Analysis of Wikipedia-based Corpora for Question Answering

Tomasz Jurczyk[1], **Amit Deshmane**[2], **Jinho D. Choi**[1]

[1]Mathematics and Computer Science, Emory University
[2]Infosys Ltd.
{*tomasz.jurczyk,jinho.choi*}*@emory.edu*
amit_deshmane@infosys.com

Abstract

This paper gives comprehensive analyses of corpora based on Wikipedia for several tasks in question answering. Four recent corpora are collected, WIKIQA, SELQA, SQUAD, and INFOBOXQA, and first analyzed intrinsically by contextual similarities, question types, and answer categories. These corpora are then analyzed extrinsically by three question answering tasks, answer retrieval, selection, and triggering. An indexing-based method for the creation of a silver-standard dataset for answer retrieval using the entire Wikipedia is also presented. Our analysis shows the uniqueness of these corpora and suggests a better use of them for statistical question answering learning.

1 Introduction

Question answering (QA) has been a blooming research field for the last decade. Selection-based QA implies a family of tasks that find answer contexts from large data given questions in natural language. Three tasks have been proposed for selection-based QA. Given a document, *answer extraction* (Shen and Klakow, 2006; Sultan et al., 2016) finds answer phrases whereas *answer selection* (Wang et al., 2007; Yih et al., 2013; Yu et al., 2014; Wang et al., 2016) and *answer triggering* (Yang et al., 2015; Jurczyk et al., 2016) find answer sentences instead, although the presence of the answer context is not assumed within the provided document for answer triggering but it is for the other two tasks. Recently, various QA tasks that are not selection-based have been proposed (Reddy and Bandyopadhyay, 2006; Hosseini et al., 2014; Jauhar et al., 2016; Sachan et al., 2016); however, selection-based QA remains still important because of its practical value to real applications (e.g., IBM Watson, MIT START).

Several datasets have been released for selection-based QA. Wang et al. (2007) created the QASENT dataset consisting of 277 questions, which has been widely used for benchmarking the answer selection task. Feng et al. (2015) presented INSURANCEQA comprising 16K+ questions on insurance contexts. Yang et al. (2015) introduced WIKIQA for answer selection and triggering. Jurczyk et al. (2016) created SELQA for large real-scale answer triggering. Rajpurkar et al. (2016) presented SQUAD for answer extraction and selection as well as for reading comprehension. Finally, Morales et al. (2016) provided INFOBOXQA for answer selection.

These corpora make it possible to evaluate the robustness of statistical question answering learning. Although all of these corpora target on selection-based QA, they are designed for different purposes such that it is important to understand the nature of these corpora so a better use of them can be made. In this paper, we make both intrinsic and extrinsic analyses of four latest corpora based on Wikipedia, WIKIQA, SELQA, SQUAD, and INFOBOXQA. We first give a thorough intrinsic analysis regarding contextual similarities, question types, and answer categories (Section 2). We then map questions in all corpora to the current version of English Wikipedia and benchmark another selection-based QA task, answer retrieval (Section 3). Finally, we present an extrinsic analysis through a set of experiments cross-testing these corpora using a convolutional neural network architecture (Section 4).[1]

2 Intrinsic Analysis

Four publicly available corpora are selected for our analysis. These corpora are based on Wikipedia, so more comparable than the others, and have already been used for the evaluation of several QA systems.

[1] All our resources are publicly available : `anonymous_url`

	WikiQA	SelQA	SQuAD	InfoboxQA
Source	Bing search queries	Crowdsourced	Crowdsourced	Crowdsourced
Year	2015	2016	2016	2016
(AE, AS, AT)	(O, O, O)	(X, O, O)	(O, O, X)	(X, O, X)
$(q, c, c/q)$	(1,242, 12,153, 9.79)	(7,904, 95,250, 12.05)	(**98,202**, 496,167, 5.05)	(15,271, 271,038, **17.75**)
(w, t)	(386,440, 30,191)	(3,469,015, 44,099)	(19,445,863, **115,092**)	(5,034,625, 8,323)
(μ_q, μ_c)	(6.44, 25.36)	(11.11, 25.31)	(11.33, 27.86)	(9.35, 9.22)
$(\Omega_q, \Omega_a, \Omega_f)$	(46.72, **11.05**, 16.96)	(32.79, 16.98, 20.19)	(32.27, 12.15, **16.54**)	(**26.80**, 35.70, 28.09)

