Answer validation through robust logical inference

Ingo Glöckner

Intelligent Information and Communication Systems (IICS)
University of Hagen, 58084 Hagen, Germany
iglockner@web.de

Abstract. The paper features MAVE, a knowledge-based system for answer validation through deep linguistic processing and logical inference. A relaxation loop is used to determine a robust indicator of logical entailment. The system not only validates answers directly, but also gathers evidence by proving the original question and comparing results with the answer candidate. This method boosts recall by up to 13%. Additional indicators signal false positives. The filter increases precision by up to 14.5% without compromising recall of the system.

1 Introduction

MAVE (MultiNet-based answer verification), a system for validating results of question-answering (QA) systems, was developed as a testbed for robust knowledge processing in the MultiNet paradigm [1]. The goal was that of showing that deep NLP and logical inference (plus relaxation) are robust enough to handle the validation task successfully.

2 System description

2.1 Overview of MAVE

A validation candidate p submitted to MAVE consists of a hypothesis string constructed by substituting the answer of the QA system into the original question, the snippet with the relevant text passage and also of the original question. Processing of input p involves these steps: a) *Preprocessing*, i.e. application of correction rules based on regular expressions. This correction step is necessary since the mechanism for generating hypotheses often results in syntactically ill-formed hypotheses. Typical errors like *'in in Lillehammer'* are easily eliminated, however. b) The preprocessed inputs are subjected to a *deep linguistic analysis* by the WOCADI parser [2] which results in a MultiNet representation [1]. This process includes disambiguation of word meanings, semantic role labeling, a facticity labeling, and coreference resolution. c) The post-processing involves a *normalization* of the generated semantic networks. A list of about 1,000 synonyms from the computational lexicon HaGenLex [3] and additional 525 hand-crafted synonyms are used for replacing lexical concepts with a canonical synset representative. Not only the query and snippet, but also the known facts and implicative rules are automatically normalized in this way. d) Next is *query construction*, i.e. translation of the generated semantic network into a conjunction of literals. Notice that two logical queries are constructed one for the hypothesis and one for the original question. e) *Robust entailment proving* is then tried for the query constructed from the hypothesis and the query constructed from the question (see next section). f) Additional non-logical indicators are extracted, which help detect trivial answers, wrong analyses of the NLP components and other sources of error. g) The results of the logical and non-logical indicators are aggregated into a validity score from which a YES/NO decision is obtained by applying a threshold.

2.2 Robust knowledge processing in MAVE

Apart from the logical representation of the snippet, the knowledge of MAVE comprises 2,484 basic facts most of which describe lexico-semantic relationships. 103 implicative rules define key properties of the MultiNet relations and knowledge on biographic topics like birth of a person. The MAVE prover uses iterative deepening to find answers. Performance is optimized by term indexing and literal sorting under a least-effort heuristics [4]. To achieve robust logical inference, the prover is embedded in a *relaxation loop*. The prover keeps track of the longest prefix $L_1, \ldots, L_k$ of the literal list for which a proof was found. When a full proof fails, the literal L_{k+1} will be skipped and a new proof of the smaller fragment begins. By subsequently removing 'critical' literals, this process always finds a (possibly empty) query fragment provable from the given knowledge. The number of skipped literals serves as a numeric indicator of entailment robust against slight errors in representations and also against gaps in the knowledge.

2.3 Indicators and scoring function

MAVE uses the following validity indicators:

- `hypo-triviality` – *Lemma repetition ratio*. This criterion measures the ratio of lemmata which occur an even number of times in the hypothesis. Trivial hypotheses like *'Gianni Versace was Gianni Versace'* result in a high repetition score.
- `hypo-num-sentences` – *Number of sentences in the hypothesis.* The parser might wrongly consider a hypothesis as consisting of two sentences, which indicates an erroneous NL analysis.
- `hypo-collapse` – *Query extraction control index.* This indicator relates the size of the constructed query to the number of content words in the hypothesis. A small value indicates a failure of semantic analysis which produced an incomplete query.
- `hypo-lexical-focus` and `question-lexical-focus` – *Hint at potential false positive.* This criterion indicates that the logical hypothesis representation (representation of original question) is too permissive due to a systematic parser error.
- `hypo-missing-constraints` and `question-missing-constraints` – *Control index for numerical constraints.* Apart from regular words, the hypothesis (original question) can also contain numbers. These numbers must show up as restrictions in the constructed logical representations (e.g. a restriction of the year to *1994*), but are occasionally dropped by the parser, as detected by these indicators.
- `hypo-failed-literals` and `question-failed-literals` – *Number of non-provable literals.* The relaxation loop into which the prover is embedded determines

the number of non-provable literals in the hypothesis representation (representation of the original question). A value of 0 means strict logical entailment.

– `hypo-delta-concepts` – *Number of answer words in the hypothesis.* The indicator then counts the number of content-bearing tokens and in the hypothesis which stem from the original answer of the QA system and not from the question.
– `hypo-delta-matched` – *Number of matched answer words.* Based on the proof of the logical question representation over the snippet and the known origin of literals in one of the sentences of the snippet, MAVE identifies that sentence in the snippet which contains most literals supporting the proof. The content-bearing lexemes and numbers in the answer part of the hypothesis are matched against the lexical concepts and numbers in the MultiNet representation of the central sentence.

