Science Question Answering using Instructional Materials

Mrinmaya Sachan **Avinava Dubey** **Eric P. Xing**
School of Computer Science
Carnegie Mellon University
{mrinmays, akdubey, epxing}@cs.cmu.edu

Abstract

We provide a solution for elementary science tests using instructional materials. We posit that there is a hidden structure that explains the correctness of an answer given the question and instructional materials and present a unified max-margin framework that learns to find these hidden structures (given a corpus of question-answer pairs and instructional materials), and uses what it learns to answer novel elementary science questions. Our evaluation shows that our framework outperforms several strong baselines.

1 Introduction

We propose an approach for answering multiple-choice elementary science tests (Clark, 2015) using the science curriculum of the student and other domain specific knowledge resources. Our approach learns latent *answer-entailing structures* that align question-answers with appropriate snippets in the curriculum. The student curriculum usually comprises of a set of textbooks. Each textbook, in-turn comprises of a set of chapters, each chapter is further divided into sections – each discussing a particular science concept. Hence, the answer-entailing structure consists of selecting a particular textbook from the curriculum, picking a chapter in the textbook, picking a section in the chapter, picking a few sentences in the section and then aligning words/multi-word expressions (mwe's) in the hypothesis (formed by combining the question and an answer candidate) to words/mwe's in the picked sentences. The answer-entailing structures are further refined using external domain-specific knowledge resources such as science dictionaries, study guides and semi-structured tables (see Figure 1). These domain-specific knowledge resources can be very useful forms of knowledge representation as shown in previous works (Clark et al., 2016).

Alignment is a common technique in many NLP applications such as MT (Blunsom and Cohn, 2006), RTE (Sammons et al., 2009; MacCartney et al., 2008; Yao et al., 2013; Sultan et al., 2014), QA (Berant et al., 2013; Yih et al., 2013; Yao and Van Durme, 2014; Sachan et al., 2015), etc. Yet, there are three key differences between our approach and alignment based approaches for QA in the literature: (i) We incorporate the curriculum hierarchy (i.e. the book, chapter, section bifurcation) into the latent structure. This helps us jointly learn the retrieval and answer selection modules of a QA system. Retrieval and answer selection are usually designed as isolated or loosely connected components in QA systems (Ferrucci, 2012) leading to loss in performance – our approach mitigates this shortcoming. (ii) Modern textbooks typically provide a set of review questions after each section to help students understand the material better. We make use of these review problems to further improve our model. These review problems have additional value as part of the latent structure is known for these questions. (ii) We utilize domain-specific knowledge sources such as study guides, science dictionaries or semi-structured knowledge tables within our model.

The joint model is trained in max-margin fashion using a latent structural SVM (LSSVM) where the answer-entailing structures are latent. We train and evaluate our models on a set of 8^{th} grade science problems, science textbooks and multiple domain-specific knowledge resources. We achieve superior performance vs. a number of baselines.

