# Filtering and Fusion of Question-Answering Streams by Robust Textual Inference

**Ingo Glöckner**
FernUniversität in Hagen, 58084 Hagen, Germany
iglockner@web.de

## Abstract

The paper presents an answer validation system based on robust textual inference. The system is evaluated on the AVE06 test set for German. Ablation experiments are used to determine the contribution of functional components and knowledge sources to the achieved validation quality. We demonstrate the utility of the approach for filtering answers and for the fusion of several answer sources into a single most trusted answer.
*Topics:* Answer Validation, Robust Textual Inference, Information Fusion/Integration, Evaluation.

## 1 Introduction

Question answering (QA) systems like COGEX [10], QUETAL [2] or InSicht [6] show impressive performance in answering factual questions over large corpora of unrestricted text. Though these systems use some limited form of inference, the integration of powerful reasoning capabilities into QA systems is still a challenging task, and naive application of a theorem prover to the knowledge base with several millions of facts, which is needed to describe the contents of all documents, will likely not result in acceptable performance. The QA system must thus find a trade-off between exploration of sufficiently many potentially relevant document passages, and exploitation of each potentially relevant answer candidate (e.g. by means of a logical entailment check). An apparent solution is provided by a two-stage approach: a fast high-recall stage based on simple techniques (which serves as a generator for answer candidates) coupled with a subsequent answer validation filter which analyzes the answer candidates more deeply in order to find the intended answers.

Due to the proliferation of QA techniques and implemented QA systems which complement each other, it makes sense to build multi-source QA systems like that of Ahn et al [1] which combine several question-answer streams into an integrated QA result. Again, answer validation is the key technique for selecting the correct answers from such an ensemble of question-answer streams.

The Answer Validation Exercise (AVE) of the CLEF 2006 system evaluation campaign has introduced annotated test collections and an evaluation methodology for answer validation systems [11]. The validation exercise consists in the verification of answers based on a description of the test case by a hypothesis (answer in affirmative form), supporting text snippet, and possibly the original question.

- Question: *"In welchem Jahr ereignete sich die Katastrophe von Tschernobyl?"* (En: 'In which year occurred the Chernobyl disaster?')
- Answer: *"Am 26. April 1986."* ('On April 26, 1986')
- Hypothesis: *"Die Katastrophe von Tschernobyl ereignete sich am 26. April 1986."* (En: 'The Chernobyl disaster occurred on April 26, 1986.')
- Snippet: *"In Tschernobyl ereignete sich am 26. April 1986 die grösste Katastrophe in der zivilen Nutzung der Atomenergie."* ('The worst disaster in civil use of nuclear power occurred on April 26, 1986 in Chernobyl.')
- Result: YES, confidence: 1.0.

The main part of the paper is organized as follows: It first introduces MAVE (Multinet-based Answer Verification), the best-performing[1] system in the CLEF 2006 answer validation exercise for German, and explains its use of indicators and robust textual inference for solving answer validation tasks. The system in its current form is then evaluated on the AVE 2006 test set, and the results of various experiments are reported which determine the effects of normalizing synonymous terms; of covering nominalizations; of resolving anaphoric references; of modelling the relationships between situations and subsituations; of using false-positive tests; of gathering evidence from the original question; and of the robust proving technique based on a relaxation loop for the logical prover. We demonstrate the utility of MAVE as a filter for QA streams and finally assess its suitability for fusing collections of answer candidates into a single preferred QA result.

## 2 System description

**Preprocessing** The hypotheses in the AVE06 test set result from a simple slot filling method which does not ensure proper German syntax (e.g. by adapting inflection). Consequently hypotheses are very often syntactically and semantically defective (e.g. *"im Jahre 26. April 1986"*, En: 'in the year April 26, 1986'). Snippet extraction by the QA systems

[1] Three systems took part in AVE 2006 for German, and MAVE produced the best two runs of the five submitted by all participants.

typically ignores sentence, phrase, or even word boundaries. Many snippets also contain artefacts of named entity recognizers, i.e. collapsed terms like ‘Gianni_Versace’. MAVE uses regular correction patterns for hypotheses and snippets to gain robustness against such errors in the inputs. Thus, the defective calendar date is corrected to “am 26. April 1986” (En: ‘on April 26, 1986’), and the collapsed name is expanded to ‘Gianni Versace’, restoring blank characters.

