Table 3: Detailed performance of TEQUILA-enabled systems on TempQuestions and ComplexQuestions.

TempQuestions (1,271 questions)	Aggregate results			Explicit constraint			Implicit constraint			Temporal answer			Ordinal constraint		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
AQQU [6]	24.6	48.0	27.2	27.6	60.7	31.1	12.9	34.9	14.5	26.1	33.5	27.4	28.4	**57.4**	32.7
AQQU+TEQUILA	**36.0***	42.3	**36.7***	**43.8***	53.8	**44.6***	**29.1***	34.7	**29.3***	27.3*	29.6	**27.7***	**38.0***	41.3	**38.6***
QUINT [2]	27.3	**52.8**	30.0	29.3	**60.9**	32.6	25.6	**54.4**	27.0	25.2	**38.2**	27.3	21.3	54.9	26.1
QUINT+TEQUILA	33.1*	44.6	34.0*	41.8*	51.3	42.2*	13.8	43.7	15.7	**28.6***	34.5	**29.4***	37.0*	42.2	37.7*
ComplexQuestions (341 questions)	**Aggregate results**			**Explicit constraint**			**Implicit constraint**			**Temporal answer**			**Ordinal constraint**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Bao et al. [4]	34.6	48.4	35.9	41.1	53.2	41.9	26.4	36.5	**27.0**	18.6	**40.2**	22.3	31.1	**60.8**	36.1
AQQU [6]	21.5	50.0	23.3	25.0	**60.1**	28.4	11.2	31.2	11.4	19.6	35.7	19.2	22.2	54.9	25.3
AQQU+TEQUILA	**36.2***	45.9	**37.5***	**41.2***	54.7	**43.5***	**27.5***	32.6	**27.0***	29.5*	32.1	29.9*	40.2*	45.1	40.8*
QUINT [2]	22.0	**50.3**	24.5	24.7	54.7	27.5	18.8	**47.9**	19.0	16.6	37.5	20.7	20.9	51.3	26.0
QUINT+TEQUILA	29.6*	44.9	31.1*	34.6*	47.3	36.3*	12.3	42.1	13.9	**33.4***	37.5	**33.9***	**44.9***	51.6*	**45.8***

Aggregate results are averaged over the four categories. The highest value in a column for each dataset is in **bold**. An asterisk (*) indicates statistical significance of TEQUILA-enabled systems over their standalone counterparts, under the 2-tailed paired t-test at $p < 0.05$ level.

5 RELATED WORK

QA has a long tradition in IR and NLP, including benchmarking tasks in TREC, CLEF, and SemEval. This has predominantly focused on retrieving answers from textual sources. The recent TREC CAR (complex answer retrieval) resource [10], explores multi-faceted passage answers, but information needs are still simple. In IBM Watson [12], structured data played a role, but text was the main source for answers. Question decomposition was leveraged, for example, in [12, 20, 29] for QA over text. However, re-composition and reasoning over answers works very differently for textual sources [20], and are not directly applicable for KB-QA. Compositional semantics of natural language sentences has been addressed by [16] from a general linguistic perspective. Although applicable to QA, existing systems support only specific cases of composite questions.

KB-QA is a more recent trend, starting with [7, 8, 11, 24, 27]. Most methods have focused on simple questions, whose SPARQL translations contain only a single variable (and a few triple patterns for a single set of qualifying entities). For popular benchmarks like WebQuestions [7], the best performing systems use templates and grammars [1, 2, 6, 19, 29], leverage additional text [21, 26], or learn end-to-end with extensive training data [15, 26, 28]. These methods do not cope well with complex questions. Bao et al. [4] combined rules with deep learning to address a variety of complex questions.

6 CONCLUSION

Understanding the compositional semantics of complex questions is an open challenge in QA. We focused on temporal question answering over KBs, as a major step for coping with an important slice of information needs. Our method showed boosted performance on a recent benchmark, and outperformed a state-of-the-art baseline on general complex questions. Our work underlines the value of building reusable modules that improve several KB-QA systems.

REFERENCES

[1] A. Abujabal, R. Saha Roy, M. Yahya, and G. Weikum. 2018. Never-Ending Learning for Open-Domain Question Answering over Knowledge Bases . In *WWW*.

