Sanity Check: A Strong Alignment and Information Retrieval Baseline for Question Answering

Vikas Yadav
School of Information
University of Arizona
vikasy@email.arizona.edu

Rebecca Sharp
Department of Computer Science
University of Arizona
bsharp@email.arizona.edu

Mihai Surdeanu
Department of Computer Science
University of Arizona
msurdeanu@email.arizona.edu

ABSTRACT

While increasingly complex approaches to question answering (QA) have been proposed, the true gain of these systems, particularly with respect to their expensive training requirements, can be inflated when they are not compared to adequate baselines. Here we propose an unsupervised, simple, and fast alignment and information retrieval baseline that incorporates two novel contributions: a *one-to-many alignment* between query and document terms and *negative alignment* as a proxy for discriminative information. Our approach not only outperforms all conventional baselines as well as many supervised recurrent neural networks, but also approaches the state of the art for supervised systems on three QA datasets. With only three hyperparameters, we achieve 47% P@1 on an 8th grade Science QA dataset, 32.9% P@1 on a Yahoo! answers QA dataset and 64% MAP on WikiQA.

CCS CONCEPTS

• **Information systems → Question answering**;

KEYWORDS

Semantic alignment; Answer re-ranking; Question answering

ACM Reference Format:
Vikas Yadav, Rebecca Sharp, and Mihai Surdeanu. 2018. Sanity Check: A Strong Alignment and Information Retrieval Baseline for Question Answering. In *SIGIR '18: 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8-12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210142

1 INTRODUCTION

Question answering (QA), i.e., finding short answers to natural language questions, is a challenging task that is an important step towards natural language understanding [6]. With the recent and widespread success of deep architectures in natural language processing (NLP) tasks [27], more and more QA tasks have been approached with deep learning and in many cases the state of the art for a given question set is held by a neural architecture (e.g., Tymoshenko et al. [22] for WikiQA [24]). However, with these architectures becoming the expectation, comparisons to *strong* baselines often are neglected and thus allow us to lose sight of the true gain of these complex architectures, especially relative to their steep training costs.

Here we introduce a strong alignment and information retrieval (IR) baseline that is simple, completely unsupervised, and trivially tuned. Specifically, the contributions of this work are:

(1) We propose an unsupervised alignment and IR approach that features one-to-many alignments to better control for context, as well as negative alignments as a proxy for discriminative information. We show that depending on the statistics of the given question set, different ratios of these components provide the best performance, but that this tuning can be accomplished with only three hyperparameters.

(2) We demonstrate that our approach yields near state-of-the-art performance on three separate QA tasks, outperforming all baselines, and, more importantly, several more complex, supervised systems. These results suggest that, contrary to recent literature, unsupervised approaches that rely on simple bag-of-word strategies remain powerful contenders on QA tasks, and, minimally, should inform stronger QA baselines. The code to reproduce the results in this paper is publicly available[1].

2 RELATED WORK

Information retrieval (IR) systems [e.g., 19] have served as the standard baseline for QA tasks [16, 21, *inter alia*]. However, the lack of lexical overlap in many QA datasets between questions and answers [1, 7, 28], makes standard IR approaches that rely on strict lexical matching less applicable. Several IR systems have been modified to use distributional similarity to align query terms to the most similar document term for various tasks, including document matching [13], short text similarity [10], and answer selection [4]. However, using only a *single* most similar term can lead to spurious matches, e.g., with different word senses. Here we expand on this by allowing a *one-to-many* mapping between a question term and similar answer terms to better represent how on-context a given answer candidate is.

Negative information has also been show to be useful in answer sentence selection [23, 25]. We also include negative information in the form of *negative alignments* to aid in distinguishing correct answers from close competitors.

Several QA approaches have used similar features for establishing strong baseline systems, [e.g., 17, 21]. These systems are conceptually related to our work, but they are supervised and employed on different tasks, so their results are not directly comparable to our unsupervised system.

SIGIR'18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210142

[1] https://github.com/clulab/releases/tree/master/sigir2018-sanitycheck

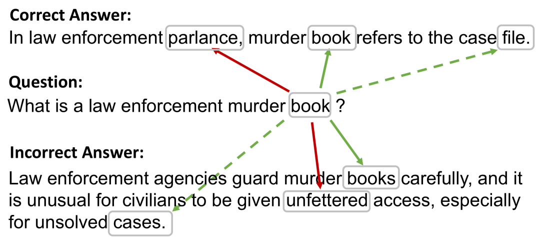

Figure 1: Example of our alignment approach for mapping terms in the question to the most similar and different terms in the candidate answer. Highest-ranked alignments are shown with a solid green arrow, second-highest with a dashed green arrow, and lowest-ranked alignments are shown with a red arrow.