**Deep linguistic analysis** The WOCADI parser [5] is used to construct a MultiNet representation [8] of the NL input (i.e. question, hypothesis, and snippet text). WOCADI also extracts features of the text at various levels: lemmata, results of lexical analysis (alternative and actual readings), and coreference information about possible antecedents of anaphora.

**Postprocessing** The refinement of the semantic representation constructed by the parser comprises the following steps.

- Assimilation (anaphor resolution) based on the coreference assignment determined by WOCADI;
- Normalization of synonyms: Synonymy relationships between lexemes are used for replacing all lexical constants in the knowledge representation with canonical representations (e.g. `ereignen.1.1` ↦ `eintreffen.1.2` for En: ‘occur’). More than 1,000 synonymy relationships from HaGenLex plus 525 additional synonyms and near-synonyms for frequent terms in the QA@CLEF corpus are used for normalizing lexical constants in MultiNet graphs and logical axioms.
- Normalisation and simplification of the semantic representation by transformation rules, e.g. elimination of modalities like *“können”* (En: ‘can’) and *“müssen”* (En: ‘must ’) from the representation of queries. Fine-grained information about the facticity status of all discourse entities is already provided by the formalism (see [8]) but not yet used in the first prototype of MAVE.
- Construction of logical query (pattern of query literals with cardinality constraints for numeric expressions).

Example: *“Die Katastrophe von Tschernobyl ereignete sich am 26. April 1986.”* (En: ‘The Chernobyl disaster occurred on April 26, 1986’):

```
((subs X1 "katastrophe.1.1") (exp !FOCUS X1)
(temp !FOCUS X2) (subs !FOCUS "eintreffen.1.2")
(attch X3 X1) (attr X3 X4) (sub X3 "stadt.1.1")
(sub X4 "name.1.1") (val X4 "tschernobyl.0")
(attr X2 X5) (attr X2 X6) (attr X2 X7)
(val X5 (X8 (card 26))) (sub X5 "tag.1.1")
(val X6 (X9 (card 4))) (sub X6 "monat.1.1")
(val X7 (X10 (card 1986))) (sub X7 "jahr.1.1"))
```

The example shows the normalisation of `ereignen.1.1` to the synonym `eintreffen.1.2`. There are three numerical constraints, e.g. `(val X7 (X10 (card 1986)))` which enforces a cardinal value of 1986 for `X10`.

**Indicator extraction** A total of 46 indicators is computed from the parsing results of WOCADI and subsequent logical inference and matching. However, only 9 indicators were actually used in the first MAVE prototype (for details see [3]):

- `hypo-triviality` Lemma repetition rate. Helps detect trivial hypotheses like *“Gianni Versace is Gianni Versace”* where typically `hypo-triviality` $\geq 0.8$.
- `hypo-num-sentences` Sentence count for hypothesis. A value other than one indicates a parsing error.
- `hypo-collapse` Size of query pattern divided by the number of content words in the hypothesis. If `hypo-collapse` $< 0.7$, the parser has likely dropped parts of the sentence in the semantical representation.
- `hypo-missing-constraints` Numbers in the text must be expressed by numerical constraints in the generated query. The indicator counts expected constraints not found in the query pattern due to parsing errors.
- `hypo-failed-literals` The prover is embedded into a *relaxation loop*. If a proof fails, the longest provable prefix of the sorted literal list is determined, the next (failed) literal is removed from the question pattern and a new proof is attempted with the simplified query. In this way, one obtains a proof and an answer substitution as well as a (possibly empty) list of failed literals for every query. The indicator counts the number of skipped literals. 0 means provability in a strict sense.
- `question-failed-literals` In order to make MAVE more robust against syntactic errors in hypotheses, a proof of the original question is also attempted if the question is known. The indicator measures the number of non-provable question literals obtained in this case.[2]
- `question-missing-constraints` counts the missing numerical constraints in the logical query generated from the original question of the QA system.
- `hypo-delta-concepts` If a proof of the original question succeeds, the hypothesis will be compared to the answer found by MAVE in order to validate its correctness. As a prerequisite, `hypo-delta-concepts` counts the number of content words and numbers in the answer part of the hypothesis (i.e. the part of the hypothesis overlapping with the original question is excluded from counting). Example: For *“The Chernobyl disaster occurred on April 26, 1986”*, the answer part (‘hypothesis delta’) is *“on April 26, 1986”*. Three concepts or numbers are extracted for subsequent matching: `26, ("april.1.1"|4), 1986`.
- `hypo-delta-matched` The proof of the original question as determined by MAVE involves literals which contain discourse referents from the snippet representation. Each discourse entity gets a vote which it equally distributes over all sentences whose representation contains the discourse entity after assimilation. Thus, if ‘Chernobyl’ occurs in the proof of the question (as a