Table 1: Comparisons between the four corpora for answer selection. Note that both WikiQA and SelQA provide separate annotation for answer triggering, which is not shown in this table. The SQuAD column shows statistics excluding the evaluation set, which is not publicly available. AE/AS/AT: annotation for answer extraction/selection/triggering, q/c: # of questions/answer candidates, w/t: # of tokens/token types, $\mu_{q/c}$: average length of questions/answer candidates, $\Omega_{q/a}$: macro average in % of overlapping words between question-answer pairs normalized by the questions/answers lengths, Ω_f: $(2 \cdot \Omega_q \cdot \Omega_a)/(\Omega_q + \Omega_a)$.

WikiQA (Yang et al., 2015) comprises questions selected from the Bing search queries, where user click data give the questions and their corresponding Wikipedia articles. The abstracts of these articles are then extracted to create answer candidates. The assumption is made that if many queries lead to the same article, it must contain the answer context; however, this assumption fails for some occasions, which makes this dataset more challenging. Since the existence of answer contexts is not guaranteed in this task, it is called answer triggering instead of answer selection.

SelQA (Jurczyk et al., 2016) is a product of five annotation tasks through crowdsourcing. It consists of about 8K questions where a half of the questions are paraphrased from the other half, aiming to reduce contextual similarities between questions and answers. Each question is associated with a section in Wikipedia where the answer context is guaranteed, and also with five sections selected from the entire Wikipedia where the selection is made by the Lucene search engine. This second dataset does not assume the existence of the answer context, so can be used for the evaluation of answer triggering.

SQuAD (Rajpurkar et al., 2016) presents 107K+ crowdsourced questions on 536 Wikipedia articles, where the answer contexts are guaranteed to exist within the provided paragraph. It contains annotation of answer phrases as well as the pointers to the sentences including the answer phrases; thus, it can be used for both answer extraction and selection. This corpus also provides human accuracy on those questions, setting up a reasonable upper bound for machines. To avoid overfitting, the evaluation set is not publicly available although system outputs can be evaluated by their provided script.

InfoboxQA (Morales et al., 2016) gives 15K+ questions based on the infoboxes from 150 articles in Wikipedia. Each question is crowdsourced and associated with an infobox, where each line of the infobox is considered an answer candidate. This corpus emphasizes the gravity of infoboxes, which summary arguably the most commonly asked information about those articles. Although the nature of this corpus is different from the others, it can also be used to evaluate answer selection.

Analysis

All corpora provide datasets/splits for answer selection, whereas only (WikiQA, SQuAD) and (WikiQA, SelQA) provide datasets for answer extraction and answer triggering, respectively. SQuAD is much larger in size although questions in this corpus are often paraphrased multiple times. On the contrary, SQuAD's average candidates per question (c/q) is the smallest because SQuAD extracts answer candidates from paragraphs whereas the others extract them from sections or infoboxes that consist of bigger contexts. Although InfoboxQA is larger than WikiQA or SelQA, the number of token types (t) in InfoboxQA is smaller than those two, due to the repetitive nature of infoboxes.

All corpora show similar average answer candidate lengths (μ_c), except for InfoboxQA where each line in the infobox is considered a candidate. SelQA and SQuAD show similar average question lengths (μ_q) because of the similarity between their annotation schemes. It is not surprising that WikiQA's average question length is the smallest, considering their questions are taken from search queries. InfoboxQA's average question length is relatively small, due to the restricted information that can be asked from the infoboxes. InfoboxQA and WikiQA show the least question-answer word overlaps over questions and answers (Ω_q and Ω_a in Table 1), respectively. In terms of the F1-score for overlapping words (Ω_f), SQuAD gives the least portion of overlaps between question-answer pairs