3 APPROACH

Our unsupervised QA approach is designed to robustly estimate how relevant a candidate answer is to a given question. We do this by utilizing both positive and negative one-to-many alignments to approximate context. Specifically, our approach operates in three main steps:

(1) Preprocessing: We first pre-process both the question and its candidate answers using NLTK [2] and retain only non-stopword lemmas for each. We additionally calculate the inverse document frequency (idf) of each query term, q_i locally using:

$$idf(q_i) = \log \frac{N - docfreq(q_i) + 0.5}{docfreq(q_i) + 0.5} \tag{1}$$

where N is the number of questions and $docfreq(q_i)$ is the number of questions that contain q_i.

(2) Alignment: Next we perform the one-to-many alignment between the terms in the question, Q, and the candidate answer, A. For $q_i \in Q$, we rank the terms $a_j \in A$ by their similarity to q_i as determined by cosine similarity using off-the-shelf 300-dim Glove word embeddings [18], which were not trained on any of the datasets used here. For each q_i we find the ranked top K^+ most similar terms in A, $\{a^+_{q_i,1}, a^+_{q_i,2}, ..., a^+_{q_i,K^+}\}$ as well as the K^- least similar terms, $\{a^-_{q_i,1}, a^-_{q_i,2}, ..., a^-_{q_i,K^-}\}$. For example, in Figure 1, *book* in the question is aligned with *book* and *files* in the correct answer and with *book* and *case* (after preprocessing) in the incorrect answer as positive alignments.

(3) Candidate Answer Scoring: We then use these alignments along with the idfs of the question terms to find the score for each candidate answer, $s(Q, A)$, based on the weighted sums of the individual term alignment scores, such that:

$$s(Q, A) = \sum_{i=1}^{N} idf(q_i) \cdot align(q_i, A) \tag{2}$$

$$align(q_i, A) = pos(q_i, A) + \lambda \cdot neg(q_i, A) \tag{3}$$

$$pos(q_i, A) = \sum_{k=1}^{K^+} \frac{1}{k} \cdot a^+_{q_i,k} \tag{4}$$

$$neg(q_i, A) = \sum_{k=1}^{K^-} \frac{1}{k} \cdot a^-_{q_i,k} \tag{5}$$

where N is the number of question terms, $align(q_i, A)$ is the alignment score between the question term, q_i and the answer candidate, A, and λ is the weight for the negative information. $pos(q_i, A)$ and $neg(q_i, A)$ represent the scores for the one-to-many alignments for the most and least similar terms respectively. Importantly, the only hyperparameters involved are: K^+, K^-, and λ.

The intuition behind this formula is that by aggregating several alignments (i.e., through summing), the model can approximate context. In terms of the example in Figure 1, the secondary alignments for *book* help discern that the correct answer is more on-context than the incorrect answer (i.e., *book* is more similar to *file* than it is to *cases*). Further, the negative alignments cause candidate answers with more off-context terms to be penalized more (as with *book* and *unfettered*). These negative alignment penalties serve as an inexpensive proxy for discriminative learning.

4 EXPERIMENTS

4.1 Data

We evaluate our approach on three distinct datasets:[2]

WikiQA:[3] a dataset created by Yang et al. [24] for open-domain QA consisting of Bing queries and corresponding answer sentences taken from Wikipedia articles. The set is divided into train/dev/test partitions with 1040, 140 and 293 questions respectively.

Yahoo! Answers[4] **(YA):** 10,000 *How* questions, each with a community-chosen best answer.[5] We use the same 50-25-25 train/dev/test partitions as Jansen et al. [8].

8th Grade Science (ScienceQA): a set of multiple-choice science exam questions, each with four candidate answers. We use the same 2500/800 train/test split as [20]. For better comparison with previous work, here we modify the approach slightly to score candidate answers against the same external knowledge base (KB) of short flash-card style texts from StudyStack[6] and Quizlet[7] as was used by Sharp et al. [20]. Specifically, we first build IR queries from the question combined with each of the multiple-choice answers, and use there queries to retrieve the top five documents from the KB for each answer candidate. We then score each of these documents, as described in Section 3, using the combined question and answer candidate in place of Q and each of the five documents in place of A. The score for the answer candidate is then the sum of these five document scores.