[2] Notice that all question-related indicators will be ignored when only hypothesis and snippet are specified.

binding of a variable of a literal) and if Chernobyl is referenced in two sentences of the snippet, then a vote of 0.5 will be added to the accumulated score of both sentences. This will be done for all discourse entities referenced in the proof of the original question, and that sentence which receives the highest accumulated vote will be chosen as the most important sentence of the snippet for answering the question. The syno-normalised lexical constants and numbers from the 'hypothesis delta' are then matched with the syno-normalised concepts and numbers in this main sentence, resulting in the number `hypo-delta-matched` of successful matches.

**Aggregation of indicators** An *indicator evaluation language* was implemented for aggregating indicators. Indicators are represented as variables which are automatically bound to the appropriate values of a given AVE test case. The language also offers means for coping with undefined values of indicators. Disregarding such undefined cases for simplicity (the exact formula is explained in [3]), the selection criterion of MAVE can be expressed as:

$$\begin{aligned}
&\texttt{accept}_{\theta} \equiv \neg\texttt{false-positive} \wedge \\
&\quad (\texttt{h-evidence} \vee \texttt{q-evidence}) \\
&\texttt{false-positive} \equiv (\texttt{hypo-triviality} \geq \texttt{0.8}) \vee \\
&\quad (\texttt{hypo-num-sentences} = \texttt{2}) \\
&\texttt{h-evidence} \equiv \neg\texttt{hypo-false-positive} \wedge \\
&\quad (\texttt{hypo-missing-constraints} \\
&\quad + \texttt{hypo-failed-literals} \leq \theta) \\
&\texttt{hypo-false-positive} \equiv (\texttt{hypo-collapse} < \texttt{0.7}) \\
&\texttt{q-evidence} \equiv \neg\texttt{question-false-positive} \wedge \\
&\quad (\texttt{question-failed-literals} \\
&\quad + \texttt{question-missing-constraints} \\
&\quad + \texttt{hypo-delta-concepts} - \texttt{hypo-delta-matched} \leq \theta) \\
&\texttt{question-false-positive} \equiv (\texttt{hypo-delta-matched} = \texttt{0})
\end{aligned}$$

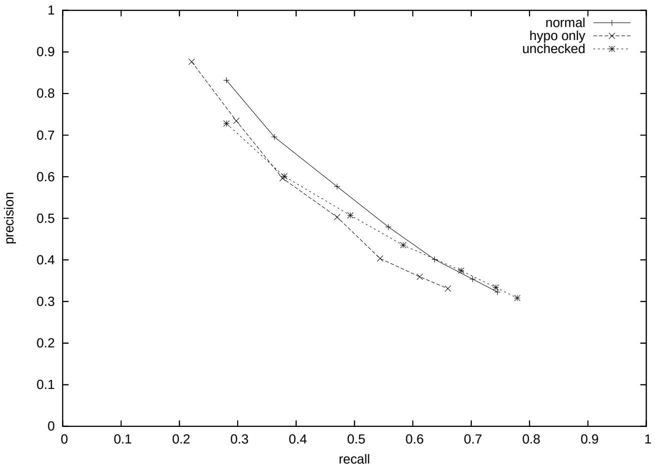
The acceptance criterion thus depends on the allowable number of differences $\theta$ from perfect provability/matching.

**Knowledge resources** The background KB of MAVE comprises 2,484 basic facts (edges of the MultiNet graph) involving 4,666 lexical constants:

- approx. 2,000 lexico-semantic relationships from HaGenLex [7] (e.g. describing nominalizations)
- 500 additional facts (e.g. tables of relationships like *'Iceland' – 'Icelandic' – 'Icelander'*)

MAVE currently uses 103 implicative rules (38 relation-defining axioms and 65 axioms involving lexical constants). The first type of axioms is concerned with the proper treatment of nominalizations and the transfer of local and temporal embedding between situations and subsituations, for example. The second type handles meta verbs like *"stattfinden"* (En: 'take place') which are frequently used for describing events. It is also utilized for a rudimentary treatment of typical QA@CLEF topics like birth and death of a person, being awarded a prize, or membership in organizations.

