overlooked by organizers. Many of the discovered terms were used for this year's conference.

1 Introduction

When submitting a paper for peer review to keyword-dependent conferences such as AAAI, authors must choose the keywords that best represent their paper from a fixed list. The senior program committee (SPC) is chosen to represent the high-level keywords and then they are asked to suggest PC members in their subfield, who are assigned papers for review by the author-selected keywords (sometimes indirectly by conference software that narrows down the abstracts that a PC member examines during the paper bidding period). Additionally, the conference co-chairs often use the high-level keywords to populate the program committee, and community members sometimes use them to identify papers of interest. Keywords thus play an important role in shaping conferences. Yet usually only a small set of conference organizers are responsible for generating the keyword list, and they typically draw heavily on the previous year's keywords. As a result, the keyword set may not accurately represent the distribution of papers currently being submitted to the conference; e.g., the keywords may lag behind the papers that they are intended to represent due to *concept drift*.

In contrast, a completely data driven approach to finding keywords would be to cluster the previous year's submissions and then derive (manually or automatically) a keyword for each cluster. This approach of gleaning keywords directly from submitted papers would allow one to find words that best fit the submissions, thus sidestepping problems of concept drift, but it may not produce a set of keywords that best reflects the conferences' needs. For example, data-derived keywords may not be able to account for some of the social or legacy factors in conference keyword determination. Indeed, it is important that keywords are guided by the community at large to align with *a priori* shared preferences regarding topics of interest; significant knowledge and effort have gone into expert-defined keywords over the years and we would like to leverage this information. In this paper, we outline our experience in attempting to take a data-driven, human-guided approach to defining a new set of keywords for AAAI 2014.

A broad overview to the strategy we took is as follows. Our dataset comprises the papers submitted to the AAAI 2013 main track. For each paper we have the abstract, title and one or more high-level keywords selected by the authors during submission from a fixed list (see column 1 in Table 2). We apply constraint-based clustering to the abstracts of these submitted papers, where the constraints are defined between AAAI 2013 keywords and reflect the preferences of the research community. (Recall that we know which keywords are associated with each 2013 submission.) More specifically, we elicited preferences regarding the 2013 keywords from seven PC members – a larger, more diverse group than relying on the current co-chairs – in order to weight the keyword-level constraints. These *community-sourced constraints* were elicited both with respect to both individual keywords (e.g., "*natural language processing* is too broad and should be broken up") and *pairs* of keywords ("*knowledge-based systems* could be merged

with *knowledge representation and reasoning*").

To leverage these preferences to inform our clustering, we incorporated keyword-level constraints via a semi-supervised clustering method (Preston et al. 2010) that used class labels (in our case keywords) to form probabilistic (soft) *must* and *cannot-link* constraints between pairs of instances (abstracts). These constraints induced a clustering that attempts to balance (1) the fit to the observed data and (2) agreement with community opinion concerning keywords. Next we manually derived keywords from the resulting clusters. We then crowdsourced our evaluation to the AAAI 2014 PC to compare the old and the new keywords. Finally, the results were given to this year's program chairs who created a final list of keywords, many of which were terms generated by this data-driven process.

In this work we have leveraged existing techniques and methods to generate a data-driven, community-sourced set of keywords within a constrained budget of time. We have made several practical decisions in our choice of methods, which could no doubt be refined and made more principled. We view this exercise as a novel demonstration of the potential that hybrid data-driven/community-sourced approaches have in terms of generating informative conference keywords; our key contribution is to argue for adopting such an approach in place of the outmoded practice of a completely manual process, and to provide a methodological starting point to this end. Another contribution of this case study is the elucidation of challenges to adopting such an approach.

2 Constraint Elicitation

Our aim is to jointly leverage last year's submissions, existing keywords and community opinion to generate a new set of keywords. As an operational realization of 'community opinion,' we solicited class-level constraint matrices with respect to last year's keywords directly from seven AAAI community members, several of whom were program chairs of previous AI conferences. By acquiring multiple sets of constraints we were able to downweight individual contributor bias (e.g., due to possible myopia stemming from his or her research focus). Ideally, even more preference sets would be collected, but practical time constraints – we had fewer than three months from the receipt of the data until the new set of keywords was needed for AAAI 2014 – precluded this possibility here.