	WIKIQA	SELQA	SQUAD
$(\rho, \gamma_c, \gamma_p), t \geq 0.3$	(92.00, 1,203, 96.86)	(90.00, 7,446, 94.28)	(100.00, 93,928, 95.61)
$(\rho, \gamma_c, \gamma_p), t \geq \mathbf{0.4}$	(**94.00**, **1,139**, **91.71**)	(**94.00**, **7,133**, **90.31**)	(**100.00**, **93,928**, **95.61**)
$(\rho, \gamma_c, \gamma_p), t \geq 0.5$	(100.00, 1,051, 84.62)	(98.00, 6,870, 86.98)	(100.00, 93,928, 95.61)
$k = (1, \mathbf{5}, 10, 20)$	$(4.39, \mathbf{12.47}, 16.59, 22.39)$	$(20.01, \mathbf{34.07}, 40.29, 46.40)$	$(19.90, \mathbf{35.08}, 40.96, 46.74)$

Table 2: Statistics of the silver-standard dataset (first three rows) and the accuracies of answer retrieval in % (last row). ρ: robustness of the silver-standard in %, $\gamma_{c/p}$: #/% of retrieved silver-standard passages (coverage).

although WIKIQA comes very close.

Fig. 1 shows the distributions of seven question types grouped deterministically from the lexicons. Although these corpora have been independently developed, a general trend is found, where the *what* question type dominates, followed by *how* and *who*, followed by *when* and *where*, and so on.

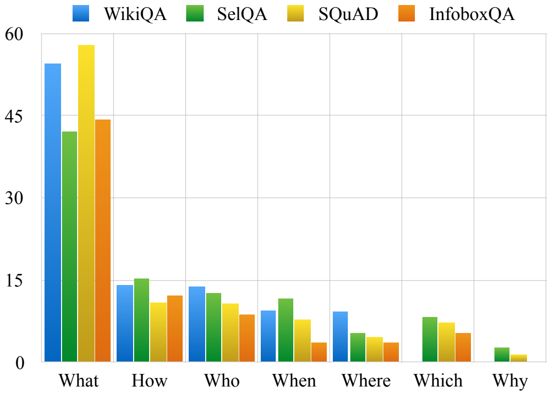

Figure 1: Distributions of question types in %.

Fig. 2 shows the distributions of answer categories automatically classified by our Convolutional Neural Network model trained on the data distributed by Li and Roth (2002).[2] Interestingly, each corpus focuses on different categories, *Numeric* for WIKIQA and SELQA, *Entity* for SQUAD, and *Person* for INFOBOXQA, which gives enough diversities for statistical learning to build robust models.

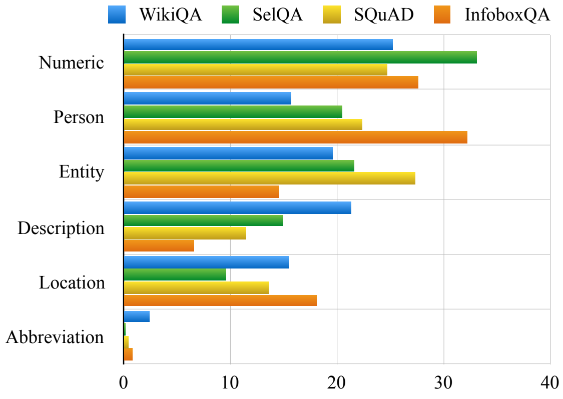

Figure 2: Distributions of answer categories in %.

3 Answer Retrieval

This section describes another selection-based QA task, called *answer retrieval*, that finds the answer context from a larger dataset, the entire Wikipedia. SQUAD provides no mapping of the answer contexts to Wikipedia, whereas WIKIQA and SELQA provide mappings; however, their data do not come from the same version of Wikipedia. We propose an automatic way of mapping the answer contexts from all corpora to the same version of Wikipeda[3] so they can be coherently used for answer retrieval.

