Choosing the right answer – combining answer validation and information fusion for improved answer selection

Ingo Glöckner

Intelligent Information and Communication Systems Group (IICS)
FernUniversität in Hagen, 58084 Hagen, Germany

CLEF 2007 Workshop, Budapest, 20 September 2007

Outline

1. Motivation
2. Combining validation and aggregation
3. Extending the scope of aggregation
4. Experimental results (AVE07 runs and ablation experiments)
5. Conclusion and future work

The problem of answer selection

Answer selection is one of the key components of a QA system. We assume the following description of the selection task (see AVE '07):

- Start from a **validation set** of **validation items** (q, a, w) with question q, answer string α and supporting text passage or 'witness text' w.
- For **answer selection**, determine the 'best' validation item v and label v as SELECTED or REJECTED.
- For a complete **answer validation**, also mark the remaining items as VALIDATED or REJECTED.

Different methods for answer selection:

- Redundancy based (Which answer is mentioned most often?)
- Aggregation based (Which answer shows highest joint evidence?)
- Multi-stream (Which answer was found by most QA systems?)
- RTE (Which answers are justified by the witness text?)

Starting point: The MAVE answer validator/selector

MAVE was developed for AVE06 and subsequently extended to support answer selection.
Basic system architecture of MAVE:

a) Deep linguistic analysis (construction of semantic representation, coreference resolution)
b) Normalization of synonyms
c) Constructing logical hypothesis from parse of question and answer
d) Robust entailment check by theorem prover embedded in a relaxation loop
e) Aggregation – determine joint evidence for an answer supported by several validation items
f) Computation of final validation score from aggregated evidence, false-positive tests, and answer or witness prior
g) Select/reject best answer and validate/reject remaining answers depending on selection and validation thresholds.

Deep linguistic analysis

- WOCADI (Hartrumpf 2003), a **robust parser for German**, is used for deep linguistic analysis of question, answer, and supporting text passage.
- Syntactic-semantic analysis results in a **semantic representation** expressed in the MultiNet formalism (Helbig 2006), a variant of semantic networks specifically suited for natural-language processing.
- Parser handles intrasentential **coreference resolution** and provides MAVE with the necessary information for treating intersentential anaphora.
- Postprocessing includes a **synonym normalization** by replacing all lexical concepts with canonical synset representatives. This process is based on 48,991 synsets (synonym sets) for 111,436 lexical constants.

Direct hypothesis construction

Question and answer together express a *hypothesis* to be checked against the supporting text passage:

- Question: *'At which age did Elvis Presley die?'*
- Answer: *'43.'*
- Hypothesis: *'Elvis Presley died at the age of 43.'*

Problem: Construction of hypothesis in textual form difficult for highly inflecting languages like German.
Solution: MAVE avoids construction of a textual hypothesis by directly building a logical hypothesis from the meaning representation of question and answer.
Additional benefit: Proper combination of question and answer is often hard to express linguistically (but straightforward on logical level).

Logic-based entailment test

Goal: Robustness of knowledge processing against knowledge gaps and errors of semantic analysis
Method: Embed prover in constraint relaxation loop.

- Sort literals of query according to least-effort heuristics, i.e. literals with least alternatives to be checked (highest likelihood of failure) come first
- If a proof fails, determine longest provable prefix of the ordered query and drop first non-provable literal.
- Skip literals until a proof of the remaining query succeeds. Use number of skipped literals as robust indicator of entailment strength.

Prover supported by 10,000 lexical-semantic facts (mainly specifying nominalizations), and 109 implicative rules already used in AVE ’06.

Determining entailment error levels

In practice, the logic-based entailment criterion is combined with several other **indicators of entailment errors** (see paper):

$$\begin{aligned}\texttt{synth-err-count} &= \texttt{match-err-count} \\ &+ \texttt{synth-name-conflicts} + \texttt{synth-dropped-names} \\ &+ \min(\texttt{synth-failed-literals} + \texttt{synth-msg-constraints} \\ &+ \texttt{synth-proof-facticity} + 2\,\texttt{synth-nonbound-focus} \\ &+ \texttt{synth-nonbound-vars}, 3)\end{aligned}$$

The criterion is only defined if question, answer and supporting text can be parsed; otherwise simple lexical overlap `match-err-count` is used as a replacement.

The **overlap measure** 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.