We asked participating individuals to provide feedback for each of the 2013 keywords (column 1 of Table 2). Specifically, we asked if they thought each keyword was: (a) too general (and hence should be broken up into multiple keywords), or (b) coherent as-is. Respondents could also (c) express indifference, i.e., say "let the data speak for itself, I have no opinion." Additionally, we solicited *pairwise* constraints between keywords. Thus for every pair, respondents were asked to express whether the constituent keywords were likely to occur for the same paper (or if he or she had no opinion either way). We requested that all responses be mapped to an arbitrarily defined numerical scale between -1 and 1. See Figure 1 caption for details.

One benefit to acquiring multiple constraint sets is that they afford insight into agreement concerning existing keywords. Figure 1 shows a histogram plot of the pairwise constraint values in the constraint matrix C for each individual keyword and for each pair of keywords. The level of agreement regarding pairwise constraints varies widely. For example, in the case of constraint entry $C_{3,10}$ (*heuristic search and optimization* and *reasoning about plans, processes and actions*), there is near-complete agreement amongst all seven respondents that these should be merged. By contrast, there is a uniform distribution in responses concerning cell $C_{3,6}$ (*heuristic search and optimization* and *machine learning*), indicating that there is no consensus of opinion. Similarly, agreement regarding values on the diagonal also varies. For example, respondents were either neutral or believed that *multidisciplinary topics* ($C_{8,8}$) could be split across clusters. By contrast, there is a strong consensus that all papers with the keyword *robotics* ($C_{12,12}$) should be clustered together, i.e., that *robotics* is a coherent keyword.

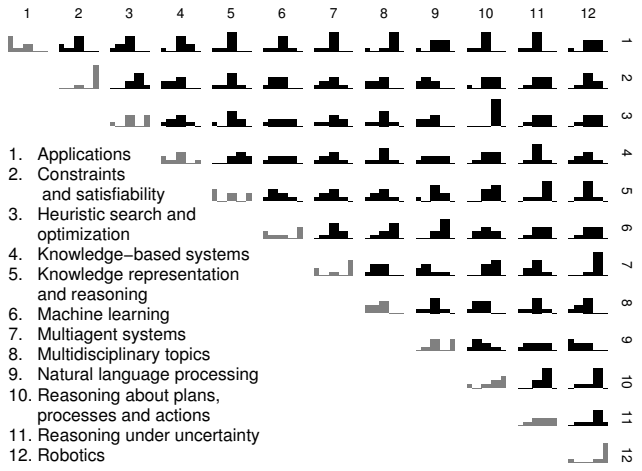

Figure 1: Histograms of the pairwise constraint values for twelve AAAI 2013 keywords as provided by seven domain experts. The x-axis captures the elicited preference from $\{-1, -0.5, 0, 0.5, 1\}$; the y-axis represents the respondent count ([0,7]). For pairwise constraints ($i \neq j$), -1 indicates that the corresponding pair should *not* be grouped together, 0 expresses indifference, and 1 indicates that the pair *should* be grouped together. For the diagonal (grey) elements, -1 indicates that the corresponding topic is too broad, 0 communicates indifference, and 1 suggests that the keyword is good as-is. Note that ± 0.5 are less confident versions of ± 1.

3 Constraint-Based Clustering

We now describe how we leveraged the elicited constraints just described to inform our clustering (and hence the generated keyword set). In short, we used a variant of (soft) constraint-based clustering.

As illustrated in Figure 1, in some cases reviewers do not form a consensus. Instead of giving equal sway to constraint contributions that effectively cancel out, we propose a method for downweighting constraints on which the domain experts disagree. To emphasize the constraints about

which PC members agreed, we weighted constraint values inversely proportional to the uniformity in respondent agreement. Operationally, we accomplished this via (the absolute value of) *skewness*, a measure of the degree of asymmetry of a distribution around its mean (Doane and Seward 2011), as calculated in Equation 1, where $\bar{x}$ is the mean and σ is the standard deviation of the data set $S = \{x_1, x_2, \ldots, x_n\}$.

$$\frac{n}{(n-1)(n-2)} \Sigma_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^3 \quad (1)$$

3.1 Multilabel Data

Most constraint-based clustering work focuses on singly-labeled data, which would correspond in our case to one keyword per submission. Under this assumption, the application of constraints to a given pair of instances is trivial: for a paper d_i with keyword k_i and a paper d_j with keyword k_j, look up the cell (k_i, k_j) in the constraint matrix (C_{k_i,k_j}). This value can then be used to modify the distance between instances d_i and d_j during clustering.

