JHUAPL TAC-KBP2013 Slot Filler Validation System

I-Jeng Wang, Edwina Liu, Cash Costello, and Christine Piatko
Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd.
Laurel, MD 20723
i-jeng.wang@jhuapl.edu

Abstract

In this paper we present a constrained optimization approach to aggregate results from multiple slot fillers taking into account the confidence values generated by individual slot fillers. The results obtained from aggregation were used to validate the individual runs. We demonstrated that the proposed aggregation approach led to a significant performance improvement over individual runs for single-value slots.

1 Introduction

The TAC KBP Slot Filler task requires that each participant provide a confidence value associated with each slot response. Since no further information is provided regarding how the confidence values were generated, it is not clear whether this additional feature provides any utility in analyzing, predicting, or further filtering of the SF results. Results from (Tamang, Chen, and Ji, 2012) based on 2012 TAC KBP SF data suggested that confidence values are informative as system features for validation and may be used for slot-filling system combination. In this paper, we present a constrained optimization framework for aggregating the confidence values produced by 2013 TAC KBP SF submissions. The question we intend to address is, without prior information on how the confidence values were estimated, is it possible to produce a system that aggregates the outputs of individual systems to improve the performance while taking into account the confidence values. In this paper, we provide a brief summary of the aggregation approach and present initial results based on 2013 data. The preliminary experimental results obtained using 2013 SF data have been encouraging. Using a subset of TAC KBP SF submissions (48 runs in total), we demonstrated that the proposed aggregation approach led to significant performance improvement over individual runs for single-value slots. We observed a more moderate performance improvements for list-value slots as measured by the F-score.

2 Technical Approach

To take into account the confidence values provided by slot fillers, we consider the problem of aggregating probabilistic evidence. We will refer to each instance of a slot value and its associated confidence as a *probabilistic evidence*, denoted by (E, c), where

- E is a specific slot value extracted from documents by a slot filler; and
- $c \in [0, 1]$ is a nonnegative real number that represents the confidence of the slot filler on E.

Specifically, given a set of probabilistic evidence for an entity from a collection of slot fillers

$$\{(E_1, c_1(j))\}_{j=1}^{N_1}, \ldots, \{(E_M, c_1(j))\}_{j=1}^{N_M}, \quad (1)$$

where

- $E_1, \ldots, E_M$ are M distinct values produced by the slot fillers;
- N_i is the number of times the value E_i is extracted by a slot filler; and

- $c_i(j) \in [0, 1]$ denotes the j-th confidence value associated with assertion E_i produced by a slot filler.

Our goal is to aggregate these "raw" confidence values produced by individual slot fillers to arrive at a single *aggregated* confidence value $x_i \in [0, 1]$ for each slot value E_i, where $i = 1, \ldots, M$. The validation of individual slot fillers is then carried out based on the aggregated confidence values.

We accomplish aggregation by solving a constraint optimization problem:

$$\min_{0 \le x_i \le 1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} w_{ij} \left(x_i - c_i(j)\right)^2, \quad \text{s.t. } g(\boldsymbol{x}) \le 0, \tag{2}$$

where $\boldsymbol{x} \triangleq [x_1, x_2, \ldots, x_M]^T$, x_i denotes the aggregated confidence for E_i, $w_{ij} \ge 0$ is a non-negative weight assigned to each instance of $c_i(j)$, and $g(\boldsymbol{x}) \le 0$ is the constraint on the confidence values. Note that, without the constraint, the optimization defined by (2) simply reduces to independent averaging of confidences associated with each distinct slot value. The quadratic distance function used in (2) can be replaced by any strictly proper scoring function (Gneiting and Raftery, 2007) to improve the robustness, for example the Huber's function (Huber, 1964). Following the approach proposed in (Predd, Osherson, Kulkarni, and Poor, 2008; Wang, Kulkarni, and Osherson, 2011), we define the constraint $g(\boldsymbol{x}) \le 0$ such that the x_s's are "probabilistically coherent," that is, there exists a probability space where x_i is the true probability associated with the events E_i is true. A partial theoretical justification of this approach is provided in (Predd, Seiringer, Lieb, Osherson, Poor, and Kulkarni, 2009), where the theoretical connection between the probabilistic coherence of forecasts and their non-domination by rival forecasts with respect to any proper scoring rule was established.

Our validation system first aggregates the outputs of the slot fillers to arrive at a single aggregated confidence value for each distinct slot value produced by any individual slot filler. The validation is then carried out by either treating the slot value with highest aggregated confidence as the truth for single-value slots, or using the values with aggregated confidence exceeding a threshold as the correct responses for the list-value slots.

