

5 Evaluation

Both qualitative and quantitative evaluation of the integration of surface text-based and knowledge-based methods for Q/A is imposed. Quantitatively, Table 3 summarizes the scores obtained when only shallow methods were employed, in contrast with the results when knowledge-based methods were integrated. We have separately measured the effect of the integration of the knowledge-based methods at question processing and answer processing level. We have also evaluated the precision of the system when both integrations were implemented. The results were the first five answers returned within 250 bytes of text, when approximatively half million TREC documents are mined. We have used the 200 questions from TREC-8, and the correct answers provided by NIST. The performance was measured both with the NIST scoring method employed in the TREC-8 and by simply assigning a score of 1 for the question having a correct answer, regardless of its position.

	Percentage of correct answers in top 5 returns	NIST score
<i>Text-surface-based</i>	77.7%	64.5%
<i>Knowledge-based</i>	83.2%	71.5%
<i>Question Processing (only)</i>		
<i>Text-surface-based only with Answer Justification</i>	77.7%	73%
<i>Knowledge-based Question Processing with Answer Justification</i>	89.5%	84.75%

Table 3: Accuracy performance

When using the NIST scoring method to evaluate an individual answer, we used only six values: (1, .5, .33, .25, .2, 0), representing the score the answer's question obtains. If the first answer is correct, it obtains a score of 1, if the second one is correct, it is scored with .5, if the third one is correct, the score becomes .33, if the fourth is correct, the score is .25 and if the fifth one is correct, the score is .2. Otherwise, it is scored with 0. No credit is given if multiple answers are correct. Table 3 shows that both knowledge-based methods enhanced the precision, regardless of the scoring method.

To further evaluate the contribution of the justification option, we evaluated separately the precision of the prover for those questions for which the surface-text-based methods of our system, when operating alone, cannot find correct answers. We had 45 TREC-8 questions for which the evaluation of the prover was performed. Table 4 summarizes the accuracy of the prover.

	Proven correct	Proven incorrect	Precision
Incorrect answers (no knowledge)	3	210	98.5%
Correct answers (KB-based)	127	5	96.2%
Incorrect answers (KB-based)	4	38	90.04%

Table 4: Prover performance

Qualitatively, we find that the integration of knowledge-based methods is very beneficial. Table 2 illustrates the correct answer obtained with these methods, in contrast to the incorrect answer provided when only the shallow techniques are applied.

6 Conclusions

We believe that the performance of a Q/A system depends on the knowledge sources it employs. In this paper we have presented the effect of the integration of knowledge derived from question taxonomies and produced by answer justifications on the Q/A precision. Our knowledge-based methods are lightweight, since we do not generate precise semantic representations of questions or answers, but mere approximations determined by syntactic dependencies. Furthermore, our prover operates on very simple logical representations, in which syntactic and semantic ambiguities are completely ignored. Nevertheless, we have shown that these approximations are functional, since we implemented a prover that justifies answers with high precision. Similarly, our knowledge-based question processing is a mere combination of word class information and syntactic dependencies.

References

- Michael Collins. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, ACL-96*, pages 184–191, 1996.
- Christiane Fellbaum (Ed). WordNet - An Electronic Lexical Database. MIT Press, 1998.
- Jerry R. Hobbs. Discourse and Inference. Unpublished manuscript, 1986.
- Jerry R. Hobbs. Overview of the TACITUS Project. In *Computational Linguistics*, 12:(3), 1986.
- Jerry Hobbs, Mark Stickel, Doug Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63, pages 69–142, 1993.
- Wendy Lehnert. The processing of question answering. Lawrence Erlbaum Publishers, 1978.
- Dan Moldovan, Sanda Harabagiu, Marius Paşca, Rada Mihalcea, Richard Goodrum, Roxana Girju and Vasile Rus. Lasso: a tool for surfing the answer net. In *Proceedings of TREC-8*, 1999.
- Ellen Riloff and Rosie Jones. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence, AAAI-99*, 1999.