However, conference papers are often associated with multiple keywords. Furthermore, the constraints between the various pairs of keywords belonging to any two documents d_i and d_j may not agree: some of these pairwise constraints may be positive and others negative. We can allow these to cancel out, but this would overlook the fact that not all keywords are likely to be equally representative of a given paper.

To calculate constraints representative of the topics in a document, we must therefore move from binary labels to probabilistic keyword assignments. Thus instead of a value $\in \{0,1\}$ indicating the presence (or absence) of keyword m on a document d_i, we instead assume d_i comprises a mixture over keywords $k_{i,m}$ such that $\Sigma_{m=1}^{n_i} k_{i,m} = 1$, where n_i is the number of keywords associated with d_i.

As a practical means of accomplishing this, we estimated document-keyword mixture weights from the original labeled data via a Naïve Bayes model (Rish 2001), treating each label independently. That is, we trained a separate classifier for each of the twelve labels (one-vs-all) and used 10-fold cross validation to predict the probability that each (held out) document was labeled with each keyword. For a document d_i, its unnormalized value for the label $k_{i,m}$ was taken as the predicted probability that this document belonged to $k_{i,m}$ as opposed to all other classes. We then normalized the probabilities of the keyword assignments across all twelve labels for each document such that they summed to 1, which we call $p_{i,m}$ for the probability that the keyword k_m belongs to document d_i.

We can then calculate the total pairwise constraint $t_{i,j}$ between any two documents d_i with keywords $\{k_{i,1}, k_{i,2} \ldots, k_{i,n_i}\}$ and d_j with keywords $\{k_{j,1}, k_{j,2} \ldots k_{j,n_j}\}$ as follows.

$$t_{i,j} = \Sigma_{m=1}^{n_i} \Sigma_{l=1}^{n_j} w_{k_{i,m},k_{j,l}} C_{k_{i,m},k_{j,l}} p_{i,m} p_{i,l} \quad (2)$$

where $C_{i,j} \in [-1, 1]$ and $w_{i,j}$ is the skew weight for constraint $C_{i,j}$ (Equation 1). Intuitively, this crude topic mixture estimation provides a mechanism with which to weight the 'constraint contributions.' A more elegant approach would be to explicitly model documents as an ad-mixture of topics and is a topic for future work.

3.2 Clustering

Spectral Clustering Algorithm We chose a spectral clustering approach for two reasons: its handling of high-dimensionality feature spaces and its amenability to additional constraints. Briefly, spectral clustering methods operate on the similarity matrix of a dataset in lieu of the original feature space (Ng et al. 2002; Von Luxburg 2007). The similarity matrix S is an $n \times n$ matrix, where n is the number of instances, and S_{ij} represents some measure of "similarity" between instances i and j. This similarity can be a distance measure (such as the Euclidean distance between instances), a measure of text similarity (such as cosine similarity), or any arbitrary encoding of pairwise instance similarity (Strehl, Ghosh, and Mooney 2000). In spectral clustering one partitions the eigenvalues of the similarity matrix, e.g., via a normalized cuts algorithm (Dhillon, Guan, and Kulis 2004). Here we leveraged the Meila-Shi algorithm (Meila and Shi 2001) to take the eigenvectors with the k largest eigenvalues of the matrix $P = D^{-1}S$, and then clustered points in this space using k-means.

Feature Representation Our documents (abstracts of the submissions to AAAI 2013) included titles, abstracts, and author-specified free-text keywords. We encoded these using a standard bag of words representation, including unigrams and bigrams (Tan, Wang, and Lee 2002). Tokens from the titles and author-specified keywords were included as separate features. We used an English stoplist and removed terms that occurred fewer than 5 times in the corpus or were shorter than three characters. In total we used 3,275 unique features.

We experimented with six different feature representations: counts-based bag of words, binary bag of words, and topic modeling representations with 25, 50, 75, and 100 topics.[1] To choose the best representation, we clustered our documents using each candidate feature space and then attempted to rediscover these cluster labels using an SVM. The feature space with the fewest misclassifications was the counts-based bag of words.