Figure 1: Recall-precision graph of MAVE, $\theta \in \{0,\dots,6\}$



| $\theta$ | **Precision** | **Recall** | **F measure** | **Accuracy** |
|---|---|---|---|---|
| 0 | 0.8319 | 0.2805 | 0.4195 | 0.8051 |
| 1 | 0.6957 | 0.3626 | 0.4767 | 0.8001 |
| 2 | 0.5764 | 0.4703 | 0.5179 | 0.7802 |
| 3 | 0.4793 | 0.5581 | 0.5157 | 0.7368 |
| 4 | 0.4011 | 0.6374 | 0.4923 | 0.6700 |
| 5 | 0.3538 | 0.7025 | 0.4706 | 0.6031 |
| 6 | 0.3227 | 0.7450 | 0.4503 | 0.5434 |

Table 1: Results of MAVE for the AVE-2006 test set

| $\theta$ | **Precision** | **Recall** | **F measure** | **Accuracy** |
|---|---|---|---|---|
| 0 | 0.8764 | 0.2210 | 0.3529 | 0.7966 |
| 1 | 0.7343 | 0.2975 | 0.4234 | 0.7966 |
| 2 | 0.5964 | 0.3768 | 0.4618 | 0.7795 |
| 3 | 0.5030 | 0.4703 | 0.4861 | 0.7504 |
| 4 | 0.4034 | 0.5439 | 0.4632 | 0.6835 |
| 5 | 0.3594 | 0.6119 | 0.4528 | 0.6287 |
| 6 | 0.3310 | 0.6601 | 0.4409 | 0.5797 |

Table 2: MAVE results based on hypothesis proofs only

## 3 Evaluation of answer filtering with MAVE

The AVE 2006 test set used for evaluating the MAVE system comprises 1,443 test cases (353 YES, 1,053 NO, 37 UNKNOWN). The system runs of MAVE are evaluated for:

$$\texttt{recall} = \frac{\texttt{\#computedYES}}{\texttt{\#YES}} \qquad \texttt{precision} = \frac{\texttt{\#correctYES}}{\texttt{\#computedYES}}$$

$$\texttt{f-measure} = \frac{\texttt{2recall}\cdot\texttt{precision}}{\texttt{recall}+\texttt{precision}}$$

and also for `accuracy`, i.e. relative frequency of correct decision. Fig. 1 summarizes the recall-precision scores of MAVE for different numbers of admissible mismatches (these results are slightly different from those reported in the original AVE task [3] because the MAVE system, the parser, and also the lexicon have evolved in the meantime):

- The results labeled *normal* are obtained by the `accept` criterion shown above; see Table 1 for details.

| $\theta$ | **Precision** | **Recall** | **F measure** | **Accuracy** |
|---|---|---|---|---|
| 0 | 0.7279 | 0.2805 | 0.4049 | 0.7930 |
| 1 | 0.6009 | 0.3796 | 0.4653 | 0.7809 |
| 2 | 0.5073 | 0.4929 | 0.5000 | 0.7525 |
| 3 | 0.4355 | 0.5836 | 0.4988 | 0.7055 |
| 4 | 0.3742 | 0.6827 | 0.4835 | 0.6337 |
| 5 | 0.3338 | 0.7422 | 0.4605 | 0.5633 |
| 6 | 0.3083 | 0.7790 | 0.4418 | 0.5057 |

Table 3: MAVE results without false positive tests

- *hypo only* shows the results obtained when omitting the `q-evidence` part, i.e. when not using information gathered from the original question (see Table 2).
- The *unchecked* results are obtained when dropping all false positive tests from the 'full' criterion (cf. Table 3).

The following observations can be made.

