Discourse Complements Lexical Semantics for Non-factoid Answer Reranking

Peter Jansen and **Mihai Surdeanu**
University of Arizona
Tucson, AZ, USA
{pajansen,msurdeanu}
@email.arizona.edu

Peter Clark
Allen Institute for Artificial Intelligence
Seattle, WA, USA
peterc@allenai.org

Abstract

We propose a robust answer reranking model for non-factoid questions that integrates lexical semantics with discourse information, driven by two representations of discourse: a shallow representation centered around discourse markers, 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 experimentally demonstrate that the discourse structure of non-factoid answers provides information that is complementary to lexical semantic similarity between question and answer, improving performance up to 24% (relative) over a state-of-the-art model that exploits lexical semantic similarity alone. We further demonstrate excellent domain transfer of discourse information, suggesting these discourse features have general utility to non-factoid question answering.

1 Introduction

Driven by several international evaluations and workshops such as the Text REtrieval Conference (TREC)[1] and the Cross Language Evaluation Forum (CLEF),[2] the task of question answering (QA) has received considerable attention. However, most of this effort has focused on factoid questions rather than more complex non-factoid (NF) questions, such as manner, reason, or causation questions. Moreover, the vast majority of QA models explore only local linguistic structures, such as syntactic dependencies or semantic role frames, which are generally restricted to individual sentences. This is problematic for NF QA, where questions are answered not by atomic facts, but by larger cross-sentence conceptual structures that convey the desired answers. Thus, to answer NF questions, one needs a model of what these answer structures look like.

Driven by this observation, our main hypothesis is that the discourse structure of NF answers provides complementary information to state-of-the-art QA models that measure the similarity (either lexical and/or semantic) between question and answer. We propose a novel answer reranking (AR) model that combines lexical semantics (LS) with discourse information, driven by two representations of discourse: a shallow representation centered around discourse markers and surface text information, and a deep one based on the Rhetorical Structure Theory (RST) discourse framework (Mann and Thompson, 1988). To the best of our knowledge, this work is the first to systematically explore within- and cross-sentence structured discourse features for NF AR. The contributions of this work are:

1. We demonstrate that modeling discourse is greatly beneficial for NF AR for two types of NF questions, manner (*"how"*) and reason (*"why"*), across two large datasets from different genres and domains – one from the community question-answering (CQA) site of Yahoo! Answers[3], and one from a biology textbook. Our results show statistically significant improvements of up to 24% on top of state-of-the-art LS models (Yih et al., 2013).
2. We demonstrate that both shallow and deep discourse representations are useful, and, in general, their combination performs best.
3. We show that discourse-based QA models using inter-sentence features considerably out-

[1] http://trec.nist.gov
[2] http://www.clef-initiative.eu
[3] http://answers.yahoo.com

perform single-sentence models when answers span multiple sentences.

4. We demonstrate good domain transfer performance between these corpora, suggesting that answer discourse structures are largely independent of domain, and thus broadly applicable to NF QA.

2 Related Work

The body of work on factoid QA is too broad to be discussed here (see, e.g., the TREC workshops for an overview). However, in the context of LS, Yih et al. (2013) recently addressed the problem of answer sentence selection and demonstrated that LS models, including recurrent neural network language models (RNNLM), have a higher contribution to overall performance than exploiting syntactic analysis. We extend this work by showing that discourse models coupled with LS achieve the best performance for NF AR.

The related work on NF QA is considerably more scarce, but several trends are clear. First, most NF QA approaches tend to use multiple similarity models (information retrieval or alignment) as features in discriminative rerankers (Riezler et al., 2007; Higashinaka and Isozaki, 2008; Verberne et al., 2010; Surdeanu et al., 2011). Second, and more relevant to this work, all these approaches focus either on bag-of-word representations or linguistic structures that are restricted to single sentences (e.g., syntactic dependencies, semantic roles, or standalone discourse cue phrases).

Answering *how* questions using a single discourse marker, *by*, was previously explored by Prager et al. (2000), who searched for *by* followed by a present participle (e.g. *by *ing*) to elevate answer candidates in a ranking framework. Verberne et al. (2011) extracted 47 cue phrases such as *because* from a small collection of web documents, and used the cosine similarity between an answer candidate and a bag of words containing these cue phrases as a single feature in their reranking model for non-factoid *why* QA. Extending this, Oh et al. (2013) built a classifier to identify causal relations using a small set of cue phrases (e.g., *because* and *is caused by*). This classifier was then used to extract instances of causal relations in answer candidates, which were turned into features in a reranking model for Japanese *why* QA.

