RAVE: A Fast Logic-Based Answer Validator

Ingo Glöckner

Intelligent Information and Communication Systems (IICS),
FernUniversität in Hagen, 58084 Hagen, Germany,
ingo.gloeckner@fernuni-hagen.de

Abstract. RAVE (Real-time Answer Validation Engine) is a logic-based answer validator/selector designed for real-time question answering. Instead of proving a hypothesis for each answer, RAVE uses logic only for checking if a considered passage supports a correct answer at all. In this way parsing of the answers is avoided, yielding low validation/selection times. Machine learning is used for assigning local validation scores based on logical and shallow features. The subsequent aggregation of these local scores strives to be robust to duplicated information in the support passages. To achieve this, the effect of aggregation is controlled by the lexical diversity of the support passages for a given answer.

1 Description of the Validation Task

The Answer Validation Exercise (AVE) [1] introduces a test set of validation items $i \in \mathcal{I}$ comprising the question q_i, answer candidate a_i and supporting snippet s_i. Let $\mathcal{Q} = \{q_i : i \in \mathcal{I}\}$ be the set of all questions and $\mathcal{I}_q = \{i \in \mathcal{I} : q_i = q\}$ the set of validation items for a question $q \in \mathcal{Q}$. The validator must assign a validation decision $v_i \in \{\mathit{REJECTED}, \mathit{SELECTED}, \mathit{VALIDATED}\}$ and confidence score $c_i \in [0, 1]$ to each $i \in \mathcal{I}$. At most one answer per question can be selected. Answers can only be validated if an answer was selected as best answer.

2 The RAVE Validator

The input to the validator comprises a question together with answer candidates for the question and the supporting text snippets, as represented by $\mathcal{I}_q$. Let $A_q = \{a_i : q_i = q\}$ denote the set of answer candidates for q in the test set. The AVE 2008 test set is redundancy-free, i.e. for each $a \in A_q$, there is only one item $i \in \mathcal{I}_q$ supporting a. As the basis for aggregation, the IRSAW system [2] was used to actively search for additional supporting snippets for each of the answer candidates. Answers were clustered into groups of minor variants with the same 'answer key' $\kappa(a_i)$ by applying a simplification function κ (which drops accents etc.) For example, $\kappa(\textit{in the year 2001}) = 2001$. The result is a set of auxiliary items $i \in \mathcal{I}'_q$ with $q_i = q$ and $\kappa(a_i) = \kappa(a_j)$ for some original answer $a_j \in A_q$, and supporting snippet s_i for a_i found by IRSAW. The original and auxiliary validation items are joined into the enhanced validation pool $\mathcal{I}^*_q = \mathcal{I}_q \cup \mathcal{I}'_q$.

Validation starts with a *deep linguistic analysis* of the question, using the NLP toolchain of the IICS. For snippets, the parse is fetched from the pre-analyzed document collections. The *question classification* identifies the descriptive core of the question, the expected answer type and question category. *Sanity tests* eliminate trivial answers and non-informative answers to definition questions [3]. Failure of temporal restrictions and incompatibility of measurement units is also detected. For the remaining answers, *shallow features*[1] and *logic-based features* are computed. This involves proving the question from the snippet in a relaxation loop.[2] RAVE then extracts the number of proven literals and other features from the prover results; about half of all features are logic-based [3, 5].

Machine learning is used to assign a local evidence score $\eta_i \in [0,1]$ to $i \in \mathcal{I}_q^*$, estimating the probability that answer a_i is correct judging from snippet s_i [5].

Intuitively, the plausibility of an answer candidate increases when multiple passages support the answer. But multiple copies of the same snippet should not increase the aggregated score $\gamma(a)$ since they do not provide independent evidence. Hence let $K_q = \{\kappa(a_i) : i \in \mathcal{I}_q\}$ be the set of answer keys (normalized answers) for a given question. Let $\mathcal{I}_{q,k} = \{i \in \mathcal{I}_q : \kappa(a_i) = k\}$ be the set of support items for answer key $k \in K_q$. For $i \in \mathcal{I}_q$, let Ω_i be the set of term occurrences[3] in snippet s_i. For a term occurrence $\omega \in \Omega_i$, let $t(\omega)$ be the corresponding term. Let $T_i = \{t(\omega) : \omega \in \Omega_i\}$ be the set of passage terms and $T^k = \bigcup \{T_i : i \in \mathcal{I}_{q,k}\}$ the set of all terms in any passage for $k \in K_q$. We abbreviate

$$\mu(k,t) = \min\left\{(1-\eta_i)^{\nu_{i,t}} : i \in \mathcal{I}_{q,k},\ t \in T_i\right\}, \quad \nu_{i,t} = \tfrac{\mathrm{occ}(t,i)}{|\Omega_i|},$$

where $\mathrm{occ}(t,i) = |\{\omega \in \Omega_i : t(\omega) = t\}|$ is the occurrence count of term t in snippet i, and η_i is the local evidence score. The aggregated support for $k \in K_q$ is then given by $\gamma(k) = 1 - \prod_{t \in T^k} \mu(k,t)$ and for extracted answers by $\gamma(a) = \gamma(\kappa(a))$.