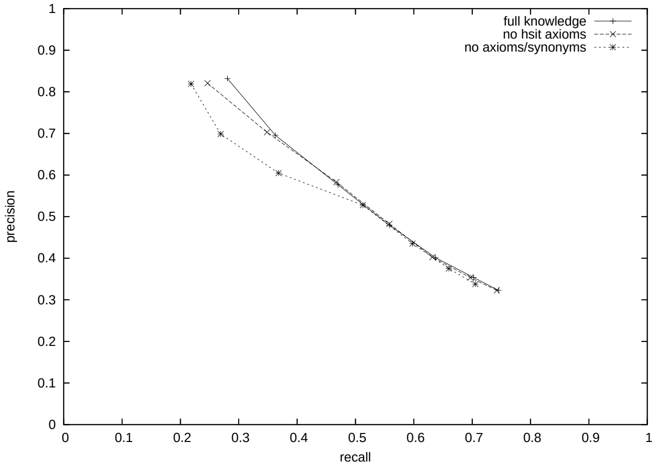
- The background knowledge of MAVE is much too sparse. Only 22.1% of hypotheses can be proved in the usual strict sense (*hypo only* with $\theta = 0$ in Table 2).
- The method for robust entailment checking, which skips non-provable literals in a relaxation loop, turns out to be very effective in making MAVE absorb a few of these knowledge gaps. For a modest number of $\theta = 1$ through $\theta = 3$ 'imperfections', there is a strong gain in recall which clearly outweighs the loss in precision.
- The sensitivity of MAVE against ill-formed hypotheses is alleviated by also using evidence gathered from the proof of the original question. Adding the `q-evidence` criterion, which does not require a parse of the hypothesis, further increases recall in the *normal* runs by up to 9.35% for $\theta = 2$. Precision decreases slightly when using question data but remains high.
- The false positive tests increase precision by up to 10.4% (for $\theta = 0$) without compromising recall of the system (see Table 3 without false positive tests vs. original results in Table 1 which are filtered by the tests).
- The best results were obtained by the normal `accept` criterion described above which uses `h-evidence`, `q-evidence`, and all false positive checks. Judging from the f-measure, the best result was obtained for $\theta = 2$. This run achieved a recall of 47.0% and a precision of 57.6% (f-measure: 0.5179).

In order to determine the relative contribution of the functional components and knowledge sources of MAVE on the achieved filter quality, further ablation experiments were carried out whose main results are summarized in Fig. 3.

- The results labeled 'full knowledge' were computed with the full knowledge base as described above.
- MAVE uses the HSIT relation to model situational composition: `(hsit H P)` means that situation `H` has a situational part `P`, or alternatively, that `P` is a *perspective* on the situation `H`. For example, the following entailment from `verleihen.1.1` (En: 'to award') to `bekommen.1.1` (En: 'get something') also expresses that getting the award is a subsituation of the awarding:

```
(subs V "verleihen.1.1")∧(ornt V X)
  ∧(obj V Y)→∃B(subs B "bekommen.1.1")
  ∧(exp B X)∧(obj B Y)∧(hsit V B).
```

Figure 2: Dependency of MAVE on knowledge sources



| $\theta$ | **Precision** | **Recall** | **F measure** | **Accuracy** |
|---|---|---|---|---|
| 0 | 0.8191 | 0.2181 | 0.3445 | 0.7916 |
| 1 | 0.6985 | 0.2691 | 0.3885 | 0.7873 |
| 2 | 0.6047 | 0.3683 | 0.4577 | 0.7809 |
| 3 | 0.5277 | 0.5127 | 0.5201 | 0.7624 |
| 4 | 0.4351 | 0.5977 | 0.5036 | 0.7041 |
| 5 | 0.3752 | 0.6601 | 0.4784 | 0.6387 |
| 6 | 0.3379 | 0.7054 | 0.4569 | 0.5789 |

Table 4: MAVE results without using axioms or synonyms

| $\theta$ | **Precision** | **Recall** | **F measure** | **Accuracy** |
|---|---|---|---|---|
| 0 | 0.8208 | 0.2465 | 0.3791 | 0.7973 |
| 1 | 0.7029 | 0.3484 | 0.4659 | 0.7994 |
| 2 | 0.5830 | 0.4674 | 0.5189 | 0.7824 |
| 3 | 0.4828 | 0.5581 | 0.5177 | 0.7390 |
| 4 | 0.4025 | 0.6317 | 0.4917 | 0.6721 |
| 5 | 0.3534 | 0.6969 | 0.4690 | 0.6038 |
| 6 | 0.3227 | 0.7422 | 0.4498 | 0.5441 |

