Spinning Straw into Gold: Using Free Text to Train Monolingual Alignment Models for Non-factoid Question Answering

Rebecca Sharp[1], Peter Jansen[1], Mihai Surdeanu[1], and **Peter Clark[2]**
[1] University of Arizona, Tucson, AZ, USA
[2] Allen Institute for Artificial Intelligence, Seattle, WA, USA
{bsharp,pajansen,msurdeanu}@email.arizona.edu
peterc@allenai.org

Abstract

Monolingual alignment models have been shown to boost the performance of question answering systems by "bridging the lexical chasm" between questions and answers. The main limitation of these approaches is that they require semistructured training data in the form of question-answer pairs, which is difficult to obtain in specialized domains or low-resource languages. We propose two inexpensive methods for training alignment models solely using free text, by generating artificial question-answer pairs from discourse structures. Our approach is driven by two representations of discourse: a shallow sequential representation, and a deep one based on Rhetorical Structure Theory. We evaluate the proposed model on two corpora from different genres and domains: one from Yahoo! Answers and one from the biology domain, and two types of non-factoid questions: manner and reason. We show that these alignment models trained directly from discourse structures imposed on free text improve performance considerably over an information retrieval baseline and a neural network language model trained on the same data.

1 Introduction

Question Answering (QA) is a challenging task that draws upon many aspects of NLP. Unlike search or information retrieval, answers infrequently contain lexical overlap with the question (e.g. *What should we eat for breakfast? – Zoe's Diner has good pancakes*), and require QA models to draw upon more complex methods to bridge this "lexical chasm" (Berger et al., 2000). These methods range from robust shallow models based on lexical semantics, to deeper, explainably-correct, but much more brittle inference methods based on first order logic.

Berger et al. (2000) proposed that this "lexical chasm" might be partially bridged by repurposing statistical machine translation (SMT) models for QA. Instead of translating text from one language to another, these monolingual alignment models learn to translate from question to answer[1], learning common associations from question terms such as *eat* or *breakfast* to answer terms like *kitchen, pancakes, or cereal*.

While monolingual alignment models have enjoyed a good deal of recent success in QA (see related work), they have expensive training data requirements, requiring a large set of aligned in-domain question-answer pairs for training. For low-resource languages or specialized domains like science or biology, often the only option is to enlist a domain expert to generate gold QA pairs – a process that is both expensive and time consuming. All of this means that only in rare cases are we accorded the luxury of having enough high-quality QA pairs to properly train an alignment model, and so these models are often underutilized or left struggling for resources.

Making use of recent advancements in discourse parsing (Feng and Hirst, 2012), here we address this issue, and investigate whether alignment models for QA can be trained from artificial question-answer pairs generated from discourse structures imposed on free text. We evaluate our methods on two corpora, generating alignment models for an open-domain community QA task using Gigaword[2], and for a biology-domain QA task using a biology textbook.

[1] In practice, alignment for QA is often done from answer to question, as answers tend to be longer and provide more opportunity for association (Surdeanu et al., 2011).

[2] LDC catalog number LDC2012T21

The contributions of this work are:

1. We demonstrate that by exploiting the discourse structure of free text, monolingual alignment models can be trained to surpass the performance of models built from expensive in-domain question-answer pairs.
2. We compare two methods of discourse parsing: a simple sequential model, and a deep model based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). We show that the RST-based method captures within and across-sentence alignments and performs better than the sequential model, but the sequential model is an acceptable approximation when a discourse parser is not available.
3. We evaluate the proposed methods on two corpora, including a low-resource domain where training data is expensive (biology).
4. We experimentally demonstrate that monolingual alignment models trained using our method considerably outperform state-of-the-art neural network language models in low resource domains.