Consider the answer $a = \kappa(a) = 42$ as to the age of death of Elvis, supported by *"Elvis died at 42."* ($i = 1$), $\eta_1 = \frac{65}{81} \approx 0.802$ and *"Elvis (42) dead"* ($i = 2$), score $\eta_2 = \frac{37}{64} \approx 0.578$. Then $T_1 = \{\mathit{Elvis}, \mathit{died}, \mathit{at}, \mathit{42}\}$, $T_2 = \{\mathit{Elvis}, \mathit{42}, \mathit{dead}\}$, and $|\Omega_1| = 4$, $|\Omega_2| = 3$ (number of words). Now $\mathrm{occ}(t,1) = 1$ and $\nu_{1,t} = \frac{1}{4}$ for all $t \in T_1$, $\nu_{1,t} = 0$ else; similarly $\mathrm{occ}(t,2) = 1$ and $\nu_{2,t} = \frac{1}{3}$ for $t \in T_2$, $\nu_{2,t} = 0$ else. Thus $\mu(\mathit{42}, \mathit{Elvis}) = \mu(\mathit{42}, \mathit{died}) = \mu(\mathit{42}, \mathit{at}) = \mu(\mathit{42}, \mathit{42}) = (\frac{16}{81})^{\frac{1}{4}} = \frac{2}{3}$ and $\mu(\mathit{42}, \mathit{dead}) = (\frac{27}{64})^{\frac{1}{3}} = \frac{3}{4}$, i.e. $\gamma(\mathit{42}) = 1 - (\frac{2}{3})^4 \cdot \frac{3}{4} = \frac{23}{27} \approx 0.852$.

The effect of aggregation on $\gamma(a)$ is strongest if two aggregated passages have no terms in common; there is no increase at all if a passage is encountered repeatedly. After aggregation, the auxiliary support items in $\mathcal{I}_q'$ are dropped.

The final selection score depends on the aggregated score $\gamma(a_i)$ and the justification strength η_i of the snippet: $\sigma_i = \eta_i\, \gamma(a_i) / \max\{\eta_j : j \in \mathcal{I}_q, \kappa(a_i) = \kappa(a_j)\}$.[4] Based on the assignment of final selection scores σ_i, the system determines a choice of $i^{\mathrm{opt}} \in \mathcal{I}_q$ which maximizes σ_i, i.e. $\sigma_{i^{\mathrm{opt}}} = \max\{\sigma_i : i \in \mathcal{I}_q\}$. The chosen i^{opt} is marked as $v_{i^{\mathrm{opt}}} = \mathit{SELECTED}$ if $\sigma_{i^{\mathrm{opt}}} \geq \theta_{\mathrm{sel}}$, where $\theta_{\mathrm{sel}} \in [0,1]$

[1] Examples are lexical overlap and compatibility of expected/found answer types

[2] RAVE uses the same background knowledge as its predecessor MAVE [4].

[3] A term occurrence is a pair (t,i) where t is a term and i the position in the passage.

[4] Due to a bug, the submitted runs used $j \in \mathcal{I}_q^*$ instead of the intended $j \in \mathcal{I}_q$.

Table 1. Results of RAVE in the AVE 2008 (plus additional experiments)

model	f-meas	prec	recall	qa-acc	s-rate	model	f-meas	prec	recall	qa-acc	s-rate
Run1	0.39	0.33	0.49	0.32	0.61	WF.75	0.47	0.45	0.50	0.25	0.48
Run2	0.29	0.25	0.34	0.23	0.44	WF1	0.47	0.44	0.50	0.26	0.50
RF	0.45	0.43	0.48	0.26	0.50	WQ.75	0.45	0.37	0.58	0.33	0.63
RQ	0.44	0.36	0.56	0.34	0.65	WQ1	0.45	0.36	0.58	0.34	0.65

is the selection threshold; otherwise i^{opt} is marked as $v_{i^{\text{opt}}} = \textit{REJECTED}$. If $\theta_{\text{sel}} = 0$, then the best validation item for a question is always selected; this maximizes selection rate. In experiments aiming at high F-score, a threshold of $\theta_{\text{sel}} = 0.23$ was chosen.[5] The non-best items $i \in \mathcal{I}_q \backslash \{i^{\text{opt}}\}$ are classified as follows: if $v_{i^{\text{opt}}} = \textit{REJECTED}$, then $v_i = \textit{REJECTED}$ for all $i \in \mathcal{I}_q \setminus \{i^{\text{opt}}\}$ as well. If a selection has been made, i.e. $v_{i^{\text{opt}}} = \textit{SELECTED}$, then we set $v_i = \textit{VALIDATED}$ if $\sigma_i \geq \theta_{\text{val}}$ and $v_i = \textit{REJECTED}$ otherwise, where $\theta_{\text{val}} = 0.23$ is the decision threshold for non-best items. The confidence into this decision is $c_i = \sigma_i$ if $v_i = \textit{SELECTED}$ or $v_i = \textit{VALIDATED}$, and $c_i = 1 - \sigma_i$ if $v_i = \textit{REJECTED}$.