Table 5: MAVE results without HSIT axioms

MAVE uses three basic HSIT axioms for describing the transfer of temporal embedding, spatial embedding, and circumstantials between situations and subsituations:

```
(hsit H P)∧(temp H T) → (temp P T)
(hsit H P)∧(loc H L) → (loc P L)
(hsit H P)∧(circ S P) → (circ S H).
```

Examples: If the wedding takes place on the weekend, then the marriage ceremony takes place on the weekend (transfer of temporal embedding); if the wedding takes

| $\theta$ | **Precision** | **Recall** | **F measure** | **Accuracy** |
|---|---|---|---|---|
| 0 | 0.8305 | 0.2776 | 0.4161 | 0.8044 |
| 1 | 0.6944 | 0.3541 | 0.4690 | 0.7987 |
| 2 | 0.5810 | 0.4674 | 0.5181 | 0.7817 |
| 3 | 0.4841 | 0.5609 | 0.5197 | 0.7397 |
| 4 | 0.4025 | 0.6374 | 0.4934 | 0.6714 |
| 5 | 0.3548 | 0.7025 | 0.4715 | 0.6046 |
| 6 | 0.3231 | 0.7450 | 0.4507 | 0.5441 |

Table 6: MAVE results without CHEA axioms

place in Munich, then the marriage ceremony takes place in Munich (transfer of spatial embedding); if the bride faints at the marriage ceremony, then the bride faints at the wedding (transfer of circumstantials). As shown by Table 5 (compared to baseline Table 1), the three HSIT axioms contribute up to 3.4% of the recall of MAVE.

- The results labeled 'no axioms/synonyms' are obtained in the extreme case that the knowledge base of MAVE is not used at all. (Accidentally, it made no significant difference to the results whether synonym normalization was activated in this case or not). When no axioms are used, then the proof of the logical query reduces to a direct matching of the query literals with the nodes and edges in the MultiNet graph. The robust proving technique based on the relaxation loop is very successful in this case. As shown by Table 4 for $\theta = 3$, MAVE even performs best in terms of the f-measure when no knowledge is used at all. However, this does not mean that direct matching without using any background knowledge measures up with knowledge-intensive robust proving. As we shall see below, the knowledge-based method clearly outperforms simple graph matching when it comes to the fusion of QA streams. The results of simple matching are also worse in the high precision range of the recall-precision graph.

Table 6 shows the effects of dropping the so-called CHEA axioms[3] which link the representations of verbs and their nominalizations (e.g. "abschrecken" vs. "Abschreckung", En: 'deter' vs. 'deterrence'). Compared to the baseline in Table 1, one gets a minor change of at most 0.85% (at $\theta = 1$) for the AVE test set. The effect of coreference resolution becomes visible from the ablation results in Table 7 which does not resolve referring expressions. There is only a minor contribution of coreference resolution for the AVE test, which amounts to 1.42% of recall (for $\theta = 1$) or less. The effect is weak since snippets are usually short and answers usually depend on a single sentence. On average, only 0.6 anaphoric references are resolved per snippet.

## 4 Using MAVE for answer stream fusion

Let us now consider the use of MAVE for filtering several streams of answer candidates, returning the union of all valid cases. The resulting filter will be useful (a) if the system eliminates most wrong answers (i.e. if precision of the system is high), and (b) if the recall is high enough that for each question, at least one correct answer will pass the validity test.

| $\theta$ | **Precision** | **Recall** | **F measure** | **Accuracy** |
|---|---|---|---|---|
| 0 | 0.8276 | 0.2720 | 0.4094 | 0.8030 |
| 1 | 0.6989 | 0.3484 | 0.4650 | 0.7987 |
| 2 | 0.5734 | 0.4646 | 0.5133 | 0.7788 |
| 3 | 0.4792 | 0.5552 | 0.5144 | 0.7368 |
| 4 | 0.4011 | 0.6374 | 0.4923 | 0.6700 |
| 5 | 0.3553 | 0.7025 | 0.4719 | 0.6053 |
| 6 | 0.3227 | 0.7450 | 0.4503 | 0.5434 |