2 Related Work

Lexical semantic models have shown promise in bridging Berger et al.'s (2000) "lexical chasm." In general, these models can be classified into alignment models (Echihabi and Marcu, 2003; Soricut and Brill, 2006; Riezler et al., 2007; Surdeanu et al., 2011; Yao et al., 2013) which require structured training data, and language models (Jansen et al., 2014; Sultan et al., 2014; Yih et al., 2013), which operate over free text. Here, we close this gap in resource availability by developing a method to train an alignment model over free text by making use of discourse structures.

Discourse has been previously applied to QA to help identify answer candidates that contain explanatory text (e.g. Verberne et al. (2007)). Jansen et al. (2014) proposed a reranking model that used both shallow and deep discourse features to identify answer structures in large answer collections across different tasks and genres. Here we use discourse to impose structure on free text to create inexpensive knowledge resources for monolingual alignment. Our work is conceptually complementary to that of Jansen et al. – where they explored largely unlexicalized discourse structures to identify explanatory text, we use discourse to learn lexicalized models for semantic similarity.

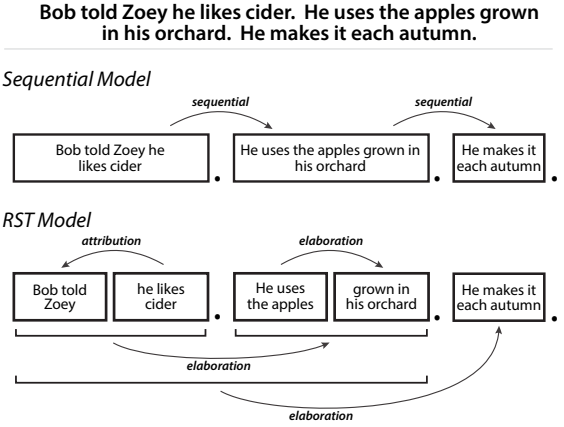

Figure 1: An example of the alignments produced by the two discourse models. The sequential model aligns pairs of consecutive sentences, capturing intersentence associations such as *cider–apples*, and *orchard–autumn*. The RST model generates alignment pairs from participants in all (binary) discourse relations, capturing both intrasentence and intersentence alignments, including *apples–orchard, cider–apples*, and *cider–autumn*.

Our work is conceptually closest to that of Hickl et al. (2006), who created artificially aligned pairs for textual entailment. Taking advantage of the structure of news articles, wherein the first sentence tends to provide a broad summary of the article's contents, Hickl et al. aligned the first sentence of each article with its headline. By making use of automated discourse parsing, here we go further and impose alignment structure over an entire text.

3 Approach

A written text is not simply a collection of sentences, but rather a flowing narrative where sentences and sentence elements depend on each other for meaning – a concept known as cohesion (Halliday and Hasan, 2014). Here we examine two methods for generating alignment training data from free text that make use of cohesion: a shallow method that uses only intersentence structures, and a deep method that uses both intrasentence and intersentence structures. We additionally attempt to separate the contribution of discourse from that of alignment in general by comparing these models against a baseline alignment model which aligns sentences at random.

The first model, the sequential discourse model (SEQ), considers that each sentence continues the

narrative of the previous one, and creates artificial question-answer pairs from all pairs of consecutive sentences. Thus, this model takes advantage of intersentence cohesion by aligning the content words[3] in each sentence with the content words in the following sentence. For example, in the passage in Figure 1, this model would associate *cider* in the first sentence with *apples* and *orchard* in the second sentence.

The second model uses RST to capture discourse cohesion both within and across sentence boundaries. We extracted RST discourse structures using an in-house parser (Surdeanu et al., 2015), which follows the architecture introduced by Hernault et al. (2010) and Feng and Hirst (2012). The parser first segments text into elementary discourse units (EDUs), which may be at sub-sentence granularity, then recursively connects neighboring units with binary discourse relations, such as *Elaboration* or *Contrast*.[4] Our parser differs from previous work with respect to feature generation in that we implement all features that rely on syntax using solely dependency syntax. For example, a crucial feature used by the parser is the dominance relations of Soricut and Marcu (2003), which capture syntactic dominance between discourse units located in the same sentence. While originally these dominance relations were implemented using constituent syntax, we provide an equivalent implementation that relies on dependency syntax. The main advantage to this approach is speed: the resulting parser performs at least an order of magnitude faster than the parser of Feng and Hirst (2012).

