Combining Logic and Aggregation for Answer Selection

Ingo Glöckner

Prakt. Informatik VII, University of Hagen, 58084 Hagen, Germany

Abstract. MAVE (Multinet-based Answer Verification) is a system for answer validation which combines logic-based techniques and aggregation for identifying the correct answers in given sets of answer candidates. The paper explains the basic concepts underlying MAVE and also presents ablation studies which reveal the contribution of the proposed methods to the achieved quality of selection.

1 Introduction

Answer selection is one of the key components of a QA system. The selection task can be described as follows: a) Start from a *validation set* of *validation items* (q, α, w) with question q, answer string α and supporting text passage or 'witness text' w. b) For *answer selection*, determine the 'best' validation item v and label v as SELECTED or REJECTED. c) For a complete *answer validation*, also mark the remaining items as VALIDATED or REJECTED. While machine learning and approaches to recognizing textual entailment are popular choices for answer validation – see [1] for an overview of the techniques used in the Answer Validation Exercise (AVE) 2007 – it is more typical of QA systems to exploit redundancy by aggregating evidence (e.g. by selecting the most frequent answers). It is not obvious how the logic-oriented techniques for answer validation and the redundancy-based techniques for selection must be combined in order to maximize selection quality. The paper describes a possible approach which proved effective in the AVE 2007. The contribution of the main techniques is also investigated.

2 System Description

This section explains the system architecture of MAVE (see [2] for details).

Deep linguistic analysis. WOCADI [3], a robust parser for German, is used for deep linguistic analysis of question, answer, and supporting text passage. This step also involves coreference resolution. The postprocessing of parsing results includes a *synonym normalization* by replacing all lexical concepts with canonical synset representatives. It is based on 48,991 synsets (synonym sets) for 111,436 lexical constants.

Hypothesis construction. Question and answer together express a *hypothesis* to be checked against the supporting text. Thus, *'At which age did Elvis die?'* and the answer *'43'* constitute the hypothesis *'Elvis died at the age of 43.'* Forming a textual hypothesis is hard for highly inflecting languages like German. MAVE avoids this problem by directly building a logical hypothesis from the logical analysis of question and answer.

Robust entailment test. The answer is logically validated by proving the hypothesis from the logical representation of the supporting text. Robustness against knowledge gaps and errors of semantic analysis is achieved by embedding the prover in a relaxation loop which repeatedly skips literals of the hypothesis until a proof of the remaining query succeeds. The skipped literal count is used as the entailment indicator. The prover utilizes 10,000 lexical-semantic facts and 109 implicative rules from AVE 2006.

Fallback entailment test. The logic-based criterion is only defined if question, answer and supporting text can be parsed; otherwise only simple lexical overlap is used as a replacement. This overlap-based fallback method uses synonym normalization and expands concepts by applying lexical-semantic relations (currently 27,814 nominalizations of verbs and 15,052 nominalizations of adjectives). Scope of the matching is restricted to sentences, with the best sentence determining the result of matching.

Assigning failure probabilities. Comparable scores are needed for ranking and aggregation, but the logic-based error counts and overlap-based counts (to which MAVE backs off when NL analysis fails) are incommensurable. The system thus switches from error counts to the probability that the answer is correct (or wrong) given the error count. The probabilities $\mathit{errProb} = P(\mathit{supported\ answer\ incorrect} | \mathit{error\ count\ of\ validation\ item} = k)$ for both types of error count are estimated from the AVE 2007 training set.

Aggregating evidence. MAVE aggregates the available evidence when several text passages support the same answer. Obviously an answer is logically justified if it is logically justified from at least one supporting text, i.e. $\mathit{multErrProb} = \prod_c \mathit{errProb}_c$ (assuming independence), where c ranges over all validation items supporting the same answer. This approach proved too optimistic when there is a lot of redundancy, though, so that MAVE now uses a more stable pragmatic criterion (with $\mathit{minErrProb} = \min_c \mathit{errProb}_c$), viz $\mathit{combinedErrProb} = (\mathit{errProb} + \mathit{multErrProb} + \mathit{minErrProb})/3$.