In terms of discourse parsing, Verberne et al. (2007) conducted an initial evaluation of the utility of RST structures to *why* QA by evaluating

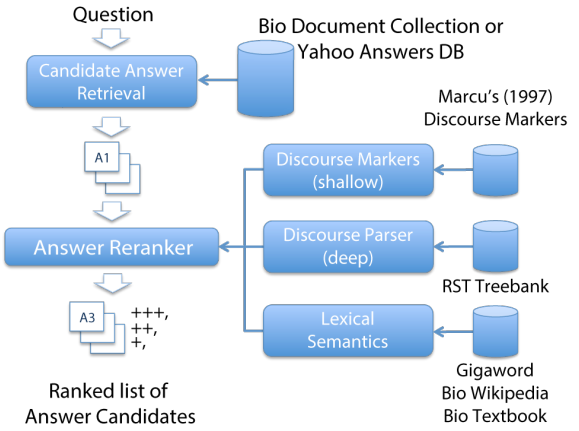

Figure 1: Architecture of the reranking framework for QA.

performance on a small sample of seven WSJ articles drawn from the RST Treebank (Carlson et al., 2003). They later concluded that while discourse parsing appears to be useful for QA, automated discourse parsing tools are required before this approach can be tested at scale (Verberne et al., 2010). Inspired by this previous work and recent work in discourse parsing (Feng and Hirst, 2012), our work is the first to systematically explore structured discourse features driven by several discourse representations, combine discourse with lexical semantic models, and evaluate these representations on thousands of questions using both in-domain and cross-domain experiments.

3 Approach

The proposed answer reranking component is embedded in the QA framework illustrated in Figure 1. This framework functions in two distinct scenarios, which use the same AR model, but differ in the way candidate answers are retrieved:

CQA: In this scenario, the task is defined as reranking all the user-posted answers for a particular question to boost the community-selected best answer to the top position. This is a commonly used setup in the CQA community (Wang et al., 2009).[4] Thus, for a given question, all its answers are fetched from the answer collection, and an initial ranking is constructed based on the cosine similarity between theirs and the question's lemma vector representations, with lemmas weighted using *tf.idf* (Ch. 6, (Manning et al., 2008)).

[4] Although most of these works use shallow textual features and focus mostly on meta data, e.g., number of votes for a particular answer. Here we use no meta data and rely solely on linguistic features.

Traditional QA: In this scenario answers are dynamically constructed from larger documents (Pasca, 2001). We use this setup to answer questions from a biology textbook, where each section is indexed as a standalone document, and each paragraph in a given document is considered as a candidate answer. We implemented the document indexing and retrieval stage using Lucene[5]. The candidate answers are scored using a linear interpolation of two cosine similarity scores: one between the entire parent document and question (to model global context), and a second between the answer candidate and question (for local context).[6] Because the number of answer candidates is typically large (e.g., equal to the number of paragraphs in the textbook), we return the N top candidates with the highest scores.

These answer candidates are then passed to the answer reranking component, the focus of this work. AR analyzes the candidates using more expensive techniques to extract discourse and LS features (detailed in §4), and these features are then used in concert with a learning framework to rerank the candidates and elevate correct answers to higher positions. For the learning framework, we used SVM^{rank}, a variant of Support Vector Machines for structured output adapted to ranking problems.[7] In addition to these features, each reranker also includes a single feature containing the score of each candidate, as computed by the above candidate retrieval (CR) component.[8]

4 Models and Features

We propose two separate discourse representation schemes – one shallow, centered around discourse markers, and one deep, based on RST.

4.1 Discourse Marker Model

The discourse marker model (DMM) extracts cross-sentence discourse structures centered around a discourse marker. This extraction process is illustrated in the top part of Figure 2. These structures are represented using three components: (1) A **discourse marker** from Daniel Marcu's list (see Appendix B in Marcu (1997)), that serves as a divisive boundary between sentences. Examples of these markers include *and, in, that, for, if, as, not, by,* and *but*; (2) **two marker arguments**, i.e., text segments before and after the marker, labeled to indicate if they are related to the question text or not; and (3) **a sentence range** around the marker, which defines the length of these segments (e.g., ± 2 sentences). For example, a marker feature may take the form of: `QSEG BY OTHER SR2`, which means that the the marker *by* has been detected in an answer candidate. Further, the text preceeding *by* matches text from the question (and is therefore labeled `QSEG`), while the text after *by* differs considerably from the question text, and is labeled `OTHER`. In this particular example, the scope of this similarity matching occurs over a span of ± 2 sentences around the marker.

Note that our marker arguments are akin to EDUs in RST, but, in this shallow representation, they are simply constructed around discourse markers and bound by an arbitrary sentence range.