Probabilistic error model

Problem:

- Logic-based error levels and overlap-based error levels are incommensurable.
- Need **comparable validation scores** from perspective of individual validation item.
- Error levels also not directly suitable for aggregation.

Solution:

- Abstract from error levels by assigning corresponding failure probabilities.
- Interested not in error count but rather in probability that the supported answer is correct (or wrong) given this error count: P(supported answer incorrect|error-level of validation item)
- Obtain probability estimate `err-prob` for validation items from a training set (the AVE07 development set in this case).

Aggregated validation score

Motivation: Aggregate evidence when several text passages support the same answer.
Use probabilistic approach: The answer is logically justified if it is logically justified from at least one supporting text. Assuming independence,

$$\texttt{mult-err-prob} = \prod_c \texttt{err-prob}_c.$$

where c ranges over all validation items supporting the same answer (Glöckner, Hartrumpf and Leveling 2007).
$\Rightarrow$ Too optimistic when there is a lot of redundancy.
Pragmatic choice:

$$\texttt{combined-err-prob} = (\texttt{err-prob} + \texttt{mult-err-prob} + \texttt{min-err-prob})/3$$

where $\texttt{min-err-prob} = \min_c \texttt{err-prob}_c$.

Witness quality factor

Problem: Aggregated error probability coincides for all validation items supporting the same answer.
$\Rightarrow$ Use **soft preferences** on 'good' supporting text passages (Glöckner, Hartrumpf and Leveling 2007):

$$\begin{aligned}\texttt{wn-heuristic-quality} = {} & c(0.2, \texttt{wn-occurrences}) \\ & \cdot c(0.2, \texttt{wn-parse-quality}) \cdot c(0.2, \texttt{producer-score}) \\ & \cdot c(0.1, \texttt{wn-num-sentences}) \cdot c(0.1, \texttt{wn-num-chars}) \\ & \cdot c(0.3, \texttt{wn-special-chars}) \cdot c(0.2, \texttt{wn-qn-focusing}) \\ & \cdot c(0.2, \texttt{wn-relativizing-words})\end{aligned}$$

where $c(w, x) = 1 - w + wx$ is a weighting function.
These criteria are **non-aggregable** since they refer to individual validation items.

Answer quality factor

MAVE uses various sanity checks for eliminating false positives. Additional terms are used to punish overlong answers, incomplete answers (due to being too short) and non-grammatical answers. The false-positive tests and soft preferences on answer quality are combined as follows,

$$\begin{aligned}\texttt{aw-heuristic-quality} = {} & c(0.1, \texttt{aw-incompleteness}) \\ & \cdot c(0.1, \texttt{aw-overlength}) \cdot c(0.2, \texttt{aw-parse-quality}) \\ & \cdot c(1.0, \texttt{aw-not-trivial}) \cdot c(0.5, \texttt{aw-significant-def}) \\ & \cdot c(0.6, \texttt{aw-not-circular}) \cdot c(1.0, \texttt{aw-eat-fat-compat})\end{aligned}$$

based on the weighting function $c(w, x) = 1 - w + wx$.
These criteria do not depend on the supporting text passage and are therefore **not aggregable**.

False-positive tests

The sanity checks currently performed by MAVE comprise:

- Check for trivial answers entailed by the query:
  'Who is Gianni Versace?' – *'Versace'*.
- Check for circular answers with definiendum part of the definiens:
  'Who is the inventor of the car?' – *'The inventor of the modern car.'*
  'What is the Eiffel tower?' – *'The Eiffel tower in Paris.'*
- Check for non-informative answers to definition questions which contain isolated nomina agentis or role terms:
  'Who is Bill Gates?' – *'The founder.'*
  'Who is Vitali Klitschko?' – *'The brother.'*
  Recognition based on list of 2,856 nomina agentis and role terms.
- Check for mismatch of expected vs. actual answer type.
  'When did Google publish the Google Web API?' – *'a software'*
  Important since relaxation proof might skip the answer type information.

Computing the total validation score

Final validation score used for selection and validation decisions:

$$\texttt{validation-score} = \texttt{aw-heuristic-quality} \cdot (\texttt{wn-quality} + \texttt{bonus-wn-quality})/2,$$

where `wn-quality` abbreviates

$$\texttt{wn-quality} = \texttt{wn-heuristic-quality} \cdot (1 - \texttt{combined-err-prob}),$$

and `bonus-wn-quality` is the maximal `wn-quality` achieved by any validation item supporting the considered answer.