In Scheme notation, the imperfection score `imp-score`(p) of p is given by:

```
(max (* (max (>= hypo-triviality 0.8) (= hypo-num-sentences 2)) 1000)
  (min (max (* (max (< hypo-collapse 0.7) hypo-lexical-focus) 1000)
            (+ hypo-missing-constraints hypo-failed-literals))
       (max (* question-lexical-focus (= hypo-delta-matched 0) 1000)
            (+ question-failed-literals question-missing-constraints
               (- hypo-delta-concepts hypo-delta-matched)))))
```

For details of handling undefined values, see [4]. A threshold θ is used to determine YES/NO decisions, i.e. p is accepted if `imp-score`$(p) \leq \theta$ and rejected otherwise.

3 Evaluation

Two runs were submitted to AVE 2006. The recall-oriented first run allowed $\theta = 2$ mismatches for `YES` decisions. The precision-oriented second run allowed $\theta = 1$ mismatch.

Table 1 lists the results of MAVE in the answer validation exercise. 'Accuracy' is ratio of correct decision. Although the approach does not search for a minimal set of failed literals, but only follows a 'longest partial proof' strategy, the failed literal counts are significant with respect to validity of entailment: There is a decrease in precision if more imperfections θ (usually skipped literals) are allowed, but also a boost in recall. Moreover the two submitted runs represent the best compromise. Assuming perfect provability ($\theta = 0$) loses too much recall, while $\theta \geq 3$ mismatches considerably lower precision. MAVE also resorts to question proofs for validating hypotheses. This mechanism is intended to increase recall when the hypothesis is syntactically ill-formed. The hypothesis is then compared with the answer determined by the proof of the question in a way which does not assume a parse of the hypothesis. Table 2 (left) lists the results when no question-related indicators are used. For $\theta = 1$, recall decreases by 10.2 percent points. For $\theta = 2$, question indicators even contribute 13.1 percent points to the achieved recall. Precision is lowered somewhat by using the additional question data but remains on a high level. The definition of `imp-score`(p) involves false-positive indicators (like `hypo-triviality`, `hypo-collapse` or `question-lexical-focus`). Table 2 (right) shows the results of MAVE without these indicators. Precision drops by 14.5 percent points when false positive indicators are ignored, thus proving their effectiveness. Recall is not altered significantly, i.e. the criteria are also sufficiently specific.

θ **(run)**	precision	recall	f-measure	accuracy
0	0.8584	0.2820	0.4245	0.8132
1 (run #2)	0.7293	0.3837	0.5029	0.8146
2 (run #1)	0.5839	0.5058	0.5421	0.7912
3	0.4786	0.5843	0.5262	0.7429
4	0.4024	0.6831	0.5065	0.6747
5	0.3592	0.7413	0.4839	0.6136

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

θ	precision	recall	f-measure	accuracy
0	0.9452	0.2006	0.3309	0.8018
1	0.8362	0.2820	0.4217	0.8111
2	0.6898	0.3750	0.4859	0.8061
3	0.5510	0.4709	0.5078	0.7770
4	0.4450	0.5407	0.4882	0.7230
5	0.3900	0.6134	0.4768	0.6712

θ	precision	recall	f-measure	accuracy
0	0.7132	0.2820	0.4042	0.7969
1	0.6000	0.4012	0.4808	0.7884
2	0.5043	0.5116	0.5079	0.7578
3	0.4238	0.5901	0.4933	0.7038
4	0.3652	0.6890	0.4773	0.6314
5	0.3329	0.7471	0.4606	0.5724

Table 2. Results for hypothesis proofs only (left), omitting false positive tests (right)

4 Conclusion

MAVE achieved the best results for German in the AVE 2006 exercise. This success is mainly due to the robust proving strategy which tolerates a few mismatches and gaps in the coded knowledge. More ablation studies on the contribution of individual system components to the achieved filtering quality can be found in [5]. A successful application of MAVE for improving answer quality by validating and merging results in a multi-stream QA setting is described in [6].

References

1. Helbig, H.: Knowledge Representation and the Semantics of Natural Language. Springer, Berlin (2006)
2. Hartrumpf, S.: Hybrid Disambiguation in Natural Language Analysis. Der Andere Verlag, Osnabrück, Germany (2003)
3. Hartrumpf, S., Helbig, H., Osswald, R.: The semantically based computer lexicon HaGenLex. Traitement automatique des langues **44**(2) (2003) 81–105
4. Glöckner, I.: University of Hagen at QA@CLEF 2006: Answer validation exercise. In: Working Notes for the CLEF06 Workshop, Alicante, Spain (2006)
5. Glöckner, I.: Filtering and fusion of question-answering streams by robust textual inference. In: Proc. of KRAQ'07, Hyderabad/India (2007)
6. Glöckner, I., Hartrumpf, S., Leveling, J.: Logical validation, answer merging and witness selection – A study in multi-stream question answering. In: Proc. of RIAO-07. (2007)