Adding Constraints Spectral clustering provides a convenient method for injecting constraints via the aforementioned similarity matrix. First, we scaled all feature values to the [0,1] interval. Then each entry in the similarity matrix can be calculated as follows.

$$\hat{S}_{i,j} = S_{i,j} + \lambda t_{i,j} \quad (3)$$

where $S_{i,j}$ is a distance measure between documents d_i and d_j, λ is a parameter than controls the relative weight given to

[1] We used the counts-based bag of words representation as input to Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) and then encoded each document via the inferred proportions of each topic.

the constraints, and $t_{i,j}$ is the aggregate pairwise constraint value between d_i and d_j (Equation 2).

In practice, λ is a multiplier for each constraint value added to the similarity matrix. Thus a higher λ places more emphasis on the constraints and less emphasis on the distances between points. We set λ using the data by selecting a value that maximized the estimated log-likelihood of held-out documents under a simple generative model. For the purpose of setting λ we make a simplifying assumption that each document belongs to a single cluster (the cluster corresponding to the keyword with the highest probability in the feature space). Specifically, we assumed a mixture over multinomials wherein each cluster corresponds to a mixture component. We estimated the parameters of each component's corresponding multinomial using maximum likelihood. We calculated the log-likelihood of a given clustering Z by summing the estimated log likelihood of each document d_i given its most likely keyword/cluster z_{d_i} and associated multinomial parameter estimates. Thus we have:

$$\begin{aligned}\mathscr{LL}(Z|D) &= \Sigma_d log(Pr(d|Z)) + log(Pr(Z)) \\ &= \Sigma_d[log(Pr(d|z_d)) + log(\hat{\Pi}_{z_d})] \\ &= \Sigma_d[\Sigma_{w \in d} log(\hat{\Theta}_w^{z_d}) + log(\hat{\Pi}_{z_d})]\end{aligned} \tag{4}$$

where $\hat{\Pi}_{z_d}$ is the maximum likelihood estimate for the probability of the cluster (component) containing document d and $\hat{\Theta}_w^{z_d}$ is the estimate for the probability of word w in document d given the dominant cluster z_d. We therefore selected λ to maximize this estimated log likelihood, i.e.,:

$$\lambda^* = \arg\max_\lambda \mathscr{LL}(Z|D) \tag{5}$$

Under this simple model, we calculated the average log likelihood of held out documents for values of λ in the range $[0, 5]$ across clusterings including $[15, 22]$ components. Both of these ranges were deemed as *a priori* 'reasonable' (see the following section for a discussion regarding the number of components, i.e., keywords). The best value under this criterion for λ was 2. This value for λ was not sensitive to the number of components. Note that because setting $\lambda = 0$ ignores constraints, the higher log likelihood for $\lambda > 0$ demonstrates empirically the benefit of incorporating constraints.

Choosing the Number of Clusters Selecting the number of clusters (components in a mixture model) is generally a tricky problem (Milligan and Cooper 1985; Sugar and James 2003). However, here we have domain expertise to aid us in narrowing down our options to a suitable range (e.g., clearly having hundreds of keywords would not be helpful to conference participants or PC members; having 2 would be equally unhelpful). The authors (including Dr. Brodley, who has chaired or co-chaired large CS conferences, and thus has relevant domain expertise) selected reasonable bounds for the number of keywords. We started with the baseline of the previous year's number of keywords (12) and added a bit to this in light of the observation that elicited preferences suggested that many keywords be broken up. This resulted in selecting a lower bound of 15. An upper bound of 22 was then chosen, somewhat arbitrarily but reflecting intuition that more than 22 keywords would start to get unwieldy.

Because we were able to designate a rather narrow range for the number of keywords (clusters), we did not experiment with nonparametric methods (Teh et al. 2006) to select the number of components (keywords). Instead, as a practical strategy we opted to select a number within this range that maximized the likelihood of held out data under the simple mixture of multinomials model described above. Specifically, we conducted cross-validation, holding out a subset of the data with which to calculate the log likelihood of each held-out document given the clustering. (Again we assigned each test point to its dominant cluster.) We performed five-fold cross validation five times for each value of k in the identified range and averaged the log likelihoods for the held-out documents. This procedure suggested 21 as best value for k within the range of interest.

3.3 Summary of Approach

To briefly recapitulate: for the constraint matrix input to the clustering algorithm, we used the means of the community-informed constraints weighted by their (absolute) skews. To calculate the pairwise constraints between documents (which are associated with multiple keywords), we used Naïve Bayes to infer a distribution over keywords (i.e., a mixture) for each document. Finally, for the parameters of the spectral clustering algorithm (λ and k), we used the values determined above: λ=2 and k=21.