Argument Labels: We label marker arguments based on their similarity to question content. If text before or after a marker out to a given sentence range matches the entire text of the question (with a cosine similarity score larger than a threshold), that argument takes on the label `QSEG`, or `OTHER` otherwise. In this way the features are only partially lexicalized with the discourse markers. Argument labels indicate only if lemmas from the question were found in a discourse structure present in an answer candidate, and do not speak to the specific lemmas that were found. We show in §5 that these lightly lexicalized features perform well in domain and transfer between domains. We explore other argument labeling strategies in §5.7.

Feature Values: Our reranking framework uses real-valued features. The values of the discourse features are the mean of the similarity scores (e.g., cosine similarity using *tf.idf* weighting) of the two marker arguments and the corresponding question. For example, the value of the `QSEG BY QSEG SR1` feature in Figure 2 is the average of the cosine similarities of the question text with the answer texts before/after *by* out to a distance of one sentence before/after the marker.

It is important to note that these discourse features are more expressive than features based on discourse markers alone (Higashinaka and Isozaki, 2008; Verberne et al., 2010). First,

[5] `http://lucene.apache.org`

[6] We empirically observed that this combination of scores performs better than using solely the cosine similarity between the answer and question.

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

[8] Including these scores as features in the reranker model is a common strategy that ensures that the reranker takes advantage of the analysis already performed by the CR model.

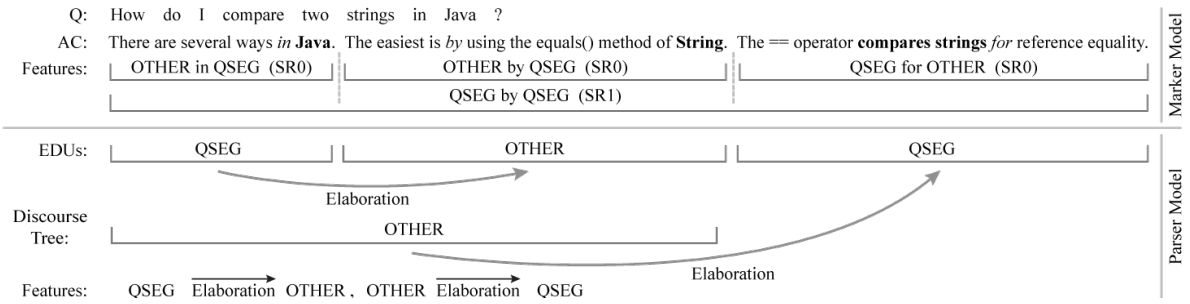

Figure 2: **Top:** Example feature generation for the discourse marker model, for one question (Q) and one answer candidate (AC). Answer candidates are searched for discourse markers (*italic*) and question word matches (**bold**), which are used to generate features both within-sentence (SR0), and ± 1 sentence (SR1). The actual DMM exhaustively generates features for all markers and all sentence ranges. Here we show just a few for brevity. **Bottom:** Example feature generation for the discourse parser model using the output of an actual discourse parser. The DPM creates one feature for each individual discourse relation.

the argument sequences used here capture cross-sentence discourse structures. Second, these features model the intensity of the match between the text surrounding the discourse structure and the question text using both the assigned argument labels and the feature values.

4.2 Discourse Parser Model

The discourse parser model (DPM) is based on the RST discourse framework (Mann and Thompson, 1988). In RST, the text is segmented into a sequence of non-overlapping fragments called elementary discourse units (EDUs), and binary discourse relations recursively connect neighboring units. Most relations are *hypotactic*, where one of the units in the relation (the *nucleus*) is considered more important than the other (the *satellite*). A few relations are *paratactic*, where both participants have equal importance. In the bottom part of Figure 2, we show hypotactic relations as directed arrows, from the nucleus to the satellite. In this work, we construct the RST discourse trees using the parser of Feng and Hirst (2012).

Relying on a proper discourse framework facilitates the modeling of the numerous implicit relations that are not driven by discourse markers (see Ch. 21 in Jurafsky and Martin (2009)). However, this also introduces noise because discourse analysis is a complex task and discourse parsers are not perfect. To mitigate this, we used a simple feature generation strategy, which creates one feature for each individual discourse relation by concatenating the relation type with the labels of the discourse units participating in it. To this end, for every relation, we extract the entire text dominated by each of its arguments, and we generate labels for the two participants in the relation using the same strategy as the DMM (based on the similarity with the question content). Similar to the DMM, these features take real values obtained by averaging the cosine similarity of the arguments with the question content.[9] Fig. 2 shows several such features, created around two RST `Elaboration` relations, indicating that the latter sentences expand on the information at the beginning of the answer. Other common relations include `Attribution`, `Contrast`, `Background`, and `Evaluation`.