For the submitted system, we considered only constraints motivated by the mutually exclusion property of the slot values. In the following, we first describe our approach to defining the constraint $g(\boldsymbol{x}) \le 0$ and weights w_{ij} for the single-value slots in Sections 2.1. The approach is extended in Section 2.2 to address list-value slots. We describe a possible extension to incorporate a propositional relationship motivated by the slot hierarchy in Section 4.

2.1 Approach for Single-Value Slots

Consider a single-value slot associated with an entity with $E_1, \ldots, E_k$ as the possible values. Then the events that any E_i is true are mutually exclusive and the total sum of their probabilities should not exceed 1. We will abuse the notation to use $P(E_i)$ to denote the probability that E_i is the true value for the slot. Therefore the following constraint resulting from this mutual exclusion (ME) property should be incorporated into the optimization problem (2):

$$P(E_1) + P(E_2) + \cdots + P(E_k) \le 1. \tag{3}$$

The choice of weights w_{ij} in the optimization problem (2) will determine how much influence a particular probabilistic evidence will have on the aggregation. In an idealized setting where feedback on prior performance of individual slot fillers is available, appropriate weights can be derived for individual slot fillers to reflect their anticipated relative performance.[1] Each slot-filler specific weight is then assigned to all probabilistic evidences produced by the corresponding slot filler.

A straightforward way to define weights based on the feedback is to use the inverse ranking of the average precision previously achieved by individual slot fillers as the weights. Alternatively, we may derive weights based on the confidence values produced by the slot fillers to take into account the quality of confidence estimation associated with individual slot fillers. Let $\mathcal{E} = \{(E, c)\}$ denote a finite set of probabilistic evidences produced by a slot filler. Assume that feedback is provided such that we can

[1] One may further assign different weights for different slot types per each slot filler if sufficient data are available.

evaluate whether each E is true or false deterministically. Such feedback can be written as a binary valued function $f\colon \mathcal{F} \to \{0,1\}$, where $\mathcal{F}$ is the set of slot values evaluated and

$$f(E) = \begin{cases} 1, & \text{the assertion } E \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$$

Then the empirical average "penalty" for the slot filler derived from the feedback on $\mathcal{E}$ can be defined as

$$\rho \triangleq \frac{\sum_{(E,c)\in\mathcal{E}} \left(f(E) - c\right)^2}{|\mathcal{E}|}, \tag{4}$$

where $|\mathcal{E}|$ denotes the cardinality of $\mathcal{E}$. A weight for the slot filler can then be defined as $1 - \rho$. We will refer to this weight design as the quadratic penalty weights.

Since information on prior performance of the slot fillers submitted to 2013 TAC KBP SF task is not available (we do have data from 2012, but no association could be made between the slot fillers from two separate years), we resorted to "bootstrapping" to derive weights for our validation system based only on the 2013 data. Specifically, we use the results from unweighted aggregation (that is, aggregation with uniform weights) as the surrogate feedback to derive weights that were then used in the subsequent weighted aggregation.

2.2 Extension to List-Value Slots

To generalize the ME constraint to the list-value slots, we estimated the expected number of correct values for each slot from TAC KBP 2012 SF data and used it to bound the total probability in place of the unity upper bound in (3). Specifically, let $E_1, \ldots, E_n$ be the distinct slot values produced by the collection of 2012 slot fillers for a list-value slot and n_c is the number of correct values among them. We compute the average of the ratios n_c/n across all entities as an estimate of the *collective precision* for the slot type achieved by the set of slot fillers. We assume that a similar rate of correctness is achieved by the 2013 slot fillers for each slot type (same slot types are defined for both years). An upper bound on the total probability is derived by multiplying the estimated collective precision (from 2012 data) to the number of distinct slot value produced by the 2013 slot fillers.

After running the aggregation with the total probability bound, we further filter the slot values by thresholding their associated aggregated confidences. The threshold is derived from the 2012 data for each list-value slot type to minimize the average error achieved by slot values with aggregated confidence exceeding the threshold. Uniform weights were used for aggregation of list-value slots.

3 Experiment Results with TAC KBP 2013 SF Submissions

We only consider runs for which the submitted confidence values are deemed "meaningful" (as claimed by the submission); in total, 48 runs were included in the aggregation.