4 From Clusters to Keywords

Using the above approach, we induced a clustering over all abstracts submitted to AAAI 2013 (both accepted and rejected).[2] The output of this clustering approach is a hard assignment of documents to clusters. Our assumption was that these clusters map to (new) keywords. We acknowledge that this violates the reality of the situation; i.e., that articles will often span multiple topics. Our assumption here is that forcing a hard assignment will result in abstracts being assigned to clusters corresponding to the most dominant topics therein. Furthermore, such a hard assignment makes the next step – mapping clusters to keywords – an easier process.

Specifically, one needs to map clusters (or the papers comprising them) to keywords. This was by far the most manual step in our keyword generation process. Two of the authors collaboratively determined this mapping, using the following information to select cluster names:

- The top 20 words appearing in the titles and keywords of all $d_i \in z_j$;
- The top 20 topics[3] assigned to all $d_i \in z_j$;

[2]This set excluded the "special track" papers from *Artificial Intelligence and the Web*, *Cognitive Systems*, *Senior Members*, *Computational Sustainability* and *Artificial Intelligence, Robotics, and AI* Subarea Spotlights.

[3]Topics are the second tier of more fine-grained, conference-defined keywords also selected by authors.

	Clusters																				
AAAI 2013 Keyword	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Applications	0.9	1.2	**7.9**	0.0	**14.1**	0.5	0.2	0.0	1.6	0.0	1.0	0.0	3.3	5.8	4.0	1.1	2.1	0.8	0.0	0.6	**5.4**
Constraints . . .	0.0	0.9	0.3	0.0	0.3	0.4	0.0	0.0	1.3	**9.9**	1.9	0.0	0.0	0.0	0.8	1.1	0.9	**8.9**	0.4	2.2	1.0
Heuristic Search . . .	0.3	2.6	1.4	0.0	0.3	2.0	0.0	0.5	**11.4**	0.0	2.3	0.0	0.3	0.4	0.8	2.5	1.2	8.6	0.6	2.6	0.4
Knowledge Based . . .	0.0	0.2	0.0	0.0	0.2	1.0	0.0	0.0	0.0	0.3	0.4	0.0	**11.2**	0.0	0.0	0.1	**11.0**	0.0	0.0	0.0	0.8
Knowledge Representation	1.7	0.4	0.0	0.0	0.6	1.3	0.0	0.0	0.0	8.2	2.3	0.0	0.3	1.5	0.3	1.2	7.7	0.5	3.4	3.1	3.2
Machine Learning	4.2	**6.4**	2.1	**23.0**	8.5	2.4	0.5	5.8	7.3	0.0	**3.0**	**22.0**	4.5	2.7	2.4	0.7	2.1	0.0	**6.4**	1.4	2.5
Multiagent Systems	3.5	1.4	5.9	0.0	3.6	1.4	**29.0**	0.3	0.8	0.7	1.7	0.0	0.5	7.4	**4.8**	3.4	1.4	0.3	1.7	2.9	4.2
Multidisciplinary Topics	0.6	2.3	0.2	0.0	0.5	0.5	0.0	0.0	0.0	0.1	0.9	0.0	1.5	**14.4**	0.0	0.5	3.6	0.0	1.3	0.0	3.4
NLP	0.4	0.0	0.3	0.0	2.0	1.8	0.0	**17.3**	0.6	0.5	0.9	0.0	5.0	2.7	0.4	0.3	3.9	0.6	0.5	0.3	0.7
Reasoning about Plans . . .	0.0	1.0	0.2	0.0	0.6	0.7	0.0	0.0	0.7	0.7	1.1	0.0	0.4	1.4	0.7	**5.6**	1.1	0.2	0.5	**15.3**	1.5
Reasoning under . . .	**11.5**	0.0	2.8	0.0	0.3	**5.4**	0.3	0.0	0.9	0.5	1.8	0.0	0.8	0.9	3.9	1.2	1.1	0.9	0.4	2.7	1.3
Robotics	0.0	3.5	0.0	0.0	0.0	0.4	0.0	0.0	6.5	0.1	0.7	0.0	0.3	0.7	0.0	0.3	0.9	0.3	2.7	0.0	0.6
Total number of papers	23.0	20.0	21.0	23.0	31.0	18.0	30.0	24.0	31.0	21.0	18.0	22.0	28.0	38.0	18.0	18.0	37.0	21.0	18.0	31.0	25.0
Entropy	1.5	2.0	1.6	0.0	1.5	2.2	0.2	0.7	1.7	1.2	2.4	0.0	1.7	1.8	1.8	2.0	2.1	1.3	1.9	1.7	2.2

Table 1: A matrix showing the occurrences of AAAI 2013 keywords in each of the new clusters, as calculated from the Naive Bayes keyword proportion estimates described in Section 3.1. Highest-occurring keywords for each cluster shown in bold.