4.3 Lexical Semantics Model

Inspired by the work of Yih et al. (2013), we include lexical semantics in our reranking model. Several of their proposed models rely on proprietary data; here we focus on LS models that rely on open-source data and frameworks. In particular, we use the recurrent neural network language model (RNNLM) of Mikolov et al. (2013; 2010). Like any language model, a RNNLM estimates the probability of observing a word given the preceding context, but, in this process, it learns word embeddings into a latent, conceptual space with a fixed number of dimensions. Consequently, related words tend to have vectors that are close to each other in this space.

We derive two LS measures from these vectors, which are then are included as features in the reranker. The first is a measure of the overall LS similarity of the question and answer can-

[9] We investigated more complex features, e.g., by exploring depths of two and three in the discourse tree, and also models that relied on tree kernels over these trees, but none improved upon this simple representation. This suggests that, in the domains explored here, there is a degree of noise introduced by the discourse parser, and the simple features proposed here are the best strategy to avoid overfitting on it.

didate, which is computed as the cosine similarity between the two composite vectors of the question and the answer candidate. These composite vectors are assembled by summing the vectors for individual question (or answer candidate) words, and re-normalizing this composite vector to unit length. Both this overall similarity score, as well as the average pairwise cosine similarity between each word in the question and answer candidate, serve as features.

5 Experiments

5.1 Data

To test the utility of our approach, we experimented with the two QA scenarios introduced in §3 using the following two datasets:

Yahoo! Answers Corpus (YA): Yahoo! Answers[10] is an open domain community-generated QA site, with questions and answers that span formal and precise to informal and ambiguous language. Due to the speed limitations of the discourse parser, we randomly drew 10,000 QA pairs from the corpus of *how* questions described by Surdeanu et al. (2011) using their filtering criteria, with the additional criterion that answers had to contain at least four community-generated answers, one of which was voted as the top answer. The number of answers to each question ranged from 4 to over 50, with the average 9.[11]

Biology Textbook Corpus (Bio): This corpus focuses on the domain of cellular biology, and consists of 185 *how* and 193 *why* questions handcrafted by a domain expert. Each question has one or more gold answers identified in Campbell's Biology (Reece et al., 2011), a popular undergraduate text. The entire biology text (at paragraph granularity) serves as the possible set of answers. Note that while our system retrieves answers at paragraph granularity, the expert was not constrained in any way during the annotation process, so gold answers might be smaller than a paragraph or span multiple paragraphs. This complicates evaluation metrics on this dataset (see §5.3).

[10] `http://answers.yahoo.com`

[11] Note that our experimental setup, i.e., reranking all the answers provided for each question, is different from that of Surdeanu et al. For each question, they retrieved candidate answers from all answers voted as best for some question in the collection. The setup in this paper, commonly used in the CQA community (Wang et al., 2009), is more relevant here because it includes both high and low quality answers.

For the YA CQA corpora, 50% of QA pairs were used for training, 25% for development, and 25% for test. Because of the small size of the Bio corpus, it was evaluated using 5-fold cross-validation, with three folds for training, one for development, and one for test.

The following additional resources were used:

Discourse Markers: A set of 75 high-frequency[12] single-word discourse markers were extracted from Marcu's (1997) list of cue phrases, and used for feature generation in DMM. These discourse markers are extremely common in the answer corpora – for example, the YA corpus contains an average of 7 markers per answer.

Discourse Trees: We generated all discourse trees using the parser of Feng and Hirst (2012). For YA, we parsed entire answers. For Bio, we parsed individual paragraphs. Note that, because these domains are considerably different from the RST Treebank, the parser fails to produce a tree on a large number of answer candidates: 6.2% for YA, and 41.1% for Bio. In these situations, we constructed artificial discourse trees using a right-attachment heuristic and a single relation label `X`.

Lexical Semantics: We trained two different RNNLMs for this work. First, for the YA experiments we trained an open-domain RNNLM using the entire Gigaword corpus of approximately 4G words.[13] For the Bio experiments, we trained a domain specific RNNLM over a concatenation of the textbook and a subset of Wikipedia specific to biology. The latter was created by extracting: (a) pages matching a word/phrase in a glossary of biology (derived from the textbook); plus (b) pages hyperlinked from (a) that are also tagged as being in a small set of (hand-selected) biology-related categories. The combined dataset contains 7.7M words. For all RNNLMs we used 200-dimensional vectors.

5.2 Hyper Parameter Tuning

The following hyper parameters were tuned using grid search to maximize P@1 on each development partition: (a) the segment matching thresholds that determine the minimum cosine similarity between an answer segment and a question for the segment to be labeled `QSEG`; and (b)

[12] We selected all cue phrases with more than 100 occurrences in the Brown corpus.