Table 7: MAVE results without coreference resolution

| $\theta$ | **Precision** | **Recall** | **(Relative)** |
|---|---|---|---|
| 0 | 0.8269 | 0.215 | 0.3525 |
| 1 | 0.7397 | 0.270 | 0.4426 |
| 2 | 0.6392 | 0.310 | 0.5082 |
| 3 | 0.5965 | 0.340 | 0.5574 |
| 4 | 0.5588 | 0.380 | 0.6230 |
| 5 | 0.5135 | 0.380 | 0.6230 |
| 6 | 0.4968 | 0.385 | 0.6311 |
| 7 | 0.4785 | 0.390 | 0.6393 |
| 8 | 0.4759 | 0.395 | 0.6475 |
| 9 | 0.4706 | 0.400 | 0.6557 |

Table 8: Answer fusion based on MAVE with full information

The AVE06 test set contains 1,443 potential answers for the 200 QA@CLEF questions. Thus on average, 7 answer candidates are available for each question. The test set only contains correct answers for 122 questions, however. We thus wondered how many questions would still be answered correctly by MAVE after filtering by validity and forming the union of filtered results. It turned out that using the standard `accept` criterion, MAVE still finds at least one correct answer for 68 questions ($\theta = 2$) or 78 questions ($\theta = 3$). Simple matching without logical inference finds a correct answer for 75 of the questions (again for $\theta = 3$).

Rather than forming the union of validated answers, an alternative idea is that of selecting a single best answer from the validated choices. This method promises a tremendous reduction of data presented to the user, who is only shown the single most trusted answer, or no answer at all if no high-quality answer exists. The idea can be applied to answer streams originating from a single source or several QA systems.

For a test case $p$ (hypothesis and snippet for a given question) and the expected quality (maximal number of acceptable mismatches) $\theta \in \mathbb{N}$, we define the imprecision score of $p$ as

$$\texttt{imprec} = \min\{\theta' \in \{0,\ldots,\theta\} : \texttt{accept}_{\theta'} \text{ holds for } p\}.$$

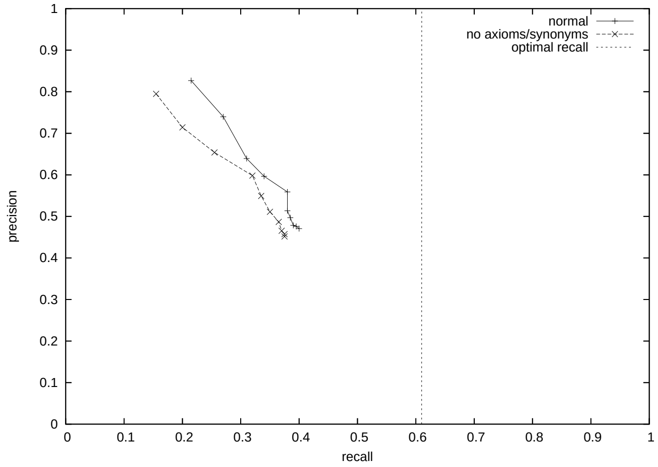
In the case that no such $\theta'$ exists, we let $\texttt{imprec} = \theta + 1$.

Now suppose that $m$ answer candidates to a question $q$ are given. The idea is that of choosing the single best answer, i.e. that answer which minimizes the `imprec` score. If the best `imprec` score for a question exceeds the given quality limit $\theta$, then the answers will be rejected (i.e. no valid answer has been found for the question); otherwise the best answer will be returned as the result of the combined system.

An experiment has been made as to the number of questions in the test set which will be correctly answered by

[3] CHEA (change from event to abstractum) is the MultiNet relation used to describe the semantic analogon of a nominalization.

Figure 3: Recall-precision graph for answer stream fusion



| $\theta$ | **Precision** | **Recall** | **(Relative)** |
|---|---|---|---|
| 0 | 0.7949 | 0.155 | 0.2541 |
| 1 | 0.7143 | 0.200 | 0.3279 |
| 2 | 0.6538 | 0.255 | 0.4180 |
| 3 | 0.5981 | 0.320 | 0.5246 |
| 4 | 0.5492 | 0.335 | 0.5492 |
| 5 | 0.5109 | 0.350 | 0.5738 |
| 6 | 0.4867 | 0.365 | 0.5984 |
| 7 | 0.4654 | 0.370 | 0.6066 |
| 8 | 0.4573 | 0.375 | 0.6148 |
| 9 | 0.4518 | 0.375 | 0.6148 |