Each paragraph in Wikipedia is first indexed by Lucene using {1,2,3}-grams, where the paragraphs are separated by WikiExtractor[4] and segmented by NLP4J[5] (28.7M+ paragraphs are indexed). Each answer sentence from the corpora in Table 2 is then queried to Lucene, and the top-5 ranked paragraphs are retrieved. The cosine similarity between each sentence in these paragraphs and the answer sentence is measured for n-grams, say $n_{1,2,3}$. A weight is assigned to each n-gram score, say $\lambda_{1,2,3}$, and the weighted sum is measured: $t = \sum_{i=1}^{3} \lambda_i \cdot n_i$. The fixed weights of $\lambda_{1,2,3} = (0.25, 0.35, 0.4)$ are used for our experiments, which can be improved.

If there exists a sentence whose $t \geq \theta$, the paragraph consisting of that sentence is considered the silver-standard answer passage. Table 2 shows how robust these silver-standard passages are based on human judgement (ρ) and how many passages are collected (γ) for $\theta = [0.3, 0.5]$, where the human judgement is performed on 50 random samples for each case. For answer retrieval, a dataset is created by $\theta = 0.4$, which gives $\rho \geq 94\%$ accuracy and $\gamma_p > 90\%$ coverage, respectively.[6] Finally, each question is queried to Lucene and the top-k paragraphs are retrieved from the entire Wikipedia. If the answer sentence exists within those retrieved paragraphs according to the silver-standard, it is considered correct.

[2] Our CNN model shows 95.20% accuracy on their test set.

[3] enwiki-20160820-pages-articles.xml.bz2

[4] github.com/attardi/wikiextractor

[5] github.com/emorynlp/nlp4j

[6] SQUAD mapping was easier than the others because it was based on a more recent version of Wikipedia.

Trained on	Evaluated on											
	WikiQA			SelQA			SQuAD			InfoboxQA		
	MAP	MRR	F1	MAP	MRR	F1	MAP	MRR	F1	MAP	MRR	F1
WikiQA	**65.54**	**67.41**	13.33	53.47	54.12	8.68	73.16	73.72	11.26	30.85	30.85	-
SelQA	49.05	49.64	**24.30**	82.72	83.70	**48.66**	77.22	78.04	44.70	63.13	63.13	-
SQuAD	58.17	58.53	19.35	81.15	82.27	42.88	88.84	89.69	**44.93**	63.24	63.24	-
InfoboxQA	45.17	45.43	-	53.48	54.25	-	65.27	65.90	-	**79.44**	**79.44**	-
W+S+Q	56.40	56.51	-	**83.19**	**84.25**	-	88.78	89.65	-	62.53	62.53	-
W+S+Q+I	60.19	60.68	-	82.88	83.97	-	**88.92**	**89.79**	-	70.81	70.81	-

Table 3: Results for answer selection and triggering in % trained and evaluated across all corpora splits. The first column shows the training source, and the other columns show the evaluation sources. W: WikiQA, S: SelQA, Q: SQuAD, I: InfoboxQA.

4 Extrinsic Analysis

4.1 Answer Selection

Answer selection is evaluated by two metrics, mean average precision (MAP) and mean reciprocal rank (MRR). The bigram CNN introduced by Yu et al. (2014) is used to generate all the results in Table 3, where models are trained on either single or combined datasets. Clearly, the questions in WikiQA are the most challenging, and adding more training data from the other corpora hurts accuracy due to the uniqueness of query-based questions in this corpus. The best model is achieved by training on W+S+Q for SelQA; adding InfoboxQA hurts accuracy for SelQA although it gives a marginal gain for SQuAD. Just like WikiQA, InfoboxQA performs the best when it is trained on only itself. From our analysis, we suggest that to use models trained on WikiQA and InfoboxQA for short query-like questions, whereas to use ones trained on SelQA and SQuAD for long natural questions.

4.2 Answer Retrieval

Finding a paragraph that includes the answer context out of the entire Wikipedia is an extremely difficult task (1/28.7M). The last row of Table 2 shows results from answer retrieval. Given $k = 5$, SelQA and SQuAD show about 34% and 35% accuracy, which are reasonable. However, WikiQA shows a significantly lower accuracy of 12.47%; this is because the questions in WikiQA is about twice shorter than the questions in the other corpora such that not enough lexicons can be extracted from these questions for the Lucene search.