[13] LDC catalog number LDC2012T21

#	Model/Features	P@1	P@1 Impr.	MRR	MRR Impr.
	YA Corpus				
1	Random Baseline	14.29		26.12	
2	CR Baseline	19.57		43.14	
3	CR + DMM	24.05*	+23%	46.40*	+8%
4	CR + DPM	24.29*	+24%	46.81*	+9%
5	CR + DMM + DPM	**24.81***	**+27%**	**47.10***	**+9%**
6	CR + LS Baseline	26.57		49.31	
7	CR + LS + DMM	29.29*	+10%	50.99*	+3%
8	CR + LS + DPM	28.73*	+8%	50.77*	+3%
9	CR + LS + DMM + DPM	**30.49***	**+15%**	**51.89***	**+5%**
	Bio HOW				
10	CR Baseline	24.12		32.90	
11	CR + DMM	29.88*	+24%	38.88*	+18%
12	CR + DPM	28.93*	+20%	37.75*	+15%
13	CR + DMM + DPM	**30.43***	**+26%**	**39.28***	**+19%**
14	CR + LS Baseline	25.35		33.79	
15	CR + LS + DMM	30.09*	+19%	39.04*	+16%
16	CR + LS + DPM	28.50	+12%	37.58*	+11%
17	CR + LS + DMM + DPM	**30.68***	**+21%**	**39.44***	**+17%**
	Bio WHY				
18	CR Baseline	28.62		38.25	
19	CR + DMM	38.01*	+33%	46.39*	+21%
20	CR + DPM	38.62*	+35%	46.85*	+23%
21	CR + DMM + DPM	**39.36***	**+38%**	**47.64***	**+25%**
22	CR + LS Baseline	31.73		39.89	
23	CR + LS + DMM	38.60*	+22%	46.41*	+16%
24	CR + LS + DPM	**39.45***	**+24%**	47.38*	+19%
25	CR + LS + DMM + DPM	39.32*	+24%	**47.86***	**+20%**

Table 1: Overall results across three datasets. The improvements in each section are computed relative to their respective baseline (CR or CR + LS). Bold font indicates the best score in a given column. * indicates that a score is significantly better ($p < 0.05$) than the score of the corresponding baseline. All significance tests were implemented using one-tailed non-parametric bootstrap resampling using 10,000 iterations.

SVM^{rank}'s regularization parameter C. For all experiments, the sentence range parameter (SRx) for DMM ranged from 0 (within sentence) to ± 3 sentences.[14]

5.3 Evaluation Metrics

For YA, we used the standard implementations for P@1 and mean reciprocal rank (MRR) (Manning et al., 2008). In the Bio corpus, because answer candidates are not guaranteed to match gold annotations exactly, these metrics do not immediately apply. We adapted them to this dataset by weighing each answer by its overlap with gold answers, where overlap is measured as the highest F1 score between the candidate and a gold answer. Thus, P@1 reduces to this F1 score for the top answer. For MRR, we used the rank of the candidate with the highest overlap score, weighed by the inverse of the rank. For example, if the best answer for a question appears at rank 2 with an F1 score of 0.3, the corresponding MRR score is $0.3/2$.

[14]This was only limited to reduce the combinatorial expansion of feature generation, and in principle could be set much broader.

5.4 Overall Results

Table 1 analyzes the performance of the proposed reranking model on the three datasets and against two baselines. The first baseline sorts the candidate answers in descending order of the scores produced by the candidate retrieval (CR) module. The second baseline (CR + LS) trains a reranking model without discourse, using just the CR and LS features. For YA, we include an additional baseline that selects an answer randomly. We list multiple versions of the proposed reranking model, broken down by the features used. For Bio, we retrieved the top 20 answer candidates in CR. At this setting, the oracle performance (i.e., the performance with perfect reranking of the 20 candidates) was 69.6% P@1 for Bio HOW, and 72.3% P@1 for Bio WHY. These relatively low oracle scores, which serve as a performance ceiling for our approach, highlight the difficulty of the task. For YA, we used all answers provided for each given question. For all experiments we used a linear SVM kernel.[15]

Examining Table 1, several trends are clear. Both discourse models significantly increase both P@1 and MRR performance over all baselines broadly across genre, domain, and question types. More specifically, DMM and DPM show similar performance benefits when used individually, but their combination generally outperforms the individual models, illustrating the fact that the two models capture related but different discourse information. This is a motivating result for discourse analysis, especially considering that the discourse parser was trained on a domain different from the corpora used here.