Determining additional quality factors. The aggregated error probability coincides for all validation items supporting the same answer. A heuristic quality factor *wnHeuristicQual* for preferences on ‘good’ supporting texts is used for tie-breaking [2]. The heuristic quality factor for answers, *awHeuristicQual*, also includes sanity criteria: a) a test for trivial answers entailed by the query: *‘Who is Gianni Versace?’ – ‘Versace’*; b) a circularity test: *‘Who is the inventor of the car?’ – ‘The inventor of the modern car’*; c) a test for other non-informative definitions, e.g. isolated nomina agentis or role terms: *‘Who is Vitali Klitschko?’ – ‘The brother.’* d) a check for mismatch of expected vs. actual answer type. *‘When did Google publish the Google Web API?’ – ‘a software’*. All of these criteria do not depend on the supporting text and are thus not aggregable.

Total validation score. The final score used in selection and validation decisions is $\mathit{validationScore} = \mathit{awHeuristicQual} \cdot (\mathit{wnQual} + \mathit{bonusWnQual})/2$, where $\mathit{wnQual} = \mathit{wnHeuristicQual} \cdot (1 - \mathit{combinedErrProb})$, and the term *bonusWnQual* denotes the maximal *wnQual* achieved by any validation item supporting the considered answer. The bonus term utilizes that there are typically few ‘unsupported’ answers compared to wrong ones. Thus correctness of an answer, as judged from the best-supporting snippet, also hints at the validity of other validation items for this answer.

Decision rules for selection and validation. MAVE uses separate thresholds for selection/rejection of the best answer and validation/rejection of remaining alternatives: a) SELECT the validation item with highest *validationScore* in the test set if $\mathit{validationScore} \geq \mathit{fSelThresh}$, REJECT otherwise. b) VALIDATE the remaining validation items in the test set if $\mathit{validationScore} \geq \mathit{fValThresh}$, REJECT otherwise. Separate thresholds make sense because factual questions often have only one correct answer. This suggests a stricter criterion for alternatives. In order to extract thresholds for optimal f-measure, MAVE chooses *fSelThresh* and *fValThresh* with *fValThresh* $\geq$ *fSelThresh* such as to maximize f-measure over the training set. For optimal selection rate, MAVE always selects the best given answer (*fSelThresh* $= 0$), and the threshold *fValThresh* for non-best answers is chosen to maximize f-measure on the training set.

Extending the scope of aggregation. Aggregation of evidence needs not be restricted to validation items which support identical answers. The *cluster method* applies a simplification function σ to the answer strings which converts to lowercase, removes accents, eliminates stopwords, etc. Scope of aggregation for the considered answer α is then extended to all validation items supporting the same answer cluster, i.e. an answer α' with $\sigma(\alpha) = \sigma(\alpha')$. This method works well for answer variants, but there is no simple extension to inclusions. In this case aggregation based on containment of cluster keys $\sigma(\alpha) \sqsubseteq \sigma(\alpha')$ will sometimes be incorrect ($\alpha =$ *'a prophet'* vs. $\alpha' =$ *'a false prophet'*). An alternative method called *evidence reassignment* (ERA) restructures the validation set by assigning each piece of evidence to all answers potentially supported by it. This process must be backed by methods for spotting answer variants and inclusions. Following the reassignment of supporting text passages, validation and aggregation of evidence is performed only for identical answers. In this way ERA copes with non-monotonic NL constructions. Consider the validation item (which should be rejected): $v = ($*'Who is Di Mambro?'*, *'a prophet'*$, w, 1)$, where $w =$ *'…self-proclaimed prophet Di Mambro…'*, and the last component $o = 1$ marks the item as non-generated. A simple method for spotting inclusions might wrongly propose $v' = ($*'Who is Di Mambro?'*, *'a false prophet'*$, w', 1)$, with $w' =$ *'…Di Mambro, the false prophet,…'*, as including v. ERA avoids the false conclusion that Di Mambro is a prophet since it forms a new item $v'' = ($*'Who is Di Mambro?'*, *'a prophet'*$, w', 0)$ replacing v'. Thus a proof that Di Mambro is a prophet fails, and v is indeed rejected.