Table 1 summarizes the performance of the aggregated results following our approach versus the best, median, and the worst performance of individual runs in terms of precision, recall, and F-score. Note that the best (or median or the worst) performance for each metric is defined for the metric independently across all runs. That is, the three performances across each row were not achieved by a single individual run. Performance of three aggregation methods were included. Results from simple averaging (or equivalently, aggregation without constraint) are also included as the baseline. All three aggregation methods achieve significantly higher recalls than individual runs (a consequence of aggregating results from all runs) while maintaining a high precision. The overall F-scores also represent clear improvements from individual runs. The results clearly indicate that simple averaging of confidence values is not a viable approach to aggregation of slot filler outputs. Table 2 summarizes the performance for list-value slots. We observe slight increases of recall and F-score from individual runs. However, these improvements were marginal relative to what was achieved for the single-value slots.

In the following, we present more detailed analysis of the experimental results summarized in Tables 1 and 2 to provide further insight into the performance of our aggregation approach.

3.1 Analysis of Single-Value Slots

To better analyze the effect of the aggregation approach, we compare its performance against each

	Precision	Recall	F-score
Ind. (Best)	0.95	0.4765	0.607
Ind. (Median)	0.6888	0.2971	0.4097
Ind. (Worst)	0	0	0
Aggr-Uniform*	0.7729	0.7706	0.7717
Aggr-Weight-1*	0.7670	0.7647	0.7658
Aggr-Weight-2	0.7965	0.7941	0.7953
Aggr-Averaging	0.4543	0.4529	0.4536

Table 1: Summary of Overall Performance for Single-Value Slots (Aggr-Weight-1: quadratic penalty weights; Aggr-Weight-2: inverse ranking weights; * denotes submitted system).

	Precision	Recall	F-score
Ind. (Best)	0.5132	0.2703	0.2955
Ind. (Median)	0.2482	0.1329	0.1709
Ind. (Worst)	0.0323	0.007	0.0135
Aggregated	0.4499	0.3067	0.3647

Table 2: Summary of Overall Performance for List-Value Slots.

individual run *only* on the slots that the particular run attempted to fill. The analysis allows us to decouple the relative precision between the two techniques from the impact of their different coverage (recall). Figure 1 plots the ratios of the number of correct values produced by each individual run to the number of correct values achieved by the aggregation approach (with uniform weights), only for the slots that the run filled. The plot indicates that the aggregation achieves better accuracy than all but one individual run (run 18_1). Figure 2 depicts the relative coverage between the aggregation and individual runs as characterized by the ratios of the total number of slots filled. As expected, the ratios are all less than 1 since the aggregation will provide a response for a slot as long as it was filled by any individual run. In particular, the individual run 18_1 that achieves higher accuracy than the aggregation filled about half of the slots. Hence the performance achieved by the aggregation is more significant than one may perceive based on the standard precision-recall metric.

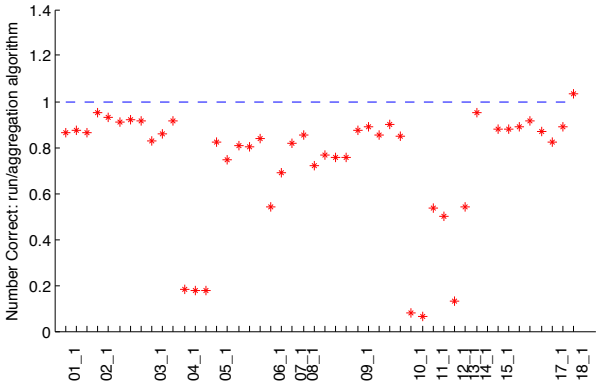

Figure 1: Relative Precision versus Individual Runs for Single-Value Slots.

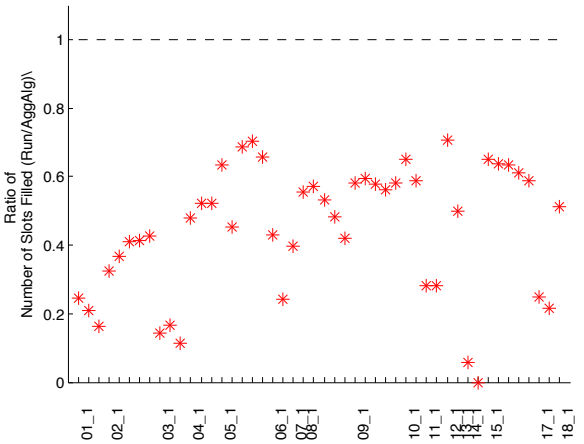

Figure 2: Relative Coverage versus Individual Runs for Single-Value Slots.

3.2 Impacts of Weight Design

Two approaches to bootstrapping weight designs were evaluated in our experiments: quadratic penalty weights and inverse ranking weights (see Section 2.1 for description of the approaches). As illustrated by Table 1, the quadratic penalty weights achieved essentially the same as the uniform weights while the inverse ranking approach resulted in minor improvements. As seen in Figure 3, the quadratic penalty weights computed by bootstrapping do not "spread out" sufficiently to have an impact on the aggregation. In contrast, the inverse ranking weights are distributed linearly and hence may have more significant effect. Since the performance of aggregation with uniform weights already achieved very

strong performance, we should not expect much further improvement from the bootstrapping-based weight designs.