4.2 Baselines

We compare against the following baselines:

BM25: We choose the candidate answer with the highest BM25 score [19], using the default values for the hyperparameters.

IDF Weighted Word Count: We also compare against baselines from previous work based on *tf-idf*. For WikiQA this is the IDF weighted word count baseline of Yang et al. [24] and in YA this is the CR baseline of Jansen et al. [8]. In YA we also compare against the stronger supervised CR + LS baseline of Jansen et al. [8], which combines *tf-idf* features with lexical semantic features into a linear SVM.

[2]As our approach is unsupervised, we tuned our two hyperparameters on the training and development partitions of each dataset.
[3]https://www.microsoft.com/en-us/download/details.aspx?id=52419
[4]http://answers.yahoo.com
[5]The questions are filtered to have at least 4 candidate answers, with an average of 9.
[6]https://www.studystack.com/
[7]https://quizlet.com/

Dataset	K^+	K^-	λ	Q:A
WikiQA	5	1	0.4	1:4
ScienceQA	1	1	0.4	2:1
Yahoo QA	3	0	–	1:5

Table 1: Tuned values for hyperparameters along with approximate ratios of average number of terms in questions versus answers across the dataset.

Learning constrained latent representations (LCLR): For WikiQA, we also compare against LCLR [25], which was used as a strong baseline by Yang et al. [24] to accompany the WikiQA dataset. LCLR uses rich lexical semantic information including synonyms, antonyms, hypernyms, and a vector space model for semantic word similarity.

Single-Alignment (One-to-one): We use only the single highest scoring pair of query word and answer word, i.e., $K^+ = 1$ and $K^- = 0$. This baseline has been used for other NLP tasks, such as document matching [12] and sentence similarity [10].

One-to-all: We additionally compare against a model without an alignment threshold, reducing Equation 3 to:

$$align(q_i, A) = \sum_{k=1}^{m} \frac{1}{k} \cdot cosSim(q_i, a^+_{q_i,k}) \tag{6}$$

where m is the number of words in the answer candidate.

4.3 Supervised Model Comparisons:

For each QA dataset, we compare against previous supervised systems.

WikiQA: For WikiQA, we compare against strong RNN and attention based QA systems. Jurczyk et al. [9] use multiple RNN models with attention pooling. Both Yin et al. [26] and dos Santos et al. [5] use similar approaches of attention layers over CNN's and RNN's. Miller et al. [15] use key value memory networks using Wikipedia as the knowledge base and Tymoshenko et al. [22] employed a hybrid of Tree Kernals and CNNs.

YA: For the YA dataset, Jansen et al. [8] use discourse, lexical semantic, and IR features in a linear SVM. Fried et al. [7] also use a linear SVM but with higher-order alignment features (i.e., "multi-hop" alignment). Bogdanova and Foster [3] used learned representations of questions and answers in a feed-forward NN and Liu et al. [14] use explicit features alongside recurrent NN representations of questions and answers.

ScienceQA: We compare against the most recently published works for this dataset. Khot et al. [11] employ Integer Linear Programming with a knowledge base of tuples to select the correct answers. Sharp et al. [20] use a combination of learned and explicit features in a shallow NN to simultaneously rerank answers and their justifications.

4.4 Tuning

As described in Section 3, our proposed model has just 3 hyperparameters: K^+, the number of positive alignments for each question term; K^-, the number of negative alignments; and λ, the weight assigned to the negative information. We tuned each of these on development and show the selected values in Table 1.

We hypothesize that these empirically determined best values for the hyperparameters are correlated with the ratio between the average length of questions and answers across the dataset[8] (also shown in Table 1). That is, in the question sets where answers tend to be several times longer than questions, more alignments per question term were useful. This is in direct contrast with the Science dataset, where questions are typically twice as long as answers.

#	Supervised	Model	MAP
1	No	Wgt Word Cnt [24]	50.99
2	Yes	LCLR [24]	59.93
3	No	Our model (One-to-one)	62.77
4	No	Our model (One-to-all)	60.91
5	Yes	Yang et al. [24] CNN+Cnt	65.20
6	Yes	Jurczyk et al. [9]RNN-1way	66.64
7	Yes	Jurczyk et al. [9] RNN-Attention_pool	67.47
8	Yes	dos Santos et al. [5]	68.86
9	Yes	Yin et al. [26]	69.21
10	Yes	Miller et al. [15]	70.69
11	Yes	Tymoshenko et al. [22]	72.19
12	No	Our final model	64.02*†

Table 2: Performance on the WikiQA dataset, measured by mean average precision (MAP), for other baselines (both supervised and unsupervised), recent supervised systems, and finally our approach. * and † indicate that the difference between the model and the One-to-one and One-to-all baselines (respectively) is statistically significant ($p < 0.05$), as determined through a one-tailed bootstrap resampling test with 10,000 iterations.