[2] A. Abujabal, M. Yahya, M. Riedewald, and G. Weikum. 2017. Automated template generation for question answering over knowledge graphs. In *WWW*.

[3] J. F. Allen. 1990. Maintaining knowledge about temporal intervals. In *Readings in qualitative reasoning about physical systems*. Elsevier.

[4] J. Bao, N. Duan, Z. Yan, M. Zhou, and T. Zhao. 2016. Constraint-based question answering with knowledge graph. In *COLING*.

[5] J. Bao, N. Duan, M. Zhou, and T. Zhao. 2014. Knowledge-based question answering as machine translation. In *ACL*.

[6] H. Bast and E. Haussmann. 2015. More Accurate Question Answering on Freebase. In *CIKM*.

[7] J. Berant, A. Chou, R. Frostig, and P. Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *EMNLP*.

[8] Q. Cai and A. Yates. 2013. Large-scale Semantic Parsing via Schema Matching and Lexicon Extension. In *ACL*.

[9] D. Diefenbach, V. Lopez, K. Singh, and P. Maret. 2017. Core techniques of question answering systems over knowledge bases: A survey. In *Knowledge and Information systems*.

[10] L. Dietz and B. Gamari. 2017. TREC CAR: A Data Set for Complex Answer Retrieval. In *TREC*.

[11] A. Fader, L. Zettlemoyer, and O. Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *KDD*.

[12] D. A. Ferrucci et al. 2012. This is Watson. In *IBM Journal of R&D*.

[13] Z. Jia, A. Abujabal, R. Saha Roy, J. Strötgen, and G. Weikum. 2018. TempQuestions: A Benchmark for Temporal Question Answering. In *HQA*.

[14] E. Kuzey, V. Setty, J. Strötgen, and G. Weikum. 2016. As Time Goes By: Comprehensive Tagging of Textual Phrases with Temporal Scopes. In *WWW*.

[15] H. Li, C. Xiong, and J. Callan. 2017. Natural Language Supported Relation Matching for Question Answering with Knowledge Graphs. In *KG4IR@SIGIR*.

[16] P. Liang, M. I. Jordan, and D. Klein. 2011. Learning Dependency-Based Compositional Semantics. In *ACL*.

[17] D. Metzler, R. Jones, F. Peng, and R. Zhang. 2009. Improving Search Relevance for Implicitly Temporal Queries. In *SIGIR*.

[18] A. Moschitti et al. 2017. Question Answering and Knowledge Graphs. In *Exploiting Linked Data and Knowledge Graphs in Large Organisations*.

[19] S. Reddy, M. Lapata, and M. Steedman. 2014. Large-scale semantic parsing without question-answer pairs. In *TACL*.

[20] E. Saquete, J. L. Vicedo, P. Martínez-Barco, R. Muñoz, and H. Llorens. 2009. Enhancing QA Systems with Complex Temporal Question Processing Capabilities. *J. Artif. Int. Res.* (2009).

[21] D. Savenkov and E. Agichtein. 2016. When a Knowledge Base Is Not Enough: Question Answering over Knowledge Bases with External Text Data. In *SIGIR*.

[22] A. Setzer. 2002. *Temporal information in Newswire articles: An annotation scheme and corpus study*. Ph.D. Dissertation. University of Sheffield.

[23] J. Strötgen and M. Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *SemEval*.

[24] C. Unger, L. Bühmann, J. Lehmann, A. N. Ngomo, D. Gerber, and P. Cimiano. 2012. Template-based question answering over RDF data. In *WWW*.

[25] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. 2016. Towards universal paraphrastic sentence embeddings. (2016).

[26] K. Xu, S. Reddy, Y. Feng, S. Huang, and D. Zhao. 2016. Question answering on freebase via relation extraction and textual evidence. *ACL*.

[27] M. Yahya, K. Berberich, S. Elbassuoni, M. Ramanath, V. Tresp, and G. Weikum. 2012. Natural language questions for the web of data. In *EMNLP*.

[28] W. Yih, M. Chang, X. He, and J. Gao. 2015. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In *ACL*.

[29] P. Yin, N. Duan, B. Kao, J. Bao, and M. Zhou. 2015. Answering Questions with Complex Semantic Constraints on Open Knowledge Bases. In *CIKM*.