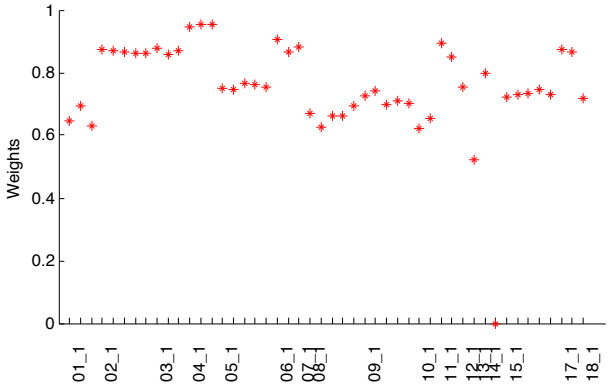

Figure 3: Quadratic Penalty Weights for Runs.

3.3 Analysis of List-Value Slots

As summarized in Table 2, the aggregation approach led to minor improvements from the individual runs for list-valued slots as measured by recall and F-score with a moderate degradation in precision. As in the case of single-value slots, we also analyze the relative precision of the aggregation method against each individual runs in Figure 4. It is clear that the aggregation approach is not as effective for the list-value slots as for their single-value counterparts as six of the individual runs achieved higher accuracy than the aggregation results. Further analysis is needed to assess whether alternative approaches to the estimation of total probability bound and/or the tuning of the filtering threshold may lead to better performance.

4 Conclusion and Future Work

In this paper we present a constrained optimization approach to aggregate outputs of a collection of slot fillers taking into account the confidence values estimated by the individual slot fillers. The approach assumes no prior information on the specific methods used by the individual slot fillers to compute confidences and does not use any other features associated with each slot value. Using a subset of TAC KBP SF submissions (48 runs in total), we demonstrated that the proposed aggregation approach led

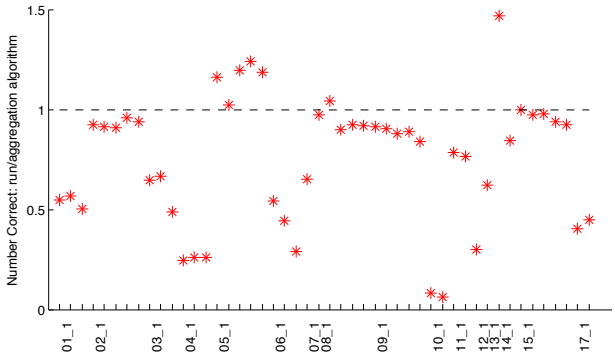

Figure 4: Relative Precision versus Individual Runs for List-Value Slots.

to significant performance improvement over individual runs for single-value slots. We observed a more moderate performance improvement for list-value slots as measured by the F-score. It is expected that the performance of any aggregation scheme will likely depend on the size and the collective quality (in terms of slot filling and confidence estimation) of the individual slot fillers included. Nevertheless, we believe that the preliminary results obtained for the 2013 data are very encouraging.

The aggregation approach discussed in the paper is a general framework that permits incorporation of prior knowledge on probabilistic dependency beyond the simple ME property used here. For example, inequality constraints can be derived according to propositional relationships motivated by known slot hierarchy (e.g., `per:city_of_birth`, versus `per:country_of_birth`).

References

S. Tamang, Z. Chen, and H. Ji, "CUNY BLENDER TAC-KBP2012 entity linking system and slot filling validation system," *Proceeding of the 5th Text Analysis Conference (TAC 2012)*, 2012.

T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, March 2007.

P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, 35, pp.73–101, 1964.

J. B. Predd, D. N. Osherson, S. R. Kulkarni, and H.V Poor. 2008. "Aggregating probabilistic forecasts from

incoherent and abstaining experts." *Decision Analysis*, vol. 5, no. 4, pp. 117–189.

G. Wang, S.R. Kulkarni, H.V. Poor, and D.N. Osherson, "Aggregating large sets of probabilistic forecast by weighted coherent adjustment," *Decision Analysis*, vol. 8, no. 2, pp. 128–14, June 2011.

J. B. Predd, R. Seiringer, E.H. Lieb, D.N. Osherson, H.V Poor, and S.R. Kulkarni, "Probabilistic coherence and proper scoring rules," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4786–4792, Oct. 2009.