2 Method

Science QA as Textual Entailment: First, we

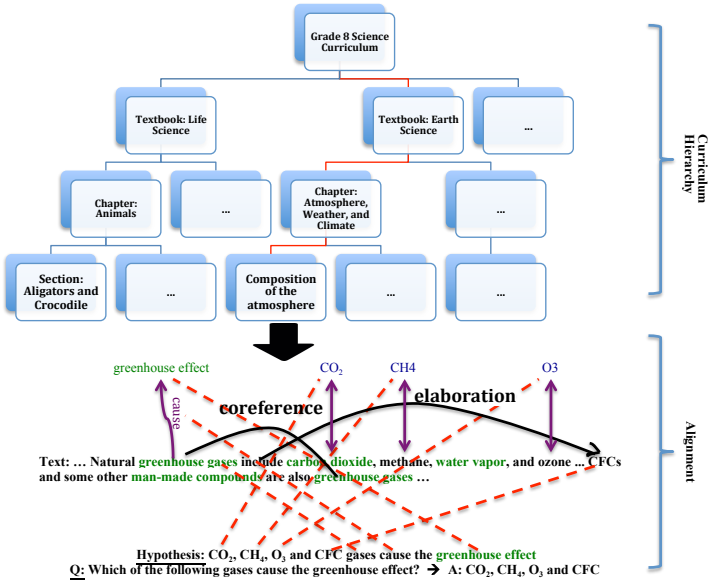

Figure 1: An example *answer-entailing structure*. The answer-entailing structure consists of selecting a particular textbook from the curriculum, picking a chapter in the textbook, picking a section in the chapter, picking sentences in the section and then aligning words/mwe's in the hypothesis (formed by combining the question and an answer candidate) to words/mwe's in the picked sentences or some related "knowledge" appropriately chosen from additional knowledge stores. In this case, the relation (greenhouse gases, cause, greenhouse effect) and the equivalences (e.g. carbon dioxide = CO_2) – shown in violet – are hypothesized using external knowledge resources. The dashed red lines show the word/mwe alignments from the hypothesis to the sentences (some word/mwe are not aligned, in which case the alignments are not shown), the solid black lines show coreference links in the text and the RST relation (elaboration) between the two sentences. The picked sentences do not have to be contiguous sentences in the text. All mwe's are shown in green.

consider the case when review questions are not used. For each question $q_i \in Q$, let $A_i = \{a_{i1}, \ldots, a_{im}\}$ be the set of candidate answers to the question [1]. We cast the science QA problem as a textual entailment problem by converting each question-answer candidate pair $(q_i, a_{i,j})$ into a hypothesis statement h_{ij} (see Figure 1)[2]. For each question q_i, the science QA task thereby reduces to picking the hypothesis $\hat{h}_i$ that has the highest likelihood of being entailed by the curriculum among the set of hypotheses $\mathbf{h}_i = \{h_{i1}, \ldots, h_{im}\}$ generated for that question. Let $h_i^* \in \mathbf{h}_i$ be the correct hypothesis corresponding to the correct answer.

Latent Answer-Entailing Structures help the model in providing evidence for the correct hypothesis. As described before, the structure depends on: (a) snippet from the curriculum hierarchy chosen to be aligned to the hypothesis, (b) external knowledge relevant for this entailment, and (c) the word/mwe alignment. The snippet from the curriculum to be aligned to the hypothesis is determined by walking down the curriculum hierarchy and then picking a set of sentences from the section chosen. Then, a subset of relevant external knowledge in the form of triples and equivalences (called knowledge bits) is selected from our reservoir of external knowledge (science dictionaries, cheat sheets, semi-structured tables, etc). Finally, words/mwe's in the hypothesis are aligned to words/mwe's in the snippet or knowledge bits. Learning these alignment edges helps the model determine which semantic constituents should be compared to each other. These alignments are also used to generate more effective features. The choice of snippets, choice of the relevant external knowledge and the alignments in conjunction form the latent answer-entailing structure. Let $\mathbf{z}_{ij}$ represent the latent structure for the question-answer candidate pair $(q_i, a_{i,j})$.

Max-Margin Approach: We treat science QA as a structured prediction problem of ranking the hypothesis set $\mathbf{h}_i$ such that the correct hypothesis is at the top of this ranking. We learn a scoring function $S_{\mathbf{w}}(h, \mathbf{z})$ with parameter $\mathbf{w}$ such that the score of the correct hypothesis h_i^* and the corresponding best latent structure $\mathbf{z}_i^*$ is higher than the score of the other hypotheses and their corresponding best latent structures. In fact, in a max-margin fashion, we want that $S_{\mathbf{w}}(h_i^*, \mathbf{z}_i^*) > S(h_{ij}, \mathbf{z}_{ij}) + 1 - \xi_i$ for all $h_j \in \mathbf{h} \setminus h^*$ for some slack ξ_i. Writing the relaxed max margin formulation:

$$\min_{||\mathbf{w}||} \frac{1}{2}||\mathbf{w}||_2^2 + C \sum_i \max_{\mathbf{z}_{ij}, h_{ij} \in \mathbf{h}_i \setminus h_i^*} S_{\mathbf{w}}(h_{ij}, \mathbf{z}_{ij}) + \Delta(h_i^*, h_{ij}) - C \sum_i S_{\mathbf{w}}(h_i^*, \mathbf{z}_i^*) \quad (1)$$