Motivation for using the bonus term:

- Compared to wrong candidates, there are typically very few 'unsupported' answers.
- Correctness of an answer (as judged from its best-supporting text passage) is therefore a strong indication for validity of the considered validation item.

Decision rules for selection and validation

Factual questions often have only one correct answer. Alternative answers should be accepted only if 'really convincing'.
$\Rightarrow$ Use **separate thresholds** for selection/rejection of best answer and validation/rejection of alternatives:

- SELECT the validation item with highest `validation-score` in the test set if `validation-score` $\geq$ `f-sel-thresh`, REJECT otherwise.
- VALIDATE the remaining validation items in the test set if `validation-score` $\geq$ `f-val-thresh`, REJECT otherwise.

Thresholds obtained from development set, depending on objective:

F — Select `f-sel-thresh` and `f-val-thresh` with `f-val-thresh` $\geq$ `f-sel-thresh` such as to maximize f-measure over the development set

Q — Use `f-sel-thresh` $= 0$ for selection in order to maximize selection rate (always selects best answer). `f-val-thresh` is then chosen to maximize f-measure on the development set.

Extending the scope of aggregation

Aggregation of evidence needs not be restricted to validation items which support identical answers.

C — **Cluster method** (Glöckner, Hartrumpf and Leveling 2007)

- Apply a simplification function σ to the answer strings which converts to lowercase, removes accents, eliminates stopwords, etc.
- Scope of aggregation for considered answer α extended to all validation items supporting the same answer cluster, i.e. an answer α' with $\sigma(\alpha) = \sigma(\alpha')$.

Good for answer variants, but no simple extension to inclusions.
Aggregation based on containment of cluster keys $\sigma(\alpha) \sqsubseteq \sigma(\alpha')$ will sometimes be incorrect ($\alpha =$ *'a prophet'* vs. $\alpha' =$ *'a false prophet'*).

E — **Evidence reassignment**

- Restructure 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-answer relationships (variants, lexical/sequential inclusion, logical entailment).
- Then validate and aggregate evidence only for identical answers.

How ERA treats non-local/non-monotonic phenomena

- Considered validation item (which should be rejected):

$$v = (\text{`Who is Di Mambro?'}, \text{`a prophet'}, w, 1),$$

  where $w =$ *'. . . self-proclaimed prophet Di Mambro. . . '*, and the last component $o = 1$ marks the item as non-generated.
- Simple method for spotting answer-answer relations based on lexical inclusion wrongly proposes

$$v' = (\text{`Who is Di Mambro?'}, \text{`a false prophet'}, w', 1),$$

  with $w' =$ *'. . . Di Mambro, the false prophet,. . . '*, as including v.
- Subsequent aggregation would thus support false conclusion that Di Mambro is a prophet.
- Reassigning evidence by building validation items solves problem:

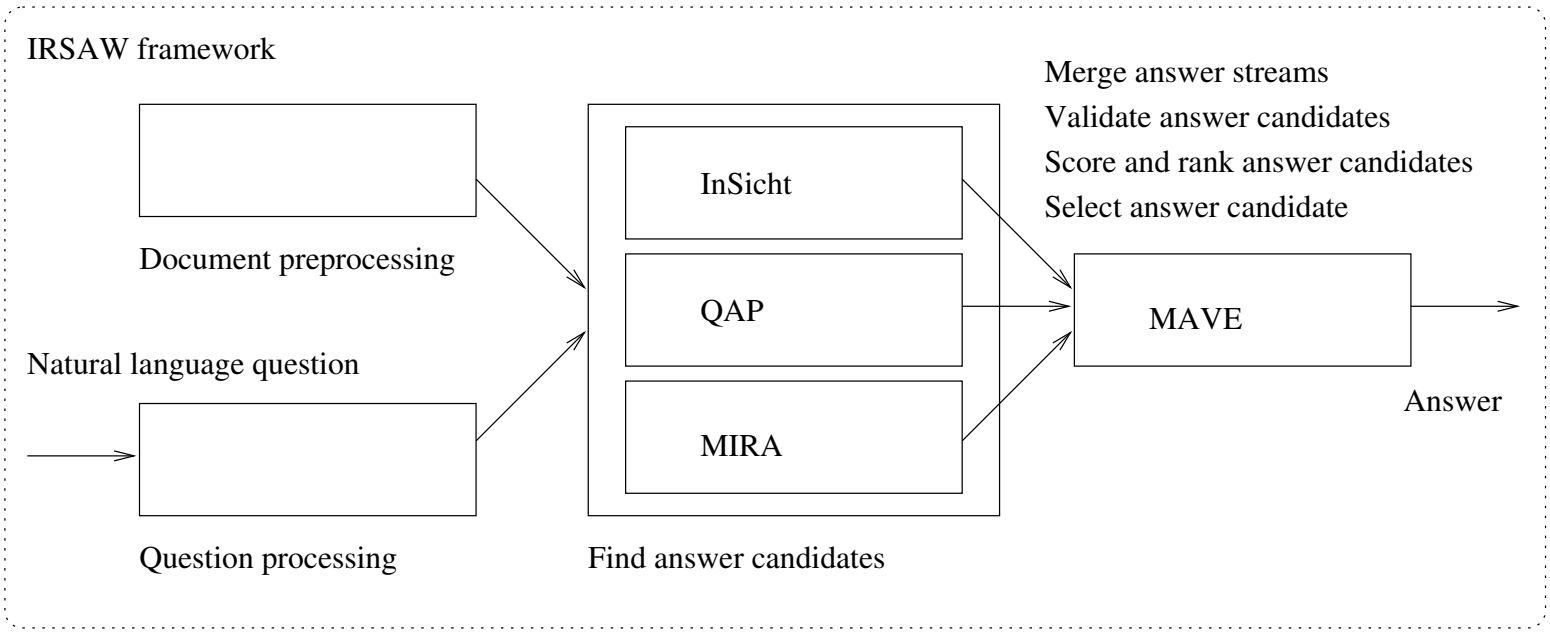
$$v'' = (\text{`Who is Di Mambro?'}, \text{`a prophet'}, w', 0).$$

  A proof that Di Mambro is a prophet now fails (assuming our coding of non-intersective adjectives), and v is indeed rejected.

Active enhancement of validation sets

Lack of redundancy in the AVE07 development and test set motivated the **active enhancement of validation sets** in MAVE:

- precondition for aggregation
- helps compensate brittleness of deep NLP approach



12,837 answer candidates with 30,432 supporting text passages searched for inclusions with respect to the AVE07 answers.

$\Rightarrow$ Test set enhanced by 2,320 **auxiliary validation items**.

AVE07 results and current performance of MAVE

The AVE07 test set for German comprises 282 answers for 113 questions (67 VALIDATED, 197 REJECTED, 18 UNKNOWN).
The random selection baseline is as high as 51.91%.

Table: Results of MAVE in the AVE07 and reference results of current system

model	f-meas	f-gain	prec	recall	qa-acc	sel-rate
CF (Run1)	0.72	0.79	0.61	0.90	0.48	0.89
CF*	0.73	0.81	0.62	0.90	0.49	0.90
CQ*	0.70	0.74	0.56	0.94	0.50	0.93
EF*	0.73	0.82	0.62	0.91	0.50	0.92
EQ (Run2)	0.68	0.69	0.54	0.94	0.50	0.93
EQ*	0.69	0.71	0.55	0.93	0.50	0.93

CF (clustering + maximize f-measure), CQ (clustering + maximize qa-accuracy), EF (ERA + maximize f-measure); EQ (ERA + maximize qa-accuracy), * (current reference results after further debugging)

Effect of enhancing the validation sets

Table: Results of MAVE without enhancement of validation sets

model	f-meas	f-gain	prec	recall	qa-acc	sel-rate
EF*	0.73	0.82	0.62	0.91	0.50	0.92
PCF	0.68	0.67	0.61	0.76	0.41	0.75
PCF+	0.68	0.68	0.60	0.78	0.42	0.77
EQ*	0.69	0.71	0.55	0.93	0.50	0.93
PCQ	0.66	0.63	0.53	0.88	0.48	0.89
PCQ+	0.67	0.66	0.54	0.90	0.48	0.89

- PCF (plain validation sets, optimizing f-measure): loses 5% in f-measure compared to EF* and 17% in selection rate.
- PCQ (plain validation sets, selection-oriented) loses 3% of f-measure and 4 percent of selection rate compared to EQ*.