#	Supervised	Model	P@1
1	No	BM25	18.60
2	No	CR [8]	19.57
3	Yes	CR + LS [8]	26.57
4	No	Our model (One-to-one)	28.41
5	No	Our model (One-to-all)	20.17
6	Yes	Jansen et al. [8]	30.49
7	Yes	Fried et al. [7]	33.01
8	Yes	Bogdanova and Foster [3]	37.17
9	Yes	Liu et al. [14]	38.74
10	No	Our final model	32.93*†

Table 3: Performance on the Yahoo! Answers dataset, measured by precision-at-one (P@1). Significance is indicated as described in Table 2.

#	Supervised	Model	P@1
1	No	BM25	39.75
2	No	Our model (One-to-one)	46.38
3	No	Our model (One-to-all)	34.13
4	Yes	Khot et al. [11]	46.17
5	Yes	Sharp et al. [20]	53.30
6	No	Our final model	47.00†

Table 4: Performance on the 8th grade science dataset, measured by precision-at-one (P@1). Significance is indicated as in Table 2.

5 RESULTS AND DISCUSSION

We evaluate our approach on three distinct QA datasets: WikiQA, Yahoo! Answers (YA), and an 8th grade science dataset (ScienceQA). These results are shown in Tables 2, 3 and 4. In Science and Yahoo! QA, our approach significantly outperforms BM25 ($p < 0.05$)[9], demonstrating that incorporating lexical semantic alignments between question terms and answer terms (i.e., going beyond strict lexical overlap) is beneficial for QA. LCLR [24, 25] is considered to be a stronger baseline for WikiQA, and our model outperforms it by +4.10% MAP.

Further, we compare our full model with both a single alignment approach (i.e., one-to-one) as well as a maximal alignment (i.e., one-to-all) approach. In all datasets, our full model performed better

[8] Length statistics were calculated after stop word removal.
[9] All statistical significance determined through one-tailed bootstrap resampling with 10,000 iterations.

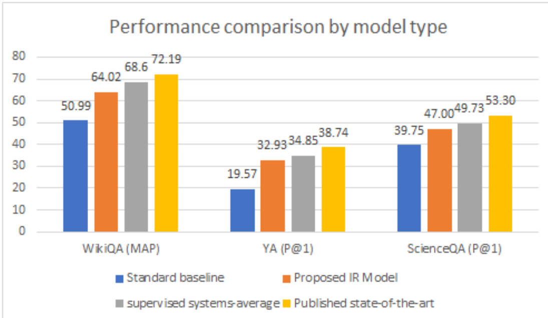

Figure 2: Histogram depicting the performance across three QA datasets of the standard baselines (shown in blue), our proposed unsupervised model (orange), the average of recently proposed supervised systems (grey), and the current state of the art (yellow). Notably, our model exceeds the standard baselines and approaches the mean of the supervised systems in all three datasets.

than the single alignment approach; in both WikiQA and YA this difference was significant ($p < 0.05$). Our full model was also significantly better than the one-to-all baseline in all models ($p < 0.05$). This demonstrates that including additional context in the form of multiple alignments is useful, but that there is a "Goldilocks" zone, and going beyond that is detrimental. We note that while the negative alignment boosted performance, in none of the datasets was its contribution significant individually.

Perhaps more interestingly, despite its simplicity, lack of parameters, and completely unsupervised nature, our approach either beats or approaches many much more complex supervised systems with steep training costs (e.g., attention-based RNNs), showing we can come closer to bridging the performance gap between simple baselines and complex systems using straightforward approaches, as illustrated in Figure 2. We suspect that our proposed approach would also be complementary to several of the more complex systems (particularly those without IR components [e.g. 11]), which would allow for additional gains through ensembling.

6 CONCLUSION

We introduced a fast and simple, yet strong, unsupervised baseline approach for QA that uses pre-trained word embeddings to produce one-to-many alignments between question and answer words, capturing both positive and negative alignments. Despite its simplicity, our approach considerably outperforms all current baselines, as well as several complex, supervised systems, approaching state-of-the-art performance on three QA tasks. Our work suggests that simple alignment strategies remain strong contenders for QA, and that the QA community would benefit from such stronger baselines for more rigorous analyses.