Active enhancement of validation sets. The lack of redundancy in the AVE 2007 test set suggested actively generating redundancy by adding more supporting text passages. Thus three QA systems were run on the AVE 2007 questions (see [2]). These runs produced 12,837 answer candidates with 30,432 supporting passages, which were then searched for inclusions with respect to the AVE 2007 answers. In this way, the original test set with 282 validation items was enhanced by 2,320 auxiliary validation items.

3 Evaluation

MAVE was evaluated on the AVE 2007 test set for German [1]; see Table 1 which also lists the reference results of the current version of MAVE. Here, CF means clustering of answers and optimizing thresholds for f-measure, CQ means clustering and optimizing for qa-accuracy, EF means ERA method and optimizing for f-measure, EQ means ERA optimizing for qa-accuracy, and * marks the current results of MAVE. after further debugging. A series of ablation results is shown in Table 2. Extracting additional supporting text passages had a positive effect: The PCF run (no enhancement, optimizing f-measure) loses 5% in f-measure compared to EF* and 17% in selection rate. The PCQ run (plain validation sets, selection-oriented) also loses 3% of f-measure and 4% of selection rate compared to EQ*. The use of separate thresholds for selecting the best answer and for accepting alternatives is also effective: EJF (ERA with joint threshold, optimizing for f-measure) loses 5% of f-measure and 7% of selection rate compared to

Table 1. AVE 2007 results of MAVE and reference results of current system. The metrics shown are f-measure, precision, recall, qa-accuracy (correct selection per question) and selection rate (correct selection per question with an answer in the test set)

model	f-meas	prec	recall	qa-acc	sel-rate
CF (Run1)	0.72	0.61	0.90	0.48	0.89
CF*	0.73	0.62	0.90	0.49	0.90
CQ*	0.70	0.56	0.94	0.50	0.93
EF*	0.73	0.62	0.91	0.50	0.92
EQ (Run2)	0.68	0.54	0.94	0.50	0.93
EQ*	0.69	0.55	0.93	0.50	0.93

Table 2. Results without enhancing validation sets (PCF, PCQ), with joint thresholds for selection/validation (EJF, EJQ), without sanity checks (E–F, E–Q), without logical features (LEF, LCF, LEQ, LCQ) and without lexical-semantic relations (KCF, KCQ)

model	f-meas	prec	recall	qa-acc	sel-rate
PCF	0.68	0.61	0.76	0.41	0.75
PCQ	0.66	0.53	0.88	0.48	0.89
EJF	0.68	0.55	0.88	0.46	0.85
EJQ	0.45	0.29	0.97	0.50	0.93
E–F	0.68	0.55	0.90	0.48	0.89
E–Q	0.65	0.51	0.91	0.49	0.90

model	f-meas	prec	recall	qa-acc	sel-rate
LEF	0.72	0.59	0.94	0.49	0.90
LCF	0.72	0.59	0.93	0.48	0.89
KCF	0.56	0.44	0.78	0.42	0.77
LEQ	0.68	0.53	0.96	0.50	0.92
LCQ	0.68	0.53	0.96	0.50	0.92
KCQ	0.55	0.43	0.79	0.42	0.79

EF*. EJQ (ERA with joint threshold, optimizing for qa-accuracy) keeps a selection rate of 0.93, but suffers a loss of f-measure by 24% compared to EQ*. The sanity checks of MAVE show a positive contribution of 5% f-measure comparing E–F to EF* (or 4% comparing E–Q to EQ*). Table 2 also reveals a very small positive effect of logical inference and a strong positive effect of lexical-semantic knowledge. The success of the fallback method which lacks a structural comparison might be an artifact of the AVE 2007 setup: since the answers originate from state-of-the-art QA systems with their own tests for structural match, repeating such a test once again is no longer effective.

4 Conclusion and Future Work

The MAVE system demonstrates how answer selection rates can be boosted by combining techniques for answer validation and for leveraging redundancy. Current work aims at *real time answer validation*, i.e. managing system resources in such a way that the best answer is also found under pre-defined time constraints.

References

1. Peñas, A., Rodrigo, Á., Verdejo, F.: Overview of the answer validation exercise 2007. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (2007)
2. Glöckner, I.: University of Hagen at QA@CLEF 2007: Answer validation exercise. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (2007)
3. Hartrumpf, S.: Hybrid Disambiguation in Natural Language Analysis. Der Andere Verlag, Osnabrück, Germany (2003)