- The top 20 user-chosen keywords assigned to all $d_i \in z_j$;
- A matrix of the occurrences of the old (AAAI 2013) keywords in each cluster, as shown in Table 1; and
- The titles of each $d_i \in z_j$.

The authors inspected the above information associated with each cluster individually and ascribed to it the most representative keyword they could. (Mappings from the clusters to keywords can be seen below in Table 2, with cluster numbers in parentheses after each new keyword.) If a cluster did not seem to have a coherent theme, it was determined to be a "junk" cluster and was assigned no keyword. From the 21 original clusters, we ended up with 18 named clusters and three "junk" clusters (clusters 1, 6 and 11 in Table 1). The data supports the impressionistic determination of these as "junk" clusters, as they are among the highest-entropy clusters with respect to the old keywords.

The authors did their utmost to faithfully interpret the data, i.e., to assign keywords that matched the papers in each cluster. But we readily acknowledge that having a few individuals manually map clusters of article submissions to keywords is not ideal, as in some cases it required a fair amount of "reading the tea leaves." Methods have been proposed to automatically label topics (Mei, Shen, and Zhai 2007; Lau et al. 2011; Hulpus et al. 2013), but these are not yet mature enough (in our view) for the present application.

The effects of the constraint matrix can be seen in the final clustering. For example, recall that PC members strongly agreed that both *applications* and *multidisciplinary topics* should be merged; in the resulting clustering, these are indeed combined into one *application* cluster. Similarly, PC members agreed that *machine learning* ought to be multimodal (i.e., was too broad), and that topic was split here into two clusters. Furthermore, *robotics*, which individuals expressly did not want merged with any other cluster, did not. *Machine learning* and *NLP* had a consensus "merge" signal and a joint cluster indeed emerged. By contrast, there was no consensus on whether *NLP* was multimodal; as a result, the data spoke for itself and three large *NLP* clusters emerged.

[4]Reasoning about plans, processes, & actions.

5 Crowd-Sourced Evaluation

The premise of our evaluation is that new keywords are 'better' than existing keywords insofar as the community prefers them. To assess this, we solicited community input the evaluation much as we did with the constraints, though from a much larger pool of respondents. We asked the 497 members of the AAAI 2014 PC (including the SPC) to evaluate the new set of keywords, compared to the existing set, under the blind headings of "List 2" and "List 1", respectively.

Evaluation 1: We asked half of the PC to choose the list of keywords that would best help them narrow down the papers for choosing their bids. We provided them with the previous year's list of keywords (12) and our new list of keywords resulting from this experiment (18). The email was as follows.

> "As an SPC/PC member which of these two sets of keywords would be best for helping you narrow down the papers that you would like to examine in order to determine your bids?"

Evaluation 2: We asked the other half of the PC to rank the top three keywords from a provided list that they think would best represent their last submission to AAAI (or IJCAI). This list was the union of the old and new keywords for a total of 27 (due to an overlap of three keywords between the two lists). The request was as follows.

> "If you were to submit a paper on your research this year or last to AAAI (or IJCAI) which keyword(s) would you choose from this list that best represents the topic of your paper? Please choose only 3 and provide the ranking."