4.3 Answer Triggering

The results of $k = 5$ from the answer retrieval task in Section 4.2 are used to create the datasets for answer triggering, where about 65% of the questions are not expected to find their answer contexts from the provided paragraphs for SelQA and SQuAD and 87.5% are not expected for WikiQA. Answer triggering is evaluated by the F1 scores as presented in Table 3, where three corpora are cross validated. The results on WikiQA are pretty low as expected from the poor accuracy on the answer retrieval task. Training on SelQA gives the best models for both WikiQA and SelQA. Training on SQuAD gives the best model for SQuAD although the model trained on SelQA is comparable. Since the answer triggering datasets are about 5 times larger than the answer selection datasets, it is computationally too expensive to combine all data for training. We plan to find a strong machine to perform this experiment in near future.

5 Related work

Lately, several deep learning approaches have been proposed for question answering. Yu et al. (2014) presented a CNN model that recognizes the semantic similarity between two sentences. Wang and Nyberg (2015) presented a stacked bidirectional LSTM approach to read words in sequence, then outputs their similarity scores. Feng et al. (2015) applied a general deep learning framework to non-factoid question answering. Santos et al. (2016) introduced an attentive pooling mechanism that led to further improvements in selection-based QA.

6 Conclusion

We present a comprehensive comparison study of the existing corpora for selection-based question answering. Our intrinsic analysis provides a better understanding of the uniqueness or similarity between these corpora. Our extrinsic analysis shows the strength or weakness of combining these corpora together for statistical learning. Additionally, we create a silver-standard dataset for answer retrieval and triggering, which will be publicly available. In the future, we will explore different ways of improving the quality of our silver-standard datasets by fine-tuning the hyper-parameters.

References

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying Deep Learning to Answer Selection: A Study and An Open Task. In *IEEE Workshop on Automatic Speech Recognition and Understanding*. pages 813–820.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 523–533.

Sujay Kumar Jauhar, Peter Turney, and Eduard Hovy. 2016. Tables as semi-structured knowledge for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 474–483. http://www.aclweb.org/anthology/P16-1045.

Tomasz Jurczyk, Michael Zhai, and Jinho D. Choi. 2016. SelQA: A New Benchmark for Selection-based Question Answering. In *Proceedings of the 28th International Conference on Tools with Artificial Intelligence*. ICTAI'16.

Xin Li and Dan Roth. 2002. Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*. COLING '02, pages 1–7.

Alvaro Morales, Varot Premtoon, Cordelia Avery, Sue Felshin, and Boris Katz. 2016. Learning to answer questions from wikipedia infoboxes. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1930–1935. https://aclweb.org/anthology/D16-1199.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* .

Rami Reddy Nandi Reddy and Sivaji Bandyopadhyay. 2006. Dialogue based question answering system in telugu. In *Proceedings of the Workshop on Multilingual Question Answering*. Association for Computational Linguistics, pages 53–60.

Mrinmaya Sachan, Kumar Dubey, and Eric Xing. 2016. Science question answering using instructional materials. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 467–473. http://anthology.aclweb.org/P16-2076.

Cícero Nogueira dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *CoRR* abs/1602.03609. http://arxiv.org/abs/1602.03609.

Dan Shen and Dietrich Klakow. 2006. Exploring correlation of dependency relation paths for answer extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 889–896.

Md Arafat Sultan, Vittorio Castelli, and Radu Florian. 2016. A joint model for answer sentence ranking and answer extraction. *Transactions of the Association for Computational Linguistics* 4:113–125.

Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 707–712. http://www.aclweb.org/anthology/P15-2116.

Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL'07, pages 22–32.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence Similarity Learning by Lexical Decomposition and Composition. *arXiv* arXiv:1602.07019.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WIKIQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP'15, pages 2013–2018.

Wen-tau Yih, Ming-Wei Chang, Christopher Meek, and Andrzej Pastusiak. 2013. Question answering using enhanced lexical semantic models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. ACL'13, pages 1744–1753.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep Learning for Answer Sentence Selection. In *Proceedings of the NIPS Deep Learning Workshop*.