$\Rightarrow$ Strong positive effect of active validation.

- PCF+ and PCQ+ use enhanced training set and plain test set.

$\Rightarrow$ Smaller training set for estimating parameters has few effect.

Effect of split thresholds for best answer and remaining alternatives

Table: Results of MAVE using a joint threshold for selection and validation

model	f-meas	f-gain	prec	recall	qa-acc	sel-rate
EF*	0.73	0.82	0.62	0.91	0.50	0.92
EJF	0.68	0.68	0.55	0.88	0.46	0.85
EQ*	0.69	0.71	0.55	0.93	0.50	0.93
EJQ	0.45	0.11	0.29	0.97	0.50	0.93

- EJF (ERA with joint threshold, optimizing for f-measure) loses 5% of f-measure and 7% of selection rate compared to EF*.
- EJQ (ERA with joint threshold, optimizing for qa-accuracy) maintains a selection rate of 0.93, but suffers a drastic loss of f-measure by 24% compared to EQ*.

$\Rightarrow$ Strong positive effect of using separate thresholds for selecting the best answer and for accepting alternative answers.

Effect of sanity checks

Table: Results of MAVE without false-positive tests

model	f-meas	f-gain	prec	recall	qa-acc	sel-rate
EF*	0.73	0.82	0.62	0.91	0.50	0.92
ESTF	0.69	0.70	0.56	0.90	0.49	0.90
E–F	0.68	0.68	0.55	0.90	0.48	0.89
EQ*	0.69	0.71	0.55	0.93	0.50	0.93
ESTQ	0.66	0.63	0.52	0.91	0.50	0.92
E–Q	0.65	0.61	0.51	0.91	0.49	0.90

- Each filter has a consistent positive but small effect (not shown)
- When deactivating all filters, f-measure drops by 5% comparing E–F to EF* (or 4% comparing E–Q to EQ*). Selection rate too drops by 3% compared to the reference.
- The ESTF and ESTQ runs show that the two variants of triviality filtering (`aw-significant-def` and `aw-not-trivial`) contribute most to the overall effect of sanity checking.

Effect of logic prover and lexical-semantic knowledge

Table: Results of MAVE without using logic-based features

model	f-meas	f-gain	prec	recall	qa-acc	sel-rate
EF*	0.73	0.82	0.62	0.91	0.50	0.92
LEF	0.72	0.79	0.59	0.94	0.49	0.90
CF*	0.73	0.81	0.62	0.90	0.49	0.90
LCF	0.72	0.78	0.59	0.93	0.48	0.89
KCF	0.56	0.39	0.44	0.78	0.42	0.77
EQ*	0.69	0.71	0.55	0.93	0.50	0.93
LEQ	0.68	0.68	0.53	0.96	0.50	0.92
CQ*	0.70	0.74	0.56	0.94	0.50	0.93
LCQ	0.68	0.69	0.53	0.96	0.50	0.92
KCQ	0.55	0.37	0.43	0.79	0.42	0.79

- L disables logic-based features, K lexical-semantic relations.
- Very small positive effect of logical inference, but strong positive effect of lexical-semantic knowledge.

Conclusion

Ablation studies explain positive results of MAVE in the AVE '07:

a) active enhancement of the validation set (separation into official and auxiliary items needed by ERA, but also useful for treating non-grammatical answers);
b) clustering of validation items for answer variants or use of the more flexible ERA method which handles inclusions of answers by reassigning evidence to all compatible validation items;
c) integration of a large repository of lexical-semantic relations;
d) use of various sanity tests for eliminating false positives;
e) use of separate thresholds for the selection of the best answer and for acceptance/rejection of the remaining alternatives
f) addition of robust fallback solutions for the logic-based features.

Success of the ablation experiment without any structural matching is intriguing: Ease of the task (random selection at 52%)? Answers in the test set already checked for structural match? Moreover two of the QA systems used for enhancing the validation set apply structure-sensitive methods.

Future work

- Improve robust inferential entailment check, e.g. by considering more relaxation paths
- Use Machine Learning rather than hand-coding weighting criteria
- Investigate real-time answer validation – manage system resources such as to find the best answer under time constraints
- Approach real logical question answering in the DFG-funded project *LogAnswer* – use inference prior to answer extraction.