Importantly, we generate artificial alignment pairs from this imposed structure by aligning the governing text (nucleus) with its dependent text (satellite).[5] Turning again to the example in Figure 1, this RST-based model captures additional alignments that are both intrasentence, e.g., *apples–orchard*, and intersentence, e.g., *cider–autumn*.

[3]In pilot experiments, we found that aligning only nouns, verbs, adjectives, and adverbs yielded higher performance.

[4]The RST parser performs better on relations which occur more frequently. We use only relations that occurred at least 1% of the time. This amounted to six relations: *elaboration*, *attribution*, *background*, *contrast*, *same-unit*, and *joint*. Using all relations slightly improves performance by 0.3% P@1.

[5]Pilot experiments showed that this direction of alignment performed better than aligning from satellite to nucleus.

4 Models and Features

We evaluate the contribution of these alignment models using a standard reranking architecture (Jansen et al., 2014). The initial ranking of candidate answers is done using a shallow candidate retrieval (CR) component.[6] Then, these answers are reranked using a more expressive model that incorporates alignment features alongside the CR score. As a learning framework we use SVM^{rank}, a Support Vector Machine tailored for ranking.[7] We compare this alignment-based reranking model against one that uses a state-of-the-art recurrent neural network language model (RNNLM) (Mikolov et al., 2010; Mikolov et al., 2013), which has been successfully applied to QA previously (Yih et al., 2013).

Alignment Model: The alignment matrices were generated with IBM Model 1 (Brown et al., 1993) using GIZA++ (Och and Ney, 2003), and the corresponding models were implemented as per Surdeanu et al. (2011) with a global alignment probability. We extend this alignment model with features from Fried et al. (In press) that treat each (source) word's probability distribution (over destination words) in the alignment matrix as a distributed semantic representation, and make use the Jensen-Shannon distance (JSD)[8] between these conditional distributions. A summary of all these features is shown in Table 1.

RNNLM: We learned word embeddings using the `word2vec` RNNLM of Mikolov et al. (2013), and include the cosine similarity-based features described in Table 1.

5 Experiments

We tested our approach in two different domains, open-domain and cellular biology. For consistency we use the same corpora as Jansen et al. (2014), which are described briefly here.

Yahoo! Answers (YA): Ten thousand open-domain *how* questions were randomly chosen from the Ya-

[6]We use the same cosine similarity between question and answer lemmas as Jansen et al. (2014), weighted using *tf.idf*.

[7]`http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

[8]Jensen-Shannon distance is based on Kullback-Liebler divergence but is a distance metric (finite and symmetric).

	Feature Group	Feature Descriptions
Alignment Models	Global Alignment Probability	p(Q|A) according to IBM Model 1 (Brown et al., 1993)
	Jenson-Shannon Distance (JSD)	Pairwise JSDs were found between the probability distribution of each content word in the question and those in the answer. The **mean, minimum, and maximum JSD values** were used as features. Additionally, composite vectors were formed which represented the entire question and the entire answer and the **overall JSD** between these two vectors was also included as a feature. See Fried et. al (In press) for additional details.
RNNLM	Cosine Similarity	Similar to Jansen et al. (2014), we include as features the **maximum and average pairwise cosine similarity** between question and answer words, as well as the **overall similarity** between the composite question and answer vectors.

Table 1: Feature descriptions for alignment models and RNNLM baseline.

hoo! Answers[9] community question answering corpus and divided: 50% for training, 25% for development, and 25% for test. Candidate answers for a given question are selected from the corresponding answers proposed by the community (each question has an average of 9 answers).

Biology QA (Bio): 183 *how* and 193 *why* questions in the cellular biology domain were hand-crafted by a domain expert, and paired with gold answers in the Campbell's Biology textbook (Reece et al., 2011). Each paragraph in the textbook was considered as a candidate answer. As there were few questions, five fold cross-validation was used with three folds for training, one for development, and one for test.