Lexical semantic features increase performance for all settings, but demonstrate far more utility to the open-domain YA corpus. This disparity is likely due to the difficulty in assembling LS training data at an appropriate level for the biology corpus, contrasted with the relative abundance of large scale open-domain lexical semantic resources. For the YA corpus, where lexical semantics showed the most benefit, simply adding

[15]The performance of all models can ultimately be increased by using more sophisticated learning frameworks, and considering more answer candidates in CR (for Bio). For example, SVMs with polynomial kernels of degree two showed approximately half a percent (absolute) performance gain over the linear kernel. However, this came at the expense of an experiment runtime about an order of magnitude larger. Experiments with more answer candidates in Bio showed similar trends to the results reported.

Q	How does myelination affect action potentials?
$A_{baseline}$	The major selective advantage of myelination is its space efficiency. A myelinated axon 20 microns in diameter has a conduction speed faster than that of a squid giant axon [...]. Furthermore, more than 2,000 of those myelinated axons can be packed into the space occupied by just one giant axon.
A_{rerank}	A nerve impulse travels [...] to the synaptic terminals by propagation of a series action potentials along the axon. The speed of conduction increases [...] with myelination. Action potentials in myelinated axons jump between the nodes of Ranvier, a process called saltatory conduction.

Table 2: An example question from the Biology corpus where the correct answer is elevated to the top position by the discourse model. $A_{baseline}$ is the top answer proposed by the CR + LS baseline, which is incorrect, whereas A_{rerank} is the correct answer boosted to the top after reranking. [...] indicates non-essential text that was removed for space.

Range	Bio HOW	Bio WHY	YA
CR + LS + DMM + DPM			
within-sentence	+0.8%	+8.4%	+13.1%
full model	+21.0%*	+23.9%*	+14.8%

Table 3: Relative P@1 performance increase over the CR + LS baseline for a model containing only intra-sentence features, compared to the full model.

LS features to the CR baseline increases baseline P@1 performance from 19.57 to 26.57, a +36% relative improvement. Most importantly, comparing lines 5 and 9 with their respective baselines (lines 2 and 6, respectively) indicates that LS is largely orthogonal to discourse. Line 5, the top-performing model with discourse but without LS outperforms the CR baseline by +5.24 absolute P@1 improvement. Similarly, line 9, the top-performing model that combines discourse with LS has a +5.69 absolute P@1 improvement over the CR + LS baseline. That this absolute performance increase is nearly identical indicates that LS features are complementary to and additive with the full discourse model. Indeed, an analysis of the questions improved by discourse vs. LS (line 5 vs. 6) showed that the intersection of the two sets is low (approximately a third of each set).

Finally, while the discourse models perform well for HOW or *manner* questions, performance on Bio WHY corpus suggests that *reason* questions are particularly amenable to discourse analysis. Relative improvements on WHY questions reach +38% (without LS) and +24% (with LS), with absolute performance on these non-factoid questions jumping from 28% to nearly 40% P@1.

Table 2 shows one example where discourse helps boost the correct answer to the top position. In this example, the correct answer contains multiple `Elaboration` relations that are both cross sentence (e.g., between the first two sentences) and intra-sentence (e.g., between the first part of the second sentence and the phrase "with myelination"). Model features associated with `Elaboration` relations are ranked highly by the learned model. In contrast, the answer preferred by the baseline contains mostly `Joint` relations, which "represent the lack of a rhetorical relation between the two nuclei" (Mann and Thompson, 1988) and have very small weights in the model.

5.5 Intra vs. Inter-sentence Features

To tease apart the relative contribution of discourse features that occur only within a single sentence versus features that span multiple sentences, we examined the performance of the full model when using only intra-sentence features, i.e., `SR0` features for DMM, and features based on discourse relations where both EDUs appear in the same sentence for DPM, versus the full intersentence models. The results are shown in Table 3.

For the Bio corpus where answer candidates consist of entire paragraphs of a biology text, overall performance is dominated by inter-sentence discourse features. Conversely, for YA, a large proportion of performance comes from features that span only a single sentence. This is caused by the fact that YA answers are far shorter and of variable grammatical quality, with 39% of answer candidates consisting of only a single sentence, and 57% containing two or fewer sentences. All in all, this experiment emphasizes that modeling both intra- and inter-sentence discourse (where available) is beneficial for non-factoid QA.

5.6 Domain Transfer

Because these discourse models appear to capture high-level information about answer structures, we hypothesize that the models should make use of many of the same discourse features, even when training on data from different domains. Table 4 shows that of the highest-weighted SVM features learned when training models for HOW questions on YA and Bio, many are shared (e.g., 56.5% of the features in the top half of both DPMs are shared), suggesting that a core set of discourse features may be of utility across domains.