We use 0-1 cost, i.e. $\Delta(h_i^*, h_{ij}) = \mathbb{1}(h_i^* \neq h_{ij})$ If the scoring function is convex then this objective is in concave-convex form and hence can be

[1] Candidate answers may be pre-defined, as in multiple-choice QA, or may be undefined but easy to extract with a degree of confidence (e.g., by using a pre-existing system)

[2] We use a set of question matching/rewriting rules to achieve this transformation. The rules match each question into one of a large set of pre-defined templates and applies a unique transformation to the question & answer candidate to achieve the hypothesis. Code provided in the supplementary.

solved by the concave-convex programming procedure (CCCP) (Yuille and Rangarajan, 2003). We assume the scoring function to be linear:$S_{\mathbf{w}}(h, \mathbf{z}) = \mathbf{w}^T \psi(h, \mathbf{z})$. Here, $\psi(h, \mathbf{z})$ is a feature map discussed later. The CCCP algorithm essentially alternates between solving for $\mathbf{z}_i^*$, $\mathbf{z}_{ij}$ $\forall j$ s.t. $h_{ij} \in \mathbf{h}_i \setminus h_i^*$ and $\mathbf{w}$ to achieve a local minima. In the absence of information regarding the latent structure $\mathbf{z}$ we pick the structure that gives the best score for a given hypothesis i.e. $\arg\max_z S_{\mathbf{w}}(h, z)$. The complete procedure is given in the supplementary.

Inference and knowledge selection: We use beam search with a fixed beam size (5) for inference. We infer the textbook, chapter, section, snippet and alignments one by one in this order. In each step, we only expand the five most promising (given by the current score) substructure candidates so far. During inference, we select top 5 knowledge bits (triples, equivalences, etc.) from the knowledge resources that could be relevant for this question-answer. This is done heuristically by picking knowledge bits that explain parts of the hypothesis not explained by the chosen snippets.

Incorporating partially known structures: Now, we describe how review questions can be incorporated. As described earlier, modern textbooks often provide review problems at the end of each section. These review problems have value as part of the answer-entailing structure (textbook, chapter and section) is known for these problems. In this case, we use the formulation (equation 1) except that the max over $\mathbf{z}$ for the review questions is only taken over the unknown part of the latent structure.

Multi-task Learning: Question analysis is a key component of QA systems. Incoming questions are often of different types (counting, negation, entity queries, descriptive questions, etc.). Different types of questions usually require different processing strategies. Hence, we also extend of our LSSVM model to a multi-task setting where each question q_i now also has a pre-defined associated type t_i and each question-type is treated as a separate task. Yet, parameters are shared across tasks,which allows the model to exploit the commonality among tasks when required. We use the MTLSSVM formulation from Evgeniou and Pontil (2004) which was also used in a reading comprehension setting by Sachan et al. (2015). In a nutshell, the approach redefines the LSSVM feature map and shows that the MTLSSVM objective takes the same form as equation 1 with a kernel corresponding to the feature map. Hence, one can simply redefine the feature map and reuse LSSVM algorithm to solve the *MTLSSVM*.