Alignment Corpora: To train the alignment models we generated alignment pairs from two different resources: Annotated Gigaword (Napoles et al., 2012) for YA, and the textbook for Bio. Each was discourse parsed with the RST discourse parser described in Section 3, which is implemented in the FastNLPProcessor toolkit[10], using the MaltParser[11] for syntactic analysis.

5.1 Results and Discussion

Figure 2 shows the performance of the discourse models against the number of documents used to train the alignment model.[12] We used the standard implementation for P@1 (Manning et al., 2008) with the adaptations for Bio described in Jansen et al. (2014). We address the following questions.

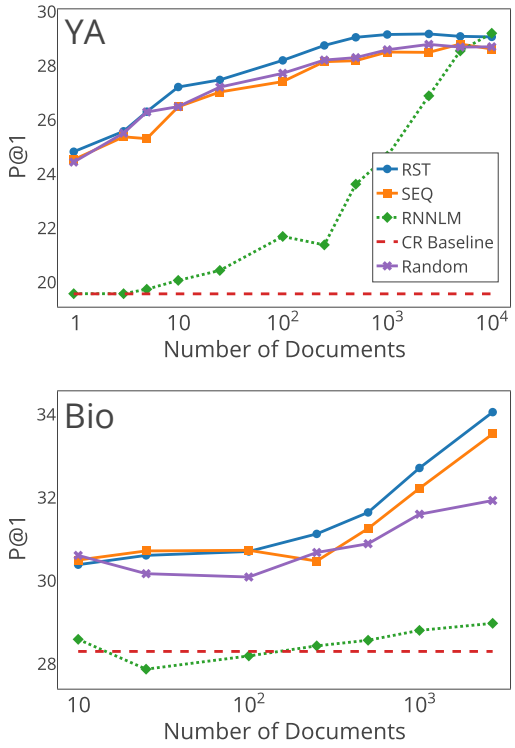

Figure 2: Overall performance for the two discourse-based alignment models, compared against the CR baseline, random baselines, and a RNNLM-based reranker. The x axis indicates the number of training documents used to construct all models. Each point represents the average of 10 samples of training documents.

How does the performance of the RST and SEQ models compare? Comparing the two principal alignment models, the RST-based model significantly outperforms the SEQ model by about 0.5% P@1 in both domains ($p < 0.001$ for Bio and $p < 0.01$ for YA)[13]. This shows that deep discourse anal-

[9] http://answers.yahoo.com

[10] http://github.com/sistanlp/processors

[11] http://www.maltparser.org/

[12] For space reasons the graph for Bio *how* is not shown, but the pattern is essentially identical to Bio *why*.

[13] All reported statistics were performed at the endpoints, i.e., when all training data is used, using bootstrap resampling with

ysis (as imperfect as it is today) is beneficial.

How does the performance of the RST model compare to a model trained on in-domain pairs? Both the RST and SEQ results for YA are *higher* than that of an alignment model trained on explicit in-domain question-answer pairs. Fried et. al (In press) trained an identical alignment model using approximately 65k QA pairs from the YA corpus, and report a performance of 27.24% P@1, or nearly 2 points lower than our model trained using 10,000 Gigaword documents. This is an encouraging result, which further demonstrates that: (a) discourse analysis can be exploited to generate artificial semi-structured data for alignment, and (b) the sequential model, which also outperforms Fried et. al, can be used as a reasonable proxy for discourse when a parser is not available.

How does the performance of the RST model compare to previous work? Comparing our work to Jansen et al. (2014), the most relevant prior work, we notice two trends. First, our discourse-based alignment models outperform their CR + RNNLM model, which peaks at 26.6% P@1 for YA and 31.7% for Bio *why*. While some of this difference can be assigned to implementation differences (e.g., we consider only content words for both alignment and RNNLM, where they used all words), this result again emphasizes the value of our approach. Second, the partially lexicalized discourse structures used by Jansen et. al to identify explanatory text in candidate answers perform better than our approach, which relies solely on lexicalized alignment. However, we expect that our two approaches are complementary, because they address different aspects of the QA task (structure vs. similarity).