To test the generality of these features, we performed a transfer study where the full model was trained and tuned on the open-domain YA corpus, then evaluated as is on Bio HOW. This is

Model	Top 10%	Top 25%	Top 50%
DMM	20.2%	33.2%	49.4%
DPM	22.2%	39.1%	56.5%

Table 4: Percentage of top features with the highest SVM weights that are shared between Bio HOW and YA models.

a somewhat radical setup, where the target corpus has both a different genre (formal text vs. CQA) and different domain (biology vs. open domain). These experiments were performed in several groups: both with and without LS features, as well as using either a single SVM or an ensemble model that linearly interpolates the predictions of two SVM classifiers (one each for DMM and DPM).[16] The results are summarized in Table 5.

The transferred models always outperform the baselines, but only the ensemble model's improvement is statistically significant. This confirms existing evidence that ensemble models perform better cross-domain because they overfit less (Domingos, 2012; Hastie et al., 2009). The ensemble model without LS (third line) has a nearly identical P@1 score as the equivalent in-domain model (line 13 in Table 1), while slightly surpassing in-domain MRR performance. To the best of our knowledge, this is one of the most striking demonstrations of domain transfer in answer ranking for non-factoid QA, and highlights the generality of these discourse features in identifying answer structures across domains and genres.

The results of the transferred models that include LS features are slightly lower, but still approach statistical significance for P@1 and are significant for MRR. We hypothesize that the limited transfer observed for models with LS compared to their counterparts without LS is due to the disparity in the size and utility of the biology LS training data compared to the open-domain LS resources. The open-domain YA model learns to place more weight on LS features, which are unable to provide the same utility in the biology domain.

5.7 Integrating Discourse and LS

So far, we have treated LS and discourse as distinct features in the reranking model, However, given that LS features greatly improve the CR baseline, we hypothesize that a natural extension

[16]The interpolation parameter was tuned on the YA development corpus. The in-domain performance of the ensemble model is similar to that of the single classifier in both YA and Bio HOW so we omit these results here for simplicity.

Model/Features	P@1	P@1 Impr.	MRR	MRR Impr.
Transfer: YA → Bio HOW				
CR Baseline	24.12		32.90	
CR + DMM + DPM	27.13	+13%	36.36†	+11%
(CR + DMM) ∪ (CR + DPM)	**30.10***	**+25%**	**39.62***	**+20%**
CR + LS Baseline	25.35		33.79	
CR + LS + DMM + DPM	25.79	+2%	35.58	+5%
(CR + LS + DMM) ∪ (CR + LS + DPM)	**29.54**†	**+17%**	**38.68***	**+15%**

Table 5: Transfer performance from YA to Bio HOW for single classifiers and ensembles (denoted with a ∪). † indicates approaching statistical significance with $p = 0.07$ or 0.06.

to the discourse models would be to make use of LS similarity (in addition to the traditional information retrieval similarity) to label discourse segments. For example, for the question *"How do cells replicate?"*, answer discourse segments containing LS associates of *cell* and *replicate*, e.g., *nucleus, membrane, genetic*, and *duplicate*, should be considered as related to the question (i.e., be labeled QSEG). We implemented two such models, denoted DMM$_{LS}$ and DPM$_{LS}$, by replacing the component that assigns argument labels with one that relies on LS. Specifically, as in §4.3, we compute the cosine similarity between the composite LS vectors of the question text and each marker argument (in DMM) or EDU (in DPM), and label the corresponding answer segment QSEG if this score is higher than a threshold, or OTHER otherwise. This way, the DMM and DPM features jointly capture discourse structures and semantic similarity between answer segments and question.

To test this, we use the YA corpus, which has the best-performing LS model. Because we are adding two new discourse models, we now tune four segment matching thresholds, one for each of the DMM, DPM, DMM$_{LS}$, and DPM$_{LS}$ models.[17] The results are shown in Table 6. These results demonstrate that incorporating LS in the discourse models further increases performance for all configurations, nearly doubling the relative performance benefits over models that do not integrate LS and discourse (compare with lines 6–9 of Table 1). For example, the last model in the table, which combines four discourse representations, improves P@1 by 24%, whereas the equivalent model without this integration (line 9 in Table 1) outperforms the baseline by only 15%.

[17]These hyperparameters were tuned on the development corpus, and were found to be stable over broad ranges.

Model Features	P@1	P@1 Impr.	MRR	MRR Impr.
CR + LS Baseline	26.57		49.31	
CR + LS + DMM + DMM$_{LS}$	32.41*	+22%	53.55*	+9%
CR + LS + DPM + DPM$_{LS}$	31.21*	+18%	52.50*	+7%
CR + LS + DMM + DPM + DMM$_{LS}$ + DPM$_{LS}$	**32.93***	**+24%**	**53.91***	**+9%**

Table 6: YA results with integrated discourse and LS.