Table 9: Answer fusion (no axioms or synomyms used)

MAVE when selecting the single best answer for each question. Fig. 3 is a summary of the evaluation results for answer stream fusion (best answer selection) with full knowledge vs. no knowledge, shown in Tables 8 and 9, respectively. Notice that 'recall' is determined with respect to the total of 200 QA@CLEF questions. Since only 122 of these were answered correctly by any of the systems in the AVE test set, recall of MAVE answer fusion is necessarily bounded by 61%. By contrast, the 'relative' recall column in the tables denotes the proportion of correctly answered questions compared to the 122 questions answered by the optimal combined system. The figure demonstrates a clear advantage of the knowledge-intensive method normally used by MAVE, as compared to direct matching without supporting background knowledge.

## 5 Conclusion

We have presented MAVE, a prototypical answer verification system for German. The basic ideas underlying MAVE, in particular the use of a relaxation loop, the idea of gathering evidence from a proof of the original question, and the tests for false positives, were shown to be useful in ablation experiments. The simple relaxation method works surprisingly well, although it only skips query literals (in the order determined by the least effort heuristic of the prover) and does not search for an optimal (smallest) number of skipped literals. A comparison to powerful graph matching methods like [9] would be instructive. The ablation tests demonstrate a positive effect of the axioms for describing transfer of situational embeddings. There is obvious potential for an extension by techniques for temporal reasoning [4]. The remaining ablation tests, i.e. running MAVE without axioms for interpreting nominalizations and without coreference resolution, show almost no effect on the performance of the system. A similar point can even be made when MAVE is run without any background knowledge at all. It thus appears that advanced methods (like coreference resolution) will only be needed as soon as the questions posed to QA systems and their logical relationship to the queried documents become more complex, too. In an experiment on using MAVE for the fusion of alternative answers by selecting the best validated answer, MAVE answered 80 questions correctly. This is a promising first result compared to the 71 correct answers found by the best individual system. The most urgent desideratum for MAVE is the integration of lexico-semantic resources like Germanet, however. A fuller account of lexico-semantic relations is one of the prerequisites for increasing recall levels of the system.

## References

[1] D. Ahn, V. Jijkoun, K. Müller, M. de Rijke, S. Schlobach, and G. Mishne. Making stone soup: Evaluating a recall-oriented multi-stream question answering system for Dutch. In *Proc. of CLEF 2004*. Springer, 2005.

[2] A. Frank, H.-U. Krieger, F. Xu, H. Uszkoreit, B. Crysmann, B. Jörg, and U. Schäfer. Querying structured knowledge sources. In *Proceedings of AAAI-05, Workshop on Question Answering in Restricted Domains*, 2005.

[3] I. Glöckner. University of Hagen at QA@CLEF 2006: Answer validation exercise. In *Working notes of the CLEF06 workshop*, Alicante, Spain, 2006.

[4] S. Harabagiu and C. Bejan. Question answering based on temporal inference. In *AAAI-2005 Workshop on Inference for Textual Question Answering*, Pittsburgh, PA, 2005.

[5] S. Hartrumpf. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany, 2003.

[6] S. Hartrumpf. Question answering using sentence parsing and semantic network matching. In *Proc. of CLEF 2004*, pages 385–392, Bath, England, Sept. 2004.

[7] S. Hartrumpf, H. Helbig, and R. Osswald. The semantically based computer lexicon HaGenLex. *Traitement automatique des langues*, 44(2):81–105, 2003.

[8] H. Helbig. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin, 2006.

[9] M. Marneffe, B. MacCartney, T. Grenager, D. Cer, A. Rafferty, and C. D. Manning. Learning to distinguish valid textual entailments. In *Proc. 2nd PASCAL Challenges Workshop*, 2006.

[10] D. Moldovan, C. Clark, S. Harabagiu, and S. Maiorano. COGEX: A logic prover for question answering. In *Proc. NAACL-HLT 2003*, 2003.

[11] A. Peñas, A. Rodrigo, V. Sama, and F. Verdejo. Overview of the answer validation exercise 2006. In *Working notes of the CLEF06 workshop*, Alicante, Spain, 2006.