2013 Keywords	1	2	3	∑
Applications	5	7	15	27
Constraints & satisfiability	3	5	3	11
Heuristic search & optimiz.	11	7	5	23
Knowledge-based systems	2	8	1	11
Knowledge rep. & reasoning	5	6	10	31
Machine learning	26	23	16	65
Multiagent systems	10	16	4	30
Multidisciplinary topics	0	1	7	8
Natural-language processing	4	1	2	7
Reasoning about plans . . . [4]	3	9	5	17
Reasoning under uncertainty	7	10	7	24
Robotics	11	4	9	24
Totals:	97	97	84	278

Proposed Keywords	1	2	3	∑
Applications (21)	5	7	15	27
Cognitive modeling (19)	7	3	4	14
Computational social choice (7)	5	0	2	7
Game theory (3)	8	7	2	17
Heuristic search & optimization (9)	11	7	5	23
Humans & AI (14)	9	9	10	28
Knowledge rep. & constraint satisf. (10)	3	2	2	7
Matrix fact. & semi-supervised ML (12)	3	1	1	5
Machine learning (methods) (4)	26	23	16	65
Multiagent planning (16)	2	6	3	11
Multiagent reinforcement learning (15)	0	3	3	6
Networks (5)	2	0	6	8
NLP & knowledge representation (17)	2	2	1	5
NLP & machine learning (8)	2	4	0	6
NLP & text mining (13)	1	0	6	7
Planning methods (20)	9	9	9	27
Search & constraint satisfaction (18)	4	3	10	17
Vision & robotics (2)	2	6	1	9
Totals:	101	92	96	289

2014 Keywords
AI & the Web*
Applications
Cognitive modeling
Cognitive systems*
Computational sustainability & AI*
Game theory & economic paradigms
Heuristic search & optimization
Human computation & crowd-sourcing*
Humans & AI
Knowledge rep. & reasoning
Machine learning applications
Multiagent systems
NLP & knowledge representation
NLP & machine learning
NLP & text mining
Novel machine learning algorithms
Planning & scheduling
Reasoning under uncertainty
Robotics
Search & constraint satisfaction
Vision

Table 2: AAAI 2013 keywords; the proposed keywords (number in parentheses is the associated cluster in Table 1); and the final list of keywords as determined from the results by the 2014 program chairs. Keywords with an asterisk are from previous year's special tracks, which are part of the main track of AAAI 2014. The numbers after the 2013 keywords and the proposed keywords are the results of part of the crowd-sourced evaluation described in Section 5.

5.1 Results

Evaluation 1: There were 170 respondents, 86 of which (50.5%) chose the AAAI 2013 list and 84 (49.5%) of which chose our proposed list. We followed up with a second survey asking those who selected the AAAI 2013 list why they preferred it. The question and answers were as follows, where List 1 refers to the AAAI 2013 list and List 2 refers to our proposed list.

"What is the top reason you chose List 1 over List 2?

1. List 1 had more uniform granularity than List 2
2. The formatting of List 2 was messed up in my email
3. I thought List 1 was overall more representative of the entire field of AI
4. My first choice topic for bidding was not in List 2
5. I was familiar with List 1
6. Other (please explain)"

Of those contacted, 40 (46.5%) responded. A majority (55%) responded (1), while about a quarter (22.5%) selected (3). Of the remainder, 7.5% chose (4), 5% chose (5), and 10% chose (6), with the reason being similar to (4) or (3).

Evaluation 2: We received 160 responses; the complete breakdown of responses, including keyword counts per rank, can be seen in Table 2. 163 (36%) of the keywords chosen came exclusively from AAAI 2013 list, 174 (38.5%) came exclusively from our proposed list, and 115 (25.5%) belonged to both lists. Note that this is with the *machine learning* and *machine learning methods* merged after many respondents commented that there was no distinction.

5.2 Determining the Final List

Ultimately, keyword selection is not only about data but also inclusiveness and history. The final list, shown as "2014 Keywords" in Table 2 was selected by the 2014 program chairs. They elected to keep the most popular keywords from each of the two lists with respect to Evaluation 2 also taking into consideration the size of their corresponding clusters. For example, *NLP* garnered 7 total votes, whereas *NLP and knowledge representation*, *NLP and machine learning*, and *NLP and text mining* garnered a total of 18 votes all together and the size of each of the latter three clusters was 37, 24, and 28 respectively. Given the small size of the stand-alone NLP cluster, the program chairs chose to keep Machine Learning and NLP bundled together. They also elected not to keep keywords with few votes such as *computational social choice* and *matrix factorization and semi-supervised ML*, which were likely unpopular as PC members found them to be too narrow and elected for broader keyword choices such as *machine learning* to describe their papers. Some 2013 keywords were far more popular than the new keywords, e.g., *knowledge representation and reasoning* was a more popular than *knowledge representation and constraint satisfaction*. In addition the program chairs chose to rename some clusters, e.g., *planning methods* was renamed *planning and scheduling* and *game theory* was renamed *game theory and economic paradigms*.