How do the RST and SEQ models compare to the non-alignment baselines? In Bio, both the RST and SEQ alignment models significantly outperform the RNNLM and CR baselines ($p < 0.001$). In YA, the RST and SEQ models significantly outperform the CR baseline ($p < 0.001$), and though they considerably outperform the the RNNLM baseline for most training document sizes, when all 10,000 documents are used for training, they do not perform better. This shows that alignment models are more robust to little training data, but RNNLMs catch up when considerable data is available.

How does the SEQ model compare to the RND baseline? In Bio, the SEQ model significantly outperforms the RND baseline ($p < 0.001$) but in YA it does not. This is likely due to differences in the size of the document which was randomized. In YA, the sentences were randomized within Gigaword articles, which are relatively short (averaging 19 sentences), whereas in Bio the randomization was done at the textbook level. In practice, as document size decreases, the RND model approaches the SEQ model.

Why does performance plateau in YA and not in Bio? With Bio, we exploit all of the limited in-domain training data, and continue to see performance improvements. With YA, however, performance asymptotes for the alignment models when trained beyond 10,000 documents, or less than 1% of the Gigaword corpus. Similarly, when trained over the entirety of Gigaword (two orders of magnitude more data), our RNNLM improves only slightly, peaking at approximately 30.5% P@1 (or, a little over 1% P@1 higher). We hypothesize that this limitation comes from failing to take context into account. In open domains, alignments such as *apple – orchard* may interfere with those from different contexts, e.g., *apple – computer*, and add noise to the answer selection process.

6 Conclusion

We propose two inexpensive methods for training alignment models using solely free text, by generating artificial question-answer pairs from discourse structures. Our experiments indicate that these methods are a viable solution for constructing state-of-the-art QA systems for low-resource domains, or languages where training data is expensive and/or limited. Since alignment models have shown utility in other tasks (e.g. textual entailment), we hypothesize that these methods for creating inexpensive and highly specialized training data could be useful for tasks other than QA.

Acknowledgments

We thank the Allen Institute for AI for funding this work.

10,000 iterations.

References

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.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Abdessamad Echihabi and Daniel Marcu. 2003. A noisy-channel approach to question answering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 16–23. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the Association for Computational Linguistics*.

Daniel Fried, Peter Jansen, Gustave Hahn-Powell, Mihai Surdeanu, and Peter Clark. In press. Higher-order lexical semantic models for non-factoid answer reranking. *Transactions of the Association for Computational Linguistics*.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in english*. Routledge.

H. Hernault, H. Prendinger, D. duVerle, and M. Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.

Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing textual entailment with lccs groundhog system. In *Proceedings of the Second PASCAL Challenges Workshop*.

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)*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pages 95–100, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

J.B. Reece, L.A. Urry, M.L. Cain, S.A. Wasserman, and P.V. Minorsky. 2011. *Campbell Biology*. Pearson Benjamin Cummings.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 464–471, Prague, Czech Republic.

Radu Soricut and Eric Brill. 2006. Automatic question answering using the web: Beyond the factoid. *Journal of Information Retrieval - Special Issue on Web Information Retrieval*, 9(2):191–206.

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*.

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

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383.

Mihai Surdeanu, Thomas Hicks, and Marco A. Valenzuela-Escárcega. 2015. Two practical rhetorical structure theory parsers. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL): Software Demonstrations*.

Susan Verberne, Lou Boves, Nelleke Oostdijk, Peter-Arno Coppen, et al. 2007. Discourse-based answering of why-questions. *Traitement Automatique des Langues, Discours et document: traitements automatiques*, 47(2):21–41.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Semi-markov phrase-based monolingual alignment. In *Proceedings of EMNLP*.

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)*.