Features: Our feature vector $\psi(h, \mathbf{z})$ decomposes into five parts, where each part corresponds to a part of the answer-entailing structure. For the first part, we index all the textbooks and score the top retrieved textbook by querying the hypothesis statement. We use tf-idf and BM25 scorers resulting in two features. Then, we find the jaccard similarity of bigrams and trigrams in the hypothesis and the textbook to get two more features for the first part. Similarly, for the second part we index all the textbook chapters and compute the tf-idf, BM25 and bigram, trigram features. For the third part we index all the sections instead. The fourth part has features based on the text snippet part of the answer-entailing structure. Here we do a deeper linguistic analysis and include features for matching local neighborhoods in the snippet and the hypothesis: features for matching bigrams, trigrams, dependencies, semantic roles, predicate-argument structure as well as the global syntactic structure: a tree kernel for matching dependency parse trees of entire sentences (Srivastava and Hovy, 2013). If a text snippet contains the answer to the question, it should intuitively be similar to the question as well as to the answer. Hence, we add features that are the element-wise product of features for the text-question match and text-answer match. Finally, we also have features corresponding to the RST (Mann and Thompson, 1988) and coreference links to enable inference across sentences. RST tells us that sentences with discourse relations are related to each other and can help us answer certain kinds of questions (Jansen et al., 2014). For example, the "cause" relation between sentences in the text can often give cues that can help us answer "why" or "how" questions. Hence, we add additional features - conjunction of the rhetorical structure label from a RST parser and the question word - to our feature vector. Similarly, the entity and event co-reference relations allow us to reason about repeating entities or events. Hence, we replace an entity/event mention with their first mentions if that results into a greater score. For the alignment part, we induce features based on word/mwe level similarity of aligned words:

(a) Surface-form match (Edit-distance), and (b) Semantic word match (cosine similarity using SENNA word vectors (Collobert et al., 2011) and "Antonymy" 'Class-Inclusion' or 'Is-A' relations using Wordnet). Distributional vectors for mwe's are obtained by adding the vector representations of comprising words (Mitchell and Lapata, 2008). To account for the hypothesized knowledge bits, whenever we have the case that a word/mwe in the hypothesis can be aligned to a word/mwe in a hypothesized knowledge bit to produce a greater score, then we keep the features for the alignment with the knowledge bit instead.

Negation Negation is a concern for our approach as facts usually align well with their negated versions. To overcome this, we use a simple heuristic. During training, if we detect negation using a set of simple rules that test for the presence of negation words ("not", "n't", etc.), we flip the partial order adding constraints that require that the correct hypothesis to be ranked below all the incorrect ones. During test phase if we detect negation, we predict the answer corresponding to the hypothesis with the lowest score.

3 Experiments

Dataset: We used a set of 8^{th} grade science questions released as the training set in the Allen AI Science Challenge[3] for training and evaluating our model. The dataset comprises of 2500 questions. Each question has 4 answer candidates, of which exactly one is correct. We used questions 1-1500 for training, questions 1500-2000 for development and questions 2000-2500 for testing. We also used publicly available 8^{th} grade science textbooks available through $ck12.org$. The science curriculum consists of seven textbooks on Physics, Chemistry, Biology, Earth Science and Life Science. Each textbook on an average has 18 chapters, and each chapter in turn is divided into 12 sections on an average. Also, as described before, each section, on an average, is followed by 3-4 multiple choice review questions (total 1369 review questions). We collected a number of domain specific science dictionaries, study guides, flash cards and semi-structured tables (Simple English Wiktionary and Aristo Tablestore) available online and create triples and equivalences used as external knowledge.

[3]https://www.kaggle.com/c/the-allen-ai-science-challenge/

Question Category	Example
Questions without context:	Which example describes a learned behavior in a dog?
Questions with context:	When athletes begin to exercise, their heart rates and respiration rates increase. At what level of organization does the human body coordinate these functions?
Negation Questions:	A teacher builds a model of a hydrogen atom. A red golf ball is used for a proton, and a green golf ball is used for an electron. Which is not accurate concerning the model?

Table 1: Example questions for *Qtype* classification

Baselines: We compare our framework with ten baselines. The first two baselines (*Lucene* and *PMI*) are taken from Clark et al. (2016). The *Lucene* baseline scores each answer candidate a_i by searching for the combination of the question q and answer candidate a_i in a lucene-based search engine and returns the highest scoring answer candidate. The *PMI* baseline similarly scores each answer candidate a_i by computing the pointwise mutual information to measure the strength of the association between parts of the question-answer candidate combine and parts of the CK12 curriculum. The next three baselines, inspired from Richardson et al. (2013), retrieve the top two CK12 sections querying $q+a_i$ in *Lucene* and score the answer candidates using these documents. The *SW* and *SW+D* baselines match bag of words constructed from the question and the answer answer candidate to the retrieved document. The *RTE* baseline uses textual entailment (Stern and Dagan, 2012) to score answer candidates as the likelihood of being entailed by the retrieved document. Then we also tried other approaches such as the *RNN* approach described in Clark et al. (2016), *Jacana aligner* (Yao et al., 2013) and two neural network approaches, *LSTM* (Hochreiter and Schmidhuber, 1997) and *QANTA* (Iyyer et al., 2014) They form our next four baselines. To test if our approach indeed benefits from jointly learning the retrieval and the answer selection modules, our final baseline *Lucene+LSSVM Alignment* retrieves the top section by querying $q + a_i$ in *Lucene* and then learns the remaining answer-entailment structure (alignment part of the answer-entailing structure in Figure 1) using a LSSVM.