3 Evaluation

The results of RAVE in the AVE 2008 (and further reference results) are shown in Table 1, using column labels *f-meas* (F-score), *prec* (precision), *recall*, *qa-acc* (qa-accuracy, i.e. correct selections divided by number of questions), and *s-rate* (selection rate, i.e. successful selections divided by optimal selections). Since it was not clear from the QA@CLEF 2008 guidelines if full-sentence answers to definition questions would be accepted or not, *Run1* was configured to accept such answers while *Run2* was configured to reject them; obviously the former policy was the intended one. The table also lists the results for the current validator after correcting minor bugs. The letter 'R' refers to the standard method of RAVE for combining scores using σ_i, 'F' means the use of $\theta_{\text{sel}} = 0.23$ for F-score oriented runs, and 'Q' means use of $\theta_{\text{sel}} = 0$ for runs aiming at qa-accuracy.

A weighted average $\sigma_i^{(\lambda)} = \lambda\gamma(a_i) + (1-\lambda)\eta_i$ for $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$ was also tried, see WFλ and WQλ runs for the best results so obtained. Some extra experiments were carried out with the aggregation model of RAVE replaced by 'best evidence' aggregation $\gamma'(k) = \max\{\eta_i : i \in \mathcal{I}_{q,k}\}$ (this decreased F-score by 5-6% and selection rate by 4%), or 'independent evidence' aggregation $\gamma''(k) = 1 - \prod_{i \in \mathcal{I}_{q,k}}(1 - \eta_i)$ (this method was only slightly worse, but it shows a spurious increase of aggregation for duplicated passages). As to support pool enhancement, experiments using $\mathcal{I}_q$ rather than $\mathcal{I}_q^*$ showed a drop of F-score by 5% compared to RF (6% for RQ); without any aggregation, the loss was 8%.

Finally the system was run with the prover switched off: selection rate of RF then dropped by 6% (RQ: 15%), but F-score increased by up to 5%. This

[5] The threshold results from the parameters used for cost-sensitive learning.

contradicts experience from experiments on CLEF07 data [5]. A possible reason is that 8 of the 11 runs for German were produced by QA systems that used RAVE for validation. The false positives that passed the validation by RAVE as part of these QA systems are likely to stay undetected when applying the validator again in the AVE. The use of a different classifier in the shallow-only run means an independence benefit that might explain the increased F-score.

Since RAVE is designed for real-time QA, processing times are also of interest. For the RQ method, validation took 126 ms per question (9.35 ms per validation item) on a standard PC. The extra effort for applying the prover to a validation item suitable for logical processing was an average 5.4 ms.

4 Conclusion

The main objective for the current work was that of making logic-based answer validation applicable in a real-time QA context. To achieve this, RAVE uses logic only for passage validation. The observed processing times confirm that this makes a fast validation possible. Sometimes a given snippet provides useful information but may not be disclosed for licensing reasons or because the user does not understand the language of the passage. The validator therefore supports 'auxiliary' passages which contribute to aggregation but are never shown. The aggregation model of RAVE is designed to be robust to replicated content. Experiments confirm its superiority over two alternative approaches.

RAVE achieved acceptable performance in the AVE (especially considering its departure from a full-fledged RTE-style answer validation), but it was outperformed by the best individual QA system with a selection rate of 0.73 (RAVE: 0.61 in *Run1*, and 0.65 in RQ after debugging). However, as a part of real QA systems, RAVE uses features not available in the AVE test set. For example, it normally considers the retrieval score of the passage retrieval system and a producer score assigned by each QA source [2]. Experiments on the CLEF07 data show that adding these features increases the F-score by up to 8%. Moreover, RAVE is normally customized by learning separate models for each QA source, which further improves results.

References

1. Rodrigo, A., Peñas, A., Verdejo, F.: Overview of the Answer Validation Exercise 2008. This volume
2. Hartrumpf, S., Glöckner, I., Leveling, J.: Efficient question answering with question decomposition and multiple answer streams. This volume
3. Glöckner, I., Pelzer, B.: Combining logic and machine learning for answering questions. This volume
4. Glöckner, I.: Filtering and fusion of question-answering streams by robust textual inference. In: Proceedings of KRAQ'07, Hyderabad, India (2007)
5. Glöckner, I.: Towards logic-based question answering under time constraints. In: Proc. 2008 IAENG Int. Conf. on Artificial Intelligence and Applications. (2008)