5.8 Error Analysis

We performed an error analysis of the full QA model (CR + LS + DMM + DPM) across the entire Bio corpus (lines 17 and 25 from Table 1). We chose the Bio setup for this analysis because it is more complex than the CQA one: here gold answers may have a granularity completely different from what the system choses as best answers (in our particular case, the QA system is currently limited to answers consisting of single paragraphs, whereas gold answers may be of any size).

Here, 94 of the 378 Bio HOW and WHY questions have improved answer scores, while 36 have reduced performance relative to the CR baseline. Of these 36 questions where answer scores decreased, nearly two thirds were directly related to the paragraph granularity of the candidate answer retrieval (see §5.1):

Same Subsection (50%): In these cases, the model selected an on-topic answer paragraph in the same subsection of the textbook as a gold answer. Often times this paragraph directly preceded or followed the gold answer.

Answer Window Size (14%): Here, both the CR and full model chose a paragraph containing a different gold answer. However, as discussed, gold answers may unevenly straddle paragraph boundaries, and the paragraph chosen by the model happened to have a somewhat lower overlap with its gold answer than the one chosen by the baseline.

Similar Topic (25%): The model chose a paragraph that had a similar topic to the question, but doesn't answer the question. These are challenging errors, often associated with short questions (e.g. *"How does HIV work?"*) that provide few keywords. In these cases, discourse features tend to dominate, and shift the focus towards answers that have many discourse structures deemed relevant. For example, for the above question, the model chose a paragraph containing many discourse structures positively correlated with high-quality answers, but which describes the origins of HIV instead of how the virus enters a cell.

Similar Words, Different Topic (8%): The model chose a paragraph that had many of the same words as the question, but is on a different topic. For example, for the question *"How are fossil fuels formed, and why do they contain so much energy?"*, the model selected an answer that mentions fossil fuels in a larger discussion of human ecological footprints. Here, the matching of both keywords and discourse structures shifted the answer towards a different, incorrect topic.

Finally, in one case (3%), the model identified an answer paragraph that contained a gold answer, but was missed by the domain expert annotator.

In summary, this analysis suggests that, for the majority of errors, the QA system selects an answer that is both topical and adjacent to a gold answer selected by the domain expert. This suggests that most errors are minor and are driven by current limitations of our answer boundary selection mechanism, rather than the inherent limitations of the discourse model.

6 Conclusions

This work focuses on two important aspects of answer reranking for non-factoid QA: similarity between question and answer content, and answer structure. While the former has been addressed with a variety of lexical-semantic models, the latter has received little attention. Here we show how to model answer structures using discourse and how to integrate the two aspects into a holistic framework. Empirically we show that modeling answer discourse structures is complementary to modeling lexical semantic similarity and that the best performance is obtained when they are tightly integrated. We evaluate the proposed approach on multiple genres and question types and obtain benefits of up to 24% relative improvement over a strong baseline that combines information retrieval and lexical semantics. We further demonstrate that answer discourse structures are largely independent of domain and transfer well, even between radically different datasets.

This work is open source and available at: `http://nlp.sista.arizona.edu/releases/acl2014`.

Acknowledgements

We thank the Allen Institute for Artificial Intelligence for funding this work. We would also like to thank the three anonymous reviewers for their helpful comments and suggestions.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer Academic Publishers.

Pedro Domingos. 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10).

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer.

Ryuichiro Higashinaka and Hideki Isozaki. 2008. Corpus-based question answering for why-questions. In *Proceedings of the Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, pages 418–425, Hyderabad, India.

Dan Jurafsky and James H. Martin. 2009. *Speech and Language Processing, Second Edition*. Prentice Hall.

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.

Daniel Marcu. 1997. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. thesis, University of Toronto.

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)*.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra- and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1733–1743, Sofia, Bulgaria, August. Association for Computational Linguistics.

Marius Pasca. 2001. *High-Performance, Open-Domain Question Answering from Large Text Collections*. Ph.D. thesis, Southern Methodist University.

John Prager, Eric Brown, Anni Coden, and Dragomir Radev. 2000. Question-answering by predictive annotation. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 184–191, New York, NY, USA. ACM.

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.

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.

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.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2010. What is not in the bag of words for why-qa? *Computational Linguistics*, 36(2):229–245.

Suzan Verberne, Hans Halteren, Daphne Theijssen, Stephan Raaijmakers, and Lou Boves. 2011. Learning to rank for why-question answering. *Inf. Retr.*, 14(2):107–132, April.

Xin-Jing Wang, Xudong Tu, Dan Feng, and Lei Zhang. 2009. Ranking community answers by modeling question-answer relationships via analogical reasoning. In *Proceedings of the Annual ACM SIGIR Conference*.

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)*.