Task Classification for Multitask Learning: We explore two simple question classification schemes. The first classification scheme classifies questions based on the question word (what, why, etc.). We call this *Qword* classification. The second scheme is based on the type of the question asked and classifies questions into three coarser categories: (a) questions without context,

(b) questions with context and (c) negation questions. This classification is based on the observation that many questions lay down some context and then ask a science concept based on this context. However, other questions are framed without any context and directly ask for the science concept itself. Then there is a smaller, yet, important subset of questions that involve negation that also needs to be handled separately. Table 1 gives examples of this classification. We call this classification *Qtype* classification[4].

Results: We compare variants of our method[5] where we consider our modification for negation or not and multi-task LSSVMs. We consider both kinds of task classification strategies and joint training (JT). Finally, we compare our methods against the baselines described above. We report accuracy (proportion of questions correctly answered) in our results. Figure 2 shows the results. First, we can immediately observe that all the LSSVM models have a better performance than all the baselines. We also found an improvement when we handle negation using the heuristic described above[6]. MTLSSVMs showed a boost over single task LSSVM. *Qtype* classification scheme was found to work better than *Qword* classification which simply classifies questions based on the question word. The multi-task learner could benefit even more if we can learn a better separation between the various strategies needed to answer science questions. We found that joint training with review questions helped improve accuracy as well.

Feature Ablation: As described before, our feature set comprises of five parts, where each part corresponds to a part of the answer-entailing structure – textbook ($\mathbf{z}_1$), chapter ($\mathbf{z}_2$), section ($\mathbf{z}_3$), snippets ($\mathbf{z}_4$), and alignment ($\mathbf{z}_5$). It is interesting to know the relative importance of these parts in our model. Hence, we perform feature ablation on our best performing model - *MTLSSVM(QWord, JT)* where we remove the five feature parts one by one and measure the loss in accuracy. Figure 3 shows that the choice of section and alignment are important components of our model. Yet, all components are important and removing any of them will result in a loss of accuracy. Finally, in order to understand the value of external knowledge resources (K), we removed the component that induces and aligns the hypothesis with knowledge bits. This results in significant loss in performance, estabishing the efficacy of adding in external knowledge via our approach.

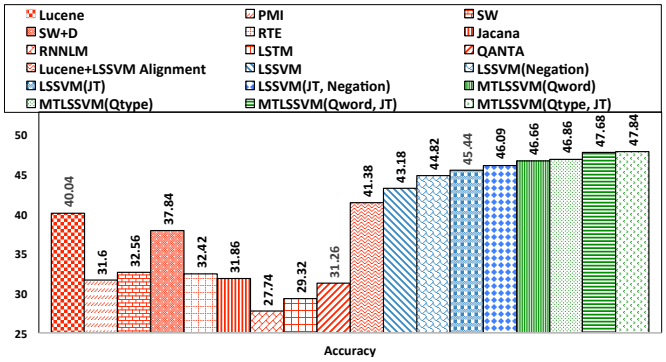

Figure 2: Variations of our method vs several baselines on the Science QA dataset. Differences between the baselines and LSSVMs, the improvement due to negation, the improvements due to multi-task learning and joint-learning are significant ($p < 0.05$) using the two-tailed paired T-test.