REFERENCES

[1] Adam Berger, Rich Caruana, David Cohn, Dayne Freytag, and Vibhu Mittal. 2000. Bridging the Lexical Chasm: Statistical Approaches to Answer Finding. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval.* Athens, Greece.

[2] Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions.* Association for Computational Linguistics, Stroudsburg, PA, USA, 69–72. https://doi.org/10.3115/1225403.1225421

[3] Dasha Bogdanova and Jennifer Foster. 2016. This is how we do it: Answer Reranking for Open-domain How Questions with Paragraph Vectors and Minimal Feature Engineering. In *HLT-NAACL 2016.*

[4] Rishav Chakravarti, Jiri Navratil, and Cicero Nogueira dos Santos. 2017. Improved Answer Selection with Pre-Trained Word Embeddings. *arXiv preprint arXiv:1708.04326* (2017).

[5] Cícero Nogueira dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive Pooling Networks. (2016). arXiv:1602.03609

[6] Oren Etzioni. 2011. Search needs a shake-up. *Nature* 476, 7358 (2011), 25–26.

[7] Daniel Fried, Peter Jansen, Gustave Hahn-Powell, Mihai Surdeanu, and Peter Clark. 2015. Higher-order Lexical Semantic Models for Non-factoid Answer Reranking. *Transactions of the Association for Computational Linguistics* 3 (2015), 197–210.

[8] Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse Complements Lexical Semantics for Non-factoid Answer Reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL).*

[9] Tomasz Jurczyk, Michael Zhai, and Jinho D Choi. 2016. SelQA: A New Benchmark for Selection-based Question Answering. In *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on.* IEEE, 820–827.

[10] Tom Kenter and Maarten De Rijke. 2015. Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management.* ACM, 1411–1420.

[11] Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering Complex Questions Using Open Information Extraction. In *Proceedings of Association for Computational Linguistics (ACL).*

[12] Been Kim, Julie A. Shah, and Finale Doshi-Velez. 2015. Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction. In *Neural Information Processing Systems (NIPS).*

[13] Sun Kim, Nicolas Fiorini, W John Wilbur, and Zhiyong Lu. 2017. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. *Journal of biomedical informatics* 75 (2017), 122–127.

[14] Qun Liu, Jennifer Foster, Dasha Bogdanova, and Daria Dzendzik. 2017. If You Can't Beat Them Join Them: Handcrafted Features Complement Neural Nets for Non-Factoid Answer Reranking. In *Association for Computational Linguistic European Chapter.*

[15] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-Value Memory Networks for Directly Reading Documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 1400–1409.

[16] Dan Moldovan and Mihai Surdeanu. 2002. On the role of information retrieval and information extraction in question answering systems. In *Information Extraction in the Web Era.* Springer, 129–147.

[17] Piero Molino, Luca Maria Aiello, and Pasquale Lops. 2016. Social question answering: Textual, user, and network features for best answer prediction. *ACM Transactions on Information Systems (TOIS)* 35, 1 (2016), 4.

[18] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 1532–1543.

[19] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[20] Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Marco A Valenzuela-Escárcega, Peter Clark, and Michael Hammond. 2017. Tell Me Why: Using Question Answering as Distant Supervision for Answer Justification. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017).* 69–79.

[21] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to Rank Answers to Non-Factoid Questions from Web Collections. *Computational Linguistics* 37, 2 (2011), 351–383.

[22] Kateryna Tymoshenko, Daniele Bonadiman, and Alessandro Moschitti. 2017. Ranking Kernels for Structures and Embeddings: A Hybrid Preference and Classification Model. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP).*

[23] Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence Similarity Learning by Lexical Decomposition and Composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.* The COLING 2016 Organizing Committee, 1340–1349.

[24] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering.. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* 2013–2018.

[25] 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).*

[26] Wenpeng Yin, Hinrich Scütze, Bing Xiang, and Bowen Zhou. 2016. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. *Transactions of the Association for Computational Linguistics* 4 (2016), 259–272.

[27] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2017. Recent trends in deep learning based natural language processing. *arXiv preprint arXiv:1708.02709* (2017).

[28] Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning Continuous Word Embedding with Metadata for Question Retrieval in Community Question Answering. In *Associations for Computational Linguistics, 2015.*