6 Conclusions

We have described a first attempt at taking a data-driven, community-sourced approach to generating keywords for conferences such as AAAI. Using crowd-sourced constraints concerning last year's keywords and human-in-the-loop guidance in drawing up the results, we generated a set of keywords that are at least as popular with AAAI PC members as last year's set was, and that included some terms that were apparently overlooked in past years. Indeed, based on the results of our crowd-sourced evaluation, a number

of our keywords made their way into the AAAI 2014 set, and we are optimistic that this improved the bidding process. Thus we were able to demonstrate the value of augmenting more traditional, historically-based keyword generation with a data-driven, crowd-sourced strategy. Additionally, this case study uncovered two new challenges for class-level, constraint-based clustering: how to handle multilabel data and how to incorporate constraints when there is no expert consensus.

The approach taken here provides what we view as a promising proof of concept for a more data-driven and community-engaged process. But the methods we leveraged can certainly benefit from a number of refinements. Specifically, we would like to explore a more holistic approach using a fully generative mixture model that directly incorporates constraints. Another potential extension would be to specify a nonparametric model, thus providing a means to select the number of keywords in a more principled fashion.

This work also reveals further potentially avenues for community participation in the keyword selection process. For example, the process of deriving keyword names from the clusters of papers took two hours of expert time, which may have been insufficient. We believe we can effectively leverage the community's expertise in titling the clusters in the future to produce a more consensus-driven set of keywords.

Data

We have placed the 2013 and 2014 data in the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Note that due to privacy reasons the 2013 data includes only the accepted abstracts, titles, and keyword information from the main track, whereas the experiments in this paper were conducted on all papers submitted to the main track in 2013. The 2014 data includes the accepted abstracts titles and keyword information for all tracks (the main track and the special tracks). More information about these datasets can be found at the repository.

Acknowledgements

We would like to thank Marie desJardins, Kiri Wagstaff and Roni Khardon for providing helpful feedback on an earlier version of this paper. We thank Carol Hamilton of AAAI for assembling the data and answering questions as to the use of the keywords in 2013. Jingjing Liu provided the code for the constrained spectral clustering and AAAI 2014 cochair Peter Stone helped determine the final set of keywords we used for AAAI 2014. Finally, we thank the members of the 2014 AAAI PC for helping with the crowd sourced evaluation and in particular we wish to thank Adele Howe, Kevin Small, Robert Holte, Marie desJardins, Kiri Wagstaff, Manuela Veloso, and Michael Littman for providing the constraints found in Figure 1.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research (JMLR)* 3:993–1022.

Dhillon, I. S.; Guan, Y.; and Kulis, B. 2004. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 551–556. ACM.

Doane, D. P., and Seward, L. E. 2011. Measuring skewness: A forgotten statistic. *Journal of Statistics Education* 19(2):1–18.

Hulpus, I.; Hayes, C.; Karnstedt, M.; and Greene, D. 2013. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 465–474. ACM.

Lau, J. H.; Grieser, K.; Newman, D.; and Baldwin, T. 2011. Automatic labelling of topic models. In *ACL*, volume 2011, 1536–1545.

Mei, Q.; Shen, X.; and Zhai, C. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 490–499. ACM.

Meila, M., and Shi, J. 2001. A random walks view of spectral segmentation. In *AI and Statistics (AISTATS)*.

Milligan, G. W., and Cooper, M. C. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50(2):159–179.

Ng, A. Y.; Jordan, M. I.; Weiss, Y.; et al. 2002. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems* 2:849–856.

Preston, D. R.; Brodley, C. E.; Khardon, R.; Sulla-Menashe, D.; and Friedl, M. 2010. Redefining class definitions using constraint-based clustering: An application to remote sensing of the earth's surface. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 823–832. ACM.

Rish, I. 2001. An empirical study of the naive bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, volume 3, 41–46.

Strehl, A.; Ghosh, J.; and Mooney, R. 2000. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 58–64.

Sugar, C. A., and James, G. M. 2003. Finding the number of clusters in a dataset. *Journal of the American Statistical Association* 98(463).

Tan, C.-M.; Wang, Y.-F.; and Lee, C.-D. 2002. The use of bigrams to enhance text categorization. *Information Processing & Management* 38(4):529–546.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association* 101(476).

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416.