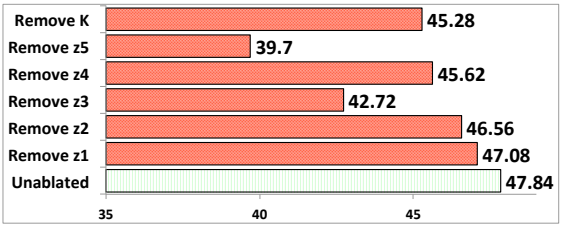

Figure 3: Ablation on *MTLSSVM(Qword, JT) model*

4 Conclusion

We addressed the problem of answering 8^{th} grade science questions using textbooks, domain specific dictionaries and semi-structured tables. We posed the task as an extension to textual entailment and proposed a solution that learns latent structures that align question answer pairs with appropriate snippets in the textbooks. Using domain specific dictionaries and semi-structured tables, we further refined the structures. The task required handling a variety of question types so we extended our technique to multi-task setting. Our technique showed improvements over a number of baselines. Finally, we also used a set of associated review questions, which were used to gain further improvements.

[4]We wrote a set of question matching rules (similar to the rules used to convert question answer pairs to hypotheses) to achieve this classification

[5]We tune the SVM regularization parameter C on the development set. We use Stanford CoreNLP, the HILDA parser (Feng and Hirst, 2014), and jMWE (Kulkarni and Finlayson, 2011) for linguistic preprocessing

[6]We found that the accuracy over test questions tagged by our heuristic as negation questions went up from 33.64 percent to 42.52 percent and the accuracy over test questions not tagged as negation did not decrease significantly

References

[Berant et al.2013] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1533–1544.

[Blunsom and Cohn2006] Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 65–72. Association for Computational Linguistics.

[Clark et al.2016] Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, and Peter Turney. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of AAAI*.

[Clark2015] Peter Clark. 2015. Elementary School Science and Math Tests as a Driver for AI:Take the Aristo Challenge! In *Proceedings of IAAI*.

[Collobert et al.2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

[Evgeniou and Pontil2004] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117.

[Feng and Hirst2014] Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521.

[Ferrucci2012] David A Ferrucci. 2012. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4):1–1.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Iyyer et al.2014] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of Empirical Methods in Natural Language Processing*.

[Jansen et al.2014] 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 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 977–986.

[Kulkarni and Finlayson2011] Nidhi Kulkarni and Mark Alan Finlayson. 2011. jmwe: A java toolkit for detecting multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 122–124. Association for Computational Linguistics.

[MacCartney et al.2008] Bill MacCartney, Michel Galley, and Christopher D Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the conference on empirical methods in natural language processing*, pages 802–811.

[Mann and Thompson1988] William C Mann and Sandra A Thompson. 1988. {Rhetorical Structure Theory: Toward a functional theory of text organisation}. *Text*, 3(8):234–281.

[Mitchell and Lapata2008] Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 236–244.

[Richardson et al.2013] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

[Sachan et al.2015] Mrinmaya Sachan, Avinava Dubey, Eric P Xing, and Matthew Richardson. 2015. Learning answer-entailing structures for machine comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[Sammons et al.2009] M. Sammons, V. Vydiswaran, T. Vieira, N. Johri, M. Chang, D. Goldwasser, V. Srikumar, G. Kundu, Y. Tu, K. Small, J. Rule, Q. Do, and D. Roth. 2009. Relation alignment for textual entailment recognition. In *TAC*.

[Srivastava and Hovy2013] Shashank Srivastava and Dirk Hovy. 2013. A walk-based semantically enriched tree kernel over distributed word representations. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1411–1416.

[Stern and Dagan2012] Asher Stern and Ido Dagan. 2012. Biutee: A modular open-source system for recognizing textual entailment. In *Proceedings of the ACL 2012 System Demonstrations*, pages 73–78.

[Sultan et al.2014] Arafat Md Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, pages 219–230.

[Yao and Van Durme2014] Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 956–966. Association for Computational Linguistics.

[Yao et al.2013] Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. A lightweight and high performance monolingual word aligner. In *ACL (2)*, pages 702–707.

[Yih et al.2013] Wentau 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*.

[Yuille and Rangarajan2003] A. L. Yuille and Anand Rangarajan. 2003. The concave-convex procedure. *Neural Comput.*